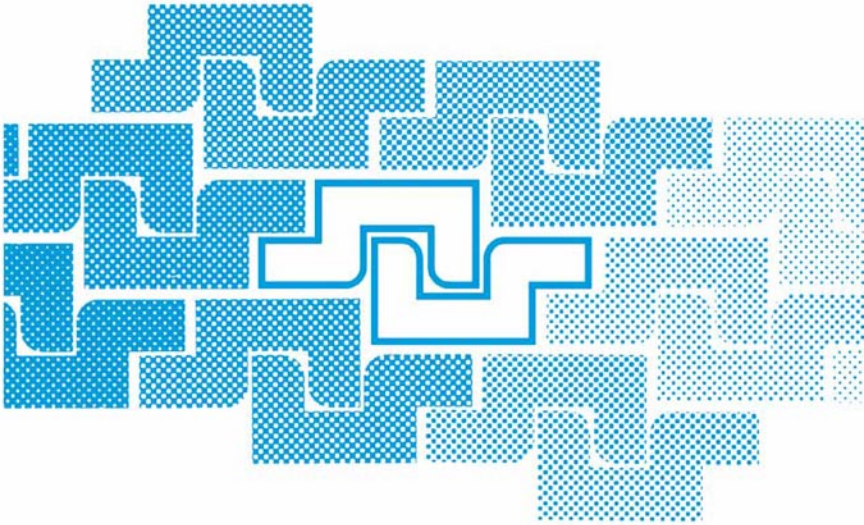


Structural Theory of Automata, Semigroups, and Universal Algebra

Edited by

Valery B. Kudryavtsev and
Ivo G. Rosenberg

NATO Science Series

II. Mathematics, Physics and Chemistry – Vol. 207

Structural Theory of Automata, Semigroups, and Universal Algebra

NATO Science Series

A Series presenting the results of scientific meetings supported under the NATO Science Programme.

The Series is published by IOS Press, Amsterdam, and Springer (formerly Kluwer Academic Publishers) in conjunction with the NATO Public Diplomacy Division.

Sub-Series

I. Life and Behavioural Sciences	IOS Press
II. Mathematics, Physics and Chemistry	Springer (formerly Kluwer Academic Publishers)
III. Computer and Systems Science	IOS Press
IV. Earth and Environmental Sciences	Springer (formerly Kluwer Academic Publishers)

The NATO Science Series continues the series of books published formerly as the NATO ASI Series.

The NATO Science Programme offers support for collaboration in civil science between scientists of countries of the Euro-Atlantic Partnership Council. The types of scientific meeting generally supported are "Advanced Study Institutes" and "Advanced Research Workshops", and the NATO Science Series collects together the results of these meetings. The meetings are co-organized by scientists from NATO countries and scientists from NATO's Partner countries — countries of the CIS and Central and Eastern Europe.

Advanced Study Institutes are high-level tutorial courses offering in-depth study of latest advances in a field.
Advanced Research Workshops are expert meetings aimed at critical assessment of a field, and identification of directions for future action.

As a consequence of the restructuring of the NATO Science Programme in 1999, the NATO Science Series was re-organized to the four sub-series noted above. Please consult the following web sites for information on previous volumes published in the Series.

http://www.nato.int/science
http://www.springeronline.com
http://www.iospress.nl

Series II: Mathematics, Physics and Chemistry – Vol. 207

Structural Theory of Automata, Semigroups, and Universal Algebra

edited by

Valery B. Kudryavtsev
Department of Mathematical Theory of Intelligent Systems,
Faculty of Mechanics and Mathematics,
M.V. Lomonosov Moscow State University,
Moscow, Russia

and

Ivo G. Rosenberg
University of Montreal,
Quebec, Canada

Technical Editor:

Martin Goldstein
Department of Mathematics and Statistics,
University of Montreal,
Quebec, Canada

Published in cooperation with NATO Public Diplomacy Division

Proceedings of the NATO Advanced Study Institute on
Structural Theory of Automata, Semigroups and Universal Algebra
Montreal, Quebec, Canada
7–18 July 2003

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN-10 1-4020-3816-X (PB)
ISBN-13 978-1-4020-3816-7 (PB)
ISBN-10 1-4020-3815-1 (HB)
ISBN-13 978-1-4020-3815-0 (HB)
ISBN-10 1-4020-3817-8 (e-book)
ISBN-13 978-1-4020-3817-4 (e-book)

Published by Springer,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

www.springeronline.com

Printed on acid-free paper

Printed in the Netherlands.

Table of Contents

Preface

In the summer of 2003 the Department of Mathematics and Statistics of the University of Montreal was fortunate to host the NATO Advanced Study Institute "Structural theory of Automata, Semigroups and Universal Algebra" as its 42nd Séminaire des mathématiques supérieures (SMS), a summer school with a long tradition and well-established reputation. This book contains the contributions of most of its invited speakers.

It may seem that the three disciplines in the title of the summer school cover too wide an area while its three parts have little in common. However, there was a high and surprising degree of coherence among the talks. Semigroups, algebras with a single associative binary operation, is probably the most mature of the three disciplines with deep results. Universal Algebra treats algebras with several operations, e.g., groups, rings, lattices and other classes of known algebras, and it has borrowed from formal logics and the results of various classes of concrete algebras. The Theory of Automata is the youngest of the three. The Structural Theory of Automata essentially studies the composition of small automata to form larger ones. The role of semigroups in automata theory has been recognized for a long time but conversly automata have also influenced semigroups. This book demonstrates the use of universal algebra concepts and techniques in the structural theory of automata as well as the reverse influences.

J. Almeida surveys the theory of profinite semigroups which grew from finite semigroups and certain problems in automata. There arises a natural algebraic structure with an interplay between topological and algebraic aspects. Pseudovarieties connect profinite semigroups to universal algebras. L. N. Shevrin surveys the very large and substantial class of special semigroups, called epigroups. He presents them as semigroups with the unary operator of pseudo-inverse and studies some nice decompositions and finiteness conditions.

A. Letichevsky studies transition systems, an extension of automata, behaviour algebras and other structures. He develops a multifaceted theory of transition systems with many aspects. J. Dassow studies various completeness results for the algebra of sequential functions on $\{0, 1\}$, essentially functions induced by automata or logical nets. In particular, he investigates completeness with respect to an equivalence relation on the algebra. V. B. Kudryavtsev surveys various completeness and expressibility problems and results starting from the completeness (primality) criterion in the propositional calculus of many-valued logics (finite algebras) to delayed algebras and automata functions. T. Hikita and I. G. Rosenberg study the week completeness of finite delayed algebras situated between universal algebras and automata. The relational counterpart of delayed clones is based on infinite sequences of relations. All the corresponding maximal clones are described except for those determined by sequences of equivalence relations or by sequences of binary central relations.

In the field of Universal Algebra J. Berman surveys selected results on the structure of free algebraic systems. His focus is on decompositions of free algebras into simpler components whose interactions can be readily determined. P. Idziak studies the G-spectrum of a variety, a sequence whose k-th term is the number of k-generated algebras in the variety. Based on commutator and tame congruence theory the at most polynomial and at most exponential G-spectra of some locally finite varieties are described. M. Jackson studies the syntactic semigroups. He shows how to efficiently associate a syntactic semigroup (monoid) with a finite set of identities to a semigroup (monoid) with a finite base of identities and finds a language-theoretic equivalent of the above finite basis problem. K. Kaarli and L. Márki

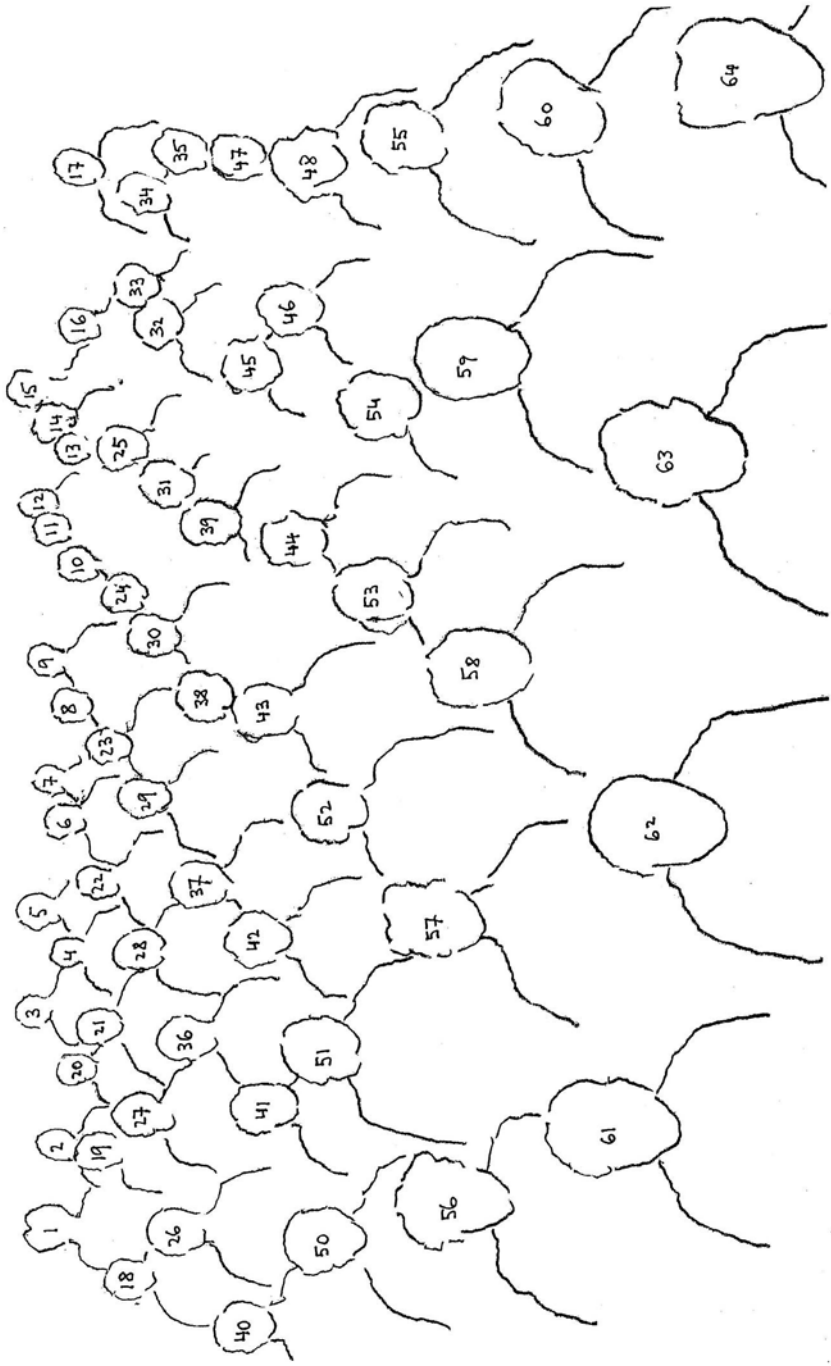
survey endoprimal algebras, i.e. algebras whose term operations comprise all operations admitting a given monoid of selfmaps as their endomorphism monoid. First they present the connection to algebraic dualisability and then characterize the endoprimal algebras among Stone algebras, Kleene algebras, abelian groups, vector spaces, semilattices and implication algebras. A. Krokhin, A. Bulatov and P. Jevons investigate the constraint satisfaction problem arising in artificial intelligence, databases and combinatorial optimization. The algebraic counterpart of this relational problem is a problem in clone theory. The paper studies the computational complexity aspects of the constraint satisfaction problem in clone terms. R. McKenzie and J. Snow present the basic theory of commutators in congruence modular varieties of algebras, an impressive machinery for attacking diverse problems in congruence modular varieties.

It is fair to state that we have met our objective of bringing together specialists and ideas in three neighbouring and closely interrelated domains. To all who helped to make this SMS a success, lecturers and participants alike, we wish to express our sincere thanks and appreciation. Special thanks go to Professor Gert Sabidussi for his experience, help and tireless efforts in the preparation and running of the SMS and, in particular, to Ghislaine David, its very efficient and charming secretary, for the high quality and smoothness with which she handled the organization of the meeting. We also thank Professor Martin Goldstein for the technical edition of this volume.

Funding for the SMS was provided in the largest part by NATO ASI Program with additional support from the Centre de recherches mathématiques of the Université de Montréal and from the Université de Montréal. To all three organizations we would like to express our gratitude for their support.

Ivo G. Rosenberg

x


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64

Key to group picture

1 Hartmann
2 Kun
3 Dinu
4 Vertesi
5 McNulty
6 Salehi
7 Vernikov
8 Piirainen
9 Semigrodskikh
10 Descalço
11 Kudryavtseva
12 Ferreira
13 Silva
14 Maroti
15 Gouveia da Costa
16 F. Almeida
17 Idziak
18 Kaarli
19 Kühr
20 Magnifo Kahou
21 Kuchmei
22 Snow
23 Jackson
24 Kambites
25 Kirsten
26
27 Rachunek
28 Márki
29
30 Halas
31 Ondrusch
32 Kufleitner
33 Mäurer
34 Püschmann
35 Gorachinova
36 Vladislavlev
37 Galatenko
38 Uvarov
39
40 Seif
41 Legault
42 J.-L. David
43 Szabó
44 Chajda
45 Fleischer
46 Sezinando
47 Fearnley
48 Gomes
49 J. Almeida
50 Haddad
51
52 Hikita
53 G. David
54 Sabidussi
55 Dassow
56 McKenzie
57 Krokhin
58 Volkov

59 Shevrin

60 Letichevsky

61 Berman

62 Freivalds

63 Steinby

64 Rosenberg

Participants

Filipa ALMEIDA
Centro de Álgebra
Universidade de Lisboa
Av. Prof. Gama Pinto, 2
P-1649-003 Lisboa
Portugal
filipasda@hotmail.com

Sergiy BORODAY
Centre de recherche informatique de Montréal
550, rue Sherbrooke O. bureau 100
Montréal, QC H3A 1B9
Canada
sboroday@crim.ca

Ivan CHAJDA
Department of Algebra and Geometry
Palacky University, Olomouc
Tomkova 40
779 00 Olomouc
Czech Republic
chajda@risc.upol.cz

Luis A. A. DESCALÇO
Departamento de Matemática
Universidade de Aveiro
P-3810-193 Aveiro
Portugal
luisd@mat.ua.pt

Petrisor Liviu DINU
Faculty of Mathematics
University of Bucharest
14 Academiei St.
70190 Bucharest
Romania
ldinu@funinf.cs.unibuc.ro

Anne FEARNLEY
Dép. de mathématiques et de statistique
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal, QC H3C 3J7
Canada
fearnley@dms.umontreal.ca

Marco A. Semana FERREIRA
Escola Superior de Tecnologia de Viseu
Campus Politécnico de Repeses
P-3504-510 Viseu
Portugal
mferreira@math.estv.ipv.pt

Isidore FLEISCHER
Centre de recherches mathématiques
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal, QC H3C 3J7
Canada
fleischi@crm.umontreal.ca

Rusins FREIVALDS
Inst. of Mathematics and Computer Science
University of Latvia
Raina bulvaris 29
Riga, LV-1459
Latvia
rusins.freivalds@mii.lu.lv

Alexei GALATENKO
Department of MaTIS
Moscow State University
Vorobjevy Gory
119899 Moscow
Russia
agalat@msu.ru

Igor GOLDBERG
Department of Mathematics and Mechanics
Ural State University
Pr. Lenina 51
620083 Ekaterinbourg
Russia
goldberg@skbkontur.ru

Gracinda M. GOMES
Centro de Álgebra
Universidade de Lisboa
Av. Prof. Gama Pinto, 2
P-1649-003 Lisboa
Portugal
ggomes@cii.fc.ul.pt

Lidja GORACHINOVA-ILIEVA
Pedagogical Faculty "Gotse Delchev"
G. Delchev 89
2000 Shtip
Macedonia
fildim@mt.net.mk

Nicolas GORSE
Département d'informatique
et de recherche opérationnelle
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal, QC H3C 3J7
Canada
gorsen@iro.umontreal.ca

Alfredo GOUVEIA DA COSTA
Departamento de Matemática
Universidade de Coimbra
Apartado 3008
P-3001-454 Coimbra
Portugal
amgcd@mat.uc.pt

Vera GROUNSKAIA
Ul'yanovsk State University
Dimitrova 4
433507 Ul'yanovskaya obl.
Dimitrovgrad
Russia
grunska@vinf.ru

Lucien HADDAD
Dept. of Mathematics and Computer Science
Royal Military College of Canada
P.O. Box 17000 Station Forces
Kingston, ON K7K 7B4
Canada
haddad-l@rmc.ca

Radomir HALAS
Department of Algebra and Geometry
Palacky University, Olomouc
Tomkova 40
779 00 Olomouc
Czech Republic
halas@risc.upol.cz

Miklós HARTMANN
Bolyai Institute
University of Szeged
Aradi vértanúk tere 1
6720 Szeged
Hungary
mota@petra.hos.u-szeged.hu

Mark KAMBITES
Department of Mathematics
University of York
Heslington
York YO10 5DD
UK
mek100@york.ac.uk

Christopher KERMOVANT
Département d'informatique
et de recherche opérationnelle
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal, QC H3C 3J7
Canada

Daniel KIRSTEN
LIAFA
Université Paris 7
Case 7014
Place Jussieu
F-75251 Paris Cedex 05
France
kirsten@liafa.jussieu.fr

Vladimir KUCHMEI
Institute of Pure Mathematics
Fac. of Mathematics and Computer Science
University of Tartu
J. Liivi 2-218
Estonia
kucmei@math.ut.ee

Ganna KUDRYAVTSEVA
Faculty of Mechanics and Mathematics
National Taras Shevchenko University of Kiev
64 Volodymyrs'ka St.
01033 Kiev
Ukraine
anjaua@yahoo.com

Manfred KUFLEITNER
Institut für Formale Methoden der Informatik
University of Stuttgart
Universitätsstraße 38
D-70569 Stuttgart
Germany
manfred.kufleitner@informatik.uni-stuttgart.de

Jan KÜHR
Department of Algebra and Geometry
Palacky University, Olomouc
Tomkova 40
779 00 Olomouc
Czech Republic
kuhr@inf.upol.cz

Gabor KUN
Department of Algebra and Number Theory
Eötvös University
Pázmány Péter sétény 1/c.
H-1117 Budapest
Hungary
kungabor@cs.elte.hu

Éric LACOURSIÈRE
Dép. d'informatique et de génie logiciel
Université Laval
Québec, QC G1K 7P4
Canada
eric.lacoursiere@ift.ulaval.ca

Florence Laure MAGNIFO KAHOU
Dép. de mathématiques et de statistique
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal, QC H3C 3J7
Canada
magnifo@dms.umontreal.ca

Miklos MAROTI
Institute for Software Integrated Systems
Vanderbilt University
Box 1829, Station B
Nashville, TN 37235
USA
mmaroti@math.vanderbilt.edu

Ina MÄURER
Institut für Algebra
Technische Universität Dresden
01062 Dresden
Germany
maeurer@math.tu-dresden.de

Pierre MCKENZIE
Département d'informatique
et de recherche opérationnelle
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal, QC H3C 3J7
Canada
mckenzie@iro.umontreal.ca

George MCNULTY
Department of Mathematics
University of South Carolina
1523 Greene Str.
Columbia, SC 29208
USA
mcnulty@math.sc.edu

Nicole ONDRUSCH
Institut für Formale Methoden der Informatik
University of Stuttgart
Universitätsstraße 38
D-70569 Stuttgart
Germany
ondrusch@informatik.uni-stuttgart.de

Ville PIIRAINEN
Turku Centre for Computer Science
Lemminkäisenkatu 14 A, 5th floor
FIN-20520 Turku
Finland
ville.piirainen@utu.fi

Simon PILETTE
Département d'informatique
et de recherche opérationnelle
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal, QC H3C 3J7
Canada
simon.pilette@umontreal.ca

Svetlana PLESCHEVA
Department of Mathematics and Mechanics
Ural State University
Pr. Lenina 51
620083 Ekaterinbourg
Russia
plescheva81@mail.ru

Ulrike PÜSCHMANN
Institut für Algebra
Technische Universität Dresden
01062 Dresden
Germany
u.pueschmann@gmx.net

Jiri RACHUNEK
Department of Algebra and Geometry
Palacky University, Olomouc
Tomkova 40
779 00 Olomouc
Czech Republic
rachunek@risc.upol.cz

Gert SABIDUSSI
Dép. de mathématiques et de statistique
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal, QC H3C 3J7
Canada
sab@dms.umontreal.ca

Saeed SALEHIPOURMEHR
Turku Centre for Computer Science
Lemminkäisenkatu 14 A, 5th floor
FIN-20520 Turku
Finland
saeed@cs.utu.fi

Steve R. SEIF
Department of Mathematics
University of Louisville
Louisville, KY 40292
USA
swseif@louisville.edu

Victor SELIVANOV
Department of Computer Science
Novosibirsk State Pedagogical University
Viluiskaya 28
630126 Novosibirsk
Russia
vseliv@nspu.ru

Alexander SEMIGRODSKIKH
Department of Mathematics and Mechanics
Ural State University
Pr. Lenina 51
620083 Ekaterinbourg
Russia
alexander.semigrodskikh@usu.ru

Helena SEZINANDO
Departamento de Matemática
Universidade de Lisboa
Rua E. de Vasconcelos Bl. C, piso 13
P-1749-016 Lisboa
Portugal
mhelena@ptmat.fc.ul.pt

Nelson SILVA
School of Mathematics and Statistics
University of St Andrews
North Haugh
Fife KY16 9SS
UK
nelson@mcs.st-and.ac.uk

Csaba SZABÓ
Department of Algebra and Number Theory
Eötvös University
Pázmány Péter sétény 1/c.
H-1117 Budapest
Hungary
csaba@cs.elte.hu

Dmitri UVAROV
Department of MaTIS
Moscow State University
Vorobjevy Gory
119899 Moscow
Russia
dima@uvarov.ru

Matt VALERIOTE
Department of Mathematics and Statistics
McMaster University
1280 Main St. West
Hamilton, Ontario L8S 4K1
Canada
matt@math.mcmaster.ca

Victor VLADISLAVLEV
Faculty of Mechanics and Mathematics
Moscow State University
Vorobjevy Gory
119899 Moscow
Russia
vladisla@mcst.ru

Boris M. VERNIKOV
Department of Mathematics and Mechanics
Ural State University
Pr. Lenina 51
620083 Ekaterinbourg
Russia
boris.vernikov@usu.ru

Mikhail VOLKOV
Department of Mathematics and Mechanics
Ural State University
Pr. Lenina 51
620083 Ekaterinbourg
Russia
mikhail.volkov@usu.ru

Vera VERTESI
Department of Algebra and Number Theory
Eötvös University
Pázmány Péter sétény 1/c.
H-1117 Budapest
Hungary
wera13@cs.elte.hu

Contributors

Jorge ALMEIDA
Departamento de Matemática Pura
Universidade do Porto
Rua do Campo Alegre, 687
4169-007 Porto
Portugal
jalmeida@fc.up.pt

Joel BERMAN
Department of Mathematics, Statistics and Computer Science
University of Illinois at Chicago
851 South Morgan
Chicago, IL 60607
USA
jberman@uic.edu

Andrei BULATOV
School of Computer Science
Simon Fraser University
Burnaby, BC V5A 1S6
Canada

Jürgen DASSOW
Fakultät für Informatik
Otto-von-Guericke-Universität
PSF 4120
D-39016 Magdeburg
Germany
dassow@iws.cs.uni-magdeburg.de

Teruo HIKITA
Department of Computer Science
Meiji University
1-1-1 Higashimita
Tama-ku, Kawasaki 214-8571
Japan
hikita@cs.meiji.ac.jp

Paweł M. IDZIAK
Computer Science Department
Jagiellonian University
Nawojki 11
PL-30-072 Kraków
Poland
idziak@ii.uj.edu.pl

Marcel JACKSON
Department of Mathematics
La Trobe University
3086 Victoria
Australia
m.g.jackson@latrobe.edu.au

Peter JEAVONS
Computing Laboratory
University of Oxford
Oxford OX1 3QD
UK

Kalle KAARLI
Institute of Pure Mathematics
University of Tartu
50090 Tartu
Estonia
kaarli@math.ut.ee

László MÁRKI
A. Rényi Institute of Mathematics
Hungarian Academy of Sciences
Pf. 127
H-1364 Budapest
Hungary
marki@renyi.hu

Andrei KROKHIN
Department of Computer Science
University of Durham
Durham, DH1 3LE
UK
andrei.krokhin@durham.ac.uk

Valery B. KUDRYAVTSEV
Department of MaTIS
Moscow State University
Vorobjevy Gory
119899 Moscow
Russia
kudryavtsev@lsili.ru

Alexander LETICHEVSKY
Glushkov Institute of Cybernetics
National Academy of Sciences of Ukraine
252601 Kiev
Ukraine
al@letichevsky.kiev.ua

Ralph MCKENZIE
Department of Mathematics
Vanderbilt University
Nashville, TN 37240
USA
mckenzie@math.vanderbilt.edu

Ivo G. ROSENBERG
Dép. de mathématiques et de statistique
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal, QC H3C 3J7
Canada
rosenb@dms.umontreal.ca

Lev N. SHEVRIN
Department of Mathematics and Mechanics
Ural State University
Lenina 51
620083 Ekaterinburg
Russia
Lev.Shevrin@usu.ru

John SNOW
Department of Mathematics
Concordia University
Seward, NE 68434
USA
john.snow@cune.org

Magnus STEINBY
Department of Mathematics
University of Turku
FIN-20014 Turku
Finland
steinby@utu.fi

Profinite semigroups and applications

Jorge ALMEIDA

Departamento de Matemática Pura, Faculdade de Ciências
Universidade do Porto
Rua do Campo Alegre, 687, 4169-007 Porto
Portugal

Notes taken by Alfredo COSTA

Abstract

Profinite semigroups may be described briefly as projective limits of finite semigroups. They come about naturally by studying pseudovarieties of finite semigroups which in turn serve as a classifying tool for rational languages. Of particular relevance are relatively free profinite semigroups which for pseudovarieties play the role of free algebras in the theory of varieties. Combinatorial problems on rational languages translate into algebraic-topological problems on profinite semigroups. The aim of these lecture notes is to introduce these topics and to show how they intervene in the most recent developments in the area.

1 Introduction

With the advent of electronic computers in the 1950's, the study of simple formal models of computers such as automata was given a lot of attention. The aims were multiple: to understand the limitations of machines, to determine to what extent they might come to replace humans, and later to obtain efficient schemes to organize computations. One of the simplest models that quickly emerged is the finite automaton which, in algebraic terms, is basically the action of a finitely generated free semigroup on a finite set of states and thus leads to a finite semigroup of transformations of the states [48, 61].

In the 1960's, the connection with finite semigroups was first explored to obtain computability results [79] and in parallel a decomposition theory of finite computing devices inspired by the theory of groups and the complexity of such decompositions [51, 52], again led to the development of a theory of finite semigroups [21], which had not previously merited any specific attention from specialists on semigroups.

In the early 1970's, both trends, the former more combinatorial and more directly concerned with applications in computer science, the latter more algebraic, continued to flourish with various results that nowadays are seen as pioneering. In the mid-1970's, S. Eilenberg, in part with the collaboration of M. P. Schützenberger and B. Tilson [35, 36] laid the foundations for a theory which was already giving signs of being potentially quite rich. One of the cornerstones of their work is the notion of a pseudovariety of semigroups and a correspondence between such pseudovarieties and varieties of rational languages which provided a systematic framework and a program for the classification of rational languages.

V. B. Kudryavtsev and I. G. Rosenberg (eds.),
Structural Theory of Automata, Semigroups, and Universal Algebra, 1–45.

The next ten years or so were rich in the execution of Eilenberg's program [53, 64, 65] which in turn led to deep problems such as the identification of the levels of J. Brzozowski's concatenation hierarchy of star-free languages [29] while various steps forward were taken in the understanding of the Krohn-Rhodes group complexity of finite semigroups [73, 71, 47].

In the beginning of the 1980's, the author was exploring connections of the theory of pseudovarieties with Universal Algebra to obtain information on the lattice of pseudovarieties of semigroups and to compute some operators on pseudovarieties (see [3] for results and references). The heart of the combinatorial work was done by manipulating identities and so when J. Reiterman [70] showed that it was possible to define pseudovarieties by pseudoidentities, which are identities with an enlarged signature whose interpretation in finite semigroups is natural, this immediately appeared to be a powerful tool to explore. Reiterman introduced pseudoidentities as formal equalities of implicit operations, and defined a metric structure on sets of implicit operations but no algebraic structure. There is indeed a natural algebraic structure and the interplay between topological and algebraic structure turns out to be very rich and very fruitful.

Thus, the theory of finite semigroups and applications led to the study of profinite semigroups, particularly those that are free relative to a pseudovariety. These structures play the role of free algebras for varieties in the context of profinite algebras, which already explains the interest in them. When the first concrete new applications of this approach started to appear (see [3] for results and references), other researchers started to consider it too and nowadays it is viewed as an important tool which has found applications across all aspects of the theory of pseudovarieties.

The aim of these notes is to introduce this area of research, essentially from scratch, and to survey a significant sample of the most important recent developments. In Section 2 we show how the study of finite automata and rational languages leads to study pseudovarieties of finite semigroups and monoids, including some of the key historical results.

Section 3 explains how relatively free profinite semigroups are found naturally in trying to construct free objects for pseudovarieties, which is essentially the original approach of B. Banaschewski [26] in his independent proof that pseudoidentities suffice to define pseudovarieties. The theory is based here on projective limits but there are other alternative approaches [3, 7]. Section 3 also lays the foundations of the theory of profinite semigroups which are further developed in Section 4, where the operational aspect is explored. Section 4 also includes the recent idea of using iteration of implicit operations to produce new implicit operations. Subsection 4.3 presents for the first time a proof that the monoid of continuous endomorphisms of a finitely generated profinite semigroup is profinite so that implicit operations on finite monoids also have natural interpretations in that monoid.

The remaining sections are dedicated to a reasonably broad survey, without proofs, of how the general theory introduced earlier can be used to solve problems. Section 5 sketches the proof of I. Simon's characterization of piecewise testable languages in terms of the solution of the word problem for free pro-$\mathbf{J}$ semigroups. Section 6 presents an introduction to the notion of tame pseudovarieties, which is a sophisticated tool to handle decidability questions which extends the approach of C. J. Ash to the "Type II conjecture" of J. Rhodes, as presented in the seminal paper [22]. The applications of this approach can be found in Sections 7 and 8 in the computation of several pseudovarieties obtained by applying natural operators to known pseudovarieties. The difficulty in this type of calculation is that it is known that those operators do not preserve decidability [1, 72, 24]. The notion of tameness came about

precisely in trying to find a stronger form of decidability which would be preserved or at least guarantee decidability of the operator image [15].

Finally, Section 9 introduces some very recent developments in the investigation of connections between free profinite semigroups and Symbolic Dynamics. The idea to explore such connections eventually evolved from the need to build implicit operations through iteration in order to prove that the pseudovariety of finite p-groups is tame [6]. Once a connection with Symbolic Dynamics emerged several applications were found but only a small aspect is surveyed in Section 9, namely that which appears to have a potential to lead to applications of profinite semigroups to Symbolic Dynamics.

2 Automata and languages

An abstraction of the notion of an *automaton* is that of a semigroup S acting on a set Q, whose members are called the *states* of the automaton. The action is given by a homomorphism $\varphi : S \to \mathcal{B}_Q$ into the semigroup of all binary relations on the set Q, which we view as acting on the right. If all binary relations in $\varphi(S)$ have domain Q, then one talks about a *complete* automaton, as opposed to a *partial* automaton in the general situation. If all elements of $\varphi(S)$ are functions, then the automaton is said to be *deterministic*. The semigroup $\varphi(S)$ is called the *transition semigroup* of the automaton. In some contexts it is better to work with monoids, and then one assumes the acting semigroup S to be a monoid and the action to be given by a monoid homomorphism φ.

Usually, a set of generators A of the acting semigroup S is fixed and so the action homomorphism φ is completely determined by its restriction to A. In case both Q and A are finite sets, the automaton is said to be *finite*. Of course the restriction that Q is finite is sufficient to ensure that the transition semigroup of the automaton is finite.

To be used as a recognition device, one fixes for an automaton a set I of *initial* states and a set F of *final* states. Moreover, in Computer Science one is interested in recognizing sets of words (or strings) over an alphabet A, so that the acting semigroup is taken to be the semigroup A^+ freely generated by A, consisting of all non-empty words in the letters of the *alphabet* A. The *language recognized* by the automaton is then the following set of words:

$$L = \{w \in A^+ : \varphi(w) \cap (I \times F) \neq \emptyset\}. \tag{2.1}$$

If the empty word 1 is also relevant, then one works instead in the monoid context and one considers the free monoid A^*, the formula (2.1) for the language recognized being then suitably adapted. Whether one works with monoids or with semigroups is often just a matter of personal preference, although there are some instances in which the two theories are not identical. Most results in these notes may be formulated in both settings and we will sometimes switch from one to the other without warning. Parts of the theory may be extended to a much a more general universal algebraic context (see [3, 7] and M. Steinby's lecture notes in this volume).

For an example, consider the automaton described by Fig. 1 where we have two states, 1 and 2, the former being both initial and final, and two acting letters, a and b, the action being determined by the two partial functions associated with a and b, respectively $\bar{a} : 1 \mapsto 2$ and $\bar{b} : 2 \mapsto 1$. The language of $\{a, b\}^*$ recognized by this automaton consists of all words of the form $(ab)^k$ with $k \geq 0$ which are *labels* of paths starting and ending at state 1. This is the submonoid generated by the word ab, which is denoted $(ab)^*$.

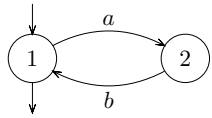


Figure 1

In terms of the action homomorphism, the language L of (2.1) is the inverse image of a specific set of binary relations on Q. We say that a language $L \subseteq A^+$ is *recognized* by a homomorphism $\psi : A^+ \to S$ into a semigroup S if there exists a subset $P \subseteq S$ such that $L = \psi^{-1}P$ or, equivalently, if $L = \psi^{-1}\psi L$. We also say that a language is *recognized* by a finite semigroup S if it is recognized by a homomorphism into S. By the very definition of recognition by a finite automaton, every language which is recognized by such a device is also recognized by a finite semigroup.

Conversely, if $L = \psi^{-1}\psi L$ for a homomorphism $\psi : A^+ \to S$ into a finite semigroup, then one can construct an automaton recognizing L as follows: for the set of states take S^1, the monoid obtained from S by adjoining an identity if S is not a monoid and S otherwise; for the action take the composition of ψ with the *right regular representation*, namely the homomorphism $\varphi : A^+ \to \mathcal{B}_{S^1}$ which sends each word w to right translation by $\psi(w)$, that is the function $s \mapsto s\psi(w)$. This proves the following theorem and, by adding the innocuous assumption that ψ is onto, it also shows that every language which is recognized by a finite automaton is also recognized by a finite complete deterministic automaton with only one initial state (the latter condition being usually taken as part of the definition of deterministic automaton).

2.1 Theorem (Myhill [61]) *A language L is recognized by a finite automaton if and only if it is recognized by a finite semigroup.*

In particular, the complement $A^+ \setminus L$ of a language $L \subseteq A^+$ recognized by a finite automaton is also recognized by a finite automaton since a homomorphism into a finite semigroup recognizing a language also recognizes its complement.

A language $L \subseteq A^*$ is said to be *rational* (or *regular*) if it may be expressed in terms of the empty language and the languages of the form $\{a\}$ with $a \in A$ by applying a finite number of times the binary operations of taking the union $L \cup K$ of two languages L and K or their concatenation $LK = \{uv : u \in L, v \in K\}$, or the unary operation of taking the submonoid L^* generated by L; such an expression is called a *rational expression* of L. For example, if letters stand for elementary tasks a computer might do, union and concatenation correspond to performing tasks respectively in parallel or in series, while the star operation corresponds to iteration. The following result makes an important connection between this combinatorial concept and finite automata. Its proof can be found in any introductory text to automata theory such as Perrin [63].

2.2 Theorem (Kleene [48]) *A language L over a finite alphabet is rational if and only if it is recognized by some finite automaton.*

An immediate corollary which is not evident from the definition is that the set of rational languages $L \subseteq A^*$ is closed under complementation and, therefore it constitutes a Boolean subalgebra of the algebra $\mathcal{P}(A^+)$ of all languages over A.

Rational languages and finite automata play a crucial role in both Computer Science and current applications of computers, since many very efficient algorithms, for instance for dealing with large texts use such entities [34]. This already suggests that studying finite semigroups should be particularly relevant for Computer Science. We present next one historical example showing how this relevance may be explored.

The *star-free* languages over an alphabet A constitute the smallest Boolean subalgebra closed under concatenation of the algebra of all languages over A which contains the empty language and the languages of the form $\{a\}$ with $a \in A$. In other words, this definition may be formulated as that of rational languages but with the star operation replaced by complementation. *Plus-free* languages $L \subseteq A^+$ are defined similarly.

On the other hand we say that a finite semigroup S is *aperiodic* if all its subsemigroups which are groups (in this context called simply *subgroups*) are trivial. Equivalently, the cyclic subgroups of S should be trivial, which translates in terms of universal laws to stating that S should satisfy some identity of the form $x^{n+1} = x^n$.

The connection between these two concepts, which at first sight have nothing to do with each other, is given by the following remarkable theorem.

2.3 Theorem (Schützenberger [79]) *A language over a finite alphabet is star-free if and only if it is recognized by a finite aperiodic monoid.*

Eilenberg [36] has given a general framework in which Schützenberger's theorem becomes an instance of a general correspondence between families of rational languages and finite monoids. To formulate this correspondence, we first introduce some important notions.

The *syntactic congruence* of a subset L of a semigroup S is the largest congruence ρ_L on S which *saturates* L in the sense that L is a union of congruence classes. The existence of such a congruence may be easily established even for arbitrary subsets of universal algebras [3, Section 3.1]. For semigroups, it is easy to see that it is the congruence ρ_L defined by $u \,\rho_L\, v$ if, for all $x, y \in S^1$, $xuy \in L$ if and only if $xvy \in L$, that is if u and v appear as factors of members of L precisely in the same context. The quotient semigroup S/ρ_L is called the *syntactic semigroup* of L and it is denoted $\operatorname{Synt} L$; the natural homomorphism $S \to S/\rho_L$ is called the *syntactic homomorphism* of L.

The syntactic semigroup $\operatorname{Synt} L$ of a rational language $L \subseteq A^+$ is the smallest semigroup S which recognizes L. Indeed all semigroups of minimum size which recognize L are isomorphic. To prove this, one notes that a homomorphism $\psi : A^+ \to S$ recognizing L may as well be taken to be onto, in which case S is determined up to isomorphism by a congruence on A^+, namely the *kernel* congruence $\ker \psi$ which identifies two words if they have the same image under ψ. The assumption that ψ recognizes L translates in terms of this congruence by stating that $\ker \psi$ saturates L and so $\ker \psi$ is contained in ρ_L. Noting that rationality really played no role in the argument, this proves the following result where we say that a semigroup S *divides* a semigroup T and we write $S \prec T$ if S is a homomorphic image of some subsemigroup of T.

2.4 Proposition *A language $L \subseteq A^+$ is recognized by a semigroup S if and only if* $\operatorname{Synt} L$ *divides S.*

The syntactic semigroup of a rational language L may be effectively computed from a rational expression for the language. Namely, one can efficiently compute the *minimal automaton* of L [63], which is the complete deterministic automaton recognizing L with the minimum number of states; the syntactic semigroup is then the transition semigroup of the minimal automaton.

Given a finite semigroup S, one may choose a finite set A and an onto homomorphism $\varphi : A^+ \to S$: for instance, one can take $A = S$ and let φ be the homomorphism which extends the identity funtion $A \to S$. For each $s \in S$, let $L_s = \varphi^{-1}s$. Since φ is an onto homomorphism which recognizes L_s, there is a homomorphism $\psi_s : S \to \operatorname{Synt} L_s$ such that the composite function $\psi_s \circ \varphi : A^+ \to \operatorname{Synt} L_s$ is the syntactic homomorphism of L_s. The functions ψ_s induce a homomorphism $\psi : S \to \prod_{s\in S} \operatorname{Synt} L_s$ which is injective since $\psi_s(t) = \psi_s(s)$ means that there exist $u, v \in A^+$ such that $\varphi(u) = s$, $\varphi(v) = t$ and $u \ \rho_{L_s} \ v$, which implies that $v \in L_s$ since $u \in L_s$ and so $t = s$. As we did at the beginning of the section, we may turn $\varphi : A^+ \to S$ into an automaton which recognizes each of the languages L_s and from this any proof of Kleene's Theorem will yield a rational expression for each L_s. Hence we have the following result.

2.5 Proposition *For every finite semigroup S one may effectively compute rational languages $L_1, \ldots, L_n$ over a finite alphabet A which are recognized by S and such that S divides $\prod_{i=1}^n \operatorname{Synt} L_i$.*

It turns out there are far too many finite semigroups for a classification up to isomorphism to be envisaged [78]. Instead, from the work of Schützenberger and Eilenberg eventually emerged [36, 37] the classification of classes of finite semigroups called *pseudovarieties*. These are the (non-empty) closure classes for the three natural algebraic operators in this context, namely taking homomorphic images, subsemigroups and finite direct products. For example, the classes A, of all finite aperiodic semigroups, and G, of all finite groups, are pseudovarieties of semigroups.

On the language side, the properties of a language may depend on the alphabet on which it is considered. To take into account the alphabet, one defines a *variety of rational languages* to be a correspondence $\mathcal{V}$ associating to each finite alphabet A a Boolean subalgebra $\mathcal{V}(A^+)$ of $\mathcal{P}(A^+)$ such that

(1) if $L \in \mathcal{V}(A^+)$ and $a \in A$ then the *quotient* languages $a^{-1}L = \{w : aw \in L\}$ and $La^{-1} = \{w : wa \in L\}$ belong to $\mathcal{V}(A^+)$ (*closure under quotients*);

(2) if $\varphi : A^+ \to B^+$ is a homomorphism and $L \in \mathcal{V}(B^+)$ then the inverse image $\varphi^{-1}L$ belongs to $\mathcal{V}(A^+)$ (*closure under inverse homomorphic images*).

For example, the correspondence which associates with each finite alphabet the set of all plus-free languages over it is a variety of rational languages. The correspondence between varieties of rational languages and pseudovarieties is easily described in terms of the syntactic semigroup as follows:

- associate with each variety of rational languages $\mathcal{V}$ the pseudovariety V generated by all syntactic semigroups $\operatorname{Synt} L$ with $L \in \mathcal{V}(A^+)$;

- associate with each pseudovariety V of finite semigroups the correspondence

$$\begin{aligned}\mathcal{V} : A \mapsto \mathcal{V}(A^+) &= \{L \subseteq A^+ : \operatorname{Synt} L \in \mathsf{V}\} \\ &= \{L \subseteq A^+ : L \text{ is recognized by some } S \in \mathsf{V}\}\end{aligned}$$

Since intersections of non-empty families of pseudovarieties are again pseudovarieties, pseudovarieties of semigroups constitute a complete lattice for the inclusion ordering. Similarly, one may order varieties of languages by putting $\mathcal{V} \leq \mathcal{W}$ if $\mathcal{V}(A^+) \subseteq \mathcal{W}(A^+)$ for every finite alphabet A. Then every non-empty family of varieties $(\mathcal{V}_i)_{i\in I}$ admits the infimum $\mathcal{V}$ given by $\mathcal{V}(A^+) = \bigcap_{i\in I} \mathcal{V}_i(A^+)$ and so again the varieties of rational languages constitute a complete lattice.

2.6 Theorem (Eilenberg [36]) *The above two correspondences are mutually inverse isomorphisms between the lattice of varieties of rational languages and the lattice of pseudovarieties of finite semigroups.*

Schützenberger's Theorem provides an instance of this correspondence, but of course this by no means says that that theorem follows from Eilenberg's Theorem. See M. V. Volkov's lecture notes in this volume and Section 5 for another important "classical" instance of Eilenberg's correspondence, namely Simon's Theorem relating the variety of so-called piecewise testable languages with the pseudovariety J of finite semigroups in which every principal ideal admits a unique element as a generator. See Eilenberg [36] and Pin [65] for many more examples.

2.7 Example An elementary example which is easy to treat here is the correspondence between the variety $\mathcal{N}$ of finite and cofinite languages and the pseudovariety N of all finite nilpotent semigroups. We say that a semigroup S is *nilpotent* if there exists a positive integer n such that all products of n elements of S are equal; the least such n is called the *nilpotency index* of S. The common value of all sufficiently long products in a nilpotent semigroup must of course be zero. If the alphabet A is finite, the finite semigroup S is nilpotent with nilpotency index n, and the homomorphism $\varphi : A^+ \to S$ recognizes the language L, then either φL does not contain zero, so that L must consist of words of length smaller than n which implies L is finite, or φL contains zero and then every word of length at least n must lie in L, so that the complement of L is finite.

Since N is indeed a pseudovariety and the correspondence $\mathcal{N}$ associating with a finite alphabet A the set of all finite and cofinite languages $L \subseteq A^+$ is a variety of rational languages, to prove the converse it suffices, by Eilenberg's Theorem, to show that every singleton language $\{w\}$ over a finite alphabet A is recognized by a finite nilpotent semigroup. Now, given a finite alphabet A and a positive integer n, the set I_n of all words of length greater than n is an ideal of the free semigroup A^+ and the Rees quotient A^+/I_n, in which all words of I_n are identified to a zero element, is a member of N. If $w \notin I_n$, that is if the length $|w|$ of w satisfies $|w| \leq n$, then the quotient homomorphism $A^+ \to A^+/I_n$ recognizes $\{w\}$. Hence we have $\mathsf{N} \leftrightarrow \mathcal{N}$ via Eilenberg's correspondence.

Eilenberg's correspondence gave rise to a lot of research aimed at identifying pseudovarieties of finite semigroups corresponding to combinatorially defined varieties of rational languages and, conversely, varieties of rational languages corresponding to algebraically defined pseudovarieties of finite semigroups.

Another aspect of the research is explained in part by the different character of the two directions of Eilenberg's correspondence. The pseudovariety V associated with a variety $\mathcal{V}$ of rational languages is defined in terms of generators. Nevertheless, Proposition 2.5 shows how to recover from a given semigroup $S \in \mathsf{V}$ an expression for S as a divisor of a product of generators so that a finite semigroup S belongs to V if and only if the languages computed from S according to Proposition 2.5 belong to $\mathcal{V}$.

On the other hand, if we could effectively test membership in V, then we could effectively determine if a rational language $L \subseteq A^+$ belongs to $\mathcal{V}(A^+)$: we would simply compute the syntactic semigroup of L and test whether it belongs to V, the answer being also the answer to the question of whether $L \in \mathcal{V}(A^+)$. This raises the most common problem encountered in finite semigroup theory: given a pseudovariety V defined in terms of generators, determine whether it has a decidable membership problem. A pseudovariety with this property is said to be *decidable*. Since for instance for each set P of primes, the pseudovariety consisting of all finite groups G such that the prime factors of $|G|$ belong to P determines P, a simple counting argument shows that there are too many pseudovarieties for all of them to be decidable. For natural constructions of undecidable pseudovarieties from decidable ones see [1, 24].

For the reverse direction, given a pseudovariety V one is often interested in natural and combinatorially simple generators for the associated variety $\mathcal{V}$ of rational languages. These generators are often defined in terms of Boolean operations: for each finite alphabet A a "natural" generating subset for the Boolean algebra $\mathcal{V}(A^+)$ should be identified. For instance, a language $L \subseteq A^+$ is *piecewise testable* if and only if it is a Boolean combination of languages of the form $A^*a_1A^* \cdots a_nA^*$ with $a_1, \ldots, a_n \in A$. We will run again into this kind of question in Subsection 3.3 where it will be given a simple topological formulation.

3 Free objects

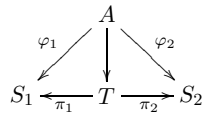
A basic difficulty in dealing with pseudovarieties of finite algebraic structures is that in general they do not have free objects. The reason is quite simple: free objects tend to be infinite.

As a simple example, consider the pseudovariety N of all finite nilpotent semigroups. For a finite alphabet A and a positive integer n, denoting again by I_n the set of all words of length greater than n, the Rees quotient A^+/I_n belongs to N. In particular, there are arbitrarily large A-generated finite nilpotent semigroups and therefore there can be none which is free among them. In general, there is an A-generated free member of a pseudovariety V if and only if up to isomorphism there are only finitely many A-generated members of V, and most interesting pseudovarieties of semigroups fail this condition.

In universal algebraic terms, we could consider the free objects in the variety generated by V. This variety is defined by all identities which are valid in V and for instance for N there are no such nontrivial semigroup identities: in the notation of the preceding paragraph, A^+/I_n satisfies no nontrivial identities in at most $|A|$ variables in which both sides have length at most n. This means that if we take free objects in the algebraic sense then we lose a lot of information since in particular all pseudovarieties containing N will have the same associated free objects.

Let us go back and try to understand better what is meant by a free object. The idea is to take a structure which is just as general as it needs to be in order to be more general than all A-generated members of a given pseudovariety V. Let us take two A-generated members

of V, say given by functions $\varphi_i : A \to S_i$ such that $\varphi_i(A)$ generates S_i $(i = 1, 2)$. Let T be the subsemigroup of the product generated by all pairs of the form $(\varphi_1(a), \varphi_2(a))$ with $a \in A$. Then T is again an A-generated member of V and we have the commutative diagram



where $\pi_i : T \to S_i$ is the projection on the ith component. The semigroup T is therefore more general than both S_1 and S_2 as an A-generated member of V and it is as small as possible to satisfy this property. We could keep going on doing this with more and more A-generated members of V but the problem is that we know, by the above discussion concerning N, that in general we will never end up with one member of V which is more general than all the others. So we need some kind of limiting process. The appropriate construction is the projective (or inverse) limit which we proceed to introduce in the somewhat wider setting of topological semigroups.

3.1 Profinite semigroups

By a *directed set* we mean a poset in which any two elements have a common upper bound. A subset C of a poset P is said to be *cofinal* if, for every element $p \in P$, there exists $c \in C$ such $p \leq c$.

By a *topological semigroup* we mean a semigroup S endowed with a topology such that the semigroup operation $S \times S \to S$ is continuous. Fix a set A and consider the category of *A-generated topological semigroups* whose objects are the mappings $A \to S$ into topological semigroups whose images generate dense subsemigroups, and whose morphisms $\theta : \varphi \to \psi$, from $\varphi : A \to S$ to $\psi : A \to T$, are given by continuous homomorphisms $\theta : S \to T$ such that $\theta \circ \varphi = \psi$. Now, consider a *projective system* in this category, given by a directed set I of indices, for each $i \in I$ an object $\varphi_i : A \to S_i$ in our category of A-generated topological semigroups and, for each pair $i, j \in I$ with $i \geq j$, a *connecting* morphism $\psi_{i,j} : \varphi_i \to \varphi_j$ such that the following conditions hold for all $i, j, k \in I$:

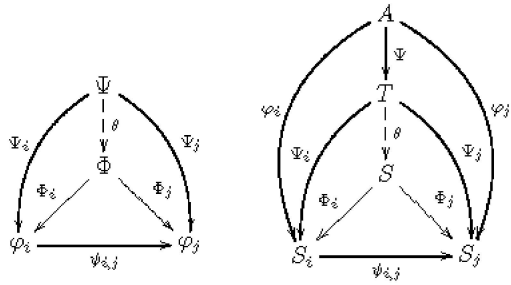
- $\psi_{i,i}$ is the identity morphism on φ_i;
- if $i \geq j \geq k$ then $\psi_{j,k} \circ \psi_{i,j} = \psi_{i,k}$.

The *projective limit* of this projective system is an A-generated topological semigroup $\Phi : A \to S$ together with morphisms $\Phi_i : \Phi \to \varphi_i$ such that for all $i, j \in I$ with $i \geq j$, $\psi_{i,j} \circ \Phi_i = \Phi_j$ and, moreover, the following universal property holds:

> *For any other A-generated topological semigroup $\Psi : A \to T$ and morphisms $\Psi_i : \Psi \to \varphi_i$ such that for all $i, j \in I$ with $i \geq j$, $\psi_{i,j} \circ \Psi_i = \Psi_j$ there exists a morphism $\theta : \Psi \to \Phi$ such that $\Phi_i \circ \theta = \Psi_i$ for every $i \in I$.*

The situation is depicted in the following two commutative diagrams of morphisms and

mappings respectively:



The uniqueness up to isomorphism of such a projective limit is a standard diagram chasing exercise. Existence may be established as follows.

Consider the subsemigroup S of the product $\prod_{i\in I} S_i$ consisting of all $(s_i)_{i\in I}$ such that, for all $i, j \in I$ with $i \geq j$,

$$\varphi_{i,j}(s_i) = s_j \tag{3.1}$$

endowed with the induced topology from the product topology. To check that S provides a construction of the projective limit, we first claim that the mapping $\Phi : A \to S$ given by $\Phi(a) = (\varphi_i(a))_{i\in I}$ is such that $\Phi(A)$ generates a dense subsemigroup T of S. Indeed, since the system is projective, to find an approximation $(t_i)_{i\in I} \in T$ to an element $(s_i)_{i\in I}$ of S given by $t_{i_j} \in K_{i_j}$ for a clopen set $K_{i_j} \subseteq S_{i_j}$ containing s_{i_j} with $j = 1, \ldots, n$, one may first take $k \in I$ such that $k \geq i_1, \ldots, i_n$. Then, by the hypothesis that the subsemigroup T_k of S_k generated by $\varphi_k(A)$ is dense, there is a word $w \in A^+$ which in T_k represents an element of the open set $\bigcap_{j=1}^n \psi_{k,i_j}^{-1} K_{i_j}$ since this set is non-empty as s_k belongs to it. This word w then represents an element $(t_i)_{i\in I}$ of T which is an approximation as required.

It is now an easy exercise to show that the projections $\Phi_i : S \to S_i$ have the above universal property. Note that since each of the conditions (3.1) only involves two components and $\varphi_{i,j}$ is continuous, S is a closed subsemigroup of the product $\prod_{i\in I} S_i$. So, by Tychonoff's Theorem, if all the S_i are compact semigroups, then so is S. We assume Hausdorff's separation axiom as part of the definition of compactness.

Recall that a topological space is *totally disconnected* if its connected components are singletons and it is *zero-dimensional* if it admits a basis of open sets consisting of clopen (meaning both closed and open) sets. See Willard [93] for a background in General Topology.

A finite semigroup is always viewed in this paper as a topological semigroup under the discrete topology. A *profinite semigroup* is defined to be a projective limit of a projective system of finite semigroups in the above sense, that is for some suitable choice of generators. The next result provides several alternative definitions of profinite semigroups.

3.1 Theorem *The following conditions are equivalent for a compact semigroup S:*

(1) *S is profinite;*

(2) *S is residually finite as a topological semigroup;*

(3) *S is a closed subdirect product of finite semigroups;*

(4) *S is totally disconnected;*

(5) *S is zero-dimensional.*

Proof By the explicit construction of the projective limit we have (1) $\Rightarrow$ (2) while (2) $\Rightarrow$ (3) is easily verified from the definitions. For (3) $\Rightarrow$ (1), suppose that $\Phi : S \to \prod_{i\in I} S_i$ is an injective continuous homomorphism from the compact semigroup S into a product of finite semigroups and that the factors are such that, for each component projection $\pi_j : \prod_{i\in I} S_i \to S_j$ the mapping $\pi_j \circ \Phi : S \to S_j$ is onto. We build a projective system of S-generated finite semigroups by considering all onto mappings of the form $\Phi_F : S \to S_F$ where F is a finite subset of I and $\Phi_F = \pi_F \circ \Phi$ where $\pi_F : \prod_{i\in I} S_i \to \prod_{i\in F} S_i$ denotes the natural projection; the indexing set is therefore the directed set of all finite subsets of I, under the inclusion ordering, and for the connecting homomorphisms we take the natural projections. It is now immediate to verify that S is the projective limit of this projective system of finite S-generated semigroups.

Since a product of totally disconnected spaces is again totally disconnected, we have (3) $\Rightarrow$ (4). The equivalence (4) $\Leftrightarrow$ (5) holds for any compact space and it is a well-known exercise in General Topology [93].

Up to this point in the proof, the fact that we are dealing with semigroups rather than any other variety of universal algebras really makes no essential difference. To complete the proof we establish the implication (5) $\Rightarrow$ (2), which was first proved by Numakura [62]. Given two distinct points $s, t \in S$, by zero-dimensionality they may be separated by a clopen subset $K \subseteq S$ in the sense that s lies in K and t does not. Since the syntactic congruence ρ_K saturates K, the congruence classes of s and t are distinct, that is the quotient homomorphism $\varphi : S \to \mathrm{Synt}\, K$ sends s and t to two distinct points. Hence, to prove (2) it suffices to show that $\mathrm{Synt}\, K$ is finite and φ is continuous, which is the object of Lemma 3.3 below. □

As an immediate application we obtain the following closure properties for the class of profinite semigroups.

3.2 Corollary *A closed subsemigroup of a profinite semigroup is also profinite. The product of profinite semigroups is also profinite.*

The following technical result has been extended in [2] to a universal algebraic setting in which syntactic congruences are determined by finitely many terms. See [32] for the precise scope of validity of the implication (5) $\Rightarrow$ (1) in Theorem 3.1 and applications in Universal Algebra.

We say that a congruence ρ on a topological semigroup is *clopen* if its classes are clopen.

3.3 Lemma (Hunter [44]) *If S is a compact zero-dimensional semigroup and K is a clopen subset of S then the syntactic congruence ρ_K is clopen, and therefore it has finitely many classes.*

Proof The proof uses nets, sequences indexed by directed sets which play for general topological spaces the role played by usual sequences for metric spaces [93]. Let $(s_i)_{i\in I}$ be a convergent net in S with limit s. We should show that there exists $i_0 \in I$ such that, whenever $i \geq i_0$, we have $s_i \,\rho_K\, s$. Suppose on the contrary that for every $j \in I$ there exists $i \geq j$ such that s_i is not in the same ρ_K-class as s. The set Λ consisting of all $i \in I$ such that

s_i is not in the ρ_K-class of s is then a cofinal subset of I which determines a subnet $(s_i)_{i\in\Lambda}$ converging to s from outside the ρ_K-class of s.

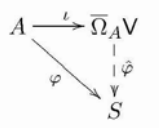
Now we use the characterization of syntactic congruences on semigroups: for each $i \in \Lambda$ there exist $x_i, y_i \in S^1$ such that the products $x_i s_i y_i$ and $x_i s y_i$ do not both lie in K. Note that if a directed set is partitioned into two parts, one of them must contain a cofinal subset. Hence for some cofinal subset of indices M we must have all of the $x_i s_i y_i$ with $i \in M$ lying in K or all of them lying in the complement $S \setminus K$, while the opposite holds for $x_i s y_i$. By compactness and taking subnets, we may as well assume that the nets $(x_i)_{i\in M}$ and $(y_i)_{i\in M}$ converge, say to x and y, respectively. By continuity of multiplication in S, the nets $(x_i s_i y_i)_{i\in M}$ and $(x_i s y_i)_{i\in M}$ both converge to xsy. Since both K and $S \setminus K$ are closed, it follows that xsy must lie in both K and $S \setminus K$, which is absurd. Hence the classes of ρ_K are open and, since they form a partition of a compact space, there can only be finitely many of them and they are also closed. □

3.2 Relatively free profinite semigroups

For a pseudovariety V, we say that a profinite semigroup S is *pro-*V if it is a projective limit of members of V. In view of Theorem 3.1 and its proof, this condition is equivalent to the profinite semigroup being a subdirect product of members of V and also to being *residually* V in the sense that for all distinct $s_1, s_2 \in S$ there exists a continuous homomorphism $\varphi : S \to T$ such that $T \in \mathsf{V}$ and $\varphi(s_1) \neq \varphi(s_2)$.

Let us go back to the construction of free objects for a pseudovariety V. For a generating set A the idea was to take the projective limit of all A-generated members of V. For set theoretical reasons this is inconvenient since there are too many such semigroups but nothing is lost in considering only representatives of isomorphism classes and that is what we do. So, let V_0 be a set containing a representative from each isomorphism class of A-generated members of V. The set V_0 determines a projective system by taking the unique connecting homomorphisms with respect to the choice of generators. The projective limit of this system is denoted $\overline{\Omega}_A\mathsf{V}$.

3.4 Proposition *The profinite semigroup $\overline{\Omega}_A\mathsf{V}$ has the following universal property: the natural mapping $\iota : A \to \overline{\Omega}_A\mathsf{V}$ is such that, for every mapping $\varphi : A \to S$ into a pro-V semigroup there exists a unique continuous homomorphism $\hat{\varphi} : \overline{\Omega}_A\mathsf{V} \to S$ such that $\hat{\varphi} \circ \iota = \varphi$ as depicted in the following diagram:*



Proof Since pro-V semigroups are subdirect products of members of V, it suffices to consider the case when S itself lies in V. Without loss of generality, we may assume that S is generated by $\varphi(A)$. Then S is isomorphic, as an A-generated semigroup, to some member of V_0 and so we may further assume that $S \in \mathsf{V}_0$. But then, from our explicit construction of the projective limit as a closed subsemigroup of a direct product, it suffices to take $\hat{\varphi}$ to be the projection $\overline{\Omega}_A\mathsf{V} \to S$ into the component corresponding to φ. □

Since, by the usual diagram chasing there is up to isomorphism at most one A-generated pro-V semigroup with the above universal property, we conclude that $\overline{\Omega}_A\mathsf{V}$ does not depend, up to isomorphism, on the choice of V_0. We call $\overline{\Omega}_A\mathsf{V}$ the *free pro-*V *semigroup on* A. A profinite semigroup is said to be *relatively free* if it is of the form $\overline{\Omega}_A\mathsf{V}$ for some set A and some pseudovariety V.

By construction, $\overline{\Omega}_A\mathsf{V}$ is an A-generated topological semigroup so that the subsemigroup $\Omega_A\mathsf{V}$ generated by the image of ι is dense in $\overline{\Omega}_A\mathsf{V}$, which explains the line over the capital omega. From Proposition 3.4 it follows that $\Omega_A\mathsf{V}$ is the free semigroup in the variety generated by V.

The mapping ι is injective provided V is not the trivial pseudovariety consisting of singleton semigroups. Hence we will often identify the elements of A with their images under ι.

3.3 Recognizable subsets

The following result characterizes the subsets of a pro-V semigroup which are recognized by members of V. The reader may wish to compare it with Hunter's lemma.

3.5 Proposition *Let S be a pro-V semigroup and let $K \subseteq S$. Then the following conditions are equivalent:*

(1) *there exists a continuous homomorphism $\varphi : S \to F$ such that $F \in \mathsf{V}$ and $K = \varphi^{-1}\varphi K$;*

(2) *K is clopen;*

(3) *the syntactic congruence ρ_K is clopen.*

In particular, all these conditions imply that the syntactic semigroup $\operatorname{Synt} K$ *belongs to* V.

Proof Assuming the existence of a function φ satisfying (1), we deduce that K is clopen since it is the inverse image under a continuous function of a clopen set. For the converse, suppose K is clopen and let $S \hookrightarrow \prod_{i\in I} S_i$ be a subdirect product of a family of members of V. Then K may be expressed as $K = S \cap (K_1 \cup \cdots \cup K_n)$, where each K_ℓ is a product of the form $\prod_{i\in I} X_i$ with $X_i \subseteq S_i$ and $X_i = S_i$ for all but finitely many indices. Let J be the (finite) set of all exceptional indices with $\ell = 1, \ldots, n$ and consider the projection $\varphi : S \to \prod_{i\in J} S_i$. Then it is routine to check that φ is a continuous homomorphism satisfying the required conditions. This establishes the equivalence (1) $\Leftrightarrow$ (2).

If K is clopen then ρ_K is clopen by Hunter's Lemma. This proves (2) $\Rightarrow$ (3) and for the converse it suffices to recall that K is saturated by ρ_K.

Finally, assuming (1), K is recognized by a semigroup from V. Since the syntactic semigroup $\operatorname{Synt} K$ divides every semigroup which recognizes K by Proposition 2.4, $\operatorname{Synt} K$ belongs to V since V is closed under taking divisors. □

We note that the assumption that $\operatorname{Synt} K$ belongs to V for a subset K of a pro-V semigroup S does not suffice to deduce that K is clopen, as the following example shows. Take S to be the Cantor set and consider the left-zero multiplication $st = s$ on S. Although one could easily show it directly, by Theorem 3.1 S is profinite and hence it is pro-LZ for the pseudovariety LZ of all finite left-zero semigroups. Let K be a subset of S. Then a simple calculation shows that the syntactic congruence ρ_K consists of two classes, namely K and its

complement. Hence Synt K belongs to LZ for an arbitrary subset $K \subseteq S$ while K does need to be clopen.

We say that a subset L of a semigroup S is V-*recognizable* if there exists a homomorphism $\varphi : S \to F$ into some $F \in \mathsf{V}$ such that $L = \varphi^{-1}\varphi L$. Proposition 3.5 leads to the following topological characterization of V-recognizable subsets of $\Omega_A\mathsf{V}$.

3.6 Theorem *The following conditions are equivalent for a subset $L \subseteq \Omega_A\mathsf{V}$:*

(1) *L is V-recognizable;*

(2) *the closure $K = \overline{L} \subseteq \overline{\Omega}_A\mathsf{V}$ is open and $L = K \cap \Omega_A\mathsf{V}$;*

(3) *$L = K \cap \Omega_A\mathsf{V}$ for some clopen $K \subseteq \overline{\Omega}_A\mathsf{V}$.*

Proof To prove (1) $\Rightarrow$ (2), suppose L is recognized by a homomorphism $\varphi : \Omega_A\mathsf{V} \to F$ such that $F \in \mathsf{V}$ and $L = \varphi^{-1}\varphi L$. By the universal property of $\overline{\Omega}_A\mathsf{V}$, there exists a unique continuous homomorphism $\hat{\varphi} : \overline{\Omega}_A\mathsf{V} \to F$ extending φ. Then $K = \hat{\varphi}^{-1}\varphi L$ is open and satisfies $K \cap \Omega_A\mathsf{V} = L$. Since $\Omega_A\mathsf{V}$ is dense in $\overline{\Omega}_A\mathsf{V}$ so is L dense in K, which shows K has the required properties for (2).

The implication (2) $\Rightarrow$ (3) is trivial, so it remains to show (3) $\Rightarrow$ ((1)). Suppose (3) holds. By Proposition 3.5 there exists a continuous homomorphism $\psi : \overline{\Omega}_A\mathsf{V} \to F$ such that $F \in \mathsf{V}$ and $K = \psi^{-1}\psi K$. Let φ be the restriction of ψ to $\Omega_A\mathsf{V}$. Then we have $L = \Omega_A\mathsf{V} \cap K = \Omega_A\mathsf{V} \cap \psi^{-1}\psi K = \varphi^{-1}\psi K$ and so L is V-recognizable. □

Another application of Proposition 3.5 is the following result.

3.7 Proposition *The image of a pro-V semigroup under a continuous homomorphism into a profinite semigroup is pro-V and it belongs to V if it is finite.*

Proof Let $\varphi : S \to T$ be a continuous homomorphism with S pro-V and T profinite. Since T is a subdirect product of finite semigroups, it suffices to consider the case where T is finite and φ is onto and show that $T \in \mathsf{V}$. The sets $K_t = \varphi^{-1}t$, with $t \in T$, are clopen subsets of S. By Proposition 3.5, there is for each $t \in T$ a continuous homomorphism $\psi_t : S \to F_t$ such that $F_t \in \mathsf{V}$ and $\psi_t^{-1}\psi_t K_t = K_t$. The induced homomorphism $\psi : S \to F$, where $F = \prod_{t\in T} F_t$, has a kernel $\ker\psi$ which is contained in $\ker\varphi$ and so T divides F, which shows that indeed $T \in \mathsf{V}$. □

The hypothesis in Proposition 3.7 that the continuous homomorphism assumes values in a profinite semigroup cannot be removed as the following example shows. We take again S to be the Cantor set under left-zero multiplication. It is well-known that the unit interval $T = [0, 1]$ is a continuous image of S and so it is also a continuous homomorphic image if we endow it with the left-zero multiplication. But of course T is not zero-dimensional and therefore it is not profinite.

We have seen in Section 2 that one is often interested in describing a variety of rational languages $\mathcal{V}$ by giving a set of generators for the Boolean algebra $\mathcal{V}(A^+)$ for each finite alphabet A. We now aim to characterize this property in topological terms.

3.8 Proposition *The following conditions are equivalent for a family $\mathcal{F}$ of V-recognizable subsets of $\Omega_A\mathsf{V}$, where $\overline{\mathcal{F}} = \{\overline{L} : L \in \mathcal{F}\}$:*

(1) *$\mathcal{F}$ generates the Boolean algebra of all V-recognizable subsets of $\Omega_A\mathsf{V}$;*

(2) *$\overline{\mathcal{F}}$ generates the Boolean algebra of all clopen subsets of $\overline{\Omega}_A\mathsf{V}$;*

(3) *$\overline{\mathcal{F}}$ suffices to separate points of $\overline{\Omega}_A\mathsf{V}$.*

Proof Note that, for subsets $L, L_1, L_2 \subseteq \Omega_A\mathsf{V}$, we have $\overline{L_1 \cup L_2} = \overline{L_1} \cup \overline{L_2}$ and, by Theorem 3.6, in case L is V-recognizable, the closure $K = \overline{L}$ is clopen with $K \cap \Omega_A\mathsf{V} = L$ and so $\overline{\Omega_A\mathsf{V} \setminus L} = \overline{\Omega}_A\mathsf{V} \setminus \overline{L}$. Hence a Boolean expression for L in terms of elements of $\mathcal{F}$ gives rise to a Boolean expression for $\overline{L}$ in terms of elements of $\overline{\mathcal{F}}$ and vice versa. This proves the equivalence (1) $\Leftrightarrow$ (2).

Assume (2) and let $s, t \in S$ be two distinct points. Since the topology of $\overline{\Omega}_A\mathsf{V}$ is zero-dimensional, there exists a clopen subset $K \subseteq \overline{\Omega}_A\mathsf{V}$ such that $s \in K$ and $t \notin K$. By assumption, K admits a Boolean expression in terms of the closures of the elements of $\mathcal{F}$ and therefore it admits an expression as a finitary union of finitary intersections of members of $\overline{\mathcal{F}}$ and their complements. At least one term of the union must contain s and none of them can contain t, and so we may avoid taking the union. Similarly, we may avoid taking the intersection, which shows that there is an element of $\overline{\mathcal{F}}$ which contains one of s and t but not the other. This proves (2) $\Rightarrow$ (3).

It remains to show that (3) $\Rightarrow$ (2). Let $\overline{\mathcal{F}}'$ be the family of all elements of $\overline{\mathcal{F}}$ together with their complements. Given a closed subset $C \subseteq \overline{\Omega}_A\mathsf{V}$ and $s \in \overline{\Omega}_A\mathsf{V} \setminus C$, for each $t \in C$ there exists $K_t \in \overline{\mathcal{F}}'$ such that $s \notin K_t$ and $t \in K_t$. Then the K_t constitute an open cover of the closed set C and so there are $t_1, \ldots, t_n \in C$ such that $K = K_{t_1} \cup \cdots \cup K_{t_n}$ contains C but not s. This shows that we may separate points from closed sets using finitary unions of elements of $\overline{\mathcal{F}}'$.

Let C now be a clopen subset of $\overline{\Omega}_A\mathsf{V}$. For each $s \in C$ we can find a member K_s of the Boolean algebra generated by $\overline{\mathcal{F}}$ containing s with empty intersection with the closed set $\overline{\Omega}_A\mathsf{V} \setminus C$, that is such that $K_s \subseteq C$. Then the compact set C is covered by the open sets K_s with $s \in C$ and so there exist $s_1, \ldots, s_m$ such that $K = K_{s_1} \cup \cdots \cup K_{s_m}$ covers C and $K \subseteq C$, that is $C = K$. This shows that C belongs to the Boolean algebra generated by $\overline{\mathcal{F}}'$. □

3.4 Metric structure

We end this section with a brief reference to a natural metric on finitely generated profinite semigroups. Let S be a profinite semigroup. Define, for $u, v \in S$,

$$d(u,v) = \begin{cases} 2^{-r(u,v)} & \text{if } u \neq v, \\ 0 & \text{otherwise,} \end{cases} \tag{3.2}$$

where $r(u,v)$ denotes the minimum cardinality of a finite semigroup T such that there exists a continuous homomorphism $\varphi : S \to T$ with $\varphi(u) \neq \varphi(v)$. Note that d is an *ultrametric* in the sense that $d : S \times S \to [0, +\infty)$ is a function satisfying the following conditions:

- $d(u,v) = 0$ if and only if $u = v$;
- $d(u,v) = d(v,u)$;

- $d(u,w) \leq \max\{d(u,v), d(v,w)\}$.

The latter condition is trivial if any two of the three elements $u, v, w \in S$ coincide and otherwise, taking logarithms, we deduce that it is equivalent to the inequality $r(u,w) \geq \min\{r(u,v), r(v,w)\}$ which follows from the trivial fact that if $\varphi(u) = \varphi(v)$ and $\varphi(v) = \varphi(w)$ for a function $\varphi : S \to T$ then $\varphi(u) = \varphi(w)$. We call d the *natural metric* on S.

3.9 Proposition *For a profinite semigroup S, the topology of S is contained in the topology induced by the natural metric and the two topologies coincide in the case where S is finitely generated.*

Proof We denote by $B_\varepsilon(u)$ the open ball $\{v \in S : d(u,v) < \varepsilon\}$. Given a clopen subset K of S, by Proposition 3.5 there exists a continuous homomorphism $\varphi : S \to T$ into a finite semigroup T such that $K = \varphi^{-1}\varphi K$. Now, for $t \in T$, the ball $B_{2^{-|T|}}(t)$ is contained in $\varphi^{-1}(t)$ and so K is a finite union of open balls.

Next assume that S is finitely generated. Consider the open ball $B = B_{2^{-n}}(u)$. Observe that up to isomorphism there are only finitely many semigroups with at most n elements. Since S is finitely generated, there are only finitely many kernels of continuous homomorphisms from S into semigroups with at most n elements and so their intersection is a clopen congruence on S. It follows that there exists a continuous homomorphism $\varphi : S \to T$ into a finite semigroup T such that $\varphi(u) = \varphi(v)$ if and only if $r(u,v) > n$. Hence $B = \varphi^{-1}\varphi B$ so that B is open in the topology of S. □

We observe that the natural metric d is such that the multiplication is contracting in the sense that the following additional condition is satisfied:

- $d(uv, wz) \leq \max\{d(u,w), d(v,z)\}$.

The completion $\hat{S}$ of a topological semigroup S whose topology is induced by a contracting metric inherits a semigroup structure where the product of two elements $s, t \in \hat{S}$ is defined by taking any sequences $(s_n)_n$ and $(t_n)_n$ converging respectively to s and t and noting that $(s_n t_n)_n$ is a Cauchy sequence whose limit st does not depend on the choice of the two sequences. This gives $\hat{S}$ the structure of a topological semigroup.

In the case of a relatively free profinite semigroup $\overline{\Omega}_A\mathsf{V}$, by Proposition 3.7 the finite continuous homomorphic images of $\overline{\Omega}_A\mathsf{V}$ are the A-generated members of V. Moreover, by Proposition 3.4 every homomorphism from $\Omega_A\mathsf{V}$ to a member of V has a unique continuous homomorphic extension to $\overline{\Omega}_A\mathsf{V}$. We define the *natural metric* on $\Omega_A\mathsf{V}$ to be the restriction to $\Omega_A\mathsf{V}$ of the natural metric on $\overline{\Omega}_A\mathsf{V}$ and we observe that, by the preceding remarks, this is equivalent to defining the natural metric directly on $\Omega_A\mathsf{V}$ by the formula (3.2) where now $r(u,v)$ denotes the minimum cardinality of a semigroup $T \in \mathsf{V}$ such that there exists a homomorphism $\varphi : \Omega_A\mathsf{V} \to T$ with $\varphi(u) \neq \varphi(v)$.

We are thus led to an alternative construction of $\overline{\Omega}_A\mathsf{V}$.

3.10 Theorem *For a finite set A, the completion of the semigroup $\Omega_A\mathsf{V}$ with respect to the natural metric is a profinite semigroup isomorphic to $\overline{\Omega}_A\mathsf{V}$.*

Proof By Proposition 3.9, $\overline{\Omega}_A\mathsf{V}$ is a metric space under the natural metric, and its restriction to the dense subspace $\Omega_A\mathsf{V}$ is the natural metric of $\Omega_A\mathsf{V}$. By results in General Topology [93, Theorem 24.4], it follows that the metric space $(\overline{\Omega}_A\mathsf{V}, d)$ is the completion of $(\Omega_A\mathsf{V}, d)$.

It remains to show that the multiplication of the completion as defined above coincides with the multiplication of $\overline{\Omega}_A\mathsf{V}$ which follows from the continuity of multiplication in $\overline{\Omega}_A\mathsf{V}$. □

4 The operational point of view

It is well known from Universal Algebra that a term w in a free algebra F on n generators (the variables) in a variety $\mathcal{V}$ induces an n-ary operation on the members S of $\mathcal{V}$ basically by substituting the arguments for the variables and operating in S [31]. This may be formulated by taking the unique extension of the variable evaluation to a homomorphism $\varphi : F \to S$ and then computing the image $\varphi(w)$. This formulation can be suitably applied to relatively free profinite semigroups, which is the starting point for this section.

4.1 Implicit operations

Given $w \in \overline{\Omega}_A\mathsf{V}$ and a pro-V semigroup S, there is a natural interpretation of w as an operation on S namely the mapping $w_S : S^A \to S$ which sends a function $\varphi : A \to S$ to $\hat{\varphi}(w)$ where $\hat{\varphi} : \overline{\Omega}_A\mathsf{V} \to S$ is the unique continuous homomorphism such that $\hat{\varphi} \circ \iota = \varphi$, where in turn $\iota : A \to \overline{\Omega}_A\mathsf{V}$ denotes the generating function associated with $\overline{\Omega}_A\mathsf{V}$ as the free pro-V semigroup on A.

4.1 Proposition *The function w_S as defined above is continuous and if $f : S \to T$ is a continuous homomorphism between two pro-V semigroups then the following diagram commutes*

$$\begin{array}{ccc} S^A & \xrightarrow{w_S} & S \\ {\scriptstyle f^A}\downarrow & & \downarrow{\scriptstyle f} \\ T^A & \xrightarrow{w_T} & T \end{array} \tag{4.1}$$

where $f^A(\varphi) = f \circ \varphi$ for $\varphi \in S^A$.

Proof We first prove the commutativity of the diagram (4.1). Consider the diagram

$$\begin{array}{ccc} A & \xrightarrow{\iota} & \overline{\Omega}_A\mathsf{V} \\ {\scriptstyle\varphi}\downarrow & \overset{\hat{\varphi}}{\swarrow} & \downarrow{\scriptstyle \widehat{f\circ\varphi}} \\ S & \xrightarrow[f]{} & T \end{array}$$

By the universal property of $\overline{\Omega}_A\mathsf{V}$ there is a unique continuous homomorphism $\widehat{f \circ \varphi}$ such that the diagram commutes. Since $f \circ \hat{\varphi}$ also has this property, it follows that $\widehat{f \circ \varphi} = f \circ \hat{\varphi}$. Hence we have

$$w_T(f^A(\varphi)) = w_T(f \circ \varphi) = \widehat{f \circ \varphi}(w) = (f \circ \hat{\varphi})(w) = f(\hat{\varphi}(w)) = f(w_S(\varphi)).$$

which establishes commutativity of diagram (4.1).

To prove continuity of the natural interpretation w_S, let $K \subseteq S$ be a clopen subset. By Proposition 3.5 there exists a continuous homomorphism $f : S \to T$ such that $T \in \mathsf{V}$ and $K = f^{-1}fK$. By commutativity of diagram (4.1) we have

$$w_S^{-1}K = w_S^{-1}f^{-1}fK = (f^A)^{-1}w_T^{-1}fK.$$

Since w_T is continuous, as T is finite, and f^A is also continuous, we conclude that $w_S^{-1}K$ is clopen. Hence w_s is continuous. □

We say that the operation w *commutes with the homomorphism* $f : S \to T$ if the diagram (4.1) commutes. An operation $w = (w_S)_{S\in\mathsf{V}}$ with an interpretation $w_S : S^A \to A$ on each $S \in \mathsf{V}$ is called an *A-ary implicit operation on* V if it commutes with every homomorphism $f : S \to T$ between members of V. The natural interpretation provides a representation

$$\begin{aligned}\Theta : \overline{\Omega}_A\mathsf{V} &\to \{A\text{-ary implicit operations on } \mathsf{V}\}\\ w &\mapsto (w_S)_{S\in\mathsf{V}}\end{aligned}$$

Since $\overline{\Omega}_A\mathsf{V}$ is residually V, given distinct $u, v \in \overline{\Omega}_A\mathsf{V}$ there exists $S \in \mathsf{V}$ and a continuous homomorphism $\varphi : \overline{\Omega}_A\mathsf{V} \to S$ such that $\varphi(u) \neq \varphi(v)$, that is $u_S(\varphi\circ\iota) \neq v_S(\varphi\circ\iota)$, and so we have $u_S \neq v_S$, which shows that Θ is injective.

4.2 Theorem *The mapping Θ is a bijection.*

Proof Let $w = (w_S)_{S\in\mathsf{V}}$ be an A-ary implicit operation on V. We exhibit an element $s \in \overline{\Omega}_A\mathsf{V}$ such that $\Theta(s) = w$. For this purpose, we take a specific representation of $\overline{\Omega}_A\mathsf{V}$ as a projective limit of members of V namely as the projective limit of a projective system containing one representative from each isomorphism class of A-generated members of V: $\varphi_i : A \to S_i$ $(i \in I)$ with connecting morphisms $\psi_{i,j} : \varphi_i \to \varphi_j$ $(i \geq j)$. Let $s_i = w_{S_i}(\varphi_i)$. Since w is an implicit operation, a simple calculation shows that $\psi_{i,j}(s_i) = s_j$ whenever $i \geq j$. Hence $(s_i)_{i\in I}$ determines an element s of the projective limit $\overline{\Omega}_A\mathsf{V}$.

It remains to show that $\Theta(s) = w$, that is $s_T = w_T$ for every $T \in \mathsf{V}$. Let $\varphi \in T^A$. Since both s and w are implicit operations, we may as well assume that $\varphi(A)$ generates T. Hence up to isomorphism $\varphi : A \to T$ is one of the $\varphi_i : A \to S_i$ and so we may assume that the two mappings coincide. Then we have

$$s_T(\varphi) = s_{S_i}(\varphi_i) = \widehat{\varphi}_i(s) = s_i = w_{S_i}(\varphi_i) = w_T(\varphi)$$

where the middle step comes from the observation that $\widehat{\varphi}_i$ is the projection on the ith component. This completes the proof of the equality $\Theta(s) = w$. □

In view of Theorem 4.2 we will from here on identify members of $\overline{\Omega}_A\mathsf{V}$ with A-ary implicit operations on V. It is this operational point of view that explains the capital omega Ω in the notation for free pro-V semigroups. Starting from an implicit operation on V we realize that it has a natural extension to an operation on all pro-V semigroups which commutes with continuous homomorphisms. Treating a positive integer n as a set with n elements, we may speak of *n-ary implicit operations*.

Recall that $\Omega_A\mathsf{V}$ denotes the subsemigroup of $\overline{\Omega}_A\mathsf{V}$ generated by the image of the natural generating mapping $\iota : A \to \overline{\Omega}_A\mathsf{V}$. Note that natural interpretation of $\iota(a)$ for $a \in A$ is given by $\varphi \in S^A \mapsto \varphi(a)$, that is the projection on the a-component if we view S^A as a product. Hence $\Omega_A\mathsf{V}$ corresponds under the above mapping Θ to the semigroup of A-ary implicit operations generated by these component projections, that is the semigroup terms over A as interpreted in V. The elements of $\Omega_A\mathsf{V}$ are also called *explicit operations*.

For a class $\mathcal{C}$ of finite semigroups, denote by $\mathsf{V}(\mathcal{C})$ the pseudovariety generated by $\mathcal{C}$. In case $\mathcal{C} = \{S_1, \ldots, S_n\}$, we may write $\mathsf{V}(S_1, \ldots, S_n)$ instead of $\mathsf{V}(\mathcal{C})$. Note that $\mathsf{V}(S_1, \ldots, S_n) = \mathsf{V}(S_1 \times \cdots \times S_n)$.

4.3 Proposition *Let S be a finite semigroup, $\mathsf{V} = \mathsf{V}(S)$, and let A be a finite set. Then there is an embedding $\overline{\Omega}_A\mathsf{V} \hookrightarrow S^{S^A}$ and so $\overline{\Omega}_A\mathsf{V}$ is finite and $\overline{\Omega}_A\mathsf{V} = \Omega_A\mathsf{V}$.*

Proof Define the mapping $\Phi : \overline{\Omega}_A\mathsf{V} \to S^{S^A}$ by sending each $w \in \overline{\Omega}_A\mathsf{V}$ to its natural interpretation $w_S : S^A \to S$. Since implicit operations commute with homomorphisms, if two implicit operations $u, v \in \overline{\Omega}_A\mathsf{V}$ coincide in S then they must also coincide in all of V, which consists of divisors of finite products of copies of S. Hence our mapping is injective and the rest of the statement follows immediately. □

If V and W are pseudovarieties with $\mathsf{V} \subseteq \mathsf{W}$, then an implicit operation $w \in \overline{\Omega}_A\mathsf{W}$ determines an implicit operation $w|_{\mathsf{V}} \in \overline{\Omega}_A\mathsf{V}$ by restriction: $(w_S)_{S\in\mathsf{W}} \mapsto (w_S)_{S\in\mathsf{V}}$. The mapping

$$\begin{aligned} \overline{\Omega}_A\mathsf{W} &\to \overline{\Omega}_A\mathsf{V} \\ w &\mapsto w|_{\mathsf{V}} \end{aligned}$$

is called the *natural projection*. In terms of the construction of the projective limit, this is indeed a projection which is obtained by disregarding all components in the product $\prod_{i\in I} S_i$ corresponding to A-generated members of W which are not in V. This proves the following result.

4.4 Proposition *For pseudovarieties V and W with $\mathsf{V} \subseteq \mathsf{W}$, the natural projection $\overline{\Omega}_A\mathsf{W} \to \overline{\Omega}_A\mathsf{V}$ is an onto continuous homomorphism.*

4.2 Pseudoidentities

By a V-*pseudoidentity* we mean a formal equality $u = v$, with $u, v \in \overline{\Omega}_A\mathsf{V}$ for some finite set A. If $u_S = v_S$ then we say that the pseudoidentity $u = v$ *holds* in a given pro-V semigroup S, or that S *satisfies* $u = v$, and we write $S \models u = v$. Note that $S \models u = v$ for $u, v \in \overline{\Omega}_A\mathsf{V}$ if and only if, for every continuous homomorphism $\varphi : \overline{\Omega}_A\mathsf{V} \to S$, the equality $\varphi(u) = \varphi(v)$ holds. For a subclass $\mathcal{C} \subseteq \mathsf{V}$, we also write $\mathcal{C} \models u = v$ if $S \models u = v$ for every $S \in \mathcal{C}$. The following result is immediate from the definitions.

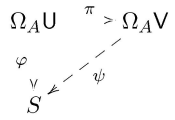
4.5 Lemma *Let V and W be pseudovarieties with $\mathsf{V} \subseteq \mathsf{W}$ and let $\pi : \overline{\Omega}_A\mathsf{W} \to \overline{\Omega}_A\mathsf{V}$ be the natural projection. Then, for $u, v \in \overline{\Omega}_A\mathsf{W}$, we have $\mathsf{V} \models u = v$ if and only if $\pi(u) = \pi(v)$.*

For a set Σ of V-pseudoidentities, we denote by $[\![\Sigma]\!]_{\mathsf{V}}$ (or simply by $[\![\Sigma]\!]$ if V is understood from the context) the class of all $S \in \mathsf{V}$ which satisfy all pseudoidentities from Σ. From the fact that implicit operations on V commute with homomorphisms between members of V, it follows easily that $[\![\Sigma]\!]$ is a pseudovariety contained in V. The converse is also true.

4.6 Theorem (Reiterman [70]) *A subclass V of a pseudovariety W is a pseudovariety if and only if it is of the form $\mathsf{V} = [\![\Sigma]\!]_{\mathsf{W}}$ for some set Σ of W-pseudoidentities.*

Proof Let V be a pseudovariety contained in W and let Σ denote the set of all W-pseudoidentities $u = v$ satisfied by V with $u, v \in \overline{\Omega}_A\mathsf{W}$ and $A \subseteq X$, where X is a fixed countably infinite set. Then we have $\mathsf{V} \subseteq [\![\Sigma]\!]$ and we claim that equality holds. Let $\mathsf{U} = [\![\Sigma]\!]$ and let $S \in \mathsf{U}$. Then there exists $A \subseteq X$ and an onto continuous homomorphism $\varphi : \overline{\Omega}_A\mathsf{U} \to S$. Let $\pi : \overline{\Omega}_A\mathsf{U} \to \overline{\Omega}_A\mathsf{V}$ be the natural projection. By Lemma 4.5, if $u, v \in \overline{\Omega}_A\mathsf{U}$ are such that

$\pi(u) = \pi(v)$ then $\mathsf{V} \models u = v$ and so $u = v$ is a pseudoidentity from Σ so that $S \models u = v$ and $\varphi(u) = \varphi(v)$. This show that $\ker \pi \subseteq \ker \varphi$ and therefore there exists a unique homomorphism $\psi : \overline{\Omega}_A\mathsf{V} \to S$ such that the following diagram commutes:



We claim that ψ is continuous. Indeed, given a subset $K \subseteq S$, by continuity of φ the set $\varphi^{-1}K$ is closed and therefore by continuity of π, $\psi^{-1}K = \pi\varphi^{-1}K$ is closed. Hence ψ is an onto continuous homomorphism. It follows that $S \in \mathsf{V}$ by Proposition 3.7. Hence $\mathsf{V} = [\![\Sigma]\!]_{\mathsf{W}}$. □

There are by now many proofs of this result. It is only fair to mention Banaschewski's proof [26] which was obtained independently of Reiterman's proof and which suggested looking at sets of implicit operations as algebraic-topological structures, a viewpoint which proved to be very productive.

In these notes from hereon we will always take the W of Theorem 4.6 to be the pseudovariety S of all finite semigroups, that is all pseudoidentities will be S-pseudoidentities. A set Σ of pseudoidentities such that $\mathsf{V} = [\![\Sigma]\!]$ is called a *basis of pseudoidentities* for V. The pseudovariety V will be called *finitely based* if it admits a finite basis of pseudoidentities.

To give examples illustrating Reiterman's Theorem, we now describe some important unary implicit operations on finite semigroups. There are several equivalent ways to describe them so we will choose one which is economical in the sense that it requires essentially no verification. For a finite semigroup S, $s \in S$, and $k \in \mathbb{Z}$, the sequence $(s^{n!+k})_n$ becomes constant for n sufficiently large, namely $n > \max\{|k|, |S|\}$ suffices. Hence, in a profinite semigroup S, for $s \in S$ and $k \in \mathbb{Z}$, the sequence $(s^{n!+k})_n$ converges; we denote its limit $s^{\omega+k}$. In particular, we have implicit operations $x^{\omega+k} \in \overline{\Omega}_1\mathsf{S}$ where x is the free generator of $\overline{\Omega}_1\mathsf{S}$.

Note that, in a finite semigroup S, for given $s \in S$, there must be some repetition in the powers $s, s^2, s^3, \ldots$ and so there exist minimal positive integers k, ℓ such that $s^k = s^{k+\ell}$. Let n be the unique integer such that $k \leq n < k + \ell$ and ℓ divides n. Then the powers $s^k, s^{k+1}, \ldots, s^{k+\ell-1}$ constitute a cyclic group with idempotent s^n and generator s^{n+1}, whose inverse is s^{2n-1}. Since $s^{m!} = s^n$ for all $m \geq n$, it follows that s^ω is the unique idempotent which is a power of s (with positive exponent) and $s^{\omega-1}$ is the inverse of $s^{\omega+1}$ in the maximal subgroup with idempotent s^ω. Hence, in a profinite group G, we have the equality $g^{\omega-1} = g^{-1}$.

From the above unary implicit operations and multiplication one may already easily construct lots of implicit operations such as $x^\omega y^\omega$, $(x^{\omega+1}y^{\omega+1})^\omega$, and the *commutator* $[x, y] = x^{\omega-1}y^{\omega-1}x^{\omega+1}y^{\omega+1}$. These examples illustrate how implicit operations are composed: given m-ary implicit operations $w_1, \ldots, w_n \in \overline{\Omega}_m\mathsf{V}$ and an n-ary implicit operation $v \in \overline{\Omega}_n\mathsf{V}$, the natural interpretation of v in $\overline{\Omega}_m\mathsf{V}$ allows us to define $v(w_1, \ldots, w_n) \in \overline{\Omega}_m\mathsf{V}$ to be the operation $v_{\overline{\Omega}_m\mathsf{V}}(w_1, \ldots, w_n)$. In particular, multiplication of implicit operations is obtained by applying the binary explicit operation x_1x_2 to two given operations. An important property of composition is that it is continuous.

4.7 Proposition *Composition of implicit operations of fixed arity as defined above is a con-*

tinuous function

$$\begin{aligned}\overline{\Omega}_n\mathsf{V} \times (\overline{\Omega}_m\mathsf{V})^n &\to \overline{\Omega}_m\mathsf{V}\\ (v, w_1, \ldots, w_n) &\mapsto v(w_1, \ldots, w_n)\end{aligned}$$

Proof Given a clopen subset $K \subseteq \overline{\Omega}_m\mathsf{V}$, by Proposition 3.5 there exists a continuous homomorphism $\varphi : \overline{\Omega}_m\mathsf{V} \to S$ such that $S \in \mathsf{V}$ and $K = \varphi^{-1}\varphi K$. Let $\mathsf{W} = \mathsf{V}(S)$. Then φ factors through the natural projection $\pi_m : \overline{\Omega}_m\mathsf{V} \to \overline{\Omega}_m\mathsf{W}$ and so we may as well assume that $S = \overline{\Omega}_m\mathsf{W}$. Consider the following diagram where $\pi_n : \overline{\Omega}_n\mathsf{V} \to \overline{\Omega}_n\mathsf{W}$ is also a natural projection and the horizontal arrows represent composition of implicit operations defined for each of the pseudovarieties V and W as above:

$$\begin{array}{ccc}\overline{\Omega}_n\mathsf{V} \times (\overline{\Omega}_m\mathsf{V})^n & \longrightarrow & \overline{\Omega}_m\mathsf{V}\\ \downarrow{\scriptstyle \pi_n\times(\pi_m)^n} & & \downarrow{\scriptstyle \pi_m}\\ \overline{\Omega}_n\mathsf{W} \times (\overline{\Omega}_m\mathsf{W})^n & \longrightarrow & \overline{\Omega}_m\mathsf{W}\end{array}$$

The commutativity of the diagram follows from the fact that implicit operations commute with continuous homomorphisms between pro-V semigroups:

$$\begin{aligned}\pi_m(v(w_1, \ldots, w_n)) &= \pi_m(v_{\overline{\Omega}_m\mathsf{V}}(w_1, \ldots, w_n))\\ &= v_{\overline{\Omega}_m\mathsf{W}}(\pi_m(w_1), \ldots, \pi_m(w_n)) = \pi_n(v)(\pi_m(w_1), \ldots, \pi_m(w_n))\end{aligned}$$

Since the bottom line is continuous, as it is a mapping between discrete spaces, it follows that the inverse image by the top line of the clopen set K is again clopen. □

The calculus of implicit operations of the form $x^{\omega+k}$ under the operations of multiplication and composition is quite simple.

4.8 Lemma *The set of unary implicit operations of the form $x^{\omega+k}$, with $k \in \mathbb{Z}$, constitutes a ring whose addition is multiplication of implicit operations and whose multiplication is composition of implicit operations. It is isomorphic to the ring $\mathbb{Z}$ of integers.*

Proof The result is immediate from the following equalities where we use continuity of composition as given by Proposition 4.7:

$$\begin{aligned}x^{\omega+k}x^{\omega+\ell} &= \lim_{n\to\infty} x^{n!+k}x^{n!+\ell} = \lim_{n\to\infty} x^{2(n!)+k+\ell} = \lim_{n\to\infty} x^{n!+k+\ell} = x^{\omega+k+\ell}\\ (x^{\omega+k})^{\omega+\ell} &= \lim_{n\to\infty} (x^{n!+k})^{n!+\ell} = \lim_{n\to\infty} x^{(n!)^2+(k+\ell)(n!)+k\ell} = \lim_{n\to\infty} x^{n!+k\ell} = x^{\omega+k\ell}\end{aligned}$$

□

4.9 Remark Let $\hat{\mathbb{Z}}$ be the profinite completion of the ring $\mathbb{Z}$ of integers. It may be obtained as the completion of $\mathbb{Z}$ with respect to the metric d of Subsection 3.4 defined similarly in the language of rings. Since $\mathbb{Z}$ is the free commutative ring on one generator and it is residually finite, $\hat{\mathbb{Z}}$ is isomorphic to $\overline{\Omega}_1\mathsf{R}$ for the pseudovariety R of finite commutative rings. It follows that the non-explicit unary implicit operations constitute a ring under multiplication and composition which is isomorphic to the completion $\hat{\mathbb{Z}}$. This completion can be easily seen to be the direct product of the p-adic completions $\mathbb{Z}_p$ of the ring $\mathbb{Z}$ as p runs over all prime numbers.

We are now ready to present some examples regarding Reiterman's Theorem.

4.10 Examples

(1) The pseudovariety A of all finite aperiodic semigroups is defined by the pseudoidentity $x^{\omega+1} = x^\omega$ since all subgroups of a semigroup are trivial if and only if all its cyclic subgroups are trivial.

(2) The pseudovariety N of all finite nilpotent semigroups is defined by the pseudoidentities $x^\omega y = yx^\omega = x^\omega$ which we abbreviate as $x^\omega = 0$.

(3) The pseudovariety G of all finite groups is defined by the pseudoidentities $x^\omega y = yx^\omega = y$ which we naturally abbreviate as $x^\omega = 1$.

(4) For a prime p, the pseudovariety G_p of all finite p-groups cannot be defined by pseudoidentities involving only products and the operation $x^{\omega-1}$. Indeed, since $\mathsf{G} \models x^{\omega-1} = x^{-1}$, all such pseudoidentities may be viewed as ordinary identities of group words and it is well known that free groups are residually G_p [28].

4.3 Iteration of implicit operations

The last example from the previous subsection shows that one needs a richer language of implicit operations than that provided by multiplication plus the unary operations of the form $x^{\omega+k}$ to define some pseudovarieties in terms of pseudoidentities. A powerful tool for constructing implicit operations is infinite iteration of composition. In this subsection, we develop a more general framework where not only infinite iteration but arbitrary implicit operations may be performed, namely we show that the monoid of continuous endomorphisms of a finitely generated profinite semigroup is profinite. Our arguments are not the most economical but the end result is the best which is presently known.

For a topological semigroup S, denote by $\operatorname{End} S$ the monoid of its continuous endomorphisms. Note that $\operatorname{End} S$ is a subset of the set of functions S^S. In general it is a delicate question which topology to consider on a function space. For S^S the two most natural alternatives are the product topology, also known as the *topology of pointwise convergence*, and the *compact-open topology*, for which a subbase consists of the sets of functions of the form

$$V(K,U) = \{f \in S^S : f(K) \subseteq U\}$$

where $K \subseteq S$ is compact and $U \subseteq S$ is open [93]. These topologies retain their names when the induced topologies are considered on subspaces of S^S. Note that a subbase for the product topology is given by the sets of the form $V(\{s\}, U)$ with $s \in S$ and $U \subseteq S$ open and so the pointwise convergence topology is contained in the compact-open topology.

For the sequel, we require the following very simple yet very useful observation.

4.11 Lemma *Let S be a profinite semigroup and let d be the natural metric on S. Then every $f \in \operatorname{End} S$ is a contracting function in the sense that, for all $s_1, s_2 \in S$,*

$$d(f(s_1), f(s_2)) \leq d(s_1, s_2). \tag{4.2}$$

Proof If $s_1 = s_2$ then the inequality (4.2) is obvious since both sides are zero. Otherwise, let $n = r(s_1, s_2)$, as defined in Subsection 3.4. Given a continuous homomorphism $\varphi : S \to T$ into a finite semigroup with $|T| < n$, the composite $\varphi \circ f : S \to T$ is again a continuous homomorphism and so, by definition of $r(s_1, s_2)$, we have $\varphi(f(s_1)) = \varphi(f(s_2))$. Hence $r(f(s_1), f(s_2)) \geq n$, which proves (4.2). □

The pointwise convergence topology has the advantage of being easier to handle in terms of convergence since a net converges in it if and only if it converges pointwise. The following result is an illustration of this fact.

4.12 Lemma *If S is a finitely generated profinite semigroup then* $\operatorname{End} S$ *is compact with respect to the pointwise convergence topology.*

Proof For each $s, t \in S$, the set $\{\varphi \in S^S : \varphi(st) = \varphi(s)\varphi(t)\}$ is closed since the equation that defines it only involves three components of the product, namely the components indexed by st, s, and t and the restriction on those three components, as a subset of S^3, is the graph of multiplication, which is assumed to be continuous. Hence the monoid of (not necessarily continuous) endomorphisms of S is closed in S^S with respect to the pointwise convergence topology.

Next, let $(f_i)_{i \in I}$ be a net in $\operatorname{End} S$ converging to some $f \in S^S$. Given $s, t \in S$, then by continuity of the natural metric d, we have $d(f(s), f(t)) = \lim_{i \in I} d(f_i(s), f_i(t))$ while $d(f_i(s), f_i(t)) \leq d(s, t)$ by Lemma 4.11. Hence $d(f(s), f(t)) \leq d(s, t)$ which, in view of Proposition 3.9, shows that f is continuous. Since f is an endomorphism of S by the preceding paragraph, we conclude that $f \in \operatorname{End} S$. Thus $\operatorname{End} S$ is a closed subset of the compact space S^S and, therefore, $\operatorname{End} S$ is compact. □

On the other hand, it is well known that for instance for locally compact S, the compact-open topology on $\operatorname{End} S$ implies continuity of the evaluation mapping

$$\begin{aligned} \varepsilon : (\operatorname{End} S) \times S &\to S \\ (f, s) &\mapsto f(s) \end{aligned}$$

Indeed, for an open subset $U \subseteq S$ and $(f, s) \in (\operatorname{End} S) \times S$ such that $f(s) \in U$, since f is continuous and S is locally compact, there is some compact neighbourhood K of s such that $f(K) \subseteq U$. Then $(V(K, U) \cap \operatorname{End} S) \times K$ is a neighbourhood of (f, s) in $(\operatorname{End} S) \times S$ which is contained in $\varepsilon^{-1}(U)$. See [93] for more general results and background.

The comparison between the two topologies on $\operatorname{End} S$ is given by the following result.

4.13 Proposition *Let S be a finitely generated profinite semigroup. Then the pointwise convergence and compact-open topologies coincide on* $\operatorname{End} S$.

Proof It remains to show that every set of the form $V(K, U) \cap \operatorname{End} S$ with $K \subseteq S$ compact and $U \subseteq S$ open is open in the pointwise convergence topology of $\operatorname{End} S$. Without loss of generality, since S is zero-dimensional we may assume that U is clopen. Consider the open cover $U \cup (S \setminus U)$. By Proposition 3.9, this is also an open cover in the topology induced by the natural metric d. By Lebesgue's Covering Lemma [93, Theorem 22.5], there exists $\delta > 0$ such that every subset of S of diameter less than δ is contained in either U or $S \setminus U$.

Consider the cover of K by the open balls $B_\delta(s)$ of radius δ centered at points $s \in K$. By Proposition 3.9, these balls are open in the topology of S and so, since K is compact, there are finitely many points $s_1, \ldots, s_n \in K$ such that

$$K \subseteq \bigcup_{i=1}^{n} B_\delta(s_i). \tag{4.3}$$

Let $f \in V(K,U) \cap \operatorname{End} S$ and let

$$W = \bigcap_{i=1}^{n} V(\{s_i\}, B_\delta(f(s_i))) \cap \operatorname{End} S.$$

Then W is an open set of $\operatorname{End} S$ in the pointwise convergence topology which contains f. Hence it suffices to show that $W \subseteq V(K,U)$. Let $g \in W$ and $s \in K$. By (4.3), there exists $i \in \{1, \ldots, n\}$ such that $d(s, s_i) < \delta$. By Lemma 4.11, we have $d(g(s), g(s_i)) < \delta$. On the other hand, since $g \in V(\{s_i\}, B_\delta(f(s_i)))$, we obtain $d(g(s_i), f(s_i)) < \delta$. Since d is an ultrametric, it follows that $d(g(s), f(s_i)) < \delta$. Finally, as $f(s_i) \in U$ since $f \in V(K,U)$, we conclude by the choice of δ that $g(s) \in U$, which shows that $g \in V(K,U)$. □

For the case of finitely generated relatively free profinite semigroups, the following result was already observed in [19] as a consequence of a result from [6]. We provide here a direct proof of the more general case without the assumption of relative freeness.

4.14 Theorem *Let S be a finitely generated profinite semigroup. Then $\operatorname{End} S$ is a profinite monoid under the pointwise convergence topology and the evaluation mapping is continuous.*

Proof By Proposition 4.13, the pointwise convergence and compact-open topologies coincide on $\operatorname{End} S$. Hence the evaluation mapping $\varepsilon : (\operatorname{End} S) \times S \to S$ is continuous.

Since S is totally disconnected, so is the product space S^S and its subspace $\operatorname{End} S$. Moreover, $\operatorname{End} S$ is compact by Lemma 4.12. Hence, by Theorem 3.1, to prove that $\operatorname{End} S$ is a profinite monoid it suffices to show that it is a topological monoid, that is that the composition $\mu : (\operatorname{End} S) \times (\operatorname{End} S) \to \operatorname{End} S$ is continuous, for which we show that, for every convergent net $(f_i, g_i) \to (f, g)$ in $(\operatorname{End} S) \times (\operatorname{End} S)$, we have $f_i \circ g_i \to f \circ g$. For the pointwise convergence topology the latter means $f_i(g_i(s)) \to f(g(s))$ for every $s \in S$. This follows from $f_i \to f$ together with $g_i(s) \to g(s)$ since the evaluation mapping is continuous. □

Thus, if S is a finitely generated profinite semigroup then the group $\operatorname{Aut} S$ of its continuous automorphisms is a profinite group since it is a closed subgroup of $\operatorname{End} S$ (cf. Corollary 3.2). In particular, the group $\operatorname{Aut} G$ of continuous automorphisms of a finitely generated profinite group G is profinite for the pointwise convergence topology, a result which is useful in profinite group theory [76]. Moreover, since one can find in [76] examples of profinite groups whose groups of continuous automorphisms are not profinite, we see that the hypothesis that S is finitely generated cannot be removed from Theorem 4.14.

Now, for a finite set A, $\overline{\Omega}_A\mathsf{S}$ is a finitely generated profinite semigroup and so $\operatorname{End}\overline{\Omega}_A\mathsf{S}$ is a profinite monoid. Since $\overline{\Omega}_A\mathsf{S}$ is a free profinite semigroup on the generating set A, continuous endomorphisms of $\overline{\Omega}_A\mathsf{S}$ are completely determined by their restrictions to A. For $\overline{\Omega}_n\mathsf{S}$ we may then choose to represent an element $\varphi \in \operatorname{End}\overline{\Omega}_n\mathsf{S}$ by the n-tuple $(\varphi(x_1), \ldots, \varphi(x_n))$ where $x_1, \ldots, x_n$ are the component projections. When n is small, we may write $x, y, z, t, \ldots$ or $a, b, c, d, \ldots$ instead of $x_1, x_2, x_3, x_4, \ldots$ respectively.

In the case $n = 1$, take $\varphi \in \operatorname{End}\overline{\Omega}_1\mathsf{S}$ determined by the 1-tuple (x^m). Then φ has an ω-power in $\operatorname{End}\overline{\Omega}_1\mathsf{S}$ and we may consider the operation $\varphi^\omega(x)$ which we denote x^{m^ω} since

$$x^{m^\omega} = \varphi^\omega(x) = \lim_{k\to\infty} \varphi^{k!}(x) = \lim_{k\to\infty} x^{m^{k!}}.$$

4.15 Examples

(1) It is now an easy exercise, which we leave to the reader, to show that, for a prime p, $\mathsf{G}_p = [\![x^{p^\omega} = 1]\!]$.

(2) Let G_{nil} denote the pseudovariety of all finite nilpotent groups. Consider the endomorphism $\varphi \in \operatorname{End}\overline{\Omega}_2\mathsf{S}$ defined by the pair $([x, y], y)$ where $[x, y]$ denotes the commutator defined earlier. Denote by $[x, {}_\omega y]$ the implicit operation $\varphi^\omega(x)$. Then it follows from a theorem of Zorn [94, 77] that $\mathsf{G}_{\text{nil}} = [\![\, [x, {}_\omega y] = 1]\!]$.

(3) Let G_{sol} denote the pseudovariety of all finite solvable groups. Let $\varphi \in \operatorname{End}\overline{\Omega}_3\mathsf{S}$ be defined by the triple $([yxy^{\omega-1}, zxz^{\omega-1}], y, z)$ and let $w = \varphi^\omega(x)$, which is a ternary implicit operation. Using J. Thompson's list of minimal non-solvable simple groups, arithmetic geometry and computer algebra and geometry, Bandman et al [27] have recently established that $\mathsf{G}_{\text{sol}} = [\![w(x^{\omega-2}y^{\omega-1}x, x, y) = 1]\!]$. Since this provides a two-variable pseudoidentity basis for G_{sol}, as an immediate corollary one obtains the Thompson-Flavell Theorem stating that a finite group is solvable if and only if its 2-generated subgroups are solvable. The fact that the pseudovariety G_{sol} is finitely based, and therefore it may be defined by a single pseudoidentity of the form $v = 1$, was previously proved by Lubotzky [56].

Having established the foundations of the theory of profinite semigroups, the remainder of these notes is dedicated to surveying some results in the area, which are meant to introduce the reader to recent developments and reveal the richness and depthness of the already existing theory. Most proofs will be omitted. Naturally, this survey will not be exhaustive and it unavoidably reflects the author's personal preferences and tastes.

5 Free pro-J semigroups and Simon's Theorem

For details and proofs of the results in this section, see [3].

Recall that J denotes the pseudovariety consisting of all finite semigroups in which every principal ideal admits a unique element as its generator. The letter J comes from Green's relation $\mathcal{J}$ which relates two elements of a semigroup if they generate the same principal ideal. Hence J consists of all finite $\mathcal{J}$-trivial semigroups, that is finite semigroups in which the relation $\mathcal{J}$ is trivial. It is an exercise to show that

$$\mathsf{J} = [\![(xy)^\omega = (yx)^\omega, x^{\omega+1} = x^\omega]\!] = [\![(xy)^\omega x = (xy)^\omega = y(xy)^\omega]\!]. \tag{5.1}$$

Since $\mathsf{J} \supseteq \mathsf{N}$, J satisfies no nontrivial semigroup identities and so $\Omega_A\mathsf{J} \simeq A^+$ is the free semigroup on A.

We recall next a theorem which was already mentioned in Section 2. A language $L \subseteq A^+$ is said to be *piecewise testable* if it is a Boolean combination of languages of the form

$$A^*a_1A^* \cdots a_nA^*. \tag{5.2}$$

The word $a_1 \cdots a_n$ is said to be a *subword* of an element $w \in \overline{\Omega}_A\mathsf{S}$ if w belongs to the closure of the language (5.2).

By taking Boolean combinations of languages of the form (5.2) we may obtain, for each positive integer N, the classes of the congruence $\sim_N$ on A^+ which identifies two words if they have precisely the same subwords of length at most N. Conversely, a language of the form (5.2) is saturated by the congruence $\sim_N$. Hence a language $L \subseteq A^+$ is piecewise testable if and only if it is saturated by some congruence $\sim_N$.

The language (5.2) is J-recognizable: in view of the preceding paragraph, this can be proved by noting that for words u and v, the words $(uv)^N$ and $(vu)^N$ are in the same $\sim_N$-class, and the same holds for the words u^{N+1} and u^N, which establishes that the quotient semigroup $A^+/\sim_N$ satisfies the first set of pseudoidentities in (5.1).

5.1 Theorem (Simon [80]) *A language $L \subseteq A^+$ over a finite alphabet is piecewise testable if and only if it is J-recognizable.*

In terms of Eilenberg's correspondence, Simon's Theorem says that the variety of languages associated with J is the variety of piecewise testable languages. In topological terms, Simon's Theorem says that the closures in $\overline{\Omega}_A\mathsf{J}$ of the languages of the form (5.2) suffice to separate points (cf. Proposition 3.8). In other words, implicit operations over J can be distinguished by looking at their subwords. This is certainly not true for implicit operations over S as for instance $(xy)^\omega$ and $(yx)^\omega$ have the same subwords.

So, it should be possible to prove Simon's Theorem by developing a good understanding of the structure of the finitely generated free pro-J semigroups $\overline{\Omega}_A\mathsf{J}$. This program was carried out by the author in the mid-1980's motivated by a question raised by I. Simon as to whether $\overline{\Omega}_A\mathsf{J}$ is countable.

Recall that an element s of a semigroup S is said to be *regular* if there exists $t \in S$ such that $sts = s$; such an element t is called a *weak inverse* of s. Then $u = tst$ is such that $sus = s$ and $usu = u$; an element $u \in S$ satisfying these two equalities is called an *inverse* of s. Semigroups in which every element has a unique inverse are called *inverse semigroups*. They are precisely the regular semigroups of partial bijections of sets and play an important role in various applications [54].

If s and t are inverses in a semigroup S then st is an idempotent and $s \mathcal{J} st$ and so regular elements are $\mathcal{J}$-equivalent to idempotents. Hence in a $\mathcal{J}$-trivial semigroup the regular elements are the idempotents.

For words $u, v \in A^+$, one can use the pseudoidentities in (5.1) to deduce that J satisfies the pseudoidentity $u^\omega = v^\omega$ if and only if u and v contain the same letters.

Consider the semilattice $\mathcal{P}(A)$ of subsets of A under union. Since it is $\mathcal{J}$-trivial, the function $A^+ \to \mathcal{P}(A)$ which associates with a word u the set of letters occurring in it extends uniquely to a continuous homomorphism $c : \overline{\Omega}_A\mathsf{J} \to \mathcal{P}(A)$. We call it the *content* function. The result in the preceding paragraph may then be stated by saying that an idempotent in $\overline{\Omega}_A\mathsf{J}$ of the form u^ω is completely characterized by its content $c(u^\omega) = c(u)$. Since every element $v \in \overline{\Omega}_A\mathsf{J}$ is the limit of a sequence of words $(v_n)_n$ and we may assume that $(c(v_n))_n$ is constant, if v is idempotent then

$$v = v^n = v^{n!} = \lim_{n\to\infty} v^{n!} = v^\omega = \lim_{n\to\infty} v_n^\omega = \lim_{n\to\infty} u^\omega = u^\omega$$

where u is any word with $c(u) = c(v)$.

Now, idempotents of $\overline{\Omega}_A\mathsf{J}$ may be characterized by the property that whenever a word u is a subword then so is u^n for every $n \geq 1$. This leads to the following result.

5.2 Theorem *Every element of $\overline{\Omega}_A\mathsf{J}$ admits a factorization of the form $u_0 v_1^\omega u_1 \cdots v_n^\omega u_n$ where the u_i, v_i are words, with the u_i possibly empty and each v_i with no repeated letters. Furthermore, one may arrange for the following to hold:*

- *no u_i ends with a letter occurring in v_{i+1};*
- *no u_i starts with a letter occurring in v_i;*
- *if u_i is the empty word, then the contents $c(v_i)$ and $c(v_{i+1})$ are not comparable under inclusion.*

Theorem 5.2 answers Simon's question: for A finite, the semigroup $\overline{\Omega}_A\mathsf{J}$ is countable and it is in fact generated by A together with its $2^{|A|} - 1$ idempotents. This suggests viewing $\overline{\Omega}_A\mathsf{J}$ as an algebra of type $(2, 1)$ under multiplication and the ω-power. The semigroup $\overline{\Omega}_A\mathsf{J}$ becomes a free algebra in the variety of algebras of this type generated by J and Theorem 5.2 suggests a canonical form for terms in this free algebra. First, Theorem 5.2 already implies that every such term is equal in $\overline{\Omega}_A\mathsf{J}$ to a term of the form described in the theorem. To distinguish terms in canonical form, one may use subwords and thus prove at the same time Simon's Theorem.

5.3 Theorem *Two terms in the form described in Theorem* 5.2 *are distinct in $\overline{\Omega}_A\mathsf{J}$ if and only if they have distinct sets of subwords.*

One may more precisely bound the length of the subwords which are necessary to distinguish two such terms. Recently, in work whose precise connection with the above remains to be determined, Simon [81] has proposed a very efficient algorithm to distinguish two words by their subwords. For more on the significance in Mathematics and Computer Science of Simon's Theorem, see the lecture notes of M. V. Volkov in this volume.

6 Tameness

This section goes deeper into decidability problems for pseudovarieties. It grows out of [16, 15, 7]. The reader is referred to those publications for details.

Let Σ be a finite system of equations of the form $u = v$ with $u, v \in X^+$. For example such a system might be

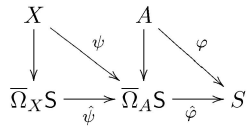
$$\begin{cases} xy = z, \\ zx = ty. \end{cases}$$

We impose on the variables *rational constraints*: for each $x \in X$, we choose a rational language $L_x \subseteq A^+$ where A is another alphabet. A *solution* of the system in an A-generated profinite semigroup $\varphi : A \to S$ is a mapping $\psi : X \to \overline{\Omega}_A\mathsf{S}$ such that

(1) $\psi(x) \in \overline{L_x}$ for every variable $x \in X$,

(2) $\hat{\varphi} \circ \hat{\psi}(u) = \hat{\varphi} \circ \hat{\psi}(v)$ for every equation $(u = v) \in \Sigma$,

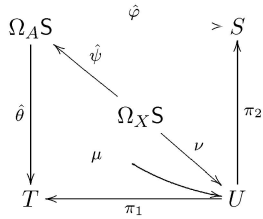
where the various mappings are depicted in the following commutative diagram.



One may compute from the constraints a finite semigroup T and a homomorphism $\theta : A^+ \to T$ recognizing all of them. Let U be the closed subsemigroup of $T \times S$ generated by all elements of the form $\mu(a) = (\theta(a), \varphi(a))$ with $a \in A$, and let $\mu : \overline{\Omega}_A\mathsf{S} \to U$ be the induced continuous homomorphism. Then the above conditions (1) and (2) may be formulated in terms of the composite $\nu = \mu\hat{\psi}$ by stating the following, where $\pi_1 : U \to T$ and $\pi_2 : U \to S$ are the component projections:

(1) $\pi_1\nu(x) \in \pi_1\mu L_x$ for every variable $x \in X$;

(2) $\pi_2\nu(u) = \pi_2\nu(v)$ for every equation $(u = v) \in \Sigma$.



In particular, if S is finite then one can test effectively whether a solution exists in S.

The semigroup U is an example of what is called a "relational morphism". More generally, a *relational morphism* between two topological semigroups S and T is a relation $\tau : S \to T$ with domain S which is a closed subsemigroup of the product $S \times T$. A continuous homomorphism and the inverse image of an onto continuous homomorphism are relational morphisms and every relational morphism may be obtained by composition of two such relational morphisms. Relational morphisms for monoids are defined similarly.

The following is a compactness result whose proof may be obtained by following basically the same lines as in the proof of the equivalence (1) $\Leftrightarrow$ (2) in Proposition 3.5.

6.1 Theorem *The following conditions are equivalent for a finite system Σ of equations with rational constraints over the finite alphabet A:*

(1) *Σ has a solution in every A-generated semigroup from V;*

(2) *Σ has a solution in every A-generated pro-V semigroup;*

(3) *Σ has a solution in $\overline{\Omega}_A\mathsf{V}$.*

A system satisfying the conditions of Theorem 6.1 is said to be V-*inevitable*. A decidability property for a pseudovariety V with respect to a given recursively enumerable set $\mathcal{C}$ of such systems is whether there is an algorithm to decide whether a given $\Sigma \in \mathcal{C}$ is V-inevitable. Before examining the relevance of such a property, we introduce a few more precise notions.

An *implicit signature* is a set σ of implicit operations (over S) which contains multiplication. It is viewed as an enlarged algebraic language for which profinite semigroups immediately inherit a natural structure by giving the chosen implicit operations their natural interpretation. Note that the subalgebra $\Omega_A^\sigma \mathsf{V}$ of $\overline{\Omega}_A \mathsf{V}$ generated by A is precisely the free σ-algebra in the variety generated by V.

The signature $\kappa = \{\_ \cdot \_,\ \_^{\omega-1}\}$ is called the *canonical signature* since most implicit operations which are commonly used are terms in its language. For finite groups, it becomes the natural signature, with multiplication and inversion. In particular, since free groups are residually finite, $\Omega_A^\kappa \mathsf{G}$ is the free group on A.

Say V is $\mathcal{C}$-*tame* if there is an implicit signature σ such that

(1) σ is recursively enumerable;

(2) the operations in σ are computable;

(3) the word problem for $\Omega_A^\sigma \mathsf{V}$ is decidable;

(4) for every V-inevitable Σ there is a solution $\psi : X \to \overline{\Omega}_A \mathsf{S}$ for $\overline{\Omega}_A \mathsf{V}$ which takes its values in $\Omega_A^\sigma \mathsf{S}$.

Under the above conditions, we may also say that V is $\mathcal{C}$-tame *with respect to* σ. In case $\mathcal{C}$ consists of all finite systems over a fixed countable alphabet, then we say that V is *completely tame* if it is $\mathcal{C}$-tame.

6.2 Theorem *Let $\mathcal{C}$ be a recursively enumerable set of finite systems of equations with rational constraints and suppose V is a $\mathcal{C}$-tame pseudovariety. Then it is decidable whether a given $\Sigma \in \mathcal{C}$ is V-inevitable.*

Proof Let V be $\mathcal{C}$-tame with respect to an implicit signature σ. To prove the theorem it suffices to effectively enumerate those $\Sigma \in \mathcal{C}$ that are V-inevitable and those that are not.

One can start by enumerating all systems $\Sigma \in \mathcal{C}$. Since V is $\mathcal{C}$-tame with respect to σ, if Σ is V-inevitable then there is a solution $\psi : X \to \overline{\Omega}_A \mathsf{S}$ for Σ in $\overline{\Omega}_A \mathsf{V}$ that takes its values in $\Omega_A^\sigma \mathsf{S}$. The candidates for such solutions can be effectively enumerated in parallel with the systems, as σ is recursively enumerable, the constraints can be effectively tested by computing operations in their syntactic semigroups, and the equations can be effectively tested by using a solution of the word problem for $\Omega_A^\sigma \mathsf{V}$. This provides an effective enumeration of all V-inevitable systems in $\mathcal{C}$.

To enumerate those systems in $\mathcal{C}$ that are not V-inevitable, we try out pairs of systems Σ in $\mathcal{C}$ together with candidates for A-generated semigroups $S \in \mathsf{V}$. We already observed that under these conditions one can test effectively whether a solution exists in S and if turns out it does not, we output the system as one that is not V-inevitable. □

One class of systems which the author has introduced for the study of semidirect products of pseudovarieties (cf. Section 7) consists of systems associated with finite directed graphs: to each vertex and edge in the graph one associates a variable and an equation $xy = z$ is written if y is the variable corresponding to an edge which goes from the vertex corresponding to x to the vertex corresponding to z. We will call a pseudovariety *graph-tame* if it is tame with respect to systems of equations arising in this way.

Here are some examples of tame pseudovarieties.

6.3 Example The pseudovarieties N and J are completely tame with respect to the canonical signature κ. Both results follow from the knowledge of the structure of the corresponding relatively free profinite semigroups and the solution of their word problems. In the case of N this is quite simple. In fact, $\overline{\Omega}_A\mathsf{N}$ is obtained from the free semigroup $\Omega_A\mathsf{N} \simeq A^+$ by adjoining a zero element, which is topologically the one-point compactification for the discrete topology on $\Omega_A\mathsf{N}$. Hence a κ-term is zero in $\overline{\Omega}_A\mathsf{N}$ if and only if it involves the $(\omega-1)$-power. Assuming a finite system has a solution in $\overline{\Omega}_A\mathsf{N}$, given by a function $\psi : X \to \overline{\Omega}_A\mathsf{S}$, we modify ψ on those variables x for which $\psi(x)$ is not explicit as follows. From Theorem 6.1, it follows that there exists a factorization $\psi(x) = uv^\omega w$ with $u, v, w \in \overline{\Omega}_A\mathsf{S}$. Since the constraint for x translates into a condition of the form $\psi(x)$ *belongs to a given clopen subset of* $\overline{\Omega}_A\mathsf{S}$ and $\psi(x)$ is zero in $\overline{\Omega}_A\mathsf{N}$, we may replace u, v, w by words. This changes $\psi(x)$ to a κ-term by maintaining a solution in $\overline{\Omega}_A\mathsf{N}$. See [7] for details.

6.4 Theorem (Almeida and Delgado [13]) *The pseudovariety* Ab *of all finite Abelian groups is completely tame with respect to* κ*.*

The proof of Theorem 6.4 amounts to linear algebra over the profinite completion $\hat{\mathbb{Z}}$ of the ring of integers, which we have already observed to be isomorphic with $\overline{\Omega}_1\mathsf{G}$ under multiplication and composition. From work of Steinberg [83] it follows that the pseudovariety Com of finite commutative semigroups is also competely tame with respect to κ.

6.5 Theorem (Ash [22]) *The pseudovariety* G *of all finite groups is graph-tame with respect to* κ*.*

Theorem 6.5 is considered one of the deepest results in finite semigroup theory. In its original version, it was proved by algebraic-combinatorial methods in a somewhat different language; see [5, 12] for a translation to the language of these notes. An independent proof of the case of 1-vertex graphs, which is already quite nontrivial was obtained using profinite group theory where the result translates to a conjecture which had been proposed by Pin and Reutenauer [66], namely the following statement, where by the *profinite topology* of the free group we mean the induced topology from the free profinite group or equivalently the topology whose open subgroups are those of finite index.

6.6 Theorem (Ribes and Zalesskiĭ [74]) *The product of finitely many finitely generated subgroups of the free group is closed in the profinite topology of the free group.*

Theorem 6.6 in turn generalizes a theorem of M. Hall [38] which is the case of just one subgroup. The interest in these results will be explained in more detail in Section 7. Finally, it follows from a result of Coulbois and Khélif [33] that G is not completely tame with respect to κ. At present it is not known whether G is completely tame with respect to some implicit signature.

There are some surprising connections of graph-tameness of G with other areas of Mathematics and in fact the result was rediscovered in disguise in Model Theory. We say that a class $\mathcal{R}$ of relational structures of the same type has the *finite extension property for partial automorphisms* (*FEPPA* for shortness) if for every finite $R \in \mathcal{R}$ and every set P of partial automorphisms of R, if there exists an extension $S \in \mathcal{R}$ of R for which every $f \in P$ extends to a total automorphism of S, then there exists such an extension $S \in \mathcal{R}$ which is finite. For a class $\mathcal{R}$ of relational structures, let $\operatorname{Excl}\mathcal{R}$ denote the class of all structures S for which there

is no homomorphism $R \to S$ with $R \in \mathcal{R}$ where by a *homomorphism* of relational structures of the same type we mean a function that preserves the relations in the forward direction. Now, we have the following remarkable statement.

6.7 Theorem (Herwig and Lascar [41]) *For every finite set $\mathcal{R}$ of finite structures of a finite relational language, Excl$\mathcal{R}$ satisfies the FEPPA.*

Herwig and Lascar recognized this result as extending Ribes and Zalesskiĭ's Theorem and provided a general translation into a property of the free group. Delgado and the author [11, 12] in turn recognized that property of the free group as being equivalent to Ash's Theorem.

The following two examples provide another two applications of Ash's Theorem which additionally use results from the theory of regular semigroups. See the quoted papers for appropriate references.

6.8 Example Let OCR be the class consisting of all finite semigroups S which are unions of their subgroups (in which case we say S is *completely regular*) and in which the products of idempotents are again idempotents (in which case S is said to be *orthodox*). In terms of pseudoidentities, we have $\mathsf{OCR} = [\![x^{\omega+1} = x,\ (x^\omega y^\omega)^2 = x^\omega y^\omega]\!]$. Trotter and the author [17] have used graph-tameness of G to show that OCR is also graph-tame with respect to the canonical signature κ.

6.9 Example Let $\mathsf{CR} = [\![x^{\omega+1} = x]\!]$ be the pseudovariety of all finite completely regular semigroups. Trotter and the author [18] have reduced the graph-tameness of CR to a property which was apparently stronger than graph-tameness of G but K. Auinger has observed that the methods of [11, 12] apply to show that the property in question follows from the graph-tameness of G.

Our final example comes from [6] which led the author to explore connections with dynamical systems.

The pseudovariety G_p of all finite p-groups is not graph-tame with respect to the signature κ since Steinberg and the author [15] have observed that if it were then G_p would be definable by identities in the signature κ, which we have already observed to be impossible. Nevertheless, based on work of Ribes and Zalesskiĭ [75], Margolis, Sapir and Weil [57], and Steinberg [84], the author has proved the following.

6.10 Theorem *It is possible to enlarge κ to an infinite signature σ so that G_p is graph-tame with respect to this signature* [6].

The added implicit operations are those of the form $\varphi^{\omega-1}(v_i)$ where $\varphi \in \operatorname{End} \overline{\Omega}_n \mathsf{S}$ is defined by the n-tuple $(w_1, \ldots, w_n)$ and the v_i and w_i are κ-terms such that $\mathsf{G}_p \models v_i = w_i$ and the determinant of the matrix $(|w_i|_{x_j})_{i,j}$ is invertible in $\mathbb{Z}/p\mathbb{Z}$. Here for a κ-term w and a letter x, $|w|_x$ is the integer obtained by viewing w as a group word and counting the signed number of occurrences of x in w. So, for example, if φ is given by the pair $((xy)^{\omega-1}yx^\omega, x^3yx^{\omega-1})$ then we get $\det \left(\begin{smallmatrix} -1 & 0 \\ 2 & 1 \end{smallmatrix}\right) = -1$. We do not know if G_p is completely tame with respect to this signature.

7 Categories, semigroupoids and semidirect products

Tilson [90] introduced pseudovarieties of categories as the foundations of an approach to the calculation of semidirect products of pseudovarieties of semigroups which had emerged earlier from work of Knast [50], Straubing [86], and Thérien and Weiss [88, 92]. A *pseudovariety of categories* is defined to be a class of finite categories which is closed under taking finite products and "divisors". A *divisor* of a category C is a category D for which there exists a category E and two functors: $E \to C$, which is injective on Hom-sets, and $E \to D$, which is onto and also injective when restricted to objects.

Jones [45] and independently Weil and the author [20] have extended the profinite approach to the realm of pseudovarieties of categories. Thus one can talk of relatively free profinite categories, implicit operations, and pseudoidentities. Instead of an unstructured set, to generate a category one takes a directed graph. Thus relatively free profinite categories are freely generated by directed graphs, implicit operations act on graph homomorphisms from fixed directed graphs into categories, and pseudoidentities are written over finite directed graphs. The free profinite category on a graph Γ will be denoted $\overline{\Omega}_\Gamma \mathsf{Cat}$. A pseudoidentity $(u = v; \Gamma)$ over the graph Γ is given by two coterminal morphisms $u, v \in \overline{\Omega}_\Gamma \mathsf{Cat}$. Examples will be presented shortly.

The morphisms from an object in a category C into itself constitute a monoid which is called a *local submonoid* of C. On the other hand, every monoid M may be viewed as a category by adding a virtual object and considering the elements of M as the morphisms, which are composed as they multiply in M. Note that the notion of division of categories applied to monoids is equivalent to the notion of division of monoids as introduced earlier.

For a pseudovariety V of finite monoids, $g\mathsf{V}$ denotes the (*global*) pseudovariety of categories generated by V and $\ell\mathsf{V}$ denotes the class of all finite categories whose local submonoids lie in V. Note that $g\mathsf{V}$ and $\ell\mathsf{V}$ are respectively the smallest and the largest pseudovarieties of categories whose monoids are those of V. The pseudovariety V is said to be *local* if $g\mathsf{V} = \ell\mathsf{V}$.

If Σ is a basis of monoid pseudoidentities for V, then the members of Σ may be viewed as pseudoidentities over (virtual) 1-vertex graphs; the resulting set of category pseudoidentities defines $\ell\mathsf{V}$. So, $\ell\mathsf{V}$ is easy to "compute" in terms of a basis of pseudoidentities. In general $g\mathsf{V}$ is much more interesting in applications and also much harder to compute. Since the problem of computing $g\mathsf{V}$ becomes simple if V is local, this explains the interest in locality results.

With appropriate care, the theory of pseudovarieties of categories may be extended to pseudovarieties of *semigroupoids*, meaning categories without the requirement for local identities [20]. Again we will move from one context to the other without further warning.

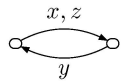
7.1 Examples

(1) The pseudovariety $\mathsf{Sl} = [\![xy = yx,\ x^2 = x]\!]$ of finite semilattices is local as was proved by Brzozowski and Simon [30].

(2) Every pseudovariety of groups is local [86, 90].

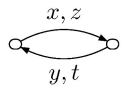
(3) The pseudovariety $\mathsf{Com} = [\![xy = yx]\!]$ is not local and its global is defined by the

pseudoidentity $xyz = zyx$ over the following graph:



This was proved by Thérien and Weiss [88].

(4) The pseudovariety J of finite $\mathcal{J}$-trivial semigroups is not local. Its global is defined by the pseudoidentity $(xy)^\omega xt(zt)^\omega = (xy)^\omega(zt)^\omega$ over the following graph:



This result which has many important applications is also considered to be rather difficult. It was discovered and proved by Knast [50]. A proof using the structure of the free pro-J semigroups can be found in [3]. The application that motivated the calculation of $g\,\mathsf{J}$ was the identification of dot-depth one languages [49] according to a natural hierarchy of plus-free languages introduced by Brzozowski [29]. The work of Straubing [86] was also concerned with the same problem. The computation of levels 2 and higher of this hierarchy remains an open problem.

(5) The pseudovariety $\mathsf{DA} = [\![\, ((xy)^\omega x)^2 = (xy)^\omega x \,]\!]$ of all finite semigroups whose regular elements are idempotents is local [4]. This result, which was proved using profinite techniques, turns out to have important applications in temporal logic [89]. See [87] for further relevance of the pseudovariety DA in various aspects of Computer Science.

See also [4] for further references to locality results.

Let S and T be semigroups and let $\varphi : T^1 \to \operatorname{End} S$ be a monoid homomorphism. For $t \in T^1$ and $s \in S$, denote $\varphi(t)(s)$ by ${}^t s$. Then the formula

$$(s_1, t_1)(s_2, t_2) = (s_1\ {}^{t_1}s_2, t_1t_2)$$

defines an associative multiplication on the set $S \times T$; the resulting semigroup is called a *semidirect product* of S and T and it is denoted $S *_\varphi T$ or simply $S * T$. Given two pseudovarieties of semigroups V and W, we denote by $\mathsf{V} * \mathsf{W}$ the pseudovariety generated by all semidirect products of the form $S * T$ with $S \in \mathsf{V}$ and $T \in \mathsf{W}$, which we also call the *semidirect product* of V and W. It is well known that the semidirect product of pseudovarieties is associative, see for instance [90] or [3].

The semidirect product is a very powerful operation. The following is a decomposition result which deeply influenced finite semigroup theory.

7.2 Theorem (Krohn and Rhodes [51]) *Every finite semigroup lies in one of the alternating semidirect products*

$$\mathsf{A} * \mathsf{G} * \mathsf{A} * \cdots * \mathsf{G} * \mathsf{A}. \tag{7.1}$$

Since the pseudovarieties of the form (7.1) form a chain, every finite semigroup belongs to most of them. The least number of factors G which is needed for the pseudovariety (7.1) to contain a given semigroup S is called the *complexity* of S. Although various announcements have been made of proofs that this complexity function is computable, at present there is yet no correct published proof. This has been over the past 40 years a major driving force in the development of finite semigroup theory, the original motivations being again closely linked with Computer Science namely aiming at the effective decomposition of automata and other theoretical computing devices [35, 36].

One approach to compute complexity is to study more generally the semidirect product operation and try to devise a general method to "compute" $\mathsf{V} * \mathsf{W}$ from V and W. What is meant by *computing a pseudovariety* is to exhibit an algorithm to test membership in it; in case such an algorithm exists, we will say, as we have done earlier, that the pseudovariety is *decidable*. So a basic question for the semidirect product and other operations on pseudovarieties defined in terms of generators is whether they preserve decidability. The difficulty in studying such operations lies in the fact that the answer is negative for most natural operations, including the semidirect product [1, 72, 24].

Let $\tau : S \to T$ be a relational morphism of monoids. Tilson [90] defined an associated category D_τ as follows: the objects are the elements of T; the morphisms from t to tt' are equivalence classes $[t, s', t']$ of triples $(t, s', t') \in T \times \tau$ under the relation which identifies the triples (t_1, s'_1, t'_1) and (t_2, s'_2, t'_2) if $t_1 = t_2$, $t_1 t'_1 = t_2 t'_2$, and for every s such that $(s, t_1) \in \tau$ we have $ss'_1 = ss'_2$; composition of morphisms is defined by the formula

$$[t, s_1, t_1]\,[tt_1, s_2, t_2] = [t, s_1 s_2, t_1 t_2].$$

The *derived semigroupoid* of a relational morphism of semigroups is defined similarly. The following is the well-known *Derived Category Theorem*.

7.3 Theorem (Tilson [90]) *A finite semigroup S belongs to $\mathsf{V} * \mathsf{W}$ if and only if there exists a relational morphism $\tau : S \to T$ with $T \in \mathsf{W}$ and $D_\tau \in g\mathsf{V}$.*

By applying the profinite approach, Weil and the author [20] have used the Derived Category Theorem to describe a basis of pseudoidentities for $\mathsf{V} * \mathsf{W}$ from a basis of semigroupoid pseudoidentities for $g\mathsf{V}$. This has come to be known as the *Basis Theorem*. Unfortunately, there is a gap in the argument which was found by J. Rhodes and B. Steinberg in trying to extend the approach to other operations on pseudovarieties and which makes the result only known to be valid in case W is locally finite or $g\mathsf{V}$ has *finite vertex-rank* in the sense that it admits a basis of pseudoidentities in graphs using only a bounded number of vertices. Although a counterexample was at one point announced for the Basis Theorem, at present it remains open whether it is true in general. Here is the precise statement of the "Basis Theorem":

> *Let V and W be pseudovarieties of semigroups and let $\{(u_i = v_i; \Gamma_i) : i \in I\}$ be a basis of semigroupoid pseudoidentities for $g\mathsf{V}$. For each pseudoidentity $u_i = w_i$ over the finite graph Γ_i one considers a labeling $\lambda : \Gamma_i \to (\overline{\Omega}_A \mathsf{S})^1$ of the graph Γ_i such that*
>
> (1) *the labels of edges belong to $\overline{\Omega}_A \mathsf{S}$;*

(2) *for every edge* $e : v_1 \to v_2$, W *satisfies the pseudoidentity* $\lambda(v_1)\lambda(e) = \lambda(v_2)$.

The labeling λ *extends to a continuous category homomorphism* $\hat{\lambda} : \overline{\Omega}_\Gamma \mathsf{Cat} \to (\overline{\Omega}_A \mathsf{S})^1$. *Let* z *be the label of the initial vertex for the morphisms* u_i, w_i *and consider the semigroup pseudoidentity* $z\,\hat{\lambda}(u_i) = z\,\hat{\lambda}(v_i)$. *Then the "Basis Theorem" is the assertion that the set of all such pseudoidentities constitutes a basis for* $\mathsf{V} * \mathsf{W}$.

In turn Steinberg and the author [15] have used the Basis Theorem to prove the following result which explains the interest in establishing graph-tameness of pseudovarieties. Say that a pseudovariety is *recursively definable* if it admits a recursively enumerable basis of pseudoidentities in which all the intervening implicit operations are computable.

7.4 Theorem (Almeida and Steinberg [15]) *If* V *is recursively enumerable and recursively definable and* W *is graph-tame, then the membership problem for* $\mathsf{V} * \mathsf{W}$ *is decidable provided* $g\,\mathsf{V}$ *has finite vertex-rank or* W *is locally finite.*

Moreover, we have the following result which was meant to handle the iteration of semidirect products. Let B_2 denote the syntactic semigroup of the language $(ab)^+$ over the alphabet $\{a, b\}$.

7.5 Theorem (Almeida and Steinberg [15]) *Let* $\mathsf{V}_1, \ldots, \mathsf{V}_n$ *be recursively enumerable pseudovarieties such that* $B_2 \in \mathsf{V}_1$, *and each* V_i *is tame. If the Basis Theorem holds then the semidirect product* $\mathsf{V}_1 * \cdots * \mathsf{V}_n$ *is decidable through a "uniform" algorithm depending only on algorithms for the factor pseudovarieties.*

Since B_2 belongs to A, in view of Ash's Theorem this would prove computability of the Krohn-Rodes complexity of finite semigroups once the Basis Theorem would be settled and a proof that A is graph-tame would be obtained. The latter has been announced by J. Rhodes but a written proof has been withdrawn since the gap in the proof of the Basis Theorem has been found.

8 Other operations on pseudovarieties

We make a brief reference in this section to two other famous results as well as some problems involving other operations on pseudovarieties of semigroups.

Given a semigroup S, one may extend the multiplication to an associative operation on subsets of S by putting $PQ = \{st : s \in P,\ t \in Q\}$. The resulting semigroup is denoted $\mathcal{P}(S)$. For a pseudovariety V of semigroups, let $\mathcal{P}\,\mathsf{V}$ denote the pseudovariety generated by all semigroups of the form $\mathcal{P}(S)$ with $S \in \mathsf{V}$. The operator $\mathcal{P}$ is called the *power operator* and it has been extensively studied. If V consists of finite semigroups each of which satisfies some nontrivial permutation identity or, equivalently, if V is contained in the pseudovariety

$$\mathsf{Perm} = [\![\, x^\omega yzt^\omega = x^\omega zyt^\omega \,]\!],$$

then one can find in [3] a formula for $\mathcal{P}\,\mathsf{V}$. Otherwise, it is also shown in [3] that $\mathcal{P}^3\,\mathsf{V} = \mathsf{S}$. These results were preceded by similar results of Margolis and Pin [58] in the somewhat easier case of monoids, where permutativity becomes commutativity.

One major open problem involving the power operator is the calculation of $\mathcal{P}\,\mathsf{J}$: it is well-known that this pseudovariety corresponds to the variety of dot-depth 2 languages in Straubing's hierarchy of star-free languages [67].

Another value of the operator $\mathcal{P}$ which has deserved major attention is $\mathcal{P}\,\mathsf{G}$. Before presenting results about this pseudovariety, we introduce another operator. Given two pseudovarieties V and W, their *Mal'cev product* $\mathsf{V}\,@m\,\mathsf{W}$ is the class of all finite semigroups S such that there exists a relational morphism $\tau : S \to T$ into $T \in \mathsf{W}$ such that, for every idempotent $e \in T$, we have $\tau^{-1}(e) \in \mathsf{V}$. It is an exercise to show that $\mathsf{V}\,@m\,\mathsf{W}$ is a pseudovariety.

The analog of the "Basis Theorem" for the Mal'cev product is the following.

8.1 Theorem (Pin and Weil [68]) *Let* V *and* W *be two pseudovarieties of semigroups and let* $\{u_i(x_1,\ldots,x_{n_i}) = v_i(x_1,\ldots,x_{n_i}) : i \in I\}$ *be a basis of pseudoidentities for* V*. Then* $\mathsf{V}\,@m\,\mathsf{W}$ *is defined by the pseudoidentities of the form* $u_i(w_1,\ldots,w_{n_i}) = v_i(w_1,\ldots,w_{n_i})$ *with* $i \in I$ *and* $w_j \in \overline{\Omega}_A\mathsf{S}$ *such that* $\mathsf{W} \models w_1 = \cdots = w_{n_i} = w_{n_i}^2$*.*

Call a pseudovariety W *idempotent-tame* if it is $\mathcal{C}$-tame for the set $\mathcal{C}$ of all systems of the form $x_1 = \cdots = x_n = x_n^2$. Applying the same approach as for semidirect products, we deduce that if V is decidable and W is idempotent-tame, then $\mathsf{V}\,@m\,\mathsf{W}$ is decidable [7].

For a pseudovariety H of groups, let BH denote the pseudovariety consisting of all finite semigroups in which regular elements have a unique inverse and whose subgroups belong to H. Finally, for a pseudovariety V, let $\mathcal{E}\,\mathsf{V}$ denote the pseudovariety consisting of all finite semigroups whose idempotents generate a subsemigroup which belongs to V.

We have the following chain of equalities

$$\mathcal{P}\,\mathsf{G} = \mathsf{J} * \mathsf{G} = \mathsf{J}\,@m\,\mathsf{G} = \mathsf{BG} = \mathcal{E}\,\mathsf{J} \tag{8.1}$$

The last equality is elementary as is the inclusion ($\subseteq$) in the second equality even for any pseudovariety of groups H in the place of G. The first and third equalities were proved by Margolis and Pin [59] using language theory. Using Knast's pseudoidentity basis for $g\,\mathsf{J}$, the inclusion ($\supseteq$) in the second equality was reduced by Henckell and Rhodes [40] to what they called the *pointlike conjecture* which is equivalent to the statement that G is $\mathcal{C}$-tame with respect to the signature κ where $\mathcal{C}$ is the class of systems associated with finite directed graphs with only two vertices and all edges coterminal. Hence Ash's Theorem implies the pointlike conjecture and the sequence of equalities (8.1) is settled.

Another problem which led to Ash's Theorem was the calculation of Mal'cev products of the form $\mathsf{V}\,@m\,\mathsf{G}$. For a finite semigroup S, define the *group-kernel* of S to consist of all elements $s \in S$ such that, for every relational morphism $\tau : S \to G$ into a finite group, we have $(s,1) \in \tau$. Then it is easy to check that S belongs to $\mathsf{V}\,@m\,\mathsf{G}$ if and only if $K(S) \in \mathsf{V}$. J. Rhodes conjectured that there should be an algorithm to compute $K(S)$ and proposed a specific procedure that should produce $K(S)$: start with the set E of idempotents of S and take the closure under multiplication in S and *weak conjugation*, namely the operation that, for a pair of elements $a, b \in S$, one of which is a weak inverse of the other, sends s to asb. This came to be known as the *Type II conjecture* and it is equivalent to Ribes and Zalesskiĭ's Theorem. Therefore, as was already observed in Section 6, Ash's Theorem also implies the Type II conjecture. See [39] for further information on the history of this conjecture.

B. Steinberg, later joined by K. Auinger, has done extensive work on generalizing the equalities (8.1) to other pseudovarieties of groups. This culminated in their recent papers [23,

25] where they completely characterize the pseudovarieties H of groups for which respectively the equalities $\mathcal{P}\mathsf{H} = \mathsf{J} * \mathsf{H}$ and $\mathsf{J} * \mathsf{H} = \mathsf{J} \mathbin{@} \mathsf{H}$ hold. On the other hand, Steinberg [85] has observed that results of Margolis and Higgins [42] imply that the inclusion $\mathsf{J} \mathbin{@} \mathsf{H} \subsetneq \mathsf{BH}$ is strict for every proper subpseudovariety $\mathsf{H} \subsetneq \mathsf{G}$ such that $\mathsf{H} * \mathsf{H} = \mathsf{H}$. Going in another direction, Escada and the author [14] have used profinite methods to show that several pseudovarieties V satisfy the equation $\mathsf{V} * \mathsf{G} = \mathcal{E}\mathsf{V}$. Hence the cryptic line (8.1) has been a source of inspiration for a lot of research.

Another result involving the two key pseudovarieties J and G is the decidability of their join $\mathsf{J} \vee \mathsf{G}$. This was proved independently by Steinberg [83] and Azevedo, Zeitoun and the author [10] using Ash's Theorem and the structure of free pro-J semigroups. Previously, Trotter and Volkov [91] had shown that $\mathsf{J} \vee \mathsf{G}$ is not finitely based.

9 Symbolic Dynamics and free profinite semigroups

We have seen that relatively free profinite semigroups are an important tool in the theory of pseudovarieties of semigroups. Yet very little is known about them in general, in particular for the finitely generated free profinite semigroups $\overline{\Omega}_A\mathsf{S}$. In this section we survey some recent results the author has obtained which reveal strong ties between Symbolic Dynamics and the structure of free profinite semigroups. See [9, 8] for more detailed surveys and [19] for related work.

Throughout this section let A be a finite alphabet. The additive group $\mathbb{Z}$ of integers acts naturally on the set $A^{\mathbb{Z}}$ of functions $f : \mathbb{Z} \to A$ by translating the argument: $(n \cdot f)(m) = f(m+n)$. The elements of $A^{\mathbb{Z}}$ may be viewed as *bi-infinite words* on the alphabet A. Recall that a *symbolic dynamical system* (*or subshift*) *over* A is a non-empty subset $\mathcal{X} \subseteq A^{\mathbb{Z}}$ which is topologically closed and stable under the natural action of $\mathbb{Z}$ in the sense that it is a union of orbits.

The *language* $L(\mathcal{X})$ of a subshift $\mathcal{X}$ consists of all finite factors of members of $\mathcal{X}$, that is words of the form $w[n, n+k] = w(n)w(n+1)\cdots w(n+k)$ with $n, k \in \mathbb{Z}$, $k \geq 0$, and $w \in \mathcal{X}$. It is easy to characterize the languages $L \subseteq A^*$ that arise in this way: they are precisely the *factorial* (closed under taking factors) and *extensible* languages ($w \in L$ implies that there exist letters $a, b \in A$ such that $aw, wb \in L$). We say that the subshift $\mathcal{X}$ is *irreducible* if for all $u, v \in L(\mathcal{X})$ there exists $w \in A^*$ such that $uwv \in L(\mathcal{X})$.

A subshift $\mathcal{X}$ is said to be *sofic* if $L(\mathcal{X})$ is a rational language. The subshift $\mathcal{X}$ is called a *subshift of finite type* if there is a finite set W of *forbidden words* which characterize $L(\mathcal{X})$ in the sense that $L(\mathcal{X}) = A^* \setminus (A^*WA^*)$; equivalently, the syntactic semigroup $\operatorname{Synt} L(\mathcal{X})$ is finite and satisfies the pseudoidentities $x^\omega yx^\omega zx^\omega = x^\omega zx^\omega yx^\omega$ and $x^\omega yx^\omega yx^\omega = x^\omega yx^\omega$ [3, Section 10.8].

The mapping $\mathcal{X} \mapsto L(\mathcal{X})$ transfers structural problems on subshifts to combinatorial problems on certains types of languages. But, from the algebraic-structural point of view, the free monoid A^* is a rather limited entity where combinatorial problems have often to be dealt in an ad hoc way. So, why not go a step forward to the profinite completion $\overline{\Omega}_A\mathsf{M} = (\overline{\Omega}_A\mathsf{S})^1$, where the interplay between algebraic and topological properties is expected to capture much of the combinatorics of the free monoid? We propose therefore to take this extra step and associate with a subshift $\mathcal{X}$ the closed subset $\overline{L(\mathcal{X})} \subseteq \overline{\Omega}_A\mathsf{M}$.

For example, in the important case of sofic subshifts, by Theorem 3.6 we recover $L(\mathcal{X})$ by

taking $\overline{L(\mathcal{X})} \cap A^*$. It turns out that the same is true for arbitrary subshifts so that the extra step does not lose information on subshifts but rather provides a richer structure in which to work.

Here are a couple of recent preliminary results following this approach.

9.1 Theorem *A subshift $\mathcal{X} \subseteq A^{\mathbb{Z}}$ is irreducible if and only if there is a unique minimal ideal $J(\mathcal{X})$ among those principal ideals of $\overline{\Omega}_A\mathsf{M}$ generated by elements of $\overline{L(\mathcal{X})}$ and its elements are regular.*

By a *topological partial semigroup* we mean a set S endowed with a continuous partial associative multiplication $D \to S$ with $D \subseteq S \times S$. Such a partial semigroup is said to be *simple* if every element is a factor of every other element. The structure of simple compact partial semigroups is well known: they are described by topological *Rees matrix semigroups* $\mathcal{M}(I, G, \Lambda, P)$, where I and Λ are compact sets, G is a compact group, and $P : Q \to G$ is a continuous function with $Q \subseteq \Lambda \times I$ a closed subset; as a set, $\mathcal{M}(I, G, \Lambda, P)$ is the Cartesian product $I \times G \times \Lambda$; the partial multiplication is defined by the formula

$$(i, g, \lambda)\,(j, h, \mu) = (i, g\,P(\lambda, j)\,h, \mu)$$

in case $P(\lambda, j)$ is defined and the product is left undefined otherwise. The group G is called the *structure group* and the function P is seen as a partial $\Lambda \times I$-matrix which is called the *sandwich matrix*.

It is well known that a regular $\mathcal{J}$-class J of a compact semigroup is a simple compact partial semigroup [43]. The structure group of J is a profinite group which is isomorphic to all maximal subgroups that are contained in J. In particular, for an irreducible subshift $\mathcal{X}$, there is an associated simple compact partial subsemigroup $J(\mathcal{X})$ of $\overline{\Omega}_A\mathsf{M}$. We denote by $G(\mathcal{X})$ the corresponding structure group which is a profinite group by Corollary 3.2.

Let $\mathcal{X} \subseteq A^{\mathbb{Z}}$ and $\mathcal{Y} \subseteq B^{\mathbb{Z}}$ be subshifts over two finite alphabets. A *conjugacy* is a function $\varphi : \mathcal{X} \to \mathcal{Y}$ which is a topological homeomorphism that commutes with the action of $\mathbb{Z}$ in the sense that for all $f \in A^{\mathbb{Z}}$ and $n \in \mathbb{Z}$, we have $\varphi(n \cdot f) = n \cdot \varphi(f)$. If there is such a conjugacy, then we say that $\mathcal{X}$ and $\mathcal{Y}$ are *conjugate*. By a *conjugacy invariant* we mean a structure $I(\mathcal{X})$ associated with each subshift $\mathcal{X} \subseteq A^{\mathbb{Z}}$ from a given class such that, if $\varphi : \mathcal{X} \to \mathcal{Y}$ is a conjugacy, then $I(\mathcal{X})$ and $I(\mathcal{Y})$ are isomorphic structures.

Let $\mathcal{X} \subseteq A^{\mathbb{Z}}$ and $\mathcal{Y} \subseteq B^{\mathbb{Z}}$ be subshifts. By a *sliding block code* we mean a function $\psi : \mathcal{X} \to \mathcal{Y}$ such that $\psi(w)(n) = \Psi(w[n-r, n+s])$ where $\Psi : A^{r+s+1} \cap L(\mathcal{X}) \to B$ is any function. The following diagram gives a pictorial description of this property and explains its name: each letter in the image $\psi(w)$ of $w \in \mathcal{X}$ is obtained by sliding a window of length $r + s + 1$ along w.

$$\begin{array}{c} \cdots\ a_{n-r-1}\ \boxed{a_{n-r}\,a_{n-r+1}\,\cdots\,a_{n+s-1}\,a_{n+s}}\ a_{n+s+1}\ \cdots \\ \Big\downarrow \Psi \\ \cdots\ b_{n-1}\ \boxed{b_n}\ b_{n+1}\ \cdots \end{array}$$

A sliding block code is said to be *invertible* if it is a bijection, which implies its inverse is also a sliding block code. It is well known from Symbolic Dynamics that the conjugacies are

the invertible sliding block codes (see for instance [55]). This implies that if a subshift $\mathcal{X}$ is conjugate to an irreducible subshift then $\mathcal{X}$ is also irreducible.

A major open problem in Symbolic Dynamics is if it is decidable whether two subshifts of finite type are conjugate and it is well known that it suffices to treat the irreducible case. Hence, the investigation of invariants seems to be worthwhile.

9.2 Theorem *For irreducible subshifts, the profinite group $G(\mathcal{X})$ is a conjugacy invariant.*

By a *minimal subshift* we mean one which is minimal with respect to inclusion. Minimal subshifts constitute another area of Symbolic Dynamics which has deserved a lot of attention. It is easy to see that minimal subshifts are irreducible.

Say that an implicit operation w is *uniformly recurrent* if every factor $u \in A^+$ of w is also a factor of every sufficiently long factor $v \in A^+$ of w.

9.3 Theorem *A subshift $\mathcal{X} \subseteq A^{\mathbb{Z}}$ is minimal if and only if the set $\overline{L(\mathcal{X})}$ meets only one nontrivial $\mathcal{J}$-class. Such a $\mathcal{J}$-class is then regular and it is completely contained in $\overline{L(\mathcal{X})}$. (This $\mathcal{J}$-class is then $J(\mathcal{X})$.) The $\mathcal{J}$-classes that appear in this way are those that contain uniformly recurrent implicit operations or, equivalently the $\mathcal{J}$-classes that contain non-explicit implicit operations and all their regular factors.*

To gain further insight, it seems worthwhile to compute the profinite groups of specific subshifts. One way to produce a wealth of examples is to consider *substitution subshifts*. We say that a continuous endomorphism $\varphi \in \operatorname{End}\overline{\Omega}_A\mathsf{S}$ is *primitive* if, for all $a, b \in A$, there exists n such that a is a factor of $\varphi^n(b)$, and that it is *finite* if $\varphi(A) \subseteq A^*$.

Given $\varphi \in \operatorname{End}\overline{\Omega}_A\mathsf{S}$ and a subpseudovariety $\mathsf{V} \subseteq \mathsf{S}$, φ induces a continuous endomorphism $\varphi' \in \operatorname{End}\overline{\Omega}_A\mathsf{V}$ namely the unique extension to a continuous endomorphism of the mapping which sends each $a \in A$ to $\pi\varphi(a)$, where $\pi : \overline{\Omega}_A\mathsf{S} \to \overline{\Omega}_A\mathsf{V}$ is the natural projection. In case $\varphi(A) \subseteq \Omega_A^\sigma\mathsf{S}$ for an implicit signature σ, the restriction of φ' to $\Omega_A^\sigma\mathsf{V}$ is an endomorphism of this σ-algebra. So, in particular, if φ is finite then it induces an endomorphism of the free group $\Omega_A^\kappa\mathsf{G}$.

The first part of the following result is well known in Symbolic Dynamics [69].

9.4 Theorem *Let $\varphi \in \operatorname{End}\overline{\Omega}_A\mathsf{S}$ be a finite primitive substitution and let $\mathcal{X}_\varphi \subseteq A^{\mathbb{Z}}$ be the subshift whose language $L(\mathcal{X}_\varphi)$ consists of all factors of $\varphi^n(a)$ ($a \in A$, $n \geq 0$). Then the following properties hold:*

(1) *the subshift $\mathcal{X}_\varphi$ is minimal;*

(2) *if φ induces an automorphism of the free group, then $G(\mathcal{X}_\varphi)$ is a free profinite group on $|A|$ free generators.*

To test whether an endomorphism ψ of the free group $\Omega_A^\kappa\mathsf{G}$ is an automorphism, it suffices to check whether the subgroup generated by $\psi(A)$ is all of $\Omega_A^\kappa\mathsf{G}$. There is a well-known algorithm to check this property, namely Stallings' folding algorithm applied to the "flower automaton", whose petals are labeled with the words $\psi(A)$ [82, 46].

9.5 Example The *Fibonacci substitution* φ given by the pair (ab, a) is finite and primitive and therefore it determines a subshift $\mathcal{X}_\varphi$. Moreover, φ is invertible in the free group $\Omega_2^\kappa\mathsf{G}$

since we may easily recover the generators a and b from their images ab and a using group operations: the substitution given by the pair $(b, b^{\omega-1}a)$ is the inverse of φ in the free group. By Theorem 9.4, the group $G(\mathcal{X}_\varphi)$ is a free profinite group on two free generators. At present we do not know the precise structure of the compact partial semigroup $J(\mathcal{X}_\varphi)$. This example has been considerably extended to minimal subshifts which are not generated by substitutions, namely to Sturmian subshifts and even to Arnoux-Rauzy subshifts [9, 8].

9.6 Example For the substitution φ given by the pair (ab, a^3b), one can show that the group $G(\mathcal{X}_\varphi)$ is not a free profinite group.

Acknowledgments

This work was supported, in part, by *Fundação para a Ciência e a Tecnologia* (FCT) through the *Centro de Matemática da Universidade do Porto*, and by the FCT project POCTI/32817/MAT/2000, which is partially funded by the European Community Fund FEDER. The author is indebted to Alfredo Costa for his comments on preliminary versions of these notes.

References

[1] D. Albert, R. Baldinger, and J. Rhodes, The identity problem for finite semigroups (the undecidability of), *J. Symbolic Logic* **57** (1992), 179–192.

[2] J. Almeida, Residually finite congruences and quasi-regular subsets in uniform algebras, *Portugal. Math.* **46** (1989), 313–328.

[3] J. Almeida, *Finite Semigroups and Universal Algebra*, World Scientific, Singapore, 1995, english translation.

[4] J. Almeida, A syntactical proof of locality of DA, *Int. J. Algebra Comput.* **6** (1996), 165–177.

[5] J. Almeida, Hyperdecidable pseudovarieties and the calculation of semidirect products, *Int. J. Algebra Comput.* **9** (1999), 241–261.

[6] J. Almeida, Dynamics of implicit operations and tameness of pseudovarieties of groups, *Trans. Amer. Math. Soc.* **354** (2002), 387–411.

[7] J. Almeida, Finite semigroups: an introduction to a unified theory of pseudovarieties, in: *Semigroups, Algorithms, Automata and Languages* (G. M. S. Gomes, J.-E. Pin, and P. V. Silva, eds.), World Scientific, Singapore, 2002.

[8] J. Almeida, Profinite structures and dynamics, *CIM Bulletin* **14** (2003), 8–18.

[9] J. Almeida, Symbolic dynamics in free profinite semigroups, *RIMS Kokyuroku* **1366** (2004), 1–12.

[10] J. Almeida, A. Azevedo, and M. Zeitoun, Pseudovariety joins involving $\mathcal{J}$-trivial semigroups, *Int. J. Algebra Comput.* **9** (1999), 99–112.

[11] J. Almeida and M. Delgado, Sur certains systèmes d'équations avec contraintes dans un groupe libre, *Portugal. Math.* **56** (1999), 409–417.

[12] J. Almeida and M. Delgado, Sur certains systèmes d'équations avec contraintes dans un groupe libre—addenda, *Portugal. Math.* **58** (2001), 379–387.

[13] J. Almeida and M. Delgado, Tameness of the pseudovariety of abelian groups, Tech. Rep. CMUP 2001-24, Univ. Porto (2001), to appear in *Int. J. Algebra Comput.*

[14] J. Almeida and A. Escada, On the equation $V * G = EV$, *J. Pure Appl. Algebra* **166** (2002), 1–28.

[15] J. Almeida and B. Steinberg, On the decidability of iterated semidirect products and applications to complexity, *Proc. London Math. Soc.* **80** (2000), 50–74.

[16] J. Almeida and B. Steinberg, Syntactic and global semigroup theory, a synthesis approach, in: *Algorithmic Problems in Groups and Semigroups* (J. C. Birget, S. W. Margolis, J. Meakin, and M. V. Sapir, eds.), Birkhäuser, 2000.

[17] J. Almeida and P. G. Trotter, Hyperdecidability of pseudovarieties of orthogroups, *Glasgow Math. J.* **43** (2001), 67–83.

[18] J. Almeida and P. G. Trotter, The pseudoidentity problem and reducibility for completely regular semigroups, *Bull. Austral. Math. Soc.* **63** (2001), 407–433.

[19] J. Almeida and M. V. Volkov, Subword complexity of profinite words and subgroups of free profinite semigroups, Tech. Rep. CMUP 2003-10, Univ. Porto (2003), to appear in *Int. J. Algebra Comput.*

[20] J. Almeida and P. Weil, Profinite categories and semidirect products, *J. Pure Appl. Algebra* **123** (1998), 1–50.

[21] M. Arbib, *Algebraic Theory of Machines, Languages and Semigroups*, Academic Press, New York, 1968.

[22] C. J. Ash, Inevitable graphs: a proof of the type II conjecture and some related decision procedures, *Int. J. Algebra Comput.* **1** (1991), 127–146.

[23] K. Auinger and B. Steinberg, On power groups and embedding theorems for relatively free profinite monoids, *Math. Proc. Cambridge Phil. Soc.* To appear.

[24] K. Auinger and B. Steinberg, On the extension problem for partial permutations, *Proc. Amer. Math. Soc.* **131** (2003), 2693–2703.

[25] K. Auinger and B. Steinberg, The geometry of profinite graphs with applications to free groups and finite monoids, *Trans. Amer. Math. Soc.* **356** (2004), 805–851.

[26] B. Banaschewski, The birkhoff theorem for varieties of finite algebras, *Algebra Universalis* **17** (1983), 360–368.

[27] T. Bandman, G.-M. Greuel, F. Grunewald, B. Kunyavskiĭ, G. Pfister, and E. Plotkin, Engel-like identities characterizing finite solvable groups, Tech. rep. (2003), available from arXiv.org e-Print archive at http://arxiv.org/abs/math/0303165.

[28] G. Baumslag, Residual nilpotence and relations in free groups, *J. Algebra* **2** (1965), 271–282.

[29] J. A. Brzozowski, Hierarchies of aperiodic languages, *RAIRO Inf. Théor. et Appl.* **10** (1976), 33–49.

[30] J. A. Brzozowski and I. Simon, Characterizations of locally testable events, *Discrete Math.* **4** (1973), 243–271.

[31] S. Burris and H. P. Sankappanavar, *A Course in Universal Algebra*, no. 78 in Grad. Texts in Math., Springer, Berlin, 1981.

[32] D. M. Clark, B. A. Davey, and M. Jackson, Standard topological algebras and syntactic congruences, Tech. rep. (2003).

[33] T. Coulbois and A. Khélif, Equations in free groups are not finitely approximable, *Proc. Amer. Math. Soc.* **127** (1999), 963–965.

[34] M. Crochemore and W. Rytter, *Text algorithms*, The Clarendon Press Oxford University Press, New York, 1994, with a preface by Zvi Galil.

[35] S. Eilenberg, *Automata, Languages and Machines*, vol. A, Academic Press, New York, 1974.

[36] S. Eilenberg, *Automata, Languages and Machines*, vol. B, Academic Press, New York, 1976.

[37] S. Eilenberg and M. P. Schützenberger, On pseudovarieties, *Advances in Math.* **19** (1976), 413–418.

[38] M. Hall, A topology for free groups and related groups, *Ann. Math.* **52** (1950), 127–139.

[39] K. Henckell, S. Margolis, J.-E. Pin, and J. Rhodes, Ash's type II theorem, profinite topology and malcev products. Part I, *Int. J. Algebra Comput.* **1** (1991), 411–436.

[40] K. Henckell and J. Rhodes, The theorem of Knast, the PG=BG and Type II Conjectures, in: *Monoids and Semigroups with Applications* (J. Rhodes, ed.), World Scientific, Singapore, 1991.

[41] B. Herwig and D. Lascar, Extending partial automorphisms and the profinite topology on free groups, *Trans. Amer. Math. Soc.* **352** (2000), 1985–2021.

[42] P. M. Higgins and S. W. Margolis, Finite aperiodic semigroups with commuting idempotents and generalizations, *Israel J. Math.* **116** (2000), 367–380.

[43] K. H. Hofmann and P. S. Mostert, *Elements of compact semigroups*, Charles E. Merrill, Columbus, Ohio, 1966.

[44] R. P. Hunter, Certain finitely generated compact zero-dimensional semigroups, *J. Austral. Math. Soc., Ser. A* **44** (1988), 265–270.

[45] P. R. Jones, Profinite categories, implicit operations and pseudovarieties of categories, *J. Pure Appl. Algebra* **109** (1996), 61–95.

[46] I. Kapovich and A. Myasnikov, Stallings foldings and subgroups of free groups, *J. Algebra* **248** (2002), 608–668.

[47] J. Karnofsky and J. Rhodes, Decidability of complexity one-half for finite semigroups, *Semigroup Forum* **24** (1982), 55–66.

[48] S. C. Kleene, Representations of events in nerve nets and finite automata, in: *Automata Studies* (C. E. Shannon, ed.), vol. 3-41, Princeton University Press, Princeton, N.J., 1956, reprinted in [60].

[49] R. Knast, A semigroup characterization of dot-depth one languages, *RAIRO Inf. Théor. et Appl.* **17** (1983), 321–330.

[50] R. Knast, Some theorems on graph congruences, *RAIRO Inf. Théor. et Appl.* **17** (1983), 331–342.

[51] K. Krohn and J. Rhodes, Algebraic theory of machines. I. Prime decomposition theorem for finite semigroups and machines, *Trans. Amer. Math. Soc.* **116** (1965), 450–464.

[52] K. Krohn and J. Rhodes, Complexity of finite semigroups, *Ann. of Math. (2)* **88** (1968), 128–160.

[53] G. Lallement, *Semigroups and Combinatorial Applications*, Wiley, New York, 1979.

[54] M. V. Lawson, *Inverse Semigroups: the Theory of Partial Symmetries*, World Scientific, Singapore, 1998.

[55] D. Lind and B. Marcus, *An introduction to symbolic dynamics and coding*, Cambridge University Press, Cambridge, 1996.

[56] A. Lubotzky, Pro-finite presentations, *J. Algebra* **242** (2001), 672–690.

[57] S. Margolis, M. Sapir, and P. Weil, Closed subgroups in pro-V topologies and the extension problem for inverse automata, *Int. J. Algebra Comput.* **11** (2001), 405–445.

[58] S. W. Margolis and J.-E. Pin, Minimal noncommutative varieties and power varieties, *Pacific J. Math.* **111** (1984), 125–135.

[59] S. W. Margolis and J.-E. Pin, Varieties of finite monoids and topology for the free monoid, in: *Proc. 1984 Marquette Semigroup Conference*, Marquette University, Milwaukee, 1984.

[60] E. F. Moore, ed., *Sequential Machines: Selected Papers*, Addison-Wesley, Reading, Mass., 1964.

[61] J. Myhill, Finite automata and the representation of events, Tech. Rep. 57624, Wright Air Development Command (1957).

[62] K. Numakura, Theorems on compact totally disconnetced semigroups and lattices, *Proc. Amer. Math. Soc.* **8** (1957), 623–626.

[63] D. Perrin, Finite automata, in: *Handbook of Theoretical Computer Science* (J. van Leeuwen, ed.), vol. B: Formal Models and Semantics, Elsevier, Amsterdam, 1990.

[64] J.-E. Pin, Variétés de langages et variétés de semigroupes, Ph.D. thesis, Univ. Paris 7 (1981).

[65] J.-E. Pin, *Varieties of Formal Languages*, Plenum, London, 1986, english translation.

[66] J.-E. Pin and C. Reutenauer, A conjecture on the Hall topology for the free group, *Bull. London Math. Soc.* **23** (1991), 356–362.

[67] J.-E. Pin and H. Straubing, Monoids of upper triangular matrices, in: *Semigroups: structure and universal algebraic problems* (G. Pollák, ed.), North-Holland, Amsterdam, 1985.

[68] J.-E. Pin and P. Weil, Profinite semigroups, mal'cev products and identities, *J. Algebra* **182** (1996), 604–626.

[69] M. Queffélec, *Substitution Dynamical Systems—Spectral Analysis*, vol. 1294 of *Lect. Notes in Math.*, Springer-Verlag, Berlin, 1987.

[70] J. Reiterman, The Birkhoff theorem for finite algebras, *Algebra Universalis* **14** (1982), 1–10.

[71] J. Rhodes, Kernel systems — a global study of homomorphisms on finite semigroups, *J. Algebra* **49** (1977), 1–45.

[72] J. Rhodes, Undecidability, automata and pseudovarieties of finite semigroups, *Int. J. Algebra Comput.* **9** (1999), 455–473.

[73] J. Rhodes and B. Tilson, Improved lower bounds for the complexity of finite semigroups, *J. Pure Appl. Algebra* **2** (1972), 13–71.

[74] L. Ribes and P. A. Zalesskiĭ, On the profinite topology on a free group, *Bull. London Math. Soc.* **25** (1993), 37–43.

[75] L. Ribes and P. A. Zalesskiĭ, The pro-p topology of a free group and algorithmic problems in semigroups, *Int. J. Algebra Comput.* **4** (1994), 359–374.

[76] L. Ribes and P. A. Zalesskiĭ, *Profinite Groups*, no. 40 in Ergeb. Math. Grenzgebiete 3, Springer, Berlin, 2000.

[77] D. J. S. Robinson, *A Course in Theory of Groups*, no. 80 in Grad. Texts in Math., Springer-Verlag, New York, 1982.

[78] S. Satoh, K. Yama, and M. Tokizawa, Semigroups of order 8, *Semigroup Forum* **49** (1994), 7–30.

[79] M. P. Schützenberger, On finite monoids having only trivial subgroups, *Inform. and Control* **8** (1965), 190–194.

[80] I. Simon, Piecewise testable events, in: *Proc. 2nd GI Conf.*, vol. 33 of *Lect. Notes in Comput. Sci.*, Springer, Berlin, 1975.

[81] I. Simon, Words distinguished by their subwords, Tech. rep., Univ. S. Paulo (2003).

[82] J. R. Stallings, Topology of finite graphs, *Inventiones Mathematicae* **71** (1983), 551–565.

[83] B. Steinberg, On pointlike sets and joins of pseudovarieties, *Int. J. Algebra Comput.* **8** (1998), 203–231.

[84] B. Steinberg, Inevitable graphs and profinite topologies: some solutions to algorithmic problems in monoid and automata theory, stemming from group theory, *Int. J. Algebra Comput.* **11** (2001), 25–71.

[85] B. Steinberg, A note on the equation $PH = J*H$, *Semigroup Forum* **63** (2001), 469–474.

[86] H. Straubing, Finite semigroup varieties of the form $V * D$, *J. Pure Appl. Algebra* **36** (1985), 53–94.

[87] P. Tesson and D. Thérien, Diamonds are forever: the variety da, in: *Semigroups, Algorithms, Automata and Languages* (G. M. S. Gomes, J.-E. Pin, and P. V. Silva, eds.), World Scientific, Singapore, 2002.

[88] D. Thérien and A. Weiss, Graph congruences and wreath products, *J. Pure Appl. Algebra* **36** (1985), 205–215.

[89] D. Thérien and T. Wilke, Over words, two variables are as powerful as one quantifier alternation, in: *STOC '98 (Dallas, TX)*, ACM, New York, 1999, 234–240.

[90] B. Tilson, Categories as algebra: an essential ingredient in the theory of monoids, *J. Pure Appl. Algebra* **48** (1987), 83–198.

[91] P. G. Trotter and M. V. Volkov, The finite basis problem in the pseudovariety joins of aperiodic semigroups with groups, *Semigroup Forum* **52** (1996), 83–91.

[92] A. Weiss and D. Thérien, Varieties of finite categories, *RAIRO Inf. Théor. et Appl.* **20** (1986), 357–366.

[93] S. Willard, *General Topology*, Addison-Wesley, Reading, Mass., 1970.

[94] M. Zorn, Nilpotency of finite groups (abstract), *Bull. Amer. Math. Soc.* **42** (1936), 485–486.

The structure of free algebras

Joel BERMAN

Department of Mathematics, Statistics and Computer Science
University of Illinois at Chicago
851 South Morgan, Chicago, IL 60607
USA

Abstract

This article is a survey of selected results on the structure of free algebraic systems obtained during the past 50 years. The focus is on ways free algebras can be decomposed into simpler components and how the number of components and the way the components interact with each other can be readily determined. A common thread running through the exposition is a concrete method of representing a free algebra as an array of elements.

1 Introduction

Let $\mathcal{V}$ be a variety (or equational class) of algebras. By $\mathbf{F}_{\mathcal{V}}(X)$ we denote the free algebra for $\mathcal{V}$ freely generated by the set X. The algebra $\mathbf{F}_{\mathcal{V}}(X)$ may be defined as an algebra in $\mathcal{V}$ generated by X that has the universal mapping property: For every $\mathbf{A} \in \mathcal{V}$ and every function $v : X \to A$ there is a homomorphism $h : \mathbf{F}_{\mathcal{V}}(X) \to \mathbf{A}$ that extends v, i.e., $h(x) = v(x)$ for all $x \in X$. It is known that $\mathbf{F}_{\mathcal{V}}(X)$ exists for every variety $\mathcal{V}$ and is unique up to isomorphism.

In this article we investigate the structure of free algebras in varieties. We present a very concrete, almost tactile, representation of $\mathbf{F}_{\mathcal{V}}(X)$ and some algebras closely related to $\mathbf{F}_{\mathcal{V}}(X)$ as a rectangular array of elements. The rows of such an array are indexed by the elements of the algebra and the columns are indexed by a set U of valuations, where a valuation is any function v from X to an algebra $\mathbf{A} \in \mathcal{V}$ for which the set $v(X)$ generates all of $\mathbf{A}$. This array is denoted $\mathbf{Ge}(X, U)$.

In Section 1 we prove some general results about $\mathbf{Ge}(X, U)$. We are interested in finding small tractable sets U of valuations that can be used to represent $\mathbf{F}_{\mathcal{V}}(X)$. Several such U are described. The section also contains a detailed description of how a software package developed by R. Freese and E. Kiss can be used to explicitly construct the array $\mathbf{Ge}(X, U)$ for finitely generated varieties $\mathcal{V}$ and finite X. The remaining three sections deal with decompositions of $\mathbf{F}_{\mathcal{V}}(X)$ into well-behaved substructures, and how these substructures are organized among themselves in a systematic way. In Section 2 the substructures considered are congruence classes of the kernel of a canonical homomorphism of $\mathbf{F}_{\mathcal{V}}(X)$ onto $\mathbf{F}_{\mathcal{W}}(X)$ for $\mathcal{W}$ a well-behaved and well-understood subvariety of $\mathcal{V}$. In Section 3 varieties $\mathcal{V}$ are described for which $\mathbf{F}_{\mathcal{V}}(X)$ can be decomposed into overlapping syntactically defined substructures that are very homogeneous and code, in various ways, families of finite ordered sets. Because of the way these subsets fit together, an inclusion-exclusion type of argument can be used to

V. B. Kudryavtsev and I. G. Rosenberg (eds.),
Structural Theory of Automata, Semigroups, and Universal Algebra, 47–76.

determine the cardinality of $\mathbf{F}_{\mathcal{V}}(X)$. The final section deals with direct decompositions of $\mathbf{F}_{\mathcal{V}}(X)$ into a product of directly indecomposable factors. Some specific general conditions on a variety $\mathcal{V}$ are provided that allow for the determination of the structure of the directly indecomposable factors and of the exact multiplicity of each factor in the product.

For general facts about varieties, free algebras and universal algebra used in this paper the reader may consult [8] or [18]. For the most part we follow the notation and terminology found there.

For a set X of variables and a variety $\mathcal{V}$ of similarity type τ, a typical term of type τ built from X is denoted $t(x_1, \ldots, x_n)$, where the x_i are elements of X. If n and X are clear from the context we simply write t. The set of variables that actually appear in a term t is denoted $\mathrm{var}(t)$. For a term $t(x_1, \ldots, x_n)$ and an algebra $\mathbf{A}$ in $\mathcal{V}$, by $t^{\mathbf{A}}$ we denote the the term operation on $\mathbf{A}$ corresponding to t. Thus $t^{\mathbf{A}} : A^n \to A$ and for $a_1, \ldots, a_n \in A$, $t^{\mathbf{A}}(a_1, \ldots, a_n)$ is the element of A obtained by interpreting t to the fundamental operations of $\mathbf{A}$ and applying the resulting operation to $a_1, \ldots, a_n$. The universe of $\mathbf{F}_{\mathcal{V}}(X)$ is denoted $\mathrm{F}_{\mathcal{V}}(X)$ and we use the following notation for elements of $\mathrm{F}_{\mathcal{V}}(X)$: For $x \in X$ we write $\overline{x}$ for the element of $\mathrm{F}_{\mathcal{V}}(X)$ obtained by applying $x^{\mathbf{F}_{\mathcal{V}}(X)}$ to the generator x of $\mathbf{F}_{\mathcal{V}}(X)$ and for a term $t(x_1, \ldots, x_n)$ we write $\overline{t}$ for $t^{\mathbf{F}_{\mathcal{V}}(X)}(\overline{x}_1, \ldots, \overline{x}_n)$. Note that $x = \overline{x}$ for every $x \in X$, every element of $\mathbf{F}_{\mathcal{V}}(X)$ is of the form $\overline{t}$ for some term t, and that for distinct terms s and t we can have $\overline{s} = \overline{t}$. A crucial fact about free algebras used repeatedly throughout this paper is that $\mathcal{V}$ satisfies the identity $s \approx t$ for terms s and t if and only if the elements $\overline{s}$ and $\overline{t}$ are equal in $\mathrm{F}_{\mathcal{V}}(X)$. Written more succinctly this is the familiar

$$\mathcal{V} \models s \approx t \iff \overline{s} = \overline{t}.$$

See, for example, [8, §11] for further details.

For an algebra $\mathbf{A}$ and a set X we view each element v in the direct power $\mathbf{A}^X$ as a function $v : X \to A$. Given $v \in \mathbf{A}^X$ we denote the algebra $\mathbf{A}$ by $\mathbf{Alg}(v)$. If $v \in \mathbf{A}^X$ and $t(x_1, \ldots, x_n)$ is a term in the variables of X, then $v(t)$ denotes $t^{\mathbf{A}}(v(x_1), \ldots, v(x_n))$. We say v is a *valuation* if $v(X)$ generates the algebra $\mathbf{Alg}(v)$. The set of all valuations from X to an algebra $\mathbf{A}$ is denoted $\mathrm{val}(X, \mathbf{A})$. For $\mathcal{K}$ a class of algebras, $\mathrm{val}(X, \mathcal{K})$ denotes the collection of all $v \in \mathrm{val}(X, \mathbf{A})$ for $\mathbf{A} \in \mathcal{K}$. Valuations will play a central role in our exposition.

If an algebra $\mathbf{B}$ is the direct product $\prod_{j \in J} \mathbf{B}_j$, then we denote by pr_j the projection homomorphism from $\mathbf{B}$ onto $\mathbf{B}_j$. For $K \subseteq J$ the projection of $\mathbf{B}$ onto $\prod_{j \in K} \mathbf{B}_j$ is denoted pr_K.

Let $\mathcal{K}$ be a set of algebras of the same similarity type, X a set, and U a subset of $\bigcup(\mathbf{A}^X : \mathbf{A} \in \mathcal{K})$. For $x \in X$ let $\overline{x} \in \prod_{v \in U} \mathbf{Alg}(v)$ denote the element given by $pr_v(\overline{x}) = v(x)$ for all $v \in U$. We let $\overline{X} = \{\overline{x} : x \in X\}$. The subalgebra of $\prod_{v \in U} \mathbf{Alg}(v)$ generated by $\overline{X}$ is denoted $\mathbf{Ge}(X, U)$.

1.1 Lemma *Let $\mathcal{V}$ be a variety and $U \subseteq \bigcup(\mathbf{A}^X : \mathbf{A} \in \mathcal{V})$. If $\mathrm{val}(X, \mathcal{V}) \subseteq U$, then $\boldsymbol{Ge}(X, U)$ is isomorphic to $\mathbf{F}_{\mathcal{V}}(X)$.*

Proof The algebra $\mathbf{Ge}(X, U)$ is generated by $\overline{X}$. The set U contains valuations to separate the elements of X, so X and $\overline{X}$ have the same cardinality. So it suffices to show that $\mathbf{Ge}(X, U)$ has the universal mapping property for $\overline{X}$ over $\mathcal{V}$. Let w be any function from $\overline{X}$ into an algebra $\mathbf{A} \in \mathcal{V}$. Let $v \in \mathrm{val}(X, \mathrm{Sg}^{\mathbf{A}}(w(X)))$ be defined by $v(x) = w(\overline{x})$ for every $\overline{x} \in \overline{X}$. By assumption $v \in U$. Then pr_v is a homomorphism from $\mathbf{Ge}(X, U)$ into $\mathbf{A}$ that extends w since $pr_v(\overline{x}) = v(x) = w(\overline{x})$. □

1.2 Definition For a variety $\mathcal{V}$ and a set X, we call $U \subseteq \mathrm{val}(X, \mathcal{V})$ *free for* $\mathcal{V}$ if $\mathbf{Ge}(X, U)$ is isomorphic to $\mathbf{F}_{\mathcal{V}}(X)$. The set U is called *independent* for $\mathcal{V}$ if $\mathbf{Ge}(X, U)$ is isomorphic to $\prod_{v \in U} \mathbf{Alg}(v)$.

We are very much interested in concrete representations of $\mathbf{Ge}(X, U)$ when U is free for $\mathcal{V}$. To this end we first consider sufficient conditions on a set $U \subseteq \mathrm{val}(X, \mathcal{V})$ that force U to be independent.

1.3 Lemma *Let $\mathcal{V}$ be a variety and suppose $U \subseteq \mathrm{val}(X, \mathcal{V})$ is such that for every pair of terms s and t for which $\mathcal{V} \not\models s \approx t$ there exists $v \in U$ such that $v(s) \neq v(t)$. Then $\boldsymbol{Ge}(X, U) \cong \mathbf{F}_{\mathcal{V}}(X)$, that is, U is free for $\mathcal{V}$.*

Proof By virtue of the previous lemma it suffices to show that the projection homomorphism pr_U from $\mathbf{Ge}(X, \mathrm{val}(X, \mathcal{V}))$ onto $\mathbf{Ge}(X, U)$ is one-to-one. If $s(\overline{x}_1, \ldots, \overline{x}_n)$ and $t(\overline{x}_1, \ldots, \overline{x}_n)$ are distinct elements of $\mathbf{Ge}(X, \mathrm{val}(X, \mathcal{V}))$, then $\mathcal{V} \not\models s \approx t$. So there exists $v \in U$ for which $v(s) \neq v(t)$. Then $pr_v(s(\overline{x}_1, \ldots, \overline{x}_n)) = s(v(x_1), \ldots, v(x_n)) = v(s) \neq v(t) = t(v(x_1), \ldots, v(x_n)) = pr_v(s(\overline{x}_1, \ldots, \overline{x}_n))$. Thus, $pr_U(s(\overline{x}_1, \ldots, \overline{x}_n)) \neq pr_U(t(\overline{x}_1, \ldots, \overline{x}_n))$. □

1.4 Corollary *Let $\mathcal{V}$ be a variety and X a set.*

(1) *If $\mathcal{V}_{\mathrm{SI}}$ is the class of (finitely generated) subdirectly irreducible algebras in $\mathcal{V}$ and if $U = \mathrm{val}(X, V_{\mathrm{SI}})$, then U is free for $\mathcal{V}$.*

(2) *If $\mathcal{V}$ is generated by the finite algebras $\mathbf{A}_1, \ldots, \mathbf{A}_m$ and if $U = \bigcup(\mathbf{A}_i^X : 1 \leq i \leq m)$, then U is free for $\mathcal{V}$ and $|\mathbf{F}_{\mathcal{V}}(n)| \leq \prod_{i=1}^{m} |A_i|^{|A_i|^n}$.*

The second part of this Corollary is presented in Birkhoff's 1935 paper [7].

1.5 Definition Let $\mathcal{V}$ be an arbitrary variety and X a set. Given two valuations v and v' with $\mathbf{A} = \mathbf{Alg}(v)$ and $\mathbf{A}' = \mathbf{Alg}(v')$ algebras in $\mathcal{V}$, we say that v and v' are *equivalent*, written $v \sim v'$, if there exist homomorphisms $h : \mathbf{A} \to \mathbf{A}'$ and $h' : \mathbf{A}' \to \mathbf{A}$ such that $v' = hv$ and $v = h'v'$. For any set $U \subseteq \mathrm{val}(X, \mathcal{V})$ let $E(U)$ denote any transversal of $\sim$ over elements of U. That is, $E(U)$ is any subset of U that consists of exactly one valuation taken from each equivalence class of $\sim$.

1.6 Lemma *If U is any set of valuations, then $\boldsymbol{Ge}(X, U) \cong \boldsymbol{Ge}(X, E(U))$.*

Proof The set $E(U)$ is a subset of U, and the projection $pr_{ev(U)} : \mathbf{Ge}(X, U) \to \mathbf{Ge}(X, E(U))$ is an onto homomorphism. It suffices to show $pr_{E(U)}$ is one-to-one. Let $\overline{s} \neq \overline{t}$ in $\mathbf{Ge}(X, U)$. So there is a $v \in U$ for which $v(s) \neq v(t)$. Let $v' \in E(U)$ be such that $v \sim v'$. Then there is a homomorphism h' for which $v = h'v'$. If $v'(s) = v'(t)$, then $v(s) = v(t)$, which is impossible. So $v'(s) \neq v'(t)$ and hence $pr_{E(U)}(\overline{s}) \neq pr_{E(U)}(\overline{t})$. □

1.7 Corollary *If a set U of valuations is free for $\mathcal{V}$, then so is $E(U)$.*

Henceforth in this paper, unless otherwise indicated, X will denote the set $\{x_1, \ldots, x_n\}$. Under this convention $\mathbf{F}_{\mathcal{V}}(X)$ and $\mathbf{F}_{\mathcal{V}}(n)$ are the same.

A variety $\mathcal{V}$ is *locally finite* if every finitely generated algebra in $\mathcal{V}$ is finite. For a locally finite variety, if $f(n)$ is the cardinality of $\mathbf{F}_{\mathcal{V}}(n)$, then every at most n generated algebra in

$\mathcal{V}$ has cardinality at most $f(n)$. The number of valuations v from a set X of size n to a particular algebra $\mathbf{A}$ is bounded above by $|A|^n$, which is at most $f(n)^n$.

Let $\mathcal{V}$ be a locally finite variety, $X = \{x_1, \ldots, x_n\}$, and $U \subseteq \mathrm{val}(X, \mathcal{V})$. We present a concrete representation of $\mathbf{Ge}(X, U)$ as a rectangular array of elements from $\mathbf{Alg}(v)$ for v ranging over U. Each $v \in U$ determines a column in this array and each element of $\mathbf{Ge}(X, U)$ determines a row. The first n rows are indexed by $\overline{x}_1, \ldots, \overline{x}_n$. If $t(x_1, \ldots, x_n)$ is any term, then $\overline{t}$ denotes $t^{\mathbf{Ge}(X,U)}(\overline{x}_1, \ldots, \overline{x}_n)$. Since the $\overline{x}_i$ generate $\mathbf{Ge}(X, U)$, every element of $\mathbf{Ge}(X, U)$ is of the form $\overline{t}$ for some term t. For $v \in U$ and $\overline{t} = t(\overline{x}_1, \ldots, \overline{x}_n) \in \mathbf{Ge}(X, U)$, the entry in row $\overline{t}$ and column v is $v(t) \in \mathbf{Alg}(v)$. Note that $v(t) = t^{\mathbf{Alg}(v)}(v(x_1), \ldots, v(x_n))$. In column v, every element of $\mathbf{Alg}(v)$ appears at least once since v is a valuation. Moreover, column v codes the projection pr_v of $\mathbf{Ge}(X, U)$ onto the algebra $\mathbf{Alg}(v)$.

In the special case that $\mathcal{V}$ is generated by a finite algebra $\mathbf{A}$ and $U = \mathbf{A}^X$, then U is free for $\mathcal{V}$ by Corollary 1.4, the number of columns of $\mathbf{Ge}(X, U)$ is $|A|^n$, and the number of rows of $\mathbf{Ge}(X, U)$ is $|\mathrm{F}_{\mathcal{V}}(n)|$.

We next present three examples of how $\mathbf{Ge}(X, U)$ may be used to describe the structure of free algebras.

1.8 Example Let $\mathcal{S}$ be the variety of semilattices. We show how to find a normal form for an arbitrary term and how to use this normal form to determine the structure of free semilattices. If $t(x_1, \ldots, x_n)$ is any semilattice term, then by associativity we may ignore the parentheses and write t as a string of variables. By commutativity we may sort the variables in the string by increasing subscripts. Idempotence allows us to conclude $\mathcal{S} \models t \approx x_{i_1} x_{i_2} \ldots x_{i_k}$ where $1 \leq i_1 < i_2 < \cdots < i_k \leq n$ and $\mathrm{var}(t) = \{x_{i_1}, x_{i_2}, \ldots, x_{i_k}\}$. Thus a concrete representation of $\mathbf{F}_{\mathcal{S}}(X)$ as a join semilattice is the set of nonvoid subsets of X with the operation of union. In particular $|\mathrm{F}_{\mathcal{S}}(X)| = 2^n - 1$.

The variety $\mathcal{S}$ is generated by the 2-element semilattice $\mathbf{S}_2$ with universe $\{0, 1\}$. Without loss of generality $\mathbf{S}_2$ is a join semilattice. The set $\mathrm{val}(X, \mathbf{S}_2)$ has $2^n - 2$ elements. Thus, $\mathbf{F}_{\mathcal{S}}(X)$ is isomorphic to $\mathbf{Ge}(X, \mathrm{val}(X, \mathbf{S}_2))$. If we view $\mathbf{Ge}(X, \mathrm{val}(X, \mathbf{S}_2))$ as an array, then it has $2^n - 2$ columns and $2^n - 1$ rows. If $U \subseteq \mathrm{val}(X, \mathbf{S}_2)$ consists of the valuations v_i for $1 \leq i \leq n$ where $v_i(x_j) = 1$ if and only if $i = j$, then U is free for $\mathcal{S}$. To see this, it suffices to show that the projection pr_U is one-to-one. Suppose $\overline{t} = t(\overline{x}_1, \ldots, \overline{x}_n)$ and $\overline{s} = s(\overline{x}_1, \ldots, \overline{x}_n)$ are distinct elements of $\mathbf{Ge}(X, \mathrm{val}(X, \mathbf{S}_2))$. Let v be any valuation into $\mathbf{S}_2$ that separates $\overline{s}$ and $\overline{t}$, with say $v(t) = 1$. So there exists $x_i \in \mathrm{var}(t) - \mathrm{var}(s)$. Then $v_i(x_i) = 1 = v_i(t)$ while $v_i(s) = 0$. So a valuation v_i in U separates $\overline{t}$ and $\overline{s}$. Hence U is free for $\mathcal{S}$. The array for $\mathbf{Ge}(X, U)$ has n columns and $2^n - 1$ rows. A cardinality argument shows that no proper subset of U is free for $\mathcal{S}$. Although U is free for $\mathbf{S}$, it is not independent since there is no $\overline{t}$ having $v(t) = 0$ for all v. However, if $\mathcal{S}$ were the variety of join semilattices with constant 0, then U would be both free and independent for $\mathcal{S}$ with $\mathbf{F}_{\mathcal{S}}(n) \cong \mathbf{S}_2^n$ for $\mathbf{S}_2$ the 2-element join semilattice with constant 0.

1.9 Example Let $\mathcal{B}$ be the variety of Boolean algebras. This variety is generated by the 2-element Boolean algebra $\mathbf{B}_2 = \langle \{0, 1\}, \wedge, \vee, ', 0, 1 \rangle$, which is the only subdirectly irreducible algebra in the variety. Note that $\mathrm{val}(X, \mathbf{B}_2) = \mathbf{B}_2^X$. For $v \in \mathrm{val}(X, \mathbf{B}_2)$ let t_v be the term $x_1^{v(x_1)} \wedge \cdots \wedge x_n^{v(x_n)}$, where $x_i^0 = x_i'$ and $x_i^1 = x_i$. Then $v(t_v) = 1$ but $w(t_v) = 0$ for every valuation $w \neq v$. The array for $\mathbf{Ge}(X, \mathrm{val}(X, \mathbf{B}_2))$ has 2^n columns. By considering joins of the 2^n different t_v we see that there are 2^{2^n} rows. So $\mathrm{val}(X, \mathbf{B}_2)$ is free for $\mathcal{B}$ and is independent

but has no proper subset that is free for $\mathcal{B}$. Thus $\mathbf{F}_{\mathcal{B}}(n) \cong \mathbf{B}_2^n$. The t_v are sometimes called *minterms*. The array for $\mathbf{Ge}(X, \mathrm{val}(X, \mathbf{B}_2))$ may be viewed as the collection of all truth tables for Boolean terms involving the variables from X. The only difference being that in truth tables the result of applying all valuations to a given term is given as a column vector whereas in the array for $\mathrm{val}(X, \mathbf{B}_2)$ this is presented as a row vector.

1.10 Example Let $\mathbf{C}_2$ be the 2-element implication algebra $\langle \{0, 1, \}, \to \rangle$ and $\mathcal{I}$ the variety it generates. In $\mathbf{C}_2$ we have $1 \to 0 = 0$ and $a \to b = 1$ otherwise. Algebras in $\mathcal{I}$ have an order defined on them by $x \leq y$ if and only if $x \to y = 1$. The term operation $(x_1 \to x_2) \to x_2$ is the join operation for this order. We consider $\mathbf{Ge}(X, \mathbf{C}_2^X)$, which is isomorphic to $\mathbf{F}_{\mathcal{I}}(X)$. For $\overline{t} \in \mathbf{Ge}(X, \mathbf{C}_2^X)$ we have $\overline{x}_i \leq \overline{t}$ if and only if $v(x_i) \leq v(t)$ for all $v \in \mathbf{C}_2^X$. For $1 \leq i \leq n$ let $\overline{T}_i = \{\overline{t} \in \mathbf{Ge}(X, \mathbf{C}_2^X) : \overline{x}_i \leq \overline{t}\}$. Then $\overline{T}_i$ is a subuniverse of $\mathbf{Ge}(X, \mathbf{C}_2^X)$ since $y \leq x \to y$ holds for all elements in algebras in $\mathcal{I}$. Moreover, every $\overline{t}$ is in at least one $\overline{T}_i$. Let U_i consist of all $v \in \mathbf{C}_2^X$ for which $v(x_i) = 0$. The array for $\mathbf{Ge}(X, U_i)$ has 2^{n-1} columns. For $\overline{t} \in \mathbf{Ge}(X, U_i)$ and $v \in U_i$ we have that $v(t \to x_i) = t'$ for $'$ the complementation operation on $\mathbf{B}_2$. Thus, on $\mathbf{Ge}(X, U_i)$ we have term operations that are the Boolean operations $\vee$ and $'$. The element $\overline{x}_i$ serves as the Boolean 0 here. So $\mathbf{Ge}(X, U_i)$ contains all Boolean term operations that can be generated by the $n-1$ elements $\overline{x}_1, \ldots, \overline{x}_{i-1}, \overline{x}_{i+1}, \ldots, \overline{x}_n$. Thus $\overline{T}_i$ contains $2^{2^{n-1}}$ elements. If $\overline{\mathbf{T}}_i$ is the algebra with universe $\overline{T}_i$, then $\overline{\mathbf{T}}_i$ and $(\mathbf{C}_2)^{2^{n-1}}$ are isomorphic. From the facts that $\mathbf{Ge}(X, \mathbf{C}_2^X) = \overline{T}_1 \cup \cdots \cup \overline{T}_n$, all the $\overline{\mathbf{T}}_i$ are isomorphic, and $|\overline{T}_1 \cap \cdots \cap \overline{T}_k| = 2^{2^{n-k}}$ for every $1 \leq k \leq n$, it follows from a standard inclusion-exclusion argument that

$$|\mathbf{F}_{\mathcal{I}}(n)| = |\,\mathbf{Ge}(X, \mathbf{C}_2^X)| = \sum_{k=1}^{n} (-1)^k \binom{n}{k} 2^{2^{n-k}}.$$

In the previous examples we considered a finite algebra $\mathbf{A}$ that generates a variety $\mathcal{V}$ and we determined the free algebra $\mathbf{F}_{\mathcal{V}}(X)$ by means of an analysis of the array $\mathbf{Ge}(X, U)$ for some $U \subset \mathbf{A}^X$ that is free for $\mathcal{V}$. The algebras $\mathbf{A}$ considered have a 2-element universe and a particularly transparent structure. For more complicated finite algebras $\mathbf{A}$, it is usually more difficult to analyze $\mathbf{Ge}(X, U)$. However, the array for $\mathbf{Ge}(X, U)$ can, in principle, be computed mechanically since it is a subalgebra of $\mathbf{A}^U$ generated by the row vectors $\overline{x}_1, \ldots, \overline{x}_n$. The natural algorithm for finding a subuniverse of an algebra generated by a given generating set can be implemented as a computer program.

Emil W. Kiss and Ralph Freese have developed a software package for doing computations in universal algebra. The package is the Universal Algebra Calculator (UAC) and is freely available from either author's webpage [11]. The menu of programs allows the user to compute all congruence relations of an algebra, view and manipulate the congruence lattice, find factor algebras for a given congruence relation, and answer questions about the algebra that arise from commutator theory and tame congruence theory. Two programs in the package are extremely useful for work on the structure of free algebras. One is a program that with the input of a finite algebra $\mathbf{A}$ and positive integer n, determines the free algebra $\mathbf{F}_{\mathcal{V}}(n)$ for $\mathcal{V}$ the variety generated by $\mathbf{A}$. The second takes as input a finite collection of finite algebras $\mathbf{A}_1, \ldots, \mathbf{A}_m$ of the same similarity type and a set of vectors $\overline{z}_1, \ldots, \overline{z}_n$, with each $\overline{z}_i \in \mathbf{A}_1 \times \cdots \times \mathbf{A}_m$, and produces as output the subalgebra of $\mathbf{A}_1 \times \cdots \times \mathbf{A}_m$ generated by $\overline{z}_1, \ldots, \overline{z}_n$. This program can of course be used to compute $\mathbf{Ge}(X, U)$. These programs are especially useful for analyzing examples and for conducting computer experiments on specific

algebras. We next present a detailed example of how the Universal Algebra Calculator might be used in this experimental manner to investigate free algebras in a finitely generated variety.

1.11 Example For every finite poset $\langle P, \leq \rangle$ with a top element 1 define a finite algebra $\mathbf{P} = \langle P, \cdot \rangle$ by

$$x \cdot y = \begin{cases} 1 & \text{if } x \leq y \\ y & \text{otherwise.} \end{cases}$$

Let $\mathcal{V}$ be the variety generated by all such $\mathbf{P}$. The general problem is to determine the structure of $\mathbf{F}_{\mathcal{V}}(n)$ for all positive n. A more specific problem is to determine the free algebras in a variety generated by a single algebra $\mathbf{P} \in \mathcal{V}$.

If $\mathbf{A}$ is any algebra in $\mathcal{V}$, then it is not hard to show that the binary relation $\leq$ on A given by $x \leq y$ if and only if $x \cdot y = 1$ is an order relation. Note that although 1 is not in the similarity type of $\mathcal{V}$, the term $x \cdot x$ will serve in its place.

If $\mathbf{P}$ is the algebra arising from the two element chain $0 < 1$, then $\mathbf{P} = \mathbf{C}_2$ as in Example 1.10, and we have described there the structure of the free algebras in the variety generated by $\mathbf{C}_2$. So let us next consider the algebra arising from the 3-element chain $2 < 0 < 1$. If $\mathbf{C}_3$ denotes this algebra, then the operation $x \cdot y$ on $\mathbf{C}_3$ is given by the following table.

$\cdot$	0	1	2
0	1	1	2
1	0	1	2
2	1	1	1

We can use this table as input for the Universal Algebra Calculator. The UAC can then calculate the free algebra on two free generators for the variety $\mathcal{V}$ generated by this input algebra. A tabular printout for the result of this computation is given in Table 1.

The algebra $\mathbf{F}_{\mathcal{V}}(2)$ is represented as a subset of $(\mathbf{C}_3)^{3^2}$. The elements are represented as 9-tuples with the two generators having the label G and numbered 0 and 1. The other elements in the free algebra are labeled with the letter V and are numbered 2 through 13. The binary operation on the algebra is denoted `f0`. Information about how each generated element is obtained is explicitly given. Thus, in the last two lines of the table we see that element 13 is obtained by applying `f0` to elements 4 and 5. A printout of this data in a more streamlined form is given in Table 2. Here we see a representation of $\mathbf{F}_{\mathcal{V}}(2)$ as we view the array $\mathbf{Ge}(X, U)$ for X the two generating elements and U the 9 functions $\{0, 1, 2\}^X$. This illustrates how the UAC program gives a concrete representation of the array $\mathbf{Ge}(X, U)$ for an algebra $\mathbf{A}$ and a set $U \subseteq \mathbf{A}^X$.

A first approach to understanding the structure of $\mathbf{F}_{\mathcal{V}}(2)$ would be to draw the order relation on the 14 elements of the algebra. This can be done using Table 2 by ordering the vectors coordinatewise by $0 < 1$. The diagram shows that the two generators are the only minimal elements of this order, that is, every element is greater than or equal to at least one generator. This is reminiscent of the situation for the free algebras in the variety $\mathcal{I}$ generated by $\mathbf{C}_2$ described in Example 1.10. The entire ordered set is quite complicated, but if we look at a single generator and all of the elements greater than or equal to it, we get the ordered set drawn in Figure 1. This ordered set is $\mathbf{C}_2{}^3$ with two additional elements labelled 8 and

```
. .   The 2 generators:
. .   G    0: ( 0, 0, 0, 1, 1, 1, 2, 2, 2)
. .   G    1: ( 0, 1, 2, 0, 1, 2, 0, 1, 2)
. . > Newly generated elements:
. . > V    2: ( 1, 1, 1, 1, 1, 1, 1, 1, 1)
. .           = f0( 0, 0)
. . > V    3: ( 1, 0, 1, 1, 1, 1, 2, 2, 1)
. .           = f0( 1, 0)
. . > V    4: ( 1, 1, 2, 0, 1, 2, 1, 1, 1)
. .           = f0( 0, 1)
. . > V    5: ( 0, 1, 0, 1, 1, 1, 1, 1, 2)
. .           = f0( 3, 0)
. . > V    6: ( 0, 0, 1, 1, 1, 1, 2, 2, 2)
. .           = f0( 4, 0)
. . > V    7: ( 0, 1, 2, 0, 1, 2, 1, 1, 2)
. .           = f0( 3, 1)
. . > V    8: ( 0, 1, 1, 1, 1, 1, 0, 1, 2)
. .           = f0( 4, 1)
. . > V    9: ( 1, 1, 0, 1, 1, 1, 1, 1, 1)
. .           = f0( 6, 0)
. . > V   10: ( 1, 0, 0, 1, 1, 1, 2, 2, 1)
. .           = f0( 8, 0)
. . > V   11: ( 1, 1, 2, 0, 1, 2, 0, 1, 1)
. .           = f0( 5, 1)
. . > V   12: ( 1, 1, 1, 1, 1, 1, 0, 1, 1)
. .           = f0( 7, 1)
. . > V   13: ( 0, 1, 1, 1, 1, 1, 1, 1, 2)
. .           = f0( 4, 5)
```

Table 1

12. Note that in $\mathbf{F}_{\mathcal{I}}(2)$ the ordered set of elements above either generator is order isomorphic to $\mathbf{C}_2^3$. So a natural question to ask is how do 8 and 12 differ from the other eight elements that are above the generator 0?

By means of Table 1 we can represent each element in terms of the two generators 0 and 1. The element 8 is (01)1 and the element 12 is ((10)1)1. The other elements in Figure 1 all have the property that each of them can be written as a term involving the generators 0 and 1 with 0 being the rightmost variable that appears. Thus $13 = (01)((01)0)$. The elements 8 and 12, at least in the representation given by Table 1, have rightmost variable 1. One can argue that in any representation of 8 and 12 as a term operation t applied to 0 and 1 the rightmost variable of t will be 1. This can be done by observing that if v is any valuation of X into $\mathbf{C}_3$ and $t(x_1, \ldots, x_n)$ is any term with rightmost variable x_i, then $v(t) = v(x_i)$ or $v(t) = 1$.

For the valuation v that sends generator 0 to 2 and generator 1 to 0, an examination of

```
[ Creating an algebra in a subproduct.
 . [ Reading vectorlist file 'C:\algebras\join3f2.uni'.
 . . < Number of vectors: 14
 . .   Length of vectors: 9
 . .   V    0: ( 0, 0, 0, 1, 1, 1, 2, 2, 2)
 . .   V    1: ( 0, 1, 2, 0, 1, 2, 0, 1, 2)
 . .   V    2: ( 1, 1, 1, 1, 1, 1, 1, 1, 1)
 . .   V    3: ( 1, 0, 1, 1, 1, 1, 2, 2, 1)
 . .   V    4: ( 1, 1, 2, 0, 1, 2, 1, 1, 1)
 . .   V    5: ( 0, 1, 0, 1, 1, 1, 1, 1, 2)
 . .   V    6: ( 0, 0, 1, 1, 1, 1, 2, 2, 2)
 . .   V    7: ( 0, 1, 2, 0, 1, 2, 1, 1, 2)
 . .   V    8: ( 0, 1, 1, 1, 1, 1, 0, 1, 2)
 . .   V    9: ( 1, 1, 0, 1, 1, 1, 1, 1, 1)
 . .   V   10: ( 1, 0, 0, 1, 1, 1, 2, 2, 1)
 . .   V   11: ( 1, 1, 2, 0, 1, 2, 0, 1, 1)
 . .   V   12: ( 1, 1, 1, 1, 1, 1, 0, 1, 1)
 . .   V   13: ( 0, 1, 1, 1, 1, 1, 1, 1, 2)
 . ] End of reading vectors.
 . [ Reading algebra 'C:\algebras\join3.alg'.
```

Table 2

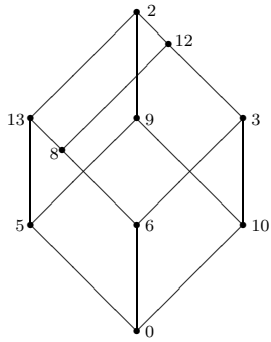


Figure 1

the appropriate column in Table 1 shows that $v(8) = 0$. So 0 could not be the rightmost variable for any representation of element 8. A similar argument works for element 12. This

analysis suggests that for a given generator x_i, rather than look at the set of elements that are greater than or equal to x_i, we look at those elements that can be written in some way with a term having x_i as its rightmost variable. There are $8 = 2^3$ elements representable with terms having 0 as the rightmost variable, and by symmetry there are 8 elements having 1 as rightmost variable, and two elements, 2 and 13, that can be written with either variable as a rightmost variable. We have $14 = 8 + 8 - 2$ and appear to have a decomposition of $\mathbf{F}_{\mathcal{V}}(2)$ into well-structured overlapping blocks.

We next want to find a small set of valuations that can be used to separate those elements of $\mathbf{F}_{\mathcal{V}}(n)$ that have a representation by means of a term with rightmost variable x_i. In the case of $\mathbf{F}_{\mathcal{I}}(n)$ we considered only those valuations v that have $v(x_i) = 0$. Could this same condition work for the variety $\mathcal{V}$? We experiment with $\mathbf{F}_{\mathcal{V}}(2)$. There are three valuations with $v(x_1) = 0$. Let U be this set of valuations and consider $\mathbf{Ge}(\{x_1, x_2\}, U)$. The UAC program allows us to generate a subalgebra of product by inputting the algebra, the number of coordinates, and the generating vectors. Thus we can generate the array corresponding to $\mathbf{Ge}(\{x_1, x_2\}, U)$. This is given in Table 3. We see that there are 10 elements generated. Those vectors that contain a 2 cannot be written as a term with rightmost variable x_1. This leaves $8 = 2^3$ vectors that contain only 0 and 1. For any such vector v we have $(v \cdot 0) \cdot 0 = v$, so all 2^3 of these vectors can be represented by terms in which the rightmost variable is x_1. So the three valuations in U serve to separate those elements that can be represented by terms with rightmost variable x_1, and the number of such elements is $\{0, 1\}^{|U|}$.

```
This algebra is the subalgebra of the 3-th power of the algebra
C:\algebras\hilbert3.alg generated by:
     0 : [0, 0, 0]
     1 : [0, 1, 2]
 . > The generated subpower contains 10 vector(s).
 > The generated universe is:
     0 : [0, 0, 0]
     1 : [0, 1, 2]
     2 : [1, 1, 1]
     3 : [1, 0, 1]
     4 : [1, 1, 2]
     5 : [0, 1, 0]
     6 : [0, 0, 1]
     7 : [0, 1, 1]
     8 : [1, 1, 0]
     9 : [1, 0, 0]
```

Table 3

On the basis of these computer experiments with $\mathbf{F}_{\mathcal{V}}(2)$ we form this conjecture: For $\mathbf{F}_{\mathcal{V}}(n)$ the subalgebra $\mathbf{T}_i$, whose universe consists of those elements that can be written using a term having rightmost variable x_i, is isomorphic to $\mathbf{C}_2^m$, where m is the number of $v \in \mathbf{C}_3^X$ for which $v(x_i) = 0$. Thus $\mathbf{T}_i \cong (\mathbf{C}_2)^{2^{n-1}}$. If this conjecture were true, then an inclusion-exclusion argument as used for the variety $\mathcal{V}$ of implication algebras could be used to find

the cardinality of $\mathbf{F}_{\mathcal{V}}(n)$.

We can use the UAC software to check the conjecture for $n = 3$. If we just run the program we discover that $\mathbf{F}_{\mathcal{V}}(3)$ has 1514 elements. The algebra is too big, however, to carry out the analysis as we did for $n = 2$. We can compute with the program the array corresponding to $\mathbf{Ge}(X, U)$ for $X = \{x_1, x_2, x_3\}$ and U consisting of all $v \in \mathbf{C}_3^X$ with $v(x_1) = 0$. Note that $|U| = 9$. The computation shows that there are 558 rows in the array $\mathbf{Ge}(X, U)$. An examination shows that 46 of these rows contain the element 2, which leaves 512 rows in $\{0, 1\}^9$. Thus, there are at least 2^9 elements of $\mathbf{F}_{\mathcal{V}}(3)$ in T_1. Likewise if we let U be those v for which $v(x_1) = 0$ and $v(x_2) = 0$, then we see that there are at least eight elements in $T_1 \cap T_2$. We also compute that $T_1 \cap T_2 \cap T_3$ has at cardinality at least 2. By symmetry we have $|T_1| = |T_2| = |T_3|$. An inclusion-exclusion argument shows that we have counted $3 * 2^9 - 3 * 2^3 + 2^1 = 1514$ elements. Since 1514 is the cardinality of $\mathbf{F}_{\mathcal{V}}(3)$, we see that T_i does indeed have 2^{3^2} elements, as predicted by the conjecture. So we have used the UAC software to verify the conjecture for $n = 3$. In Section 3 we will give a proof of the conjecture in a more general context. The proof will essentially be a formalization of the analysis that we used in our computer experiments in this example.

2 Structure via decomposition

We wish to understand the structure of free algebras in a variety. For some varieties the free algebras have a fairly transparent structure. For example, the structure of free semilattices is reasonably clear and is presented in Example 1.8. But for most varieties the structure of the free algebras is too complex to discern in a single view. In some cases one can decompose the free algebras into manageable blocks, and describe how the blocks fit together. If the structure of the blocks is clear and if the manner in which the blocks are related to one another is understood, then a good analysis of the structure of the free algebras in the variety may be provided. In this section we consider a standard and useful method for such decompositions of free algebras. The basic situation is that we are interested in the free algebras in a variety $\mathcal{V}$ and we know that $\mathcal{V}$ has a subvariety $\mathcal{W}$ for which we have some understanding of the algebras $\mathbf{F}_{\mathcal{W}}(n)$. Let $g : \mathbf{F}_{\mathcal{V}}(n) \to \mathbf{F}_{\mathcal{W}}(n)$ be the canonical map that sends each generator x_i to x_i. Then the kernel of g is a congruence on $\mathbf{F}_{\mathcal{V}}(n)$ whose congruence classes partition $\mathbf{F}_{\mathcal{V}}(n)$. We want $\mathcal{V}$ and $\mathcal{W}$ in which the structure of $\mathbf{F}_{\mathcal{W}}(n)$ and information about the individual congruence classes of the kernel of g may be used to describe $\mathbf{F}_{\mathcal{V}}(n)$. We present several concrete examples of this approach.

2.1 Example The analysis of the free bands by J. A. Green and D. Rees [13] is an excellent example of this method. Let $\mathcal{V}$ be the variety of bands, that is, the variety of semigroups given by the idempotent law $x^2 \approx x$. The variety $\mathcal{S}$ of semilattices, which is the class of commutative bands, is a subvariety of $\mathcal{V}$. Let γ be the kernel of the canonical map $g : \mathbf{F}_{\mathcal{V}}(n) \to \mathbf{F}_{\mathcal{S}}(n)$ in which $g(x_i) = x_i$ for $1 \leq i \leq n$. Recall that for any term t, we let $\mathrm{var}(t)$ denote the set of variables that appear in t. It is easily seen that for $\overline{s} \in \mathrm{F}_{\mathcal{S}}(n)$ the congruence class $g^{-1}(\overline{s})$ consists of those $\overline{t} \in \mathrm{F}_{\mathcal{V}}(n)$ for which $\mathrm{var}(t) = \mathrm{var}(s)$. Moreover, $g^{-1}(\overline{s})$ is a subuniverse of $\mathbf{F}_{\mathcal{V}}(n)$. It can be argued that for $\overline{a}, \overline{b}$ and $\overline{c}$ in $\mathbf{F}_{\mathcal{V}}(n)$, if $(\overline{a}, \overline{c}) \in \gamma$ and $\mathrm{var}(b) \subseteq \mathrm{var}(a) = \mathrm{var}(c)$, then $\overline{a}\overline{b}\overline{c}$ and $\overline{a}\overline{c}$ are in the same γ class. This means that each γ class is a rectangular band, that is, the elements in the class satisfy the identity $xyx \approx x$. Thus, the classes of γ partition $\mathbf{F}_{\mathcal{V}}(n)$ into rectangular bands and these classes interact together as elements of the free

semilattice on n generators. Further structure of the γ classes is presented in [13]. It is shown that if $\overline{t} \in \mathbf{F}_{\mathcal{V}}(n)$ with $|\operatorname{var}(t)| = i$, then

$$|\overline{t}/\gamma| = \prod_{j=1}^{i} (i-j+1)^{2^j},$$

and hence the variety of bands is locally finite with

$$|\mathrm{F}_{\mathcal{V}}(n)| = \sum_{i=1}^{n} \binom{n}{i} \prod_{j=1}^{i} (i-j+1)^{2^j}.$$

In this type of decomposition in which $\mathbf{F}_{\mathcal{V}}(n)$ is partitioned by a canonical map onto $\mathbf{F}_{\mathcal{W}}(n)$, it is critical that the structure of $\mathbf{F}_{\mathcal{W}}(n)$ be well understood. In the band example the map is onto a free semilattice, which as we have seen has a very transparent structure. Free finitely generated Boolean algebras are also well understood and so for many of the varieties of algebras arising in algebraic logic that contain Boolean algebras as a subvariety, this method of describing the structure of free algebras has been used with some success.

The method may also be used for locally finite varieties of lattices. If $\mathcal{V}$ is any nontrivial variety of lattices, then the variety $\mathcal{D}$ of distributive lattices is a subvariety of $\mathcal{V}$. The free distributive lattice has been the object of intensive study and its structure is understood, although not as thoroughly as that of free semilattices or free Boolean algebras. Thus, an analysis of the partition of $\mathbf{F}_{\mathcal{V}}(n)$ induced by the kernel of the canonical map $g : \mathbf{F}_{\mathcal{V}}(n) \to \mathbf{F}_{\mathcal{D}}(n)$ might reveal the structure of this free lattice.

We first discuss some of the known structure of the free distributive lattice $\mathbf{F}_{\mathcal{D}}(X)$ for $X = \{x_1, \ldots, x_n\}$. There is a normal form for distributive lattice terms, as every element can be written as $\bigvee_i (\bigwedge X_i)$ where the X_i are pairwise incomparable subsets of X. This is, of course, the well-known conjunctive normal form. The lattice $\mathbf{F}_{\mathcal{D}}(X)$ has a concrete representation as the lattice of down-sets in the ordered set consisting of all proper nonvoid subsets of X. An equivalent representation is the set of proper nonvoid anti-chains in the ordered set $\{0,1\}^X$. Yet another representation is as the set of all n-ary operations on $\{0,1\}$ that preserve the order relation $0 < 1$. The ordered set of join-irreducible elements of $\mathbf{F}_{\mathcal{D}}(X)$ is order isomorphic to the set of proper nonvoid subsets of X ordered by inclusion.

Despite all this structural information the problem of actually determining the size of $\mathbf{F}_{\mathcal{D}}(n)$ remains a difficult one. An 1897 paper of Dedekind gives $|\mathrm{F}_{\mathcal{D}}(4)| = 166$. In the years that have followed the cardinalities of free distributive lattices on n free generators have been determined, but with the time interval between successive values of n being 15–20 years. Thus, the exact values of $\mathrm{F}_{\mathcal{V}}(n)$ are known only for $n \leq 8$, with value for $n = 8$ being a 23 decimal digit number found in 1991. This suggests that unlike the case for bands, for lattice varieties the decomposition of $\mathbf{F}_{\mathcal{V}}(n)$ into congruence classes of the canonical map onto $\mathbf{F}_{\mathcal{D}}(n)$ will not provide a usable general formula for $|\mathrm{F}_{\mathcal{V}}(n)|$ as a function of n.

Nonetheless, the method has been used to determine the structure and cardinality of free lattices on a small number of generators in some varieties of lattices. The next paragraphs present some of the results from [6] where such an analysis has been performed.

Let $\mathcal{V}$ be any variety of lattices and let $g : \mathbf{F}_{\mathcal{V}}(n) \to \mathbf{F}_{\mathcal{D}}(n)$ be the homomorphism in which $g(x_i) = x_i$ for all variables $x_i \in X$. The kernel of g is the congruence relation γ. For any lattice term p we wish to describe the class $\overline{p}/\gamma$. It can be argued that g is a bounded

homomorphism in the sense that the interval $\overline{p}/\gamma$ has a top and bottom element. An argument in [6] shows that if p_* is the disjunctive normal form of p and p^* is the conjunctive normal form, then $\overline{p}/\gamma$ is the interval $[\overline{p_*}, \overline{p^*}]$. Let U be all those valuations $v \in \mathbf{A}^X$, as $\mathbf{A}$ ranges over the subdirectly irreducible algebras in $\mathcal{V}$, with $v(p_*) \neq v(p^*)$. Then the interval $\overline{p}/\gamma$ is lattice isomorphic to a sublattice of $\prod_{v\in U} \mathbf{Alg}(v)$ with the embedding given by $pr_v(\overline{q}) = v(q)$ for every $\overline{q} \in \overline{p}/\gamma$. The structure of $\overline{p}/\gamma$ is analyzed by determining the collection of join-irreducible elements in this interval lattice. Note that in any finite lattice L, the set $J(L)$ of join-irreducible elements generates all of L. The following result of R. Wille [22] is useful here.

Let the finite lattice L be a subdirect product of the lattices L_i $(i \in I)$. Then $J(L) = \bigcup_{i\in I}\{\bigwedge pr_i^{-1}(c) : c \in J(L_i)\}$.

With this result and the fact that $\overline{p}/\gamma$ is the interval $[\overline{p_*}, \overline{p^*}]$, one can, in principle, determine the structure of each congruence class $\overline{p}/\gamma$ if one knows $\bigwedge pr_v^{-1}(c)$ where v is an arbitrary valuation in U and c is a join-irreducible element of $\mathbf{Alg}(v)$. We present a concrete example illustrating this, using the variety $\mathcal{V}$ generated by the 5-element non-modular lattice $\mathbf{N}$.

Let the elements of the lattice $\mathbf{N}$ be $0, 1, 2, 3, 4$ with $2 < 3$, $4 \vee 2 = 4 \vee 3 = 1$ and $4 \wedge 2 = 4 \wedge 3 = 0$. This lattice is generated by the set $\{2, 3, 4\}$ and any generating set must contain this set of elements. If $\mathcal{V}$ is the lattice variety generated by $\mathbf{N}$, then the only subdirectly irreducible lattices in $\mathcal{V}$ are $\mathbf{N}$ and the 2-element chain $\mathbf{C}_2$ with elements $0 < 1$. So the variety $\mathcal{V}$ covers the variety of distributive lattices in the lattice of all lattice varieties. Now $\mathbf{C}_2$ is a homomorphic image of $\mathbf{N}$, so if p and q are lattice terms in the variables X and if v is a valuation to $\mathbf{C}_2$ for which $v(p) \neq v(q)$, then there is a $w \in \mathbf{N}^X$ for which $w(p) \neq w(q)$. So there is a set $U \subseteq \mathbf{N}^X$ for which $\mathbf{F}_\mathcal{V}(X)$ is lattice isomorphic to $\mathbf{Ge}(X, U)$.

In order to apply Wille's result we need a description of $\bigwedge pr^{-1}(c)$ for c a join-irreducible element of N. The following is proved in [6]:

Let $h : \mathbf{F}_\mathcal{V}(X) \to \mathbf{N}$ be an onto homomorphism. For $i \in N$ define $\overline{i} = \bigwedge\{x \in X : h(x) \geq i\}$. Then h is a bounded homomorphism with a lower bound given by $h(\overline{p}) \geq i$ if and only if $\overline{p} \geq \overline{i} \wedge (\overline{2} \vee \overline{4})$ in $\mathbf{F}_\mathcal{V}(X)$.

We view $\mathbf{F}_\mathcal{V}(X)$ as $\mathbf{Ge}(X, U)$ and we let v be any valuation in U. Then the projection pr_v is the homomorphism of $\mathbf{F}_\mathcal{V}(X)$ onto $\mathbf{N}$ that extends v. For a nonsingleton congruence class $\overline{p}/\gamma$ we have that a typical join-irreducible element will be of the form $p_* \vee (\overline{i} \wedge (\overline{2} \vee \overline{4}))$ where i is join-irreducible in $\mathbf{N}$. If p is a lattice term for which $g(p_*) = g(p^*)$, then the class $\overline{p}/\gamma$ is a singleton since $\mathcal{D} \models p_* \approx p^*$. If $v \in \mathbf{N}^X$ is such that $v(p_*) \neq v(p^*)$, then $\{v(p_*), v(p^*)\} = \{2, 3\}$ since $\mathbf{N}/\theta$ is a distributive lattice for θ the congruence relation generated by identifying 2 and 3. So for every $\overline{q} \in \overline{p}/\gamma$ we have $v(q) \in \{2, 3\}$ since $v(p_*) \leq v(q) \leq v(p^*)$. Hence the entire congruence class $\overline{p}/\gamma$ is embedded in a product of copies of the 2-element chain $2 < 3$. This implies that $\overline{p}/\gamma$ is a distributive lattice. Any finite distributive lattice is uniquely determined by its ordered set of join-irreducible elements.

We consider the simplest case in which $|X| = 3$ and $g : \mathbf{F}_\mathcal{V}(3) \to \mathbf{F}_\mathcal{D}(3)$. The cardinality of $\mathbf{F}_\mathcal{D}(3)$ is 18. So $\gamma = \ker g$ has 18 congruence classes. An examination of the 18 elements shows that 11 have the property that $p_* = p^*$ and so their γ classes are singletons. The remaining seven classes $\overline{p}/\gamma$ have as p_* either the lower median term $(x_1 \wedge x_2) \vee (x_1 \wedge x_3) \vee (x_2 \wedge x_3)$ or

the six duals or permutations of $x_1 \vee (x_2 \wedge x_3)$. There are six valuations of $\{x_1, x_2, x_3\}$ into $\mathbf{N}$, so $\mathbf{F}_{\mathcal{V}}(X)$ is in $\mathbf{N}^6$ and each class $\overline{p}/\gamma$ is a sublattice of the distributive lattice $\mathbf{C}_2^6$. The join-irreducible elements in the γ class of the lower median term can be shown to form a six-element antichain. So this class is lattice isomorphic to $(\mathbf{C}_2)^6$. Each of the other six congruence classes has a two-element antichain for the ordered set of join-irreducible elements. So each of these congruence classes is isomorphic to $\mathbf{C}_2 \times \mathbf{C}_2$. So all the 18 congruence classes are Boolean lattices of cardinality 1, 4 or 64. Thus, the cardinality of $\mathbf{F}_{\mathcal{V}}(3)$ is $11 + 64 + 6 * 4 = 99$. The classes interact with one another as in the 18-element free distributive lattice. This analysis provides a reasonably good description of the the structure of $\mathbf{F}_{\mathcal{V}}(3)$.

If we next consider $\mathbf{F}_{\mathcal{V}}(4)$, then the basic analysis is the same. The distributive lattice $\mathbf{F}_{\mathcal{D}}(4)$ has 166 elements. Of the 166 congruence classes of γ there are 26 for which $p_* = p^*$. The 140 classes that are not singletons can be grouped by duality and permutation of variables into 12 families. For each of these12 the ordered set of join-irreducibles can be determined. These ordered sets range in size from 6 to 36 elements. Unlike the case for $\mathbf{F}_{\mathcal{V}}(3)$, these ordered sets are not particularly well-behaved. Even though the distributive lattices that have these ordered sets as their set of join-irreducible elements can be found, the structure of each class is not very transparent. So although the cardinality of $\mathbf{F}_{\mathcal{V}}(4)$ can be determined by this method, it is $540,792,672$, the structure of this lattice can only be described in fairly general terms because of the lack of regularity in the structure of the congruence classes of γ.

If we consider the variety $\mathcal{W}$ generated by the 5-element modular nondistributive lattice, then a similar analysis leads to similar results. The lattice $\mathbf{F}_{\mathcal{W}}(3)$ has 28 elements and each of the nonsingleton classes of γ is well-behaved. But for $\mathbf{F}_{\mathcal{W}}(4)$, although the method of mapping down to $\mathbf{F}_{\mathcal{D}}(4)$ allows us to find the cardinality, which is $19,982$, the structure of each class of γ is far from transparent.

One difficulty with the method used in Example 2.1, and with the more general technique of partitioning $\mathbf{F}_{\mathcal{V}}(X)$ into congruence classes determined by the canonical map of $\mathbf{F}_{\mathcal{V}}(X)$ onto $\mathbf{F}_{\mathcal{W}}(X)$ for a subvariety $\mathcal{W}$ of $\mathcal{V}$, is that the congruence classes are not 'free' in any obvious sense. We conclude this section with a discussion of a method in which $\mathbf{F}_{\mathcal{V}}(X)$ is decomposed into congruence classes, and each class is a free object in a variety intimately related to $\mathcal{V}$.

Let $\mathcal{W}$ be an arbitrary variety in which there are no constant symbols. As in Example 2.1 we consider an equivalence relation $\overset{v}{\sim}$ defined on a free algebra $\mathbf{F}_{\mathcal{W}}(X)$ by $\overline{p} \overset{v}{\sim} \overline{q}$ if and only if $\mathrm{var}(p) = \mathrm{var}(q)$. When is $\overset{v}{\sim}$ a congruence relation? A sufficient condition is that if $\mathcal{W} \models s \approx t$, then $\mathrm{var}(s) = \mathrm{var}(t)$. Identities of this form are called *regular* and a variety that is presented by regular identities is called a *regular variety*. If $\overset{v}{\sim}$ is a congruence relation of $\mathbf{F}_{\mathcal{W}}(X)$, then the quotient algebra $\mathbf{S} = \mathbf{F}_{\mathcal{W}}(X)/\overset{v}{\sim}$ can be represented as an algebra whose universe is the set of nonvoid subsets of X and if f is an k-ary operation symbol, then $f^{\mathbf{S}}(S_1, \ldots, S_k) = S_1 \cup \cdots \cup S_k$ for $S_1, \ldots, S_k \subseteq X$. So $\mathbf{S}$ is term equivalent to a join semilattice. We let $\mathcal{S}$ denote the variety of join semilattices but presented with the similarity type of $\mathcal{W}$. Then the canonical homomorphism $g : \mathbf{F}_{\mathcal{W}}(X) \to \mathbf{F}_{\mathcal{S}}(X)$ determined by $g(\overline{x}_i) = \overline{x}_i$ is such that the congruence relation $\ker g$ is $\overset{v}{\sim}$. Note that each congruence class is in fact a subuniverse of $\mathbf{F}_{\mathcal{W}}(X)$. The classes of $\ker g$ interact with each other according to the join semilattice structure on $\mathbf{F}_{\mathcal{S}}(X)$. This interaction is very well-behaved and transparent since it can be viewed as the operation of union on the set of nonvoid subsets of X. We describe a general situation in which the individual classes of $\ker g$ will have a well-behaved internal

structure as well. In this situation the structure of $\mathbf{F}_{\mathcal{V}}(X)$ admits a reasonable description.

If $\mathcal{V}$ is an arbitrary variety, then we can form the *regularization of* $\mathcal{V}$, denoted $R(\mathcal{V})$, which is the variety axiomatized by all the regular identities of $\mathcal{V}$. The variety $\mathcal{V}$ is a subvariety of $R(\mathcal{V})$. We investigate the structure of the free algebras of $R(\mathcal{V})$ by considering the canonical map of $\mathbf{F}_{R(\mathcal{V})}(X)$ onto $\mathbf{F}_{\mathcal{V}}(X)$.

Let $\mathcal{V}$ be a finitely generated variety that is not regular and that has no constant symbols in its similarity type. Since $\mathcal{V}$ is not regular, there is an $n \geq 2$ and there are terms s and t with $\mathrm{var}(s) = \{x_1, \ldots, x_n\}$ and $\mathrm{var}(t) = \{x_1, \ldots, x_i\}$ for $i < n$ such that $\mathcal{V} \models s \approx t$. Identifying $x_1, \ldots, x_i$ with x and identifying $x_{i+1}, \ldots, x_n$ with y, we have $\mathcal{V} \models s(x, \ldots, x, y \ldots, y) \approx t(x, \ldots, x)$. If we restrict our investigation to the case that $\mathcal{V} \models t(x, \ldots, x) \approx x$, then in this situation we then have a binary term $p(x, y) = s(x, \ldots, x, y, \ldots, y)$ for which $\mathcal{V} \models p(x, y) \approx x$ witnesses that $\mathcal{V}$ is not a regular variety.

The following theorem of J. Płonka [19] gives the structural decomposition of free algebras in a variety that is the regularization of an irregular variety.

> *Let $\mathcal{V}$ be a variety, with no constant symbols, that satisfies an irregular identity of the form $p(x, y) \approx x$. Let $\mathcal{W} = R(\mathcal{V})$ be the regularization of $\mathcal{V}$. Then for $\overline{p} \in \mathbf{F}_{\mathcal{W}}(X)$ the congruence class of $\overline{p}/\overset{v}{\sim}$ is a subalgebra of $\mathbf{F}_{\mathcal{W}}(X)$ that is isomorphic to $\mathbf{F}_{\mathcal{V}}(\mathrm{var}(p))$.*

An immediate corollary of this theorem is that if $\mathcal{V}$ is also locally finite then

$$|\mathbf{F}_{R(\mathcal{V})}(n)| = \sum_{i=1}^{n} \binom{n}{i} |\mathbf{F}_{\mathcal{V}}(i)|.$$

We present a constructive proof of Płonka's theorem in the case that the variety $\mathcal{V}$ is finitely generated. We do this by representing $\mathbf{F}_{R(\mathcal{V})}(X)$ as $\mathbf{Ge}(X, U)$ for a suitable set U of valuations.

If $\mathbf{A}$ is an algebra and $0 \notin A$, then $\mathbf{A}^*$ denotes the algebra having the same similarity type as $\mathbf{A}$, with universe $A \cup \{0\}$, and operations given by

$$f^{\mathbf{A}^*}(a_1, \ldots, a_k) = \begin{cases} 0 & \text{if } 0 \in \{a_1, \ldots, a_k\} \\ f^{\mathbf{A}}(a_1, \ldots, a_k) & \text{otherwise.} \end{cases}$$

The element 0 is called an *absorbing element* of $\mathbf{A}^*$.

Let $\mathbf{A}$ be a finite algebra with no constant symbols in its similarity type and suppose $\mathbf{A} \models p(x, y) \approx x$ for a term p with $\mathrm{var}(p) = \{x, y\}$. If $\mathcal{V}$ is the variety generated by $\mathbf{A}$, then $\mathcal{V}$ satisfies the hypotheses of Płonka's theorem. By results of H. Lakser, R. Padmanabhan, and C. Platt [17], the variety $R(\mathcal{V})$ is generated by the algebra $\mathbf{A}^*$. Therefore, the algebra $\mathbf{F}_{R(\mathcal{V})}(X)$ is isomorphic to $\mathbf{Ge}(X, U)$ where U consists of all $v : X \to \mathbf{A}^*$. Let Z be any nonvoid subset of X. Without loss of generality we let $Z = \{x_1, \ldots, x_m\}$. Form

$$U_Z = \{v \in U : v(x_i) \in A \text{ for all } x_i \in Z\}.$$

For $1 \leq i \leq m$ let y_i denote the term $p(x_i, p(x_1, p(x_2, p(\ldots, p(x_{m-1}, x_m) \ldots))))$. Then $\mathcal{V} \models x_i \approx y_i$ and $\mathrm{var}(y_i) = Z$. For $v \in U_Z$ and $x_i \in Z$ we have $v(x_i) = v(y_i)$. If $w \in U - U_Z$, then there is an $x_j \in Z$ for which $w(x_j) = 0$. Hence $w(y_i) = 0$ for all $1 \leq i \leq m$. Let

t be any term with $\mathrm{var}(t) = Z$. We write t as $t(x_1, \ldots, x_m)$. For all $v \in U_Z$ we have $v(t) = v(t(x_1, \ldots, x_m)) = v(t(y_1, \ldots, y_m)) \in A$. For all $w \in U - U_Z$ we have $w(t) = 0$. From these observations it follows that the congruence class and subuniverse $\overline{t}/\overset{v}{\sim}$ is generated by $\overline{y}_1, \ldots, \overline{y}_m$. Moreover, $\overline{t}/\overset{v}{\sim}$ can be embedded in $\mathbf{Ge}(X, U_Z)$ since $w(s) = w(t)$ for every $w \in U - U_Z$ and s having $\mathrm{var}(s) = Z$. Now $\overline{y}_i$ and $\overline{x}_i$ agree on all $v \in U_Z$. So the subalgebra of $\mathbf{Ge}(X, U_Z)$ generated by $\overline{y}_1, \ldots, \overline{y}_m$ is isomorphic to the subalgebra of $\mathbf{Ge}(X, U_Z)$ generated by $Z = \{x_1, \ldots, x_m\}$. This latter algebra is isomorphic to $\mathbf{Ge}(Z, A^Z)$. Since $\mathbf{Ge}(Z, A^Z) \cong \mathbf{F}_{\mathcal{V}}(Z)$ we conclude that $\overline{t}/\overset{v}{\sim}$ is isomorphic to $\mathbf{F}_{\mathcal{V}}(Z)$.

For example, let $\mathcal{B}$ be the variety of Boolean algebras considered in Example 1.9. The variety $\mathcal{B}$ is generated by the 2-element Boolean algebra $\mathbf{B}_2$. We can eliminate the constant symbols 0 and 1 by using the unary terms $x_1 \wedge x_1'$ and $x_1 \vee x_1'$ in their stead. If $p(x, y)$ denotes the term $(x \vee y) \wedge x$, then $\mathcal{B} \models p(x, y) \approx x$. If $\mathcal{W}$ denotes the regularization $R(\mathcal{B})$ of Boolean algebras, then as seen in the previous paragraph, $\mathcal{W}$ is generated by the 3-element algebra obtained by adjoining an absorbing element to $\mathbf{B}_2$. Each congruence class $\overline{t}/\overset{v}{\sim}$ in $\mathbf{F}_{\mathcal{W}}(X)$ is isomorphic to the free Boolean algebra $\mathbf{F}_{\mathcal{B}}(\mathrm{var}(t))$. For $\overline{s}, \overline{t} \in \mathbf{F}_{\mathcal{W}}(X)$ and term $q(x, y)$, we have $q(\overline{s}, \overline{t})/\overset{v}{\sim} = \overline{r}/\overset{v}{\sim}$ where r is any term with $\mathrm{var}(r) = \mathrm{var}(s) \cup \mathrm{var}(t)$. As observed in Example 1.9 $|\mathbf{F}_{\mathcal{B}}(m)| = 2^{2^m}$. Thus,

$$\mathbf{F}_{R(\mathcal{B})}(n) = \sum_{i=1}^{n} \binom{n}{i} 2^{2^i}.$$

3 Structure via inclusion-exclusion

In the previous section we investigated the structure of a free algebra by decomposing it into the disjoint blocks of a congruence relation that is the kernel of a homomorphism onto a free algebra in a particular subvariety. In this section we decompose a free algebra into overlapping canonically defined subalgebras. The homogeneity of these subalgebras allows for the inclusion-exclusion principle to be used to express the cardinality of the free algebra in terms of the cardinalities of these subalgebras. We present some examples of varieties in which the subalgebras have some interesting structure of their own and for which we can either determine their cardinality or else express the cardinality in terms of the cardinalities of some sets of some familiar combinatorial structures.

Throughout this section we let $X = \{x_1, \ldots, x_n\}$ be a finite set of variables. For a language (or similarity type) $\mathcal{L}$ let $\mathrm{T}_{\mathcal{L}}(X)$ denote the set of all terms that can be built from X using operation symbols from $\mathcal{L}$. We often write $\mathrm{T}(n)$ for $\mathrm{T}_{\mathcal{L}}(X)$. In this section $\mathcal{L}$ will always be a language in which there are no constant symbols. Since a nullary constant operation can always be represented by a constant unary operation, this restriction on $\mathcal{L}$ does not result in any loss of generality.

3.1 Definition For $t \in \mathrm{T}_{\mathcal{L}}(X)$ we define the *right-most variable* $\mathrm{rv}(t)$ *of* t inductively as follows:

$$\mathrm{rv}(t) = \begin{cases} x_i & \text{if } t = x_i \in X \\ \mathrm{rv}(t_m) & \text{if } t = f(t_1, \ldots, t_m) \text{ for } f \in \mathcal{L} \text{ and } t_1, \ldots, t_m \in \mathrm{T}_{\mathcal{L}}(X). \end{cases}$$

For $1 \leq i \leq n$, let $\mathrm{T}_i(n) = \{t \in \mathrm{T}(n) : \mathrm{rv}(t) = x_i\}$ and let $\overline{\mathrm{T}}_i(n)$ be the set of elements

$\{\overline{t} \in \mathbf{F}_{\mathcal{V}}(n) : t \in \mathrm{T}_i(n)\}$. It is easily seen that each $\overline{\mathrm{T}}_i(n)$ is a subuniverse of $\mathbf{F}_{\mathcal{V}}(X)$; we write $\overline{\mathbf{T}}_i(n)$ for the corresponding subalgebra of $\mathbf{F}_{\mathcal{V}}(n)$.

Note that the algebras $\overline{\mathbf{T}}_i(n)$ need not be disjoint. We use the $\overline{\mathbf{T}}_i(n)$ in our decomposition of $\mathbf{F}_{\mathcal{V}}(X)$. The intersections of the $\overline{\mathrm{T}}_i(n)$ will also play a role. We write $\overline{\mathrm{T}}_{\leq \ell}(n)$ for $\bigcap_{1 \leq i \leq \ell} \overline{\mathrm{T}}_i(n)$. Each $\overline{\mathrm{T}}_{\leq \ell}(n)$ is a subuniverse and $\overline{\mathbf{T}}_{\leq \ell}(n)$ denotes the corresponding subalgebra.

All the $\overline{\mathbf{T}}_i(n)$ are isomorphic to each other. Every permutation of X extends to an automorphism of $\mathbf{F}_{\mathcal{V}}(X)$. From this it follows that if $1 \leq i_1 < \cdots < i_\ell \leq n$, then $\overline{\mathbf{T}}_{i_1}(n) \cap \overline{\mathbf{T}}_{i_2}(n) \cap \cdots \cap \overline{\mathbf{T}}_{i_\ell}(n) \cong \overline{\mathbf{T}}_{\leq \ell}(n)$.

3.2 Theorem *Let $\mathcal{V}$ be a locally finite variety in a similarity type that has no constant symbols. For every positive integer n we have*

$$|\mathbf{F}_{\mathcal{V}}(n)| = \sum_{\ell=1}^{n} (-1)^{\ell-1} \binom{n}{\ell} |\overline{\mathrm{T}}_{\leq \ell}(n)|.$$

Proof First note that $\mathcal{V}$ is locally finite, and hence $\mathrm{F}_{\mathcal{V}}(n)$ is a finite set. Every element of $\mathbf{F}_{\mathcal{V}}(n)$ is of the form $\overline{t}$ for some $t \in \mathrm{T}(n)$. There are no constant symbols in the similarity type of $\mathcal{V}$ and so t has a right–most variable $\mathrm{rv}(t)$, that is, $t \in \mathrm{T}_i(n)$ for some i. Thus,

$$\mathrm{F}_{\mathcal{V}}(n) = \bigcup_{1 \leq i \leq n} \overline{\mathrm{T}}_i(n).$$

Applying the inclusion-exclusion principle, we have

$$|\mathbf{F}_{\mathcal{V}}(n)| = \sum \{(-1)^{\ell-1} |\overline{\mathrm{T}}_{i_1}(n) \cap \overline{\mathrm{T}}_{i_2}(n) \cap \cdots \cap \overline{\mathrm{T}}_{i_\ell}(n)| : 1 \leq i_1 < \cdots < i_\ell \leq n\}.$$

As we have observed, $\overline{\mathbf{T}}_{i_1}(n) \cap \overline{\mathbf{T}}_{i_2}(n) \cap \cdots \cap \overline{\mathbf{T}}_{i_\ell}(n)$ is isomorphic to $\overline{\mathbf{T}}_{\leq \ell}(n)$ so

$$|\overline{\mathrm{T}}_{i_1}(n) \cap \overline{\mathrm{T}}_{i_2}(n) \cap \cdots \cap \overline{\mathrm{T}}_{i_\ell}(n)| = |\overline{\mathrm{T}}_{\leq \ell}(n)|.$$

For each $1 \leq \ell \leq n$ there are $\binom{n}{\ell}$ sets of the form $\overline{\mathrm{T}}_{i_1}(n) \cap \overline{\mathrm{T}}_{i_2}(n) \cap \cdots \cap \overline{\mathrm{T}}_{i_\ell}(n)$, each of size $|\overline{\mathrm{T}}_{\leq \ell}(n)|$, thereby yielding the desired formula for $|\mathbf{F}_{\mathcal{V}}(n)|$. □

Although the formal appearance of Theorem 3.2 is appealing, it is not immediately clear if the result has any content. For example, in any variety of groups or lattices, $\overline{\mathrm{T}}_i(n) = \mathrm{F}_{\mathcal{V}}(n)$ for every $1 \leq i \leq n$. For such varieties the theorem tells us nothing. For other varieties, although the $\overline{\mathrm{T}}_i(n)$ are proper subsets of $\mathrm{F}_{\mathcal{V}}(n)$, the structure of the subalgebra $\overline{\mathbf{T}}_i(n)$ is difficult to determine. Nonetheless, there are varieties in which the algebras $\overline{\mathbf{T}}_i(n)$ have an interesting describable structure and for which Theorem 3.2 can be used to provide significant information about the cardinalities of finitely generated free algebras. In what follows we present two varieties in which this is the case. Each is based on an algebraic coding of ordered sets as presented in [3, 4].

3.3 Definition Let $P = \langle P, \leq, 1 \rangle$ be a ordered set with a top element 1. We form the algebra $\mathbf{H}(P) = \langle P, \cdot \rangle$ with universe P and binary operation $\cdot$ given by

$$x \cdot y = \begin{cases} 1 & \text{if } x \leq y \\ y & \text{otherwise,} \end{cases}$$

and the algebra $\mathbf{J}(P) = \langle P, \cdot \rangle$ with universe P and binary operation $\cdot$ given by

$$x \cdot y = \begin{cases} y & \text{if } x \leq y \\ 1 & \text{otherwise.} \end{cases}$$

By $\mathcal{H}$ and $\mathcal{J}$ we denote the varieties generated by all $\mathbf{H}(P)$ and all $\mathbf{J}(P)$, where P ranges over all ordered sets with a top element.

The variety $\mathcal{H}$ is a subvariety of the variety of all Hilbert algebras, which is the variety generated by all $\{\rightarrow, 1\}$-subreducts of the varieties of Brouwerian semilattices. Hilbert algebras have received considerable attention in the algebraic logic literature. A standard reference on the basic properties of Hilbert algebras is A. Diego's monograph [9]. A Hilbert algebra includes the constant 1 in the similarity type. However, since Hilbert algebras satisfy the identity $x \rightarrow x = 1$, this constant can be omitted from the similarity type. For $\mathcal{H}$ we write $\cdot$ for the binary operation symbol $\rightarrow$ and omit the constant 1 from the similarity type. Members of $\mathcal{J}$ are called *join algebras*. The only work on join algebras that I am aware of is [4].

It is not hard to argue that every algebra in $\mathcal{H}$ or in $\mathcal{J}$ has an equationally definable order relation. For elements in an algebra in $\mathcal{H}$ we have $x \leq y$ if and only if $xy = xx$ while for algebras in $\mathcal{J}$ the order is given by $x \leq y$ if and only if $xy = y$.

3.4 Definition An algebra $\mathbf{A}$ in $\mathcal{H}$ is called *pure* if it is of the form $\mathbf{H}(P)$ for some ordered set P having a top element. A pure algebra in $\mathcal{J}$ is defined analogously.

Both $\mathcal{H}$ and $\mathcal{J}$ are generated by their pure members. Both $\mathcal{H}$ and $\mathcal{J}$ are known to be locally finite varieties. The subdirectly irreducible algebras in each variety have been characterized: they are the pure algebras with underlying ordered set P of the form $Q \oplus 1$, where Q is an arbitrary ordered set.

There are some differences between $\mathcal{H}$ and $\mathcal{J}$. Notably, $\mathcal{H}$ is a congruence distributive variety and its type set, in the sense of tame congruence theory, is $\{\mathbf{3}\}$ while $\mathcal{J}$ satisfies no nontrivial congruence identities and its type set is $\{\mathbf{5}\}$.

Another important difference between $\mathcal{H}$ and $\mathcal{J}$ is that the variety $\mathcal{J}$ of join algebras is generated by $\mathbf{J}(C)$, where C is the 3-element chain. The only proper nontrivial subvariety of $\mathcal{J}$ is the variety of semilattices. These results are proved in [4]. The variety $\mathcal{H}$ on the other hand, is not finitely generated and has $2^{\aleph_0}$ subvarieties. The least nontrivial subvariety of $\mathcal{H}$ is the variety $\mathcal{I}$ of implication algebras discussed in Example 1.10. These results and others for $\mathcal{H}$ are also given in [4].

We investigate the structure of free algebras $\mathbf{F}_{\mathcal{H}}(n)$ and $\mathbf{F}_{\mathcal{J}}(n)$ by describing the subalgebras $\overline{\mathbf{T}}_i(n)$.

3.5 Definition Let $t \in \mathrm{T}_{\mathcal{L}}(n)$ be an arbitrary term for $\mathcal{L} = \{\cdot\}$. The binary relation $\mathrm{re}(t)$ on $\mathrm{var}(t)$ is defined inductively by:

$$\mathrm{re}(t) = \begin{cases} \{(x_i, x_i)\} & \text{if } t = x_i \\ \mathrm{re}(p) \cup \mathrm{re}(q) \cup \{(\mathrm{rv}(p), \mathrm{rv}(q))\} & \text{if } t = p \cdot q. \end{cases}$$

The transitive closure of $\mathrm{re}(t)$ is denoted $\mathrm{qo}(t)$.

A *quasi-order* on a set Z is any reflexive transitive binary relation on Z. If σ is a quasi-order on Z, then the quotient Z/σ is an ordered set in which $a/\sigma \leq b/\sigma$ if and only if $(a,b) \in \sigma$. The ordered set Z/σ has a top element if and only if there is a $z \in Z$ such that $(a,z) \in \sigma$ for all $a \in Z$. If $t \in \mathrm{T}_{\mathcal{L}}(n)$, then $\mathrm{qo}(t)$ is a quasi-order on the set $\mathrm{var}(t)$ in which $\mathrm{rv}(t)/\mathrm{qo}(t)$ is the top element of the ordered set $\mathrm{var}(t)/\mathrm{qo}(t)$.

3.6 Lemma *Let σ be any quasi-order on a set $Y \subseteq X$ such that Y/σ has a top element. Then there exists $t \in \mathrm{T}_{\mathcal{L}}(X)$ such that $\mathrm{qo}(t) = \sigma$.*

Proof Let $(x_{i_1}, x_{j_1}), (x_{i_2}, x_{j_2}), (x_{i_3}, x_{j_3}), \ldots, (x_{i_k}, x_{j_k})$ be any list of the elements of σ subject only to the constraint that x_{j_1}/σ is the top element of Y/σ. Let

$$t = (x_{i_k} x_{j_k})(\ldots(x_{i_3} x_{j_3})((x_{i_2} x_{j_2})(x_{i_1} x_{j_1}))\ldots).$$

Clearly $\mathrm{var}(t) = Y$. It is immediate from the recursive definition of the operator qo that $\mathrm{qo}(t) = \sigma$. □

We now restrict our discussion to the variety $\mathcal{J}$. The next lemma is proved by considering valuations of X into the 3-element pure join algebra $\mathbf{J}(C)$, which generates $\mathcal{J}$.

3.7 Lemma *Let $s, t \in \mathrm{T}_{\mathcal{L}}(n)$.*

(1) *$\mathcal{J} \models s \approx t$ if and only if $\mathrm{qo}(s) = \mathrm{qo}(t)$.*

(2) *$\mathcal{J} \models st \approx t$ if and only if $\mathrm{qo}(s) \subseteq \mathrm{qo}(t)$.*

We have $\overline{\mathrm{T}}_i(n) = \{\overline{t} \in \mathbf{F}_{\mathcal{J}}(n) : \mathrm{rv}(t) = x_i\}$ and that $\overline{\mathbf{T}}_i(n)$ is the subalgebra of $\mathbf{F}_{\mathcal{J}}(n)$ with universe $\overline{\mathrm{T}}_i(n)$. It can be argued that $\mathcal{J}$ is a regular variety (in the sense of the previous section), that is, if $\mathcal{J} \models s \approx t$, then $\mathrm{var}(s) = \mathrm{var}(t)$. In particular, if $\overline{s} \in \overline{\mathrm{T}}_i(n)$, then $x_i \in \mathrm{var}(s)$.

For $1 \leq i \leq n$

$$\mathrm{Q}_i(n) = \{\sigma : \exists Y \subseteq X,\ x_i \in Y,\ \sigma \text{ is a quasi-order on } Y, \text{ and } x_i/\sigma \text{ is the top element of } Y/\sigma\}.$$

For σ_1 and $\sigma_2 \in \mathrm{Q}_i(n)$ with σ_k defined on $Y_k \subseteq X$ for $k = 1, 2$, let $\sigma_1 \cdot \sigma_2$ denote the smallest quasi-order on $Y_1 \cup Y_2$ that contains both σ_1 and σ_2. Then necessarily $\sigma_1 \cdot \sigma_2 \in \mathrm{Q}_i(n)$ and $\sigma_1 \cdot \sigma_2$ is the transitive closure of $\sigma_1 \cup \sigma_2$. Let $\mathbf{Q}_i(n)$ denote the algebra $\langle \mathrm{Q}_i(n), \cdot \rangle$. We note that the algebra $\mathbf{Q}_i(n)$ is, in fact, a semilattice since the operation $\cdot$ on $\mathrm{Q}_i(n)$ is idempotent, commutative, and associative.

3.8 Theorem *The map $\overline{t} \mapsto \mathrm{qo}(t)$ is an isomorphism from $\overline{\mathbf{T}}_i(n)$ onto $\mathbf{Q}_i(n)$.*

Proof Lemmas 3.6 and 3.7 show the map is well-defined, one-to-one and onto $\mathrm{Q}_i(n)$. It remains to show that $\mathrm{qo}(s \cdot t) = \mathrm{qo}(s) \cdot \mathrm{qo}(t)$ for all $\overline{s}$ and $\overline{t}$ in $\overline{\mathrm{T}}_i(n)$. Let $\mathrm{rv}(s) = x_\ell$ and $\mathrm{rv}(t) = x_m$. By the definition of qo we have that $\mathrm{qo}(s \cdot t)$ is the transitive closure of $\mathrm{qo}(s) \cup \mathrm{qo}(t) \cup \{(x_\ell, x_m)\}$. We have $(x_\ell, x_i) \in \mathrm{qo}(s)$ since $\overline{s} \in \overline{\mathrm{T}}_i(n)$. Also $(x_i, x_m) \in \mathrm{qo}(t)$ since $t \in \overline{\mathrm{T}}_i(n)$ and $\mathrm{rv}(t) = x_m$. So (x_ℓ, x_m) is in the transitive closure of $\mathrm{qo}(s) \cup \mathrm{qo}(t)$. Hence the transitive closure of $\mathrm{qo}(s) \cup \mathrm{qo}(t) \cup \{(x_\ell, x_m)\}$ is the same as the transitive closure of $\mathrm{qo}(s) \cup \mathrm{qo}(t)$, which is $\mathrm{qo}(s) \cdot \mathrm{qo}(t)$. □

From Theorem 3.8 it follows that $\mathbf{Q}_i(n)$ is a join algebra. The induced order on $\mathrm{Q}_i(n)$ is that of containment. Therefore, $\overline{s} \leq \overline{t}$ in $\overline{\mathbf{T}}_i(n)$ if and only if $\mathrm{qo}(s) \subseteq \mathrm{qo}(t)$. Actually, from Lemma 3.7 we see that for arbitrary $\overline{s}, \overline{t} \in \overline{\mathrm{T}}_i(n)$ it is the case that $\overline{s} \leq \overline{t}$ if and only if $\mathrm{qo}(s) \subseteq \mathrm{qo}(t)$.

Let q_k denote the number of quasi-orders on $\{1, 2, \ldots, k\}$. It is known that q_k is also equal to the number of topologies on $\{1, 2, \ldots, k\}$. For example, the values of q_k for $k = 1, 2, 3, 4$ are $1, 4, 29, 355$ respectively. We let $q_0 = 1$.

3.9 Theorem

$$|\mathrm{F}_{\mathcal{J}}(n)| = \sum_{i=0}^{n-1} q_i \binom{n}{i} (2^{n-i} - 1).$$

Proof We have

$$|\mathbf{F}_{\mathcal{J}}(n)| = \sum_{\ell=1}^{n} (-1)^{(\ell-1)} \binom{n}{\ell} |\overline{\mathrm{T}}_{\leq \ell}(n)|$$

by Theorem 3.2. The cardinality of $\overline{\mathbf{T}}_{\leq \ell}(n)$ is equal to $|\mathrm{Q}_1(n) \cap \cdots \cap \mathrm{Q}_\ell(n)|$ by Theorem 3.8. Let σ be a quasi-order on a set $Y \subseteq X$ for which Y/σ has a top element. Let $Z \subseteq Y$ be be the set of all x_i for which x_i/σ is the top element in Y/σ. Then $\sigma = \tau \cup (Y \times Z)$ where τ is a quasi-order on $Y - Z$. Conversely, if $\emptyset \neq Z \subseteq Y \subseteq X$ and τ is a quasi-order on $Y - Z$, then $\sigma = \tau \cup (Y \times Z)$ is a quasi-order on Y for which Y/σ has a maximal element consisting of x_k/σ for every $x_k \in Z$. For $\sigma \in \mathrm{Q}_1(n) \cap \cdots \cap \mathrm{Q}_\ell(n)$ there are $\binom{n-\ell}{i}$ choices for the set $Y - Z$ if $|Y - Z| = i$. For a given i, with $0 \leq i \leq n - \ell$, there are q_i choices for a quasi-order on $Y - Z$. The set Z must contain $\{x_1, \ldots, x_\ell\}$ so there are $2^{n-\ell-i}$ ways to choose the remaining elements of Z. Therefore

$$|\mathrm{Q}_1(n) \cap \cdots \cap \mathrm{Q}_\ell(n)| = \sum_{i=0}^{n-\ell} q_i \binom{n-\ell}{i} 2^{n-\ell-i}.$$

This implies

$$|\mathbf{F}_{\mathcal{J}}(n)| = \sum_{\ell=1}^{n} (-1)^{(\ell-1)} \binom{n}{\ell} \left(\sum_{i=0}^{n-\ell} q_i \binom{n-\ell}{i} 2^{n-\ell-i} \right).$$

We interchange the order of summation and use the formula $\binom{n}{\ell}\binom{n-\ell}{i} = \binom{n}{i}\binom{n-i}{n-\ell-i}$ to obtain

$$|\mathbf{F}_{\mathcal{J}}(n)| = \sum_{i=0}^{n-1} q_i \binom{n}{i} \left(\sum_{\ell=1}^{n-i} (-1)^{\ell-1} \binom{n-i}{n-\ell-i} 2^{n-\ell-i} \right).$$

The inner summation simplifies to $2^{n-i} - 1$. Thus,

$$|\mathbf{F}_{\mathcal{J}}(n)| = \sum_{i=0}^{n-1} q_i \binom{n}{i} (2^{n-i} - 1).$$

□

For example, if we use the previously mentioned values of q_i for $1 \leq i \leq 4$, then we see that the values of $|\mathrm{F}_{\mathcal{J}}(n)|$ for $1 \leq n \leq 5$ are 1, 5, 28, 231 and 3031 respectively.

The formula in Theorem 3.9 involves q_n, the number of quasi-orders on an n-element set. Although no simple formula for the value of q_n is known, asymptotic estimates do exist. By means of these asymptotics we can obtain the following bounds on the cardinalities of free join algebras.

3.10 Theorem *There exists a positive constant c such that for all sufficiently large n,*

$$2^{n^2/4+n-c\lg n} \leq |\mathrm{F}_{\mathcal{J}}(n)| \leq 2^{n^2/4+n+c\lg n}.$$

We next consider the structure of free algebras $\mathbf{F}_{\mathcal{V}}(n)$ for $\mathcal{V}$ an arbitrary subvariety of $\mathcal{H}$. As with the variety $\mathcal{J}$ we provide a detailed description of the subalgebras $\overline{\mathbf{T}}_i(n)$. From this description we show that $\overline{\mathbf{T}}_{\leq\ell}(n)$ is a direct power of the 2-element implication algebra $\mathbf{C}_2$, and we determine the exponent in this direct power. So with Theorem 3.2 we can, in principle, find an expression for the cardinality of $\mathbf{F}_{\mathcal{V}}(n)$ when $\mathcal{V}$ is a subvariety of $\mathcal{H}$.

Recall that a subdirectly irreducible algebra $\mathbf{A} \in \mathcal{H}$ has an element $e \prec 1$ with $a \leq e$ for all $a \in A, a \neq 1$. The element e is irreducible in the sense that if $t^{\mathbf{A}}(a_1, \ldots, a_n) = e$, then e appears among the a_i. Hence if v is a valuation mapping to the subdirectly irreducible algebra $\mathbf{A}$, then $e \in v(X)$. We also note that the set $\{1, e\}$ is a subuniverse of $\mathbf{A}$ and the corresponding subalgebra is isomorphic to the 2-element implication algebra $\mathbf{C}_2$. The algebra $\mathbf{C}_2$ is a pure algebra in $\mathcal{H}$ since it is $\mathbf{H}(P)$ for P the 2-element chain.

The following facts about valuations into pure algebras in $\mathcal{H}$ are easily established using nothing more than definitions.

3.11 Lemma *Let v be a valuation into a pure algebra $\mathbf{A}$ in $\mathcal{H}$ and let t be a term in $\mathrm{T}_{\mathcal{L}}(n)$ with $\mathrm{rv}(t) = x_i$.*

(1) $v(t) \in \{1, v(\mathrm{rv}(t))\}$.

(2) *If $\overline{t} \in \overline{\mathrm{T}}_i(n) \cap \overline{\mathrm{T}}_j(n)$ and $v(t) \neq 1$, then $v(x_i) = v(x_j)$.*

(3) $\mathcal{H} \models \mathrm{rv}(t) \leq t$.

Let $\mathcal{V}$ be a subvariety of $\mathcal{H}$ and $\mathcal{V}_{\mathrm{SI}}$ the class of finitely generated subdirectly irreducible algebras in $\mathcal{V}$. Since $\mathcal{V}$ is locally finite, every algebra in $\mathcal{V}_{\mathrm{SI}}$ is finite. By Corollary 1.4 the set $\mathrm{val}(X, \mathcal{V}_{\mathrm{SI}})$ is free for $\mathcal{V}$. Hence $E(\mathrm{val}(X, \mathcal{V}_{\mathrm{SI}}))$ is also free for $\mathcal{V}$ by Lemma 1.7.

Now, from Lemma 3.11, if v is a valuation in $E = E(\mathrm{val}(X, \mathcal{V}_{\mathrm{SI}}))$ and $t \in \mathrm{T}_i(n)$, then $v(t) \in \{1, v(x_i)\}$. If $\overline{s}, \overline{t} \in \overline{\mathrm{T}}_i(n)$ with $\overline{s} \neq \overline{t}$, then there exists $v \in E$ with $v(x_i) \leq e$ and $\{v(s), v(t)\} = \{1, v(x_i)\}$. By composing v with an endomorphism of $\mathbf{Alg}(v)$ that maps $v(x_i)$ to e we can find a $w \in E$ for which $\{v(s), v(t)\} = \{1, e\}$.

3.12 Definition Let $\mathcal{V}$ be a subvariety of $\mathcal{H}$. We adopt the following notation:

$$E(\mathcal{V}) := E(\mathrm{val}(X, \mathcal{V}_{\mathrm{SI}})).$$

For $1 \leq \ell \leq n$,

$$E(\mathcal{V})_\ell := \{v \in E(\mathrm{val}(X, \mathcal{V}_{\mathrm{SI}})) : v(x_\ell) = e\}$$

and

$$E(\mathcal{V})_{\leq\ell} := \{v \in E(\mathrm{val}(X, \mathcal{V}_{\mathrm{SI}})) : v(x_i) = e \text{ for } 1 \leq i \leq \ell\}.$$

When the variety $\mathcal{V}$ is clear from the context or is not important we write E, E_ℓ and $E_{\leq\ell}$.

Note that $E_{\leq \ell} = E_1 \cap \cdots \cap E_\ell$ follows from the definition and that $E = E_1 \cup \cdots \cup E_n$ follows from the observation that $e \in v(X)$ for every $v \in E$.

For any subvariety $\mathcal{V}$ of $\mathcal{H}$. we know $\mathbf{F}_{\mathcal{V}}(X)$ is isomorphic to $\mathbf{Ge}(X, E(\mathcal{V}))$. Since the valuations in E_i serve to separate the elements of $\overline{\mathbf{T}}_i(n)$, we have $\overline{\mathbf{T}}_i(n)$ embedded in $\mathbf{Ge}(X, E_i)$. From Lemma 3.11(3) it follows that if $v \in E_i$ and $t \in \mathrm{T}_i(n)$, then $v(t) \geq v(x_i) \geq e$. This means that the embedding of $\overline{\mathbf{T}}_i(n)$ into $\mathbf{Ge}(X, E_i)$ is actually an embedding into $\mathbf{C}_2^{E_i}$. It can be argued that this embedding is onto $\mathbf{C}_2^{E_i}$. One way this can be done is to construct for every $v \in E_i$ a term t_v such that $v(t_v) = 1$ and $w(t_v) = e$ for every $w \in E_i$ with $w \neq v$. The construction makes use of the ordered set of elements $v(X)$ in the subdirectly irreducible algebra $\mathbf{Alg}(v)$ and is similar in spirit to that given in Lemma 3.6. That the valuations in E_i are pairwise nonequivalent is also used in this argument. Then by means of the term $(x \cdot y) \cdot y$, which behaves as a semilattice join on the ordered set $e < 1$, it is possible to form for every element c of $\mathbf{C}_2^{E_i}$ a term $t \in \mathrm{T}_i$ built from the appropriate t_v for which $w(t) = c(w)$ for all $w \in E_i$. Thus we have the following.

3.13 Lemma *For any subvariety $\mathcal{V}$ of $\mathcal{H}$ the algebra $\overline{\mathbf{T}}_{\leq \ell}(n)$ is isomorphic to $\mathbf{C}_2^{|E_{\leq \ell}|}$ where $\mathbf{C}_2 = \mathbf{H}(P)$ for P the 2-element chain $e < 1$.*

A stronger formulation of this lemma is given in [3] where the variety $\mathcal{H}$ is replaced by the variety of all Hilbert algebras.

Theorem 3.2 and Lemma 3.13 provide the following expression for the cardinality of a finitely generated free algebra $\mathbf{F}_{\mathcal{V}}(n)$ in a subvariety $\mathcal{V}$ of $\mathcal{H}$.

3.14 Theorem *For every variety $\mathcal{V} \subseteq \mathcal{H}$ and every positive integer n*

$$|\mathrm{F}_{\mathcal{V}}(n)| = \sum_{\ell=1}^{n} (-1)^{\ell-1} \binom{n}{\ell} 2^{|E_{\leq \ell}|}.$$

So, for a variety $\mathcal{V}$ in $\mathcal{H}$ the problems of describing $\overline{\mathbf{T}}_i(n)$ and of determining the cardinality of $\mathbf{F}_{\mathcal{V}}(n)$ reduce to counting the number of elements in E_i and $E_{\leq \ell}$.

For example, if $\mathcal{V}$ is the variety generated by $\mathbf{C}_3 = \mathbf{H}(P)$ for P the 3-element chain, then the only subdirectly irreducible algebras in $\mathcal{V}$ are $\mathbf{C}_2$ and $\mathbf{C}_3$. Among the valuations in $E_{\leq \ell}$ there are $2^{n-\ell}$ valuations v for which $\mathbf{Alg}(v) = \mathbf{C}_2$ and $3^{n-\ell} - 2^{n-\ell}$ with $\mathbf{Alg}(v) = \mathbf{C}_3$. So we have

$$\overline{\mathbf{T}}_{\leq \ell}(n) \cong (\mathbf{C}_2)^{3^n} \quad \text{and} \quad |\mathbf{F}_{\mathcal{V}}(n)| = \sum_{\ell=1}^{n} (-1)^{\ell-1} \binom{n}{\ell} 2^{3^n}.$$

This result may be found in [14].

We use Theorem 3.14 to determine the cardinality of the finitely generated free algebras for $\mathcal{V} = \mathcal{H}$. We first describe $E_{\leq \ell}$. Let $v \in E_{\leq \ell}$ with $\mathbf{A} = \mathbf{Alg}(v)$. The subdirectly irreducible algebra $\mathbf{A}$, which is generated by the elements e and the $v(x_j)$ for $\ell + 1 \leq j \leq n$, has at most $n - \ell$ elements strictly below e. Let $R = \{x_j \in X : v(x_j) < e\}$ and $T = \{x_j \in X : v(x_j) = 1\}$. We form a quasi-order σ_v consisting of the union of these sets:

- $R \times (X - R)$,
- a quasi-order ρ on R given by $\rho = \{(x_j, x_k) \in R^2 : v(x_j) \leq v(x_k)\}$,

- $X \times T$,
- the diagonal $\{(x_j, x_j) : x_j \in X\}$.

If $|R| = r$, then the number of ways the set R and the quasi-order ρ on R can be chosen is $\binom{n-\ell}{r} q_r$, where $0 \leq r \leq n - \ell$ and q_r is the number of quasi-orders on an r-element set. Having chosen r of the $v(x_j)$ to have value strictly less than e there are at most $2^{n-\ell-r}$ ways to choose the set T. Hence

$$|E_{\leq \ell}| \leq \sum_{r=0}^{n-\ell} \binom{n-\ell}{r} q_r 2^{n-\ell-r}.$$

This upper bound is actually obtained since for every set $R \subseteq \{x_{\ell+1}, \ldots, x_n\}$ of cardinality r and every quasi-order ρ defined on R, and for every choice of a set $T \subseteq \{x_{\ell+1}, \ldots, x_n\} - R$ with the quasi-order τ on X given by

$$\tau = \{(x_j, x_k) : x_j \in R,\ x_k \in X - R\} \cup \{(x_j, x_k) : x_j \in X,\ x_k \in T\} \cup \{(x_j, x_j) : x_j \in X\},$$

there is a subdirectly irreducible algebra $\mathbf{A} = \mathbf{H}(P)$ for P the ordered set $X/(\rho \cup \tau)$. For this $\mathbf{A}$, the element 1 is $T/(\rho \cup \tau)$, the element e is $(X - (R \cup T))/(\rho \cup \tau)$, and each $x_j/(\rho \cup \tau)$ for $x_j \in R$ is strictly below e. If $v : X \to \mathbf{A}$ is defined by $v(x_i) = x_i/(\rho \cup \tau)$, then $v(X)$ generates $\mathbf{A}$, $v \in E_{\leq \ell}$, and $\sigma_v = \rho \cup \tau$. Thus,

$$|E_{\leq \ell}| = \sum_{r=0}^{n-\ell} \binom{n-\ell}{r} q_r 2^{n-\ell-r}. \tag{3.1}$$

An application of Theorem 3.14 gives the following result from [3].

3.15 Theorem

$$|\mathrm{F}_{\mathcal{H}}(n)| = \sum_{\ell=1}^{n} (-1)^{\ell-1} \binom{n}{\ell} 2^{\sum_{r=0}^{n-\ell} \binom{n-\ell}{r} q_r 2^{n-\ell-r}}.$$

The values of $|\mathrm{F}_{\mathcal{H}}(n)|$ for $n = 1$, 2, and 3 are 2, 14, and 12266 respectively.

4 Structure via direct product decomposition

In this section we describe a wide class of varieties for which the finitely generated free algebras have a direct product decomposition into directly indecomposable algebras where the structure of these indecomposables can be described in terms of free algebras of associated varieties and where the multiplicity of each indecomposable in the product is linked in a strong way to free algebras in another associated variety.

4.1 Definition A finite nontrivial algebra $\mathbf{Q}$ is *quasiprimal* if every nontrival subalgebra of $\mathbf{Q}$ is simple and the variety generated by $\mathbf{Q}$ is congruence permutable and congruence distributive.

There is an extensive literature on quasiprimal algebras and the varieties that they generated. We mention [20] and [16] as important sources of information about quasiprimal algebras and the varieties they generate.

Let $\mathbf{Q}$ be a quasiprimal algebra and $\mathcal{Q}$ the variety that it generates. It is known that the variety $\mathcal{Q}$ is *semisimple*, that is, every subdirectly irreducible algebra in $\mathcal{Q}$ is a simple algebra. Moreover, every simple algebra in $\mathcal{Q}$ is a subalgebra of $\mathbf{Q}$. If $\mathbf{A}$ is any finite algebra in $\mathcal{Q}$, then from the congruence distributivity of $\mathcal{Q}$ it follows that the congruence lattice $\mathbf{Con}\,\mathbf{A}$ is a distributive lattice in which each element is a meet of coatoms, and therefore is a finite Boolean lattice. From the congruence permutability of $\mathcal{Q}$ we have that if α and β are complements in this lattice, then $\mathbf{A}$ is isomorphic to $\mathbf{A}/\alpha \times \mathbf{A}/\beta$. If $\alpha_1, \ldots, \alpha_r$ are the coatoms of $\mathbf{Con}\,\mathbf{A}$, then $\mathbf{A} \cong \mathbf{A}/\alpha_1 \times \cdots \times \mathbf{A}/\alpha_r$ and each $\mathbf{A}/\alpha_i$ is a subalgebra of $\mathbf{Q}$. So every finite algebra in $\mathcal{Q}$ is a direct product, in a unique way, of simple algebras. In particular for every positive integer n we may write

$$\mathbf{F}_{\mathcal{Q}}(n) = \prod_{j=1}^{r} \mathbf{B}_j, \tag{4.1}$$

where each $\mathbf{B}_j$ is a simple subalgebra of $\mathbf{Q}$ and the kernel of the projection pr_j onto $\mathbf{B}_j$ is a coatom, say α_j, of the congruence lattice of $\mathbf{F}_{\mathcal{Q}}(n)$.

We rephrase (4.1) in terms of $\mathbf{Ge}(X, R)$ for $X = \{x_1, \ldots, x_n\}$ and R a set of valuations to subalgebras of $\mathbf{Q}$. For $x_i \in X$ let $\overline{x}_i$ be the r-tuple $(\overline{x}_i(1), \ldots, \overline{x}_i(r))$ in (4.1). For each $1 \leq j \leq r$ let v_j be the function from X to B_j for which $v_j(x_i) = \overline{x}_i(j)$. Then v is a valuation to $\mathbf{B}_j$. If $R = \{v_1, \ldots, v_r\}$, then the array corresponding to $\mathbf{Ge}(X, R)$ is identical to $\prod_{j=1}^{r} \mathbf{B}_j$ in (4.1). On the other hand, let $v \in \mathbf{Q}^X$ be arbitrary. If $\mathbf{B}$ is the subalgebra of $\mathbf{Q}$ generated by $v(X)$, then v is a valuation to $\mathbf{B}$, the algebra $\mathbf{B}$ is simple, and the kernel of the homomorphic extension of v to all of $\mathbf{F}_{\mathcal{Q}}(n)$ is a coatom of $\mathbf{Con}(\mathbf{F}_{\mathcal{Q}}(n))$. So there is a $1 \leq j \leq r$ and an isomorphism $h : \mathbf{B} \to \mathbf{B}_j$ such that $h(v(x_i)) = \overline{x}_i(j) = v_j(x_i)$. Thus v and v_j are equivalent in the sense of Definition 1.5. In fact, if $v, v' \in Q^X$ are valuations that are equivalent in the sense of Definition 1.5, then the homomorphisms h and h' between $\mathbf{Alg}(v)$ and $\mathbf{Alg}(v')$ for which $v' = hv$ and $v = h'v'$ must be isomorphisms since every subalgebra of $\mathbf{Q}$ is simple. So the transversal of valuations of Definition 1.5 can be chosen to be the set $R = \{v_1, \ldots, v_r\}$.

Therefore, the structure of $\mathbf{F}_{\mathcal{Q}}(n)$ is determined by a transversal, say T, of the pairwise nonisomorphic subalgebras of $\mathbf{Q}$ and for each algebra $\mathbf{B}$ in T, the number, say $m(\mathbf{B})$, of factors in the representation in (4.1) that are isomorphic to $\mathbf{B}$. The value of $m(\mathbf{B})$ is completely determined by the subuniverses and the automorphisms of $\mathbf{B}$. Namely, consider the set of all $v \in \mathbf{B}^X$ for which $v(X)$ is not a subset of any proper subuniverse of $\mathbf{B}$, and for this set of valuations to $\mathbf{Q}$, choose a subset whose elements are pairwise inequivalent with respect to the equivalence relation of Definition 1.5. This pairwise inequivalence can be determined by only using automorphisms of $\mathbf{B}$ since $\mathbf{B}$ is a simple algebra. From T and the values of $m(\mathbf{B})$ as $\mathbf{B}$ ranges over T we have the direct product decomposition of $\mathbf{F}_{\mathcal{Q}}(n)$ given in (4.1). Note that in the case that the algebra $\mathbf{B}$ is rigid (i.e., no proper automorphisms), then $m(B)$ is just the number of $v \in \mathbf{B}^X$ for which $v(X)$ is not contained in any proper subuniverse of $\mathbf{B}$.

So by the analysis given in the previous paragraphs, the structure of finitely generated free algebras in a variety $\mathcal{Q}$ generated by a quasiprimal algebra $\mathbf{Q}$ can be described. Each is of the form $\mathbf{Ge}(X, R)$ for R a set of valuations that is free for $\mathcal{Q}$ and independent in the sense of Definition 1.2. The set R of valuations can, in principle, be determined from the

collection of subuniverses and automorphisms of $\mathbf{Q}$.

In the remainder of this section we obtain a structure theorem for free algebras that builds on the quasiprimal construction just given. The full theory is given in [2]. The presentation that we now give makes use of the $\mathbf{Ge}(X, R)$ construction. We first present a detailed illustrative example.

4.2 Example Stone algebras: The variety $\mathcal{V}$ of Stone algebras is the subvariety of the variety of all distributive pseudocomplemented lattices that satisfy the identity $x^* \vee x^{**} = 1$. The only nontrivial subvariety of Stone algebras is the variety $\mathcal{B}$ of Boolean algebras. The 3-element Stone algebra $\mathbf{S} = \langle\{0, 1, 2\}, \wedge, \vee, {}^*, 0, 1\rangle$ has the lattice structure of the 3-element chain $0 < 2 < 1$ with the pseudocomplement operation * given by $0^* = 1$, $1^* = 0$, and $2^* = 0$. It is known that the variety $\mathcal{V}$ of Stone algebras is generated by $\mathbf{S}$. Stone algebras have been thoroughly studied during the past 50 years. The structure of free Stone algebras was presented by R. Balbes and A. Horn in [1].

The only subdirectly irreducible Stone algebras are $\mathbf{S}$ and the 2-element Boolean algebra $\mathbf{B} = \langle\{0, 1\}, \wedge, \vee, {}^*, 0, 1\rangle$. Let h denote the homomorphism from $\mathbf{S}$ to $\mathbf{B}$ given by $h(0) = 0$ and $h(1) = h(2) = 1$. If α is the kernel of h, then α is a coatom of the congruence lattice of $\mathbf{S}$, the congruence lattice of $\mathbf{S}$ is $0 < \alpha < 1$ with α having $\{1, 2\}$ as its only nontrivial block.

The free algebra $\mathbf{F}_{\mathcal{V}}(n)$ is a subdirect product of copies of $\mathbf{S}$ and $\mathbf{B}$. Let U be the set of all valuations of X to $\mathbf{S}$ or $\mathbf{B}$. There are 2^n valuations to $\mathbf{B}$ and $3^n - 2^n$ valuations to $\mathbf{S}$ since every valuation to $\mathbf{S}$ must have 2 in its range. So $\mathbf{F}_{\mathcal{V}}(n)$ can be represented as the array $\mathbf{Ge}(X, U)$. We define an equivalence relation $\sim_{\mathbf{B}}$ on U by $v \sim_{\mathbf{B}} w$ if and only if $h(v(x_i)) = h(w(x_i))$ for all $x_i \in X$. Clearly, there are 2^n equivalence classes for $\sim_{\mathbf{B}}$. If $Z \subset U$ is an equivalence class of $\sim_{\mathbf{B}}$ we wish to describe the projection of $\mathbf{Ge}(X, U)$ on Z. That is, we wish to describe $\mathbf{Ge}(X, Z)$. We compute an illustrative example with the UAC program to see what is going on.

So let $n = 4$ and suppose Z consists of those valuations v for which $hv(x_1) = hv(x_2) = hv(x_3) = 1$ and $hv(x_4) = 0$. Then $v(x_4) = 0$ but $v(x_i) \in \{1, 2\}$ for $1 \leq i \leq 3$. So $|Z| = 8$. The algebra $\mathbf{Ge}(X, Z)$ is generated by $\overline{x}_1, \ldots, \overline{x}_4$. Table 4 contains the output of the UAC program's computation of the subalgebra of $\mathbf{S}^3$ generated by $\overline{x}_1$, $\overline{x}_2$, $\overline{x}_3$ and $\overline{x}_4$.

The generators $\overline{x}_1$, $\overline{x}_2$, $\overline{x}_3$ and $\overline{x}_4$ of $\mathbf{Ge}(X, Z)$ in Table 4 are labelled V0, V1, V2, and V3 respectively. The elements of each row of $\mathbf{Ge}(X, Z)$ are all in the same α class, that is, each row is constant with respect to the homomorphism h. Of the 20 rows in $\mathbf{Ge}(X, Z)$, 19 lie in $\{1, 2\}^8$ and one is in $\{0\}^8$. The operations $\wedge$ and $\vee$, when restricted to $\{1, 2\}$ behave as lattice operations on the distributive lattice $\mathbf{D}$ with $2 < 1$. The 8 valuations of $\{x_1, x_2, x_3\}$ to $\mathbf{D}$ are coded as the eight columns of $\mathbf{Ge}(X, Z)$. So the rows generated by V0, V1, and V2 by means of $\wedge$ and $\vee$ give the 18 elements of the free distributive lattice on 3 free generators. The operation * applied to $\{1, 2\}$ gives $\{0\}$, but the term ** sends $\{1, 2\}$ to $\{1\}$. So the row consisting of all 1's also appears. This accounts for the 19 rows whose elements are in the class $1/\alpha$. There can only be one row whose elements are in the singleton class $0/\alpha$. So all elements are accounted for. Therefore $\mathbf{Ge}(X, Z)$ is isomorphic to $1 \oplus \mathbf{F}_{\mathcal{D}_1}(3)$, where $\mathcal{D}_1$ is the variety of upper bounded distributive lattices.

More generally, for arbitrary n, if Z is an equivalence class of $\sim_{\mathbf{B}}$ and n_1 is the number of the elements $\overline{x}_i$ in $\mathbf{Ge}(X, Z)$ that have $v(x_i) \in 1/\alpha$ for all $v \in Z$, then as in the example the algebra $\mathbf{Ge}(X, U)$ is isomorphic to $1 \oplus \mathbf{F}_{\mathcal{D}_1}(n_1)$. So the structure of $\mathbf{Ge}(X, U)$ is quite clear. This algebra is also isomorphic to the free bounded distributive lattice on n_1 free generators,

```
. . < Number of vectors: 20
 . .   Length of vectors: 8
 . .   V    0: ( 1, 1, 1, 1, 2, 2, 2, 2)
 . .   V    1: ( 1, 1, 2, 2, 1, 1, 2, 2)
 . .   V    2: ( 1, 2, 1, 2, 1, 2, 1, 2)
 . .   V    3: ( 0, 0, 0, 0, 0, 0, 0, 0)
 . .   V    4: ( 1, 1, 2, 2, 2, 2, 2, 2)
 . .   V    5: ( 1, 2, 1, 2, 2, 2, 2, 2)
 . .   V    6: ( 1, 2, 2, 2, 1, 2, 2, 2)
 . .   V    7: ( 1, 1, 1, 1, 1, 1, 2, 2)
 . .   V    8: ( 1, 1, 1, 1, 1, 2, 1, 2)
 . .   V    9: ( 1, 1, 1, 2, 1, 1, 1, 2)
 . .   V   10: ( 1, 1, 1, 1, 1, 1, 1, 1)
 . .   V   11: ( 1, 2, 2, 2, 2, 2, 2, 2)
 . .   V   12: ( 1, 1, 1, 2, 2, 2, 2, 2)
 . .   V   13: ( 1, 1, 2, 2, 1, 2, 2, 2)
 . .   V   14: ( 1, 2, 1, 2, 1, 2, 2, 2)
 . .   V   15: ( 1, 1, 1, 1, 1, 2, 2, 2)
 . .   V   16: ( 1, 1, 1, 2, 1, 1, 2, 2)
 . .   V   17: ( 1, 1, 1, 2, 1, 2, 1, 2)
 . .   V   18: ( 1, 1, 1, 1, 1, 1, 1, 2)
 . .   V   19: ( 1, 1, 1, 2, 1, 2, 2, 2)
```

Table 4

but we write it as an ordered sum in order to emphasize the origin of the two components of the algebra as determined by those n_1 variables that are mapped to 1 and those n_2 that are mapped to 0. We shall see that this decomposition is an example of a more general one of the form $\mathbf{F}_{\mathcal{V}_0}(n_0) \cup \mathbf{F}_{\mathcal{V}_1}(n_1)$ for two varieties $\mathcal{V}_0$ and $\mathcal{V}_1$ with $n_0 + n_1 = n$.

Next we consider how the different $\mathbf{Ge}(X, Z)$ fit together. We view $\mathbf{F}_{\mathcal{S}}(n)$ as $\mathbf{Ge}(X, U)$ for U the set of all valuations to $\mathbf{S}$ and $\mathbf{B}$ and we view $\mathbf{F}_{\mathcal{B}}(n)$ as $\mathbf{Ge}(X, W)$ with $W = \mathbf{B}^X$. Let $Z_1, \ldots, Z_{2^n}$ be a list of all the classes of $\sim_{\mathbf{B}}$. Let g be the canonical onto homomorphism $g : \mathbf{F}_{\mathcal{S}}(n) \to \mathbf{F}_{\mathcal{B}}(n)$ that maps generators to generators. Since g is onto, for each $\sim_{\mathbf{B}}$ class Z there is a term $m_Z(x_1, \ldots, x_n)$ for which $v(g(m_Z)) = 1$ if $v \in Z$, and $v(g(m_Z)) = 0$ otherwise. Then for any $\overline{t} \in \mathbf{Ge}(X, Z)$ the term $p_{Z,t} = t \wedge m_Z$ is such that $v(p_{Z,t}) = v(t)$ if $v \in Z$ and $v(p_{Z,t}) = 0$ if $v \notin Z$. Let $(\overline{t}_1, \ldots, \overline{t}_{2^n}) \in \mathbf{Ge}(X, Z_1) \times \cdots \times \mathbf{Ge}(X, Z_{2^n})$ be arbitrary. Form the term

$$t(x_1, \ldots, x_n) = \bigvee_{1 \leq k \leq 2^n} p_{Z_k, t_k}.$$

For every $1 \leq k \leq n$ and every $v \in Z_k$ we have $v(t_k) = v(t)$. Therefore we conclude that $\mathbf{F}_{\mathcal{S}}(n) \cong \mathbf{Ge}(X, U) \cong \mathbf{Ge}(X, Z_1) \times \cdots \times \mathbf{Ge}(X, Z_{2^n})$.

By combining this with the characterization of $\mathbf{Ge}(X, Z)$ given in the previous paragraph

we get the result of [1]:

$$\mathbf{F}_{\mathcal{S}}(n) = \prod_{n_1=0}^{n} (1 \oplus \mathbf{F}_{\mathcal{D}_1}(n_1))^{\binom{n}{n_1}}.$$

The paper [2] grew out of an attempt to understand a number of similar examples in the literature in which the free algebra $\mathbf{F}_{\mathcal{V}}(n)$ for a locally finite variety $\mathcal{V}$ has the form

$$\mathbf{F}_{\mathcal{V}}(n) = \prod_{n_0+n_1=n} (\mathbf{F}_{\mathcal{V}_0}(n_0) \cup \mathbf{F}_{\mathcal{V}_1}(n_1))^{\binom{n}{n_1}},$$

or the more general form

$$\mathbf{F}_{\mathcal{V}}(n) = \prod_{n_1+\cdots+n_k=n} (\mathbf{F}_{\mathcal{V}_1}(n_1) \cup \cdots \cup \mathbf{F}_{\mathcal{V}_k}(n_k))^{\binom{n}{n_1,\ldots,n_k}},$$

where the $\mathcal{V}_i$ are varieties derived from $\mathcal{V}$ in some way. The general phenomenon observed is that $\mathcal{V}$ is a locally finite variety and

$$\mathbf{F}_{\mathcal{V}}(n) = \prod_i \mathbf{D}_i^{m_i},$$

with each $\mathbf{D}_i$ a directly indecomposable factor of $\mathbf{F}_{\mathcal{V}}(n)$; m_i is the multiplicity of $\mathbf{D}_i$ in the product, each $\mathbf{D}_i$ is built from some algebras in some varieties associated with $\mathcal{V}$ and the values of the m_i are determined by enumerating valuations of a particular form. The following is a description of some facts noticed in surveying these examples.

If $\mathcal{V}$ is one of the locally finite varieties examined, then the following facts were noted. The subvariety $\mathcal{V}_0$ generated by all finite simple algebras of $\mathcal{V}$ is generated by a quasiprimal algebra $\mathbf{Q}$. Typically $\mathbf{Q}$ is a familiar algebra such as the 2-element Boolean algebra $\mathbf{B}$. As we observed earlier, the free algebra $\mathbf{F}_{\mathcal{V}_0}(n)$ is of the form $\prod_{j=1}^{r} \mathbf{Q}_j$ with each $\mathbf{Q}_j$ a nontrivial subalgebra of $\mathbf{Q}$, and r the cardinality of the set R of valuations for which $\mathbf{F}_{\mathcal{V}_0}(n)$ is isomorphic to $\mathbf{Ge}(X, R)$. As is the case for any variety generated by a quasiprimal algebra, the congruence lattice of $\mathbf{F}_{\mathcal{V}_0}(n)$ is a Boolean lattice with r coatoms. If $\mathbf{F}_{\mathcal{V}}(n)$ is written as product of directly indecomposable algebras then, in the examples considered, the number of factors in this decomposition is also r, and $\mathbf{F}_{\mathcal{V}}(n) = \prod_{j=1}^{r} \mathbf{D}_j$ with each $\mathbf{D}_j$ directly indecomposable, the congruence lattice $\mathbf{Con}(\mathbf{D}_j)$ has exactly one coatom, say α_j, and $\mathbf{D}_j/\alpha_i$ is isomorphic to $\mathbf{Q}_j$. Moreover, in these examples, $\mathbf{Con}(\mathbf{F}_{\mathcal{V}}(n)) \cong \prod_{j=1}^{r} \mathbf{Con}(\mathbf{D}_j)$. Let $\mathcal{V}$ have similarity type τ. For each $q \in Q$ let τ_q denote the similarity type consisting of all terms t of type τ that satisfy $t^{\mathbf{Q}}(q, \ldots, q) = q$. We identify $\mathbf{F}_{\mathcal{V}}(n)$ with $\prod_{j=1}^{r} \mathbf{D}_j$ and let $pr_j : \mathbf{F}_{\mathcal{V}}(n) \to \mathbf{D}_j$ be the projection map and let $h_j : \mathbf{D}_j \to \mathbf{Q}_j$ be a homomorphism with kernel α_j. For $1 \leq j \leq r$ and $q \in Q$ let $X_j^q = \{x_i \in X : h_j(pr_j(\overline{x}_i)) = q\}$. Let $\mathcal{V}_q$ denote the variety of similarity type τ_q that is generated by all algebras $\mathbf{A}_j^q$ whose universe is $h_j^{-1}(q) \subseteq D_j$ and whose operations are those term operations t of $\mathbf{D}_j$ for which $t^{\mathbf{Q}}(q, \ldots, q) = q$. Then each $\mathbf{D}_j$ is built from subsets having the structure of $\mathbf{F}_{\mathcal{V}_q}(X_j^q)$ where q ranges over the elements of Q_j. In the more familiar examples we have $D_j = \bigcup_{q \in Q_j} \mathbf{F}_{\mathcal{V}_q}(X_j^q)$.

For example, let us reconsider Example 4.2 of Stone algebras. Here the only simple algebra is the 2-element Boolean algebra $\mathbf{B}$ so the variety $\mathcal{S}_0$ is the variety $\mathcal{B}$ of Boolean algebras. The free algebra $\mathbf{F}_{\mathcal{B}}(n)$ is $\mathbf{B}^{2^n}$ since every valuation to $\mathbf{B}$ is needed to separate the

elements of $\mathbf{F}_{\mathcal{B}}(n)$. So $r = 2^n$ and $\mathbf{Con}(\mathbf{F}_{\mathcal{B}}(n))$ is a Boolean algebra with r coatoms. In the decomposition of $\mathbf{F}_{\mathcal{S}}(n)$ into directly indecomposables, $\prod_{j=1}^{r} \mathbf{D}_j$, the unique coatom of the congruence lattice of $\mathbf{D}_j$ is the kernel of the homomorphism h_j onto $\mathbf{B}$ in which $h_j^{-1}(0) = \{0\}$ and $h_j^{-1}(1) = D_j - \{0\}$. The variety $\mathcal{S}$ is congruence distributive, so it follows from known results that for every $\mathbf{A} \in \mathcal{S}$, if $\mathbf{A} = \mathbf{A}_1 \times \mathbf{A}_2$, then $\mathbf{Con}\,\mathbf{A} \cong \mathbf{Con}\,\mathbf{A}_1 \times \mathbf{Con}\,\mathbf{A}_2$. The similarity types of τ_0 and τ_1 both include $\wedge, \vee, ^{**}$; while τ_0 includes the constant symbol 0 and τ_1 contains the constant symbol 1. The algebras $\mathbf{A}_j^0$ for $1 \leq j \leq r$ are all 1-element algebras and so the variety $\mathcal{S}_q$ for $q = 0$ is the trivial variety $\mathcal{T}$. For $q = 1$ the algebras $\mathbf{A}_j^1$ are term equivalent to distributive lattices with a top element 1 and the variety $\mathcal{S}_q$ for $q = 1$ is $\mathcal{D}_1$. For every j, we have $X_j^k = \{x_i : h_j(pr_j(\overline{x}_i)) = k\}$ for $k = 0, 1$. Then $\mathbf{D}_j$ has as its universe $\mathrm{F}_{\mathcal{T}}(X_j^0) \cup \mathrm{F}_{\mathcal{D}_1}(X_j^1)$.

In [2] general conditions on a variety $\mathcal{V}$ are presented and are proved to be sufficient to force the well-behaved direct product decomposition of $\mathbf{F}_{\mathcal{V}}(n)$ into indecomposables described in the previous paragraphs. We now sketch these results.

A variety $\mathcal{V}$ is said to have the *Fraser-Horn Property* (FHP) if there are no skew congruences on finite products, i.e., for all $\mathbf{A}_1, \mathbf{A}_2 \in \mathcal{V}$, every $\theta \in \mathbf{Con}(\mathbf{A}_1 \times \mathbf{A}_2)$ is a product congruence $\theta_1 \times \theta_2$ with $\theta_k \in \mathbf{Con}(\mathbf{A}_k)$ for $k = 1, 2$. If $\mathcal{V}$ has FHP, then $\mathbf{Con}(\mathbf{A}_1 \times \mathbf{A}_2) \cong \mathbf{Con}\,\mathbf{A}_1 \times \mathbf{Con}\,\mathbf{A}_2$ for all $\mathbf{A}_1$ and $\mathbf{A}_2$ in $\mathcal{V}$. Various conditions that imply FHP are known. If $\mathcal{V}$ is congruence distributive, then $\mathcal{V}$ has FHP. Fraser and Horn [10] give a Mal'cev condition equivalent to FHP and they derive as a special case the following:

> *A variety $\mathcal{V}$ has FHP if there are binary terms $+$ and $\cdot$ and elements* 0 *and* 1 *in* $\mathrm{F}_{\mathcal{V}}(3)$ *such that for all $z \in \mathrm{F}_{\mathcal{V}}(3)$, we have $z \cdot 1 = z + 0 = 0 + z = z$ and $z \cdot 0 = 0$.*

So, for example every variety of rings with unit has FHP.

A finite algebra $\mathbf{A}$ has the *Apple Property* (AP) if for all $\beta \in \mathrm{Con}\,\mathbf{A}$, if $\beta < 1_A$ is a factor congruence with $\mathbf{A}/\beta$ directly indecomposable, then the interval lattice $[\beta, 1_A]$ in $\mathbf{Con}\,\mathbf{A}$ has exactly one coatom. A variety $\mathcal{V}$ has AP if every finite algebra in $\mathbf{A}$ has AP. Equivalently, $\mathcal{V}$ has AP if every directly indecomposable finite algebra in $\mathcal{V}$ has a congruence lattice with exactly one coatom. For example, a local ring has AP.

If an algebra or a variety has both FHP and AP, then we say it has FHAP. In [6] locally finite varieties with FHAP are shown to have finitely generated free algebras with the rich direct product structure described earlier in this section.

For any locally finite variety $\mathcal{V}$ let $\mathcal{V}_0$ be the subvariety generated by all finite simple algebras in $\mathcal{V}$. We call $\mathcal{V}_0$ the *prime variety* of $\mathcal{V}$. For arbitrary $\mathcal{V}$ little can be said about $\mathcal{V}_0$. However, in the event that a locally finite variety has FHAP and satisfies one additional condition, then the prime variety is especially well-behaved.

4.3 Theorem *Let $\mathcal{V}$ be a locally finite variety with FHAP. Suppose every subalgebra of a finite simple algebra in $\mathcal{V}$ is a product of simple algebras. Then all of the following hold:*

(1) *$\mathcal{V}_0$ is congruence distributive.*

(2) *$\mathcal{V}_0$ is congruence permutable.*

(3) *Every finite member of $\mathcal{V}_0$ is a product of simple algebras.*

(4) *For every n the free algebras $\mathbf{F}_{\mathcal{V}}(n)$ and $\mathbf{F}_{\mathcal{V}_0}(n)$ have the same number of directly indecomposable factors, and if r denotes this number, then it is possible to write*

$$\mathbf{F}_{\mathcal{V}}(n) = \prod_{j=1}^{r} \mathbf{D}_j \quad \text{and} \quad \mathbf{F}_{\mathcal{V}_0}(n) = \prod_{j=1}^{r} \mathbf{Q}_j$$

with the $\mathbf{D}_j$ and $\mathbf{Q}_j$ directly indecomposable, so that if α_j is the unique coatom of $\boldsymbol{Con}(\mathbf{D}_j)$, then $\mathbf{D}_j/\alpha_j$ and $\mathbf{Q}_j$ are isomorphic.

In the examples described in the earlier part of this section, the varieties satisfy the hypotheses of this theorem and the prime variety in each case is, in fact, a variety generated by a quasiprimal algebra.

4.4 Theorem *Let $\mathcal{V}$ be a locally finite variety with FHAP. Suppose every subalgebra of a finite simple algebra in $\mathcal{V}$ is a product of simple algebras. Let $r, \mathbf{D}_j, \mathbf{Q}_j$ be as in Theorem* 4.3 *with $h_j : \mathbf{D}_j \to \mathbf{Q}_j$ a homomorphism with kernel α_j. For $\mathbf{Q}$ a finite simple algebra in $\mathcal{V}$ and $q \in Q$, let τ_q be the similarity type of all terms t for which $t^{\mathbf{Q}}(q, \dots, q) = q$. For each j with $\mathbf{Q}_j \cong \mathbf{Q}$, let $\mathbf{A}_j^q$ be the algebra with universe $h_j^{-1}(q) \subseteq D_j$ and operations those t in τ_q, and let $\mathcal{V}_q$ be the variety generated by all $\mathbf{A}_j^q$. If $X_j^q = \{x_i \in X : h_j(pr_j(\overline{x}_i)) = q\}$, then the algebra $\mathbf{A}_j^q$ contains a set that is the universe of an algebra isomorphic to $\mathbf{F}_{\mathcal{V}_q}(X_j^q)$.*

The hypotheses of Theorems 4.3 and 4.4 are robust with many different varieties satisfying them. The examples alluded to earlier all satisfy them and thus Theorem 4.4 provides a uniform explanation of the description of the directly indecomposable direct factors of the finitely generated free algebras in those varieties.

In Theorem 4.4 we have that the universe of $\mathbf{D}_j$ is the disjoint union of the universes of the $\mathbf{A}_j^q$ as q ranges over the elements of $\mathbf{Q}_j$ and that $\mathbf{A}_j^q$ contains a subalgebra isomorphic to $\mathbf{F}_{\mathcal{V}_q}(X_j^q)$. For varieties such as Stone algebras, $\mathbf{A}_j^q = \mathbf{F}_{\mathcal{V}_q}(X_j^q)$ and so $D_j = \bigcup_{q \in Q_j} \mathrm{F}_{\mathcal{V}_q}(X_j^q)$. This situation represents a lower bound for the size of the $\mathbf{D}_j$. Moreover, it is possible to provide sufficient conditions on the behavior of the fundamental operations of $\mathcal{V}$ and $\mathcal{V}_q$ that guarantee that all the directly indecomposable $\mathbf{D}_j$ have this minimal structure.

On the other hand, it is possible for the size of $\mathbf{D}_j$ to exceed this lower bound. In [2] a condition is given that if added to the hypotheses of Theorem 4.4 forces $\mathbf{A}_j^q \cong \mathbf{F}_{\mathcal{V}_q}(X)$ for every $q \in Q_j$. Thus, in this case D_j is the disjoint union of $|Q_j|$ sets, each corresponding to the universe of $\mathbf{F}_{\mathcal{V}_q}(X)$. Roughly speaking, this condition given in [2] is that for every $q, q' \in Q_j$ there is a unary term $u_{qq'}$ that maps A_j^q onto $A_j^{q'}$. Note that if this condition holds, then $|D_j| = |Q_j| \cdot |\mathrm{F}_{\mathcal{V}_0}(n)|$. An example of a variety in which this occurs is the variety $\mathcal{V}$ of rings generated by $\mathbf{Z}_{p^2}$, for p any prime. As observed earlier, a variety of rings with unit has FHP. The variety $\mathcal{V}$ is known to have AP and thus $\mathcal{V}$ has FHAP. The only simple ring in $\mathcal{V}$ is $\mathbf{Z}_p$ and so the prime variety $\mathcal{V}_0$ is the variety generated by $\mathbf{Z}_p$. So $\mathbf{F}_{\mathcal{V}_0}(n) \cong (\mathbf{Z}_p)^{p^n}$. Theorem 4.4 applies. The number of factors in the decomposition of $\mathbf{F}_{\mathcal{V}}(n)$ into directly indecomposables is the same as that for $\mathbf{F}_{\mathcal{V}_0}(n)$, which is p^n. For each $q \in Z_p$ it can be shown that the variety $\mathcal{V}_q$ is the variety generated by the p-element group in which all p elements are constants. The free algebra on n free generators for this variety is easily seen to have p^{n+1} elements. For each $q, q' \in Z_p$ the unary term $u(x) = (x + q' - q) \bmod (p)$ can serve as $u_{qq'}$. So for each directly indecomposable factor $\mathbf{D}_j$ of $\mathbf{F}_{\mathcal{V}}(n)$ we have $|D_j| = p|\mathrm{F}_{\mathcal{V}_0}(n)| = p^{n+2}$. All the $\mathbf{D}_j$ are isomorphic and thus $|\mathbf{F}_{\mathcal{V}}(n)| = (p^{n+2})^{p^n}$.

References

[1] R. Balbes and A. Horn, Stone lattices, *Duke Math J.* **37** (1970), 537–546.

[2] J. Berman and W. J. Blok, The Fraser-Horn and Apple properties, *Trans. Amer. Math. Soc.* **302** (1987), 427–465.

[3] J. Berman and W. J. Blok, Free Łukasiewicz and hoop residuation algebras, *Studia Logica* **77** (2004), 153–180.

[4] J. Berman and W. J. Blok, *Algebras defined from ordered sets and the varieties they generate*, manuscript, 2003.

[5] J. Berman and P. Idziak, Counting finite algebras in the Post varieties, *Internat. J. Algebra Comput.* **10** (2000), 323–337.

[6] J. Berman and B. Wolk, Free lattices in some small varieties, *Algebra Universalis* **10** (1980), 269–289.

[7] G. Birkhoff, On the structure of abstract algebras, *Proc. Cambridge Philos. Soc.* **31** (1935), 433-454.

[8] S. Burris and H. P. Sanappanavar, *A Course in Universal Algebra*, Springer-Verlag, New York, 1981.

[9] A. Diego, *Sur les algèbres d'Hilbert*, Collection de Logique Mathématique, Sér. A, Fasc. XXI Gauthier-Villars, Paris; E. Nauwelaerts, Louvain, 1966.

[10] G. A. Fraser and A. Horn, Congruence relations on direct products, *Proc. Amer. Math. Soc.* **26** (1970), 390–394.

[11] `http://www.cs.elte.hu/~ewkiss/software/uaprog/`

[12] G. Grätzer and A. Kisielewicz, A survey of some open problems on p_n-sequences and free spectra of algebras and varieties, in: *Universal Algebra and Quasigroup Theory* (A. Romanowska and J.D.H. Smith, eds.), Heldermann Verlag, Berlin, 1992, 57–88.

[13] J. A. Green and D. Rees, On semigroups in which $x^r = x$, *Proc. Cambridge Phil. Soc.* **58** (1952), 35–40.

[14] F. Guzmán and C. Lynch, Varieties of positive implicative BCK-algebras—subdirectly irreducible and free algebras, *Math. Japon.* **37** (1992), 27–32.

[15] D. Hobby and R. McKenzie, *The Structure of Finite Algebras*, Amer. Math. Soc., Providence, 1988.

[16] K. Kaarli and A. F. Pixley, *Polynomial Completeness in Algebraic Systems* Chapman & Hall/CRC, Boca Raton, 2001.

[17] H. Lakser, R. Padmanabhan, and C. Platt, Subdirect decomposition of Płonka sums, *Duke Math. J.* **39** (1972), 485–488.

[18] R. N. McKenzie, G. F. McNulty, and W. F. Taylor, *Algebras, Lattices, Varieties*, Vol. 1, Wadsworth & Brooks/Cole, Monterey, 1987.

[19] J. Płonka, On free algebras and algebraic decompositions of algebras from some equational classes defined by regular equations, *Algebra Universalis* **1** (1971–1972), 261–264.

[20] R. W. Quackenbush, Structure theory for equational classes generated by quasi-primal algebras, *Trans. Amer. Math. Soc.* **187** (1974), 127–145.

[21] Á. Szendrei, *Clones in Universal Algebra*, Séminaire de Mathématiques Supérieures **99**, Université de Montréal, 1986.

[22] R. Wille, Subdirecte Produkte vollständiger Verbände, *J. Reine Angew. Math.* **283–284** (1976), 53–70.

Completeness of automaton mappings with respect to equivalence relations

Jürgen DASSOW

Fakultät für Informatik
Otto-von-Guericke-Universität
PSF 4120, D–39016 Magdeburg
Germany

Dedicated to Gustav Burosch on the occasion of his 65th birthday.

Abstract

In the algebra of sequential functions which map words over $(\{0,1\})^n$ to words over $\{0,1\}$, we call a set M complete with respect to an equivalence relation if the subalgebra generated by M contains at least one element of any equivalence class. The paper summarizes the results with respect to some known completeness notions, such as the classical completeness, τ-completeness and Kleene-completeness which can be reformulated as completenesses with respect to some equivalence relations, and it presents some new results on completeness with respect to further equivalence relations. Moreover, we discuss the metric completeness which is nearly related to completeness with respect to an equivalence relation and can also be formulated in terms of equivalences.

1 Introduction

In the structural theory of automata the following two problems belong to the most important questions.

- Given an automaton $\mathcal{A}$ or an equivalent device, a measure of complexity and some methods of decomposition, decide whether or not the automaton can be decomposed into simpler automata.
- Given an automaton $\mathcal{A}$ and a set M of automata and some operations to construct new automata from given ones, decide whether or not $\mathcal{A}$ can be built from the elements of M by means of the operations.

In this paper we discuss a special variant of the second problem. This variant is known as the completeness problem:

> Given two sets F and M of automata or equivalent devices or other descriptions of certain input-output-behaviours and some operations, decide whether or not all elements of F can be generated from the elements of M by the operations.

V. B. Kudryavtsev and I. G. Rosenberg (eds.),
Structural Theory of Automata, Semigroups, and Universal Algebra, 77–108.

The most important case occurs when F consists of all automata which are of interest in this context. We shall restrict ourselves to automata which transform input words over a cartesian product of $\{0,1\}$ into words over $\{0,1\}$. The reason for this is that other cases can be described by a codification and that such automata correspond in a very natural way to logical nets which are of some importance nowadays, too. The input-output-behaviour of such automata can be described by sequential functions, and therefore we shall consider an algebra of sequential functions. The operations are formalizations of the constructions used for logical nets.

It is well-known that the completeness of this algebra is undecidable, i.e., there is no algorithm which decides the completeness of a given finite set. Moreover, the lattice of subalgebras has a very complicated structure, e.g. the number of maximal subalgebras—which can be used as a completeness criterion—is at least infinite. Thus one must consider special cases to obtain better results with respect to the decidability of completeness. One approach consists of restricting the set of automata which can occur in F and M (e.g. the restriction to t-stable automata which obtain the same state for all inputs of length $\geq t$ or the requirement that M contains all automata which realize a Boolean function lead to decidable cases, see [25, 23]). Another approach only requires that "almost all" automata/behaviours/sequential functions can be generated. Typical examples are t-completeness and Kleene-completeness where one only requires that, for any automaton, we can generate from M an automaton with the same input-output-behaviour at the first t moments and that, for any regular language R, there can be obtained at least one automaton which accepts R, respectively. Both concepts have practical importance since mostly we can only observe the input-output-behaviour at a limited number of moments, and automata are often used as acceptors.

The latter approaches can be formulated as follows: For any automaton $\mathcal{A}$ we require that we can generate from M one automaton which can be considered as a representative of $\mathcal{A}$. From the algebraic point of view this leads to equivalence relations and the requirement to generate at least one element of any equivalence class. The corresponding completeness concept is the central topic of this paper. The notions of "classical" completeness, t-completeness and Kleene-completeness can be reformulated as completeness with respect to some equivalence relation. Thus the paper contains a summary of important results in those completeness concepts studied in the seventies and eighties. Moreover, we present some further equivalence relations which are closely related to those associated with the known concepts and the related results on completeness.

Hitherto equivalence has been defined on the target set of sequential functions. It is natural to extend it to the power set. Given an equivalence relation on the set of sequential functions, we say that two sets M and M' are equivalent if, for any function $f \in M$, M' contains a function f' equivalent to f and vice versa. Based on such an equivalence a further notion of completeness is studied, the so-called metric equivalence, where—intuitively—it is required that any function can be approximated up to any accuracy.

In this paper we are mostly interested in the decidability of and in criteria for completeness with respect to some equivalence relations.

We note that our completeness with respect to equivalence relations differs from the notion studied by Blochina, Kudrjavcev and Burosch in [4] where an equivalence of sets has been introduced and one is looking for completeness criteria which hold for all sets which are equivalent.

2 The algebra of sequential functions and its completeness concepts

First we recall some notions and the associated notation used in this paper.

By $\mathbb{N}$ we denote the set of positive integers.

An alphabet is a finite non-empty set. A word p over an alphabet X is a finite sequence of elements of X written as $p = x_1x_2 \dots x_n$ with $x_i \in X$ for $1 \leq i \leq n$. λ denotes the empty word. By X^* we designate the set of all words over X including the empty word. Moreover, we set $X^+ = X^* \setminus \{\lambda\}$, i.e., X^+ is the set of all non-empty words over X. The length of a word $p \in X^*$ is the number of elements of X occurring in p and is denoted by $|p|$. The concatenation pq of two words $p \in X^*$ and $q \in X^*$ is written as the juxtaposition of p and q. For two sets $U \subseteq X^*$ and $V \subseteq X^*$, we define their product by

$$UV = \{pq \mid p \in U,\ q \in V\}.$$

The powers of a set L are defined by

$$L^1 = L \quad \text{and} \quad L^{i+1} = L \cdot L^i \quad \text{for} \quad i \geq 1.$$

We then define the Kleene-closure of L by

$$L^+ = \bigcup_{i \geq 1} L^i.$$

Let X be an arbitrary set. By id_X we denote the function mapping X to X and defined by $id_X(x) = x$ for $x \in X$. If X is obvious from the context, we only write id.

A function f which maps a cartesian power of $\{0,1\}$ into $\{0,1\}$ is called a Boolean function. By non, et, vel and eq we denote the functions which correspond to the logical negation, conjunction, disjunction and equivalence, respectively. The constant functions which map to 0 and 1 are designated by k_0 and k_1, respectively. Further, we define the Sheffer function sh by $\mathrm{sh}(x_1, x_2) = \mathrm{non}(\mathrm{et}(x_1, x_2))$.

2.1 The algebra of sequential functions

2.1 Definition A function $F : X^+ \to Y^+$ is called *sequential*, if the following conditions are satisfied:

- $|F(p)| = |p|$ for any $p \in X^+$,
- for any word $p \in X^+$, there is a function F_p such that $F(pq) = F(p)F_p(q)$ for any $q \in X^+$,
- the set $\{F_p \mid p \in X^+\}$ is finite.

Obviously, if $F : X^+ \to Y^+$ is a sequential function and $p = x_1x_2 \dots x_n$ with $x_i \in X$ for $1 \leq i \leq n$, then $F(x_1x_2 \dots x_n) = y_1y_2 \dots y_n$ for some $y_i \in Y$, $1 \leq i \leq n$. x_i and y_i, $1 \leq i \leq n$, are called the input and the output at the ith moment, respectively.

2.2 Example We consider the function $F : \{0,1\}^+ \to \{0,1\}^+$ defined by

$$F(x_1x_2\dots x_n) = y_1y_2\dots y_n \quad \text{with} \quad y_i = \begin{cases} 1 & \text{if } i \geq 2 \text{ and } x_{i-1} = x_i = 1, \\ 0 & \text{otherwise.} \end{cases}$$

F has the output 1 at the moment i if and only if the inputs at the moments $i-1$ and i are 1, i.e., if at two moments in succession the inputs are 1, the output is 1, too.

Further, let $F' : \{0,1\}^+ \to \{0,1\}^+$ be the function defined by

$$F'(x_1x_2\dots x_n) = y_1y_2\dots y_n \quad \text{with} \quad y_i = \begin{cases} 1 & \text{if } i = 1 \text{ and } x_1 = 1, \\ 1 & \text{if } i \geq 2 \text{ and } x_{i-1} = x_i = 1, \\ 0 & \text{otherwise.} \end{cases}$$

F' differs from F only in words of length 1, where F gives the output 0 for all inputs whereas F' gives the input as output.

Now assume that p ends with 0, i.e., $p \in \{0,1\}^*\{0\}$, and $q = x_1x_2\dots x_m$ is a word of length m, then

$$F(pq) = F(p)z_1z_2\dots z_m \quad \text{with} \quad z_i = \begin{cases} 1 & \text{if } i \geq 2 \text{ and } x_{i-1} = x_i = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Thus we have

$$F(pq) = F(p)F(q) \quad \text{for} \quad p \in \{0,1\}^*\{0\}.$$

Analogously, we can see that

$$F(pq) = F(p)F'(q) \quad \text{for} \quad p \in \{0,1\}^*\{1\}.$$

This proves that F is a sequential function with

$$F_p = \begin{cases} F & \text{if } p \in \{0,1\}^*\{0\}, \\ F' & \text{if } p \in \{0,1\}^*\{1\}. \end{cases}$$

Let F be a sequential function. Then we have

$$F(pqr) = F(p)F_p(qr) \quad \text{and} \quad F(pqr) = F(pq)F_{pq}(r) = F(p)F_p(q)F_{pq}(r)$$

which proves $F_p(qr) = F_p(q)F_{pq}(r)$. Therefore F_p is a sequential function, too.

We now present some devices generating sequential function.

2.3 Definition A (deterministic finite) *automaton* is a quintuple $\mathcal{A} = (X, Y, Z, z_0, \delta, \gamma)$ where

- X, Y and Z are finite non-empty sets of input and output symbols and states, respectively,
- $z_0 \in Z$ is the initial state,
- $\delta : Z \times X \to Z$ and $\gamma : Z \times X \to Y$ are each a function called the transition and output function, respectively.

We extend inductively the functions δ and γ to $Z \times X^+$ by setting

$$\delta^*(z,a) = \delta(z,a) \quad \text{and} \quad \delta^*(z,pa) = \delta(\delta^*(z,p),a)),$$
$$\gamma^*(z,a) = \gamma(z,a) \quad \text{and} \quad \gamma^*(z,pa) = \gamma^*(z,p)\gamma(\delta^*(z,p),a))$$

for $p \in X^+$ and $a \in X$. The function $F_{\mathcal{A}} : X^+ \to Y^+$ defined by $F_{\mathcal{A}}(p) = \gamma^*(z_0,p)$ for $p \in X^+$ is called the function induced by $\mathcal{A}$.

If we define the functions

$$F_p(q) = \gamma^*(\delta^*(z_0,p),q) \quad \text{for} \quad p \in X^+,$$

then we get

$$F_{\mathcal{A}}(pq) = \gamma^*(z_0,pq) = \gamma^*(z_0,p)\gamma^*(\delta^*(z_0,p),q) = F_{\mathcal{A}}(p)F_p(q).$$

Moreover, there are only finitely many different functions F_p since there are only finitely many different states $\delta^*(z_0,p)$. Thus the function induced by an automaton is a sequential function.

On the other hand, given a sequential function $F : X^+ \to Y^+$, we can construct the finite automaton

$$\mathcal{A}_F = (X, Y, \{F\} \cup \{F_p \mid p \in X^+\}, F, \delta, \gamma)$$

with

$$\delta(G,a) = G_a \quad \text{and} \quad \gamma(G,a) = G(a).$$

It is easy to prove that $F(q) = \gamma^*(F,q)$.

Hence the sets of sequential functions and of functions induced by automata coincide.

2.4 Example We illustrate the above construction of an automaton for a given sequential function starting with the function F of Example 2.2. With the notation of Example 2.2 we obtain the automaton

$$\mathcal{A} = (\{0,1\}, \{0,1\}, \{F, F'\}, F, \delta, \gamma)$$

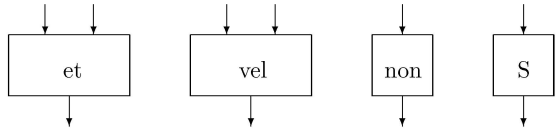
with the functions δ and γ described by the following tables (the rows correspond to states and the columns to inputs, and the value is given in the corresponding intersection):

δ	0	1
F	F	F'
F'	F	F'

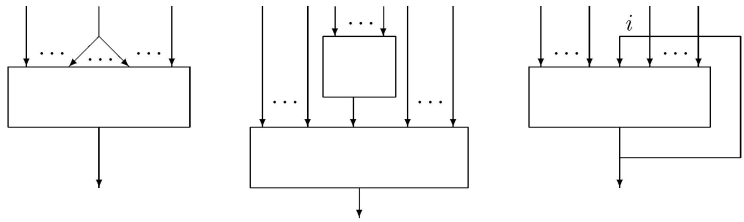
γ	0	1
F	0	0
F'	0	1

Logical nets are our second concept to generate sequential functions. We only present an informal definition of this device and the associated function.

The basic logical nets are:



and from these elements further logical nets can be constructed using iteratively the following operations:



where we require that any path from the "input" i to the output contains at least one basic logical net of type S.

Obviously, any logical net has a certain number n of inputs and one output. Now we assume that at a certain time moment t the i-th input gets the value $x_i(t) \in \{0,1\}$. Then we associate with the basic logical nets the following functions:

- the output of et at the t-th moment is $\mathrm{et}(x_1(t), x_2(t))$;
- the output of vel at the t-th moment is $\mathrm{vel}(x_1(t), x_2(t))$;
- the output of non at the t-th moment is $\mathrm{non}(x_1(t))$;
- the output of S at the first moment is 1, and
  the output of S at the t-th moment is $x_1(t-1)$ for $t \geq 2$.

Thus the elements et, vel, and non realize at any moment the corresponding Boolean functions, and S stores the input for one time moment.

Now we extend the associated function in the natural way to arbitrary logical nets.

Thus with any logical net N with n inputs we can associate a function

$$F_N : (\{0,1\}^n)^+ \to \{0,1\}^+.$$

Obviously, the output at the t-th moment depends on the inputs at the t-th moment and the values stored at that moment in the elements of type S. If we interpret the tuple of the values stored in the elements of type S as a state, then the output depends on a state and the tuple of inputs. An analogous situation holds for the values stored at a certain moment. Therefore the output and the state at the next moment can be described as transition and output functions of a finite automaton. Conversely, if we code the set of inputs and states by tuples of zeros and ones then the transition and output functions can be interpreted as Boolean functions which can be realized by logical nets (for details see [11, 18]).

2.5 Example We illustrate the idea by converting the automaton constructed in Example 2.4 into a logical net. Since $\mathcal{A}_F$ has two input symbols and two states, both sets can be coded by $\{0,1\}$. As the initial state we have to take 1 because the elements of type S have the output 1 at the first moment. Thus F corresponds to 1 and F' to 0 which results in the Boolean functions

δ	0	1
1	1	0
0	1	0

γ	0	1
1	0	0
0	0	1

or, equivalently, by taking the input as the first variable and the state as the second variable,

$$\delta(x_1, x_2) = \mathrm{non}(x_1) \quad \text{and} \quad \gamma(x_1, x_2) = \mathrm{et}(x_1, \mathrm{non}(x_2)).$$

This leads to the net given in Figure 1. The lower two basic elements realize the output function γ whereas the upper element non realizes δ (where we have omitted the feedback since δ does not depend on the second variable). Obviously, we can omit both non-elements because the double negation does not change the value, but then we will not be following the idea for the construction of a net from an automaton.

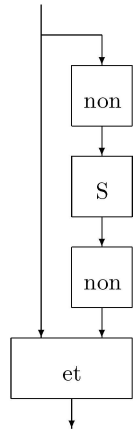


Figure 1: Logical net associated with the automaton $\mathcal{A}_F$ of Example 2.4

Combining these constructions we get the following statement.

2.6 Theorem *For a function $F : (\{0,1\}^n)^+ \to \{0,1\}^+$, the following statements are equivalent:*

(1) *F is a sequential function;*

(2) *F is a induced by some automaton;*

(3) *F is associated with some logical net.*

For a sequential function $F : (\{0,1\}^m)^* \to \{0,1\}^*$, we call m the arity of F and denote it by $\mathrm{arity}(F)$.

We designate the set of all sequential functions $F : (\{0,1\}^m)^* \to \{0,1\}^*$ (of arity m) by $\mathcal{F}^m$ and set

$$\mathcal{F} = \bigcup_{m \geq 1} \mathcal{F}^m.$$

Let F be a function of $\mathcal{F}$ of arity m. Then we have

$$F(x_1 x_2 \dots x_n) = y_1 y_2 \dots y_n$$

where

$$x_i \in \{0,1\}^m \quad \text{and} \quad y_i \in \{0,1\} \quad \text{for} \quad 1 \leq i \leq n.$$

Moreover,

$$x_i = (x_{i1}, x_{i2}, \ldots, x_{im}) \quad \text{for} \quad 1 \leq i \leq n.$$

Then, for any $i \geq 1$, there is a function φ_F^i such that

$$y_i = \varphi_F^i(x_1, x_2, \ldots, x_i) = \varphi_F^i(x_{11}, x_{12}, \ldots, x_{1m}, x_{21}, \ldots, x_{2m}, \ldots, x_{i1}, \ldots, x_{im}).$$

Further, for $1 \leq j \leq m$, we set

$$p_j = x_{1j}x_{2j}\ldots x_{nj},$$

and we call p_j the input word of the j-th variable or the j-th input. Then we can also write

$$F(p) = F(p_1, p_2, \ldots p_m).$$

For a Boolean function $f : \{0,1\}^m \to \{0,1\}$, let $F_{(f)}$ be the sequential function with

$$\begin{aligned} F_{(f)}(x_1, x_2, \ldots, x_n) &= f(x_1)f(x_2)\ldots f(x_n) \\ &= f(x_{11}, x_{12}, \ldots, x_{1m})\ldots f(x_{n1}, x_{n2}, \ldots, x_{nm}) \end{aligned}$$

or, equivalently,

$$\varphi^i_{F_{(f)}}(x_{11}, x_{12}\ldots, x_{1m}, x_{21}, \ldots, x_{2m}, \ldots, x_{i1}, \ldots, x_{im}) = f(x_{i1}, x_{i2}, \ldots, x_{im})$$

for $i \geq 1$. Obviously, $F_{(f)}$ realizes f at any moment.

For a unary Boolean function $f : \{0,1\} \to \{0,1\}$ and $a \in \{0,1\}$, we denote by $F_{(f;a)}$ the sequential function with

$$F_{(f;a)}(x_1x_2\ldots x_n) = a\ f(x_1)\ f(x_2)\ldots f(x_{n-1}).$$

$F_{(f;a)}$ realizes f with delay 1 and has output a at the first moment. Note that $F_{(\mathrm{id};1)}$ is the sequential function associated with the logical net of type S.

We now introduce some operations on the set $\mathcal{F}$, which are motivated by the definition of logical nets.

For $F \in \mathcal{F}^m$ and $G \in \mathcal{F}^k$, we set

$$\begin{aligned} (\zeta F)(p_1, p_2, \ldots, p_m) &= F(p_2, p_3, \ldots, p_m, p_1), \\ (\eta F)(p_1, p_2, \ldots, p_m) &= F(p_2, p_1, p_3, p_4, \ldots, p_m), \\ (\Delta F)(p_1, p_2, \ldots, p_{m-1}) &= F(p_1, p_1, p_2, \ldots, p_{m-1}), \\ (\nabla F)(p_1, p_2, \ldots, p_{m+1}) &= F(p_1, p_2, \ldots, p_m), \\ (F \circ G)(p_1, p_2, \ldots, p_{m+k-1}) &= F(G(p_1, p_2, \ldots p_k), p_{k+1}, \ldots, p_{k+m-1}) \end{aligned}$$

By the operation ζ we produce a cyclic shift of the inputs, and η exchanges the first two inputs. It is well-known that by these two special permutations of the inputs we can generate any permutation of the inputs. By Δ we identify two inputs, i.e., their inputs coincide at any moment. Combining ζ, η and Δ we can identify an arbitrary set of inputs. Therefore Δ can be considered as the formal description of the first operation for the construction of logical

nets. ∇ adds a variable, but the result does not depend on the input word of this variable. $\circ$ is the formal description of the substitution or superposition used as the second operation for the construction of logical nets.

We now give a formal definition of the third operation for the construction of logical nets, the feedback operation.

We say that $F \in \mathcal{F}^m$ depends on its first variable in a delayed fashion if, for $i \geq 1$, φ_F^i does not depend on x_{i1}. (This reflects the requirement that any path from the first input to the output contains at least one element of type S.)

If $F \in \mathcal{F}^m$ depends on its first variable delayed, we define $\uparrow F$ by

$$\begin{aligned}\varphi^1_{\uparrow F}(x_{12}, x_{13}, \ldots, x_{1m}) &= \varphi^1_F(x_{11}, x_{12}, x_{13}, \ldots, x_{1m}),\\ \varphi^t_{\uparrow F}(x_{11}, \ldots, x_{1m}, x_{22}, \ldots, x_{2m}, \ldots, x_{i2}, \ldots, x_{im}) &= \varphi^t_F(\varphi^1_{\uparrow F}(z_1), x_{12}, x_{13}, \ldots, x_{1m},\\ \varphi^2_{\uparrow F}(z_2), x_{22}, x_{23}, \ldots, x_{2m}, \ldots, \varphi^{t-1}_{\uparrow F}(z_{t-1}), x_{t-1,2}, \ldots, x_{t-1,m}, x_{t2}, x_{t3}, \ldots, x_{tm})&\end{aligned}$$

for $t \geq 2$ where z_i stands for the tuple $(x_{11}, \ldots, x_{1m}, x_{21}, \ldots, x_{im})$.

Now we define the algebra

$$\underline{\mathcal{F}} = (\mathcal{F}, \{\zeta, \eta, \Delta, \nabla, \circ, \uparrow\}).$$

of sequential functions which will be studied in this paper. Since we do not expect misunderstandings we shall use the notation $\mathcal{F}$ for the algebra, too, i.e., we shall not distinguish—in the notation—between the algebra and its target set. The same will also be done for all other algebras considered in this paper.

By $[M]$ we denote subalgebra generated by $M \subseteq \mathcal{F}$ in $\underline{\mathcal{F}}$.

Sometimes we are also interested in the algebra without the feedback operation, i.e., we consider the algebra

$$\underline{\mathcal{F}'} = (\mathcal{F}, \{\zeta, \eta, \Delta, \nabla, \circ\}).$$

In this case, $\langle M \rangle$ denotes the subalgebra generated by $M \subseteq \mathcal{F}$ in $\underline{\mathcal{F}'}$.

Moreover, for $k \geq 2$, we set

$$P_k = \{f \mid f : \{0, 1, \ldots, k-1\}^m \to \{0, 1, \ldots, k-1\},\ m \geq 1\}.$$

P_k is the set of all functions over a k-element set. P_2 is the set of Boolean functions. It is easy to see that the operations ζ, η, Δ, ∇ and $\circ$ can be defined in P_k, $k \geq 2$, too. Thus we get the algebra

$$\underline{P_k} = (P_k, \{\zeta, \eta, \Delta, \nabla, \circ\}).$$

2.2 Completeness in the algebra of sequential functions

In this subsection we recall the most important facts about classical completeness in $\mathcal{F}$. We start with a definition of the concept.

2.7 Definition A subset M of $\mathcal{F}$ is called *complete* (in $\mathcal{F}$) if the set $[M]$ generated by M is $\mathcal{F}$.

By the definition of logical nets, it is obvious that the set

$$M = \{F_{(\text{non})}, F_{(\text{et})}, F_{(\text{vel})}, F_{(\text{id};1)}\}$$

is complete. If we delete $F_{(\text{et})}$, then the remaining set of sequential functions is complete, too. The following theorem gives the numbers of elements which are possible for complete sets which cannot be reduced.

2.8 Theorem *For any n, there is a complete set $M = \{F_1, F_2, \ldots, F_n\}$ such that $M \setminus \{F_i\}$ is not complete.*

For a proof of Theorem 2.8, we refer to [11]. Moreover, we mention that there is a sequential function $U \in \mathcal{F}^2$ such that $\{U\}$ is complete, as shown by Buevic [1].

2.9 Definition A subalgebra M is called *maximal* (in $\mathcal{F}$), if there is no subalgebra N with $M \subset N \subset \mathcal{F}$.

It is well-known from universal algebra that the subsets generating the whole set can be characterized by the maximal subalgebras if the algebra is finitely generated. Thus we immediately get the following theorem.

2.10 Theorem *M is complete if and only if M is not contained in any maximal subalgebra.*

The cardinality of maximal subalgebras of $\mathcal{F}$ has been determined by V. B. Kudrjavcev [20].

2.11 Theorem *The cardinality of the set of maximal subalgebras is the cardinality of the set of real numbers.*

Proof For a complete proof of the theorem, we refer to [20, 11]. Here we only mention that maximal subalgebras are constructed which depend on a chosen subset of $\mathbb{N}$. Since all these maximal subalgebras are different, one gets that there are at least as many maximal subalgebras as subsets of $\mathbb{N}$.

On the other hand, since the cardinality of $\mathcal{F}$ is countably infinite, there are at most as many subsets of $\mathcal{F}$ as subsets of $\mathbb{N}$. Therefore we cannot have more maximal subalgebras as subsets of $\mathbb{N}$. □

Obviously, by Theorem 2.11, the completeness criterion of Theorem 2.10 does not provide an algorithm to decide the completeness of a finite set. Therefore we are interested in reducing the number of maximal subalgebras which have to be taken into consideration. This is done by the following theorem, but it also provides no algorithm.

2.12 Theorem *There is a countable set N of maximal subalgebras such that $M \subset \mathcal{F}$ is complete if and only if M is not contained in any algebra of N.*

Proof Let M be a finite subset of $\mathcal{F}$ which is not complete. Then $M \subseteq N(M)$ for some maximal subalgebra $N(M)$. We set

$$U = \{N(M) \mid M \text{ is a finite incomplete subset of } \mathcal{F}\}.$$

Since the number of finite subsets of $\mathcal{F}$ is countable, U is countable.

Obviously, if M is complete, then M is not contained in any maximal subalgebra of $\mathcal{F}$, and thus M is not contained in any member of U. If M is finite and incomplete, then M is contained in an element of U. Therefore a finite subset M of $\mathcal{F}$ is complete if and only if it is not contained in an element of the countable set U of maximal subalgebras. □

The completeness criteria of Theorems 2.10 and 2.12, using maximal subalgebras, do not give an algorithm to decide the completeness of a finite set. We shall now prove that there is no such algorithm at all.

2.13 Theorem *There is no algorithm to decide the completeness of a finite subset of* $\mathcal{F}$.

Proof The proof consists of a reduction to finite derivations of words in normal Post systems. Thus we start with a definition of these systems.

A normal Post system is a triple

$$G = (\{x_1, x_2, \ldots, x_n\}, \mu, \{w_1, w_2, \ldots, w_n\})$$

where

- $\{x_1, x_2, \ldots, x_n\}$ is a finite alphabet,
- μ is a positive integer,
- for $1 \leq i \leq n$, w_i is a word over $\{x_1, x_2, \ldots, x_n\}$ associated with the i-th letter x_i.

We introduce a derivation relation with respect to G as follows. For two words u and v over $\{x_1, x_2, \ldots, x_n\}$, we say that $u \Longrightarrow v$ if and only if

$$u = x_{i_1} x_{i_2} \ldots x_{i_r}, \quad r \geq \mu, \quad v = x_{i_{\mu+1}} x_{i_{\mu+2}} \ldots x_{i_r} w_{i_1},$$

i.e., we cancel the first μ letters of u and add at the end the word associated with the first letter of u. Thus any word w induces a derivation

$$w = w_0 \Longrightarrow w_1 \Longrightarrow w_2 \Longrightarrow \ldots \Longrightarrow w_i \Longrightarrow \ldots$$

Such a derivation can be finite or infinite. Finiteness occurs if $|w_i| < \mu$ for some i (since we can only perform a derivation step if the length of the word is at least μ). The following fact is well-known [17].

Fact *There is no algorithm which decides whether or not a given word w has a finite derivation with respect to a given normal Post system G.*

We now consider the alphabet $X = \{0, 1, 2, x_1, x_2, \ldots, x_n\}$ of $n + 3$ letters. Further, let $\mathcal{S}^m$ be the set of all sequential functions mapping $(X^m)^+$ into X^+, and let $\mathcal{S}$ be the union of all $\mathcal{S}^m$ taken over $m \in \mathbb{N}$. Obviously, we can form the algebra $\underline{\mathcal{S}}$ with the same operations as in $\underline{\mathcal{F}}$. We note that there also is a function $U \in \mathcal{S}^2$ such that $[\{U\}] = \mathcal{S}$ (see the remark after Theorem 2.8).

Let G be a normal Post system specified as above, and let w be an arbitrary word over $\{x_1, x_2, \ldots, x_n\}$.

We now consider the functions $F_w \in \mathcal{S}^1$, $F_G \in \mathcal{S}^1$ and $F_U \in \mathcal{S}^3$ with

$$\begin{aligned} F_w(y_1y_2\dots y_n) &= 2w1^{n-|w|-1} \quad \text{for} \quad n \geq |w|+1, \\ F_G(2^r u 1^s) &= 2^{r+\mu}v1^{s'} \quad \text{where} \quad |u| \geq \mu,\ u \Longrightarrow v,\ r+|u|+s = r+\mu+|v|+s' \\ F_G(2^r u 1^s) &= 2^{r+|u|}1^s \quad \text{where} \quad |u| < \mu, \\ F_U(2^r 1^s, q_1, q_2) &= U(q_1, q_2). \end{aligned}$$

Note that the functions F_w, F_G and F_U are not completely defined. We complete their definitions by the following two conditions:

- if p is a prefix of an input word mentioned in the conditions above, then the output is the corresponding prefix of the output;
- if t is the first moment where the input word is not a prefix of a word mentioned in the conditions above, then the output is 0 at all moments $j \geq t$.

Then F_w is a constant function which essentially produces the word w, F_G simulates a derivation step according to G, and F_U simulates the function U.

It is easy to construct automata which induce the functions F_w, F_G and F_U (using an automaton inducing U). Therefore F_w, F_G and F_U are sequential functions.

We now prove that $\{F_w, F_G, F_U\}$ is complete in $\mathcal{S}$ if and only if w has a finite derivation.

Assume that w has a finite derivation, and let w_i be the word of length smaller than μ. Then the sequential function

$$H = \underbrace{F_G \circ (F_G \circ (\dots (F_G}_{i+1 \text{ times}} \circ F_w)\dots))$$

produces as output a finite prefix of the infinite word $2^{1+i\mu+|w_i|}11\dots$ Thus $\Delta(F_U \circ H)$ coincides with the function U which is complete for $\mathcal{S}$. Thus $\{F_w, F_G, F'\}$ is complete, too.

The idea of the converse statement is an analysis of the way to obtain U from the elements of the complete set $\{F_w, F_G, F_U\}$. The last element of the corresponding net cannot be F_w or F_G since they give the output 2 or 0 at the first moment which does not hold for U. Thus the last element is F_U. Going backwards we analogously show that essentially the net has to have the form given above for U. Thus w has to have a finite derivation for w. A detailed proof of this fact is given in [11].

Thus the completeness of a finite subset of $\mathcal{S}$ cannot be decided. To transform this to $\mathcal{F}$ we code the values of X by binary words and modify F_w, F_G and F_U accordingly. □

We also recall some facts on completeness in the algebra P_2 (for the proofs we refer to [16]).

A function $f \in P_2^n$ is called *linear* if there are elements $c_0, c_1, \dots, c_n$ of $\{0,1\}$ such that

$$f(x_1, x_2, \dots, x_n) = c_0 + c_1x_1 + c_2x_2 + \dots + c_nx_n \mod 2.$$

By Lin we denote the set of all linear functions of P_2.

For $i \in \{0,1\}$, Q_i is the set of all functions f such that

$$f(x_1, x_2 \dots, x_n) = f(y_1, y_2, \dots, y_n) = i$$

implies the existence of j, $1 \leq j \leq n$, with $x_j = y_j = i$.

Further we set

$$\begin{aligned} T_0 &= \{f \mid f(0,0,\ldots,0) = 0\}, \\ T_1 &= \{f \mid f(1,1,\ldots,1) = 1\}, \\ \mathrm{Sd} &= \{f \mid f(x_1,x_2,\ldots,x_n) = \mathrm{non}(f(\mathrm{non}(x_1),\mathrm{non}(x_2),\ldots,\mathrm{non}(x_n)))\}, \\ \mathrm{Mon} &= \{f \mid x_i \leq y_i \text{ for } 1 \leq i \leq n \text{ implies } f(x_1,x_2,\ldots x_n) \leq f(y_1,y_2,\ldots,y_n)\}. \end{aligned}$$

We also say that the functions in T_0, T_1, Sd and Mon are 0-preserving, 1-preserving, self-dual and monotone, respectively.

The following completeness criteria hold.

2.14 Theorem

(1) *A subset M of P_2 is complete in P_2 if and only if it is not contained in T_0, T_1,* Lin*,* Sd *and* Mon.

(2) *For $i \in \{0,1\}$, a subset M of T_i satisfies $\langle M \rangle = T_i$ if and only if M is not contained in Q_i,* $\mathrm{Lin} \cap T_i$*,* $\mathrm{Sd} \cap T_i$*,* $\mathrm{Mon} \cap T_i$ *and* $T_0 \cap T_1$*.*

2.3 Completeness with respect to equivalence relations

In this subsection we define a modified concept of completeness which will be the subject of this paper. Although this notion has not been discussed before, most modifications of completeness which have been investigated in recent years (especially in the seventies) are covered by our concept.

2.15 Definition Let ϱ be an equivalence relation on $\mathcal{F}$.

(1) A subalgebra M of $\mathcal{F}$ is called a *ϱ-algebra* if and only if $M \cap K \neq \emptyset$ for all equivalence classes K of ϱ.

(2) A subset M of $\mathcal{F}$ is called *ϱ-complete* if and only if $[M]$ is a ϱ-algebra.

The requirement of ϱ-completeness does not require for the generation of any sequential function; it is sufficient to generate a representive of any equivalence class.

Since we are interested in obtaining algebraic completeness criteria (as given in Theorem 2.10) for ϱ-completeness too, we define a notion analogous to maximality.

2.16 Definition Let ϱ be an equivalence relation on $\mathcal{F}$. A subalgebra M of $\mathcal{F}$ is called *ϱ-maximal* if and only if the following conditions are satisfied:

- M is not a ϱ-algebra;
- any subalgebra N with $M \subset N$ is a ϱ-algebra.

We illustrate these concepts by an example.

2.17 Example We consider the equivalence relation ϱ_1 defined by

$$\varrho_1 = \{(F,G) \mid \text{arity}(F) = \text{arity}(G),\ F((0,0,\ldots,0)) = G((0,0,\ldots,0))\}.$$

It is easy to see that the equivalence classes of ϱ_1 are given by

$$K_{n,a} = \{F \mid \text{arity}(F) = n \text{ and } F((0,0,\ldots,0)) = a\},$$

where $n \in \mathbb{N}$ and $a \in \{0,1\}$.

Obviously, $\mathcal{F}$ is a ϱ_1-algebra (this statement holds for any equivalence relation). But there are further ϱ_1-algebras. Here we only present the subalgebra with the target set

$$Q = \{F \mid F_p(q) = 0^{|q|} \text{ for } |p| \geq 1\},$$

which is the set of all sequential functions where, for $t \geq 2$, φ^t_F is the constant k_0. It is easy to see that these functions form a subalgebra. Moreover, it is a ϱ_1-subalgebra since, for any Boolean function f, it contains a sequential function F with $\varphi^1_F = f$.

Since by applications of ∇ and Δ, we can obtain functions of arbitrary arity, M is ϱ_1-complete if and only if $[M]$ contains at least one F with $F((0,0,\ldots,0)) = 0$ and at least one G with $G((0,0,\ldots,0)) = 1$.

If M only contains functions F with $F((0,0,\ldots,0)) = 0$, then this holds for $[M]$ too. Therefore M contains a function G with $G((0,0,\ldots,0)) = 1$. Let G' be the sequential function obtained from G by identification of all inputs.

First let us assume that $G(1,1,\ldots,1) = 0$. Then, $\varphi^1_{G'} = \text{non}$. Thus $\varphi^1_{G' \circ G'} = \text{id}$, i.e., $(G' \circ G')(0) = 0$. Thus M is ϱ_1-complete.

We now assume that $\varphi^1_G(1,1,\ldots,1) = 1$ and that the function φ^1_G is not a constant. Then there is a tuple $(a_1,a_2,\ldots,a_r)$ with $\varphi^1_G(a_1,a_2,\ldots,a_r) = 0$. For $1 \leq i \leq r$, we define the sequential functions H_i by

$$H_i = \begin{cases} F_{(\text{id})} & \text{if } a_i = 0 \\ G' & \text{if } a_i = 1 \end{cases}$$

Then the sequential function $H \in \mathcal{F}^1$ defined by

$$H(p) = G(H_1(p), H_2(p), \ldots, H_r(p))$$

satisfies $\varphi^1_H(0) = 0$. Thus M is also ϱ_1-complete in this case.

Finally, let us assume that φ^1_G is the (r-ary) constant k_1. Then M has to contain a function F with $F((0,0,\ldots,0)) = 0$ in order to be ϱ_1-complete.

Thus we have the following decidable criterion for ϱ_1-completeness: M is ϱ_1-complete if and only if M contains at least one function G where $G((0,0,\ldots,0)) = 1$ and φ^1_G is not a constant or M contains at least a function G where φ^1_G is the constant k_1 and one F with $F((0,0,\ldots,0)) = 0$.

By this criterion (or already by the definition),

$$M_{T_0} = \{F \mid F((0,0,\ldots,0)) = 0\}$$

is an example for a ϱ_1-maximal subalgebra.

3 Completeness with respect to congruence relations

3.1 Congruence relations on $\mathcal{F}$

From the algebraic point of view congruence relations are the most interesting equivalence relations since they can be used for structural characterizations, to induce a quotient algebra and appear at least implicitly in many other concepts.

Therefore it is of great interest to study completeness with respect to congruence relations. As a first step we have to know all congruence relations on $\mathcal{F}$. In this subsection we prove that there are only four types of congruence relations.

We define the following relations on $\mathcal{F}$:

- the complete relation $\sigma_C = \{(F, G) \mid F, G \in \mathcal{F}\}$,
- the equality relation $\sigma_E = \{(F, G) \mid F = G\}$,
- the arity relation $\sigma_A = \{(F, G) \mid \text{arity}(F) = \text{arity}(G)\}$,
- for any $t \in \mathbb{N}$, the t-bounded equality relation

$$\sigma_t = \{(F, G) \mid \text{arity}(F) = \text{arity}(G) \text{ and } F(p) = G(p) \text{ for all } p \text{ with } |p| \leq t\}.$$

The easy proof of the following statement is left to the reader.

3.1 Lemma *σ_C, σ_E, σ_A and σ_t for $t \geq 1$ are congruence relations on $\underline{\mathcal{F}'}$ as well as on $\underline{\mathcal{F}}$.*

We now prove that there are no other congruence relations on $\underline{\mathcal{F}}$ and $\underline{\mathcal{F}'}$.

3.2 Lemma *Let σ be a congruence relation on $\mathcal{F}'$ such that $(F, G) \in \sigma$ holds for two sequential functions F and G of different arity. Then σ coincides with the complete relation.*

Proof We omit the proof which can be given analogously to Lemma 1 in [13] and Lemma 1 in [22]. □

3.3 Theorem

(1) *σ_C, σ_E, σ_A and σ_t for $t \geq 1$ are the only congruence relations on $\underline{\mathcal{F}'}$.*

(2) *σ_C, σ_E, σ_A and σ_t for $t \geq 1$ are the only congruence relations on $\underline{\mathcal{F}}$.*

Proof We only give the proof of the first statement. The second statement follows from Lemma 3.1 since additional operations can only delete some congruence relations.

Let σ be a congruence relation on $\mathcal{F}'$ such that σ is not the complete relation and not the equality relation. We define $t \in \mathbb{N}_0$ as the integer such that the following two conditions are satisfied:

- $(F', G') \in \sigma$ implies $F'(q) = G'(q)$ for all words q with $|q| \leq t$;
- there are two sequential functions F and G and a word p of length $t + 1$ such that $(F, G) \in \sigma$ and $F(p) \neq G(p)$.

By our assumptions on σ such a number exists. We assume that $t \geq 1$. Then σ is a refinement of the t-bounded equality since the arities of functions in relation have to coincide by Lemma 3.2. Let $r = \text{arity}(F) = \text{arity}(G)$.

Let $p = (p_1, p_2, \ldots, p_r)$ and $p_j = x_{1j}x_{2j}\ldots x_{t+1,j}$ for $1 \leq j \leq r$. For $1 \leq i \leq r$, let F_i be the sequential function of $\mathcal{F}^r$ such that

$$\varphi^k_{F_i} = \begin{cases} x_{ki} & \text{if } k \leq t+1, \\ 0 & \text{if } k > t+1. \end{cases}$$

Then we consider the sequential functions F' and G' defined by

$$F'(u) = F(F_1(u), F_2(u), \ldots, F_r(u)) \quad \text{and} \quad G'(u) = G(F_1(u), F_2(u), \ldots, F_r(u)).$$

Clearly, F' and G' are constant functions and

$$\varphi^l_{F'}(x_1, x_2, \ldots, x_l) = \varphi^l_{G'}(x_1, x_2, \ldots, x_l) \quad \text{for} \quad l \neq t+1,$$
$$y = \varphi^{t+1}_{F'}(x_1, x_2, \ldots, x_{t+1}) \neq \varphi^{t+1}_{G'}(x_1, x_2, \ldots, x_{t+1}) = z.$$

Moreover, $(F', G') \in \sigma$.

Let H be the binary sequential function such that

$$\varphi^l_H((v_{11}, v_{21}), (v_{12}, v_{22}), \ldots, (v_{1l}, v_{l2})) = \begin{cases} v_{1l} & \text{if } l \leq t, \\ v_{1l} & \text{if } l \geq t+1,\ v_{2,t+1} = y, \\ 0 & \text{if } l \geq t+1,\ v_{2,t+1} = z. \end{cases}$$

If the value of the second input at moment $t+1$ is y, then the output of H is the first input at every moment. Otherwise, the output is 0 at any moment $j \geq t+1$.

For an arbitrary sequential function $R \in \mathcal{F}^r$, we now define the sequential functions H_{1R} and H_{2R} by

$$H_{1R}(q_1, q_2, \ldots, q_r) = H(R(q_1, q_2, \ldots, q_r), F'(q_1)),$$
$$H_{2R}(q_1, q_2, \ldots, q_r) = H(R(q_1, q_2, \ldots, q_r), G'(q_1)).$$

Obviously, we have $(H_{1R}, H_{2R}) \in \sigma$. By the definition of H and $y = \varphi^{t+1}_{F'}(x_1, x_2, \ldots, x_{t+1})$, $H_{1R} = R$. H_{2R} coincides with R at the first t moments and has the output 0 at the other moments.

Now assume that $(R, S) \in \sigma_t$ for two functions of $\mathcal{F}^r$. Then R and S coincide at the first t moments. Thus $H_{2R} = H_{2S}$. Therefore we have the relations

$$R = H_{1R}, \quad (H_{1R}, H_{2R}) \in \sigma, \quad H_{2R} = H_{2S}, \quad (H_{2S}, H_{1S}) \in \sigma, \quad H_{1S} = S$$

which implies that $(R, S) \in \sigma$. Hence σ_t is a refinement of σ too.

Thus $\sigma = \sigma_t$.

If $t = 0$, then we conclude analogously that σ is the arity relation. □

We note that the relations σ_t for $t \geq 1$ are not only of interest because they are congruence relations. Practically we cannot observe an infinite behaviour of an automaton, i.e., there is a bound t and we can restrict our observations to the behaviour at the first t moments. Thus we are only interested in sequential function up to σ_t-equivalence.

3.2 Results on completeness with respect to congruences

In this subsection we discuss completeness with respect to the congruence relations determined in the preceding subsection.

By the definition of σ_E, σ_E-completeness coincides with "classical" completeness. Therefore we can refer to Section 2.3.

By the definition of σ_C-completeness, we immediately get the following results.

3.4 Theorem

(1) *Any non-empty subalgebra of $\mathcal{F}$ is a σ_C-algebra.*

(2) *Any non-empty subset of $\mathcal{F}$ is σ_C-complete.*

(3) *$\emptyset$ is the only σ_C-maximal subalgebra.*

(4) *The σ_C-completeness of a finite subset of $\mathcal{F}$ is decidable.*

By the definition of σ_A-completeness, a set M is complete if and only if, for any $n \in \mathbb{N}$, it contains at least one n-ary function. From an arbitrary m-ary function F, for any $m' < m$ and any $m'' > m$, we can obtain an m'-ary function by identification of inputs (i.e. by application of operation Δ sometimes) and an m''-ary function by application of ∇ sometimes. Thus M is σ_A-complete if and only if M contains at least one function. Therefore M is σ_A-complete iff M is σ_C-complete. Thus we get the following statements from Theorem 3.4.

3.5 Theorem

(1) *Any non-empty subalgebra of $\mathcal{F}$ is a σ_A-algebra.*

(2) *Any non-empty subset of $\mathcal{F}$ is σ_A-complete.*

(3) *$\emptyset$ is the only σ_A-maximal subalgebra.*

(4) *The σ_A-completeness of a finite subset of $\mathcal{F}$ is decidable.*

We now discuss σ_t-completeness where $t \in \mathbb{N}$. We start with a definition which extends σ_t-equivalence to sets.

3.6 Definition Let $t \in \mathbb{N}$. We say that two sets M and M' are *σ_t-equivalent* if and only if, for any two functions $F \in M$ and $F' \in M'$, there are functions $G \in M'$ and $G' \in M$ such that $(F, G) \in \sigma_t$ and $(F', G') \in \sigma_t$.

First we show that we can restrict ourselves to the algebra $\mathcal{F}'$, i.e., we do not have to take into consideration the feedback operation.

3.7 Theorem *For any set $M \subset \mathcal{F}$ and any $t \in \mathbb{N}$, $[M]$ and $\langle M \rangle$ are σ_t-equivalent.*

Proof Let $F' \in \langle M \rangle$. Since $M' = \langle M \rangle$ is a subset of $[M]$, we obtain $F' \in [M]$. Choosing $G' = F'$, we get $G' \in [M]$ and $(G', F') \in \sigma_t$, which proves that G' satisfies the requirements of Definition 3.6.

In order to show the other statement we prove that, for any $t \in \mathbb{N}$ and any sequential function F, the function

$$F_t = \underbrace{F \circ (F \circ (\dots (F}_{t-1 \text{ times}} \circ F) \dots))$$

is σ_t-equivalent to $\uparrow F$. Using this construction instead of the feedback operation we obtain for any $F \in [M]$ the corresponding $G \in \langle M \rangle$.

Let $t = 1$. Since F depends in a delayed fashion on its first variable, F and $\uparrow F$ are σ_1-equivalent. Thus the statement holds for $t = 1$.

Let $t \geq 2$. Obviously we have $\uparrow F = F \circ \uparrow F$ since we use the same input for the first variable. If we also take into consideration that F depends in a delayed fashion on its first variable we obtain that F and F_t are σ_t-equivalent.

Choosing $G = F_t$, we satisfy the requirements of Definition 3.6. □

3.8 Theorem *For any $t \in \mathbb{N}$, there is a mapping $\tau_t : \mathcal{F} \to P_{2^t}$ such that $M \subset \mathcal{F}$ is σ_t-complete if and only if $\tau_t(M) = \{\tau_t(F) \mid F \in M\}$ generates $\tau_t(\mathcal{F})$.*

Proof Obviously, it is sufficient to study the behaviour of the sequential functions at the first t moments. Thus it is sufficient to consider words of length at most t. Further, the value $F(q_1, q_2, \dots, q_n)$ where q_i is a prefix of p_i for $1 \leq i \leq n$ is determined by the corresponding prefix of $F(p_1, p_2, \dots, p_n)$. Thus it is sufficient to consider only $F(p_1, p_2, \dots, p_n)$ where, for $1 \leq i \leq n$, p_i is a word of length t.

Now we interpret a word $p = x_1 x_2 \dots x_t$ of length t over $\{0, 1\}$ as a binary representation of a natural number $\varphi(p) \in \{0, 1, \dots, 2^t - 1\}$. Then with $F \in \mathcal{F}^n$ we can associate the function $\tau_t(F) \in P_{2^t}$ where

$$\tau_t(F)(z_1, z_2, \dots, z_n) = \varphi(F(\varphi^{-1}(z_1), \varphi^{-1}(z_2), \dots, \varphi^{-1}(z_n)).$$

For a set $M \in \mathcal{F}$, we set

$$\tau_t(M) = \{\tau_t(F) \mid F \in M\}.$$

It is easy to see that, for any functions F and G,

$$\alpha(\tau_t(F)) = \tau_t(\alpha(F)) \quad \text{for} \quad \alpha \in \{\zeta, \eta, \Delta, \nabla\}, \tag{3.1}$$

$$\tau_t(F) \circ \tau_t(G) = \tau_t(F \circ G). \tag{3.2}$$

Thus

$$\tau_t(\langle M \rangle) = \langle \tau_t(M) \rangle.$$

The statement now follows. □

By Theorem 3.8 we have reduced σ_t-completeness to the generation of the special set $\tau_t(\mathcal{F})$ of functions over a set of 2^t elements by such functions. We note that $\tau_t(\mathcal{F})$ is not P_{2^t}. We consider a unary function F. Then $F(x_1 x_2 x_3 \dots x_t) = y_1 y_2 y_3 \dots y_t$ implies $F(x_1 x_2' x_3' \dots x_t') = y_1 y_2' y_3' \dots y_t')$, i.e., the value $\tau_t(F)$ cannot be chosen arbitrarily on $\varphi(x_1 x_2' x_3' \dots x_t')$.

Obviously, for any $t \in \mathbb{N}$, $\{F_{(\text{non})}, F_{(\text{et})}, F_{(\text{vel})}, F_{(\text{id};1)}\}$ is a finite σ_t-complete set. From the general statements of universal algebra, we obtain the following criterion.

3.9 Theorem *Let $t \in \mathbb{N}$. $M \subseteq \mathcal{F}$ is σ_t-complete if and only if M is not contained in any σ_t-maximal subalgebra of $\mathcal{F}$.*

Without proof we give the following statement.

3.10 Theorem *For any $t \in \mathbb{N}$, there is a finite number of σ_t-maximal subalgebras.*

It is well-known that the maximal subalgebras of P_k, $k \geq 2$, can be described by relations. Since the σ_t-completeness corresponds to the generation of a subset of P_{2^t} one can expect that the σ_t-maximal subalgebras can be described by relations, too. Therefore we recall this concept.

Let $t \in \mathbb{N}$. Let R be a k-ary relation on $\{0,1\}^t$. We say that the n-ary sequential function F preserves the relation R if, for any words $p_{ij} \in \{0,1\}^t$, $1 \leq i \leq k$, $1 \leq j \leq n$, $(p_{1j}, p_{2j}, \dots p_{kj}) \in R$ for $1 \leq j \leq n$ implies

$$(F(p_{11}, p_{12}, \dots, p_{1n}), F(p_{21}, p_{22}, \dots, p_{2n}), \dots, F(p_{k1}, p_{k2}, \dots, p_{kn})) \in R.$$

(This requirement can be interpreted as follows: Consider the words p_{ij} as elements of a (k,n)-matrix. If the columns of the matrix belong to R, then the vector formed by the values obtained by applying F to the rows belongs to R, too.) For a given relation R, let $M(R)$ be the set of all sequential functions preserving R.

In [3], for any $t \in \mathbb{N}$ and any σ_t-maximal subalgebra M, Buevic has given a relation R such that $M = M(R)$. Since we can check whether or not a function preserves a given relation, the result by Buevic and Theorem 3.9 lead to an algorithm to decide the σ_t-completeness of a given finite set.

This decidability is stated in the following theorem. However, since the complete proof of the Buevic result is very long (more than 100 pages) and cannot be given here, we give an alternative proof of the decidability of σ_t-completeness, especially, because this method can be used in many other cases, too.

3.11 Theorem *For any $t \in \mathbb{N}$, the σ_t-completeness of a finite subset of $\mathcal{F}$ is decidable.*

Proof For any set $N \subseteq P_k$, let

$$U(N) = \{f \mid f = \alpha(h \circ h'),\ f \in P_{2^t}^2,\ h \in N,\ h' \in N \cup \{\mathrm{id}\},\ \alpha \text{ is a composition of } \zeta, \eta, \Delta, \nabla\}.$$

Using the empty composition for α and $h' = \mathrm{id}$, we reproduce h. Thus $N \subseteq U(N)$. For any N, $U(N) \subseteq P_k^2$, and thus $U(N)$ contains at most k^{k^2} elements. Moreover, if N is finite, then $U(N)$ can algorithmically be determined.

Further we set

$$H = \{\tau_t(F_{(\mathrm{non})}), \tau_t(F_{(\mathrm{et})}), \tau_t(F_{(\mathrm{vel})}), \tau_t(F_S)\}.$$

Note that $\langle H \rangle = \tau_t(\mathcal{F})$.

We consider the following algorithm which takes an arbitrary finite set $M \subset \mathcal{F}$ as input:

$N_1 := \emptyset$;
$N_2 = U(\tau_t(M))$;
A: `if` $N_1 = N_2$ `then goto B` ;
$N_1 := N_2$;
$N_2 := U(N_1)$;
`goto` A;
B: `if` $H \subset N_2$ `then` Output: "*M is σ_t-complete.*" ;
Output: "*M is not σ_t-complete.*"

The algorithm produces the sequence

$$M_1 = U(\tau_t(M),\ M_2 = U(U(\tau_t(M))),\ M_3 = U(U(U(\tau_t(M))),\ M_4 = U(U(U(U(\tau_t(M)))),\dots$$

Since $M_i \subseteq M_{i+1} \subseteq P^2_{2^t}$ for $i \geq 1$, there is a j, $j \geq 1$ such that $M_j = M_{j+1}$. Thus the condition in row A is satisfied after some steps and we continue with row B.

Moreover, for $i \geq 1$, $M_i \subseteq \tau_t(\langle M\rangle)$ is valid and hence

$$M_j = \tau_t(\langle M\rangle) \cap P^2_{2^t}.$$

If the condition in row B is satisfied, then

$$\tau_t(\mathcal{F}) = \langle H\rangle \subseteq \langle M_j\rangle \subseteq \tau_t(\langle M\rangle) \subseteq \tau_t(\mathcal{F})$$

for some j, $j \geq 1$, which implies $\tau_t(\langle M\rangle) = \tau_t(\mathcal{F})$. By Theorem 3.8, M is σ_t-complete. On the other hand, if H is not a subset of $\tau_t(\langle M\rangle) \cap P^2_{2^t}$, then M cannot be complete. □

4 Kleene-completeness

In the preceding section we considered equivalence relations which are important from the algebraic point of view. Now we consider an equivalence relation which is based on the fact that automata or sequential functions can be used as acceptors for languages.

4.1 Definition A set L of words over X is called *regular* if and only if L can be obtained from $\emptyset$ and $\{x\}$, where $x \in X$, by iterated application of union, product and Kleene-closure.

As an example, we mention that, for $1 \in X$, $X^*\{1\}\{1\}$ is regular.

Now we introduce a notion of acceptance of a language by a sequential function.

4.2 Definition Let $F : X^* \to Y^*$ be a sequential function and $\emptyset \subset Y' \subset Y$. Then the language $T(F, Y')$ accepted by F and Y' is defined as the set of all words p such that $F(p) = p'y$ with $y \in Y'$.

Intuitively, we accept a word p of length t if taking p as input the output at the moment t is an element of Y'.

4.3 Example We consider the function F of Example 2.2. Then the output 1 is produced at the moment t if and only if the input at the moments $t-1$ and t was 1. Therefore $T(F, \{1\})$ is the set of all words ending with two 1's or formally

$$T(F, \{1\}) = \{0, 1\}^*\{1\}\{1\}.$$

Usually, in automata and language theory we use the notion of acceptance of languages by states, i.e. a word w is accepted by an automaton $\mathcal{A} = (X, Z, z_0, T, \delta)$ (without output), where T is a subset of Z, if $\delta^*(z_0, w) \in T$. Obviously, if—in addition—one defines the output function $\gamma : Z \times X \to \{0,1\}$ by $\gamma(z,x) = 1$ if and only if $\delta(z,x) \in T$ (and $\gamma(z,x) = 0$ otherwise), then the automaton and its sequential function accept the same set by state set T and output set $\{1\}$.

The concepts of regular languages and languages accepted by a sequential function are connected by the famous theorem of Kleene.

4.4 Theorem *A set $L \subset X^+$ is regular if and only if there are a sequential function $F : X^+ \to Y^+$ and a set Y', $\emptyset \subset Y' \subset Y$, such that $T(F, Y') = L$.*

We omit a proof of Kleene's Theorem. It is given in any textbook on the basics of theoretical computer science using acceptance by states and can be transformed easily to output acceptance.

4.5 Definition Two sequential functions $F : X^+ \to Y^+$ and $G : X^+ \to Z^+$ are *Kleene-equivalent* (written as $(F, G) \in \sigma_K$) if and only if there are sets Y' and Z' with $\emptyset \subset Y' \subset Y$ and $\emptyset \subset Z' \subset Z$ such that $T(F, Y') = T(G, Z')$.

We omit the easy proof of the following lemma.

4.6 Lemma *σ_K is an equivalence relation on $\mathcal{F}$.*

Assume that F and G are Kleene-equivalent sequential functions of $\mathcal{F}$. Then $T(F, Y) = T(G, Y')$. Since $\emptyset \subset Y \subset \{0,1\}$ and $\emptyset \subset Y' \subset \{0,1\}$, either of the sets Y and Y' has to be $\{0\}$ or $\{1\}$.

Let us assume that $Y = Y' = \{1\}$. Then, for any $t \in \mathbb{N}$, the functions φ_F^i and φ_G^i have to coincide, since they have to give 1 if and only if the input belongs to their accepted sets (which are equal). Thus $F = G$.

If we assume that $Y = \{0\}$ and $\{Y'\} = \{1\}$, then

$$\varphi_F^i(x_1, x_2, \ldots, x_n) = 0 \quad \text{if and only if} \quad \varphi_G^i(x_1, x_2, \ldots, x_n) = 1$$

for $i \in \mathbb{N}$. Thus we have

$$\varphi_F^i = (\text{non}) \circ \varphi_G^i \quad \text{for} \quad i \in \mathbb{N}$$

and

$$F = F_{(\text{non})} \circ G.$$

The other cases can be handled analogously. Thus we get the following assertion.

4.7 Lemma *Let $F \in \mathcal{F}^m$ and $G \in \mathcal{F}^m$. $(F, G) \in \sigma_K$ if and only if $F = G$ or $F = F_{(\text{non})} \circ G$ (and $G = F_{(\text{non})} \circ F$).*

We now give an example of a σ_K-algebra.

4.8 Example Let $i \in \{0,1\}$. We set

$$M_{T_i} = \{F \mid F((i, i, \ldots, i)) = i\}.$$

Obviously, $F \in M_{T_i}$ if and only if $\varphi_F^1 \in T_i$. It is easy to prove that M_{T_i} is a subalgebra of $\mathcal{F}$.

We now show that M_{T_0} is a σ_K-algebra; the proof for M_{T_1} can be given analogously.

Let R be a regular set over $\{0,1\}^n$ and let R be accepted by the sequential functions F and G by $\{0\}$ and $\{1\}$, respectively. If $(0,0,\ldots,0) \in R$, then $F \in M_{T_0}$ since $F((0,0,\ldots,0)) \in \{0\}$. If $(0,0,\ldots,0) \notin R$, then $G((0,0,\ldots,0)) \notin \{1\}$, i.e., $G \in M_{T_0}$. In both cases we have a sequential function in M_{T_0} which accepts R. Hence M_{T_0} is a σ_K-algebra.

First we determine all σ_K-algebras.

4.9 Theorem *If M is a σ_K-algebra of $\mathcal{F}$, then $M = M_{T_0}$ or $M = M_{T_1}$ or $M = \mathcal{F}$.*

Proof We only give an outline of the proof; for a complete proof we refer to [5, 9, 11].

Step 1. We consider the regular set $\{(0,1)\} \cup (\{0,1\}^2)^+\{(0,1)\}$ which is accepted by the sequential functions $F_{(f)}$ and $F_{(\mathrm{non})} \circ F_{(f)} = F_{(\mathrm{non} \circ f)}$ and $\{1\}$ and $\{0\}$, respectively, where

$$f(0,1) = 1 \quad \text{and} \quad f(x,y) = 0 \text{ for } (x,y) \neq (0,1).$$

It is easy to see that f and $\mathrm{non} \circ f$ are not monotone. Thus any σ_K-algebra M contains a sequential function $F_{(h)}$ where $h \notin \mathrm{Mon}$.

By analogous arguments one can prove that any σ_K-algebra contain sequential functions $F_{(h_i)}$, $1 \leq i \leq 5$, where $h_1 \notin \mathrm{Sd}$, $h_2 \notin \mathrm{Lin}$, $h_3 \notin Q_0$, $h_4 \notin Q_1$ and $h_5 \notin T_0 \cap T_1$, respectively. By Theorem 2.14 (2),

$$M \supseteq \{F_{(f)} \mid f \in T_i\} \tag{4.1}$$

for some $i \in \{0,1\}$.

The regular set $\{0,1\}^*\{0\}\{0,1\}$ is accepted by the functions $F_{(\mathrm{id};1)}$ and $F_{(\mathrm{non};0)}$ and $\{0,1\}^*\{1\}\{0,1\}$ is accepted by the functions $F_{(\mathrm{id};0)}$ and $F_{(\mathrm{non};1)}$. Thus we get eight possibilities depending on the i from (4.1) and the function used to accept $\{0,1\}^*\{0\}\{0,1\}$ and $\{0,1\}^*\{1\}\{0,1\}$. A detailed discussion of these cases leads to

$$M \supseteq \{F_{(f)} \mid f \in T_i\} \cup \{F_{(\mathrm{id};i)}, F_{(\mathrm{non};i)}\} \tag{4.2}$$

for some $i \in \{0,1\}$.

Step 2. We prove that, for $i \in \{0,1\}$,

$$M_{T_i} = [\{F_{(f)} \mid f \in T_i\} \cup \{F_{(\mathrm{non};i)}\}].$$

If we combine this relation with (4.2), then we get $M \supseteq M_{T_i}$ for some $i \in \{0,1\}$.

Step 3. We prove that M_{T_0} and M_{T_1} are maximal subalgebras of $\mathcal{F}$. This implies $M = M_{T_0}$ or $M = M_{T_1}$ or $M = \mathcal{F}$. □

4.10 Theorem *There is no algorithm which decides the σ_K-completeness of a finite set.*

Proof Assume that there is an algorithm to decide the σ_K-completeness for a finite set M. We show that this implies that there exists an algorithm to decide the completeness of a finite set which contradicts Theorem 2.13.

Let M be a finite set. First we decide whether M is σ_K-complete. If the answer is "no", then M is not complete (because any complete set is σ_K-complete). If M is σ_K-complete, then one of the following three cases occurs:

$$[M] = M_{T_0} \quad \text{or} \quad [M] = M_{T_1} \quad \text{or} \quad [M] = \mathcal{F}.$$

Let $i \in \{0,1\}$. Obviously, $[M] = M_{T_i}$ holds if and only if all elements of M belong to M_{T_i}. In order to check whether or not a sequential function F belongs to M_{T_i} we have only to test whether the output of F at the first moment is i if the input at the first moment is the tuple consisting of i's only. If the answer is "yes" for any function of M, then $[M] = M_{T_i}$.

Now we check whether $[M] = M_{T_0}$ or $[M] = M_{T_1}$. If the answer is "yes", for some $i \in \{0,1\}$, then M is not complete. If the answer is no in both cases, then $[M] = \mathcal{F}$ has to hold. Therefore M is complete. □

4.11 Theorem

(1) *$M \subset \mathcal{F}$ is σ_K-complete if and only if M is not contained in any σ_K-maximal subalgebra.*

(2) *The cardinality of the set of σ_K-maximal subalgebras of $\mathcal{F}$ is the cardinality of the set of real numbers.*

(3) *There is a countable set N of σ_K-maximal subalgebras of $\mathcal{F}$ such that $M \subset \mathcal{F}$ is complete if and only if M is not contained in any algebra of N.*

Proof (1) Since any σ_K-algebra is finitely generated, the statement follows from known facts of universal algebra.

(2) Any maximal subalgebra of $\mathcal{F}$ which is different from M_{T_0} and M_{T_1} is σ_K-maximal too. By Theorem 2.11, the statement follows.

(3) can be proved analogously to the proof of Theorem 2.12. □

We now discuss some special cases by restricting to special regular sets, i.e., we do not require that any regular language be accepted; we only ask for acceptance of some special classes of regular languages.

First, we take into consideration only regular languages contained in $(\{0,1\}^n)^+$, where $n \in \mathbb{N}$ is a natural number.

Let $n \geq 2$. Obviously, if $M \subseteq \mathcal{F}$ is a σ_K-algebra, then, for any regular set $R \subseteq (\{0,1\}^n)^+$, there is a sequential function in M which accepts R.

We now prove the converse statement. For a regular set $R \subseteq (\{0,1\}^2)^+$, we define the extension $R(n)$ as follows: Let $x_i = (x_{i1}, x_{i2}, \ldots, x_{in})$ for $i \geq 1$. Then the word $x_1x_2x_3 \ldots x_m \in (\{0,1\}^n)^+$ belongs to $R(n)$ if and only if $(x_{11}, x_{12})(x_{21}, x_{22}) \ldots (x_{m1}, x_{m2}) \in R$, i.e., a word belongs to $R(n)$ if the word obtained by a restriction to the first two components of any letter belongs to R.

Let $R(n)$ be an extension of $R \subseteq (\{0,1\}^2)^+$, and let F be a sequential function of $\mathcal{F}^n$ which accepts $R(n)$ by Y. Then the function F' which is obtained from F by identification of the the last $n-1$ inputs, belongs to $\mathcal{F}^2$ and accepts R by Y.

Analogously, we define the extension $R(n)$ of a regular set $R \subseteq \{0,1\}^+$. Moreover, for any $R \subseteq \{0,1\}^+$, there is a function $G \in \mathcal{F}^n$ such that G' obtained from G by identification of all inputs accepts R.

Now assume that, $M \subseteq \mathcal{F}$ is a subalgebra of $\mathcal{F}$ such that, for any regular set $R \subseteq (\{0,1\}^n)^+$, there is a sequential function in M which accepts R. Then, for any regular set $R(n)$ which is an extension of a regular set $R \subseteq \{0,1\}^+ \cup (\{0,1\}^2)^+$, there is a sequential function in M which accepts $R(n)$. Thus by the above construction, for any regular set $R \subseteq \{0,1\}^+ \cup (\{0,1\}^2)^+$, M contains a function F accepting R. Since we used only regular sets from $\{0,1\}^+ \cup (\{0,1\}^2)^+$ in the proof of Theorem 4.9, we can prove that M is a σ_K-algebra.

Therefore, for $n \geq 2$, a subset M of $\mathcal{F}$ is σ_K-complete if and only if, for any regular set $R \subseteq (\{0,1\}^n)^+$, $[M]$ contains a sequential function which accepts R. Thus we have the following statement.

For $n \geq 2$, there is no algorithm which decides for a finite set $M \subseteq \mathcal{F}$ whether or not, for any regular set $R \subseteq (\{0,1\}^n)^+$, $[M]$ contains a sequential function which accepts R.

Now let $n = 1$, i.e., we consider only languages which are contained in $\{0,1\}^+$. Obviously, in order to accept such languages we can restrict ourselves to unary sequential functions, i.e., to elements of $\mathcal{F}^1$. Moreover, using the operation we can generate only functions which depend essentially on at most one variable. Hence we here consider the semigroup $\mathcal{F}^1$ instead of $\mathcal{F}$. We mention the following two results.

There is no finite set M such that, for any regular set $R \subseteq \{0,1\}^+$, $[M]$ contains a sequential function accepting R.

This can be seen very easily. Assume that such a finite set M exists. Then we consider the set $M \cup \{F_{(non)}\}$ which is finite, too, and generates $\mathcal{F}^1$. This contradicts the fact that there is no finite system of generators for $\mathcal{F}^1$ [15].

The set of subsemigroups of $\mathcal{F}^1$, which for any regular set $R \subseteq \{0,1\}^+$ contain a sequential function accepting R, has the cardinality of the set of real numbers.

Let $M = M_{T_0} \cap \mathcal{F}^1$. Obviously, M is a subsemigroup of $\mathcal{F}^1$ and, for any regular set $R \subseteq \{0,1\}^+$, M contains a sequential function accepting R. For any $t \in \mathbb{N}$, let F_t be the sequential function of $\mathcal{F}^1$ with

$$\varphi^i_{F_t}(x_1, x_2, \ldots, x_i) = \begin{cases} \mathrm{non}(x_1) & \text{if } i < t, \\ \mathrm{non}(x_t) & \text{if } i \geq t. \end{cases}$$

For any set $I \subseteq \mathbb{N}$, we set

$$M_I = [M \cup \{F_t \mid t \in I\}].$$

By definition, M_I is a subsemigroup of $\mathcal{F}^1$ and, for any regular set $R \subseteq \{0,1\}^+$, M_I contains a sequential function accepting R. We show that $s \notin I$, $I \subset \mathbb{N}$ and $s \neq 1$ implies $F_s \notin M_I$. Assume the contrary. Then

$$F_s = G_r \circ G_{r-1} \circ \cdots \circ G_1$$

for some functions G_j, $1 \leq j \leq r$, with $G_j \in M$ or $G_j = F_t$ for some $t \in I$. If $G_j \in M$ holds for $1 \leq j \leq r$, then $G_j \in M_{T_0}$ for $1 \leq j \leq r$ and thus $F_s \in M_{T_0}$ in contrast to its construction. Let $G_k \notin M$, i.e., $G_k = F_t$ for some $t \in I$. Let

$$G = G_{k-1} \circ G_{k-2} \circ \cdots \circ G_1 \quad \text{and} \quad H = G_r \circ G_{r-1} \circ \cdots \circ G_{k+1}.$$

Then we get

$$\varphi^i_{F_s}(x_1,x_2,\dots,x_i)=\begin{cases}\varphi^i_H(\mathrm{non}(\varphi^1_G(x_1)),\mathrm{non}(\varphi^1_G(x_1)),\dots,\mathrm{non}(\varphi^1_G(x_1))) & \text{if } i<t,\\ \varphi^i_H(\mathrm{non}(\varphi^1_G(x_1)),\dots,\mathrm{non}(\varphi^1_G(x_1)),\\ \qquad \mathrm{non}(\varphi^t_G(x_1,\dots,x_t)),\dots,\mathrm{non}(\varphi^t_G(x_1,\dots,x_t))) & \text{if } i\geq t.\end{cases}$$

It is easy to see that the behaviour given on the right side of the last equation does not describe the behaviour of F_s (e.g., if $s>t$ then F_s changes its function after the s-th moment, whereas this does not hold for the functions given at the right hand side). This contradiction proves that $F_s\notin M_I$.

Let I_1 and I_2 be two subsets of $\mathbb{N}$ not containing 1. Then $I_1\neq I_2$ implies $M_{I_1}\neq M_{I_2}$. Therefore we have at least as many sets M containing a sequential function accepting R for any regular set $R\subseteq\{0,1\}^+$ as we have subsets of $\mathbb{N}\setminus\{1\}$. Hence the statement follows.

We mention that the results under the restriction to regular sets contained in $\{0,1\}^+$ strongly differ from those for arbitrary regular sets (or regular sets contained in $(\{0,1\}^n)^+$ with $n\geq 2$), for instance we have an infinite set of subalgebras accepting all regular sets over $\{0,1\}$ in contrast to only three σ_K-algebras, or any σ_K-algebra is finitely generated whereas there is no finitely generated subsemigroup accepting all regular sets of $\{0,1\}^+$.

We now restrict ourselves to strongly definite languages, i.e., to languages of the form

$$(\{0,1\}^n)^*U(\{0,1\}^n)^r\cup\bigcup_{i\in Z}(\{0,1\}^n)^i\quad\text{with}\quad n\geq 1,\ r\geq 0,\ U\subseteq\{0,1\}^n,\ Z\subseteq\{1,2,\dots,r\}.$$

(for information on definite languages we refer to [12].) We note that a strongly definite language of the above form is accepted by a sequential function F by $Y=\{y\}$ where

$$\varphi^j_F(x_1,x_2,\dots,x_j)=\begin{cases}y & \text{if } j\leq r,\ j\in Z,\\ \mathrm{non}(y) & \text{if } j\leq r,\ j\notin Z,\\ f(x_{j-r}) & \text{if } j>r.\end{cases}$$

This means that the regular set is accepted by a sequential function which corresponds to a Boolean function with delay r (and some fixed outputs at the moments before the delay is effective). Such sequential functions can be described as $(f;r;a_1,a_2,\dots,a_r)$, where r is the delay and a_j is the output at the j-th moment, $1\leq j\leq r$.

Thus it is also natural to consider only such sequential functions. Therefore we have to change the operation of substitution (or superposition) to synchronized substitution, i.e., for any n-ary function $F=(f;r;a_1,a_2\dots,a_r)$ and n functions $G_i=(g_i;s;b_{i1},b_{i2},\dots,b_{is})$ with arity m, we define $F\circ(G_1,G_2,\dots,G_n)$ by

$$F\circ(G_1,G_2,\dots,G_n)(p_1,p_2,\dots,p_m)=F(G_1(p_1,\dots p_m),G_2(p_1,\dots,p_m),\dots,G_n(p_1,\dots,p_m)).$$

Let $\mathcal{F}_d$ be the corresponding algebra, and let $\langle\langle M\rangle\rangle$ be the closure of a set M with respect to $\mathcal{F}_d$. Then we have the following result.

> *There is an algorithm which decides for a finite set $M\subset\mathcal{F}_d$ whether or not, for any strongly definite language R, there is a function in $\langle\langle M\rangle\rangle$ which accepts R.*

We omit the proof and refer to [6, 9].

However, we mention that there is a difference between our concepts and those of [14] besides our restriction to a special equivalence relation. We have to generate all functions for all delays whereas in [14] one only requires that any function can be generated with a certain delay.

If we consider the special case of a function without delay or, equivalently, languages of the form

$$(\{0,1\}^n)^*U \quad \text{with} \quad n \geq 1 \text{ and } U \subseteq \{0,1\}^n, \tag{4.3}$$

we have to consider P_2. As above we can prove that T_0, T_1 and P_2 are the only subalgebras such that, for any regular set R of the form (4.3), the subalgebra contains a function accepting R. Hence, by Theorem 2.14, we can decide for a finite set $M \subseteq P_2$ whether or not, for any regular set R of the form (4.3), $\langle M \rangle$ contains a function accepting R. For corresponding results for P_k, $k \geq 3$, we refer to [7].

Let us turn back to the consideration of equivalence relations and completeness with respect to them.

We give some variations of Kleene-equivalence which can be described as shown in Lemma 4.7 by using $F_{(\text{non})}$ in the beginning or in the beginning and the end instead of using it only in the end.

4.12 Definition

(1) $F \in \mathcal{F}^m$ and $G \in \mathcal{F}^m$ are called *negation-equivalent* (written as $(F,G) \in \sigma_N$) if and only if $F = G$ or

$$F(p_1, p_2, \ldots, p_n) = G(F_{(\text{non})}(p_1), F_{(\text{non})}(p_2), \ldots, F_{(\text{non})}(p_n)).$$

(2) $F \in \mathcal{F}^m$ and $G \in \mathcal{F}^m$ are called *dual* (written as $(F,G) \in \sigma_D$) if and only if $F = G$ or

$$F(p_1, p_2, \ldots, p_n) = F_{(\text{non})}(G(F_{(\text{non})}(p_1), F_{(\text{non})}(p_2), \ldots, F_{(\text{non})}(p_n))).$$

The duality is a well-known notion for Boolean functions (the Boolean functions which are dual to itself form the maximal subalgebra Sd of P_2).

Let us consider σ_N- and σ_D-completeness.

4.13 Theorem

(1) *$M \subset \mathcal{F}$ is σ_N-complete if and only if M is complete.*

(2) *$M \subset \mathcal{F}$ is σ_D-complete if and only if M is complete.*

Proof (1) With a set $M \subseteq \mathcal{F}$ we associate the sets

$$M' = [M] \cap \{F_{(f)} \mid f \in P_2\} \quad \text{and} \quad M'' = \{f \mid f \in P_2,\ F_{(f)} \in M\}.$$

First we note that any equivalence class contains at most two elements. If $F_{(f)}$ is an element of an equivalence class with respect to σ_N, then the other element is $F_{(g)}$ where g is the Boolean function with

$$g(y_1, y_2, \ldots, y_n) = f(\text{non}(y_1), \text{non}(y_2), \ldots, \text{non}(y_n)).$$

(Note that $f = g$ can hold.) Thus we have the following four equivalence classes with respect to σ_N:

$$K_1 = \{F_{(\mathrm{et})}, F_{(\mathrm{non} \circ \mathrm{vel})}\}, \qquad K_2 = \{F_{(\mathrm{eq})}\}, \qquad K_3 = \{F_{(k_0)}\}, \qquad K_4 = \{F_{(k_1)}\}.$$

All functions of K_1 are not contained in neither Lin nor in Sd. The functions of K_2, K_3 and K_4 do not belong to Mon, T_1 and T_0, respectively.

Now let M be a σ_N-complete set. Then M contains at least one element of each of the classes K_1, K_2, K_3 and K_4. Thus M'' is not contained in any of the maximal subalgebras T_0, T_1, Lin, Sd and Mon of P_2. By Theorem 2.14 (1), M'' is complete in P_2. Hence $F_{(\mathrm{non})}$ can be generated by the elements of M. Since M is σ_N-complete, we can generate at least one element of any equivalence class with respect to σ_N. But using $F_{(\mathrm{non})}$, we can generate the other element (if it is different), too. Thus M can generate any sequential function, i.e., M is complete.

Obviously, if M is complete, then it is σ_N-complete, too.

(2) We note that $F_{(\mathrm{non})}$ is dual to itself, i.e., the equivalence class of $F_{(\mathrm{non})}$ only consists of this function. Thus any σ_D-complete set contains $F_{(\mathrm{non})}$. As above we can prove the completeness of any σ_D-complete set. □

From Theorem 4.13 we immediately obtain the following result on the algorithmic aspect.

4.14 Corollary

(1) *There is no algorithm which decides the σ_N-completeness of a finite set.*

(2) *There is no algorithm which decides the σ_D-completeness of a finite set.*

5 Metric completeness

In this section we change the concept of equivalence. We do not consider equivalence of functions, but equivalence of sets of functions. We start with the formal definition.

5.1 Definition For $F \in \mathcal{F}^m$ and $G \in \mathcal{F}^m$, we define

$$d(F,G) = \frac{1}{t} \quad \text{if and only if} \quad \begin{aligned} &F(p) = G(p) \text{ for all } p \text{ with } |p| \leq t-1 \text{ and} \\ &F(p') \neq G(p') \text{ for some } p' \text{ with } |p'| = t \end{aligned}$$

The easy proof of the following lemma is left to the reader.

5.2 Lemma *For any $m \in \mathbb{N}$, d is a distance on $\mathcal{F}^m$.*

5.3 Definition $M \subseteq \mathcal{F}$ and $M' \subseteq \mathcal{F}$ are called *metrically equivalent* if, for any $t \in \mathbb{N}$ and any two sequential functions $F \in M$ and $F' \in M'$, there are sequential functions $G \in M'$ and $G' \in M$ such that

$$d(F,G) \leq \frac{1}{t} \quad \text{and} \quad d(F',G') \leq \frac{1}{t}.$$

Again, it is easy to prove that the following statement holds.

5.4 Lemma *The metric equivalence is an equivalence relation on the powerset of* $\mathcal{F}$.

5.5 Example For $F \in \mathcal{F}$ and $i \in \mathbb{N}$, let F_i be defined by

$$F_i(p) = \begin{cases} F(p) & \text{if } |p| \leq i \\ F(q')0^n & \text{if } |p| > i,\ p = q'q'',\ |q'| = i,\ |q''| = n \end{cases}$$

Further we define

$$Q = \{F_i \mid F \in \mathcal{F},\ i \in \mathbb{N}\}.$$

Obviously, Q is a proper subset of $\mathcal{F}$, and it is easy to see that Q is metrically equivalent to $\mathcal{F}$.

Note that Q forms a subalgebra and that Q is not finitely generated. The last fact can be seen as follows. If M is a finite subset of Q, then there is a number i such that $\varphi_G^j = k_0$ for all $G \in M$and all $j > i$. Then $\varphi_H^j = k_0$ for $j > i$ also holds for any sequential function generated by M. However, this property is not valid for Q.

We now define metric completeness.

5.6 Definition $M \subset \mathcal{F}$ is called *metrically complete* if and only if $[M]$ is metrically equivalent to $\mathcal{F}$.

In the definition of ϱ-completeness of a set M where ϱ is some equivalence relation, we have required that, for any sequential function $F \in \mathcal{F}$, $[M]$ contains a function equivalent to F. In case of metric completeness, we require that, for any sequential function F and any $t \in \mathbb{N}$, $[M]$ contains a sequential function G with $d(G, F) < 1/t$. Thus we replace equivalence by small distance. Thus it is natural to study metric completeness in connection with completeness with respect to equivalence relations. Furthermore, we shall see that metric completeness is nearly related to σ_t-completeness for $t \in \mathbb{N}$.

5.7 Theorem *There is no algorithm which decides whether or not a finite set is metrically complete.*

Proof The proof can be given by small modifications of the proof of Theorem 2.13. □

We now discuss the corresponding concept of maximal subalgebras and its use as a completeness criterion. We start with the definition of metric maximality.

5.8 Definition A subalgebra $M \subset \mathcal{F}$ is called *metrically maximal* if

- M is not metrically complete and
- $M \cup \{F\}$ is metrically complete for any $F \in \mathcal{F}$.

5.9 Theorem

(1) *Any metrically maximal subalgebra of* $\mathcal{F}$ *is* σ_t*-maximal for some* $t \in \mathbb{N}$. *For* $t \in \mathbb{N}$, *any* σ_t*-maximal subalgebra of* $\mathcal{F}$ *is a metrically maximal subalgebra.*

(2) *The set of all metrically maximal subalgebras of* $\mathcal{F}$ *has the same cardinality as* $\mathbb{N}$.

(3) *M is metrically complete if and only if M is not contained in any metrically maximal subalgebra of $\mathcal{F}$.*

Proof (1) Let M be a metrically maximal subalgebra of $\mathcal{F}$. Then there is a function $F \in \mathcal{F}$ and a number $t \in \mathbb{N}$ such that $d(F,G) \geq 1/t$ for any function $G \in M$. Hence M is not σ_t-complete. Thus M is contained in a σ_t-maximal subalgebra N. If $M \subset N$, then there is a sequential function $H \in N \setminus M$. Hence $[M \cup \{H\}] \subseteq N$. Since, obviously, N is not metrically equivalent to $\mathcal{F}$, this contradicts the metric maximality of M. Therefore $M = N$ and the first statement of (1) has been shown.

Now let M be a σ_t-maximal subalgebra for some $t \in \mathbb{N}$. First we note that M contains any function F where

$$\varphi^i_F((x_{11}, x_{12}, \ldots, x_{1n}), (x_{21}, x_{22}, \ldots, x_{2n}), \ldots, (x_{i1}, x_{i2}, \ldots, x_{in})) = x_{ik}$$

for some k and $i \leq t$. If we assume the contrary, then $[M \cup \{F\}]$ is a subalgebra of $\mathcal{F}$ with

$$M \subset [M \cup \{F\}]. \tag{5.1}$$

Moreover, M and $[M \cup \{F\}]$ are σ_t-equivalent since using F we cannot change the behaviour at the first t moments. Hence $[M \cup \{F\}]$ and $\mathcal{F}$ are not σ_t-equivalent. Taking into consideration (5.1) we get a contradiction to the definition of σ_t-maximality.

Thus M contains the functions $F_1 \in \mathcal{F}^2$, $F_2 \in \mathcal{F}^1$ and $F_3 \in \mathcal{F}^2$ with

$$\varphi^i_{F_1}((x_{11}, x_{12}), (x_{21}, x_{22}), \ldots, (x_{i1}, x_{i2})) = \begin{cases} x_{i1} & \text{if } i \leq t, \\ \mathrm{sh}(x_{i1}, x_{i2}) & \text{if } i > t, \end{cases}$$

$$\varphi^i_{F_2}(x_1, x_2, \ldots, x_i) = \begin{cases} x_i & \text{if } i \leq t, \\ x_{i-1} & \text{if } i > t, \end{cases}$$

$$\varphi^i_{F_3}((x_{11}, x_{12}), (x_{21}, x_{22}), \ldots, (x_{i1}, x_{i2})) = \begin{cases} x_{i1} & \text{if } i \leq t, \\ x_{i2} & \text{if } i > t. \end{cases}$$

Now let F be an arbitrary function not in M. Then $[M \cup \{F\}]$ is a subalgebra σ_t-equivalent to $\mathcal{F}$. Therefore, $[M \cup \{F\}]$ contains sequential functions G_1 and G_2 such that

$$\varphi^i_{G_1}((x_{11}, x_{12}), (x_{21}, x_{22}), \ldots, (x_{i1}, x_{i2})) = \mathrm{sh}(x_{i1}, x_{i2}) \quad \text{for} \quad i \leq t,$$

$$\varphi^i_{G_2}(x_1, x_2, \ldots, x_i) = \begin{cases} 0 & \text{if } i = 1, \\ x_{i-1} & \text{if } i \in \{2, 3, \ldots, t\}. \end{cases}$$

Then the functions

$$F_3(G_1(p_1, p_2), F_1(p_1, p_2)) \quad \text{and} \quad F_3(G_2(p), F_2(p))$$

belong to $[M \cup \{F\}]$ and coincide with $F_{(\mathrm{sh})}$ and $F_{(\mathrm{id};1)}$. Therefore, $[M \cup \{F\}] = \mathcal{F}$ which proves that $[M \cup \{F\}]$ is metrically equivalent to $\mathcal{F}$. Hence M is metrically maximal.

(2) follows from (1) and Theorem 3.10.

(3) We prove that M is not metrically complete if and only if is contained in some metrically maximal subalgebra.

Assume that M is not metrically complete. Then there is a function $F \in \mathcal{F}$ and a number $t \in \mathbb{N}$ such that $d(F,G) \geq 1/t$ for any function $G \in [M]$. Thus M is not σ_t-complete. By Theorem 3.9, there is a σ_t-maximal subalgebra N with $M \subseteq N$. By (1) N is metrically maximal, too. Thus M is contained in a metrically maximal subalgebra.

The converse statement is obvious. □

We now combine the notions of metric completeness and σ_K-completeness.

5.10 Definition A set M is called 1-*metrically Kleene-complete*, if $[M]$ is metrically equivalent to a σ_K-algebra.

By this definition we require that M is 1-metrically Kleene-complete if, for any regular set R and any $t \in \mathbb{N}$, $[M]$ contains a function F such that there is a function G which accepts R and satisfies $d(F,G) < 1/t$. It is natural to modify the concept such that the metric does not work on the functions but on the regular sets.

5.11 Definition For $R \in (\{0,1\}^n)^+$ and $Q \in (\{0,1\}^n)^+$, we define

$$d'(R,Q) = \frac{1}{t} \quad \text{if and only if} \quad p \in R \quad \text{if and only if} \quad p \in Q \text{ for all } p \text{ with } |p| \leq t-1 \text{ and}$$
$$p' \in R \cup Q \text{ and } p' \notin R \cap Q \text{ for some } p' \text{with } |p'| = t$$

5.12 Lemma *For any $n \in \mathbb{N}$, d' is a metric on the powerset of $(\{0,1\}^n)^+$.*

5.13 Definition M is called 2-*metrically Kleene-complete* iff, for any regular set R and any integer $t \in \mathbb{N}$, there is a sequential function F in $[M]$ which accepts a set Q with $d(R,Q) \leq 1/t$.

The following theorem relates the two notions of metric Kleene-completeness to each other.

5.14 Theorem *M is 1-metrically Kleene-complete if and only if M is 2-metrically Kleene-complete.*

Proof Let M be 1-metrically Kleene-complete. Let R be an arbitrary regular set and t an arbitrary element of $\mathbb{N}$. Since M is 1-metrically complete, $[M]$ contains a sequential function F such that there is sequential function G which accepts R by Y for some $Y \subset \{0,1\}$ and satisfies $d(F,G) < 1/t$. Let Q be the set accepted by F by Y. Obviously, $d'(R,Q) < 1/t$. Thus, for any R and any t, there is a sequential function $F \in [M]$ which accepts a regular set Q such that $d'(R,Q) < 1/t$, i.e. M is 2-metrically Kleene-complete.

In a similar way one proves the converse statement. □

5.15 Theorem

(1) *There is no algorithm which decides the 1-metric Kleene-completeness of a finite set.*

(2) *There is no algorithm which decides the 2-metric Kleene-completeness of a finite set.*

Proof (1) It is easy to see that 1-metrically Kleene-complete set M is metrically complete if and only if M contains sequential functions F and G with $F \notin M_{T_0}$ and $G \notin M_{T_1}$. Thus the decidability of 1-metric Kleene-completeness would imply the decidability of Kleene-completeness in contrast to Theorem 4.10.

(2) follows from (1) and Theorem 5.14. □

References

[1] V. A. Buevic, Construction of a universal bounded-determined function with two input variables, *Probl. kibernetiki* **15** (1965).

[2] V. A. Buevic, On the algorithmic undecidability of the A-completeness of bounded-determined mappings, *Mat. Zametki* **6** (1972), 687–697 (in Russian).

[3] V. A. Buevic, On the τ-completeness in the class of automaton mappings, *Dokl. Akad. Nauk* **252** (1980), 1037–1041 (in Russian).

[4] G. N. Blochina, W. B. Kudrjavcev, and G. Burosch, Ein Vollständigkeitskriterium bis auf eine gewise Äquivalenzrelation für eine verallgemeinerte Postsche Algebra, *Publicationes Math.* **20** (1973), 141–152.

[5] J. Dassow, Kleene-Mengen und Kleene-Vollständigkeit. *EIK* **10** (1974), 287–295.

[6] J. Dassow, Kleene-Vollständigkeit bei stark-definiten Ereignissen. *EIK* **10** (1974), 399–405.

[7] J. Dassow, Kleene-Mengen und trennende Mengen, *Math. Nachr.* **74** (1976), 89–97.

[8] J. Dassow, Some remarks on the algebra of automaton mappings, in: *Proc. Symp. Fundamentals of Computation Theory* (M. Karpinski, ed.), Lecture Notes in Comput. Sci. **56**, Springer, 1977, 78–83.

[9] J. Dassow, Ein modifizierter Vollständigkeitsbegriff in einer Algebra von Automatenabbildungen, Habilitationsschrift, Universität Rostock, 1978.

[10] J. Dassow, On the congruence lattice of algebras of automaton mappings, in: *Proc. Finite Algebras and Multiple-Valued Logic* (B. Csakany, ed.), Colloq. Math. Soc. Janos Bolyai **28**, North-Holland, 1979, 161–182.

[11] J. Dassow, *Completeness Problems in the Structural Theory of Automata*, Akademie-Verlag, Berlin, 1981.

[12] F. Gecseg and I. Peak, *Algebraic Theory of Automata*, Budapest, 1972.

[13] V. V. Gorlov, On congruence relations on closed Post classes, *Mat. Zametki* **13** (1973), 725–736 (in Russian).

[14] T. Hikita and I. G. Rosenberg, Completeness of uniformly delayed operations, in: *Structural Theory of Automata, Semigroups, and Universal Algebra*, NATO Adv. Study Inst. Ser. C, Kluwer, Dordrecht, 2005, 109–147.

[15] J. Hořejš, Durch endliche Automaten definierte Abbildungen, *Probleme der Kybernetik* **6** (1977), 285–297.

[16] S. V. Jablonskij, G. P. Gawrilow, and W. B. Kudrjawzew, *Boolesche Funktionen und Postsche Klassen*, Berlin, 1970.

[17] C. C. Kleene, *Introduction to Metemathematics*, Amsterdam, 1967.

[18] N. E. Kobrinski and B. A. Trachtenbrot, *Einführung in die Theorie der endlichen Automaten*, Berlin, 1967.

[19] M. I. Kratko, The algorithmic undecidability of the completeness problem for finite automata, *Dokl. Akad. Nauk* **155** (1964), 35–37 (in Russian).

[20] V. B. Kudryavtsev, On the cardinality of the set of precomplete sets of some functional systems related to automata, *Probl. kibernetiki* **13** (1965), 45–74 (in Russian).

[21] V. B. Kudryavtsev, S. V. Alesin, and A. S. Podkolsin, *Elements of the Theory of Automata*, Moscow, 1978 (in Russian).

[22] D. Lau, Kongruenzen auf gewissen Teilklassen von $P_{k,l}$, *Rostock. Math. Colloq.* **3** (1977), 37–44.

[23] A. A. Letichevsky, Criteria for the completeness of a class of Moore automata. *Materiali nauchnych seminarov po teoret. i prikladn. voprosam kibernetiki* **2** (1963), 1–39 (in Russian).

[24] M. Steinby, Algebraic classifications of regular tree languages, in: *Structural Theory of Automata, Semigroups, and Universal Algebra*, NATO Adv. Study Inst. Ser. C, Kluwer, Dordrecht, 2005, 381–432.

[25] B. Thalheim, Über ein Vollständigkeitskriterium für eine Klasse von Automaten (einstellige Approximationsautomaten), *Rostock. Math. Kolloq.* **18** (1981), 99–111.

Completeness of uniformly delayed operations

Teruo HIKITA

Department of Computer Science
Meiji University
1-1-1 Higashimita
Tama-ku, Kawasaki 214-8571
Japan

Ivo G. ROSENBERG*

Département de mathématiques et de statistique
Université de Montréal
C.P. 6128, Succursale Centre-ville
Montréal, Qué., H3C 3J7
Canada

Abstract

The paper reports on progress toward an effective completeness criterion for uniformly delayed multiple-valued combinatorial circuits. In view of previous work by Hikita and Nozaki, and Hikita it suffices to study periodic closed spectra. The main tool is the use of polyrelations and certain constructions on polyrelations developed by Hikita and Miličić. We were able to restrict the search to unary polyrelations and three types of binary polyrelations $\rho = (\rho_0, \rho_1, \dots)$:

(1) period 2^m, ρ_0 bounded order, $\rho_{2^{m-1}}$ its converse and $\rho_i = \iota_2 := \{(a, a) \mid a \in \mathbf{k}\}$ otherwise,

(2) every nonempty $\rho_i = \{(a, s_i(a)) \mid a \in \mathbf{k}\}$ where s_i is a permutation of $\mathbf{k}$; the permutations are interrelated,

(3) components are either (i) all equivalence relations on $\mathbf{k}$ or (ii) all central or equal to $\mathbf{k}^2$. In both cases they have strong properties in terms of intersecting cliques.

1 Introduction

This paper reports on progress toward an effective general completeness criterion for uniformly delayed k-valued combinatorial circuits (this and other rather technical concepts are fully explained in Section 2). In a historical retrospective, the topic was introduced rather early by Kudryavtsev in 1960 who defined the various basic concepts and gave an effective completeness criterion for uniformly delayed binary circuits based on precomplete

*Research partially supported by CRSNG Canada grants DGP 5407 and by JSPS Japan fellowship programs 1991 and 2003.

V. B. Kudryavtsev and I. G. Rosenberg (eds.),
Structural Theory of Automata, Semigroups, and Universal Algebra, 109–147.

classes [Ku60, Ku62]. For ordinary non-delayed circuits and logic such a criterion was given by Post [Po21] for $k = 2$, Iablonskiĭ [Ia58] for $k = 3$, and the second author for $k > 3$ [Ro65, Ro70, Ro77]. Some of Kudryavtsev's results were rediscovered by Loomis [Lo65]. Other completeness aspects for delayed circuits were studied by Biryukova and Kudryavtsev [BK70]. After this early Russian start the focus moved to Japan where Nozaki and his school took up and expanded the study of multiple-valued delayed circuits, to the extent that during the past 33 years all papers in this domain (with the lone exception of [MRR78]) were published by the Japanese school. For uniformly delayed circuits the breakthrough came in Hikita and Nozaki's 1977 paper [HN77] which reduced the problem to three more manageable types. The first case (type A) is directly solved by the primality criterion [Ro65, Ro70] while the third case (type C) was solved by Hikita in 1976 [Hi81b]. Meanwhile Hikita also completely classified the ternary case [Hi78]. The relational theory for uniformly delayed circuits was partly given by Hikita [Hi81a] and fully by Miličić [Mi84, Mi88]. It is based on infinite sequences $\rho = (\rho_0, \rho_1, \dots)$ of relations of the same arity on the alphabet $\mathbf{k} := \{0, 1, \dots, k-1\}$. For such a sequence, called a *polyrelation*, an n-ary operation f with nonnegative integer delay δ carries ρ_i^n into $\rho_{i+\delta}$ for all $i \geq 0$. This concept replaces the preservation of a single relation which is the basic concept in the nondelayed case.

We report on the results for the remaining case of periodic spectra (type B) and the corresponding periodic polyrelations. The precomplete classes obtained are rather exceptional as witnessed by the fact that they are determined by at most binary polyrelations. We have succeeded in limiting them to unary periodic polyrelations and to binary polyrelations $\rho = (\rho_0, \rho_1, \dots)$ of period p of the following three types:

(1) $p = 2^m$ $(m > 0)$, ρ_0 is a bounded partial order $\leq$ on $\mathbf{k}$, $\rho_{2^{m-1}}$ is $\geq$ and $\rho_i = \iota_2 := \{(a, a) \mid a \in \mathbf{k}\}$ for $0 < i < 2^m$, $i \neq 2^{m-1}$. Each of these polyrelations gives a precomplete class [Ku60, Ku62].

(2) Every nonempty component is of the form $\{(a, s(a)) \mid a \in \mathbf{k}\}$ where s is a permutation of $\mathbf{k}$. The permutations involved are intimately linked and the case is essentially one of a group-theoretical nature.

(3) All components $\rho_0, \dots, \rho_{p-1}$ are either (i) equivalence relations $\neq \iota_2$ or (ii) all are central or equal to $\mathbf{k}^2$ (a binary symmetric relation σ is *central* if $\iota_2 \subset \sigma \subset \mathbf{k}^2$ and $c \times \mathbf{k} \subseteq \sigma$ for some $c \in \mathbf{k}$). In both cases we have strong properties in terms of intersecting cliques.

Although the uniformity and the completeness concepts are open to discussion as to their practicality and relation to reality, the authors feel that this study is justified as a first step in this direction, and perhaps even more by the richness of the mathematical theory involved.

The financial support provided to the second author in 1990 by S.T.A. Japan Research Award, by NSERC Canada operating grant A-9128, and by FCAR Québec Subvention d'équipe Eq-0539, and the 1991 and 2003 J.S.P.S., Japan, grants as well as the hospitality of Université de Montréal, Meiji University (Tokyo), Electrotechnical Laboratory (Tsukuba) and Tsukuba College of Technology is gratefully acknowledged.

2 Preliminaries

2.1 Switching circuits are built from basic hardware components which we shall henceforth call *gates*. For simplicity each gate (Fig. 1) is a device with a single output and n inputs

where n is a positive integer. The gate receives and emits signals in the same finite alphabet which will be identified with $\mathbf{k} := \{0, 1, \dots, k-1\}$. If the signal on the i-th input is x_i $(i = 1, \dots, n)$ then the response of the gate is a unique output signal determined by the n-tuple $(x_1, \dots, x_n) \in \mathbf{k}^n$. Denoting this signal by $x_0 = f(x_1, \dots, x_n)$ we can describe the functioning of a gate by an n-ary operation f on $\mathbf{k}$ (i.e., a map $f : \mathbf{k}^n \to \mathbf{k}$). Thus each gate carries an operation f describing its behavior (also called a logic or switching function or a connective). Denote by $\mathcal{O}^{(n)}$ the set of all n-ary operations on $\mathbf{k}$ and set $\mathcal{O} := \bigcup_{n=1}^{\infty} \mathcal{O}^{(n)}$.

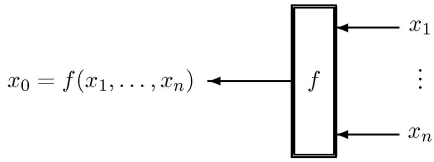


Figure 1

In reality, the physical time dependent signal $x_i(t)$ on the i-th input $(1 \le i \le n)$ and the output signal $x_0(t)$ are continuous functions of time t. The real situation may be rather complex and so we approximate it by assuming that there are time invariant delays $\delta_1, \dots, \delta_n$ such that for all $t \ge \max(\delta_1, \dots, \delta_n)$

$$x_0(t) = f(x_1(t - \delta_1), \dots, x_n(t - \delta_n)) \tag{2.1}$$

where $x_i(t)$ are maps from $[0, \infty)$ into $\mathbf{k}$ $(i = 1, \dots, n)$. Thus (2.1) means that the present output depends on the value of the i-th input δ_i time units ago $(i = 1, \dots, n)$. Such gates are called *delayed input devices* (or d-modules). For simplicity we assume that all δ_i belong to the set $\mathbb{N} = \{0, 1, \dots\}$ of nonnegative integers. The integrality assumption follows the connections to automata theory; however the delays of real gates are at best known approximately and, moreover, may depend on extremal conditions, e.g. the temperature, or on the age of the circuit. In this paper we go even further and assume that $\delta_1 = \cdots = \delta_n$. This restriction, common to most engineering literature, is far reaching, substantially limiting the composition and causing the theory to be nonalgebraic. Such a gate is called a *uniformly delayed device* (or k-module), and it is fully described by the pair (f, δ) called an operation (or function) *with delay* δ on $\mathbf{k}$. The set of all n-ary operations with delay on $\mathbf{k}$ is $\mathcal{U}^{(n)} := \mathcal{O}^{(n)} \times \mathbb{N}$, and $\mathcal{U} := \mathcal{O} \times \mathbb{N}$ is the set of all operations with delay on $\mathbf{k}$.

2.2 Switching circuits are obtained from a collection of gates by attaching outputs of certain gates to inputs of other gates. Again for simplicity we consider only the combinatorial (i.e. feedback-free) switching circuits. The simplest case is the following. We have a gate F described by $(f, \delta) \in \mathcal{U}^{(n)}$ and n gates G_i determined by $(g_i, \delta_i) \in \mathcal{U}^{(m_i)}$, $i = 1, \dots, n$. If we attach the single output of G_i to the i-th input of F, $i = 1, \dots, n$, the resulting tree-like circuit has $m := m_1 + \cdots + m_n$ external inputs and realizes the operation $h = f \otimes (g_1, \dots, g_n) \in \mathcal{O}^{(m)}$ defined by setting

$$h(x_1, \dots, x_m) \approx f(g_1(x_1, \dots, x_{m_1}), \dots, g_n(x_{m-m_n+1}, \dots, x_m))$$

where $\approx$ means that the equation holds for all $x_1, \dots, x_m \in \mathbf{k}$. The accumulated delays are $(\delta+\delta_1, \dots, \delta+\delta_1, \dots, \delta+\delta_n, \dots, \delta+\delta_n)$. It follows that the circuit will have a uniform delay if and only if $\delta_1 = \cdots = \delta_n$. In this paper we restrict ourselves to uniformly delayed circuits and so we only accept the $\otimes$-composition $(f \otimes (g_1, \dots, g_n), \delta+\delta')$ with $F := (f, \delta) \in \mathcal{U}^{(n)}$ and $G_i := (g_i, \delta') \in \mathcal{U}^{(m_i)}$ $(i = 1, \dots, n)$. Denote the resulting circuit by $F \otimes (G_1, \dots, G_n)$.

Suppose we have a circuit in the shape of a rooted tree with gates at the vertices distinct from the leaves and external inputs (not necessarily pairwise distinct) at the leaves and let the sum of the delays be constant on each branch from a leaf to the root. Working from the leaves to the root we can express the delayed function represented by the tree through repeated composition (e.g. in the situation of Fig. 2 the function is $(f \otimes (g \otimes (h, i), j), 4)$. Thus the $\otimes$-composition suffices for the description of functions associated with acyclic circuits yielding uniform delays.

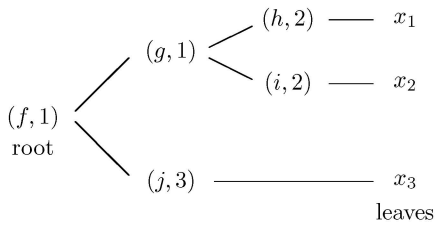


Figure 2

It should be stressed that the $\otimes$-composition can be performed if and only if the inside functions have an identical delay. This restriction differentiates our structure from universal algebra and propositional calculus of most logics which allow unrestricted composition. The structure may be described by a suitable partial algebra, but as this fact seems to have little impact on completeness, we should not dwell on it.

2.3 In what follows we need the *projection* (trivial operation or selector). This is an n-ary operation e_i^n on $\mathbf{k}$ such that $e_i^n(x_1, \dots, x_n) \approx x_i$ $(1 \le i \le n)$. Denote by $J := \{e_i^n \mid 1 \le i \le n < \aleph_0\}$ the set of all projections. For a subset V of $\mathcal{U}$ define $\langle\langle V \rangle\rangle$ as the least subset of $\mathcal{U}$ containing $F \otimes (G_1, \dots, G_n)$ whenever $F \in \langle\langle V \rangle\rangle$ and $G_i \in V \cup (J \times \{0\})$ $(i = 1, \dots, n)$ (it being understood that $G_1, \dots, G_n$ have the same delay). We have added $J \times \{0\}$ to allow arbitrary changes of variables (i.e. for $F \in \langle\langle V \rangle\rangle$ the set $\langle\langle V \rangle\rangle$ also contains each F' obtained from F by permuting (i.e. exchanging or switching) or identifying (fusing) the variables). Clearly $V \to \langle\langle V \rangle\rangle$ is a closure operator on $\mathcal{U}$. The subsets V of $\mathcal{U}$ satisfying $V = \langle\langle V \rangle\rangle$ are called *closed uniform sets*. A closed uniform set containing $J \times \{0\}$ is a *uniform clone*. We say that $V \subseteq \mathcal{U}$ is *complete* if to every $f \in \mathcal{O}$ there is $\delta \in \mathbb{N}$ such that $(f, \delta) \in \langle\langle V \rangle\rangle$. This was introduced for $k = 2$ by Kudryavtsev [Ku60, Ku62] (as *completeness in the second sense*) and captures the possibility of constructing each operation with some—possibly very large—delay. The object of this paper is to give a universal completeness criterion. Before embarking into the technical details we comment on the relation between our model and

reality.

As pointed out in 2.1 the input delayed gate is already a considerable simplification. The delays in circuits should be taken into consideration since to ignore them is tantamount to neglecting such well known phenomena as races or hazards, and therefore input delayed devices constitute the first step in the right direction. The restriction to uniformly delayed devices is all pervasive through the literature but it is not altogether clear whether it is motivated by mere convenience or rather a hard fact about today's commercially available gates. In practice, often several functions have to be represented simultaneously which could be used as an argument for uniform delays. The completeness concept is open to the obvious criticism: what is the purpose of constructing an F with an enormous delay? We chose our concept because it is the simplest and weakest completeness concept, and since it makes a nice mathematical theory.

2.4 We conclude with two minor points. Suppose we have constructed (f, δ) and it happens that f is constant and hence time-independent. Of course, there is no observable delay and we can assume that we have all (f, δ') with $\delta' \in \mathbb{N}$. This is rather academic because usually sources of constant signal are easy to get and so cheap that they can be taken for granted. However this fact is not accounted for in our model.

Finally, we stress that we are interested in sets of gates with the potential to represent any $f \in \mathcal{O}$ (with some delay and assuming an unlimited supply of each type of gate) but we ignore completely the inherent optimality issue: if f can be represented, what is the cheapest way of representing it? There is a good reason for this limitation; it is because the problem is notoriously hard, depends on present technology and labor costs and thus, to be meaningful, should be closely tailored to a very specific situation which could become obsolete within a very short time.

2.5 We conclude this section with a completeness criterion. First call $P \subseteq \mathcal{O}$ *primal* if every $f \in \mathcal{O}$ is a composition of operations from $P \cup J$ (we reserve this term for operations without delays). Put $e := e_1^1$ (i.e. $e(x) \approx x$) and for $V \subseteq \mathcal{U}$, $n > 0$, and $\delta \geq 0$, set

$$V^{(n)} := V \cap \mathcal{U}^{(n)},\ V_\delta := \{f \mid (f, \delta) \in V\},$$
$$V_{(\delta)} := \bigcup_{m=0}^{\infty} \{f \in \mathcal{O} \mid (f, m\delta) \in V \text{ for some } m \geq 0\}.$$

We have ([Ku62, Theorem 4] for $k = 2$ and [PK79, 7.3.5]; see also [MRR78, Theorem 10] for $k > 2$):

2.6 Proposition *A closed subset V of $\mathcal{U}$ is complete if and only if $e \in V_\delta$ and $V_{(\delta)}$ is primal for some $\delta \geq 0$.*

2.7 Corollary *Let V be a closed uniform set. If V is incomplete, then $F := (J \times \{0\}) \cup V$ is a uniform incomplete clone.*

Proof Clearly F is a uniform clone. Suppose to the contrary that F is complete. Then in Proposition 2.6 we have $\delta = 0$ because otherwise V would be complete. Thus F_0 is primal. Then V_0 is primal [Ro70, 3.1.3] and since V is a closed uniform set, $V_0 = \mathcal{O}$ also. This contradiction proves the statement. □

Needless to say, Proposition 2.6 hardly solves the completeness problem and thus we search for a better criterion. This will be based on sequences of relations introduced and elaborated in the next section.

2.8 A short notational remark. The symbol $\subset$ stands for strict inclusion, while $\subseteq$ means inclusion or equality. Whenever possible an n-tuple is written $x_1 \dots x_n$ instead of the more conventional $(x_1, \dots, x_n)$. The same applies to arguments of maps, functions and operations, e.g. we write fx or $fx_1 \dots x_n$ instead of $f(x)$ or $f(x_1, \dots, x_n)$. Sometimes we do not distinguish notationally an element a and the singleton $\{a\}$ writing e.g. $A \setminus a$ and $a \times A$ for $A \setminus \{a\}$ and $\{a\} \times A$.

3 Polyrelations

3.1 A subset ρ of $\mathbf{k}^h$ (i.e. a set of h-tuples over $\mathbf{k}$) is an h-ary *relation* on $\mathbf{k}$. Let ρ and σ be h-ary relations on $\mathbf{k}$ and $f \in \mathcal{O}^{(n)}$. We say that f *carries* ρ *in* σ if for every $h \times n$ matrix $A = [a_{ij}]$ over $\mathbf{k}$, whose column vectors all belong to ρ, the values of f on the rows of A form an h-tuple from σ; in symbols

$$(a_{1j}, \dots, a_{hj}) \in \rho \ (j = 1, \dots, n) \implies (f(a_{11}, \dots, a_{1n}), \dots, f(a_{h1}, \dots, a_{hn})) \in \sigma.$$

Denote by $\mathrm{Pol}(\rho, \sigma)$ the set of all $f \in \mathcal{O}$ carrying ρ into σ. We set $\mathrm{Pol}\,\rho = \mathrm{Pol}(\rho, \rho)$ and say that f *preserves* ρ if $f \in \mathrm{Pol}\,\rho$ (many other names are used in the literature; e.g., f is compatible, substitutive, homeomorphic and ρ is invariant or stable for f). In universal algebra terms "f preserves ρ" means that ρ is a subuniverse of $\langle \mathbf{k}; f \rangle^h$. We recall some basic results on clones. An h-ary relation ρ on $\mathbf{k}$ has *no systematically repeated coordinate* if for all $1 \le i < j \le h$ there exists $(a_1, \dots, a_h) \in \rho$ with $a_i \ne a_j$. A clone C on $\mathbf{k}$ is *rational* if it is of the form $\mathrm{Pol}\,\rho$ for some relation ρ on $\mathbf{k}$. Here without loss of generality we can assume that ρ has no systematically repeated coordinate.

We recall a relational construction. Let $h \ge 1$ and $\ell \ge p \ge 1$ be integers and let Γ and ρ be h-ary relations on $L = \{1, \dots, \ell\}$ and $\mathbf{k}$, respectively. Set

$$\Gamma \hookrightarrow_p \rho = \{(\varphi(1), \dots, \varphi(p)) \mid \varphi \in \mathrm{Hom}(\Gamma, \rho)\}$$

where $\mathrm{Hom}(\Gamma, \rho)$ denotes the set of relational homomorphisms from Γ into ρ, i.e. maps $\varphi : L \to \mathbf{k}$ such that $\varphi(a) = (\varphi(a_1), \dots, \varphi(a_h)) \in \rho$ whenever $a = (a_1, \dots, a_h) \in \Gamma$. An easy consequence of the results of [BKKR69] is: Let ρ and σ be h-ary and p-ary relations on $\mathbf{k}$ and let σ be without systematically repeated coordinates. Then $\mathrm{Pol}\,\rho \subseteq \mathrm{Pol}\,\sigma$ if and only if there exists $\ell \ge p$ and h-ary relation Γ on $\{1, \dots, \ell\}$ such that $\sigma = \Gamma \hookrightarrow_p \rho$. Another result from [BKKR69] is also basic. To every irrational clone (i.e., nonrational) C there exist relations $\rho_1, \rho_2, \dots$ on $\mathbf{k}$ such that $\mathrm{Pol}\,\rho_1 \supseteq \mathrm{Pol}\,\rho_2 \supseteq \dots \supseteq C = \bigcap_{i=1}^{\infty} \mathrm{Pol}\,\rho_i$. We extend this to uniform clones.

3.2 A countably infinite sequence $\rho := (\rho_0, \rho_1, \dots)$ of h-ary relations on $\mathbf{k}$ is an h-ary *polyrelation* on $\mathbf{k}$. The set of h-ary polyrelations on $\mathbf{k}$ is denoted by $\mathcal{R}_h$ and $\mathcal{R} := \bigcup_{h=1}^{\infty} \mathcal{R}_h$. We say that $(f, \delta) \in \mathcal{U}$ *preserves* an h-ary polyrelation $\rho = (\rho_0, \rho_1, \dots)$ if

$$f \in \mathrm{Pol}_\delta\, \rho := \bigcap_{i \in \mathbb{N}} \mathrm{Pol}(\rho_i, \rho_{i+\delta}).$$

We set

$$\operatorname{Pold}\rho := \bigcup_{\delta=0}^{\infty}(\operatorname{Pol}_\delta\rho\times\delta).$$

3.3 Example Let $\leq$ be an order (a reflexive, transitive and antisymmetric binary relation) on $\mathbf{k}$. Then $M := \operatorname{Pol}\leq$ is the set of $\leq$-*monotone* (also called *isotone* or *order preserving*) operations on $\mathbf{k}$ (i.e. $f\in\mathcal{O}^{(n)}$ such that $fx_1\cdots x_n\leq fy_1\cdots y_n$ whenever $x_1\leq y_1,\dots,x_n\leq y_n$). Similarly $f\in\mathcal{O}^{(n)}$ is $\leq$-*antimonotone* if $fx_1\cdots x_n\geq fy_1\cdots y_n$ whenever $x_1\leq y_1,\dots,x_n\leq y_n$. Let A be the set of $\leq$-antimonotone operations and let $\rho=(\leq,\geq,\leq,\geq,\dots)$. Then $\operatorname{Pol}_{2n}\rho=M$ and $\operatorname{Pol}_{2n+1}\rho=A$ for all $n\geq 0$.

We have [Ku62]:

3.4 Lemma *Let $\rho=(\rho_0,\rho_1,\dots)$ be a polyrelation. Then $\operatorname{Pol}_0\rho=\bigcap_{i\geq0}\operatorname{Pol}\rho_i$ and $\operatorname{Pold}\rho$ is a uniform clone.*

Proof By definition,

$$\operatorname{Pol}_0\rho=\bigcap_{i\in\mathbb{N}}\operatorname{Pol}\rho_i,$$

and therefore $J\times0\subseteq\operatorname{Pold}\rho$. It remains to prove that $\operatorname{Pold}\rho$ is a closed set. Let $f\in\operatorname{Pol}_\delta\rho$ be n-ary, let $g_j\in\operatorname{Pol}_{\delta'}\rho$ $(j=1,\dots,n)$ and let $h:=f\otimes(g_1,\dots,g_n)$ (as defined in 2.2). For $i\in\mathbb{N}$ we have $g_j[\rho_i]\subseteq\rho_{i+\delta'}$ $(j=1,\dots,n)$. Using $f[\rho_{i+\delta'}]\subseteq\rho_{i+\delta+\delta'}$ we obtain $h[\rho_i]\subseteq\rho_{i+\delta+\delta'}$. □

3.5 We recall a basic result from [Mi84]. The relation "(f,δ) preserves a polyrelation" gives rise to a Galois connection between the sets $\mathcal{U}$ (of uniformly delayed operations on $\mathbf{k}$) and $\mathcal{R}$ (of polyrelations on $\mathbf{k}$). The Galois closed subsets of $\mathcal{U}$ are the sets of the form $\operatorname{Pold}X=\bigcap_{\rho\in X}\operatorname{Pold}\rho$ with $X\subseteq\mathcal{R}$. It is shown in [Mi84] that they are exactly the uniformly closed clones; more precisely each uniform clone is either of the form $\operatorname{Pold}\rho$ from some $\rho\in\mathcal{R}$ or the intersection of a countable descending chain $\operatorname{Pold}\rho_0\supset\operatorname{Pold}\rho_1\supset\cdots$ with $\rho_i\in\mathcal{R}$ for all $i\geq0$. It is therefore of interest to characterize the Galois-closed subsets of $\mathcal{R}$. By definition they are the sets

$$\operatorname{Inv}Y:=\{\rho\in\mathcal{R}\mid\text{every }(f,\delta)\in Y\text{ preserves }\rho\}$$

with $Y\subseteq\mathcal{U}$. To describe them internally we need the following notions. Let σ be an h-ary relation on $\mathbf{k}$. If $h>1$ set

$$\begin{aligned}\zeta(\sigma)&:=\{(a_2,\dots,a_h,a_1)\mid(a_1,\dots,a_h)\in\sigma\},\\ \tau(\sigma)&:=\{(a_2,a_1,a_3,\dots,a_h)\mid(a_1,\dots,a_h)\in\sigma\},\\ \operatorname{pr}(\sigma)&:=\{(a_1,\dots,a_{h-1})\mid(a_1,\dots,a_h)\in\sigma\},\end{aligned}$$

while for $h=1$ set $\zeta(\sigma)=\tau(\sigma)=\operatorname{pr}(\sigma)=\sigma$. Further set

$$\nabla(\sigma)=\{(a_1,\dots,a_{h+1})\mid(a_1,\dots,a_h)\in\sigma\}.$$

Extend ζ, τ, pr, ∇ to polyrelations by applying them termwise. For $\rho=(\rho_0,\rho_1,\dots)$ and $\alpha\in\{\zeta,\tau,\operatorname{pr},\nabla\}$ set $\alpha(\rho)=(\alpha(\rho_0),\alpha(\rho_1),\dots)$. Further set $\operatorname{sh}(\rho)=(\rho_1,\rho_2,\dots)$. For polyrelations

$\rho = (\rho_0, \rho_1, \dots)$ and $\sigma = (\sigma_0, \sigma_1, \dots)$ set $\rho \cap \sigma = (\rho_0 \cap \sigma_0, \rho_1 \cap \sigma_1, \dots)$ (notice that $\rho \cap \sigma = (\emptyset, \emptyset, \dots)$ if ρ and σ have different arities).

We also need a partial infinitary operation on $\mathcal{R}$ called upper superposition. For $\rho^i = (\rho_0^i, \rho_1^i, \dots) \in \mathcal{R}_h$ $(i = 0, 1, \dots)$ the upper superposition of $(\rho^0, \rho^1, \dots)$ is defined wherever, for all $i \geq 0$,

$$\rho_0^i \supseteq \rho_1^{i-1} \cup \rho_2^{i-2} \cup \dots \cup \rho_i^0.$$

If this condition is met the upper superposition is the polyrelation $(\rho_0^0, \rho_0^1, \rho_0^2, \dots)$. In other words, consider in the infinite table

$$\begin{array}{cccccc} \rho_0^0 & \rho_1^0 & \rho_2^0 & \rho_3^0 & \cdots \\ \rho_0^1 & \rho_1^1 & \rho_2^1 & \rho_3^1 & \cdots \\ \rho_0^2 & \rho_1^2 & \rho_2^2 & \rho_3^2 & \cdots \\ \rho_0^3 & \rho_1^3 & \rho_2^3 & \rho_3^3 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \end{array}$$

whose rows are the given polyrelations. We take the polyrelation in the first column provided each of its entries contains all relations on the south-west to north-east diagonal starting at it.

It is shown in [Mi84] (in a slightly different formulation) that a subset X of $\mathcal{R}$ is Galois closed if and only if X is closed under ζ, τ, pr, ∇, sh, $\cap$, upper superposition and contains the constant polyrelation $\omega = (\iota_2, \iota_2, \dots)$ (where $\iota_2 = \{(x, x) \mid x \in \mathbf{k}\}$).

3.6 For a positive integer h denote by $\mathbf{E}_h$ the set of equivalence relations on $\{1, \dots, h\}$. For $\varepsilon \in \mathbf{E}_h$ put

$$\Delta_\varepsilon = \{a_1 \cdots a_h \in \mathbf{k}^h \mid (i, j) \in \varepsilon \Rightarrow a_i = a_j\};$$

i.e. Δ_ε consists of all h-tuples over $\mathbf{k}$ constant on each block of ε.

If ε has a unique nonsingleton block $\{i_1, \dots, i_q\}$ we denote Δ_ε by $\Delta_{i_1 \cdots i_q}$; e.g., Δ_{12} stands for Δ_ε where $\varepsilon = \{12, 21\} \cup \iota_2$. Similarly $\Delta_{i_1 \cdots i_b - j_1 \cdots j_c}$ denotes Δ_ε with ε having exactly two nonsingleton blocks $\{i_1, \dots, i_b\}$ and $\{j_1, \dots, j_c\}$. The relations Δ_ε are termed *diagonal*. The diagonal relations and the empty relation $\emptyset$ are called *trivial*. A nonempty subset E of $\mathbf{E}_h$ is a *clutter* (also *antichain* or *Sperner subset*) if $\theta \subset \tau$ for no $\theta, \tau \in E$. For a clutter E put $\Delta_E := \bigcup_{\varepsilon \in E} \Delta_\varepsilon$.

It is well known [Ro70, 2.8.7] that $\operatorname{Pol} \sigma = \mathcal{O}$ if and only if σ is trivial. We say that a polyrelation ρ is *proper* if $\operatorname{Pold} \rho$ is incomplete. We characterize improper polyrelations.

3.7 Proposition *A polyrelation $\rho = (\rho_0, \rho_1, \dots)$ is improper if and only if*

(1) *all ρ_i are trivial, or*

(2) *there are $\delta > 0$, $m > 0$ and trivial relations $\alpha_0, \dots, \alpha_{\delta-1}$ such that*

$$\rho_i \subseteq \rho_{i+\delta} \subseteq \dots \subseteq \rho_{i+m\delta} = \rho_{i+(m+1)\delta} = \dots = \alpha_i$$

for all $i = 0, \dots, \delta - 1$.

Proof *Necessity.* For $\delta \geq 0$ put $F_\delta := \mathrm{Pol}_\delta \rho$. First consider the case $F_0 = \mathcal{O}$. By Lemma 3.4 then $\mathrm{Pol}\,\rho_i = \mathcal{O}$ for all $i \in \mathbb{N}$, i.e., all ρ_i are trivial proving (1). Thus assume $F_0 \subset \mathcal{O}$. By Proposition 2.6 there is $\delta > 0$ such that $e \in F_\delta$ and $G := \bigcup_{n\geq 0} F_{n\delta}$ is primal (i.e., G generates $\mathcal{O}$). From $e \in F_\delta$ it follows that $\rho_i \subseteq \rho_{i+\delta}$ for all $i \in \mathbb{N}$. Since there are only 2^{k^h} h-ary relations on $\mathbf{k}$, for each fixed $0 \leq i < \delta$ the non-decreasing chain $\rho_i \subseteq \rho_{i+\delta} \subseteq \rho_{i+2\delta} \subseteq \cdots$ is stationary, i.e. there is $j_i \in \mathbb{N}$ such that $\rho_{i+l\delta} = \rho_{i+j_i\delta}$ for all $l \geq j_i$. Put $m := \max\{j_0, \ldots, j_{\delta-1}\}$, and set $\alpha_i := \rho_{m\delta+i}$ $(i = 0, \ldots, \delta - 1)$. For a fixed $0 \leq i < \delta$, all $f \in F_{n\delta}$ we have $f[\alpha_i] = f[\rho_{m\delta+i}] \subseteq \rho_{(m+1)\delta+i} = \rho_{m\delta+i} = \alpha_i$ proving $f \in \mathrm{Pol}\,\alpha_i$, $F_{n\delta} \subseteq \mathrm{Pol}\,\alpha_i$ and finally $G \subseteq \mathrm{Pol}\,\alpha_i$. Owing to G primal we have α_i trivial and (2) holds.

Sufficiency. If (1) holds then by Lemma 3.4 we have $\mathrm{Pol}_0 \rho = \mathcal{O}$ and $\mathrm{Pol}\,\rho$ is complete by Proposition 2.6. Thus suppose that (2) holds. Then $e \in \mathrm{Pol}_\delta \rho$. In view of Proposition 2.6 it suffices to show that $\mathrm{Pol}_{m\delta} \rho = \mathcal{O}$ for all $m \geq 0$. Let $0 \leq i < \delta$. If $\alpha_i = \emptyset$, then $\rho_{l\delta+i} = \emptyset$ for all $l \in \mathbb{N}$ and

$$f[\rho_{l\delta+i}] = f[\emptyset] = \emptyset = \rho_{(\ell+m)\delta+i}$$

for every $f \in \mathcal{O}$. Thus assume that $\alpha_i = \Delta_\varepsilon$ for some equivalence relation ε on $\{1, \ldots, h\}$. Then for every $f \in \mathcal{O}$ and $l \in \mathbb{N}$ we have the required

$$f[\rho_{l\delta+i}] \subseteq f[\alpha_i] \subseteq \alpha_i = \rho_{(l+m)\delta+i}.$$

□

We say that $\rho = (\rho_0, \rho_1, \ldots)$ is *periodic* if there is $\delta > 0$ such that $\rho_{i+\delta} = \rho_i$ for all $i \geq 0$. For a periodic ρ the least $\delta > 0$ with this property is the *period* of ρ and it will be denoted by p_ρ.

3.8 Corollary *Let ρ be a periodic polyrelation with period p. Then ρ is proper if and only if at least one ρ_i is nontrivial.*

3.9 Definition It will be convenient to say that a polyrelation τ on $\mathbf{k}$ *dominates* a polyrelation ρ on $\mathbf{k}$ if $\mathrm{Pol}\,\rho \subseteq \mathrm{Pol}\,\tau$. For a given polyrelation ρ on $\mathbf{k}$ denote by $[\rho]$ the set of all polyrelations on $\mathbf{k}$ dominating ρ. Notice that $[\rho]$ is the least Galois-closed subset of $\mathcal{R}$ (i.e., the least set containing ρ and ω and closed under ζ, τ, pr, ∇, sh, $\cap$ and upper superpositions, see 3.5).

We describe a construction of $\tau \in [\rho]$. Let $h > 0$ and $\ell \geq p > 0$ be integers and let $\Gamma = (\gamma, g)$ where γ is an h-ary relation on $L = \{1, \ldots, \ell\}$ and g is a map from γ into the finite nonvoid subsets of $\mathbb{N}$. Let $\rho = (\rho_0, \rho_1, \ldots)$ be an h-ary polyrelation on $\mathbf{k}$. For each $i \in \mathbb{N}$ set

$$\tau_i = \{(\varphi(1), \ldots, \varphi(p)) \mid \varphi : L \to \mathbf{k},\ \varphi(a) \in \rho_{i+j} \text{ for all } a \in \gamma \text{ and } j \in g(a)\},$$

where for $a = (a_1, \ldots, a_h) \in \gamma$ we write $\varphi(a) = (\varphi(a_1), \ldots, \varphi(a_h))$. This defines an h-ary polyrelation $\tau = (\tau_0, \tau_1, \ldots)$ on $\mathbf{k}$ denoted by $\Gamma \hookrightarrow_p \rho$.

We illustrate this construction on a few examples.

3.10 Examples

(1) Let $\ell = p = h$, $\gamma = \{(\sigma(1), \dots, \sigma(h))\}$ where σ is a permutation of $\{1, \dots, h\}$ and $g((\sigma(1), \dots, \sigma(h))) = \{0\}$. Then $\tau = \Gamma \hookrightarrow_p \rho$ satisfies for all $i \in \mathbb{N}$

$$\tau_i = \{a_1 \dots a_n \mid a_{\sigma(1)} \dots a_{\sigma(h)} \in \rho_i\}.$$

Obviously, τ_i is obtained from ρ_i by a coordinate exchange independent on i; i.e., the same coordinate switch is performed on each ρ_i.

(2) Let $h = p = 2$, $\ell = 3$, $\gamma = \{(1,3),(3,2)\}$ and $g((1,3)) = \{0\}$, $g((3,2)) = \{1\}$. For a binary polyrelation ρ we get $\tau = \Gamma \hookrightarrow_2 \rho = (\rho_0 \circ \rho_1, \rho_1 \circ \rho_2, \dots)$ where $\circ$ denotes the standard relational product or composition (i.e. $\alpha \circ \beta := \{(x,y) \mid (x,u) \in \alpha, (u,y) \in \beta$ for some $u \in \mathbf{k}\}$).

(3) Let $\ell = p = h$, $\gamma = \{(1, \dots, h)\}$ and $g((1, \dots, h)) = \{1\}$. Then $\tau = \Gamma \hookrightarrow_h \rho = (\rho_1, \rho_2, \dots)$. Obviously τ is obtained from ρ by a one-step shift to the right.

(4) Let $\ell = p = h$, $\gamma = \{(1, \dots, h)\}$ and $g((1, \dots, h)) = \{0, 1\}$. Then $\tau = (\rho_0 \cap \rho_1, \rho_1 \cap \rho_2, \dots)$.

The Definition 3.9 is justified by the following lemma which is essentially in [Mi84] and can be proved directly.

3.11 Lemma *If ρ and τ are as in Definition* 3.9*, then $\tau \in [\rho]$; i.e.,* $\operatorname{Pold}\rho \subseteq \operatorname{Pold}\tau$.

3.12 A uniform incomplete clone C on $\mathbf{k}$ is *precomplete* if every uniform clone on $\mathbf{k}$ properly containing C is complete; in other words, if C is a maximal element of the set of uniform incomplete clones on $\mathbf{k}$ ordered by containment. A clone M of ordinary (i.e. non-delayed) operations on $\mathbf{k}$ is *maximal* if $M \subset \mathcal{O}$ and $M \subset M' \subset \mathcal{O}$ for no clone M'.

The maximal clones on $\mathbf{k}$ are completely known: [Po21] for $k = 2$, [Ia58] for $k = 3$ and [Ro65, Ro70] for $k > 3$. They are of the form $\operatorname{Pol}\sigma$ where the relation σ on $\mathbf{k}$ runs through 6 families. For further use we quote a few of them: (i) proper unary relations (i.e. subsets of $\mathbf{k}$ distinct from $\emptyset$ and $\mathbf{k}$); (ii) binary relations $\{xs(x) \mid x \in \mathbf{k}\}$ where s is a fixed permutation of $\mathbf{k}$ with k/p cycles of prime length p; (iii) bounded partial orders on $\mathbf{k}$ (i.e., transitive, reflexive and antisymmetric binary relations on $\mathbf{k}$ with a least and a greatest element); (iv) proper equivalence relations on $\mathbf{k}$ (i.e., distinct from $\{(x,x) \mid x \in \mathbf{k}\}$ and $\mathbf{k}^2$), and (v) binary central relations on $\mathbf{k}$ (i.e., reflexive and symmetric relations distinct from $\mathbf{k}^2$ and such that $c \times \mathbf{k} \subseteq \sigma$ for some $c \in \mathbf{k}$).

A set Ξ of proper polyrelations is termed *generic* if each incomplete uniform clone C extends to $\operatorname{Pold}\xi$ for some $\xi \in \Xi$. Our task is to find a small generic set. The optimal generic set Ξ would have $\operatorname{Pold}\xi$ precomplete for each $\xi \in \Xi$. Such a system would provide the best general completeness criterion in the sense that for a given F we have only to test whether $F \subseteq \operatorname{Pold}\xi$ for all $\xi \in \Xi$. The essential step is Hikita and Nozaki's generic system Ξ_0 [HN77]. For a relation σ put $\sigma^* = (\sigma, \sigma, \dots)$. We say that σ^* is of *type A* if $\operatorname{Pol}\sigma$ is maximal. Proper periodic polyrelations ξ of arity at most k such that $\operatorname{Pold}\xi$ is contained in no $\operatorname{Pold}\sigma^*$ of type A are said to be of *type B*. For the last type C we need the following special binary relations. An equivalence relation on a proper subset P of $\mathbf{k}$ distinct from $\{pp \mid p \in P\}$ is called a *proper partial equivalence* on $\mathbf{k}$. Let $\iota_2 := \{aa \mid a \in \mathbf{k}\}$.

3.13 Definition A binary polyrelation $(\rho_0, \iota_2, \iota_2, \dots)$ is of *type C* if

(1) $\rho_0 = c \times \mathbf{k}$ for some $c \in \mathbf{k}$, or

(2) ρ_0 is a proper partial equivalence relation on $\mathbf{k}$, or

(3) $\rho_0 = \{ps(p) \mid p \in P\}$ where $\emptyset \neq P \subseteq \mathbf{k}$ and either

  (a) s is a permutation of P of prime order or

  (b) s is an injective map from P into $\mathbf{k}$ such that

    (i) $s(P) \neq P$ and

    (ii) $s(p) \in P \iff s(p) = p$.

Denote by Ξ_0 the set of all polyrelations of type A, B and C. The following is a major step, and probably the most important one towards the general uniform completeness criterion.

3.14 Theorem [HN77, Hi81b] *The set Ξ_0 is generic. The uniform clones $\operatorname{Pol}\xi$ of type A and C are precomplete.*

The general uniform completeness criterion based on the full list of $\aleph_0$ uniform precomplete clones, was found for $k = 2$ in [Ku62] and for $k = 3$ in [Hi78].

3.15 Definition A set Ξ of polyrelations on $\mathbf{k}$ of type B (i.e., periodic and of arity at most k) is called *B-generic* if every precomplete clone $\operatorname{Pol}\sigma$ with σ a polyrelation of type B satisfies $\operatorname{Pol}\sigma = \operatorname{Pol}\tau$ for some $\tau \in \Xi$.

4 Minimal Polyrelations

4.1 In view of Theorem 3.14 it suffices to study the B-generic sets $\mathbf{B}$ of polyrelations. Denote by Ξ_1 the set of all polyrelations of type B, i.e., ρ proper, periodic, of arity at most k and such that $\operatorname{Pol}\rho \subseteq \operatorname{Pol}\sigma^*$ for no σ^* of type A. Clearly Ξ_1 is B-generic. For a given $\rho = (\rho_0, \rho_1, \dots) \in \mathbf{B}$ denote by h_ρ and p_ρ (or briefly h and p) its arity and period.

For $h > 2$ set

$$\iota_h := \{(x_1, \dots, x_h) \in \mathbf{k}^h \mid x_i = x_j \text{ for some } 1 \leq i < j \leq h\}$$

(i.e. ι_h consists of all h-tuples over $\mathbf{k}$ whose coordinates are not all pairwise distinct). An h-ary relation σ on $\mathbf{k}$ is *totally reflexive* if $\iota_h \subseteq \sigma$. For $2 < h \leq k$ an h-ary polyrelation $\lambda = (\lambda_0, \lambda_1, \dots)$ is *totally reflexive* if at least one λ_i is totally reflexive but distinct from $\mathbf{k}^h$. The following theorem is basic for our study.

4.2 Theorem *If $\rho \in \Xi_1$ then $[\rho]$ contains no totally reflexive periodic polyrelation.*

Proof Suppose $\tau = (\tau_0, \tau_1, \dots) \in [\rho]$ is a totally reflexive h-ary polyrelation of period p. The set $T := \{i \in \mathbf{p} \mid \tau_i \supseteq \iota_h\}$ is then nonempty. For $\delta \geq 0$ put $\lambda_\delta := \bigcap_{t \in T} \tau_{t+\delta}$. First observe that $\iota_h \subseteq \lambda_0 \subseteq \tau_t$ for all $t \in T$. Taking into account that for at least one $t \in T$ we have $\tau_t \subset \mathbf{k}^h$ we obtain $\iota_h \subseteq \lambda_0 \subset \mathbf{k}^h$, hence $\lambda = (\lambda_0, \lambda_1, \dots)$ is proper by Corollary 3.8. Next $\lambda = (\lambda_0, \lambda_1, \dots) \in [\tau]$ by Lemma 3.11. Indeed choose $\ell = p = h$,

$\gamma = \{(1,\dots,h)\}$ and $g((1,\dots,h)) = T$. Then for $\Gamma = (\gamma, g)$ the corresponding polyrelation $\sigma = \Gamma \hookrightarrow_p \tau = (\sigma_0, \sigma_1, \dots)$ satisfies $\sigma_\delta = \bigcap_{t\in T} \tau_{t+\delta} = \lambda_\delta$ for all $\delta \geq 0$ (see Examples 3.10(4)). For integers x and y denote by $x \dot{+} y$ the integer z such that $z \equiv x + y \pmod p$ and $0 \leq z < p$. Further denote by s the least positive integer such that $t \dot{+} s \in T$ for all $t \in T$. Clearly $0 < s \leq p$ and it is almost immediate that s divides p. Next $\lambda_{\delta+s} = \lambda_\delta$ for all $\delta \geq 0$. This shows that the period p_λ of λ divides s. We show that $\iota_h \not\subseteq \lambda_i$ for $i = 1, \dots, s-1$. Indeed, by the minimality of s there exists $t' \in T$ such that $t' \dot{+} i \notin T$ whereby $\lambda_i = \bigcap_{t\in T} \tau_{i+t} \subseteq \tau_{i+t'} = \tau_{i \dot{+} t'}$ and $\iota_h \subseteq \lambda_i$ would imply $\iota_h \subseteq \tau_{i \dot{+} t'}$ and $i \dot{+} t' \in T$. This contradiction shows $\iota_h \not\subseteq \lambda_i$ for all $i = 1, \dots, s-1$. From this it follows that actually $p_\lambda = s$. In view of λ being proper and ρ being B-generic we get $s > 1$. Set $F := \operatorname{Pol}\lambda$.

Denote by V the set of all operations on $\mathbf{k}$ taking less than k values. Further for all $i = 0, \dots, s-1$ set

$$F_i := \operatorname{Pol}_i \lambda = \bigcap_{j=0}^{s-1} \operatorname{Pol}(\lambda_j, \lambda_{j+i})$$

(see 3.1). We prove that $F_i \subseteq V$ for $i = 1, \dots, s-1$. Suppose to the contrary that $\operatorname{im} f = \mathbf{k}$ for some n-ary $f \in F_i$. Then there exist $x^0, \dots, x^{k-1}$ in $\mathbf{k}^n$ such that $f(x^j) = j$ for all $j \in \mathbf{k}$. As $\iota_h \not\subseteq \lambda_i$ there exists $(j_1, \dots, j_h) \in \iota_h \setminus \lambda_i$. Denote by X the $h \times n$ matrix with rows $x^{j_1}, \dots, x^{j_h}$. Clearly all columns of X belong to ι_h and hence to λ_0. As $f \in F_i \subseteq \operatorname{Pol}(\lambda_0, \lambda_i)$ clearly $(y_{j_1}, \dots, y_{j_h}) \in \lambda_i$. This contradiction proves $f \in V$ and $F_i \subseteq V$.

It is known [Ro70] that for the nontrivial totally reflexive relation λ_0 there exists an at least h-ary totally reflexive relation ζ such that $\operatorname{Pol}\lambda_0 \subseteq \operatorname{Pol}\zeta$ where $\operatorname{Pol}\zeta$ is maximal in $\mathcal{O}$. We have $F_0 \subseteq \operatorname{Pol}\lambda_0 \subseteq \operatorname{Pol}\zeta$ and, in view of total reflexivity, also $F_i \subseteq V \subseteq \operatorname{Pol}\zeta$ for $i = 1, \dots, s-1$. Thus $F \subseteq \operatorname{Pol}\zeta^*$ in contradiction to $\rho \in \Xi_1$. □

A nontrivial relation is *primitive* if it is the union of diagonal relations, and a polyrelation λ is *primitive* if all λ_i are trivial or primitive and at least one λ_i is primitive.

4.3 Lemma *If $\rho \in \Xi_1$ then $[\rho]$ contains no primitive periodic polyrelation.*

Proof Suppose to the contrary that $\tau \in [\rho]$ is a primitive periodic polyrelation and set $F := \operatorname{Pol}\tau$. It is known [Ro70] that $\mathcal{O}^{(1)} \subseteq \operatorname{Pol}\sigma$ if and only if σ is trivial or primitive. Thus $F_0 = \bigcap_{i\geq 0} \operatorname{Pol}\tau_i \supseteq \mathcal{O}^{(1)}$. We prove that $F_\delta \subseteq \operatorname{Pol}\iota_k$ (recall $\iota_k := \{(a_1, \dots, a_k) \in \mathbf{k}^k \mid a_i = a_j$ for some $1 \leq i < j \leq k\}$) for all $\delta \in \mathbb{N}$. Suppose to the contrary that there exist $\delta \geq 0$ and $f \in F_\delta \setminus \operatorname{Pol}\iota_k$. Let f be n-ary. By Słupecki's criterion [Sł39], see [PK79, 5.3], the operation f depends essentially on at least two variables and takes all values from $\mathbf{k}$. Thus there exist $a_0, \dots, a_{k-1} \in \mathbf{k}^n$ such that $fa_i = i$ for all $i \in \mathbf{k}$ and hence $f(h_1(x), \dots, h_n(x)) \approx x$ for some $h_1, \dots, h_n \in \mathcal{O}^{(1)}$. As $\mathcal{O}^{(1)} \subseteq F_0$ and $f \in F_\delta$, clearly $(e, \delta) \in F$ and $e \in F_\delta$ (where $e = e_1^1 = \mathrm{id}_{\mathbf{k}}$). Let $g \in \mathcal{O}$ be arbitrary. Again by Słupecki's criterion the operation g is a composition of operations from $\{f\} \cup \mathcal{O}^{(1)}$. Using the fact that $\mathcal{O}^{(1)} \subseteq F_0$ and $e, f \in F_\delta$ we can convert this composition into a uniform one (see 2.2). It follows that $g \in F_{m\delta}$ for some $m \geq 0$. By Proposition 2.6 the closed set F is complete. As τ is primitive, at least one τ_i is nontrivial and so by Corollary 3.8 the polyrelation τ is proper. This contradicts $F = \operatorname{Pol}\tau$ incomplete and so $F_\delta \subseteq \operatorname{Pol}\iota_k$ for all $\delta \geq 0$; consequently, $F \subseteq \operatorname{Pol}\iota_k^*$ where ι_k^* is of type A, in contradiction to $\rho \in \Xi_1$. □

4.4 Our goal is to find the smallest possible B-generic system. The basic strategy is the following: given a B-generic Ξ we find $\Xi' \subset \Xi$ such that each $\rho \in \Xi$ is dominated by some $\rho' \in \Xi'$. This reduction will be done in several steps. In the first step we basically reduce the arities. For ease of presentation a proper periodic h-ary polyrelation (with $h \leq k$) ρ is *minimal* if either $h = 1$ or ρ is dominated by no proper periodic polyrelation of arity less than h. Denote by Ξ_2 the set consisting of minimal polyrelations. It is almost immediate that Ξ_2 is B-generic. First we show that Ξ_2 consists of unary and binary relations. In the remainder of the section the polyrelation $\rho = (\rho_0, \rho_1, \dots)$ denotes a fixed minimal polyrelation of arity h and period p. Put

$$\sigma_h := \{x_1 \dots x_h \in \mathbf{k}^h \mid x_i \neq x_j \text{ for all } 1 \leq i < j \leq h\}$$

and recall that $\iota_h := \mathbf{k}^h \setminus \sigma_h$. Put $\omega_h := \{x \dots x \in \mathbf{k}^h \mid x \in \mathbf{k}\}$. We use throughout the notation $\rho_i := \mu_i \cup \nu_i$ where $\mu_i := \rho_i \cap \sigma_h$ and $\nu_i := \rho_i \cap \iota_h$ $(i = 0, \dots, p-1)$.

We start out with the following technical lemmas. For an h-ary relation τ and $1 \leq i_1 < \dots < i_l \leq h$ put

$$\mathrm{pr}_{i_1 \dots i_l} \tau := \{x_{i_1} \dots x_{i_l} \mid x_1 \dots x_h \in \tau\}, \qquad \mathrm{Pr}_{i_1 \dots i_l} \tau := \mathrm{pr}_{j_1 \dots j_{h-l}} \tau$$

where $1 \leq j_1 < \dots < j_{h-l} \leq h$ and $\{i_1, \dots, i_l, j_1, \dots, j_{h-l}\} = \{1, \dots, h\}$. Denote by $\mathbf{E}_h$ the set of equivalence relations on $\{1, \dots, h\}$ and put $\xi_h := \{11, 22, \dots, hh\}$. Finally for $1 \leq i < j \leq h$ denote by Δ_{ij} the diagonal relation $\{a_1 \dots a_h \in \mathbf{k}^h \mid a_i = a_j\}$. We have:

4.5 Lemma *If $h > 2$, $m \geq 0$ and ρ_m is nonvoid, then $\rho_m \cap \Delta_{ij}$ is diagonal for all $1 \leq i < j \leq h$.*

Proof For notational simplicity let $i = 1$ and $j = h$. Set $\tau_n := \mathrm{Pr}_h(\rho_n \cap \Delta_{1h})$ for all $n \geq 0$. The $(h-1)$-ary polyrelation τ thus defined dominates ρ since in Lemma 3.11 we can choose $\ell = h$, $p = h - 1$, $\gamma = \{(1, 2, \dots, h-1, 1)\}$ and $g((1, 2, \dots, h-1, 1)) = \{0\}$. By the minimality of ρ the polyrelation τ is improper and hence every τ_n is trivial by Corollary 3.8. Suppose to the contrary that $\tau_m = \emptyset$. Set $\xi_n := \mathrm{pr}_{1h} \rho_n$ for all $n \geq 0$. Clearly $\xi_m \neq \emptyset$ due to the assumption $\rho_m \neq \emptyset$. We claim that the binary relation ξ_m is areflexive (i.e. if $xy \in \xi_m$ then $x \neq y$). Indeed, were $aa \in \xi_m$ for some $a \in \mathbf{k}$ then $(a_1, a_2, \dots, a_{h-1}, a) \in \rho_m$ for some $a_2, \dots, a_{h-1} \in \mathbf{k}$ and $(a, a_2, \dots, a_{h-1}) \in \tau_m = \emptyset$. Thus the binary polyrelation ξ dominates ρ and is proper because ξ_m is nontrivial; however, this contradicts the minimality of ρ. Thus τ_m is diagonal, and $\tau_m = \Delta_\theta$ for some equivalence relation θ on $\{1, \dots, h-1\}$. It is easy to see that $\rho_m \cap \Delta_{1h} = \Delta_{\theta'}$ where θ' is the equivalence relation on $\{1, \dots, h\}$ obtained by adjoining h to the block of θ' containing 1. □

An h-ary relation τ is *reflexive* if $\tau \cap \iota_h$ is primitive or diagonal.

4.6 Lemma *If $h > 2$ then every ρ_i is empty or reflexive.*

Proof Let $\rho_i \neq \emptyset$. Then

$$\nu_i = \rho_i \cap \iota_h = \bigcup_{1 \leq j < l \leq h} (\rho_i \cap \Delta_{jl})$$

where by Lemma 4.5 all $\rho_i \cap \Delta_{jl}$ are diagonal. □

4.7 Lemma *If $\mu_i = \rho_i \cap \sigma_h \neq \emptyset$ then $\mathrm{Pr}_l\, \rho_i = \mathbf{k}^{h-1}$ for all $l = 1, \dots, h$.*

Proof The relation $\mathrm{Pr}_l\, \rho_i$ is a trivial relation which contains a repetition-free $(h-1)$-tuple proving $\mathrm{Pr}_l\, \rho_i = \mathbf{k}^{h-1}$. □

4.8 Lemma *No minimal polyrelation has arity greater than 4.*

Proof Suppose to the contrary that ρ is a minimal polyrelation of arity $h > 4$. By Lemma 4.3 the polyrelation ρ is nonprimitive, i.e. at least one ρ_i is nontrivial and nonprimitive. From Lemma 4.6 we have $\nu_i = \Delta_E$ where E is a clutter in $\mathbf{E}_h$ (see 3.6). In view of Lemma 4.5 the relation $\rho_i \cap \Delta_{12} = \nu_i \cap \Delta_{12}$ is diagonal; i.e., $\nu_i \cap \Delta_{12} = \Delta_\theta$ for some equivalence relation $\theta \in \mathbf{E}_h$. Suppose to the contrary that $\Delta_{12} \neq \Delta_\theta$; i.e. θ is not the equivalence relation on $\{1, \dots, h\}$ whose unique nonsingleton block is $\{1, 2\}$. Then there are $1 \leq m < n \leq h$ such that $(m, n) \in \theta$ and $\{m, n\} \neq \{1, 2\}$. In view of $h > 4$ we can choose t so that $2 < t \leq h$ and $m \neq t \neq n$. By Lemma 4.7 we have $\mathrm{Pr}_t\, \rho_i = \mathbf{k}^{h-1}$; therefore

$$\mathrm{Pr}_t\, \rho_i \supseteq \Delta'_{12} := \{x_1 \dots x_{h-1} \in \mathbf{k}^{h-1} \mid x_1 = x_2\},$$

and due to $t > 2$ also $\Delta'_{12} \subseteq \mathrm{Pr}_t(\rho_i \cap \Delta_{12}) = \mathrm{Pr}_t\, \Delta_\theta$. However, this is impossible because each $a_1 \dots a_{h-1} \in \mathrm{Pr}_t\, \Delta_\theta$ has equal coordinates at places corresponding to m and n (e.g. if $t > n$ we have $a_m = a_n$). Thus $\Delta_\theta = \Delta_{12}$ and $\Delta_{12} \subseteq \rho_i$. By symmetry $\Delta_{lj} \subseteq \rho_i$ for all $1 \leq l < j \leq h$ proving $\iota_h \subseteq \rho_i$ and ρ_i totally reflexive. But this contradicts Theorem 4.2. □

Next we consider the quaternary (i.e. 4-ary) minimal relations. For $\{a, b, c, d\} = \{1, \dots, 4\}$ denote by *ab-cd* the equivalence relation on $\{1, \dots, 4\}$ with two blocks $\{a, b\}$ and $\{c, d\}$ and set $\chi := \Delta_{12-34} \cup \Delta_{13-24} \cup \Delta_{14-23}$.

4.9 Lemma *If $h = 4$ then each nonprimitive and nontrivial ρ_i contains χ.*

Proof By Lemma 4.5 the relation $\rho_i \cap \Delta_{12}$ is diagonal i.e. equals Δ_θ for some $\theta \in \mathbf{E}_4$ such that $(1, 2) \in \theta$. Applying Lemma 4.7 we get $\mathrm{Pr}_3\, \rho_i = \mathrm{Pr}_4\, \rho_i = \mathbf{k}^3$ and therefore neither 3 nor 4 can be in the block of θ containing 1 and 2. Thus $\Delta_\theta \supseteq \Delta_{12-34}$. By symmetry the same result holds for every pair $1 \leq i < j \leq 4$ proving $\rho_i \supseteq \chi$. □

To settle the case of quaternary polyrelations we need the following technical lemmas. Denote by C the set of constant operations on $\mathbf{k}$ and by ω_h the h-ary relation $\{a \dots a \mid a \in \mathbf{k}\}$.

4.10 Lemma *Let $\rho = (\rho_0, \rho_1, \dots) \in \Xi_2$ have arity $h > 2$, period p and satisfy* (i) *$[\rho]$ contains no proper binary polyrelation and* (ii) *$\rho_j = \emptyset$ or $\rho_j \supseteq \omega_h$ for all $j \in \mathbf{p}$. Then there is $0 < i < p$ such that $\Delta_\theta \subseteq \rho_i$ whenever $\theta \in \mathbf{E}_h$ has exactly two blocks and satisfies $\Delta_\theta \subseteq \rho_0$.*

Proof Suppose to the contrary that for every $0 < i < p$ there is $\theta_i \in \mathbf{E}_h$ with exactly two blocks B_{i1} and B_{i2} such that $\Delta_{\theta_i} \subseteq \rho_0 \setminus \rho_i$. We show:

Claim 1: $\Delta_{\theta_i} \cap \rho_i \subseteq \omega_h$ for all $0 < i < p$.

Proof: Suppose to the contrary that $\Delta_{\theta_i} \cap \rho_i \not\subseteq \omega_h$ for some $0 < i < p$. Choose $b_l \in B_{il}$ $(l = 1, 2)$ and set

$$\tau_j := \mathrm{pr}_{b_1 b_2}(\Delta_{\theta_i} \cap \rho_j) = \{x_{b_1} x_{b_2} \mid x_1 \dots x_h \in \Delta_{\theta_i} \cap \rho_j\}$$

for all $j \geq 0$. From $\Delta_{\theta_i} \cap \rho_i \not\subseteq \omega_h$ we get $\rho_i \neq \emptyset$ and (ii) yields $\rho_i \supseteq \omega_h$. From this it follows that τ_i is a reflexive binary relation on $\mathbf{k}$. Moreover, τ_i is distinct from $\mathbf{k}^2$ because $\Delta_{\theta_i} \not\subseteq \rho_i$. Thus $\tau := (\tau_0, \tau_1, \dots)$ is a proper binary polyrelation. Letting $a_m = \ell$ for all $m \in B_{i\ell}$ and $\ell = 1, 2$, $\gamma = \{(a_1, \dots, a_h)\}$ and $g((a_1, \dots, a_h)) = \{0\}$ we obtain from Lemma 3.11 that the polyrelation τ belongs to $[\rho]$ in contradiction to (i). This proves the claim. □

Set $F := \operatorname{Pol}\rho$ and $F_1 = \{f \mid (f, i) \in F\}$ for all $i \in \mathbb{N}$.

Claim 2: $F_i \subseteq C$ for $i = 1, \dots, p-1$.

Proof: Suppose to the contrary that there exist $0 < i < p$, $n \geq 1$ and $f \in F_i^{(n)} \setminus C$. Then $fx \neq fy$ for some $x = (x_1, \dots, x_n) \in \mathbf{k}^n$ and $y = (y_1, \dots, y_n) \in \mathbf{k}^n$. For $j = 1, \dots, h$ and $l = 1, \dots, n$ set $z_{jl} := x_l$ if $j \in B_{i1}$ and $z_{jl} := y_l$ if $j \in B_{i2}$. Clearly each column of the $h \times n$ matrix $Z = [z_{jl}]$ belongs to $\Delta_{\theta_i} \subseteq \rho_0$ and from $f \in \operatorname{Pol}(\rho_0, \rho_i)$ the values of f in the rows of Z form $b \in \rho_i$. Clearly $b \in \Delta_{\theta_i}$, and thus by Claim 1 we have $z \in \Delta_{\theta_i} \cap \rho_i \subseteq \omega_h$. This contradiction shows that $F_i \setminus C = \emptyset$; i.e., $F_i \subseteq C$ and the claim is proved. □

Claim 3: $C \subseteq F_0$.

Proof: From the assumption (ii) we get $C \subseteq \operatorname{Pol}\rho_j$ for all $0 \leq j < p$ and thus $C \subseteq \bigcap_{j \in \mathbb{N}} \operatorname{Pol}\rho_j = F_0$, proving the claim. □

Now clearly $e \in F_0$. As $\rho \in \Xi_1$, the set $F = \operatorname{Pol}\rho$ is incomplete. Now Proposition 2.6 (for $\delta = 0$) yields that F_0 is not primal. By Claim 3 also $C \subseteq F_0$. It is implicit in [Ro70] that there exists a reflexive relation σ on $\mathbf{k}$ such that $F_0 \subseteq \operatorname{Pol}\sigma$ and $\operatorname{Pol}\sigma$ is maximal in $\mathcal{O}$. Together with $F_i \subseteq C \subseteq F_0 \subseteq \operatorname{Pol}\sigma$ $(i = 1, \dots, p-1)$ we get $F \subseteq \operatorname{Pol}\sigma^*$ in contradiction to $\rho \in \Xi_1$. □

Recall that for a clutter E in $\mathbf{E}_h$ the relation $\Delta_E = \bigcup_{\epsilon \in E} \Delta_\epsilon$ was defined in 3.6.

4.11 Lemma *Let ρ satisfy the assumptions of Lemma 4.10 and let E be a clutter in $\mathbf{E}_h$ such that* (i) *each $\theta \in E$ has exactly two blocks and* (ii) *$\bigcap E$ is the least equivalence relation. Then $\Delta_E \subseteq \rho_i$ implies $\rho_i = \mathbf{k}^h$.*

Proof Put $T := \{t \in \mathbf{p} \mid \Delta_E \subseteq \rho_t\}$. Suppose to the contrary that $\rho_u \subset \mathbf{k}^h$ for some $u \in T$. For $\delta \geq 0$ put $\lambda_\delta := \bigcap_{t \in T} \rho_{\delta + t}$. Clearly $\Delta_E \subseteq \lambda_0 \subseteq \rho_u \subset \mathbf{k}^h$. Suppose that $\Delta_E \subseteq \Delta_\theta$ for some $\theta \in \mathbf{E}_h$. Then $\bigcap E \supseteq \theta$ and from (ii) we obtain that θ is the least equivalence relation on $\{1, \dots, h\}$ and $\Delta_\theta = \mathbf{k}^h$. This shows that λ_0 is nontrivial and therefore $\lambda := (\lambda_0, \lambda_1, \dots)$ is proper and $\lambda \in [\rho]$ by Lemma 3.11.

Denote by s the least positive integer such that $t \dot{+} s \in T$ for all $t \in T$ (where, as in the proof of Theorem 4.2, $t \dot{+} s$ denotes the least integer $0 \leq z < p$ such that $z \equiv t + s \pmod p$). Clearly $s \leq p$ and s divides p. Thus λ has period s and in view of $\rho \in \Xi_1$, we get $\lambda \in \Xi_1$. Now λ also satisfies the assumptions of Lemma 4.10 and therefore $\Delta_E \subseteq \lambda_i$ for some $0 < i < s$, proving $t \dot{+} i \in T$ for all $t \in T$ in contradiction to the minimality of s. Thus $\rho_u = \mathbf{k}^h$ for all $u \in T$. □

4.12 Proposition *There is no minimal quaternary polyrelation.*

Proof Suppose to the contrary that $\rho = (\rho_0, \rho_1, \dots)$ is a minimal quaternary polyrelation of period p and let E consist of the three equivalence relations on $\{1, \dots, 4\}$ with two 2-element blocks each. Applying Lemmas 4.9 and 4.11 we obtain the contradiction that every ρ_i is either primitive or trivial. □

We consider ternary polyrelations. Recall that

$$\Delta_{12} = \{aab \mid a, b \in \mathbf{k}\}, \qquad \Delta_{13} = \{aba \mid a, b \in \mathbf{k}\}, \qquad \Delta_{23} = \{abb \mid a, b \in \mathbf{k}\},$$

$\omega_3 = \{aaa \mid a \in \mathbf{k}\}$, and $\nu_i = \rho_i \cap \iota_3$ for all $i \geq 0$.

4.13 Lemma *Let $\rho \in \Xi_1$ be ternary with period p and such that* (i) *$[\rho]$ contains no proper binary polyrelation and* (ii) *$\rho_j = \emptyset$ or $\rho_j \supseteq \omega_3$ for all $0 \leq j < p$. Then for all $j \in \mathbf{p}$*

$$\nu_j \in I = \{\emptyset, \omega_3, \Delta_{12}, \Delta_{13}, \Delta_{23}\}.$$

Proof Suppose to the contrary that some ν_i is nontrivial. By Lemma 4.6 we have $\nu_i = \rho_i \cap \iota_3 = \Delta_E$ for a clutter E in $\mathbf{E}_3$ with $|E| \geq 2$. Then each $\varepsilon \in E$ has exactly two blocks and the relation $\bigcap E$ is the least equivalence relation on $\{1, 2, 3\}$. Applying Lemma 4.11 we get $\rho_i = \mathbf{k}^3$, a contradiction. Thus ν_i is trivial and $\nu_i \in I$. □

4.14 Lemma *If $\rho = (\rho_0, \rho_1, \dots)$ is a minimal ternary polyrelation then each nontrivial ρ_i has $\nu_i = \omega_3$.*

Proof Suppose to the contrary that there is a nontrivial ρ_i with $\nu_i \neq \omega_3$. By Lemma 4.6 for all $0 \leq j < p$ either $\rho_j = \emptyset$ or $\rho_j \supseteq \omega_3$. By Lemma 4.13 then $\nu_i \in \{\Delta_{12}, \Delta_{13}, \Delta_{23}\}$. By an appropriate exchange of coordinates we can get $\nu_i = \Delta_{12}$ (i.e., we apply Lemma 3.11 to $\ell = p = h = 3$, $\Gamma = (\gamma, g)$ with $\gamma = \{(a, b, c)\}$, $g((a, b, c)) = \{0\}$ where $\{a, b, c\} = \{1, 2, 3\}$ is suitably chosen). By Lemma 4.7 we have $\mathrm{Pr}_3\, \rho_i = \mathbf{k}^2$; i.e., for arbitrary $x, y \in \mathbf{k}$ there exists w so that $xyw \in \rho_i$. For $l = 0, \dots, p-1$ and $n = 3, \dots, k$ put

$$\tau_l^n = \{x_1 \dots x_n \mid x_i uv \in \rho_l \ (i = 1, \dots, n) \text{ for some } u, v \in \mathbf{k}\}.$$

Clearly the n-ary polyrelation $\tau^n = (\tau_0^n, \tau_1^n, \dots)$ thus defined belongs to $[\rho]$ (indeed, in Lemma 3.11 set $\tau^n = \Gamma \hookrightarrow_n \rho$ where $\ell = n + 2$, $p = n$, $h = 3$, $\Gamma = (\gamma, g)$ with $\gamma = \{i(n+1)(n+2) \mid i = 1, \dots, n\}$ and $g(a) = \{0\}$ for each $a \in \gamma$). By induction on $n = 3, \dots, k$ we prove that $\tau_i^n = \mathbf{k}^n$. First we prove that $\tau_i^3 \supseteq \iota_3$. Let $x, y \in \mathbf{k}$ be arbitrary. Then $xyw \in \rho_i$ for some w. Using $xyw \in \rho_i$ and $yyw \in \Delta_{12} \subseteq \rho_i$ we get $xxy \in \tau_i^3$. Similarly $xyx \in \tau_i^3$ and $xxy \in \tau_i^3$ proving $\tau_i^3 \supseteq \iota_3$. Now τ^3 is not totally reflexive by Theorem 4.2 and therefore $\tau_i^3 = \mathbf{k}^3$ for all $i \geq 0$.

Suppose that $3 \leq n < k$ and $\tau_i^n = \mathbf{k}^n$. Let $x_1, \dots, x_n \in \mathbf{k}$ be arbitrary. By the inductive assumption $x_j uv \in \rho_i$ $(j = 1, \dots, n)$ for some u,$v \in \mathbf{k}$. Thus for all $1 \leq j < l \leq n$ we have $x_1 \dots x_{l-1} x_j x_{l+1} \dots x_n \in \tau_i^{n+1}$ proving that $\iota_{n+1} \subseteq \tau_i^{n+1}$. Again by Theorem 4.2 we have $\tau_i^{n+1} = \mathbf{k}^{n+1}$ completing the induction step. Now, $\tau_i^k = \mathbf{k}^k$ means that there are $u, v \in \mathbf{k}$ such that $xuv \in \rho_i$ for all $x \in \mathbf{k}$. In particular, $vuv \in \rho_i$ shows $u = v$. However $xuu \in \rho_i$ holds only for $x = u$. This contradiction proves the lemma. □

4.15 We need the following result from [Ro70]. Recall that for a prime number q an *elementary abelian q-group* is an abelian group $\langle G; +, -, 0\rangle$ such that $qx = 0$ (where $qx = x + \dots + x$ with q summands) for all $x \in G$. Such a finite group is isomorphic to the additive structure of a vector space over the Galois field $GF(q)$. More explicitly, $G \simeq \langle \mathbf{q}^n; \oplus\rangle$ where for $x = (x_1, \dots, x_n) \in \mathbf{q}^n$ and $y = (y_1, \dots, y_n) \in \mathbf{q}^n$, $x \oplus y := (x_1 \dot{+} y_1, \dots, x_n \dot{+} y_n)$ (the mod q addition $a \dot{+} b$ is the remainder of the division of $a + b$ by q). Put $m_G^\circ := \{xyz(x - y + z) \mid x, y, z \in \mathbf{k}\}$.

4.16 Lemma *Let $\lambda = \mu_3 \cup \omega_3$ be a ternary relation on $\mathbf{k}$ where $\emptyset \neq \mu_3 \subseteq \sigma_3$. Suppose that* (i) $\operatorname{Pol}\lambda \subseteq \operatorname{Pol}\tau$ *for no binary nontrivial relation τ,* (ii) $\operatorname{Pol}\lambda \subseteq \operatorname{Pol}\tau$ *for no nontrivial at most k-ary totally reflexive relation τ and* (iii) $\operatorname{Pol}\lambda \subseteq \operatorname{Pol}\tau$ *for no ternary relation τ of the form $\mu \cup \Delta_{12}$ or $\mu \cup \Delta_{12} \cup \Delta_{13}$ with $\emptyset \neq \mu \subseteq \sigma_3$. Then there exists an elementary abelian q-group G on $\mathbf{k}$ such that $\operatorname{Pol}\lambda \subseteq \operatorname{Pol} m_G^\circ$.*

Now we can prove:

4.17 Proposition *There is no minimal ternary polyrelation.*

Proof Suppose to the contrary that there exists a minimal ternary polyrelation ρ. Let ρ_i be nontrivial. By Lemma 4.14 then $\nu_i = \omega_3$. Now by minimality, ρ_i satisfies the assumption (i) of Lemma 4.16. The assumption (ii) is true due to Theorem 4.2. Next (iii) holds on account of Lemma 4.14. From Lemma 4.16 it follows that there is a prime q, an elementary abelian q-group G on $\mathbf{k}$ and a quaternary polyrelation $\tau \in [\rho]$ such that $\tau_i = m_G^\circ$ (the construction from [Ro70] yielding m_G° gives τ when applied to the polyrelation ρ). It is easy to verify that $\tau_j \supseteq \omega_4$ whenever τ_j is nonempty (due to $\rho_j \supseteq \omega_3$ whenever $\rho_j \neq \emptyset$).

It follows that $\rho_0 \subseteq \operatorname{Pol}\tau_0$. Finally $m_G^\circ \supseteq \Delta_{12-34} \cup \Delta_{14-23}$ because $x - x + y \approx y$ and $x - y + y \approx x$). Since $E = \{\{\{1,2\},\{3,4\}\},\{\{1,3\},\{2,4\}\}\}$ is such that $\bigcap E$ is the least equivalence relation on $\{1,2,3,4\}$, we have a contradiction to Lemma 4.11. □

Summing up:

4.18 Theorem *The set Ξ_3 of proper periodic unary and binary polyrelations is B-generic.*

5 Unary Polyrelations

5.1 In this section we study the set V of proper periodic unary polyrelations on $\mathbf{k}$. Call $\rho \in V$ with period p *optimal* if

(1) $p = \min\{p_\tau \mid \tau \in V \cap [\rho]\}$ and

(2) $|\rho_0| + \cdots + |\rho_{p-1}| \leq |\lambda_0| + \cdots + |\lambda_{p-1}|$ whenever $\lambda \in V \cap [\rho]$ and $p_\lambda = p$.

Thus for optimality our primary concern is to have the shortest possible period p among the proper unary polyrelations from $[\rho]$ while our secondary concern is to have the least possible sum of the sizes of the relations $\rho_0, \ldots, \rho_{p-1}$ (amongst those with the shortest period p in $[\rho]$). In 5.2–5.8, ρ is an optimal unary polyrelation with period p. We need the following technical lemma.

5.2 Lemma *If $\tau \in V \cap [\rho]$ satisfies $\tau_i \subseteq \rho_i$ for all $i \geq 0$ and $p_\tau \leq p_\rho$ then $\tau = \rho$.*

Proof From optimality we get $p_\tau = p_0 = p$. From $\tau_i \subseteq \rho_i$ $(i = 0, \ldots, p-1)$ we have

$$|\tau_0| + \cdots + |\tau_{p-1}| \leq |\rho_0| + \cdots + |\rho_{p-1}|. \tag{5.1}$$

By optimality also $|\rho_0| + \cdots + |\rho_{p-1}| \leq |\tau_0| + \cdots + |\tau_{p-1}|$, so (5.1) holds with equality. In view of $\tau_i \subseteq \rho_i$ we get $\tau_i = \rho_i$ for $i = 0, \ldots, p-1$ and $\tau = \rho$. □

Now we have:

5.3 Lemma *The sets $\rho_0, \dots, \rho_{p-1}$ are pairwise disjoint.*

Proof Suppose to the contrary that $\rho_i \cap \rho_j \neq \emptyset$ for some $0 \leq i < j < p$. Put $r := j - i$ and $\tau_n := \rho_n \cap \rho_{n+r}$ for $n = 0, \dots, p-1$. Now $\tau_i = \rho_i \cap \rho_j \neq \emptyset$ shows τ to be proper. It can be checked directly that $\tau = \Gamma \hookrightarrow_1 \rho$ where $\ell = p = h = 1$ and $\Gamma = (\gamma, g)$ has $\gamma = \{(1)\}$ and $g((1)) = \{0, r\}$. According to Lemma 3.11 we have $\tau \in [\rho]$ and thus $\tau \in V \cap [\rho]$. Moreover, $p_\tau \leq p$ and $\tau_x \subseteq \rho_x$ for all $x \geq 0$. Applying Lemma 5.2 we get $\tau = \rho$. Thus $\rho_i \cap \rho_{i+r} = \tau_i = \rho_i$ for all $i = 0, \dots, p-1$ and so $\rho_i \subseteq \rho_{i+r}$. Thus for every $0 \leq i < p$ we have $\rho_i \subseteq \rho_{i+r} \subseteq \cdots \subseteq \rho_{i+pr} = \rho_i$ proving $\rho_i = \rho_{i+r} = \cdots = \rho_{i+(p-1)r}$. Denote by s the greatest common divisor $\gcd(p, r)$ of p and r. It is easy to check that for all $i = 1, \dots, s-1$ we have that $\rho_i = \rho_{i+s} = \cdots = \rho_{i+p-s}$ and therefore ρ is periodic with period s. The minimality of p shows $p = s = \gcd(p, r)$, a contradiction to $0 < r < p$. This proves the lemma. □

We need the following simple observation.

5.4 Lemma *Let $\rho = (\rho_0, \rho_1, \dots)$ be an h-ary periodic polyrelation on $\mathbf{k}$ of period p. Then $\rho' = (\rho_1, \rho_2, \dots)$ is periodic of period p and*

$$\operatorname{Pol} \rho = \operatorname{Pol} \rho'.$$

Proof Let $\ell = p = h$ and for $i = 1, \dots, p-1$ set $\Gamma_i = (\gamma, g_i)$ where $\gamma = \{(1, \dots, h)\}$ and $g_i((1, \dots, h)) = \{i\}$. Then by Lemma 3.11 clearly $\rho^{(i)} = \Gamma_i \hookrightarrow_h \rho$ satisfies $\operatorname{Pol} \rho \subseteq \operatorname{Pol} \rho^{(i)}$. In particular, $\rho' = \rho^{(1)}$ and $(\rho')^{p-1} = \rho$ show

$$\operatorname{Pol} \rho \subseteq \operatorname{Pol} \rho' \subseteq \operatorname{Pol}(\rho')^{(p-1)} = \operatorname{Pol} \rho.$$

□

Thus we can turn around the cycle of ρ at will without effecting the corresponding uniform clone. Set $I := \{i \in \mathbf{p} \mid \rho_i \neq \emptyset\}$. According to Lemma 5.4 henceforward we assume $0 \in I$. We have:

5.5 Lemma *$I = \{0, r, \dots, p-r\}$ for some divisor r of p.*

Proof If $I = \{0\}$ then clearly $r = p$. Thus suppose that $|I| > 1$. Denote by r the least positive integer such that $i \dot{+} r \in I$ for some $i \in I$ (where again $i \dot{+} r$ is the remainder of division of $i + r$ by p); in other words, r is the least distance between two consecutive members of I on the cycle. For $x \geq 0$ put

$$\tau_x := \begin{cases} \rho_x & \text{if } x \dot{+} r \in I, \\ \emptyset & \text{otherwise.} \end{cases} \tag{5.2}$$

Observe that the polyrelation τ is constructed in the way described in Definition 3.9. Indeed, set $\ell = 2$, $p = h = 1$, and $\Gamma = (\gamma, g)$ where $\gamma = \{(1), (2)\}$ and $g((1)) = \{0\}$, $g((2)) = \{r\}$. The corresponding $\sigma = \Gamma \hookrightarrow_1 \rho$ has

$$\sigma_x = \{\varphi(1) \mid \varphi : \{1, 2\} \to \mathbf{k},\ \varphi(1) \in \rho_x,\ \varphi(2) \in \rho_{x+r}\}$$

for all $x \geq 0$ which is (5.2). In view of $i, i \dot{+} r \in I$, we have $\tau_i = \rho_i \neq \emptyset$ and so $\tau \in V \cap [\rho]$. Clearly $p_\tau \leq p$ and $\tau_x \subseteq \rho_x$ for all $x \geq 0$. Applying Lemma 5.2 we obtain $\tau = \rho$. According

to (5.2) we have $i \in I \Rightarrow i \dot{+} r \in I$. In particular, from $0 \in I$ we get that I consists of all $j \in \mathbf{p}$ such that $j \equiv rx \pmod p$ for some integer x. Now r is a divisor of p on account of the minimality of r. Thus $I \supseteq \{0, r, \dots, p-r\}$ and again by the minimality of r we obtain the required $I = \{0, r, \dots, p-r\}$. □

The following lemma relates r and $n := p/r$. For later use it is given in a more general form.

5.6 Lemma *Let ρ be an at most binary periodic polyrelation with period p and such that $p_\sigma < p$ for no proper $\sigma \in [\rho]$. Let r be a divisor of p such that $p/r = p_1^{a_1} \dots p_l^{a_l}$ for $l \geq 1$, $p_1, \dots, p_l$ distinct primes and positive integers $a_1, \dots, a_l$. Suppose that for $I := \{0, r, 2r, \dots, p-r\}$ and $J := \mathbf{p} \setminus I$ we have either*

(1) (a) *$\emptyset \neq \rho_i \subset \mathbf{k}$ for all $i \in I$ while $\rho_i = \emptyset$ for all $j \in J$, or*

(b) *$\rho_i \neq \emptyset$ is an areflexive binary relation for all $i \in I$ while $\rho_j = \emptyset$ for all $j \in J$,*

or

(2) *$\iota_2 \subset \rho_i \subset \mathbf{k}^2$ for all $i \in I$ while $\rho_j = \iota_2$ for all $j \in J$.*

Then $r = p_1^{b_1} \cdots p_l^{b_l}$ for some $b_1, \dots, b_l \geq 0$.

Proof Put $n := p/r$. Suppose to the contrary that there exists a prime divisor d of r not dividing n. Set $r' := r/d$. For $i \geq 0$ define a polyrelation σ^i as follows. (i) If r' does not divide i put $\sigma^i = (\emptyset, \emptyset, \dots)$ in the case (1) and $\sigma^i = (\iota_2, \iota_2, \dots)$ in the case (2) (whereby σ^i is unary if ρ is unary and binary otherwise). (ii) Let $i = r'x$ for some $x \geq 0$. Denote by z the unique solution in $\mathbf{n}$ of the congruence $dz \equiv x \pmod n$. This solution exists as d is a prime and d does not divide n. Now for all $j \geq 0$ set

$$\sigma^i_j := \rho_{rz+j}. \tag{5.3}$$

We prove:

Claim 1:

$$\sigma^i_j \subseteq \sigma^{i+j}_0 \quad \text{for all } i, j \geq 0. \tag{5.4}$$

Proof: Suppose to the contrary that (5.4) does not hold for some $i, j \geq 0$. Then clearly $j > 0$, σ^i_j is nonvoid and, moreover, in the case (2) also $\sigma^i_j \supset \iota_2$ (since all the relations are then reflexive). Now r' divides i since otherwise in the case (1) we get $\sigma^i_j = \emptyset$ and in the case (2) we get $\sigma^i_j = \iota_2$. Thus $i = r'x$ for some $x \geq 0$ and $dz \equiv x \pmod n$ for a unique $z \in \mathbf{n}$ and $\sigma^i_j = \rho_{rz+j}$. According to (1) and (2) here $rz + j \equiv \ell \pmod p$ where $\ell \in I = \{0, r, \dots, p-r\}$. As r divides p, we obtain that r divides j; i.e., $j = ry$ for some $y \geq 0$. Note that from $r = r'd$

$$i + j = r'(x + dy). \tag{5.5}$$

Denote by z' the element of $\mathbf{n}$ satisfying $z' \equiv z + y \pmod n$. Now $dz' \equiv dz + dy \equiv x + dy \pmod n$. Next $z' = z + y + mn$ for some integer m. Using this and $j = ry$ we get

$$rz' = r(z + y) + mrn = rz + j + mp \equiv rz + j \pmod p. \tag{5.6}$$

Now on account of (5.5), (5.4) and (5.6) we have $\sigma^{i+j}_0 = \rho_{rz'+0} = \rho_{rz'} = \rho_{rz+j} = \sigma^i_j$ proving the claim. □

Claim 2: $\sigma^i \in [\rho]$ *for all* $i \geq 0$.

Proof: First consider the case where r' does not divide i. Then σ^i is either $(\emptyset, \emptyset, \dots)$ or $(\iota_2, \iota_2, \dots)$; hence $\operatorname{Pol} \sigma^i = \mathcal{U}$ (= the set of all uniformly delayed operations on $\mathbf{k}$) and clearly $\sigma^i \in [\rho]$. Thus let $i = r'x$ for some $x \geq 0$ and let $z \in \mathbf{n}$ satisfy $dz \equiv x \pmod{n}$. Then for $s := rz$ clearly $\sigma^i = (\rho_s, \rho_{s+1}, \dots)$, hence in view of Lemma 5.4 we have $\operatorname{Pol} \rho = \operatorname{Pol} \sigma^i$ and so $\sigma^i \in [\rho]$. □

Claim 3: $\sigma_0^{i+r'n} = \sigma_0^i$ *for all* $i \geq 0$.

Proof: If r' does not divide i then r' does not divide $i + r'n$ either and so $\sigma_0^{i+r'n} = \emptyset = \sigma_0^i$ in the case (1) while $\sigma_0^{i+r'n} = \iota_2 = \sigma_0^i$ in the case (2). Thus let $i = r'x$ for some $x \geq 0$. Denote by z the unique solution in $\mathbf{n}$ of $dz \equiv x \pmod{n}$. Then $dz \equiv x \equiv x + n \pmod{n}$ and so by (5.3)

$$\sigma_0^{i+r'n} = \sigma_0^{r'(x+n)} = \rho_{rz} = \sigma_0^{r'x} = \sigma_0^i.$$

□

The upper superposition of $\sigma^0, \sigma^1, \dots$ is $\sigma := (\sigma_0^0, \sigma_0^1, \sigma_0^2, \dots)$. In view of Claims 1 and 2, according to [Mi84] we have $\sigma \in [\rho]$.

Note that for $i = 0$ we have $0 = r0$ and $d0 \equiv 0 \pmod{n}$ and so $\sigma_0^0 = \rho_{r0} = \rho_0$ by (5.3). Since by assumption ρ_0 is nontrivial, according to Corollary 3.8 the polyrelation $\sigma \in [\rho]$ is proper. Finally $p_\sigma \leq r'n$ by Claim 3. By the definitions $r'n = (r/d)(p/r) = p/d < p$ and so $p_\sigma < p$ in contradiction to the hypothesis. Thus every divisor of r divides n and r has the required form. □

Applying Lemma 5.6 to optimal unary polyrelations and reformulating the divisibility condition we obtain:

5.7 Proposition *Let* ρ *be an optimal unary polyrelation with nontrivial* ρ_0 *and period* $p = p_1^{d_1} \cdots p_\ell^{d_\ell}$ *where* $\ell \geq 1$, $p_1, \dots, p_\ell$ *are pairwise distinct primes and* $d_1, \dots, d_\ell$ *positive integers. Then there exists* $r = p_1^{c_1} \cdots p_\ell^{c_\ell}$ *such that*

(1) *the integers* c_i *satisfy* $0 \leq c_i < d_i$ *for all* $i = 1, \dots, \ell$, *and*

(2) $\rho_0, \rho_r, \dots, \rho_{p-r}$ *are pairwise disjoint and nonempty subsets of* $\mathbf{k}$ *while* $\rho_j = \emptyset$ *for all* j *not divisible by* r.

5.8 Remarks

(1) An example of p and r in Proposition 5.7 are $p = 12$ and $r = 2$. Then $\rho_0, \rho_2, \dots, \rho_{10}$ are pairwise distinct nonempty subsets of $\mathbf{k}$ and $\rho_1, \rho_3, \dots, \rho_{11} = \emptyset$. Notice that $k \geq 6$.

(2) Suppose the number p/r of sets $\rho_0, \rho_r, \dots, \rho_{p-r}$ is a prime power q^m. Then $p = q^{n+m}$ for some $n \geq 1$.

(3) Let $k = 2$. There are two candidate subsets of $\mathbf{2}$ for the ρ_i, namely $\{0\}$ and $\{1\}$, and so $p/r = 2$ and $p = 2^{m+1}$ and $r = 2^m$. Thus we may suppose that $\rho_0 = \{0\}$, $\rho_{2^m} = \{1\}$ and $\rho_i = \emptyset$ for all $0 < i < 2^{m+1}$, $i \neq 2^m$. For every $m \geq 0$ the uniform clone $\operatorname{Pol} \rho$ is one of the $\aleph_0$ precomplete uniform clones on $\mathbf{2}$ from [Ku62].

(4) Let $k = 3$. If $p/r = 2$ we basically have the situation described in (3) above. Thus let $p/r = 3$. Then $p = 3^{m+1}$ and $r = 3^m$ and we may assume that $\rho_0 = \{0\}$, $\rho_{3^m} = \{a\}$, $\rho_{2\cdot 3^m} = \{b\}$ (where $\{a, b\} = \{1, 2\}$) and $\rho_i = \emptyset$ for $0 < i < 3^{m+1}$, $i \notin \{3^m, 2 \cdot 3^m\}$. This yields the $\aleph_0$ precomplete uniform clones on **3** from [Hi78].

6 Binary Areflexive Polyrelations

In this section we consider the set Ξ_4 of minimal binary polyrelations from Ξ_3. Recall that $\iota_2 := \{aa \mid a \in \mathbf{k}\}$, and set $\sigma_2 := \mathbf{k}^2 \setminus \iota_2$. A binary relation τ is *reflexive* (*areflexive*) if $\tau \supseteq \iota_2$ ($\tau \subseteq \sigma_2$). Let ρ be a binary polyrelation from Ξ_4 with period p.

6.1 Lemma *If ρ is a minimal binary polyrelation then every ρ_i is either reflexive or areflexive.*

Proof For every $i \geq 0$ set $\tau_i := \{x \mid xx \in \rho_i\}$. As in Lemma 4.5 the unary polyrelation $\tau = (\tau_0, \tau_1, \dots)$ satisfies $\operatorname{Pol}\rho \subseteq \operatorname{Pol}\tau$. Since by minimality $\tau = (\tau_0, \tau_1, \dots)$ is not proper, we have $\tau_i \in \{\emptyset, \mathbf{k}\}$ i.e. every ρ_i is either areflexive or reflexive. □

The *relational* (or de Morgan) *product* of binary relations λ and μ on $\mathbf{k}$ is

$$\lambda \circ \mu := \{xy \mid xu \in \lambda,\ uy \in \mu \text{ for some } u\}.$$

The *converse* (or inverse) of λ is $\breve{\lambda} := \{yx \mid xy \in \lambda\}$.

Denote by R and A the sets of $\rho \in \Xi_4$ such that each ρ_i is reflexive and areflexive, respectively. We show that the binary polyrelations may be reduced to those from $R \cup A$.

6.2 Lemma *Every minimal binary polyrelation on $\mathbf{k}$ is dominated by a binary polyrelation from $R \cup A$.*

Proof Let ρ be a proper binary periodic polyrelation on $\mathbf{k}$. There is nothing to prove if $\rho \in A$. Thus let $\iota_2 \subset \rho_i \subset \mathbf{k}^2$ for some $i \geq 0$. Taking a suitable shift of ρ, by Lemma 5.4 we may assume that $\iota_2 \subset \rho_0 \subset \mathbf{k}^2$. Put $B := \{b \in \mathbf{p} \mid \rho_b \supseteq \iota_2\}$ and $C := \mathbf{p} \setminus B$. Put $\rho'_b := \rho_b$ for $b \in B$ and $\rho'_c := \mathbf{k}^2$ for $c \in C$. Notice that $0 \in B$.

Claim: ρ' dominates ρ.

Proof: For $b \in B$ put $\sigma_i^b := \rho_{b+i}$ for all $i \geq 0$ while for $c \in C$ put $\sigma_0^c := \mathbf{k}^2$ and $\sigma_i^c := \iota_2$ for all $i > 0$. This defines $\sigma^j = (\sigma_0^j, \sigma_1^j, \dots)$ for all $j \geq 0$. For $b \in B$ the polyrelation σ^b is just a shift of ρ and so $\sigma^b \in [\rho]$. For $c \in C$ the polyrelation σ^c is improper and thus $\sigma^c \in [\rho]$. We show that $\sigma_i^j \subseteq \sigma_0^{i+j}$ for all $i, j \geq 0$. If $i + j \in C$ then $\sigma_0^{i+j} = \mathbf{k}^2 \supseteq \sigma_i^j$. Thus let $i + j \in B$. Then $\sigma_0^{i+j} = \rho_{i+j+0} = \rho_{i+j}$. If also $j \in B$ then $\sigma_i^j = \rho_{i+j} = \sigma_0^{i+j}$ and we are done. Thus let $j \in C$. Then $j > 0$ and $\sigma_i^j = \iota_2 \subseteq \rho_{i+j}$ because $i + j \in B$ and so ρ_{i+j} is reflexive. Thus $\sigma_i^j \subseteq \sigma_0^{i+j}$ for all $i, j \geq 0$. The upper superposition of $(\sigma^0, \sigma^1, \dots)$ is ρ' and so $\rho' \in [\rho]$. This proves the claim. □

Clearly $\rho'_0 = \rho_0$ is proper and so $\rho' \in R$. □

Denote by A_1 and R_1 the set of all $\rho \in A$ and $\rho \in R$ with $\rho_i \neq \mathbf{k}^2$ for all $i \in \mathbf{p}$. We show that we can consider only polyrelations from $A_1 \cup R_1$.

6.3 Lemma *Every $\rho \in A \cup R$ is dominated by some $\sigma \in A_1 \cup R_1$.*

Proof We may assume ρ_0 to be nontrivial. Put $p := p_\rho$ and $D_\rho := \{i \in \mathbf{p} \mid \rho_i = \mathbf{k}^2\}$. There is nothing to prove if $D_\rho = \emptyset$, thus let $D_\rho \neq \emptyset$ and denote by d the least element of D_ρ. Put $\tau_i := \rho_{i-d} \cap \rho_i$ for all $i \in \mathbf{p}$. Clearly $\tau_d = \rho_0 \cap \rho_d = \rho_0 \cap \mathbf{k}^2 = \rho_0$ and so τ is proper. Also $\tau \in [\rho]$ and $D_\tau := \{i \in \mathbf{p}_\sigma \mid \tau_i = \mathbf{k}^2\}$ satisfies $|D_\tau| < |D_\rho|$ due to $\tau_i \subseteq \rho_i$ for all $i \in \mathbf{p}$ and $d \in D_\rho \setminus D_\tau$. Note that $\tau \in R \cup A$. Repeating this we finally arrive at σ with $D_\sigma = \emptyset$. □

6.4 In the remainder of this section we study areflexive binary polyrelations. For $\rho \in A_1$ with period p put $I_\rho := \{i \in \mathbf{p} \mid \rho_i \neq \emptyset\}$, $J_\rho := \mathbf{p} \setminus I_\rho$ and $j_\rho := |J_\rho|$.

For a map φ from a subset D of $\mathbf{k}$ into $\mathbf{k}$ put $\varphi^\circ := \{x\varphi(x) \mid x \in D\}$. Let S denote the set of all permutations of $\mathbf{k}$. Finally denote by A_2 the set of all $\rho \in A_1$ such that $\rho_i \in S^\circ = \{s^\circ \mid s \in S\}$ for all $i \in I_\rho$. We have:

6.5 Lemma *Every $\rho \in A_1 \setminus A_2$ is dominated by some $\sigma \in R_1$.*

Proof Let $\rho \in A_1 \setminus A_2$. Suppose to the contrary that ρ is not dominated by any $\sigma \in R_1$. Put $p := p_\rho$ and $I := I_\rho$. Further for all $i \in \mathbf{p}$ put $\lambda_i := \mathrm{pr}_1\, \rho_i = \{x \in \mathbf{k} \mid xu \in \rho_i \text{ for some } u\}$. According to Lemma 4.7 we have $\lambda_i = \mathbf{k}$ for all $i \in I$. Put $\tau_i := \rho_i \circ \breve{\rho}_i = \{xy \mid xu, yu \in \rho_i \text{ for some } u\}$ for all $i \geq 0$. Since τ_i is clearly reflexive for all $i \in I$ and $\tau_i = \emptyset$ otherwise, and also $\tau \in [\rho]$ clearly the polyrelation τ is improper. Thus suppose $\tau_i \in \{\iota_2, \mathbf{k}^2\}$ for all $i \in I$. Let $\tau_j = \mathbf{k}^2$ for some $j \in I$. For $n = 3, \ldots, k$ and $i = 0, \ldots, p-1$ form

$$\lambda_i^{(n)} := \{x_1 \ldots x_n \mid x_l u \in \rho_i \ (l = 1, \ldots, n) \text{ for some } u\}.$$

It is easy to see that $\lambda_j^{(3)}$ is totally reflexive. Thus $\lambda_j^{(3)} = \mathbf{k}^3$ by Theorem 4.2. An easy induction shows that $\lambda_j^{(n)} = \mathbf{k}^n$ for $n = 3, \ldots, k$. Thus $(0, \ldots, k-1) \in \lambda_j^{(k)}$, i.e., there is a u such that $xu \in \rho_j$ for all $x \in \mathbf{k}$. This is impossible because uu does not belong to the areflexive relation ρ_j. Thus $\rho_i \circ \breve{\rho}_i = \iota_2$ for all $i \in I$. The same argument applied to $\mu_i := \breve{\rho}_i \circ \rho_i = \{xy \mid ux, uy \in \rho_i \text{ for some } u\}$ yields $\breve{\rho}_i \circ \rho_i = \iota_2$ for all $i \in I$. Now it is easy to see that $\rho_i = s_i^\circ$ for some permutation s_i of $\mathbf{k}$. Thus $\rho \in A_2$ contrary to our assumption. □

Our next strategy is the following. In the first place we try to reduce the period. Among the polyrelations with minimal period we try to increase the number of empty relations.

6.6 Definition More formally, a polyrelation $\rho \in A_2$ is *strict* if each $\lambda \in A_2 \cap [\rho]$ satisfies (i) $p_\lambda \geq p_\rho$ and (ii) $j_\lambda \leq j_\rho$ whenever $p_\lambda = p_\rho$ (where $J_\rho = \{i \in \mathbf{p}_\rho \mid \rho_i = \emptyset\}$ and $j_\rho = |J_\rho|$). Denote by V_1 the set of unary polyrelations satisfying the conditions of Proposition 5.7. Denote by A_3 the set of all strict polyrelations.

The following is almost immediate.

6.7 Lemma *The system $V_1 \cup R_1 \cup A_3$ is B-generic.*

Proof According to Proposition 5.7, Lemmas 6.2, 6.3 and 6.5 the system $V_1 \cup R_1 \cup A_2$ is B-generic and so it suffices to show that each $\rho \in A_2$ is dominated by some $\sigma \in A_3$. Put $p := p_\rho$. The set L_0 of all proper $\lambda \in [\rho]$ with $p_\lambda \leq p$ is finite (as it has less than 2^{pk^2} elements). Put

$$p^* := \min\{p_\lambda \mid \lambda \in L_0\}, \qquad L_1 := \{\lambda \in L_0 \mid p_\lambda = p^*\}.$$

Next let

$$j^* := \max\{j_\lambda \mid \lambda \in L_1\}, \qquad L_2 := \{\lambda \in L_1 \mid j_\lambda = j^*\}.$$

For $\lambda_1, \lambda_2 \in L_2$ put $\lambda_1 \leq \lambda_2$ if $\operatorname{Pold}\lambda_1 \subseteq \operatorname{Pold}\lambda_2$. The relation $\leq$ is a finite quasiorder (a reflexive and transitive relation) on L_2. Letting $\lambda_1 \sim \lambda_2$ if $\lambda_1 \leq \lambda_2 \leq \lambda_1$ we get an equivalence relation $\sim$ on L_2 whose blocks $B_1, \ldots, B_m$ may be ordered by $B_i < B_j$ if $\lambda' < \lambda''$ for some $\lambda' \in B_i$ and $\lambda'' \in B_j$. This finite order has a maximal element B_l. Now any $\sigma \in B_l$ is strict, and dominates ρ. □

In the following lemmas we find the shape of polyrelations from A_3. Recall that (i) $I_\rho := \{i \in \mathbf{p}_\rho \mid \rho_i \neq \emptyset\}$, (ii) The *converse* of a binary relation α is $\breve{\alpha} := \{ab \mid ba \in \alpha\}$, and (iii) A permutation $s \in S$ is of *order* m if $s^m(x) = x$ for all $x \in \mathbf{k}$ and m is the least integer with this property.

6.8 Lemma *If $\sigma \in A_3$ then a shift ρ of σ satisfies either*

(1) *$I_\rho = \{0, r, 2r, \ldots, p - r\}$ for some proper divisor r of p_ρ and $\rho_r \neq \breve{\rho}_0$, or*

(2) *$I_\rho = \{0, i\}$ where $0 < i < p_\rho$ and $\rho_i = \breve{\rho}_0 \neq \rho_0$.*

Proof Put $p := p_\rho$, $I := I_\rho$ and $J := J_\rho$. We have two cases: (i) Suppose there are $i \in I$ and $q \in \mathbf{p} \setminus \{0\}$ such that $i \dot{+} q \in I$ and the relational product $\rho_i \circ \rho_{i+q} \neq \iota_2$ (see Lemma 6.1). Fix q to be the least possible element of $\mathbf{p} \setminus \{0\}$ with this property. Without loss of generality we may assume that the corresponding i equals 0 (see Lemma 5.4). Put $\tau_n := \rho_n \circ \rho_{n+q}$ for all $n \geq 0$. Since $\tau_0 = \rho_0 \circ \rho_q \neq \iota_2$, we get $\tau_0 = s^\circ$ for a nonidentity permutation s of $\mathbf{k}$. Thus τ_0 is nontrivial; hence τ is proper and so $\tau \in A_3 \cap [\rho]$. Clearly $p_\tau \leq p$ and, ρ being strict, also $p_\tau = p$. It is easy to see that $J_\tau \supseteq J$; hence $j_\tau \geq j$ and so $j_\tau = j$ by the strictness of ρ, proving $J_\tau = J$. Consider $l \in J$. Then $\rho_l = \emptyset$ and consequently $\tau_l = \rho_l \circ \rho_{l+q} = \emptyset$, proving $q \dot{+} l \in J$.

For a subset A of $\mathbf{p}$ and $b \in \mathbf{p}$ put $b \dot{+} A := \{b \dot{+} a \mid a \in A\}$ (where $\dot{+}$ is the sum mod p in $\mathbf{p}$). With this notation we have $q \dot{+} J \subseteq J$. We need the following easy fact.

Fact: The following are equivalent for $A \subseteq \mathbf{p}$, $b \in \mathbf{p}$ and $c := \gcd(b, p)$ (the greatest common divisor of b and p):

(1) *$b \dot{+} A \subseteq A$,*

(2) *$b \dot{+} A = A$,*

(3) *$b \dot{+} (\mathbf{p} \setminus A) = \mathbf{p} \setminus A$, and*

(4) *$c \dot{+} A = A$.*

Put $r := \gcd(q, p)$. Applying the fact, we have $r \dot{+} J = J$ and $r \dot{+} I = I$. Set

$$Q := \{0 < i < q \mid \rho_i \neq \emptyset\} = I \cap \{1, \ldots, q - 1\}.$$

We need:

Claim 1: (i) *$\rho_l = \breve{\rho}_0$ for all $l \in Q$,* (ii) *$|Q| \leq 1$.*

Proof: (i) Let $l \in Q$. By the minimality of q and in view of $\rho_0, \rho_l \in S^\circ$ we have $\rho_0 \circ \rho_l = \iota_2$. Thus $\rho_l = \breve{\rho}_0$ as required.

(ii) Suppose there are $m, n \in Q$ with $m < n$. From (i) we have $\rho_m = \rho_n = \breve{\rho}_0$. Taking into account $0 < n-m < q$ and the minimality of q we have $\breve{\rho}_0 \circ \breve{\rho}_0 = \rho_m \circ \rho_n = \rho_m \circ \rho_{m+(n-m)} = \iota_2$ and therefore the nontrivial relation ρ_0 equals s° for some $s \in S$ of order 2. Thus $\rho_0 = \breve{\rho}_0 = s^\circ$. Now $\rho_m \circ \rho_{m+(q-m)} = \rho_m \circ \rho_q = \rho_0 \circ \rho_q = \iota_2$ where $0 < q - m < q$ and this contradicts our choice of q. □

Claim 2: $Q = \emptyset$.

Proof: By Claim 1 (ii) we have $|Q| \leq 1$. Suppose to the contrary that $|Q| = 1$. From $0 \in I$ and $r \dot{+} I = I$ we get $\{0, r, \dots, p-r\} \subseteq I$. It follows that $Q = \{r\}$ and $q = 2r$. By Claim 1 (i) we have $\rho_r = \breve{\rho}_0$. Again by the minimality of q we have $\breve{\rho}_0 \circ \rho_q = \rho_r \circ \rho_{r+(q-r)} = \iota_2$ proving $\rho_q = \rho_0$. The same argument yields $\rho_{2lr} = \rho_0$, $\rho_{(2l+1)r} = \breve{\rho}_0$ for all $l \geq 0$. From $\rho_p = \rho_0$ we get p even and so $q = 2r$ divides p. Moreover ρ has period q proving $q = p$. However, this contradicts the definition of q. Thus $Q = \emptyset$. □

Now $Q = \emptyset$ means that $q = r$ divides p and we have the case (i).

(ii) Thus suppose that for all $i, j \in I$, $i \neq j$ we have $\rho_i \circ \rho_j = \iota_2$. In particular, $\rho_l = \breve{\rho}_0$ for all $l \in I \setminus \{0\}$. If $|I| = 2$ and $\breve{\rho}_0 \neq \rho_0$ we have case (ii). We show that in all the other cases we have $\rho_i = \rho_0$ for all $i \in I$. Choose $i, j \in I$, $0 < i < j$. Then $\breve{\rho}_0 \circ \breve{\rho}_0 = \rho_i \circ \rho_j = \iota_2$ and therefore $\rho_0 = \breve{\rho}_0 = \rho_i$ for all $i \in I$. Now put $\sigma^n := \rho$ for all $n \geq 0$. Clearly $\sigma^n_m \in \{\emptyset, \rho_0\}$ and consequently $\sigma^n_m \subseteq \rho_0 = \sigma_0^{m+n}$. The upper superposition of $\sigma^0, \sigma^1, \dots$ is $(\rho_0, \rho_0, \dots)$. However, $\rho \in A_3$ is not dominated by a proper polyrelation of this type. This contradiction settles the remaining cases. □

For a divisor q of k denote by S_q the set of all $s \in S$ with k/q cycles of length q (note that each $s \in S_q$ is a fixedpoint-free permutation of order q). We have:

6.9 Lemma *Let $\rho \in A_3$. Then*

(1) *For every $i \in I_\rho$ we have $\rho_i \in S^\circ_{m_i}$ for some divisor m_i of k, and*

(2) $\rho_l \neq \rho_m \Rightarrow \rho_l \cap \rho_m = \emptyset$ *for all* $0 \leq l < m < p_\rho$.

Proof Write p and I instead of p_ρ and I_ρ. (1) Consider $i \in I$. Then $\rho_i = s_i^\circ$ for some $s_i \in S$. Denote by m_i the minimum length of cycles of s_i. Recall that for a binary relation σ the power σ^n is defined recursively by setting $\sigma^0 = \iota_2$ and $\sigma^{n+1} := \sigma^n \circ \sigma$ for $n = 0, 1, \dots$. Put $\nu_n := \{x \mid xx \in \rho_n^{m_i}\}$ for all $n \geq 0$. By minimality (see 4.4) the unary polyrelation $\nu \in [\rho]$ is trivial. From $\nu_i \neq \emptyset$ we obtain $\nu_i = \mathbf{k}$, i.e. $\rho_i = s_i^\circ$ for some $s_i \in S_{m_i}$.

(2) By way of contraposition suppose $\rho_l \cap \rho_m \neq \emptyset$ for some $0 \leq l < m < p$. Put $s := m - l$ and $\mu_n := \mathrm{pr}_1(\rho_n \cap \rho_{n+s})$ for all $n \in \mathbf{p}$. Clearly $\mu_l \neq \emptyset$. Again by minimality we have $\mu_l = \mathbf{k}$. Since ρ_l, $\rho_m \in S^\circ$, we have the required $\rho_l = \rho_m$. □

6.10 Lemma *If $\rho \in A_3$ then $\rho_i = \iota_2$ for no $i \in \mathbf{p}$.*

Proof We may suppose that ρ is of the form given in Lemma 6.8. If it satisfies (2) then $\rho_0 \neq \iota_2$ (due to $\breve{\rho}_0 \neq \rho_0$) and $\rho_i \neq \iota_2$ (due to $\rho_i = \breve{\rho}_0$). Thus let (1) hold. Put $I := I_\rho$ and $p := p_\rho$. Suppose to the contrary that $L := \{l \in I \mid \rho_l = \iota_2\}$ is nonempty. Now ρ being proper, we have $L \subset I$ and so we may assume $0 \notin L$. Denote by l and g the least and greatest element of L. Put $\sigma_n := (\rho_{n+r} \circ \rho_n) \cap \rho_n$ for all $n \geq 0$. First $\sigma_{l-r} = (\rho_l \circ \rho_{l-r}) \cap \rho_{l-r} = (\iota_2 \circ \rho_{l-r}) \cap \rho_{l-r} = \rho_{l-r} \cap \rho_{l-r} = \rho_{l-r}$. Here $l - r \notin L$, hence $\rho_{l-r} \in S^\circ \setminus \{\iota_2\}$ and therefore

σ is proper. On the other hand $\sigma_g = (\rho_{g+r} \circ \rho_g) \cap \rho_g = (\rho_{g+r} \circ \iota_2) \cap \iota_2 = \rho_{g+r} \cap \iota_2$. Here $g \dot{+} r \notin L$ by the definition and $0 \notin L$ and so by Lemma 6.9 we have $\rho_{g+r} \in S^\circ_m$ for some divisor m of k, $m \neq 1$. Thus $\sigma_{g \dot{+} r} = \emptyset$. Clearly $\sigma_n = (\rho_{n+r} \circ \emptyset) \cap \emptyset = \emptyset$ if r does not divide n. Consequently $|I_\sigma| < |I|$. Since clearly also $p_\sigma \leq p_\rho$, this contradicts the strictness of ρ. □

6.11 Lemma *Let $\rho \in A_3$. Then there exists a divisor $m > 1$ of k such that $\rho_i \in S^\circ_m$ for all $i \in I_\rho$.*

Proof We know from Lemma 6.9 that for every $i \in I := I_\rho$ we have $\rho_i \in S^\circ_{m_i}$ for some divisor m_i of k. In view of Lemma 6.10 clearly $m_i > 1$. Suppose $m_i \neq m_j$ for some $0 \leq i < j < p_\rho$. Put $\sigma_n := \rho_n^{m_i}$ for all $n \geq 0$. Then $\sigma_i = \iota_2$ while $\sigma_j \in S \setminus \iota_2$. Thus σ is proper and so clearly $\sigma \in A_3$. However, this contradicts Lemma 6.10 and therefore all the m_i are equal. □

Denote by A_4 the set of all $\rho \in A_3$ for which there is a prime divisor q of k such that $\rho_i \in S^\circ_q$ for all $i \in I_\rho$. We have:

6.12 Proposition *The set $V_1 \cup A_4 \cup R_1$ is B-generic.*

Proof Let $\rho \in A_3 \setminus A_4$. By Lemma 6.11 there is a divisor $m > 1$ of k such that $\rho_i \in S^\circ_m$ for all $i \in I_\rho$. Choose a prime divisor q of m and put $m' := m/q$. For all $n \geq 0$ put $\sigma_n := \rho_n^{m'}$. It is easy to see that $I_\sigma = I_\rho$, $p_\sigma \leq p_\rho$ and that $\sigma_i \in S^\circ_q$ for all $i \in I_\sigma$. Moreover, $\sigma \in [\rho]$ and so $\sigma \in A_4$ dominates ρ. □

The following lemma relates r and p.

6.13 Lemma *Let $\rho \in A_4$ have period p and satisfy $0 \in I = I_\rho$. Then there exists a divisor r of p such that*

(1) *$p/r = p_1^{d_1} \cdots p_\ell^{d_\ell}$ for some $\ell \geq 0$, primes $p_1 < \cdots < p_\ell$ and positive integers $d_1, \ldots, d_\ell$,*

(2) *$r = p_1^{c_1} \cdots p_\ell^{c_\ell}$ for some nonnegative integers $c_1, \ldots, c_\ell$,*

(3) *$I = \{0, r, 2r, \ldots, p - r\}$,*

(4) *$\rho_0, \rho_r, \ldots, \rho_{p-r} \in S^\circ_q$ for some prime divisor q of k.*

Proof The condition (4) is just a part of the definition of A_4. Recall that $A_4 \subseteq A_3$ and so the conclusion of Lemma 6.8 holds. Let ρ satisfy (1) from Lemma 6.8. Applying Proposition 5.7 we obtain (3). Thus let ρ satisfy (2) from Lemma 6.8. Then $I = \{0, i\}$ for some $0 < i < p$ and $\rho_i = \breve{\rho}_0 \neq \rho_0$. We need the following:

Claim: We have $p = 2i$.

Proof: Suppose to the contrary that $2i \neq p$. Put $\tau_n := \rho_n \circ (\breve{\rho}_{n+i})^q$ for all $n \geq 0$. On one hand in view of $\rho_i = \breve{\rho}_0$, we have $(\breve{\rho}_i)^q = (\breve{\rho}_0)^q = \iota_2$ and therefore $\tau_0 = \rho_0 \circ \iota_2 = \rho_0$. On the other hand, $\rho_{2i} = \emptyset$ due to $2i \neq p$ and so $\tau_i = \rho_i \circ (\breve{\emptyset})^q = \emptyset$. Since $\tau_j = \emptyset \circ (\breve{\rho}_{i+j})^q = \emptyset$ for all $j \in \mathbf{p} \setminus \{0, i\}$, we have $I_\tau = \{0\}$. Since $\tau \in [\rho]$ is proper and $p_\tau = p$, this contradicts the strictness of ρ. This proves the claim. □

Now we can apply Lemma 5.6. In our case $r = i$ and $p/r = 2$, so $r = 2^b$ for some $b \geq 0$ and consequently $p = 2i = 2^{b+1}$ proving (3) in this case. □

6.14 Consider $\rho \in A_4$ with period p and $I := I_\rho = \{0, r, \dots, p-r\}$. The case $p = 2r$ is relatively simple and will be dealt with separately later. For notational simplicity we also assume $r = 1$. It is easy to see that our results for $r = 1$ can be blown up to the case $r > 1$. We have $\rho_i = s_i^\circ$ for some $s_i \in S_q$ $(i = 0, \dots, p-1)$ where q is a prime divisor of k. We start with the following:

6.15 Lemma *Let ρ be as in* 6.14. *Then*

(1) *If $i_1, \dots, i_u, j_1, \dots, j_v \in \mathbf{p}$ and $\alpha := \rho_{i_1} \circ \cdots \circ \rho_{i_u}$, $\beta := \rho_{j_1} \circ \cdots \circ \rho_{j_v}$ are not disjoint then for all $n \geq 0$,*

$$\rho_{i_1+n} \circ \cdots \circ \rho_{i_u+n} = \rho_{j_1+n} \circ \cdots \circ \rho_{j_v+n}.$$

(2) *$\rho_0, \dots, \rho_{p-1}$ are pairwise distinct.*

Proof (1) Choose $ab \in \alpha \cap \beta$. For all $n \geq 0$ set

$$\tau_n := \rho_{i_1+n} \circ \cdots \circ \rho_{i_u+n} \circ \breve{\rho}_{j_v+n} \circ \cdots \circ \breve{\rho}_{j_1+n}; \qquad \sigma_n := \mathrm{pr}_1\{xx \mid xx \in \tau_n\}.$$

Clearly $aa \in \alpha \circ \breve{\beta} = \tau_0$, whence $a \in \sigma_0$. By the minimality of ρ, the unary polyrelation σ is improper and so $\sigma_0 = \mathbf{k}$ proving $\tau_0 = \iota_2$. From Lemma 6.10 we obtain that τ is improper. In view of $\tau_n \in S^\circ$ clearly $\tau_n = \iota_2$ for all $n \geq 0$ and (1) follows from the definition of τ_n.

(2) Suppose to the contrary that $s_i = s_j$ for some $0 \leq i < j < p$. By (1) clearly $\rho_{i \dotplus n} = \rho_{j \dotplus n}$ for all $n \geq 0$. It is easy to see that then ρ is periodic with period $d = \gcd(p, j-i) < p$, a contradiction. □

6.16 Put $T_\rho := \{s_0, \dots, s_{p-1}\}$. For $s, t \in S$ we define the product st by setting $(st)(x) := t(s(x))$.

Denote by G_ρ the subgroup $[T_\rho]$ of S generated by T_ρ. In other words, G_ρ is the set of all products $s_{i_1} \cdots s_{i_m}$ with $m > 0$ and $i_1, \dots, i_m \in \mathbf{p}$ (this is indeed a group because $s_i^{-1} = s_i^{q-1}$ for all $i \in \mathbf{p}$). Denote by e the identity permutation on $\mathbf{k}$. Further set $s'_i := s_{i \dotplus 1}$ for all $i \in \mathbf{p}$ and define a binary relation φ_ρ on G_ρ by setting

$$\varphi_\rho := \{(s_{i_1} \cdots s_{i_u}, s'_{i_1} \cdots s'_{i_u}) \mid u > 0, \quad i_1, \dots, i_u \in \mathbf{p}\}.$$

We shall often use the following lemma.

6.17 Lemma *If ρ is as in* 6.14 *then:*

(1) *$T_\rho \subseteq S_q$ for some prime divisor q of k and $G_\rho \subseteq \{e\} \cup \bigcup\{S_m \mid m > 1$ divides $k\}$.*

(2) *φ_ρ is the diagram (graph) of an automorphism f_ρ of the permutation group G_ρ.*

Proof (1) The first statement is just Lemma 6.13 (4). For the second one let $g = s_{i_1} \cdots s_{i_u} \in G_\rho \setminus \{e\}$. Put $\tau_n := \rho_{n+i_1} \circ \cdots \circ \rho_{n+i_u}$ for all $n \in \mathbf{p}$. Now $\tau_0 = s^\circ_{i_1} \circ \cdots \circ s^\circ_{i_u} = g^\circ \neq \iota_2$ and so τ is proper. Since $\tau \in A_3$, from Lemma 6.9 (1) we see that $\tau_0 \in S^\circ_m$ for some divisor m of k proving $g \in S_m$.

(2) First we prove:

Claim 1: The relation φ_ρ is the diagram of a selfmap f_ρ of G_ρ.

Proof: Let $s_{i_1} \cdots s_{i_u} = s_{j_1} \cdots s_{j_v}$ for some $i_1, \ldots, i_u, j_1, \ldots, j_v \in \mathbf{p}$. For $n = 1$ Lemma 6.15 (1) yields $s'_{i_1} \cdots s'_{i_u} = s'_{j_1} \cdots s'_{j_v}$. □

Claim 2: The map f_ρ is injective.

Proof: Let $s'_{i_1} \cdots s'_{i_u} = s'_{j_1} \cdots s'_{j_v}$. For $n = p - 1$ Lemma 6.15 (1) yields $s_{i_1} \cdots s_{i_u} = s_{j_1} \cdots s_{j_v}$. □

A direct verification shows that f_ρ is an endomorphism of G_ρ, i.e. $f_\rho(ab) = f_\rho(a)f_\rho(b)$ for all $a, b \in G_\rho$. Combining this with Claims 1 and 2 we obtain (2). □

6.18 We write $H \leq G$ to indicate that H is a subgroup of a group G. For a selfmap f of G, a subgroup H is *f-closed* if $f(H) \subseteq H$. Denote by A_5 the set of all $\rho \in A_4$ such that $\{e\}$ and G_ρ are the only f_ρ-closed subgroups of G_ρ. We have:

6.19 Lemma *Every $\rho \in A_4$ is dominated by some $\sigma \in A_5$.*

Proof Choose an inclusion-minimal member H of

$$\{X \leq G_\rho \mid X > \{e\} \text{ is } f_\rho\text{-closed}\}$$

and let $s \in H \setminus \{e\}$. By Lemma 6.17 (1) we have $s \in S_m$ for some divisor m of k with $m > 1$. Choose a prime divisor n of m and set $t := m/n$. As s consists of k/m cycles of length m, clearly $s^* := s^t$ has t cycles of length n and so $s^* \in (H \setminus \{e\}) \cap S_n$. In view of $s^* \in (H \setminus \{e\}) \subseteq G_\rho \setminus \{e\}$ we have that $s^* = s_{i_1} \cdots s_{i_u}$ for some $u > 0$ and $i_1, \ldots, i_u \in \mathbf{p}$. Set $\sigma_n := \rho_{n+i_1} \circ \cdots \circ \rho_{n+i_u}$ for all $n \geq 0$. Clearly $\sigma_0 = s^{*\circ}$, hence σ is proper and $\sigma \in [\rho] \cap A_4$ dominates ρ. Set $T^* := \{f^n_\rho(s^*) \mid n \in \mathbf{p}\}$ and notice that the permutation group G_σ is generated by T^*. Next G_σ is a subgroup of H due to $s^* \in H$, the definition of T^* and the fact that H is f_ρ-closed. From the definition of σ it follows directly that f_σ is the restriction of f_ρ to G_σ. Now G_σ is f_σ-closed by Lemma 6.17 (2) and so G_σ is f_ρ-closed. Finally by the minimality of H we have $G_\sigma = H$ and therefore $\sigma \in A_5$. □

6.20 Recall that for a prime q an *elementary Abelian q-group* G is an Abelian group $\langle G; +, -, 0\rangle$ in which every $g \in G \setminus \{0\}$ is of order q. It is well known that a finite elementary q-group is isomorphic to the additive group of a finite vector space over the Galois field $\mathrm{GF}(q)$ (also denoted $\mathbf{F}_q$). This can be described more explicitly as follows.

Let $m > 0$. For $a = (a_1, \ldots, a_m) \in \mathbf{q}^m, b = (b_1, \ldots b_m) \in \mathbf{q}^m$ put $a+b := (a_1 \oplus b_1, \ldots, a_m \oplus b_m)$ whereby $\oplus$ denotes the mod q sum on $\mathbf{q} := \{0, \ldots, q-1\}$. Now a finite Abelian elementary q-group G is isomorphic to some $\langle \mathbf{q}^m; +, \tilde{0}\rangle$ where $\tilde{0} := (0, \ldots, 0)$; in particular $|G| = q^m$ for some $m > 0$.

For the following proposition we are indebted to P. Frankl and P. P. Pálfy (personal communications 1983 and 1998).

6.21 Proposition *Let $\rho \in A_5$ and let q be a prime divisor of k such that $\rho_i \in S^\circ_q$ for all $i \geq 0$. Then G_ρ is an elementary Abelian q-group.*

Proof

Claim: G_ρ is Abelian.

Proof: Suppose to the contrary that G_ρ is noncommutative. We show:

Fact: The automorphism f_ρ is fixed-point free (i.e., the identity e is the unique fixed-point of f_ρ).

Proof: Suppose to the contrary that $f_\rho(g) = g$ for some $g \in G_\rho \setminus \{e\}$. Clearly f_ρ fixes the cyclic subgroup H of G_ρ generated by g. From Lemma 6.19 we get $H = G_\rho$, a contradiction since H is abelian. This proves the fact. □

Applying a well-known theorem (see e.g. [Go68, Theorem 10.1.2 p. 335]) we obtain that f_ρ fixes a unique Sylow q-subgroup H of G_ρ and therefore again G_ρ is a Sylow q-group. Now a q-group is nilpotent (see e.g. [Go68, Theorem 2.3.3 p. 22]) and hence $L = [G_\rho, G_\rho]$ is a proper subgroup of G_ρ. Here f_ρ fixes L, a contradiction. Thus G_ρ is abelian. □

Now the generators $s_0, \ldots, s_{p-1}$ of G_ρ belong to S_q and hence they are all of order q and so in the Abelian group G_ρ every element is of order q. Thus G_ρ is an elementary Abelian q-group. □

The remainder of this section is essentially due to P. Frankl (personal communication 1983).

6.22 By Proposition 6.21 the group G_ρ is isomorphic to the additive group of an m-dimensional vector space V_m over $GF(q)$ (whose universe is $\mathbf{q}^m$). Denote by $\varphi : g \mapsto \hat{g}$ this isomorphism from G_ρ onto V_m and for all $j \in \mathbf{n}$ set $x_j := \hat{s_j}$. We need the following.

6.23 Fact *Let Y be an $m \times m$ matrix over $\mathbf{q}$ and $e_1 := (1, 0, \ldots, 0) \in \mathbf{q}^m$. If Y is nonsingular over $GF(q)$ then $e_1 Y^h = e_1$ for some $h > 0$.*

Proof The members of the sequence $e_1 Y, e_1 Y^2, \ldots$ belong to the finite set $\mathbf{q}^m$ and so there is a repetition in the sequence; i.e., $e_1 Y^{i+h} = e_1 Y^i$ for some $i \geq 0$ and $h > 0$. Now, Y being nonsingular, we have $e_1 Y^h = e_1$. □

6.24 For $\tilde{\alpha} = (\alpha_0, \ldots, \alpha_{m-1}) \in \mathbf{q}^m$ denote by $Y_{\tilde{\alpha}}$ the $m \times m$ matrix

$$Y_{\tilde{\alpha}} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ \alpha_0 & \alpha_1 & \alpha_2 & \cdots & \alpha_{m-2} & \alpha_{m-1} \end{bmatrix}.$$

As $\det Y_{\tilde{\alpha}} = (-1)^{m-1}\alpha_0$, clearly $Y_{\tilde{\alpha}}$ is nonsingular if and only if $\alpha_0 \neq 0$. Now we have:

6.25 Lemma *Let $\rho \in A_5$ have period p and let q be a prime divisor of k such that $\rho_i \in S_q^\circ$ for all $i \in \mathbf{p}$. Then either*

(1) *$q > 2$ and there exists $1 < \alpha < q$ such that*

   (a) *p is the least positive integer solution h of $\alpha^h \equiv 1 \pmod q$ and*

   (b) *$x_j = \alpha^j x_0$ for all $j \in \mathbf{p}$, or*

(2) *There exist $1 < m < p$ and $\tilde{\alpha} = (\alpha_0, \ldots, \alpha_{m-1}) \in \mathbf{q}^m$ such that*

(a) $\alpha_0 \neq 0$,

(b) *p is the least positive integer solution h of $e_1 Y_{\tilde{\alpha}}^h = e_1$, and*

(c) *the map $s_j \mapsto e_1 Y_{\tilde{\alpha}}^j$ from $\{s_j \mid j \in \mathbf{p}\}$ to V_m extends to an isomorphism from G_ρ onto V_m.*

Proof Put $x_j := \hat{s_j}$ for all $j \in \mathbf{p}$. We start with the following:

Claim: $x_0 + \cdots + x_{p-1} = 0$.

Proof: Put $\tau_n := \rho_n \circ \rho_{n+1} \circ \cdots \circ \rho_{n+p-1}$ for all $n \in \mathbf{p}$. On account of G_ρ Abelian we have $\tau_n = \tau_0$ for all $n \in \mathbf{p}$. Were $\tau_0 \neq \iota_2$, the proper polyrelation $(\tau_0, \tau_0, \dots)$ would dominate ρ. This being impossible, $\tau_0 = \iota_2$ proving the claim. □

Now there exists a least $0 < d < p$ for which

$$\alpha_i x_i + \cdots + \alpha_{i \dotplus d} x_{i \dotplus d} = 0 \tag{6.1}$$

holds for some $i \in \mathbf{p}$ and $\alpha_i, \dots, \alpha_{i+d} \in \mathbf{p}$ with $\alpha_i \neq 0 \neq \alpha_{i+d}$ since by the claim the equation (6.4) holds for $i = 0$, $d = p - 1$ and $\alpha_0 = \cdots = \alpha_{p-1} = 1$. Taking an appropriate shift of ρ, without loss of generality we may assume that $i = 0$. Also multiplying (6.1) by the inverse of $-\alpha_d$ in the field $GF(q)$ we get $\alpha_d = -1$. Thus

$$\alpha_0 x_0 + \cdots + \alpha_{d-1} x_{d-1} - x_d = 0 \tag{6.2}$$

where $\alpha_0 \neq 0$. The automorphism f_ρ transforms (6.2) into

$$\alpha_0 x_j + \cdots + \alpha_{d-1} x_{j \dotplus (d-1)} - x_{j \dotplus d} = 0 \tag{6.3j}$$

for all $j \in \mathbf{p}$. From (6.2)

$$x_d = \alpha_0 x_0 + \cdots + \alpha_{d-1} x_{d-1}. \tag{6.4}$$

Denote by H the subspace of V_m generated by $x_0, \dots, x_{d-1}$. From (6.4) we have $x_d \in H$. An easy induction on $j = 0, \dots, p - d - 1$ based in (6.3j) shows that $x_d, \dots, x_{p-1} \in H$. Clearly V_m is generated by $x_0, \dots, x_{p-1}$ because G_ρ is generated by $s_0, \dots, s_{p-1}$; therefore $H = V_m$.

Moreover, our choice of d guarantees that the set $B := \{x_0, \dots, x_{d-1}\}$ is independent (in the vector space V_m; i.e., a linear combination of $x_0, \dots, x_{d-1}$ equals 0 if and only if all coefficients are 0) and so B is a basis of V_m. In particular, $d = m$. We have two cases.

(i) Let $m = 1$. From (6.3j) we obtain $x_j = (\alpha_0)^j x_0$ for all $j \in \mathbf{p}$. Clearly $\alpha_0 \neq 1$ and p is the least integer h such that $\alpha_0^h \equiv 1 \pmod{q}$ (the so-called index of α_0).

(ii) Let $m > 1$. Put

$$e_1 := (1, 0, \dots, 0), \dots, e_m := (0, \dots, 0, 1). \tag{6.5}$$

As B is a basis of V_m, the above isomorphism $\varphi : G_\rho \to V_m$ can be chosen so that $x_i = e_{i+1}$ for $i = 0, \dots, m - 1$. The automorphism f_ρ of G_ρ becomes a linear transformation of V_m. It is easy to check that this linear transformation sends x to $xY_{\tilde{\alpha}}$ for all $x \in \mathbf{q}^m$ (with the matrix product, and $\tilde{\alpha} = (\alpha_0, \dots, \alpha_{m-1})$ whereby x is considered as a $1 \times m$ matrix) and $Y_{\tilde{\alpha}}$ from 6.24. A direct check shows that indeed $e_j Y_{\tilde{\alpha}} = e_{j+1}$ for $j = 1, \dots, m-1$ and that (6.3j) holds for all $j \in \mathbf{p}$. Moreover, $x_j = e_1 Y_{\tilde{\alpha}}^j$ for all $j \in \mathbf{p}$ and so p is indeed the least positive integer h satisfying $e_1 Y_{\tilde{\alpha}}^h = e_1$. □

α_0	$x_0, \dots, x_{p-1}$
2	1, 2, 4, 3
3	1, 3, 4, 2
4	1, 4

Table 1: $m = 1$ and $q = 5$

6.26 Examples

(1) Let $m = 1$ and $q = 5$. All three sequences $1, x_1, \dots, x_{p-1}$ are given in Table 1.

(2) Let $m = 2$ and $q = 3$. We choose $x_0 = (1, 0)$ and $x_1 = (0, 1)$. We have six matrices

$$Y_{\tilde{\alpha}} = \begin{bmatrix} 0 & 1 \\ \alpha_0 & \alpha_1 \end{bmatrix}$$

with $\alpha_0 \in \{1, 2\}$ and $\alpha_1 \in \mathbf{3}$. The calculation of the powers of $Y_{\tilde{\alpha}}$ yields the values of p given in Table 2.

α_0	α_1	p
α_0	0	$2\alpha_0$
1	$\neq 0$	8
2	$\neq 0$	6

Table 2: $m = 2$ and $q = 3$

For example, consider the second case $\alpha_0 = 1$, $\alpha := \alpha_1 \neq 0$. Taking into account $\alpha^2 \equiv 1 \pmod 3$ the consecutive powers of $Y_{\tilde{\alpha}}$ are

$$Y_{\tilde{\alpha}}, \begin{bmatrix} 1 & \alpha \\ \alpha & 2 \end{bmatrix}, \quad \begin{bmatrix} \alpha & 2 \\ 2 & 0 \end{bmatrix}, \quad \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, \quad \begin{bmatrix} 0 & 2 \\ 2 & \alpha \end{bmatrix}, \quad \begin{bmatrix} 2 & 2\alpha \\ 2\alpha & 1 \end{bmatrix}, \quad \begin{bmatrix} 2\alpha & 1 \\ 1 & 0 \end{bmatrix}, \quad \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

The corresponding sequence $x_0, \dots, x_7$ is

$$(1, 0), \quad (0, 1), \quad (1, \alpha), \quad (\alpha, 2), \quad (2, 0), \quad (0, 2), \quad (2, 2\alpha), \quad (\alpha, 0).$$

We describe concretely the permutations s_j.

6.27 (1) Consider case (1) of Lemma 6.25. Then $s_0 \in S_q$ is arbitrary and $s_j = s_0^{\alpha^j}$ for all $j \in \mathbf{p}$.

(2) Consider case (2) of Lemma 6.25. We need:

Fact: If $0 \leq i < j < m$ and C and C' are cycles of s_i and s_j then $|C \cap C'| \leq 1$.

Proof: Let $a, b \in C \cap C'$. Then $s_i^\ell(a) = b = s_j^n(a)$ for some $\ell, n \in \mathbf{q}$. Thus $\rho_i^\ell \cap \rho_j^n$ is non-void and from Lemma 6.15 (1) we obtain $\rho_i^\ell = \rho_j^n$. Thus $s_i^\ell = s_j^n$ and this translates

into $\ell e_i = ne_j$. Recalling from the proof of Lemma 6.25 (2) that $\{e_0, \dots, e_{m-1}\}$ is a basis of the vector space V_m we obtain that the set $\{e_i, e_j\}$ is linearly independent (over $GF(q)$). Thus $\ell = n = 0$ and $b = s_i^0(a) = a$. This proves the fact. □

Recall that $s_0, \dots, s_{m-1}$ permute pairwise and that each s_i is fixed-point free and all its cycles are of length q. This and the fact shows that $\mathbf{k}$ is the disjoint union of blocks of size q^m and thus q^m divides k. Setting $l := kq^{-m}$ we can identify $\mathbf{k}$ with $\mathbf{l} \times \mathbf{q}^m$. Now for $(u, v) \in \mathbf{l} \times \mathbf{q}^m$ and $j \in \mathbf{m}$ we put $s_j(u, v) := (u, v \oplus e_{j+1})$ (where $\oplus$ denotes addition on V_m, i.e., the componentwise mod q addition on $\mathbf{q}^m$) and $e_1 := (1, 0, \dots, 0), \dots, e_{m-1} := (0, \dots, 1, 0), e_m := e_1$. Since $s_0, \dots, s_{m-1}$ generate G_ρ, each $s \in G_\rho$ respects the above blocks; in particular this holds for $s_m, \dots, s_{p-1}$. For $m \leq j < p$ and $e_1 Y_\alpha^j = (c_0, \dots, c_{m-1})$ we can set $s_j := s_0^{c_0} \cdots s_{m-1}^{c_{m-1}}$.

7 Order polyrelations

7.1 Sections 7 and 8 treat the remaining case of reflexive binary polyrelations from R_1. The set R_1 from Lemma 6.3 consists of proper binary periodic polyrelations $\rho = (\rho_0, \rho_1, \dots)$ such that $\rho_i = \mathbf{k}^2$ for no $i \geq 0$ and every $\rho_i \neq \emptyset$ is reflexive. First we eliminate $\rho_i = \emptyset$. Denote by R_2 the set of all proper binary periodic polyrelations $\rho = (\rho_0, \rho_1, \dots)$ with $\iota_2 \subseteq \rho_i \subset \mathbf{k}^2$ for all $i \geq 0$. We have:

7.2 Lemma *Every $\rho \in R_1$ is dominated by some $\sigma \in R_2$.*

Proof Let $\rho \in R_1$. Since ρ is proper, through an appropriate shift we can get $\iota_2 \subset \rho_0 \subset \mathbf{k}^2$. Put $\lambda := (\iota_2, \iota_2, \dots)$. For $j \geq 0$ put $\tau^j := (\rho_j, \rho_{j+1}, \dots)$ if $\rho_j \neq \emptyset$ and $\tau^j := \lambda$ if $\rho_j = \emptyset$. We show $\tau_i^j \subseteq \tau_0^{i+j}$ for all $i, j \geq 0$. Indeed, if $\rho^j \neq \emptyset$ then $\tau_i^j = \rho_{j+i} = \tau_0^{i+j}$ while for $\rho^j = \emptyset$ we get $\tau_i^j = \iota_2 \subseteq \tau_0^{i+j}$ because $\tau_0^{i+j} = \iota_2$ if $\rho_{i+j} = \emptyset$ and $\tau_0^{i+j} = \rho_{i+j} \supseteq \iota_2$ if $\rho_{i+j} \neq \emptyset$. Moreover, $\tau^j \in [\rho]$ for all $j \geq 0$. Indeed, if $\rho_j \neq \emptyset$ then $\operatorname{Pold} \tau^j = \operatorname{Pold} \rho$ while $\operatorname{Pold} \tau^j = \operatorname{Pold} \lambda = \mathcal{U}$ if $\rho_j = \emptyset$. Thus the upper superposition σ of $\tau^0, \tau^1, \dots$ dominates ρ. It is easy to see that $\sigma_i = \rho_i$ if $\rho_i \neq \emptyset$ and $\sigma_i = \iota_2$ if $\rho_i = \emptyset$. Thus σ belongs to R_2 and dominates ρ. □

A binary relation σ is *symmetric*, *antisymmetric* or *transitive* if $\sigma = \breve{\sigma}$, $\sigma \cap \breve{\sigma} = \iota_2$ and $\sigma^2 \subseteq \sigma$, respectively. A reflexive, antisymmetric and transitive relation is an *order*. A binary polyrelation $\lambda = (\lambda_0, \lambda_1, \dots)$ is *symmetric*, *antisymmetric*, or *transitive*, if all λ_i are reflexive as well as symmetric, antisymmetric or transitive, respectively. Denote by M the set of all binary, symmetric, proper and periodic polyrelations on $\mathbf{k}$ (i.e., proper $\rho = (\rho_0, \rho_1, \dots)$ with $\iota_2 \subseteq \rho_i \subseteq \mathbf{k}^2$ and $\rho_i = \breve{\rho}_i$ for all $i \geq 0$; note that here we allow also $\rho_i = \mathbf{k}^2$). Further denote by P the set of binary antisymmetric proper polyrelations on $\mathbf{k}$. We have:

7.3 Lemma *Every $\rho \in R_2$ is dominated by a polyrelation from $M \cup P$.*

Proof Let $\rho \in R_2 \setminus P$. Then $\rho_i \cap \breve{\rho}_i \supset \iota_2$ for some $i \geq 0$. Put $\tau_n := \rho_n \cap \breve{\rho}_n$ for all $n \geq 0$. Now $\iota_2 \subset \tau_i \subseteq \rho_i \subset \mathbf{k}^2$ and so τ is proper. Clearly τ is symmetric and so $\tau \in M$ dominates ρ. □

7.4 Let ξ be a binary reflexive and antisymmetric relation on $\mathbf{k}$. Call $x \in \mathbf{k}$ a *least* (*greatest*) element of ξ if $x \times \mathbf{k} \subseteq \xi$ ($\mathbf{k} \times x \subseteq \xi$). Notice that if ξ has a least (greatest) element then it is unique. Call ξ *bounded* if it has both a least and a greatest element.

Denote by P_1 the set of all $\rho \in P$ such that every $\rho_n \supset \iota_2$ is bounded. We have:

7.5 Lemma *Let $\sigma \in P$ satisfy $[\sigma] \cap M = \emptyset$. Then every proper binary $\rho \in [\sigma]$ belongs to P_1.*

Proof Put $p := p_\rho$ and $I := \{i \in \mathbf{p} \mid \rho_i \supset \iota_2\}$. Form $\tau_i := \rho_i \circ \breve{\rho}_i$ for all $i \geq 0$. Now $\tau \in [\sigma]$ is obviously reflexive and symmetric and $\rho_i \subseteq \tau_i$ by reflexivity. If $\tau_i \subset \mathbf{k}^2$ for at least one $i \in I$ then τ is proper in contradiction to $[\sigma] \cap M = \emptyset$. Thus $\tau_i = \mathbf{k}^2$ for all $i \in I$. For $n \geq 2$ put

$$\lambda_i^{(n)} := \{x_1 \dots x_n \mid x_1 u, \dots, x_n u \in \rho_i \text{ for some } u \in \mathbf{k}\}.$$

Now $\lambda_i^{(2)} = \tau_i = \mathbf{k}^2$, and by Theorem 4.2 an easy induction shows $\lambda_i^{(n)} = \mathbf{k}^n$ for $n = 2, \dots, k$. In particular, $\lambda_i^{(k)} = \mathbf{k}^k$, i.e., for each $i \in I$ the relation ρ_i has a greatest element l_i. It can be easily verified that for every binary polyrelation $\sigma = (\sigma_0, \sigma_1, \dots)$ the converse polyrelation $\breve{\sigma} = (\breve{\sigma}_0, \breve{\sigma}_1, \dots)$ satisfies $\operatorname{Pol} \sigma = \operatorname{Pol} \breve{\sigma}$. The same argument applied to $\breve{\rho} := (\breve{\rho}_0, \breve{\rho}_1, \dots)$ yields that for each $i \in I$ the relation ρ_i has a least element o_i. □

Denote by P_2 the set of all $\rho \in P_1$ such that every $\rho_n \supset \iota_2$ is a bounded order. We have:

7.6 Lemma *Every $\rho \in P$ is dominated by a polyrelation from $M \cup P_2$.*

Proof There is nothing to prove if $[\rho] \cap M \neq \emptyset$. Thus assume $[\rho] \cap M = \emptyset$. Then $\rho \in P_1$ by Lemma 7.5. Put $\tau_n := \rho_n^k$ ($= \rho_n \circ \dots \circ \rho_n$ with k factors) for all $n \geq 0$. Let $i \in I := \{j \in \mathbf{p}_\rho \mid \rho_j \supset \iota_2\}$ and denote by l_i the greatest element of ρ_i. Now $\tau_i \neq \mathbf{k}^2$ because $l_i x \in \tau_i$ holds only for $x = l_i$. Thus $\rho_i \subseteq \tau_i \subset \mathbf{k}^2$ proving that τ is proper. Note that all τ_i are transitive. For all $i \geq 0$ put $\xi_i := \tau_i \cap \breve{\tau}_i$. Were $\iota_2 \subset \xi_i$ for at least one $i \in I$ we would have $\xi \in [\rho] \cap M$. Thus τ_i is also antisymmetric; hence τ_i is a bounded order for all $i \in I$ proving $\tau \in P_2$. □

7.7 Next we pick an optimal polyrelation $\rho \in P_2 \setminus M$ (see Proposition 5.7). Formally, denote by P_3 the set of all $\rho \in P_2$ such that

(1) $[\rho] \cap M = \emptyset$,

(2) For every $\lambda = (\lambda_0, \lambda_1, \dots) \in [\rho] \cap P_2$,

$$p_\lambda \geq p := p_\rho \quad \text{and} \quad p_\lambda = p \implies |\lambda_0| + \dots + |\lambda_{p-1}| \geq |\rho_0| + \dots + |\rho_{p-1}|.$$

On account of periodicity and finiteness we have the following:

7.8 Lemma *Every $\rho \in P_2$ is dominated by some $\sigma \in P_3$.*

In Lemmas 7.9 and 7.10 we completely describe P_3. In Lemmas 7.9 and 7.10 ρ is always a fixed polyrelation from P_3 with period p and

$$c := |\rho_0| + \dots + |\rho_{p-1}|, \qquad I := \{i \in \mathbf{p} \mid \rho_i \supset \iota_2\}.$$

7.9 Lemma *We have $\rho_i \cap \rho_{i+d} = \iota_2$ for all $0 < d < p$.*

Proof Suppose $\rho_j \cap \rho_{j+d} \supset \iota_2$ for some $j \in \mathbf{p}$ and $0 < d < p$. Form $\tau_i = \rho_i \cap \rho_{i+d}$ for all $i \geq 0$. By Lemma 7.5 and $[\rho] \cap M = \emptyset$ we have $\tau \in P_1$. Now for all $i \geq 0$ the relation τ_i is the intersection of two orders and therefore an order. Since $\iota_2 \subset \tau_j \subseteq \rho_j \subset \mathbf{k}^2$, the polyrelation τ is proper and so $\tau \in P_2$. Clearly $p_\tau \leq p$ and in view of $\rho \in P_3$, also $p \leq p_\tau$. Thus $p_\tau = p$ and from the definition of P_3 we have $t := |\tau_0| + \cdots + |\tau_{p-1}| \geq |\rho_0| + \cdots + |\rho_{p-1}| =: c$.

Now $\rho_i \supseteq \tau_i$ for all $i \in \mathbf{p}$ and so $c \geq t$ proving $c = t$. This together with $\rho_i \supseteq \tau_i$ shows $\rho_i = \tau_i$ for all $i \in \mathbf{p}$. This yields $\rho_i \subseteq \rho_{i+d}$ for all $i \in \mathbf{p}$. Consequently $\rho_i \subseteq \rho_{i+d} \subseteq \cdots \subseteq \rho_{i+pd} = \rho_i$ and therefore $\rho_i = \rho_{i+d}$ for all $i = 0, \ldots, p-1$. However, then ρ has period d strictly less than p. This contradiction proves the lemma. □

7.10 Lemma *We have $p = 2r$, $I = \{0, r\}$, and $\rho_r = \breve{\rho}_0$.*

Proof We may assume that $0 \in I$. We start with:

Claim 1: $|I| > 1$.

Proof: Suppose $|I| = 1$. Then $I = \{0\}$. For all $n \geq 0$ set $\tau_n := \rho_n \circ \rho_{n+1} \circ \cdots \circ \rho_{n+p-1}$. It is easy to see that $\tau_n = \rho_0$ for all $n \geq 0$ and so τ is dominated by $(\rho_0, \rho_0, \ldots)$, a contradiction. □

Denote by r the least positive integer such that $j \dotplus r \in I$ for some $j \in I$.

Claim 2: $\rho_n \subseteq \breve{\rho}_{n+r}$ *for all* $n \in \mathbf{p}$.

Proof: For $i \in I$ denote by o_i and l_i the least and greatest elements of ρ_i. Let $j \in I$ satisfy $j \dotplus r \in I$. Notice that the ordered pairs $o_j l_{j+r}$ and $o_{j+r} l_j$ belong to $\rho_j \cap \rho_{j+r}$. As $\rho_j \cap \rho_{j+r} = \iota_2$ from Lemma 7.9, we obtain $o_j = l_{j+r}$ and $l_j = o_{j+r}$. Now put $\tau_n := \rho_n \cap \breve{\rho}_{n+r}$ for all $n \geq 0$. Now $o_j l_j = l_{j+r} o_{j+r} \in \tau_j$ proving that τ is proper. The argument used in the proof of Lemma 7.9 yields $\tau_n = \rho_n$ for all $n \in \mathbf{p}$. Finally $\tau_n = \rho_n \cap \breve{\rho}_{n+r} = \rho_n$ shows that $\rho_n \subseteq \breve{\rho}_{n+r}$. □

In particular, we have $\rho_0 \subseteq \breve{\rho}_r \subseteq \rho_{2r}$. Now p divides $2r$ because otherwise by Lemma 7.9 we would have $\rho_0 = \rho_0 \cap \rho_{2r} = \iota_2$ in contradiction to $0 \in I$. In view of $0 < r < p$ we have $p = 2r$. The minimality of r and $0 \in I$ show that $I = \{0, r\}$. Finally by Claim 2 we have $\rho_r \subseteq \breve{\rho}_0 \subseteq \rho_r$ proving $\rho_r = \breve{\rho}_0$. □

Now we have:

7.11 Theorem

(1) *The set P_3 consists of ρ with period $2^m > 1$ and such that ρ_0 is a bounded order, $\rho_{2^{m-1}}$ its converse and $\rho_i = \iota_2$ otherwise.*

(2) *$\Xi_5 = V_1 \cup A_5 \cup P_3 \cup M$ is B-generic.*

Proof (1) By Lemma 7.10 we have $p = 2r$ and from Lemma 5.6 (2) $r = 2^b$ for some $b \geq 0$. Now (1) follows from Lemma 7.10, and (2) follows from Lemma 6.7, Proposition 6.21, Lemmas 7.2, 7.3, 7.6 and 7.8. □

8 Reflexive Symmetric Polyrelations

8.1 In this section we study the set M of symmetric binary proper polyrelations. We need the following terminology.

Let σ be a binary reflexive and symmetric relation on $\mathbf{k}$. A subset C of $\mathbf{k}$ is a *clique* of σ if $C^2 \subseteq \sigma$. The *center* C_σ of σ is the set $\{x \in \mathbf{k} \mid x \times \mathbf{k} \subseteq \rho\}$. If $\emptyset \neq C_\sigma \subset \mathbf{k}$ the relation σ is called *central*.

Given binary reflexive relations σ_1 and σ_2 define

$$\sigma_1 * \sigma_2 := \{xy \mid xu, yv \in \sigma_1 \text{ and } uy, vx \in \sigma_2 \text{ for some } u, v \in \mathbf{k}\}.$$

It is easy to see that $\sigma_1 * \sigma_2$ is always a symmetric relation and $\sigma_1 \cup \sigma_2 \subseteq \sigma_1 * \sigma_2$ and if σ_1 and σ_2 are symmetric then $\sigma_1 * \sigma_2 = \sigma_2 * \sigma_1$. A binary polyrelation $\rho = (\rho_0, \rho_1, \dots) \in M$ is *central* if every $\rho_i \subset \mathbf{k}_2$ is central.

As in the preceding sections we choose ρ with firstly the least possible period and secondly the largest sum of relation sizes. Formally, polyrelation $\rho = (\rho_0, \rho_1, \dots) \in M$ with period p is *rich* if every $\lambda \in [\rho] \cap M$ satisfies (i) $p_\lambda \geq p$ and (ii) $|\lambda_0| + \cdots + |\lambda_{p-1}| \leq |\rho_0| + \cdots + |\rho_{p-1}|$ whenever $p_\lambda = p$. Denote by M_1 the set of rich polyrelations. A standard argument (based on finiteness) shows:

8.2 Lemma *Every polyrelation from M is dominated by some polyrelation from M_1.*

The following lemma will be used several times.

8.3 Lemma *Let $\rho \in M_1$. If $\lambda \in [\rho] \cap M$ satisfies $p_\lambda \leq p_\rho$ and $\lambda_i \supseteq \rho_i$ for all $i \in \mathbf{p}$ then $\lambda = \rho$.*

Proof First $p_\lambda = p := p_\rho$ by the definition of richness. Let $c := |\rho_0| + \cdots + |\rho_{p-1}|$ and $d := |\lambda_0| + \cdots + |\lambda_{p-1}|$. Clearly $d \geq c$ on account of $\lambda_i \supseteq \rho_i$ for all $i \in \mathbf{p}$ and $c \geq d$ by the richness of ρ and so $c = d$. Using again $\lambda_i \supseteq \rho_i$ for all $i \in \mathbf{p}$ we obtain $|\lambda_i| = |\rho_i|$ and by the same token we have $\lambda_i = \rho_i$ for all $i \in \mathbf{p}$ proving $\lambda = \rho$. □

In 8.4–8.7 $\rho = (\rho_0, \rho_1, \dots)$ is a fixed polyrelation from M_1 with period p, and such that $\iota_2 \subset \rho_0 \subset \mathbf{k}^2$. For $I \subseteq \mathbf{p}$ put $\rho_I := \bigcap_{i \in I} \rho_i$. Further let $\mathbf{J}$ consist of the nonempty subsets I of $\mathbf{p}$ such that $\rho_I \supset \iota_2$.

8.4 Lemma (1) $\mathbf{J}$ *contains all singletons $\{i\}$ ($i \in \mathbf{p}$).*

(2) *The set $\mathbf{p}$ does not belong to $\mathbf{J}$, and*

(3) $\rho_i * \rho_{i+d} = \mathbf{k}^2$ *for all $i \in \mathbf{p}$ and $0 < d < p$.*

Proof (2) Suppose to the contrary that $\mathbf{p} \in \mathbf{J}$. Then $\rho_\mathbf{p} := \rho_0 \cap \cdots \cap \rho_{p-1} \supset \iota_2$. In view of $\mathbf{k}^2 \supset \rho_0 \supseteq \rho_\mathbf{p}$ the relation $\rho_\mathbf{p}$ is a nontrivial (reflexive and symmetric) binary relation. Put $\tau_n := \rho_n \cap \cdots \cap \rho_{n+p-1}$ for all $n \geq 0$. Then ρ is dominated by the proper polyrelation $(\rho_\mathbf{p}, \rho_\mathbf{p}, \dots)$ and this contradiction proves (2).

Let $0 < d < p$. For $x \geq 0$ form $\lambda_x := \rho_x * \rho_{x+d}$. Then $\lambda \in [\rho]$ and λ is reflexive and symmetric. Moreover, $\lambda_x \supseteq \rho_x \cup \rho_{x+d}$ and taking into account that ρ is rich, from Lemma 8.3 we obtain that either λ is improper or $\lambda = \rho$. In the latter case from $\rho_x = \lambda_x \supseteq \rho_x \cup \rho_{x+d}$ we obtain $\rho_x \supseteq \rho_{x+d}$. Thus $\rho_x \supseteq \rho_{x+d} \supseteq \cdots \supseteq \rho_{x+pd} = \rho_x$; hence $\rho_x = \rho_{x+d}$ and ρ has period

$d < p$. This contradiction shows that $\lambda_x \in \{\iota_2, \mathbf{k}^2\}$ for all $x \in \mathbf{k}$. We show that $\rho_d \supset \iota_2$. If $\rho_d = \iota_2$, then $\lambda_0 = \rho_0 * \iota_2 = \rho_0$ where $\iota_2 \subset \rho_0 \subset \mathbf{k}^2$. This contradiction shows $\rho_d \supset \iota_2$ for all $d \in \mathbf{p}$ proving (1). Now for all $x \geq 0$ we have $\lambda_x = \rho_x * \rho_{x+d} \supseteq \rho_x \supset \iota_2$ and so $\lambda_x = \mathbf{k}^2$ proving (3). □

8.5 We need the following definitions and notations. Denote by $\mathbf{E}$ the set of all equivalence relations on $\mathbf{k}$ distinct from ι_2 (note that $\mathbf{k}^2 \in \mathbf{E}$) and put

$$M_2 := \{\rho \in M_1 \mid \rho_n \in \mathbf{E} \text{ for all } n \geq 0\}.$$

Let ξ be a symmetric and reflexive binary relation on $\mathbf{k}$. As usual, a *path* of length n from $i \in \mathbf{k}$ to $j \in \mathbf{k}$ in ξ is a sequence $\langle v_0, \dots, v_n\rangle$ in $\mathbf{k}$ such that $i = v_0$, $j = v_n$ and $v_i v_{i+1} \in \xi$ for all $i = 0, \dots, n-1$. The relation ξ is *connex* if for all $i, j \in \mathbf{k}$ there is a path from i to j. The relation ξ is *disconnected* if it is not connex. We need:

8.6 Lemma *If there exists $\lambda \in [\rho] \cap M$ with all $\lambda_n \supset \iota_2$ and some λ_i disconnected then ρ is dominated by some $\sigma \in M_2$.*

Proof Put $\sigma_n := (\lambda_n)^k$ ($= \lambda_n \circ \dots \circ \lambda_n$, k times) for all $n \geq 0$. It is easy to see that $\sigma_n \in \mathbf{E}$ for all $n \geq 0$ and $\sigma_i \neq \mathbf{k}^2$. Thus $\sigma \in M_2$. □

8.7 Lemma *If $\rho \in M_1 \setminus M_2$ then every ρ_i distinct from $\mathbf{k}^2$ is central.*

Proof For $i = 2, \dots, k$ and $n > 0$ put

$$\tau_n^i := \{x_1 \cdots x_i \mid x_1 u, \dots, x_i u \in \rho_n \text{ for some } u \in \mathbf{k}\}. \tag{8.1}$$

Suppose to the contrary that the polyrelation τ^2 is proper. Then clearly $\tau^2 \in [\rho] \cap M$, the period of τ^2 is at most p and $\tau_n^2 \supseteq \rho_n$ for all $n \geq 0$. Applying Lemma 8.3 we obtain $\tau^2 = \rho$ i.e. $\rho_n^2 = \rho_n \circ \breve{\rho}_n = \tau_n = \rho_n$ for all $n \geq 0$. It can be easily seen that ρ_n is an equivalence relation on $\mathbf{k}$ and so $\rho \in M_2$ contrary to our assumption. Thus τ^2 is improper. Since $\rho_n \supset \iota_2$ by Lemma 8.4 (1), we have $\tau_n^2 = \mathbf{k}^2$ for all $n \geq 0$.

By induction on $i = 2, \dots, k$ we prove that $\tau_n^i = \mathbf{k}^i$. We have just proved it for $i = 2$. Suppose $2 \leq i < k$ and $\tau_n^i = \mathbf{k}^i$ for all $n \geq 0$. It is easy to see from (8.1) that τ_n^{i+1} is totally symmetric. To show that it is also totally reflexive let $x_1, \dots, x_i \in \mathbf{k}$ be arbitrary. Then $x_1 \dots x_i \in \mathbf{k}^i = \tau_n^i$ and hence $x_1 u, \dots, x_i u \in \rho_n$ for some $u \in \mathbf{k}$. Clearly $x_1 u, \dots, x_i u \in \rho_n$ and hence from (8.1) we get $x_1 \dots x_i x_i \in \tau_n^{i+1}$. Now τ_n^{i+1} is totally symmetric and reflexive. Applying Theorem 4.2 we obtain $\tau_n^{i+1} = \mathbf{k}^{i+1}$. This concludes the induction step. In particular, we have $\tau_n^k = \mathbf{k}^k$ and so $\rho_n = \mathbf{k}^2$ or ρ_n is central. □

Denote by C the set of proper, binary and periodic polyrelations $(\rho_0, \rho_1, \dots)$ on $\mathbf{k}$ such that each $\rho_i \neq \mathbf{k}^2$ is central.

8.8 Theorem *The set $\Sigma_6 := V_1 \cup A_5 \cup P_3 \cup M_2 \cup C$ is B-generic.*

8.9 Remark It is shown in [Hi78] that for $k = 3$ and $\rho \in M_2 \cup C$ the uniformly delayed clone $\operatorname{Pold}\rho$ is never precomplete. We know that for nontrivial equivalence relations $\theta_0, \dots, \theta_{p-1}$ on $\mathbf{k}$ that are pairwise commuting (i.e., $\theta_i \circ \theta_j = \theta_j \circ \theta_i$ for all $0 \leq i < j < p$) and satisfy $\theta_0 \cap \dots \cap \theta_{p-1} = \iota_2$ the polyrelation $\theta = (\theta_0, \dots, \theta_{p-1}, \theta_0, \dots)$ with period p, the uniformly delayed clone $\operatorname{Pold}\theta$ is precomplete.

9 Conclusion

It remains to show that for each $\rho \in V_1 \cup A_5 \cup P_3$ the uniformly delayed clone $D = \operatorname{Pold} \rho$ is precomplete. This could be done directly (by showing that $D \cup \{(f, \delta)\}$ is complete for each $(f, \delta) \in \mathcal{U} \setminus D$) or by showing that $\operatorname{Pold} \sigma = \mathcal{U}$ or $\operatorname{Pold} \sigma$ is complete for each $\sigma \in [\rho]$.

The main remaining problem is the elimination within $M_2 \cup C$; i.e., finding proper periodic polyrelations ρ consisting of either (i) equivalence relations on $\mathbf{k}$, or (ii) binary central relations or $\mathbf{k}^2$ such that $\operatorname{Pold} \rho$ is precomplete. So far we have several partial results but the problem seems to be complex.

References

[Ba67] R. A. Baĭramov, On the question of functional completeness in many-valued logics, *Diskret. Analiz* **11** (1967), 3–20 (Russian).

[BW79] N. L. Biggs and A. T. White, *Permutation Groups and Combinatorial Structures*, London Math. Soc. Lecture Note Series **33**, Cambridge Univ. Press, 1979.

[BK70] L. A. Biryukova and V. B. Kudryavtsev, On completeness of functions with delays, *Problemy Kibernet.* **23** (1970), 5–25 (Russian); *Systems Theory Research* **23** (1973), 3–24.

[BKKR69] V. G. Bodnarchuk, L. A. Kaluzhnin, V. N. Kotov, and B. A. Romov, The Galois theory for Post algebras I–II, *Kibernetika* (Kiev) **3** (1969), 1–10, **5** (1969), 1–9 (Russian); *Systems Theory Research* **3**.

[Bu80] V. A. Buevich, On the t-completeness in the class of automata maps, *Dokl. Akad. Nauk. SSSR* **252** (1980), 1037–1041 (Russian).

[Da81] J. Dassow, *Completeness Problems in the Structural Theory of Automata*, Akademie-Verlag, Berlin, 1981.

[Ge68] D. Geiger, Closed systems of functions and predicates, *Pacific J. Math.* **27** (1) (1968), 95–100.

[Go68] D. Gorenstein, *Finite Groups*, second ed., Chelsea Pub., 1968, 1980.

[Hi78] T. Hikita, Completeness criterion for functions with delay defined over a domain of three elements, *Proc. Japan Acad.* **54** (1978), 335–339.

[Hi81a] T. Hikita, Completeness properties of k-valued functions with delays: Inclusions among closed spectra, *Math. Nachr.* **103** (1981), 5–19.

[Hi81b] T. Hikita, On completeness for k-valued functions with delay, in: *Finite Algebra and Multiple-Valued Logic* (B. Csákány and I. Rosenberg, eds.), *Coll. Math. Soc. János Bolyai* **28**, North-Holland, 1981, 345–371.

[HN77] T. Hikita and A. Nozaki, A completeness criterion for spectra, *SIAM J. Comput.* **6** (1977), 285–297. Corrigenda, ibid., **8** (1979), 656.

[HR98] T. Hikita and I. G. Rosenberg, Completeness for uniformly delayed circuits, a survey, *Acta Applic. Math.* **52** (1998), 49–61.

[Ia58] S. V. Iablonskiĭ, Functional constructions in the k-valued logic, *Trudy Mat. Inst. Steklov.* **51** (1958), 5–142 (Russian).

[Ib68] K. Ibuki, A study on universal logical elements , *NTT — Inst. of Electrocomm. Report*, **3747** (1968) (Japanese).

[In82] K. Inagaki, t_6-completeness of sets of delayed logic elements, *Trans. IECE Japan* **J63-D** (1982), 827–834 (Japanese).

[Kr59] R. E. Krichevskiĭ, The realization of functions by superpositions, *Problemy Kibernet.* **2** (1959), 123–138 (Russian); *Probleme der Kybernetik* **2** (1963), 139–159 (German).

[Ku60] V. B. Kudryavtsev, Completeness theorem for a class of automata without feedback couplings, *Dokl. Akad. Nauk SSSR* **132** (1960), 272–274 (Russian); *Soviet Math. Dokl.* **1** (1960), 537–539.

[Ku62] V. B. Kudryavtsev, Completeness theorem for a class of automata without feedback couplings, *Problemy Kibernet.* **8** (1962), 91–115 (Russian); *Probleme der Kybernetik* **8** (1965), 105–136 (German).

[Ku65] V. B. Kudryavtsev, The power of sets of precomplete sets for certain functional systems connected with automata, *Problemy Kibernet.* **13** (1965), 45–74 (Russian).

[Ku73] V. B. Kudryavtsev, On functional properties of logical nets, *Math. Nachr.* **55** (1973), 187–211 (Russian).

[Ku82] V. B. Kudryavtsev, *Functional Systems*, Monograph, I.M.U., Moscow, 1982 (Russian).

[Lo65] H. H. Loomis, Jr., Completeness of sets of delayed-logic devices, *IEEE Trans. Electron. Comput.* **EC-14** (1965), 157–172.

[MRR78] L. Martin, C. Reischer, and I. G. Rosenberg, Completeness problems for switching circuits constructed from delayed gates, *Proc. 8th Internat. Symp. Multiple-Valued Logic*, 1978, 142–148; *Elektron. Informationsverarb. Kybernet.*, **19** (1983), 415, 171–186 (in French); An expanded French version appeared in: Réseaux modulaires, *Monographies de l.U.Q.T.R.* **11** 1980.

[Ma59] N. M. Martin, On the completeness of sets of gating functions with delay, in: *Sympos. on Gating Algebra*, Proc. Internat. Congr. Information Processing, Paris, 1959.

[Mi84] M. Miličić, Galois connections for closed classes of functions with delays *Publ. de l'Inst. Mathématique Nouvelle Série* **36** (50) (1984), Part I: 119–124; Part II: 125–136 (Russian).

[Mi88] M. Miličić, On Galois connections for one place functions with delays, *Publ. de l'Inst. Mathématique Nouvelle Série* **44** (58) (1988), 147–150 (Russian).

[MSHMF88] M. Miyakawa, I. Stojmenović, T. Hikita, H. Machida, and R. Freivalds, Sheffer and symmetric Sheffer Boolean functions under various functional constructions, *J. Inform. Process. Cybernet. EIK*, (formerly: *Elektron. Informationsverarb. Kybernet.*) **24** (6) (1988), 251–266.

[No70a] A. Nozaki, Functional studies of automata I–II, *Sci. Papers College Gen. Ed. Univ. Tokyo* **20** (1970), 21–36, 109–121.

[No70b] A. Nozaki, Réalisation des fonctions definies dans un ensemble fini à l'aide des organes élémentaires d'entrée-sortie, *Proc. Japan Acad.* **46** (1970), 478–482.

[No72] A. Nozaki, Complete sets of switching elements and related topics, *First USA-Japan Computer Conference*, 1972, 393–396.

[No78] A. Nozaki, Functional completeness of multi-valued logical functions under uniform compositions, *Rep. Fac. Eng. Yamanashi Univ.* **29** (1978), 61–67.

[No81] A. Nozaki, Completeness criteria for a set of delayed functions with or without non-uniform compositions, in: *Finite Algebra and Multiple-Valued Logic* (B. Csákány and I. Rosenberg, eds.), *Coll. Math. Soc. János Bolyai* **28**, North-Holland, 1981, 489–519.

[No82] A. Nozaki, Completeness of logical gates based on sequential circuits, *Trans. IECE Japan* **J65-D** (1982), 171–178 (Japanese).

[PK79] R. Pöschel and L. A. Kaluzhnin, *Funktionen- und Relationenalgebren*, VEB Deutscher Verlag der Wissenschaften, Berlin, 1979.

[Po21] E. L. Post, Introduction to a general theory of elementary propositions, *Amer. J. Math.* **43** (1921), 163–185.

[Ro65] I. G. Rosenberg, La structure des fonctions de plusieurs variables sur un ensemble fini, *C.R. Acad. Sci. Paris* Sér. A–B, **260** (1965), 3817–3819.

[Ro70] I. G. Rosenberg, Über die funktionale Vollständigkeit in dem mehrwertigen Logiken (Struktur der Funktionen von mehreren Veränderlichen auf endlichen Mengen), *Rozpravy Československé Akad. Věd.*, Ser. Math. Nat. Sci., **80** (1970), 3–93.

[Ro73] I. G. Rosenberg, The number of maximal closed classes in the set of functions over a finite domain, *J. Comb. Theory* **14** (1973), 1–7.

[Ro77] I. G. Rosenberg, Completeness properties of multiple-valued logic algebras, in: *Computer Science and Multiple-Valued Logic, Theory and Applications* (D. C. Rine, ed.), North-Holland, 1977, 144–186; 2nd edition: ibid, 1984, 150–194.

[Ro83] I. G. Rosenberg, The multi-faceted completeness problem of the structural theory of automata, preprint, CRMA-1197, Université de Montréal, 1983.

[RH83] I. G. Rosenberg and T. Hikita, Completeness for uniformly delayed circuits, *Proceedings 13th Internat. Sympos. on Multiple-Valued Logic*, IEEE, Kyoto, 1983, 2–10; Proofs: Preprint, Tokyo Metropolitan Univ., 1983.

[RS84] I. G. Rosenberg and L. Szabó, Local completeness I, *Algebra Universalis* **18** (1984), 308–326.

[Sł39] J. Słupecki, A completeness criterion in many-valued propositional calculi, *C.R. Séance Soc. Sci. Varsovie* **32** (1939), 102–109 (Polish); *Studia Logica* (1972), 153–157.

[Sz86] Á. Szendrei, *Clones in universal algebras*, Les Presses de l'Université de Montréal, SMS, 1986.

Classification in finite model theory: counting finite algebras

Paweł M. IDZIAK

Computer Science Department
Jagiellonian University
Nawojki 11, PL-30-072 Kraków
Poland

Abstract

The *G-spectrum* or *generative complexity* of a class $\mathcal{C}$ of algebraic structures is the function $G_{\mathcal{C}}(k)$ that counts the number of non-isomorphic models in $\mathcal{C}$ that are generated by at most k elements. We consider the behavior of $G_{\mathcal{C}}(k)$ when $\mathcal{C}$ is a locally finite equational class (variety) of algebras and k is finite. We are interested in ways that algebraic properties of $\mathcal{C}$ lead to upper or lower bounds on generative complexity. Some of our results give sharp upper and lower bounds so as to place a particular variety or class of varieties at a precise level in an exponential hierarchy. Much of our work is motivated by a desire to know which locally finite varieties have polynomially many or singly exponentially many models, and to discover conditions that force a variety to have many models. We present characterization theorems for a very broad class of varieties including most known and well-studied types of algebras, such as groups, rings, modules, lattices. In particular, we show that a finitely generated variety of groups has singly exponentially many models if and only if it is nilpotent and has polynomially many models if and only if it is Abelian.

1 Introduction

Given a class $\mathcal{C}$ of structures two basic questions about the class $\mathcal{C}$ are:

- What are the cardinalities of structures in $\mathcal{C}$?
- How many non-isomorphic structures of a given cardinality are in $\mathcal{C}$?

The first question is the problem of determining the *spectrum* $\mathrm{Spec}(\mathcal{C})$ of the class $\mathcal{C}$, that is, the class of cardinalities that occur as sizes of structures in $\mathcal{C}$. This problem appears in all sorts of contexts in mathematics. For example, if $\mathcal{C}$ is the class of models of a first order theory in a countable language then, by the Löwenheim-Skolem theorems, every infinite cardinality is in $\mathrm{Spec}(\mathcal{C})$. Thus, $\mathrm{Spec}(\mathcal{C})$ is usually defined to be the set of sizes of finite models in $\mathcal{C}$. Let us mention here a result of R. Fagin [12] that characterizes those sets of integers that can be expressed as $\mathrm{Spec}(\mathcal{C})$, where $\mathcal{C}$ is the class of all models of a first order sentence in a first order language. They are exactly the sets that can be recognized in nondeterministic exponential time. He also proved a similar connection between nondeterministic polynomial time and what are called generalized spectra.

V. B. Kudryavtsev and I. G. Rosenberg (eds.),
Structural Theory of Automata, Semigroups, and Universal Algebra, 149–158.

The second basic question is about the *fine spectrum* of the class $\mathcal{C}$, i.e., about the function that assigns to each cardinal k the number $I(\mathcal{C}, k)$ of non-isomorphic k-element structures in $\mathcal{C}$. The problem of determining the fine spectrum of a class, either exactly or asymptotically, is a natural one, and is a problem that has been considered for various classes $\mathcal{C}$ and in various contexts over the years. Combinatorial enumeration problems, such as finding the number of k-vertex graphs in some specific class of graphs, are familiar examples of such fine spectra problems.

In Model Theory one of the fundamental topics is Shelah's classification theory. Given a first order theory T we are interested in the number of non-isomorphic models of T of a given (infinite) cardinality k. Moreover we want to classify theories with the number of non-isomorphic models of infinite cardinality k given by a prescribed function $f(k)$. Some of the results here say that not every function f can be realized. For example an early result of M. Morley says that if T is a complete theory in a countable language and T is κ-categorical for some uncountable cardinal κ then T is categorical in all uncountable powers.

When we are interested in finite models only, i.e., restricting k to be finite, and in integer valued functions f, then the situation changes dramatically. For any function $f : \omega \longrightarrow \omega$ there is a first order theory T having exactly $f(k)$ models with k elements. Simply let T be axiomatized by sentences expressing:

> *If a model has k elements then it is isomorphic with one of the $f(k)$ choices given by their diagrams.*

Even restricting the language to be finite leaves a lot of room for possible fine spectra. A function $f : \omega \longrightarrow \omega$ counts non-isomorphic models of a first order theory in a finite language if and only if $\log f(k)$ is bounded from above by a polynomial in k.

Therefore the finite part of the fine spectrum problem is much more interesting in the case of restricted theories, eg. in equational theories of algebras. Actually the terminology "fine spectrum" was introduced by W. Taylor in [31] in the case where $\mathcal{C}$ is a variety, that is, a class of algebraic structures or algebras closed under the formation of homomorphisms, subalgebras and direct products. Among the results of Taylor's paper is a characterization of those varieties that have exactly one model of size 2^k for each finite k and no other finite models. All such varieties must be generated by a 2-element algebra. Subsequent papers by R. Quackenbush [27] and D. Clark and P. Krauss [9] extended Taylor's work to n-element algebras that generate varieties having minimal fine spectra.

Despite this work on minimal fine spectra there has been relatively little success on general problems involving fine spectra of varieties. In [32] Taylor writes of fine spectra functions that "characterizations of such functions seems hopeless" and Quackenbush states in his survey on enumeration problems for ordered structures [28] that the fine spectrum problem here "is usually hopeless".

There are many reasons for the difficulties in determining the fine spectra of varieties. One reason that interests us is that algebras are often described by means of a set of generators. Once we know the generators of an algebra $\mathbf{A}$ in a variety $\mathcal{V}$ and some of the conditions that these generators satisfy, our freedom in building the rest of the model is heavily restricted. This effect is widely used in group theory where a group is usually presented by a (finite) set of generators and a set of relations the generators must obey. The constraints put on the behavior of the generators place restrictions on the structure of the entire algebra $\mathbf{A}$. However, there is in general no obvious or transparent way to determine the cardinality of

A. This makes the counting of all n-element or at most n-element algebras difficult, if at all possible, even if we content ourselves with an asymptotic estimate.

Another area of research in which fine spectra appear is in finite model theory. When investigating asymptotic probabilities in finite model theory, some of the results rely on counting all finite models, up to isomorphism, for a given theory T. This usually is not hard when T has no axioms. However there are only a few results on zero-one (or more generally on limit) laws for specific theories T. One reason is that for such counting deep insight into the structure of finite models of T is required. The counting is even more difficult if the language of T contains function symbols. Except for unary functions [23] (in which case the models behave much more like relational structures than algebras) and Abelian groups [10] (where we completely understand the structure) there are only a few other results on limit laws for algebras to report. The reader may wish to consult [4, 5, 6, 7, 8, 22] for related work.

Again, there are various reasons for the difficulties here with fine spectra for classes arising from theories in a language containing function symbols. Some important techniques for asymptotic probabilities rely on extension axioms. These techniques work perfectly well for purely relational structures when often the resulting random structure is model complete. However (randomly) adding a new element a to the universe of an algebra $\mathbf{A}$ in a variety $\mathcal{V}$ and keeping the resulting extended algebra in $\mathcal{V}$ requires a much bigger extension $\mathbf{A}^*$ of $\mathbf{A}$. The number and the behavior of all those new elements in $\mathbf{A}^*$ is fully determined by the old algebra $\mathbf{A}$ and the interaction of the new element a with the elements of A. Thus, in vector spaces, by adding a new element we actually increase the dimension.

One possible way to overcome these difficulties is to count k-generated models instead of k-element ones. This is a more tractable problem when the classes are varieties of algebras and, we believe, is the proper setting for asymptotic probabilities in algebra. Note however that these numbers are the same for purely relational languages.

Thus, we introduce the *G-spectrum*, or *generative complexity*, of a class $\mathcal{C}$, which is the function $\mathrm{G}_{\mathcal{C}}(k)$ that counts the number of non-isomorphic (at most) k-generated models in $\mathcal{C}$.

We concentrate on the case where $\mathcal{C}$ is a variety of algebras and restrict ourselves to finite k. Even in this new setting the counting remains hard. It requires an understanding of the 'generating power' in a given variety $\mathcal{V}$. This is related to the old problem of determining the *free spectrum* of $\mathcal{V}$, that is, the sizes of free algebras $\mathbf{F}_{\mathcal{V}}(k)$ in $\mathcal{V}$ with $k = 1, 2, \ldots$ free generators. This is because every k-generated algebra in $\mathcal{V}$ is isomorphic to a quotient of $\mathbf{F}_{\mathcal{V}}(k)$ by a congruence relation. However, two different congruences may give rise to isomorphic quotients. Thus, the second problem we meet here is to measure the amount of homogeneity in $\mathbf{F}_{\mathcal{V}}(k)$. One cannot hope to solve these two problems without understanding the structure of algebras from $\mathcal{V}$.

As we have already mentioned the infinite counterpart of the problem of counting non-isomorphic models is widely studied in Model Theory, and is one of the fundamental topics for Shelah's classification theory and for stability theory. Note that in the infinite realm (and for a countable language), being κ-generated and having κ elements are the same.

For example, the celebrated Vaught conjecture says that the number $I(\mathcal{C}, \omega)$ of non-isomorphic countable models of any first order theory in a countable language is either countable or 2^{ω}. In [18] and [17] B. Hart, S. Starchenko, and M. Valeriote have been able to prove this conjecture for varieties of algebras. They actually determined the possible infinite fine spectra (which, in this case, are the same as infinite G-spectra) of varieties and correlated them with algebraic properties of varieties. A characterization of locally finite varieties with

strictly less than 2^ω non-isomorphic countable models can be easily inferred from deep work on decidability done by R. McKenzie and M. Valeriote [25].

The general project of determining the fine spectra of theories has a long history, going back at least to the Łoś conjecture, which stated that the uncountable fine spectrum of any countable theory that is categorical in some uncountable cardinal must be identically equal to 1. The conjecture was confirmed by M. Morley [26] in 1965; but already in 1957, A. Ehrenfeucht [11] had provided some of the methods which were later developed by S. Shelah to show that unstable theories have many models. In Shelah's classification theory [29], the countable theories with a prescribed fine spectrum for uncountable cardinals are described and classified. Some of the results here, like the mentioned Łoś conjecture, say that not every function f can be realized as the uncountable fine spectrum of a countable theory. In fact, it is proved in [29] that restricted to sufficiently large κ, there are precisely eight possible functions $I(\mathcal{C}, \kappa)$. B. Hart, E. Hrushovski, and M. Laskowski [16] completed Shelah's classification theory by resolving a conjecture of Shelah and computing the spectrum function for small (uncountable) values of κ.

However, one cannot easily (if at all) transfer infinite methods and results into the finite world. Thus, we focus on the G-spectrum of a variety rather than its fine spectrum. We further restrict ourselves to locally finite varieties, i.e., varieties in which all finitely generated algebras are finite. This gives that the G-spectrum of the variety $\mathcal{V}$ is integer-valued. Even with this finiteness restriction G-spectra can be arbitrarily large: for any sequence $(p_k)_{k\in\omega}$ of integers there is a locally finite variety $\mathcal{V}$ of groupoids such that $G_\mathcal{V}(k) \geq p_k$ for all k [2, Example 5.9]. On the other hand, if $\mathcal{V}$ is a finitely generated variety of finite type then easily established upper bounds for the G-spectrum, free spectrum and fine spectrum of $\mathcal{V}$ are $2^{2^{2^{ck}}}$, $2^{2^{ck}}$, and 2^{ck^s}, respectively, for some constants c and s.

2 Results

The main problems that stimulate our research in this area are:

- *How does the growth of the G-spectrum of a locally finite variety $\mathcal{V}$ affect the structure of algebras in $\mathcal{V}$?*
- *In what way do algebraic properties of $\mathcal{V}$ influence the behavior of $\mathrm{G}_\mathcal{V}(k)$?*
- *For a given variety $\mathcal{V}$, can an explicit formula for $\mathrm{G}_\mathcal{V}(k)$ be found?*
- *Can the asymptotic behavior of $\mathrm{G}_\mathcal{V}(k)$ be determined?*

For some $\mathcal{V}$, the value of $\mathrm{G}_\mathcal{V}(k)$ is easily determined, as the following examples show.

2.1 Example For the variety $\mathcal{V}$ of sets, that is algebras with no fundamental operations, $\mathrm{G}_\mathcal{V}(k) = k$. In [2, Section 3] a full description of G-spectra for varieties generated by arbitrary multi-unary algebras is given.

2.2 Example Let $\mathcal{V}$ be the variety of vector spaces over a fixed field. A k-generated algebra here is a vector space of dimension at most k. So $\mathrm{G}_\mathcal{V}(k) = k + 1$. If $\mathcal{A}_p$ is the variety generated by the p-element group for a prime p, then a k-generated group in $\mathcal{A}_p$ has size p^m,

for $0 \leq m \leq k$. So $\mathrm{G}_{\mathcal{A}_p}(k) = k+1$. In [2, Example 6.7] the function $\mathrm{G}_{\mathcal{V}}(k)$ for $\mathcal{V}$ an arbitrary, finitely generated variety of Abelian groups is determined.

2.3 Example For the variety $\mathcal{B}$ of Boolean algebras we have $|\mathrm{F}_{\mathcal{B}}(k)| = 2^{2^k}$. Every k-generated member of $\mathcal{B}$ is a Boolean algebra with m atoms, $0 \leq m \leq 2^k$. Thus $\mathrm{G}_{\mathcal{B}}(k) = 1+2^k$.

2.4 Example In contrast to these varieties that have small G-spectra, [2, Section 4] shows that semilattices and distributive lattices have $\mathrm{G}_{\mathcal{V}}(k)$ that are doubly exponential functions of k. Actually it is shown that $\mathrm{G}_{\mathcal{S}}(k) = 2^{\binom{k}{\lfloor k/2 \rfloor}(1+o(1))}$ for the variety $\mathcal{S}$ of semilattices and $\mathrm{G}_{\mathcal{D}}(k) = 2^{2^k(1+o(1))}$ for the variety $\mathcal{D}$ of distributive lattices.

Moreover [2, Section 5] contains examples of locally finite varieties with arbitrarily large G-spectra.

To report the results on G-spectra we introduce some classes of functions that are helpful in describing the possible behavior of the $\mathrm{G}_{\mathcal{C}}(k)$.

2.5 Definition Let f be a real-valued function of the positive integers.

- We say f is *at most* 0*-fold exponential* if there exists a polynomial p such that $f(k) \leq p(k)$ for all k.
- For $m > 0$ the function f is *at most* m*-fold exponential* if there is an at most $(m-1)$-fold exponential function g such that $f(k) \leq 2^{g(k)}$ for all k.
- The function f is called *at least* 0*-fold exponential* if there exists a constant $c > 0$ such that $f(k) \geq ck$ for all but finitely many k.
- For $m > 0$ the function f is *at least* m*-fold exponential* if there is an at least $(m-1)$-fold exponential function g such that $f(k) \geq 2^{g(k)}$ for all but finitely many k.
- The function f is of m*-fold exponential complexity* if f is both at most and at least m-fold exponential.
- A class $\mathcal{C}$ of structures has m-fold exponential generative complexity if the function $\mathrm{G}_{\mathcal{C}}(k)$ is of m-fold exponential complexity.

We usually write polynomial, exponential, doubly exponential, and triply exponential in place of 0-fold exponential, 1-fold exponential, 2-fold exponential, and 3-fold exponential.

2.6 Definition Let $\mathcal{V}$ be a locally finite variety.

- $\mathcal{V}$ has *many models* if $\mathrm{G}_{\mathcal{V}}(k)$ is at least doubly exponential.
- $\mathcal{V}$ has *few models* if $\mathrm{G}_{\mathcal{V}}(k)$ is at most exponential.
- $\mathcal{V}$ has *very few models* if $\mathrm{G}_{\mathcal{V}}(k)$ is at most polynomial.

The research reported here was motivated by a desire to know which locally finite varieties have few and very few models, respectively. Although we have not managed to solve these problems in the most general setting we have obtained such a characterization for a very broad class of varieties including most known and well-studied types of algebras, such as groups, rings, modules, lattices. In those cases we are in the realm of congruence modular varieties where we have a better than triply exponential upper bound for G-spectra of finitely generated subvarieties.

2.7 Theorem (J. Berman and P. Idziak [2]) *If $\mathcal{V}$ is a finitely generated, congruence modular variety, then $\mathrm{G}_{\mathcal{V}}(k)$ is at most doubly exponential.*

The proofs of the results give a deep insight into the structure of locally finite varieties with few and very few models. The analysis relies heavily on two major developments of the late 70's and early 80's. One of them is *modular commutator theory*. The theory had been introduced by J. D. H. Smith [30] for congruence permutable varieties. It was further developed by J. Hagemann and Ch. Herrmann [15], H. P. Gumm [14] and R. Freese and R. McKenzie [13]. The book of Freese and McKenzie contains several important results and techniques that are extremely useful when studying congruence modular varieties. A binary operation on congruences that simultaneously generalizes the concept of a commutator $[H, K]$ of two normal subgroups H, K of a group as well as the ideal multiplication in rings is defined. The theory shows how some information about algebras or varieties can be recovered from congruence lattices endowed with this binary operation. Moreover the concept of the commutator allows us to speak about a solvable, nilpotent or Abelian congruence (or algebra) as well as about the center of an algebra or the centralizer of a congruence relation.

The second big development in universal algebra that we use is *tame congruence theory*, which was created and described in D. Hobby and R. McKenzie [19]. Tame congruence theory is a tool for studying the local structure of finite algebras. Instead of considering the whole algebra and all of its operations at once, tame congruence theory allows us to localize to small subsets on which the structure is much simpler to understand and to handle. According to this theory there are only five possible ways a finite algebra can behave locally. The local behavior must be one of the following:

1 a finite set with a group action on it;

2 a finite vector space over a finite field;

3 a two element Boolean algebra;

4 a two element lattice;

5 a two element semilattice.

Now, if from our point of view a local behavior of an algebra is 'bad' then we can often show that the algebra itself behaves 'badly'. For example, since the varieties of distributive lattices or semilattices have many models (see Example 2.4) then one can argue that in any locally finite variety with few models structures of type **4** or **5** cannot occur.

On the other hand it is not true that if the local behavior is 'good' then the global one is, as well. Several kinds of interactions between these small sets can produce fairly messy

global behavior. Such interactions often contribute to produce many models. Also the relative 'geographical layout' of those small sets can result in unpredictable phenomena.

The main results of the work in this area are stated in the following theorems. The first one gives a full characterization of locally finite varieties omitting type **1** with polynomially many models:

2.8 Theorem (P. Idziak and R. McKenzie [20]) *A locally finite variety omitting type* **1** *has very few models if and only if it is an affine variety over a ring of finite representation type.*

Very recently this characterization was extended to all locally finite varieties:

2.9 Theorem (P.Idziak, R. McKenzie, and M. Valeriote [21]) *A locally finite variety has very few models if and only if it decomposes into a varietal product of an affine variety over a ring of finite representation type, and a sequence of strongly Abelian varieties equivalent to matrix powers of varieties of G-sets, with constants, for various finite groups G.*

The work on characterizing locally finite varieties with at most singly exponentially many models is still in progress. However, for the case of finitely generated varieties, omitting type **1**, the following characterization (with more than a 100-page proof) was obtained:

2.10 Theorem (J. Berman and P. Idziak [2]) *Let $\mathcal{V}$ be a finitely generated variety omitting type* **1**. *Then $\mathcal{V}$ has few models if and only if the following conditions hold:*

(1) *$\mathcal{V}$ is congruence permutable;*

(2) *for any finite subdirectly irreducible algebra $\mathbf{A}$ in $\mathcal{V}$ with monolith μ and its centralizer $\nu = (0 : \mu)$ we have:*

   (a) *ν is the solvable radical of $\mathbf{A}$;*

   (b) *ν is comparable to all congruences of $\mathbf{A}$;*

   (c) *ν is Abelian or $\mathbf{A}$ is nilpotent;*

   (d) *the quotient $\mathbf{A}/\nu$ is either trivial or simple non-Abelian;*

(3) *the variety $\mathcal{N}$ of all nilpotent algebras in $\mathcal{V}$ has a finitely generated clone and $\mathcal{N}$ itself is generated by finitely many finite algebras each being of prime power order;*

(4) *for any finite simple non-Abelian algebra $\mathbf{S}$ in $\mathcal{V}$ the ring $\mathbf{R}_{\mathbf{S}}^{\mathcal{V}}$* (connected with the Abelian part of subdirectly irreducible algebras $\mathbf{A}$ in $\mathcal{V}$ with the quotient $\mathbf{A}/\nu$ isomorphic to $\mathbf{S}$) *is of finite representation type.*

The necessity of the conditions in the above theorems was shown by detecting more than two dozens different ways in which 'bad' local behavior can occur in an algebra $\mathbf{A}$. In each such situation we are able to produce many non-isomorphic k-generated algebras in the variety generated by $\mathbf{A}$. After detecting all those instances of 'bad' local behavior we formulate global algebraic conditions that forbid such undesirable behavior. More surprisingly we are able to show the list of conditions we obtain is actually complete, at least for finitely generated

varieties, i.e., these conditions taken together are sufficient for a variety to have few (or very few) models.

The conditions given in Theorems 2.8 and 2.9 are very simple and easily stated. The conditions involved in the second characterization (Theorem 2.10) are more complicated, although they also have a natural algebraic meaning. In both cases we know that the bound for the number of algebras implies a very transparent and manageable structure. For example, when specializing our results to groups we get the following:

- every finitely generated variety of groups has at most doubly exponentially many models;
- a finitely generated variety of groups has few models if and only if it is nilpotent;
- a finitely generated variety of groups has very few models if and only if it is Abelian.

while for commutative rings with unit our characterization reduces to:

- every finitely generated variety of rings has at most doubly exponentially many models;
- a finitely generated variety of commutative rings with unit has few models if and only if the Jacobson radical in the generating ring squares to 0;
- no nontrivial variety of rings with unit has very few models.

Also, from Theorem 2.9 one can infer an earlier result of M. Bilski [3] on semigroup varieties with very few models. Once more, it confirms nice structure theory for varieties with polynomially many models:

> *A locally finite variety $\mathcal{V}$ of semigroups has polynomially bounded G-spectrum if and only if $\mathcal{V}$ is the variety with zero multiplication (satisfies the identity $xy \approx uv$) or there is an m such that $\mathcal{V}$ is a subvariety of the join $\mathcal{L} \vee \mathcal{R} \vee \mathcal{A}_m$, where $\mathcal{L}$ is the variety of left-zero semigroups (satisfying the identity $xy \approx x$), $\mathcal{R}$ is the variety of right-zero semigroups (satisfying the identity $xy \approx y$) and $\mathcal{A}_m$ is the variety of Abelian groups of exponent m treated as semigroups (i.e., satisfies the identities $x^m \approx y^m$ and $x^{m+1} \approx x$).*

The work on this project started in the fall of 1996 during the semester on Algebraic Model Theory at the Fields Institute in Toronto.

Since then several researchers have also worked on the topic. Here are the results on G-spectra I wish to report:

- J. Berman and P. Idziak [2] most of the material described in this talk. It introduces and motivates the notion of generative complexity, provides many examples and contains a proof of Theorem 2.10.
- J. Berman and P. Idziak [1] described G-spectra of all Post varieties, i.e., varieties generated by a single two-element algebra.
- M. Bilski [3] characterized finitely generated varieties of semigroups with very few models.

- P. Idziak and R. McKenzie [20] succeeded in characterizing locally finite varieties omitting type **1** with very few models.
- P. Idziak, R. McKenzie, and M. Valeriote [21] extended the above characterization of varieties with very few models to all locally finite varieties.
- R. McKenzie [24] produced several examples of locally finite varieties with arbitrarily large free spectra and G-spectra.

References

[1] J. Berman and P. M. Idziak, Counting finite algebras in the Post varieties, *International Journal of Algebra and Computation* **10** (2000), 323–337.

[2] J. Berman and P. M. Idziak, *Generative complexity in algebra*, manuscript, 2002.

[3] M. Bilski, Generative complexity in semigroups varieties, *Journal of Pure and Applied Algebra* **165** (2001), 137–149.

[4] S. Burris, Spectrally determined first-order limit laws, in: *Logic and Random Structures* (R. Boppana and J. Lynch, eds.), DIMACS Series in Discrete Mathematics and Theoretical Computer Science **33**, Amer. Math. Soc., 1997, 33–52.

[5] S. Burris, *Number theoretic density and logical limit laws*. Mathematical Surveys and Monographs **86**. American Mathematical Society, Providence, RI, 2001.

[6] S. Burris and K. Compton, Fine spectra and limit laws. I. First order laws, *Canad. J. Math.* **49** (1997), 468–498.

[7] S. Burris, K. Compton, A. Odlyzko, and B. Richmond, Fine spectra and limit laws. II. First-order 0-1 laws, *Canad. J. Math.* **49** (1997), 641–652.

[8] S. Burris and P. Idziak, A directly representable variety has a discrete first-order law, *International J. of Algebra and Computation* **6** (1996), 269–276.

[9] D. Clark and P. Krauss, Plain para primal algebras, *Algebra Universalis* **11** (1980), 365–388.

[10] K. Compton, personal communication.

[11] A. Ehrenfeucht, On the theories categorical in power, *Fundamenta Mathematicae* **44** (1957), 241–248.

[12] R. Fagin, Generalized first-order spectra and polynomial-time recognizable sets, in: *Complexity and Computation, SIAM-AMS Proceedings* (R.Karp, ed.), **7** (1974), 43–73.

[13] R. Freese and R. McKenzie, *Commutator Theory for Congruence Modular Varieties*, London Math. Soc. Lecture Notes **125**, Cambridge Univ. Press, Cambridge, 1987.

[14] H. P. Gumm, *Geometrical methods in congruence modular varieties*, Memoirs Amer. Math. Soc. **289** (1983).

[15] J. Hagemann and C. Herrmann, A concrete ideal multiplication for algebraic systems and its relation to congruence distributivity, *Archive der Mathematik* **32** (1979), 234–245.

[16] B. Hart, E. Hrushovski, and M. Laskowski, The uncountable spectra of countable theories, *Annals of Mathematics* **152** (2000), 207–257.

[17] B. Hart, S. Starchenko and M. Valeriote, Vaught's conjecture for varieties, *Trans. Amer. Math. Soc* **342** (1994), 173–196.

[18] B. Hart and M. Valeriote, A structure theorem for strongly abelian varieties with few models, *Journal of Symbolic Logic*, **56** (1991), 832–852

[19] D. Hobby and R. McKenzie, *The Structure of Finite Algebras*, Contemporary Mathematics **76**, Amer. Math. Soc., Providence, RI, 1988.

[20] P. Idziak and R. McKenzie, Varieties with very few models, *Fundamenta Mathematicae*, **170** (2001), 53–68.

[21] P. Idziak, R. McKenzie, and M. Valeriote, The structure of locally finite varieties with polynomially many models, manuscript 2003.

[22] P. Idziak and J. Tyszkiewicz, Monadic second order probabilities in algebra. Directly representable varieties and groups, in: *Logic and Random Structures* (R. Boppana and J. Lynch, eds.), DIMACS Series in Discrete Mathematics and Theoretical Computer Science **33**, Amer. Math. Soc., 1997, 79–107.

[23] J. F. Lynch, Probabilities of first-order sentences about unary functions, *Trans. Amer. Math. Soc.* **287** (1985), 543–568.

[24] R. McKenzie, Locally finite varieties with large free spectra, *Algebra Universalis* **47** (2002), 303–318.

[25] R. McKenzie and M. Valeriote, *The Structure of Decidable Locally Finite Varieties*, Birkhäuser, Boston, 1989.

[26] M. Morley, Categoricity in power, *Transaction American Mathematical Society* **114** (1965), 514–538.

[27] R. W. Quackenbush, Algebras with minimal spectrum, *Algebra Universalis* **10** (1980), 117–129.

[28] R. W. Quackenbush, Enumeration in classes of ordered structures, in: *Ordered Sets (Banff, Alta., 1981)*, Reidel, Dordrecht-Boston, MA, 1982, 523–554.

[29] S. Shelah, *Classification Theory*, North Holland, Amsterdam, 1990.

[30] J. D. H. Smith, *Mal'cev Varieties*, Lecture Notes in Mathematics **554**, Springer-Verlag, Berlin, 1976.

[31] W. Taylor, The fine spectrum of a variety, *Algebra Universalis* **5** (1975), 263–303.

[32] W. Taylor, Equational logic, in: *Universal Algebra* (G. Grätzer, ed.), 2nd ed., Springer-Verlag, New York, 1979, 378–400.

Syntactic semigroups and the finite basis problem

Marcel JACKSON

Department of Mathematics
La Trobe University
Australia

Abstract

The finite basis problem for semigroups asks when a given finite semigroup has a finite basis for its identities. This problem is one of the most investigated in variety theory. In this note we look at some easily established equivalent decision problems.

1 Introduction

We show how one may efficiently associate with each semigroup (monoid) $\mathbf{S}$ a syntactic semigroup (monoid) $\bar{\mathbf{S}}$ such that $\mathbf{S}$ has a finite basis of identities if and only if $\bar{\mathbf{S}}$ has a finite basis of identities. Using this and a result of Sapir [14] we obtain a language theoretic equivalent to the finite basis problem for semigroups and monoids. While our techniques are completely elementary, the results may provide a useful alternative approach to the finite basis problem for finite semigroups.

1.1 Syntactic semigroups and monoids

In this section we gather togethor some basic facts concerning syntactic semigroups and monoids. For a more complete treatment, the reader should consult a book such as [6]. Recall that if $\mathbf{S}$ is a monoid, and W is a subset of S, then the *syntactic congruence* $\sim_W$ of W in $\mathbf{S}$ is given by $(a, b) \in \sim_W$ if for every $c, d \in S \cup \{1\}$, $cad \in W \Leftrightarrow cbd \in W$. (This congruence is the largest congruence for which W is a union of congruence classes.) In the case where $\mathbf{S}$ is a free semigroup (monoid), the quotient by the syntactic congruence of a subset L is called the *syntactic semigroup* (*syntactic monoid*) of L, and we denote this by $\mathrm{Syn}(L)$ (or $\mathrm{Syn}_{\mathrm{M}}(L)$, respectively). A language is *regular* (accepted by a finite state automaton) if and only if its syntactic semigroup (or monoid) is finite.

We will say that a subset $W \subseteq S$ of a semigroup $\mathbf{S}$ is a *syntactic subset* if the syntactic congruence of W in $\mathbf{S}$ is the diagonal relation. In this case, if $\nu : A^+ \to \mathbf{S}$ is a surjective homomorphism onto $\mathbf{S}$, then $\mathbf{S}$ is isomorphic to the syntactic semigroup of the language $\nu^{-1}(W)$.

If $\mathbf{S}$ is a semigroup, then for every $s \in S$, we may take the syntactic congruence of the singleton $\{s\}$; the semigroup $\mathbf{S}/\sim_{\{s\}}$ is syntactic, and for distinct elements $s, t \in S$ there is an $r \in S$ for which $s/\sim_{\{r\}} \neq t/\sim_{\{r\}}$ (choosing r to be either s or t suffices). Hence there is a subdirect embedding ν of $\mathbf{S}$ into the direct product $\prod_{s \in S} \mathbf{S}/\sim_{\{s\}}$ defined by $\nu(s) : x \mapsto$

V. B. Kudryavtsev and I. G. Rosenberg (eds.),
Structural Theory of Automata, Semigroups, and Universal Algebra, 159–167.

$x/\sim_{\{s\}}$. We note that one can construct $\mathbf{S}/\sim_{\{s\}}$ from $\mathbf{S}$ in polynomial time: to decide $x \sim_{\{s\}} y$ one must calculate uxv and uyv for each $u, v \in S^1$. Using a multi-tape Turing machine, we may encode $\mathbf{S}$ by listing the rows of its Cayley table; let n denote the number of squares this occupies (clearly $n \geq |S|^2$ and also $n \in O(|S|^2 \log_2(|S|))$ since each element can be encoded as a binary string of length at most $\log_2(|S|)$). Calculating uxv and uyv takes $O(n)$ steps, while there are $O(|S|^2)$ pairs u, v to consider. We direct the reader to [7] for details on these notions and more information on problems of computational complexity.

For a language W and a word w we define $w^{-1}W := \{u \in W : wu \in W\}$ and $Ww^{-1} := \{u \in W : uw \in W\}$. Now let $\mathcal{L}$ be a class of regular languages. For a finite alphabet A we denote by $\mathcal{L}(A)$ the set of all members of $\mathcal{L}$ whose alphabet is A. We say that $\mathcal{L}$ is a *+-variety* of languages if for every finite alphabet A the following properties hold: $\mathcal{L}(A)$ is closed under taking finite unions, intersections and complementation; for every letter $a \in A$ and every $W \in \mathcal{L}(A)$ we have $a^{-1}W \in \mathcal{L}$ and $Wa^{-1} \in \mathcal{L}$, and for every pair of finite alphabets A, B and every homomorphism $\nu : A^+ \to B^+$ we have $L \in \mathcal{L}(B)$ implies $\nu^{-1}(L) \in \mathcal{L}(A)$. If we replace free semigroups in this definition by free monoids, the corresponding notion is that of a $*$-variety.

There is a natural correspondence between +-varieties of regular languages and pseudovarieties (classes closed under finite direct products, subalgebras and homomorphic images) of finite semigroups. With each +-variety of regular languages $\mathcal{L}$, we may associate the semigroup pseudovariety generated by the syntactic semigroups of the languages in $\mathcal{L}$. Conversely with each pseudovariety of finite semigroups $\mathcal{V}$ we may associate the class of regular languages whose syntactic semigroups are in $\mathcal{V}$. This class of languages turns out to be a +-variety of languages and the two correspondences we have described are mutually inverse bijections between the lattice of semigroup pseudovarieties and the lattice of +-varieties of regular languages (this is the so-called *Eilenberg correspondence* [6]). Monoid and $*$-variety versions of this correspondence are given in the obvious way.

1.2 The finite basis problem

An *identity* of a semigroup is an expression $u \approx v$ where u and v are semigroup words. A semigroup satisfies the identity $u \approx v$ (in variables A, say) if for every homomorphism $\theta : A^+ \to \mathbf{S}$, the equality $\theta(u) = \theta(v)$ holds (written $\mathbf{S} \models u \approx v$). The set of identities of $\mathbf{S}$ over some fixed countably infinite alphabet is denoted $\mathrm{Id}(\mathbf{S})$. The class of all semigroups satisfying the identities of $\mathbf{S}$ is called the *variety* generated by $\mathbf{S}$. Equivalently, the variety of $\mathbf{S}$ is the class of all homomorphic images of subalgebras of direct powers of $\mathbf{S}$. (Similar notions hold for general algebras.) We will denote the variety of an algebra $\mathbf{S}$ by $\mathbb{V}(\mathbf{S})$, and the finite members of $\mathbb{V}(\mathbf{S})$ by $\mathbb{V}_{\mathrm{fin}}(\mathbf{S})$. When $\mathbf{S}$ is finite, $\mathbb{V}_{\mathrm{fin}}(\mathbf{S})$ coincides with the pseudovariety generated by $\mathbf{S}$. We will also write $\mathbf{S} \cong \mathbf{T}$ when $\mathrm{Id}(\mathbf{S}) = \mathrm{Id}(\mathbf{T})$.

If $\mathbf{S}$ is a semigroup without an identity element, then we let $\mathbf{S}^1$ denote the monoid obtained by adjoining an identity element. Otherwise we let $\mathbf{S}^1$ simply denote $\mathbf{S}$. A dual definition for zero elements 0 gives $\mathbf{S}^0$. The following elementary lemma is useful because we will frequently be moving between monoids and of semigroups.

1.1 Lemma *Let $\mathbf{S}$ be a monoid and $\mathbf{T}$ a semigroup such that $\mathbf{T}$ is contained in the semigroup variety generated by $\mathbf{S}$. Then $\mathbf{T}^1$ is contained in the monoid variety of $\mathbf{S}$.*

Proof Let $\mathbf{S}$ and $\mathbf{T}$ be as in the statement of the lemma. Then, working within the type

$\langle 2\rangle$, we have $\mathbf{T} \in \mathbb{HSP}(\mathbf{S})$. Therefore there are semigroups $\mathbf{T}_1, \mathbf{T}_2$ such that $\mathbf{T}_1 \in \mathbb{P}(\mathbf{S})$, $\mathbf{T}_2 \in \mathbb{S}(\mathbf{T}_1)$ and $\mathbf{T} \in \mathbb{H}(\mathbf{T}_2)$. Clearly, $\mathbf{T}_1$ is a monoid that is also contained in the monoid variety of $\mathbf{S}$. If $\mathbf{T}_2$ contains the identity element, e say, of $\mathbf{T}_1$ then $\mathbf{T}$ (a quotient of $\mathbf{T}_2$) is a monoid that is contained in the monoid variety of $\mathbf{S}$ and we are done.

Now assume that $e \notin T_2$ and let $\mathbf{T}_2'$ denote the submonoid of $\mathbf{T}_1$ obtained by adjoining the element e to T_2. Note that $\mathbf{T}_2$ may already be a monoid, but with an element other than e acting as the identity, whence $\mathbf{T}_2'$ need not be isomorphic to $\mathbf{T}_2^1$. Without loss of generality we may assume that there is a congruence θ such that $\mathbf{T}_2/\theta = \mathbf{T}$.

If $\mathbf{T}$ has an identity element 1, then choose $f \in T_2$ such that $f/\theta = 1$, and extend θ to a congruence θ' on $\mathbf{T}_2'$ by adjoining to θ the pair (e, e) and also the pairs (e, x) and (x, e) if (f, x) and (x, f) are in θ_1. This gives $\mathbf{T}_2'/\theta' \cong \mathbf{T}_2/\theta = \mathbf{T}$, showing $\mathbf{T} \in \mathbb{V}(\mathbf{S})$ as a monoid.

Otherwise, if $\mathbf{T}$ has no identity element, then we may extend θ to a congruence on $\mathbf{T}_2'$ by adjoining the pair (e, e). The resulting quotient is isomorphic to $\mathbf{T}^1$ and again lies in the monoid variety of $\mathbf{S}$. □

An equational deduction of an identity $p \approx q$ from a set of identities Σ, is a sequence of identities $p \equiv p_1 \approx p_2 \approx \cdots \approx p_n \equiv q$ such that for each $i < n$ there is an identity $u_i \approx v_i \in \Sigma$ or $v_i \approx u_i \in \Sigma$ and a semigroup substitution θ_i such that p_{i+1} is obtained from p_i by replacing a subword $\theta_i(u_i)$ with $\theta_i(v_i)$. In this case we write $\Sigma \vdash p \approx q$. If $n = 1$ we will interpret the definition of an equational deduction as saying that $\Sigma \vdash p \approx p$.

A basis for the identities of a semigroup $\mathbf{S}$ is a subset Σ of $\mathrm{Id}(\mathbf{S})$ such that $\Sigma \vdash \mathrm{Id}(\mathbf{S})$. In 1964, Oates and Powell [10] showed that if $\mathbf{G}$ is a finite group, then $\mathbf{G}$ has a finite basis of identities. However in 1966 Perkins (see [11]) showed that the semigroup $\mathbf{B}_2^1$ given by the following matrices under matrix multiplication has no finite identity basis:

$$\begin{pmatrix}0&0\\0&0\end{pmatrix}, \quad \begin{pmatrix}1&0\\0&1\end{pmatrix}, \quad \begin{pmatrix}1&0\\0&0\end{pmatrix}, \quad \begin{pmatrix}0&1\\0&0\end{pmatrix}, \quad \begin{pmatrix}0&0\\1&0\end{pmatrix}, \quad \begin{pmatrix}0&0\\0&1\end{pmatrix}.$$

In the subsequent years a great many other examples have shown that the *finite basis problem*—determining which finite semigroups admit a finite basis of identities—is very complicated. For general algebras this problem has been shown to be undecidable [8], but for semigroups the complexity of the problem remains open. For further details we direct the reader to the most recent survey of this area by Volkov [15].

The following lemma gives some well known equivalents of the finite basis property. We omit the proof, but note that the equivalence of (2) and (3) follows from Birkhoff's completeness theorem for equational logic (which says that $\Sigma \vdash p \approx q$ if and only if whenever some algebra $\mathbf{A}$ satisfies all of Σ, then also $\mathbf{A} \models p \approx q$), while the equivalence of (1) and (2) is another theorem of Birkhoff.

1.2 Lemma (Birkhoff [2]; see also [3]) *Let* $\mathbf{S}$ *be a finite algebra of finite signature. The following are equivalent:*

(1) $\mathrm{Id}(\mathbf{S})$ *is finitely based;*

(2) *there is an positive integer* n *such that the* n*-variable identities in* $\mathrm{Id}(\mathbf{S})$ *are a basis for* $\mathrm{Id}(\mathbf{S})$*;*

(3) *there is a positive integer* n *such that for all algebras* $\mathbf{A}$*, we have* $\mathbf{A} \in \mathbb{V}(\mathbf{S})$ *if and only if all* n*-generated subalgebras of* $\mathbf{A}$ *are in* $\mathbb{V}(\mathbf{S})$*.*

2 Syntactic equivalents

We are going to show that the finite basis problem for finite semigroups is equivalent to its restriction to finite syntactic semigroups. The idea is easy, so we prove it in an even more general setting.

The notion of a syntactic congruence can be extended to general algebras by defining the syntactic congruence $\sim_W$ of a subset W of an algebra $\mathbf{S}$ as the largest congruence θ for which W is a union of θ-classes (for further details see [1] for example). A *syntactic algebra* is an algebra with a *syntactic subset*—a subset whose syntactic congruence is the diagonal relation.

2.1 Proposition *If* $\mathbf{S}$ *is an algebra with a one element subalgebra then there is a syntactic algebra* $\bar{\mathbf{S}}$ *such that* $\mathbf{S} \cong \bar{\mathbf{S}}$. *If* $\mathbf{S}$ *is finite, then* $\bar{\mathbf{S}}$ *is finite.*

Proof Let $\{e\}$ be a one element subuniverse of $\mathbf{S}$. For each $t \in S \setminus \{e\}$, let $\mathbf{S}_t$ denote $\mathbf{S}/\sim_{\{t\}}$. As in Section 1.1, $\mathbf{S}$ subdirectly embeds into $\prod_{t \in S \setminus \{e\}} \mathbf{S}_t$, and so we have $\mathbf{S} \cong \prod_{t \in S \setminus \{e\}} \mathbf{S}_t$. For $s, t \in S \setminus \{e\}$, let $s_t \in \prod_{t \in S \setminus \{e\}} \mathbf{S}_t$ be defined by $s_t(r) = e/\sim_{\{r\}}$ if $r \neq t$ and $s/\sim_{\{r\}}$ otherwise. Let W be $\{s_s : s \in S\}$, let $\mathbf{T}$ be any subalgebra of $\prod_{t \in S \setminus \{e\}} \mathbf{S}_t$ containing $\{s_t : s, t \in S\}$, and let $\bar{\mathbf{S}}$ denote $\mathbf{T}/\sim_W$. Now $\bar{\mathbf{S}} \in \mathbb{V}(\mathbf{S})$ so to complete the proof it will suffice to show that each $\mathbf{S}_t$ embeds into $\bar{\mathbf{S}}$. Now for each $t \in S \setminus \{e\}$, the subalgebra of $\mathbf{T}$ on $\{s_t : s \in S\}$ is isomorphic to $\mathbf{S}_t$ and $\{s_t : s \in S\} \cap W = \{s_s\}$, a syntactic subset of $\{s_t : s \in S\}$. Hence the restriction of $\sim_W$ on $\mathbf{T}$ to $\{s_t : s \in S\}$ is trivial. Therefore $\mathbf{S}_t$ embeds into $\bar{\mathbf{S}}$, as required. □

Algebras satisfying the conditions of this proposition are reasonably common—every monoid, ring, lattice and finite semigroup has a one-element subalgebra. We note that Proposition 2.1 would follow trivially if the class of syntactic algebras was closed under the taking of direct products (as is suggested by [6, Proposition VII.1.5]); however a 2-element left zero semigroup is syntactic, while its square is not.

Unfortunately, the construction in Proposition 2.1 is not very efficient—it seems possible that $\bar{\mathbf{S}}$ could have size close to $|S|^{|S|}$. To gain greater efficiency, we now restrict our attention to semigroups.

2.2 Lemma *If* $\mathbf{S}$ *is a semigroup with zero element then one can construct a syntactic semigroup* $\bar{\mathbf{S}}$ *equationally equivalent to* $\mathbf{S}$ *in polynomial time.*

Proof We are going to follow the proof of Proposition 2.1. We can fix the zero element 0 of $\mathbf{S}$ as the one element subalgebra and then the elements $\{s_t : s, t \in S\}$ in $\prod_{t \in S \setminus \{0\}} \mathbf{S}_t$ actually form a subsemigroup, which we choose as $\mathbf{T}$.

Now $\mathbf{T}$ has fewer than $|S|^2$ elements and can clearly be constructed in polynomial time from the family $\{\mathbf{S}_t : t \in S \setminus \{0\}\}$; in fact $\mathbf{T}$ is isomorphic to the so-called 0-*direct join* of $\{\mathbf{S}_t : t \in S \setminus \{e\}\}$, formed by amalgamating these semigroups at 0 and then setting any undefined products to equal 0. As each of the $|S| - 1$ semigroups of the form $\mathbf{S}_t$ can be constructed from $\mathbf{S}$ in polynomial time (see Section 1.1) it follows that $\bar{\mathbf{S}}$ can be constructed in polynomial time from $\mathbf{S}$. □

For each $t \in S \setminus \{0\}$, let A_t be an alphabet (with $A_t \cap A_s = \emptyset$ for $s \neq t$) and $\nu_t : A_t^+ \to \mathbf{S}_t$ be a surjective morphism so that $\mathbf{S}_t$ is the syntactic semigroup of some (regular) subset $L_t \subseteq A_t^+$ with $\nu_t(L_t) = \{t/\sim_{\{t\}}\}$. One can see from the proof of Lemma 2.2 that $\bar{\mathbf{S}}$ is (isomorphic to)

the syntactic semigroup of the *disjoint* union of the languages L_t in the free semigroup over the disjoint union of the alphabets A_t.

The proof of Lemma 2.2 does not quite hold for monoids because 0 is not a one element submonoid and because the subsemigroup on $\{s_t : s, t \in S \setminus \{0\}\}$ is not a submonoid (except when $|S| = 2$). However, the submonoid of $\prod_{t \in S \setminus \{0\}} \mathbf{S}_t$ generated by $\{s_t : s, t \in S \setminus \{0\}\}$ is simply $\{s_t : s, t \in S \setminus \{0\}\} \cup \{\underline{1}\}$ (where $\underline{1}$ is the identity element). Choosing this as $\mathbf{T}$ and calculating $\bar{\mathbf{S}}$ as before, we find that each of the $\mathbf{S}_t$ are monoids that embed into $\bar{\mathbf{S}}$ as semigroups and hence are in the semigroup variety of the monoid $\bar{\mathbf{S}}$. By Lemma 1.1, they are in the monoid variety of $\bar{\mathbf{S}}$ as well. This is clearly still a polynomial time construction and hence Lemma 2.2 holds for monoids as well. In fact we do not need monoids with zero to state this result. First consider the following lemma.

2.3 Lemma *If* $\mathbf{S}$ *is a semigroup whose variety contains the variety of semilattices then* $\mathbf{S} \approxeq \mathbf{S}^0$.

Proof If $\mathbf{S} = \mathbf{S}^0$ we are done, so assume that $\mathbf{S}$ does not contain a zero element. Certainly $\mathbf{S}$ lies within the variety of $\mathbf{S}^0$. Conversely, as the variety of $\mathbf{S}$ contains the two element semilattice $\mathbf{2}$ (with universe $\{0, 1\}$), it also contains $\mathbf{S} \times \mathbf{2}$. The semigroup $\mathbf{S}^0$ is isomorphic to the Rees quotient of $\mathbf{S} \times \mathbf{2}$ by the ideal $\{(x, 0) : x \in S\}$. □

2.4 Lemma *Every finite monoid* $\mathbf{S}$ *is equationally equivalent to the syntactic monoid of some regular language that can be constructed from* $\mathbf{S}$ *in polynomial time.*

Proof By considering the one-generated submonoids of $\mathbf{S}$ it can be seen that the monoid variety of $\mathbf{S}$ fails to contain the variety of semilattice monoids if and only if $\mathbf{S}$ is a finite group. If $\mathbf{S}$ is a group then the syntactic congruence of any singleton subset of S is the diagonal relation and so $\mathbf{S}$ is the syntactic monoid of a regular language and we are done.

Now say that the monoid variety of $\mathbf{S}$ contains the variety of semilattice monoids. Then by Lemma 2.3 it can be generated by a monoid with zero element. A syntactic monoid generating the same variety is then given by the argument following Lemma 2.2. □

In [15], the question is asked as to whether or not the finite basis problem for syntactic semigroups is decidable (that is, algorithmically solvable). It is clear from Proposition 2.1 that this problem is equivalent to the general finite basis problem, but in fact we can use Lemma 2.2 to show that these problems are polynomially equivalent.

We first need the following result from [9].

2.5 Lemma (Mel'nik [9]) *A semigroup* $\mathbf{S}$ *is finitely based if and only if* $\mathbf{S}^0$ *is finitely based.*

Proof If $\mathbb{V}(\mathbf{S})$ contains the variety of semilattices then by Lemma 2.3 we have $\mathbf{S} \approxeq \mathbf{S}^0$. So we may assume throughout that $\mathbb{V}(\mathbf{S})$ does not contain the variety of semilattices.

Let us say that an identity $u \approx v$ is *homotypical* if the variables in u are the same as those in v (this is also called *regular*). The identities of $\mathbf{2}$ are exactly the class of homotypical semigroup identities and so it follows that $\mathbf{S}$ satisfies some non-homotypical identity. In this case $\mathbf{S}^0$ generates the smallest variety containing $\mathbb{V}(\mathbf{S})$ that is definable by homotypical identities. Following [9], it can be shown that $\mathbf{S}$ has a basis of the form $\Sigma \cup (x^m y^m x^m)^m \approx x^m$ for some $m \in \mathbb{N}$ and where Σ is a collection of homotypical identities. It is shown in [9] that one can find a basis for $\mathbf{S}^0$ by adjoining to Σ a finite set of homotypical identities. Thus if $\mathbf{S}$

is finitely based, then Σ can be chosen to be finite, and we get a finite basis for $\mathbf{S}^0$ (this is also discussed in [15]). Now if $\mathbf{S}$ is not finitely based, then Σ must be infinite and no finite subset of $\mathrm{Id}(\mathbf{S})$ is sufficient to derive all of Σ. As Σ is also satisfied by $\mathbf{S}^0$, and $\mathrm{Id}(\mathbf{S}^0) \subseteq \mathrm{Id}(\mathbf{S})$, it follows that $\mathbf{S}^0$ also has no finite basis of identities. □

2.6 Corollary *The finite basis problem for finite semigroups (monoids) is polynomially equivalent to its restriction to the class of finite syntactic semigroups (monoids).*

Proof The monoid version of this result follows immediately from Lemma 2.4.

Now let $\mathbf{S}$ be a finite semigroup. By Lemma 2.5, $\mathbf{S}$ is finitely based if and only if $\mathbf{S}^0$ is finitely based. By Lemma 2.2, $\mathbf{S}^0$ generates a variety equal to one generated by a single syntactic semigroup (that can be constructed in polynomial time from $\mathbf{S}^0$). □

3 A language theoretic approach

The finite basis problem for monoids has a natural analogue in terms of varieties of regular languages. Analogous statements will hold for semigroups as well.

3.1 Definition For K a class of regular languages, let $\mathcal{V}^*(K)$ denote the $*$-variety of languages generated by K and let $\mathcal{V}_n^*(K)$ denote the subclass of $\mathcal{V}^*(K)$ consisting of those languages which are in alphabets of at most n letters. We say that a $*$-variety of languages $\underline{\mathcal{V}}$ is *finitely $*$-verifiable for K languages* if there exists an $n \in \mathbb{N}$ such that for all languages $W \in K$,

$$W \in \underline{\mathcal{V}} \iff \mathcal{V}_n^*(W) \subseteq \underline{\mathcal{V}}.$$

If K is the class of all regular languages, then we say that a $*$-variety is finitely $*$-verifiable instead of finitely $*$-verifiable by K-languages.

The connection between the finite $*$-verification property for $*$-varieties of languages and the finite basis problem for monoids arises through the Eilenberg correspondence and the following result essentially due to M. Sapir.

3.2 Lemma (M. Sapir [14]) *Let $\mathcal{V}$ be a subvariety of a finitely generated variety of semigroups (monoids). Then $\mathcal{V}$ is non-finitely based if and only if for each $n \in \mathbb{N}$ there are finite semigroups (monoids, respectively) $\mathbf{S}_n \notin \mathcal{V}$ such that $\mathbf{S}_n$ satisfies all n-variable identities of $\mathcal{V}$ but not all identities of $\mathcal{V}$. Equivalently, $\mathbf{S}_n \notin \mathcal{V}$ but all n-generated subsemigroups (submonoids) of $\mathbf{S}_n$ are contained in $\mathcal{V}$.*

Proof We first discuss the semigroup case. The statement above differs from the main result of [14] only in that it allows for the possibility that $\mathcal{V}$ is not generated by any finite semigroup. All arguments in [14] except for Proposition 1 depend only on the local finiteness of $\mathcal{V}$. Proposition 1 however, is itself an extract from a more general result from [12] that concerns any locally finite variety in which all groups are finitely based. As Sapir notes in [12], this is true of any finitely generated semigroup variety (by [10]) and hence in any subvariety of such a variety. Therefore Sapir's proof in fact extends to the semigroup part of the lemma we have stated above.

Now we investigate the monoid case which again follows without significant change from Sapir's result for semigroups. As is discussed in [15], a monoid is non-finitely based (in the

type $\langle 2,0\rangle$) if and only if it is non-finitely based as a semigroup (in the type $\langle 2\rangle$). Let $\mathcal{V}'$ denote the semigroup variety generated by the members of $\mathcal{V}$ considered as semigroups, and say that $\mathcal{V}$ (whence $\mathcal{V}'$) is not finitely based. Using the semigroup version of the lemma, there are finite semigroups $\mathbf{S}_n$ such that $\mathbf{S}_n$ is not contained in $\mathcal{V}'$, while all n-generated subsemigroups of $\mathbf{S}_n$ are contained in $\mathcal{V}'$. Now consider $\mathbf{S}_n^1$, also not in $\mathcal{V}'$ and therefore not in $\mathcal{V}$ (as a monoid). As semigroups, the n-generated *submonoids* of $\mathbf{S}_n^1$ are either n-generated *subsemigroups* of $\mathbf{S}$ (that happen to be monoids) or are of the form of an n-generated subsemigroup of $\mathbf{S}_n$ with adjoined identity element. In either case, Lemma 1.1 shows that the n-generated submonoids lie in $\mathcal{V}$. Thus we have, for all $n \in \mathbb{N}$, a finite monoid not in $\mathcal{V}$ but whose n-generated submonoids are in $\mathcal{V}$. □

This result guarantees that for finite semigroups and monoids, the third condition of Lemma 1.2 can be replaced by its restriction to finite algebras. The corresponding result for general algebras appears to be an interesting open problem; see [4, 5].

3.3 Lemma *If* $\mathbf{S}$ *is a finite monoid with no finite basis of identities then for each* $n \in \mathbb{N}$ *there is a regular language* W_n *such that every* n*-generated submonoid of* $\mathrm{Syn}_{\mathrm{M}}(W_n)$ *is contained in* $\mathbb{V}(\mathbf{S})$ *but* $\mathrm{Syn}_{\mathrm{M}}(W_n) \notin \mathbb{V}(\mathbf{S})$.

Proof By Lemma 3.2, for each $n \in \mathbb{N}$ there is a finite monoid $\mathbf{S}_n$ such that every n-generated submonoid of $\mathbf{S}_n$ is contained in $\mathbb{V}(\mathbf{S})$ but $\mathbf{S}_n \notin \mathbb{V}(\mathbf{S})$.

Let us fix some arbitrary $n \in \mathbb{N}$. For each element $s \in \mathbf{S}_n$, let $\mathbf{S}_{n,s}$ denote the semigroup $\mathbf{S}_n/{\sim_{\{s\}}}$. Now each $\mathbf{S}_{n,s}$ is a quotient of $\mathbf{S}_n$ and therefore satisfies all identities satisfied by $\mathbf{S}_n$. However $\mathbb{V}(\mathbf{S})$ is generated by $\{\mathbf{S}_{n,s} : s \in S\}$ (see Section 1.1) and therefore there must be an identity of $\mathbf{S}$ that fails on $\mathbf{S}_{n,s}$ for some $s \in \mathbf{S}_n$. Therefore we have a finite syntactic monoid of a regular language, W_n say, that satisfies all identities of $\mathbf{S}_n$ (and therefore all n-variable identities of $\mathbf{S}$) but not all identities of $\mathbf{S}$. Equivalently (see Lemma 1.2), we have shown that there is a regular language W_n such that every n-generated submonoid of $\mathrm{Syn}_{\mathrm{M}}(W_n)$ lies in the monoid variety of $\mathbf{S}$ while $\mathrm{Syn}_{\mathrm{M}}(W_n)$ does not. □

The extension of Lemma 3.3 to semigroups is obvious and we omit the details.

3.4 Theorem *Let* $\mathbf{S}$ *be a finite monoid and* $\underline{\mathcal{V}}$ *denote the* $*$*-variety of languages whose syntactic monoids are finite and contained in* $\mathbb{V}(\mathbf{S})$. *Then* $\mathbf{S}$ *is finitely based if and only if* $\underline{\mathcal{V}}$ *is finitely* $*$*-verifiable.*

Proof Let $\mathbf{S}$ be a finite monoid with a finite basis of identities Σ in n variables and let $\underline{\mathcal{V}}$ denote the $*$-variety of regular languages whose syntactic monoids are in $\mathbb{V}_{\mathrm{fin}}(\mathbf{S})$. Obviously if $W \in \underline{\mathcal{V}}$ then $\mathcal{V}_n^*(W) \subseteq \underline{\mathcal{V}}$, while if W is a language not in $\underline{\mathcal{V}}$ then the Eilenberg correspondence guarantees that $\mathrm{Sym}_{\mathrm{M}}(W) \not\models \Sigma$. Hence there is an n-variable identity $u \approx v$ that fails on $\mathrm{Sym}_{\mathrm{M}}(W)$. Evidently, there is an n-generated submonoid $\mathbf{T}$ of $\mathrm{Sym}_{\mathrm{M}}(W)$ on which $u \approx v$ fails. By examining the syntactic quotients of $\mathbf{T}$ with respect to the singleton subsets of T (as in the proof of Lemma 3.3), it follows that there is $t \in T$ such that $u \approx v$ fails on $\mathbf{T}/{\sim_{\{t\}}}$, also n-generated. Let $A = \{a_1, \ldots, a_n\}$ be a set of free generators for a free monoid A^* and let $\phi : A^* \to \mathbf{T}$ be a surjective homomorphism. Then $\mathbf{T}/{\sim_{\{t\}}}$ is the syntactic monoid of the regular language $W_t := \phi^{-1}(t/{\sim_{\{t\}}})$ in the n-letter alphabet A, whence it follows that $W_t \in \mathcal{V}_n^*(W) \setminus \underline{\mathcal{V}}$. This shows that there is an $n \in \mathbb{N}$ such that for regular languages W, $W \notin \underline{\mathcal{V}} \implies \mathcal{V}_n^*(W) \not\subseteq \underline{\mathcal{V}}$, that is, $\underline{\mathcal{V}}$ is finitely $*$-verifiable.

Now assume that $\mathbf{S}$ is non-finitely based. By Lemma 3.3, for all $n \in \mathbb{N}$ there are regular languages W_n such that $\mathrm{Sym}_{\mathrm{M}}(W_n) \notin \mathbb{V}(\mathbf{S})$ but such that every n-variable identity satisfied by $\mathbf{S}$ is satisfied by $\mathrm{Sym}_{\mathrm{M}}(W_n)$. Now if $U \in \mathcal{V}_n^*(W_n)$ then $\mathrm{Sym}_{\mathrm{M}}(U)$ is n-generated and satisfies all identities of $\mathrm{Syn}_{\mathrm{M}}(W_n)$ and in particular, all n-letter identities of $\mathbf{S}$. Elementary arguments then show that $\mathrm{Sym}_{\mathrm{M}}(U) \in \mathbb{V}(\mathbf{S})$. Hence $U \in \underline{\mathcal{V}}$ so $\mathcal{V}_n^*(W_n)$ is a subclass of $\underline{\mathcal{V}}$, while W_n is not in $\underline{\mathcal{V}}$. Because $n \in \mathbb{N}$ is arbitrary, $\underline{\mathcal{V}}$ is not finitely $*$-verifiable. □

Again, trivial variations of this theorem give the corresponding result for semigroup varieties and $+$-varieties of languages.

It is interesting to note that the main proof in [14] shows that if $\mathbf{S}$ is a finite inherently non-finitely based monoid (or semigroup), then the corresponding $*$-variety of languages is not finitely $*$-verifiable for singleton languages. For example, the monoid $\mathbf{B_2^1}$ given in the introduction is known to be (isomorphic to) the syntactic monoid of the language $\{ab\}^*$ (in the alphabet $\{a, b\}$) and is also known to be inherently non-finitely based [13]. Hence $\mathcal{V}^*(\{ab\}^*)$ is not finitely $*$-verifiable for singleton languages. In fact from [11] and the proof of Theorem 3.4, we have $\{ba_1a_2 \dots a_nba_na_{n-1} \dots a_1\} \notin \mathcal{V}^*(\{ab\}^*)$ for any $n \in \mathbb{N}$, but $\mathcal{V}_n^*(\{ba_1a_2 \dots a_nba_na_{n-1} \dots a_1\}) \in \mathcal{V}^*(\{ab\}^*)$.

Let the *finite $*$-verification problem for regular expressions* denote the following decision problem: given a regular expression, decide if the corresponding language generates a finitely $*$-verifable $*$-variety. The finite $+$-verification problem is defined analogously.

3.5 Corollary *The finite basis problem for finite semigroups* (*monoids*) *is decidable if and only if the finite $+$-verification* (*$*$-verification, respectively*) *problem for regular expressions is decidable.*

Proof We discuss the monoid case; the semigroup case is almost identical. The standard proofs of the equivalence of the class of languages recognisable by finite state automata and languages corresponding to regular expressions are constructive. That is, with each finite automata, we can effectively associate a regular expression corresponding to the language recognised by the machine and vice versa. Likewise, with a syntactic monoid $\mathrm{Syn}_{\mathrm{M}}(W)$ of a language W we may construct an automata recognising W (by representing $\mathrm{Syn}_{\mathrm{M}}(W)$ as a monoid of transformations), and conversely with each automata, we may effectively construct a minimal automata recognising the same language and then construct the syntactic monoid of this language as the transition monoid of the automata. Therefore we can effectively associate a syntactic monoid $\mathbf{S}$ with a regular expression for a language recognised by $\mathbf{S}$ and vice versa.

The result now follows by combining Theorem 3.4 and its semigroup variant with Corollary 2.6. □

We note that the translation from syntactic semigroups or monoids to regular expressions is polynomial time, however the reverse may not be true. It is a well known PSPACE-complete problem to determine if a regular expression r over $\{0, 1\}$ represents the same language as $\{0, 1\}^*$; see [7]. The syntactic monoid of $\{0, 1\}^*$ (as a regular language in the alphabet $\{0, 1\}$) is trivial, however if r represents a non-empty language W distinct from $\{0, 1\}^*$, then the syntactic monoid of W is non-trivial and so $\mathrm{Syn}_{\mathrm{M}}(W)$ generates a non-trivial variety. Thus if P≠PSPACE, there can be no polynomial time method for constructing, from a regular expression r for a language W, a monoid generating the same variety as $\mathrm{Syn}_{\mathrm{M}}(W)$.

References

[1] J. Almeida, *Finite Semigroups and Universal Algebra*, Series in Algebra **3**, World Scientific Publishing, Singapore, 1994.

[2] G. Birkhoff, On the structure of abstract algebras, *Proc. Camb. Philos. Soc.* **31** (1935), 433–454.

[3] S. Burris and H. Sankappanavar, *A Course in Universal Algebra*, Graduate Texts in Mathematics **78**, Springer Verlag, New York, 1980.

[4] R. Cacioppo, Finite bases for varieties and pseudovarieties, *Algebra Universalis* **25** (1988), 263–280.

[5] R. Cacioppo, Nonfinitely based pseudovarieties and inherently nonfinitely based varieties, *Semigroup Forum* **47** (1993), 223–226.

[6] S. Eilenberg, *Automata, Languages and Machines Vol. B*, Pure and Applied Mathematics, Academic Press, New York, 1976.

[7] M. Garey and D. Johnson, *Computers and Intractability: A Guide to the theory of NP-Completeness*, W. H. Freeman, New York, 1979.

[8] R. McKenzie, Tarski's finite basis problem is undecidable, *Internat. J. Algebra Comput.* **6** (1996), 49–104.

[9] I. I. Mel'nik, On varieties and lattices of varieties of semigroups, in: *Studies in algebra No. 2* (V. V. Vagner, ed.), Izdat. Saratov. Univ., Saratov (1970), 47–57 (in Russian).

[10] S. Oates and M. B. Powell, Identical relations in finite groups, *J. Algebra* **1** (1964), 11–39.

[11] P. Perkins, Bases for equational theories of semigroups, *J. Algebra* **11** (1969), 298–314.

[12] M. V. Sapir, Problems of Burnside type and the finite basis property in varieties of semigroups, *Math. USSR Izv.* **30** (2) (1988), 295–314.

[13] M. V. Sapir, Inherently nonfinitely based finite semigroups, *Math. USSR-Sb.* **61** (1) (1988), 155–166.

[14] M. Sapir, Sur la propiété de base finie pour les pseudovariétés de semigroupes finis, *C. R. Acad. Sci. Paris (Sér. I)* **306** (1988), 795–797 (in English and French).

[15] M. V. Volkov, The finite basis problem for finite semigroups *Sci. Math. Jpn.* **53** (2001), 171–199.

Endoprimal algebras

Kalle KAARLI*

Institute of Pure Mathematics
University of Tartu
50090 Tartu
Estonia

László MÁRKI†‡

A. Rényi Institute of Mathematics
Hungarian Academy of Sciences
H-1364 Budapest, Pf. 127
Hungary

Abstract

This survey is an overview of investigations into endoprimal algebras: their general properties and connections with endoduality, as well as descriptions of endoprimal objects in various kinds of algebras.

1 Early development

An algebra is called primal if every finitary function defined on it is a term function. The most common primal algebra is the 2-element Boolean algebra. Primal algebras were introduced when studying categorical properties of the variety of Boolean algebras. Subsequently, several generalisations have been investigated: algebras in which the term functions are exactly the functions that preserve some derived structure (see e.g. [17]). Following this line, an algebra is called *endoprimal* if its term functions are precisely those functions which permute with all endomorphisms. It is convenient to call the latter functions *endofunctions* of the given algebra. Thus, an algebra is endoprimal if and only if its endofunctions are term functions.

Endoprimal algebras made their first appearance, however, in a different way: in the course of investigations into duality theory. Without using this name, B. A. Davey [2] proved in 1976 that every finite chain is endoprimal as a Heyting algebra. In 1985 B. A. Davey and H. Werner [8] proved that the Heyting algebra $\mathbf{2^2} \oplus \mathbf{1}$ is also endoprimal, and this paper marks the appearance of the name 'endoprimal'. (Here and in the sequel $A \oplus 1$ is the ordered set A

*The first author thanks the Estonian Science Foundation grant no. 5368 for support.

†The second author's research was partially supported by the Hungarian National Foundation for Scientific Research grants T34530 and T43034.

‡Both authors thank the Estonian and the Hungarian Academies of Sciences for making possible mutual visits through their exchange agreement.

V. B. Kudryavtsev and I. G. Rosenberg (eds.),
Structural Theory of Automata, Semigroups, and Universal Algebra, 169–180.

with new greatest element 1.) In a note published in 1990, A. A. L. Sangalli [22] rephrased the definition of endoprimality by using clones, and gave three examples for endoprimal algebras: primal algebras, sets of cardinality different from 2, and (arbitrary) free algebras on infinite sets.

The next result appeared in 1993, when L. Márki and R. Pöschel [20] proved that a distributive lattice is endoprimal if and only if it is not relatively complemented. This result seemed to be the end of the story, since it showed that endoprimal algebras do not share any of the 'nice' properties of (quasi)primal algebras; hence one could not expect much of their investigation. Contrary to this expectation, [20] has become the starting point of further research, and the impetus for this came from B. A. Davey.

2 General theory and links with endoduality

In the paper [3] written in 1994, B. A. Davey showed that the occurrence of endoprimal algebras in duality theory is not an accident: every endodualisable finite algebra (i.e., every finite algebra which admits a particular kind of natural duality) is endoprimal. Then B. A. Davey, M. Haviar, and H. A. Priestley [4] explained how the arguments used by L. Márki and R. Pöschel lift to a duality-theoretic setting and yield that the class of endoprimal distributive lattices coincides with that of endodualisable distributive lattices. Finally, in 1997, B. A. Davey and J. Pitkethly [7] managed to carry over the argumentation to a wide range of quasi-varieties generated by a single algebra.

Let us now take a closer look at the connection between endoprimality and endodualisability. Without defining the latter notion, we briefly explain why endoprimality of a finite algebra is an easy consequence of its endodualisability. All details can be found in the monograph [1]. Suppose we are given a finite algebra $\mathbf{M}$. On M, the underlying set of $\mathbf{M}$, a new structure $\underset{\sim}{\mathbf{M}}$ is introduced. This structure may involve finitary (partial) functions on M and finitary relations on M. For building up the duality it is essential to assume that $\underset{\sim}{\mathbf{M}}$ is equipped with the discrete topology but this is not important for our considerations. The functions and relations involved are assumed to be algebraic, that is, the functions are homomorphisms from $\mathbf{B}^n$ to $\mathbf{M}$ where $\mathbf{B}$ is a subalgebra of $\mathbf{M}$ and relations are subalgebras of some $\mathbf{M}^n$ where n is a natural number. In particular, it makes sense to consider the case $\underset{\sim}{\mathbf{M}} = \langle M; \operatorname{End}\mathbf{M}\rangle$.

One of the main questions of duality theory is: which structures $\underset{\sim}{\mathbf{M}}$ dualise $\mathbf{M}$? The First Duality Theorem (see [1]) asserts that if this happens then the morphisms of $\underset{\sim}{\mathbf{M}}^n$ to $\underset{\sim}{\mathbf{M}}$ are precisely the n-ary term functions of $\mathbf{M}$. If $\underset{\sim}{\mathbf{M}} = \langle M; \operatorname{End}\mathbf{M}\rangle$ dualises $\mathbf{M}$ then we say that $\mathbf{M}$ is endodualisable. Since clearly in this case the morphisms from $\underset{\sim}{\mathbf{M}}^n$ to $\underset{\sim}{\mathbf{M}}$ are the n-ary endofunctions of $\mathbf{M}$, it becomes obvious that endodualisable algebras are endoprimal.

Note that the finiteness of $\mathbf{M}$ is essential for the theory of natural dualities. Methods of duality theory can therefore be directly used only for constructing finite endoprimal algebras. Fortunately, there is a way to go beyond finiteness as the next theorem shows.

2.1 Theorem [7, Theorem 1.2] *Let* $\mathbf{M}$ *be an endoprimal algebra and* $\mathbf{A} \in \mathbb{ISP}(\mathbf{M})$. *If* $\mathbf{A}$ *has a retract isomorphic to* $\mathbf{M}$ *then* $\mathbf{A}$ *is endoprimal.*

It is worth emphasizing that duality theory is good for providing examples of endoprimal algebras but has very rarely been used for proving that a given algebra is not endoprimal.

Results of the latter kind can be obtained using the following basic theorem.

2.2 Theorem [1, Proposition 2.2.3] *Let* $\mathbf{M}$ *be a finite algebra and* $\mathcal{A} = \mathbb{ISP}(\mathbf{M})$. *Then* $\mathbf{M}$ *is endoprimal if and only if* $\operatorname{End}\mathbf{M}$ *yields duality on all finitely generated free algebras of* $\mathcal{A}$.

In most cases, however, non-endoprimality is proved by the explicit construction of an endofunction which is not a term function. A model argument is that used in [20]: a relatively complemented distributive lattice is not endoprimal because the ternary function $f(x, y, z) = u$, where u is the complement of y in the interval $[x \wedge y, y \vee z]$, permutes with all endomorphisms but is not a term function, since it does not preserve the order.

It was noticed already in [7] that there are finite endoprimal algebras which are not endodualisable; e.g., $\mathbf{2^2} \oplus \mathbf{1}$ regarded as a bounded semilattice is endoprimal but not endodualisable. Thus certainly, even in the case of finite algebras, duality theory is not sufficient for characterising endoprimal algebras. The problem has arisen of how far actually endoprimality is from endodualisability. Perhaps the main result in this direction is the criterion given by M. Haviar and H. A. Priestley.

2.3 Theorem [13, Theorem 6] *Assume that a finite algebra* $\mathbf{D}$ *is dualised by the structure* $\underset{\sim}{\mathbf{D}}$ *which beside* $\operatorname{End}\mathbf{D}$ *involves finitary algebraic relations* $s_1, \ldots, s_m$. *Let* $\mathbf{F}$ *be a finitely generated free algebra in* $\mathcal{D} = \mathbb{ISP}(\mathbf{D})$. *If the algebras* $s_1, \ldots, s_m$ *are retracts in* $\mathbf{F}$ *and* $\mathbf{D}$ *is a retract in a finite endoprimal algebra* $\mathbf{M}$, *then* $\mathbf{M}$ *is endodualisable.*

One can apply the theorem to prove that in the classes of distributive lattices, double Stone algebras, median algebras, and finite abelian groups endoprimal algebras are endodualisable. However, which median algebras are endoprimal seems to be an open question.

It is well known that the term functions of a finite algebra $\mathbf{M}$ can be characterised as the functions which preserve all subalgebras of the finite powers of $\mathbf{M}$. Hence, in view of the standard Galois connection between clones and relational clones the endoprimality of a finite algebra $\mathbf{M}$ means, in fact, that $\operatorname{End}\mathbf{M}$ (viewed as the set of all graphs of endomorphisms of $\mathbf{M}$) generates the relational clone of all subalgebras of finite powers of $\mathbf{M}$ or, in the language of [6], that this set *clone entails* all finitary algebraic relations on $\mathbf{M}$. By a well-known result in clone theory (cf. [19]), the latter is equivalent to the requirement that every algebraic relation on $\mathbf{M}$ can be obtained from graphs of endomorphisms using certain well-defined natural constructs like projections, relational product, etc. The other entailment comes from duality theory. Our limited space does not allow us to give a strict definition of this notion; it serves to reduce the size of the set of algebraic relations on $\mathbf{M}$ used for providing duality, so that if the structure $\mathbf{M} = \langle M; R\rangle$ dualises $\mathbf{M}$ and a relation $r \in R$ is *duality entailed* by $R \setminus r$, then the structure $\langle \widetilde{M}; R \setminus r\rangle$ dualises $\mathbf{M}$, too. In particular, if R contains all endomorphisms of $\mathbf{M}$ and we are able to show that $\operatorname{End}\mathbf{M}$ duality entails every element of $R \setminus \operatorname{End}\mathbf{M}$ then $\mathbf{M}$ is endodualisable. A list of constructs which allow us to get from a given set of relations R every relation r duality entailed by R was first presented in [5]. This list is smaller than that for clone entailment: in particular, it does not include relational product. This shows once again that endoprimality is a consequence of endodualisability but not vice versa. A detailed analysis of the difference between the two entailments on the example of Kleene algebras was given by B. A. Davey, M. Haviar, and H. A. Priestley [6].

The next step makes use of an idea coming from the study of affine completeness. Recall that an algebra is called *affine complete* if all its congruence compatible functions (i.e.,

functions which preserve all congruences of the algebra) are polynomials. Suppose we want to prove that an algebra $\mathbf{A}$ is affine complete. We take a congruence compatible function f on A and try to prove that it is a polynomial function. $\mathbf{A}$ is often given as a subdirect product of some smaller algebras $\mathbf{A}_i$, $i \in I$. Then, since f is compatible with all projection kernels, it decomposes into a system of coordinate functions $f = (f_i)_{i\in I}$. Trying to handle an endofunction f in a similar way, we face the problem that f need not in general be compatible with all congruence kernels. Still the argument goes through if all subdirect factors $\mathbf{A}_i$ are isomorphic to subalgebras of $\mathbf{A}$, and hence the projection kernels are kernels of endomorphisms of $\mathbf{A}$. This observation is the underlying idea of the paper [18] by K. Kaarli and H. A. Priestley. One of the main results of the paper shows that if an algebra has a 'good' subdirect decomposition then its endofunctions are determined by what they induce on the largest subdirect factor.

2.4 Theorem [18, Corollary 2.4] *Let $\mathbf{A}$ be a subdirect product of the algebras $\mathbf{A}_i$, $i \in I$, and assume that there is an algebra $\mathbf{M}$ satisfying the following conditions:*

(SD1) *every $\mathbf{A}_i$ is a subalgebra of $\mathbf{M}$;*

(SD2) *$\mathbf{A}$ has a subalgebra isomorphic to $\mathbf{M}$;*

(SD3) *some of the $\mathbf{A}_i$ is isomorphic to $\mathbf{M}$.*

Then for every endofunction f of $\mathbf{A}$ there exists an endofunction g of $\mathbf{M}$ such that $f_i = g|_{A_i}$ for every $i \in I$.

In fact, these conditions (SD1)–(SD3) are satisfied in several cases where the endoprimality problem has been succesfully solved. As a corollary one immediately gets that, under these conditions, endoprimality of $\mathbf{M}$ yields that of $\mathbf{A}$, a fact that actually follows also from Theorem 2.1. The usefulness of Theorem 2.4 becomes more obvious if we apply it in the case when the algebra has a majority term. Recall that a *majority function* is a ternary function f satisfying the identities

$$f(x, x, y) = f(x, y, x) = f(y, x, x) = x$$

and a *majority term* for an algebra $\mathbf{A}$ is a ternary term which induces a majority function in $\mathbf{A}$. The median function $(x \vee y) \wedge (y \vee z) \wedge (z \vee x)$ is a majority term for all algebras with lattice reduct.

The next theorem follows from Theorem 2.4 and the well-known Baker-Pixley Lemma which essentially asserts, in particular, that a function defined on a finite algebra $\mathbf{M}$ with a majority term is a term function if and only if it preserves all subalgebras of $\mathbf{M}^2$. We denote by $\mathbf{S}_2(\mathbf{M})$ the algebra whose underlying set is the set of all subalgebras of $\mathbf{M}^2$ and whose fundamental operations are the relational product $\circ$, taking the converse $^\smile$, intersection $\cap$, and two nullary operations: one corresponding to the diagonal Δ_M and another corresponding to the whole square $\mathbf{M}^2$.

2.5 Theorem [18, Proposition 3.3] *Assume that the assumptions of Theorem 2.4 are fulfilled, $\mathbf{M}$ is finite and has a majority term. If $\mathbf{S}_2(\mathbf{M})$ is generated by all graphs of homomorphisms $\mathbf{A}_i \to \mathbf{A}_j$ and all 2-fold coordinate projections A_{ij}, $i, j \in I$, then $\mathbf{A}$ is endoprimal.*

Applications of this theorem are given in the next section.

3 Enriched distributive lattices

In this section we survey the results on endoprimality of algebras having a distributive lattice reduct.

We start with the observation that the result of L. Márki and R. Pöschel [20] on distributive lattices follows easily from Theorem 2.5. Indeed, it is well known that a distributive lattice is relatively complemented if and only if it cannot be mapped homomorphically onto the 3-element chain. Now, if a distributive lattice $\mathbf{L}$ is not relatively complemented, it has a subdirect decomposition with all factors isomorphic to the 2-element chain $\mathbf{2}$ and at least one 2-fold coordinate projection isomorphic to the 3-element chain. Clearly then $\mathbf{S}_2(\mathbf{2})$ is generated by 2-fold coordinate projections of $\mathbf{L}$, hence $\mathbf{L}$ is endoprimal.

Using similar, only slightly more complicated arguments, one can derive from Theorem 2.5 the description of endoprimal non-Boolean Stone algebras, originally obtained by B. A. Davey and J. G. Pitkethly [7], and this was done by K. Kaarli and H. A. Priestley in [18]. Formally the result presented in [18] is stronger because in [7] it was required that $\mathbf{4}$ is a retract, not only a homomorphic image, but it is easy to see that in the present situations these requirements are equivalent. Recall that a Stone algebra is a bounded distributive lattice with an extra unary operation * such that the following identities are satisfied:

$$(x \wedge y)^* = x^* \vee y^*, \qquad (x \vee y)^* = x^* \wedge y^*, \qquad 0^* = 1, \qquad 1^* = 0, \tag{3.1}$$

$$x \wedge x^* = 0, \qquad x^* \vee (x^*)^* = 1. \tag{3.2}$$

3.1 Theorem [18, Theorem 5.1] *A non-Boolean Stone algebra is endoprimal if and only if it has the* 4*-element Stone chain as a homomorphic image.*

Note that all Boolean algebras are endoprimal by Theorem 2.1 as is, in fact, every algebra in the variety generated by a (finite) primal algebra. Also note that, as pointed out by B. A. Davey and J. G.Pitkethly [7], the endoprimality criterion for non-Boolean Stone algebras can be given in internal terms. Given a Stone algebra $\mathbf{S}$, an element $x \in S$ is called *dense* if $x^* = 0$. The set of all dense elements of $\mathbf{S}$ is denoted by $d(\mathbf{S})$. It is easy to see that $d(\mathbf{S})$ is a sublattice of $\mathbf{S}$. Now, it is known that the lattice $d(\mathbf{S})$ is relatively complemented if and only if $\mathbf{4}$ is not a homomorphic image of $\mathbf{S}$. Hence, a non-Boolean Stone algebra $\mathbf{S}$ is endoprimal if and only if $d(\mathbf{S})$ is not relatively complemented.

M. Haviar and H. A. Priestley [12] described finite endoprimal double Stone algebras (hence, by Theorem 2.2, also finite endodualisable double Stone algebras). It is pretty clear that one should be able to derive the same results from Theorem 2.5. Recall that a double Stone algebra is a bounded distributive lattice $\mathbf{L}$ with two extra unary operations, * and $^+$, such that the reducts $\langle \mathbf{L}; ^* \rangle$ and $\langle \mathbf{L}; ^+ \rangle$ are a Stone algebra and a dual Stone algebra, respectively. A double Stone algebra is called regular if it satisfies the condition $x \vee x^* \geq y \wedge y^+$. The core of a double Stone algebra $\mathbf{A}$ is its subset $K(\mathbf{A}) = \{x \in A \mid x^* = 0, x^+ = 1\}$. Recall that every n-element chain has a unique structure of a double Stone algebra; the latter will be denoted by $\mathbf{n}$.

3.2 Theorem [12, Theorems 6.1–6.3]

(1) *A finite non-regular double Stone algebra with non-empty core is endoprimal if and only if it has* $\mathbf{5}$ *as a retract.*

(2) *A finite non-regular double Stone algebra with empty core is endoprimal if and only if it has* $\mathbf{5} \times \mathbf{2}^2$ *as a retract.*

(3) *A finite regular non-Boolean double Stone algebra is endoprimal if and only if it has* $\mathbf{3} \times \mathbf{2}^2$ *as a retract.*

In [6], B. A. Davey, M. Haviar, and H. A. Priestley study endodualisability and endoprimality of Kleene algebras. Recall that a Kleene algebra is a bounded distributive lattice with an extra unary operation * such that beside (3.1) the identities

$$(x^*)^* = x, \qquad (x \wedge x^*) \vee (x \vee x^*) = x \vee x^*$$

are satisfied. It turns out that from the point of view of endoprimality Kleene algebras are much worse than Stone algebras, a fact strikingly different from the situation in the case of affine completeness (cf. [17]). In view of [18] the reason is the following. The variety of all Kleene algebras is generated by the 3-element Kleene chain $\mathbf{3}$ but there are Kleene algebras in which $\mathbf{3}$ is not a subalgebra. Therefore Theorem 2.5 does not apply to all Kleene algebras. It may apply, however, in quasivarieties generated by an appropriate Kleene algebra different from $\mathbf{3}$, but there are too many quasivarieties of this kind. Actually the Kleene algebras containing $\mathbf{3}$ as a subalgebra are not endoprimal. Indeed, if $\mathbf{3}$ is a subalgebra of a Kleene algebra $\mathbf{K}$ then $\mathbf{K}$ contains an element a such that $a^* = a$. Then it is easy to see that the constant function with value a is an endofunction but not a term function of $\mathbf{K}$.

The paper [6] focuses mainly on the quasivariety $\mathcal{L}$ generated by the 4-element Kleene chain $\mathbf{4}$; its results on endoprimality were improved by K. Kaarli and H. A. Priestley [18].

3.3 Theorem [6, Theorem 5.2], [18, Theorem 6.1] *A non-Boolean Kleene algebra* $\mathbf{A} \in \mathcal{L}$ *is endoprimal if it has the* 1*-generated free Kleene algebra and the 6-element Kleene chain as homomorphic images. If* $\mathbf{A}$ *is finite then this condition is also sufficient.*

We conclude the section with Heyting algebras. Recall that a Heyting algebra can be defined as a bounded distributive lattice with an extra binary operation $\to$ such that the following identities are satisfied:

$$x \to x = 1, \qquad (x \to y) \wedge y = y, \qquad x \wedge (x \to y) = x \wedge y,$$
$$x \to (y \wedge z) = (x \to y) \wedge (x \to z), \qquad (x \vee y) \to z = (x \to z) \wedge (y \to z).$$

The standard example of a Heyting algebra is a Boolean algebra with $x \to y = x' \vee y$. As we have mentioned already, Heyting algebras provided the first nontrivial examples for endoprimal algebras. It is easy to see that every chain has a unique structure of a Heyting algebra. As we did above in case of Kleene, Stone and double Stone algebras, we shall denote the n-element Heyting chain simply by $\mathbf{n}$. The following theorem is due to B. A. Davey and J. G. Pitkethly [7].

3.4 Theorem [7, Theorems 4.1 and 4.2] *If* $\mathbf{M}$ *is a finite Heyting chain or the Heyting algebra* $\mathbf{2}^2 \oplus \mathbf{1}$ *then all algebras of the variety generated by* $\mathbf{M}$ *are endoprimal.*

We see that beside the varieties generated by a primal algebra there are other ones with the property that all their members are endoprimal. It might be an interesting problem to try to describe all varieties with this property, at least under the restrictions that they are finitely generated and admit a majority term.

4 Abelian groups

Throughout this section, *group* will mean abelian group. We shall use the notations $\mathbb{Z}$, $\mathbb{Z}_n$, and $\mathbb{Q}$ for the (additive) groups (or sometimes, by abuse, for the sets or rings) of integers, integers mod n, and rational numbers, respectively. The letter p will denote an arbitrary prime number. For undefined notions and notations concerning groups we refer to [9]. Recall that the term functions of abelian groups are the functions of the form $f(x_1, \dots, x_n) = k_1x_1 + \cdots + k_nx_n$ with some integers $k_1, \dots, k_n$.

We begin by giving some easy obstacles to endoprimality of a group; these were established in [14]. All results in this section are due to K. Kaarli and L. Márki, in some cases with other co-authors.

4.1 Proposition *If a torsion-free group is p-divisible for a prime p then the group is not endoprimal.*

Indeed, in such a group the function $f(x) = x/p$ is an endofunction but not a term function.

4.2 Proposition [14, Proposition 2.1] *If* $\operatorname{End}\mathbf{A}$ *is isomorphic to a subring of* $\mathbb{Q}$ *then the group* $\mathbf{A}$ *is not endoprimal.*

In fact, if $\operatorname{End}\mathbf{A}$ is isomorphic to a subring of $\mathbb{Q}$ then $\mathbf{A}$ must be a torsion-free group, and the function

$$f(x,y) = \begin{cases} 0 & \text{if } y = 0, \\ x & \text{if } y \neq 0 \end{cases}$$

in A is an endofunction but not a term function. Consequently, no torsion-free group of rank 1 is endoprimal.

If the centre of a ring $\mathbf{R}$ consists only of integral multiples of the identity map 1_R then we say that the centre of $\mathbf{R}$ is trivial.

4.3 Proposition *A group whose endomorphism ring has non-trivial centre cannot be endoprimal.*

Indeed, any central element of the endomorphism ring is a unary endofunction, and if a central element is not trivial (i.e., not of the form nx for some $n \in \mathbb{Z}$) then it is not a term function.

It is a very useful observation that endofunctions of a direct sum of two groups are sums of endofunctions of the direct summands. If there are no non-trivial homomorphisms between these summands, then endoprimality reduces even beyond endoprimality of the summands.

4.4 Proposition [14, Corollary 2.6] *Let* $\mathbf{A}$ *be the direct sum of nonzero groups* $\mathbf{B}$ *and* $\mathbf{C}$ *with* $\operatorname{Hom}(\mathbf{B},\mathbf{C}) = \operatorname{Hom}(\mathbf{C},\mathbf{B}) = 0$. *Then* $\mathbf{A}$ *is endoprimal if and only if* $\mathbf{B}$ *and* $\mathbf{C}$ *are bounded and endoprimal.*

Let us show how boundedness (that is, having finite exponent) comes into play. If, say, $\mathbf{B}$ is unbounded then the function $f = f|_B + f|_C$ with $f|_B = 1_B$, and $f|_C = 0$, is an endofunction of $\mathbf{A}$ which is not a term function. On the other hand, if both $\mathbf{B}$ and $\mathbf{C}$ are bounded then

the Chinese Remainder Theorem applies to show that every pair of n-ary term functions of the components is induced by a common (n-ary) term.

Let us give also a 'prototype' of examples for endoprimal groups.

4.5 Theorem [14, Theorem 2.7] *Let* $\mathbf{A}$ *be the direct sum of a group* $\mathbf{B}$ *and the infinite cyclic group* $\mathbb{Z}$. *Then* $\mathbf{A}$ *is endoprimal if and only if* $\mathbf{B}$ *is not bounded.*

Heuristically, in the case when $\mathbf{B}$ is unbounded, the proof relies on the fact that $\mathbb{Z}$ maps 'everywhere' in $\mathbf{B}$ via endomorphisms. This forces both $f|_{\mathbb{Z}}$ and $f|_B$ to be term functions for any endofunction f in $\mathbf{A}$, and then, by the unboundedness of $\mathbf{B}$, these two term functions must coincide.

Here are the main results of [14].

4.6 Theorem [14, Theorem 3.1] *A torsion group* $\mathbf{A}$ *is endoprimal if and only if it is bounded, say of exponent* m, *and* $\mathbb{Z}_m \oplus \mathbb{Z}_m$ *embeds into* $\mathbf{A}$.

(For bounded groups, this was proved earlier by B. A. Davey and J. Pitkethly [7].)

To present results on the torsion-free case, we recall the definitions of some well-known notions. By $\chi(a)$ we denote the characteristic sequence of the element a (i.e., the sequence of the p-heights of a under a fixed ordering of the primes). The members of such a sequence are non-negative integers and ∞. Two characteristic sequences are said to be equivalent if they differ only at finitely many places, and not at places where ∞ occurs. The equivalence classes of characteristic sequences are called *types*. The type $\mathbf{t}(a)$ of an element a is the type containing $\chi(a)$. It is well known that all nonzero elements of a torsion-free group $\mathbf{A}$ of rank 1 have the same type. This type is denoted by $\mathbf{t}(\mathbf{A})$. Conversely, every type can be realised as the type of a torsion-free group of rank 1. Clearly, the natural (placewise) ordering of characteristic sequences carries over to an ordering of types.

4.7 Theorem [14, Theorem 4.1] *Let a torsion-free group* $\mathbf{A}$ *decompose into* $\mathbf{A} = \mathbf{B} \oplus \mathbf{C}$ *where* $\mathbf{B}$ *is nontrivial,* $\mathbf{C}$ *has rank* 1 *and its type does not contain infinity, and suppose that* $\mathbf{t}(\mathbf{C}) \leq \mathbf{t}(b)$ *for every* $b \in B$. *Then* $\mathbf{A}$ *is endoprimal.*

4.8 Theorem [14, Theorem 4.3] *A torsion-free group* $\mathbf{A}$ *of rank* 2 *is endoprimal if and only if* $\mathbf{A} = \mathbf{B} \oplus \mathbf{C}$ *where* $\mathbf{B}$ *and* $\mathbf{C}$ *are groups of rank* 1, $\mathbf{t}(\mathbf{B}) \geq \mathbf{t}(\mathbf{C})$, *and* $\mathbf{C}$ (*or, equivalently,* $\mathbf{A}$) *is not* p*-divisible for any prime* p.

For mixed groups the following was obtained in [14]. Recall that a p-group is reduced if and only if it does not contain the Prüfer group $\mathbb{Z}_{p^\infty}$.

4.9 Theorem [14, Corollary 5.5] *If the torsion part* $\mathbf{T}$ *of a mixed group* $\mathbf{A}$ *is bounded* (*in which case* $\mathbf{T}$ *is a direct summand of* $\mathbf{A}$) *then* $\mathbf{A}$ *is endoprimal if and only if* $\mathbf{A}/\mathbf{T}$ *is endoprimal.*

In [14] the endoprimality problem was also solved for groups of torsion-free rank 1 with splitting torsion part. Since this result was later generalised to the non-splitting case (see Theorem 4.12), we do not formulate it here.

Thus we see that all major results of [14] relate to groups which have a suitable direct decomposition. In view of this fact and the above results, the following classes of groups come into question for a complete description of their endoprimal members.

(1) Torsion-free groups which decompose into a direct sum of rank 1 groups.

(2) Torsion-free groups of rank 3.

(3) Mixed groups whose maximal torsion-free factor group has rank 1.

Recently we have managed to settle all these cases.

For case (1) the problem could be solved for an even larger class of groups by R. Göbel, K. Kaarli, L. Márki, and S. L. Wallutis [10]. To state the result, we need the following definitions.

A torsion-free group $\mathbf{A}$ is said to be *separable* if every finite subset of A is contained in some direct summand of $\mathbf{A}$ which decomposes into a direct sum of rank 1 groups. If $\mathbf{A}$ is a separable torsion-free group then a type τ is said to be *critical* for $\mathbf{A}$ if $\mathbf{A}$ has a rank-1 direct summand of type τ. The set of critical types for a group $\mathbf{A}$ will be denoted by $T_{\mathrm{cr}}(\mathbf{A})$.

For a set T of types we define the *type graph of* T, $\Gamma(T)$, to be the undirected graph with T as the set of vertices and edges (τ_1, τ_2) for comparable $\tau_1, \tau_2 \in T$, i.e. $\tau_1 \leq \tau_2$ or $\tau_2 \leq \tau_1$. Recall that a graph is said to be *connected* if between any two vertices there is a finite path.

4.10 Theorem [10, Theorem 9] *Let* $\mathbf{A}$ *be a torsion-free separable group and* $T = T_{\mathrm{cr}}(\mathbf{A})$ *be the set of all critical types of* $\mathbf{A}$*. Then* $\mathbf{A}$ *is endoprimal if and only if* $\mathrm{rk}(\mathbf{A}) \geq 2$*,* $\mathbf{A}$ *is not* p*-divisible for any prime* p*, and the type graph* $\Gamma(T)$ *is connected.*

So far, all endoprimal groups we have seen admit 'good' direct decompositions. This is not a must, however: in [10] it is shown that there exist endoprimal indecomposable groups of arbitrarily large cardinality. The proof relies on a construction of R. Göbel and S. Shelah [11]. On the other hand, since any proper subring $\mathbf{R}$ of $\mathbb{Q}$ can be realised as the endomorphism ring of torsion-free groups of arbitrarily large cardinality, Proposition 4.1 tells us that there are indecomposable torsion-free groups of arbitrarily large cardinality which are not endoprimal. This leads to the (sad?) conclusion that it is hopeless to find a reasonable description of all endoprimal abelian groups.

In case (2), the following was proved by K. Kaarli and K. Metsalu [16].

4.11 Theorem [16, Theorem 1.1] *Let* $\mathbf{A}$ *be a torsion-free abelian group of rank* 3*. Then one of the following cases occurs:*

(1) *the group* $\mathbf{A}$ *is endoprimal;*

(2) *the group* $\mathbf{A}$ *is* p*-divisible for some prime* p *and therefore* $\mathbf{A}$ *is not endoprimal;*

(3) $\operatorname{End}\mathbf{A}$ *is a subring of* $\mathbb{Q}$ *and therefore* $\mathbf{A}$ *is not endoprimal;*

(4) $\operatorname{End}\mathbf{A}$ *has non-trivial centre and therefore* $\mathbf{A}$ *is not endoprimal;*

(5) $\operatorname{End}\mathbf{A}$ *is an abelian group of rank* 3 *and there exist linearly independent elements* $a_1, a_2, a_3 \in A$*, endomorphisms* $\phi, \psi \in \operatorname{End}\mathbf{A}$ *and* $u \in \mathbb{Q}$ *such that* $\phi(a_2) = a_1$*,* $\phi(a_1) = \phi(a_3) = 0$*,* $\psi(a_1) = ua_1$*,* $\psi(a_2) = \psi(a_3) = 0$*; then* $\mathbf{A}$ *is not endoprimal.*

Notice that cases 2–4 present general necessary conditions for a group to be endoprimal. Hence it is 'almost true' that a torsion-free group of rank 3 is endoprimal unless there is an obvious obstacle to its endoprimality. It may be surprising that, even for these groups of small rank, decomposability is not a necessary condition of endoprimality: in [16] an example is constructed for an endoprimal indecomposable torsion-free group of rank 3.

A large part of the proof of Theorem 4.11 goes through for torsion-free groups of arbitrary finite rank, so at least 'reasonable' partial results can be hoped for in that fairly general situation.

Finally, the question of endoprimality for mixed groups whose maximal torsion-free factor group has rank 1 was settled by K. Kaarli and L. Márki [15]. The result reads as follows.

4.12 Theorem [15, Theorem 1.2] *Let* $\mathbf{A}$ *be a group with torsion part* $\mathbf{T}$ *and* P *be the set of those primes* p *for which* $\mathbf{A}/\mathbf{T}$ *is* p*-divisible. Assume* $\mathbf{A}$ *has torsion-free rank* 1. *Then* $\mathbf{A}$ *is endoprimal if and only if* $\mathbf{T}$ *is unbounded and, for every* $p \in P$*, the* p*-component of* $\mathbf{T}$ *is either not reduced or it is not a direct summand of* $\mathbf{A}$.

Notice that we also get endoprimal groups of this kind in which none of the primary parts splits off; in other words, such groups are as 'close to being indecomposable' as a mixed group of torsion-free rank 1 can be.

5 Other classes of algebras

We survey here the known endoprimality results for sets, vector spaces, semilattices, and implications algebras.

We have already mentioned Sangalli's result [22] that all but 2-element sets are endoprimal. The 2-element set is not endoprimal because the transposition of its elements is an endofunction but not a term function.

Using the duality theory approach it is easy to show that all but 1-dimensional finite vector spaces are endoprimal. B. A. Davey and J. G. Pitkethly [7] extended this result to infinite vector spaces. The underlying idea was that all vector spaces are free algebras. Another possible way to get the same result is to rewrite the proof of Theorem 4.5 for vector spaces. Thus we have the following theorem.

5.1 Theorem [7, Theorem 2.4] *A vector space is endoprimal if and only if its dimension is not* 1.

Now we formulate the result for semilattices, which again was obtained by B. A. Davey and J. G. Pitkethly [7] using a mixture of duality and non-duality theory methods.

5.2 Theorem [7, Theorem 5.2] *A non-trivial semilattice is endoprimal if and only if it is not a tree.*

Note that in this theorem we did not assume that the signature of the semilattice includes bound(s) 0, 1 or both. If there is at least one bound then the result may be slightly different. For example, a non-trivial bounded semilattice is endoprimal if and only if it is not a chain.

Endoprimal implication algebras were completely characterised in the recent paper [21] by J. G. Pitkethly. Recall that an implication algebra can be defined as a groupoid with a binary operation $\to$ satisfying the identities:

$$(x \to y) \to x = x, \qquad (x \to y) \to y = (y \to x) \to x, \qquad x \to (y \to z) = y \to (x \to z).$$

It is known that every implication algebra is a semilattice with respect to the term operation $x \vee y = (x \to y) \to y$.

5.3 Theorem [21, Theorem 2.5] *An implication algebra* $\mathbf{A}$ *is endoprimal if and only if for every* $n \geq 1$ *there exist* $a_1, \ldots, a_n \in A$ *such that the set* $\{\bigvee_{j \neq i} a_j \mid i \in \{1, \ldots, n\}\}$ *does not have a lower bound in* $\langle A; \vee \rangle$.

Note that this result cannot be obtained by using duality theory.

References

[1] D. M. Clark and B. A. Davey, *Natural Dualities for the Working Algebraist*, Cambridge University Press, Cambridge, 1998.

[2] B. A. Davey, Duality for equational classes of Brouwerian and Heyting algebras, *Trans. Amer. Math. Soc.* **221** (1976), 119–146.

[3] B. A. Davey, Dualisability in general and endodualisability in particular, in: *Logic and Algebra (Proc. Conf. Pontignano, 1994)*, (A. Ursini, P. Agliano, eds.), Lecture Notes in Pure and Appl. Math. **180**, Marcel Dekker, New York, 1996, 437–455.

[4] B. A. Davey, M. Haviar, and H. A. Priestley, Endoprimal distributive lattices are endodualisable, *Algebra Universalis* **34** (1995), 444–453.

[5] B. A. Davey, M. Haviar, and H. A. Priestley, The syntax and semantics of entailment in duality theory, *J. Symbolic Logic* **60** (1995), 1087–1114.

[6] B. A. Davey, M. Haviar, and H. A. Priestley, Kleene algebras: a case-study of clones and dualities from endomorphisms, *Acta Math. Sci. (Szeged)* **67** (2001), 77–103.

[7] B. A. Davey and J. G. Pitkethly, Endoprimal algebras, *Algebra Universalis* **38** (1997), 266–288.

[8] B. A. Davey and H. Werner, Piggyback-Dualitäten, *Bull. Austral. Math. Soc.* **32** (1985), 1–32.

[9] L. Fuchs, *Infinite Abelian Groups, I–II*, Academic Press, New York, 1970, 1973.

[10] R. Göbel, K. Kaarli, L. Márki, and S. Wallutis, Endoprimal torsion-free separable abelian groups, *J. Algebra Appl.* **3** (2004), 61–73.

[11] R. Göbel and S. Shelah, Uniquely transitive torsion-free abelian groups, in: *Rings, Modules, Algebras, and Abelian Groups (Proc. Conf. Venice, 2002)* (A. Facchini, E. Houston, and L. Salce, eds.), Lecture Notes in Pure and Appl. Math. **236**, Marcel Dekker, New York, 2004, 271–290.

[12] M. Haviar and H. A. Priestley, Endoprimal and endodualisable finite double Stone algebras, *Algebra Universalis* **42** (1999), 107–130.

[13] M. Haviar and H. A. Priestley, A criterion for a finite endoprimal algebra to be endodualisable, *Algebra Universalis* **42** (1999), 183–193.

[14] K. Kaarli and L. Márki, Endoprimal abelian groups, *J. Austral. Math. Soc. (Series A)* **67** (1999), 412–428.

[15] K. Kaarli and L. Márki, Endoprimal abelian groups of torsion-free rank 1, *Rend. Semin. Mat. Univ. Padova*, to appear.

[16] K. Kaarli and K. Metsalu, On endoprimality of torsion-free abelian groups of rank 3, *Acta Math. Hungar.* **104** (2004), 271–289.

[17] K. Kaarli and A. F. Pixley, *Polynomial Completeness in Algebraic Systems*, Chapman & Hall/CRC, Boca Raton, 2001.

[18] K. Kaarli and H. A. Priestley, Endoprimality without duality, *Algebra Universalis* **51** (2004), 361–372.

[19] L. A. Kalužnin and R. Pöschel, *Funktionen- und Relationenalgebren. Ein Kapitel der diskreten Mathematik*, Lehrbücher und Monographien aus dem Gebiete der exakten Wissenschaften, VEB Deutscher Verlag der Wissenschaften, Berlin, 1979.

[20] L. Márki and R. Pöschel, Endoprimal distributive lattices, *Algebra Universalis* **30** (1993), 272–274.

[21] J. G. Pitkethly, Endoprimal implication algebras, *Algebra Universalis* **41** (1999), 201–211.

[22] A. A. L. Sangalli, The degree of invariancy of a bicentrally closed clone, in: *Lattices, Semigroups, and Universal Algebra (Proc. Conf. Lisbon, 1988)*, Plenum Press, New York, 1990, 279–283.

The complexity of constraint satisfaction: an algebraic approach

Andrei KROKHIN

Department of Computer Science
University of Durham
Durham, DH1 3LE
UK

Andrei BULATOV

School of Computer Science
Simon Fraser University
Burnaby BC, V5A 1S6
Canada

Peter JEAVONS

Computing Laboratory
University of Oxford
Oxford OX1 3QD
UK

Notes taken by Alexander SEMIGRODSKIKH

Abstract

Many computational problems arising in artificial intelligence, computer science and elsewhere can be represented as constraint satisfaction and optimization problems. In this survey paper we discuss an algebraic approach that has proved to be very successful in studying the complexity of constraint problems.

1 Constraint satisfaction problems

The constraint satisfaction problem (CSP) is a powerful general framework in which a variety of combinatorial problems can be expressed [20, 59, 61, 79]. The aim in a constraint satisfaction problem is to find an assignment of values to the variables, subject to specified constraints. In artificial intelligence, this framework is widely acknowledged as a convenient and efficient way of modelling and solving a number of real-world problems such as planning [48] and scheduling [75], frequency assignment problems [27], image processing [63], programming language analysis [65] and natural language understanding [2]. In database theory, it has been shown that the key problem of conjunctive-query evaluation can be viewed as

V. B. Kudryavtsev and I. G. Rosenberg (eds.),
Structural Theory of Automata, Semigroups, and Universal Algebra, 181–213.

a constraint satisfaction problem [35, 54]. Furthermore, some central problems in combinatorial optimization can be represented as constraint problems [20, 30, 42, 50]. Finally, CSPs have attracted much attention in complexity theory because various versions of CSPs lie at the heart of many standard complexity classes, and because, despite their great expressiveness, they tend to avoid "intermediate" complexity; that is, they tend to be either tractable or complete for standard complexity classes [7, 8, 9, 11, 13, 20, 30, 51, 46, 73]. On a more practical side, constraint programming is a rapidly developing area with its own international journal and an annual international conference, and with new programming languages being specifically designed (see, e.g., [61]).

The standard toy example of a problem modelled as a constraint satisfaction problem is the "8-queens" problem: place eight queens on a chess board so that no queen can capture any other one [79]. One can think of the horizontals of the board as variables, and the verticals as the possible values, so that assigning a value to a variable means placing a queen on the corresponding square of the board. The fact that no queen must be able to capture any other queen can be represented as a collection of binary constraints C_{ij}, one for each pair of variables i, j, where the constraint C_{ij} allows only those pairs (k, l) such that a queen at position (i, k) cannot capture a queen at position (j, l). It is easy to see that every solution of this constraint satisfaction problem corresponds to a "legal" placing of the 8 queens.

We now give a formal definition of the general CSP.

1.1 Definition An instance of a constraint satisfaction problem is a triple $(V, D, \mathcal{C})$ where

- V is a finite set of variables,
- D is a set of values (sometimes called a domain), and
- $\mathcal{C}$ is a set of constraints $\{C_1, \ldots, C_q\}$, in which each constraint C_i is a pair $\langle s_i, \varrho_i \rangle$ with s_i a list of variables of length m_i, called *the constraint scope*, and ϱ_i an m_i-ary relation over the set D called *the constraint relation.*

The question is whether there exists a *solution* to $(V, D, \mathcal{C})$, that is, a function from V to D such that, for each constraint in $\mathcal{C}$, the image of the constraint scope is a member of the constraint relation.

Now we give some examples of natural problems and their representations as CSPs.

1.2 Example The most obvious algebraic example of a CSP is the problem of solving a system of equations: given a system of linear equations over a finite field F, does it have a solution? Clearly, in this example each individual equation is a constraint, where the variables in the equation form the scope, and the set of all tuples corresponding to solutions of this equation is the constraint relation.

1.3 Example An instance of the standard propositional 3-SATISFIABILITY problem [32, 66] is specified by giving a formula in propositional logic consisting of a conjunction of clauses, each containing three literals (that is, variables or negated variables), and asking whether there are values for the variables which make the formula true.

Suppose that $\Phi = \phi_1 \wedge \cdots \wedge \phi_n$ is such a formula, where the ϕ_i are clauses. The satisfiability question for Φ can be expressed as the instance $(V, \{0, 1\}, \mathcal{C})$ of CSP, where V is the set of

all variables appearing in the formula, and $\mathcal{C}$ is the set of constraints $\{\langle s_1, \varrho_1\rangle, \ldots, \langle s_n, \varrho_n\rangle\}$, where each constraint $\langle s_k, \varrho_k\rangle$ is constructed as follows: s_k is a list of the variables appearing in ϕ_k and ϱ_k consists of all tuples that make ϕ_k true. The solutions of this CSP instance are exactly the assignments which make the formula ϕ true. Hence, any instance of 3-SATISFIABILITY can be expressed as an instance of CSP.

Example 1.3 suggests that any instance of CSP can be represented in a logical form. Indeed, using the standard correspondence between relations and predicates, one can rewrite an instance of CSP as a first-order formula $\varrho_1(s_1) \wedge \cdots \wedge \varrho_q(s_q)$ where the ϱ_i $(1 \leq i \leq q)$ are predicates on D and $\varrho_i(s_i)$ means ϱ_i applied to the tuple s_i of variables. The question then would be whether this formula is satisfiable [74]. In this paper we will sometimes use this alternative logical form for the CSP. This form is commonly used in database theory because it corresponds so closely to conjunctive query evaluation [54], as the next example indicates.

1.4 Example A *relational database* is a finite collection of *tables*. A table consists of a *scheme* and an *instance*, where

A scheme is a finite set of *attributes*, where each attribute has an associated set of possible values, referred to as a *domain*.

An instance is a finite set of *rows*, where each row is a mapping that associates with each attribute of the scheme a value in its domain.

A standard problem in the context of relational databases is the CONJUNCTIVE QUERY EVALUATION problem [35, 54]. In this problem we are asked if a *conjunctive query* to a relational database, that is, a query of the form $\varrho_1 \wedge \cdots \wedge \varrho_n$ where the $\varrho_1, \ldots, \varrho_n$ are atomic formulas, has a solution.

A conjunctive query over a relational database corresponds to an instance of CSP by a simple translation of terms: 'attributes' have to be replaced with 'variables', 'tables' with 'constraints', 'scheme' with 'scope', 'instance' with 'constraint relation', and 'rows' with 'tuples'. Hence a conjunctive query is equivalent to a CSP instance whose variables are the variables of the query. For each atomic formula ϱ_i in the query, there is a constraint C such that the scope of C is the list of variables of ϱ_i and the constraint relation of C is the set of models of ϱ_i.

Another important reformulation of the CSP is the HOMOMORPHISM problem: the question of deciding whether there exists a homomorphism between two relational structures (see [30, 35, 54]). Let $\tau = (R_1, \ldots, R_k)$ be a signature, that is, a list of relation names with a fixed arity assigned to each name. Let $\mathcal{A} = (A; R_1^{\mathcal{A}}, \ldots, R_k^{\mathcal{A}})$ and $\mathcal{B} = (B; R_1^{\mathcal{B}}, \ldots, R_k^{\mathcal{B}})$ be relational structures of signature τ. A mapping $h : A \to B$ is called a *homomorphism* from $\mathcal{A}$ to $\mathcal{B}$ if, for all $1 \leq i \leq k$, $(h(a_1), \ldots, h(a_m)) \in R_i^{\mathcal{B}}$ whenever $(a_1, \ldots, a_m) \in R_i^{\mathcal{A}}$. In this case we write $h : \mathcal{A} \to \mathcal{B}$. To see that the HOMOMORPHISM problem is the same as the CSP, think of the elements in A as variables, the elements in B as values, tuples in the relations of $\mathcal{A}$ as constraint scopes, and the relations of $\mathcal{B}$ as constraint relations. Then, clearly, the solutions to this CSP instance are precisely the homomorphisms from $\mathcal{A}$ to $\mathcal{B}$.

We now give some more examples of well-known combinatorial problems and their representations as a CSP. For the sake of brevity, we use the homomorphism form of the CSP.

1.5 Example For any positive integer k, an instance of the Graph k-Colorability problem consists of a graph G. The question is whether the vertices of G can be coloured with k colours in such a way that adjacent vertices receive different colours.

It follows that every instance of Graph colorability can be expressed as a CSP instance where $\mathcal{A} = G$ and $\mathcal{B}$ is the complete graph on k vertices, K_k.

1.6 Example An instance of the Clique problem consists of an undirected graph G and an integer k. The question is whether G has a clique of size k (that is, a subgraph isomorphic to the complete graph K_k).

It follows that every instance of the Clique problem can be expressed as a CSP instance where $\mathcal{A}$ is K_k and $\mathcal{B}$ is the graph G.

1.7 Example An instance of the Hamiltonian Circuit problem consists of a graph $G = (V; E)$. The question is whether there is a cyclic ordering of V such that every pair of successive nodes in V is adjacent in G.

It follows that every instance of the Hamiltonian Circuit problem can be expressed as a CSP instance with $\mathcal{A} = (V; C_V, \neq_V)$ and $\mathcal{B} = (V; E, \neq_V)$, where $\neq_V$ denotes the disequality relation on V and C_V is the graph of an arbitrary cyclic permutation on V.

1.8 Example An instance of the Graph Isomorphism problem consists of two graphs, $G = (V; E)$ and $G' = (V'; E')$, with $|V| = |V'|$. The question is whether there is a bijection between V and V' such that adjacent vertices in G are mapped to adjacent vertices in G', and non-adjacent vertices are mapped to non-adjacent vertices.

It follows that every instance of the Graph Isomorphism problem can be expressed as a CSP instance with $\mathcal{A} = (V; E, \overline{E})$ and $\mathcal{B} = (V'; E', \overline{E'})$, where $\overline{E}$ denotes the set of all pairs in $\neq_V$ that are not in E.

Many other examples of well-known problems expressed as CSPs can be found further on in this paper, and also in [42].

2 Related constraint problems

As with many other computational problems, it is not only the standard version of the CSP (that is, deciding whether a CSP instance has a solution or not) which is of interest. There are many related problems that have been studied, and in this section we give a brief overview of some of these.

- **Counting Problem**
  How many solutions does a given CSP instance have?

  A standard natural problem associated with many computational decision problems [20].

- **Quantified Problem**
  Given a fully quantified instance of CSP, is it true?

  Problems of this form have provided several fundamental examples of **PSPACE**-complete problems [20, 22, 74]. Any instance of the ordinary CSP can be viewed as an instance of this problem in which all the quantifiers are existential.

- **Minimal Solution**
  Given a CSP instance and some solution to it, is there a solution that is strictly less (point-wise) than the given one?

  This problem is connected with circumscription, a framework used in artificial intelligence to formalize common-sense reasoning [53]. It was also studied as "minimal model checking" in [52].

- **Circumscriptive Inference**
  Given two CSP instances with the same set of variables, is every minimal solution to the first one also a solution to the second one?

  This is a popular problem in nonmonotonic reasoning, an area of artificial intelligence, related to the previous version of the CSP. It was studied in [51, 52].

- **Equivalence**
  Given two CSP instances, do they have the same sets of solutions?

  In database theory, this corresponds to the question of whether or not two queries are equivalent [6].

- **Isomorphism**
  Given two CSP instances, can one permute the variables in them so that they become equivalent in the above sense?

  This is a more general form of the Equivalence problem which is of interest in some contexts. The complexity of the Boolean case of this problem is classified in [7].

- **Inverse Satisfiability**
  Given a set of n-tuples, is it the set of all solutions to a CSP instance of some certain type?

  This problem is related to efficient knowledge representation issues in artificial intelligence [49].

- **Listing Problem**
  Generate all solutions of a given CSP instance.

  A standard natural problem associated with many computational decision problems [20].

- **Max CSP**
  Maximize the number of satisfied constraints in a CSP instance.

  For over-constrained problems, where it is impossible to satisfy all of the constraints, it may be appropriate to try to find a solution satisfying as many constraints as possible [31]. A number of standard optimization problems, e.g., maximum cut, can also be expressed as Max CSP problems [20, 50].

- **Maximum Solution**
  Maximize the sum of values in a solution of a CSP instance.

  Many optimization problems including maximum clique are of this form; in the Boolean case this problem is known as MAX ONES [20, 50].

- **Maximum Hamming Distance**
  Find two solutions to a CSP instance that are distinct in a maximal number of variables.

  The "world difference" in the blocks world problem from knowledge representation can be modelled in this way [21].

- **Lex Max CSP**
  Given a CSP instance where the variables are linearly ordered, find a solution that is lexicographically maximal.

  This form of CSP is used when variables in instances have priorities according to some preference list [73].

- **Unique Solution**
  Does a given instance of CSP have a unique solution?

  This problem is studied in [47]. A related problem concerning partially unique solutions (that is, solutions that are unique on some subsets of variables) was studied in [46].

3 Parameterization of the CSP

The main object of our interest is the computational complexity of constraint problems of various kinds. We refer the reader to [32, 66] for a general background in complexity theory and the definitions of standard complexity classes. In general, the standard decision-problem form of the CSP is **NP**-complete, as one can see from Example 1.3, so it is unlikely to be computationally tractable. However, certain restrictions on the form of the problems can ensure tractability, that is, solvability in polynomial time (see, e.g., [67]).

With any CSP instance one can associate two natural parameters, which represent, informally, the following two features of the instance: which variables constrain which others, and the way in which the values are constrained.

(1) The first feature (that is, which variables constrain which others) can be captured in two ways: one of these is by giving a hypergraph defined on the set of variables used in the instance, where each hyperedge consists of the set of variables appearing together in some constraint scope. The other, finer, way is by specifying the left-hand-side structure, $\mathcal{A}$, in the homomorphism form of the CSP.

(2) The second feature (that is, the way in which the values are constrained), can be captured by specifying the set of constraint relations used in the instance, or alternatively by specifying the right-hand-side structure, $\mathcal{B}$, in the homomorphism form of the CSP.

It follows from these observations that the general CSP can be restricted by fixing either the set of allowed hypergraphs (or left-hand-side structures) or else the set of allowed constraint relations (or right-hand-side structures).

The case when the set of hypergraphs is fixed has been studied in connection with databases [35, 54]. Moreover, in [36], there is a complete classification of the complexity of the CSP in the case when the set of possible left-hand-side structures is fixed, and there are no restrictions on the right-hand-side structures.

In this paper we concentrate on the case when *the set of constraint relations allowed in instances is fixed*, but there is no restriction on the form of the associated hypergraphs (or

left-hand-side structures). Let $R_D^{(n)}$ denote the set of all n-ary relations (or predicates) on a set D, and let $R_D = \bigcup_{n=1}^{\infty} R_D^{(n)}$.

3.1 Definition A *constraint language* over D is a subset Γ of R_D. The *constraint satisfaction problem over* Γ, denoted CSP(Γ), is the subclass of the CSP defined by the following property: any constraint relation in any instance must belong to Γ.

Of course, such a parameterization can also be considered for all of the related constraint problems discussed in Section 2 above.

3.2 Definition A constraint language Γ is called *globally tractable* if CSP(Γ) is tractable, and it is called *tractable* if, for every finite $\Gamma_0 \subseteq \Gamma$, CSP($\Gamma_0$) is tractable. It is called **NP**-*complete* if, for some finite $\Gamma_0 \subseteq \Gamma$, CSP($\Gamma_0$) is **NP**-complete.

Of course, every finite tractable constraint language is also globally tractable, but for infinite constraint languages this implication is not immediate (see [14, 17]), so it is technically necessary to distinguish the notions of tractability and global tractability. In fact, all known tractable constraint languages are globally tractable, and it seems plausible that the two notions coincide, though at present this is an open problem. In this paper, we will consider only the question of determining which constraint languages are tractable, and we will not make any further use of the notion of global tractability.

When the set $\Gamma \subset R_D$ is finite, let $\mathcal{B}_\Gamma$ denote the relational structure over the universe D whose relations are precisely the relations of Γ (listed in some order). Then the problem CSP(Γ) corresponds exactly with the problem Hom($\mathcal{B}_\Gamma$), defined as follows: given a structure $\mathcal{A}$ similar to $\mathcal{B}_\Gamma$ (i.e., of the same signature), is it true that $\mathcal{A} \to \mathcal{B}_\Gamma$? Note that the order in which the relations from Γ are listed in $\mathcal{B}_\Gamma$ does not affect the complexity of this problem.

We now give some examples of well-known problems expressible as CSP(Γ) for suitable sets Γ.

3.3 Example An instance of Linear Equations consists of a system of linear equations over a field.

Following Example 1.2, it is easy to see that this problem can be expressed as CSP(Γ) where Γ consists of all relations expressible by a linear equation. This problem is clearly tractable because it can be solved by a straightforward polynomial-time algorithm, such as Gaussian elimination.

Moreover, systems of equations can be considered not only over fields, but also over other algebraic structures. For example, systems of polynomial equations over a (fixed) finite group (that is, equations of the form $a_1 x_1 a_2 \cdots x_n a_{n+1} = b_1 y_1 b_2 \cdots y_m b_{m+1}$ where the a_i's and the b_i's are constants and the x_i's and y_i's are variables) are studied in [33] where it is proved that solving such systems is tractable if the underlying group is Abelian, and is **NP**-complete otherwise. This result is generalised in [64] to solving systems of equations over finite monoids: this problem is tractable if the underlying monoid is a union of groups and commutative; otherwise it is **NP**-complete. A more general setting, when systems of polynomial equations are considered over an arbitrary finite (universal) algebra, is studied in [57], which gives a generalization of the results on groups and monoids mentioned above.

3.4 Example The NOT-ALL-EQUAL SATISFIABILITY problem [32, 74] is a restricted version of the standard 3-SATISFIABILITY problem (Example 1.3) which remains **NP**-complete. In this problem the clauses are ternary, and each clause is satisfied by any assignment in which the variables of the clause do not all receive the same truth value.

This problem corresponds to the problem CSP($\{N\}$) where N is the following ternary relation on $\{0,1\}$:

$$N = \{0,1\}^3 \setminus \{(0,0,0),(1,1,1)\}.$$

3.5 Example Let $H = (V,E)$ be a finite graph. An instance of the GRAPH H-COLORING problem consists of a finite graph G. The question is whether G can be homomorphically mapped to H.

This problem precisely corresponds to the problem CSP($\{E\}$). If we consider only undirected graphs H, then the complexity of GRAPH H-COLORING has been completely characterised [39]: it is tractable if H is bipartite or contains a loop; otherwise it is **NP**-complete. However, if we allow H and G to be directed graphs, then the complexity of GRAPH H-COLORING has not yet been fully characterised. Moreover, it was shown in [30] that every problem CSP(Γ) with finite Γ is polynomial-time equivalent to GRAPH H-COLORING for some suitable directed graph H.

Following a seminal work by Schaefer in 1978 [74], many researchers have studied the following problem:

3.6 Problem Determine the complexity of a given constraint problem for *all possible* values of the parameter Γ.

Most progress has been made in the Boolean case (that is, when the set of values D is $\{0,1\}$), such problems are sometimes called "generalized satisfiability problems" [32]. Schaefer obtained a complete classification for the standard decision-problem form of the CSP over $\{0,1\}$ [74], which is described in Section 4.3, below. Over the last decade, classifications for many related Boolean constraint problems, including all of the problems mentioned in Section 2, have been completed (see references in Section 2). Some of these classifications are also described in Section 4.3.

Classifying the complexity in the non-Boolean case has proved to be a very difficult task. Three main approaches to this problem have been considered; two of them are based on the homomorphism form of the CSP.

(1) The homomorphism problem for *graphs* has been extensively studied (see, e.g., [38]), and thus one can try to develop some methods of graph theory to apply in the more general context of constraint satisfaction.

(2) The problem Hom($\mathcal{B}$) can be seen as the membership problem for the class of all relational structures $\mathcal{A}$ such that $\mathcal{A} \to \mathcal{B}$, and hence methods of finite model theory can applied to study the definability of this class in various logics (from which one can then derive information about the complexity of the problem [29]).

Elements of these two approaches are present in [23, 25, 30, 54].

In the remainder of this paper, we will discuss the third, *algebraic*, approach to the complexity classification problem. This approach has proved to be the most fruitful so far; it

has made it possible to obtain very strong complexity classification results for a wide variety of cases.

4 The finite-valued CSP

In this section we consider the case when the set of possible values for the variables in a constraint satisfaction problem is finite.

4.1 Expressive power of constraint languages

In any CSP instance some of the required relationships between variables are given explicitly in the constraints, whilst others generally arise implicitly from interactions among different constraints. For any instance in CSP(Γ), the explicit constraint relations must be elements of Γ, but there may be implicit restrictions on some subsets of the variables for which the corresponding relations are not elements of Γ, as the next example indicates.

4.1 Example Let Γ be the set containing a single binary relation, χ, over the set $\{0, 1, 2\}$, where χ is defined as follows:

$$\chi = \{(0,0), (0,1), (1,0), (1,2), (2,1), (2,2)\}.$$

One element of CSP(Γ) is the instance

$$\mathcal{P} = (\{v_1, v_2, v_3, v_4\}, \{0, 1, 2\}, \{C_1, C_2, C_3, C_4, C_5\}),$$

where $C_1 = ((v_1, v_2), \chi)$, $C_2 = ((v_1, v_3), \chi)$, $C_3 = ((v_2, v_3), \chi)$, $C_4 = ((v_2, v_4), \chi)$, $C_5 = ((v_3, v_4), \chi)$.

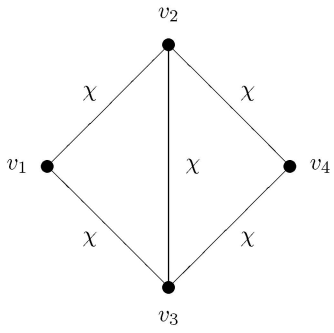


Figure 1: The CSP instance $\mathcal{P}$ defined in Example 4.1.

Note that there is no explicit constraint on the pair (v_1, v_4). However, by considering all solutions to $\mathcal{P}$, it can be shown that the possible pairs of values which can be taken by this pair of variables are precisely the elements of the relation $\chi' = \chi \cup \{(1,1)\}$.

We now define exactly what it means to say that a constraint relation can be expressed in a constraint language.

4.2 Definition A relation ϱ can be *expressed* in a constraint language Γ over D if there exists a problem instance (V, D, C) in CSP(Γ), and a list, s, of variables, such that the solutions to (V, D, C) restricted to s give precisely the tuples of ϱ.

For any constraint language Γ, the set of all relations which can be expressed in Γ will be called the *expressive power* of Γ.

The expressive power of a constraint language Γ can be characterised in a number of different ways [45]. For example, it is equal to the set of all relations that may be obtained from the relations in Γ using the *relational join* and *project* operations from relational database theory [37]. Alternatively, it can be shown to be equal to the set of relations definable by *primitive positive formulas* involving the relations of Γ and equality, which is defined as follows.

4.3 Definition For any set of relations Γ over D, the set $\langle\Gamma\rangle$ consists of all relations that can be expressed using

(1) relations from Γ, together with the binary equality relation on D (denoted $=_D$),

(2) conjunction, and

(3) existential quantification.

4.4 Example Example 4.1 demonstrates that the relation χ' belongs to the expressive power of the constraint language $\Gamma = \{\chi\}$. It is easy to deduce from the construction given in Example 4.1 that

$$\chi'(x, y) \equiv \exists u\, \exists v\, (\chi(x, u) \wedge \chi(x, v) \wedge \chi(u, v) \wedge \chi(u, y) \wedge \chi(v, y)).$$

Hence, $\chi' \in \langle\{\chi\}\rangle$.

4.2 Polymorphisms and complexity

In this section we shall explore how the notion of expressive power may be used to simplify the analysis of the complexity of the constraint satisfaction problem.

We first note that any relation that can be expressed in a language Γ can be added to Γ without changing the complexity of CSP(Γ).

4.5 Proposition *For any constraint language* Γ *and any relation* ϱ *belonging to the expressive power of* Γ, CSP($\Gamma \cup \{\varrho\}$) *is reducible in polynomial time to* CSP(Γ).

This result can be established simply by noting that, given an arbitrary problem instance in CSP($\Gamma \cup \{\varrho\}$), we can obtain an equivalent instance in CSP(Γ) by replacing each constraint C that has constraint relation ϱ with a collection of constraints that have constraint relations chosen from Γ and that together express the constraint C.

By iterating this procedure we can obtain the following corollary.

4.6 Corollary *For any constraint language* Γ, *and any finite constraint language* Γ_0, *if* Γ_0 *is contained in the expressive power of* Γ, *then* CSP(Γ_0) *is reducible to* CSP(Γ) *in polynomial time.*

Corollary 4.6 implies that for any finite constraint language Γ, the complexity of CSP(Γ) is determined, up to polynomial-time reduction, by the expressive power of Γ, and hence by $\langle\Gamma\rangle$. This raises an obvious question: how can we obtain sufficient information about the set $\langle\Gamma\rangle$ to determine the complexity of CSP(Γ)?

A very successful approach to this question has been developed in [16, 42, 44], using techniques from universal algebra [62, 70]. To describe this approach, we need to consider finitary *operations* on D. We will use $O_D^{(n)}$ to denote the set of all n-ary operations on the set D (that is, the set of mappings $f\colon D^n \to D$), and O_D to denote the set $\bigcup_{n=1}^{\infty} O_D^{(n)}$.

An operation $f \in O_D^{(n)}$ will be called *essentially unary* if there exists some i in the range $1 \leq i \leq n$, and some operation $g \in O_D^{(1)}$ such that the following identity is satisfied

$$f(x_1, x_2, \ldots, x_n) = g(x_i).$$

An essentially unary operation for which g is the identity operation is called a *projection*. Any operation (of whatever arity) which is *not* essentially unary will be called *essentially non-unary*.

Any operation on D can be extended in a standard way to an operation on tuples over D, as follows. For any operation $f \in O_D^{(n)}$, and any collection of tuples $\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_n \in D^m$, where $\vec{a}_i = (a_{i1}, \ldots, a_{im})$ $(i = 1 \ldots n)$, define $f(\vec{a}_1, \ldots, \vec{a}_n)$ by setting

$$f(\vec{a}_1, \ldots, \vec{a}_n) = (f(a_{11}, \ldots, a_{n1}), \ldots, f(a_{1m}, \ldots, a_{nm})).$$

4.7 Definition For any relation $\varrho \in R_D^{(m)}$, and any operation $f \in O_D^{(n)}$, if $f(\vec{a}_1, \ldots, \vec{a}_n) \in \varrho$ for all choices of $\vec{a}_1, \ldots, \vec{a}_n \in \varrho$, then ϱ is said to be *invariant* under f, and f is called a *polymorphism* of ϱ.

The set of all relations that are invariant under each operation from some set $C \subseteq O_D$ will be denoted $\mathsf{Inv}(C)$. The set of all operations that are polymorphisms of every relation from some set $\Gamma \subseteq R_D$ will be denoted $\mathsf{Pol}(\Gamma)$. The operators Inv and Pol form a Galois correspondence between R_D and O_D (see [70, Proposition 1.1.14]). A basic introduction to this correspondence can be found in [68], and a comprehensive study in [70].

Sets of operations of the form $\mathsf{Pol}(\Gamma)$ are known as *clones* and sets of relations of the form $\mathsf{Inv}(C)$ are known as *relational clones* [70]. Moreover, the following useful characterisation of sets of the form $\mathsf{Inv}(\mathsf{Pol}(\Gamma))$ can be found in [70].

4.8 Theorem *For every set* $\Gamma \subseteq R_D$, $\mathsf{Inv}(\mathsf{Pol}(\Gamma)) = \langle\Gamma\rangle$.

This result was combined with Corollary 4.6 to obtain the following result in [42].

4.9 Theorem *For any constraint languages* $\Gamma, \Gamma_0 \subseteq R_D$, *with* Γ_0 *finite, if* $\mathsf{Pol}(\Gamma) \subseteq \mathsf{Pol}(\Gamma_0)$, *then* CSP($\Gamma_0$) *is reducible to* CSP($\Gamma$) *in polynomial time.*

This result implies that, for any finite constraint language Γ over a finite set, the complexity of CSP(Γ) is determined, up to polynomial-time reduction, by the polymorphisms of Γ.

We now apply this result to obtain a sufficient condition for **NP**-completeness of CSP(Γ). A constraint language Γ is said to be *strongly rigid* if $\mathsf{Pol}(\Gamma)$ consists of projections only.

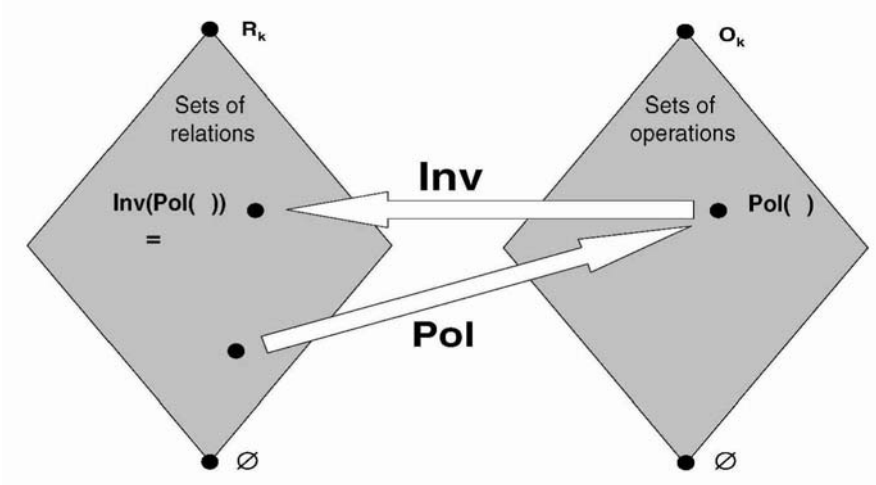


Figure 2: The operators Inv and Pol.

4.10 Proposition *If* Γ *is strongly rigid then* CSP(Γ) *is* **NP***-complete.*

This proposition follows from Theorem 4.9 by setting $\Gamma_0 = \{N\}$ (see Example 3.4), assuming $\{0,1\} \subseteq D$, and using the fact that every relation on D is invariant under any projection.

Proposition 4.10 was used in [58] to show that most non-trivial problems of the form CSP(Γ), with finite Γ, are **NP**-complete. More precisely, let $R(n,k)$ denote a random k-ary relation on the set $\{1,\ldots,n\}$, for which the probability that $(a_1,\ldots,a_k) \in R(n,k)$ is equal to $1/2$ independently for each k-tuple $(a_1,\ldots,a_k)$ where not all a_i's are equal; also, set $(a,\ldots,a) \notin R(n,k)$ for all a (this is necessary to ensure that CSP($R(n,k)$) is non-trivial). It is shown in [58] that the probability that $\{R(n,k)\}$ is strongly rigid tends to 1 as either n or k tends to infinity.

4.3 Complexity of Boolean problems

In this section we describe some of the results that have been obtained concerning the complexity of Boolean constraint problems, that is, problems over a two-valued domain.

The first result of this kind was a complete classification of the complexity of the ordinary Boolean constraint satisfaction problem obtained by Schaefer in 1978 [74]. Recall that a computational problem is called *tractable* if there is a polynomial-time algorithm deciding every instance of the problem. The class of all tractable problems is denoted **PTIME**.

4.11 Theorem *For any constraint language* $\Gamma \subseteq R_{\{0,1\}}$, CSP($\Gamma$) *is tractable when (at least) one of the following conditions holds:*

(1) *Every* ϱ *in* Γ *contains the tuple* $(0,0,\ldots,0)$.

(2) *Every* ϱ *in* Γ *contains the tuple* $(1,1,\ldots,1)$.

(3) *Every ϱ in Γ is definable by a CNF formula in which each conjunct has at most one negated variable.*

(4) *Every ϱ in Γ is definable by a CNF formula in which each conjunct has at most one unnegated variable.*

(5) *Every ϱ in Γ is definable by a CNF formula in which each conjunct has at most two literals.*

(6) *Every ϱ in Γ is definable by a system of linear equations over the two-element field.*

In all other cases CSP(Γ) *is* **NP***-complete.*

This result establishes a dichotomy for versions of this problem parameterized by the choice of constraint language: they are all either tractable or **NP**-complete. Dichotomy theorems of this kind are of particular interest because, on the one hand, they determine the precise complexity of particular constraint problems, and, on the other hand, they demonstrate that no problems of intermediate complexity can occur in this context. Note that the existence of constraint problems of intermediate complexity cannot be ruled out *a priori* due to the result [56] that if **PTIME** $\neq$ **NP** then the class **NP** contains (infinitely many pairwise inequivalent) problems which are neither tractable nor **NP**-complete.

Using the algebraic approach described in the previous sections, together with the knowledge of possible clones on a two-element set obtained in [71], Schaefer's result can be reformulated in the following much more concise form.

4.12 Theorem *For any set of relations $\Gamma \subseteq R_{\{0,1\}}$,* CSP($\Gamma$) *is tractable when* $\mathsf{Pol}(\Gamma)$ *contains any essentially non-unary operation or a constant operation. Otherwise it is* **NP***-complete.*

4.13 Example Recall the relation N over $\{0,1\}$ defined in Example 3.4. Using general results from [71], it can be shown that $\mathsf{Pol}(\{N\})$ contains essentially unary operations only, and hence, by Theorem 4.12, CSP($\{N\}$) is **NP**-complete.

Schaefer's result has inspired a series of analogous investigations for many related constraint problems, including those listed in Section 2. We will now list some complexity classification results that have recently been obtained for these problems in the Boolean case. Surprisingly, for a wide variety of such related problems it turns out that the polymorphisms of the constraint language are highly relevant to the study of the computational complexity.

4.14 Theorem *Let $\Gamma \subseteq R_{\{0,1\}}$ be a Boolean constraint language. The following facts are known to hold for constraint problems parameterized by Γ:*

- *The* ***Counting Problem*** *is tractable if* $\mathsf{Pol}(\Gamma)$ *contains the unique affine operation on $\{0,1\}$, $x - y + z$. Otherwise it is #***P***-complete* [20].
- *The* ***Quantified Problem*** *is tractable if* $\mathsf{Pol}(\Gamma)$ *contains an essentially non-unary operation. Otherwise it is* **PSPACE***-complete* [20, 22].
- *The* ***Equivalence*** *problem is tractable if* $\mathsf{Pol}(\Gamma)$ *contains an essentially non-unary operation or a constant operation. Otherwise it is* **coNP***-complete* [6].

- *The **Inverse Satisfiability** problem is tractable if* $\mathsf{Pol}(\Gamma)$ *contains an essentially non-unary operation. Otherwise it is* **coNP***-complete* [49].
- *The **Maximum Hamming Distance** problem is tractable if* $\mathsf{Pol}(\Gamma)$ *contains either a constant operation, or the affine operation and the negation operation on* $\{0,1\}$ [21].

A full description of these results requires the careful definition of the relevant complexity classes and reductions, which is beyond the scope of this paper, so we refer the reader to the cited papers for details.

4.15 Example Recall the relation N over $\{0,1\}$ defined in Example 3.4. Using general results from [71], it can be shown that $\mathsf{Pol}(\{N\})$ contains essentially unary operations only. Hence, by Theorem 4.14, we can immediately conclude that:

- Counting the number of solutions to an instance of CSP($\{N\}$) is #**P**-complete;
- Deciding whether a quantified Boolean formula, whose quantifier-free part involves only conjunctions of the predicate N, is true is **PSPACE**-complete.
- Deciding whether two instances of CSP($\{N\}$) have the same solutions is **coNP**-complete;
- Deciding whether a given set of n-tuples is the set of solutions to some instance of CSP($\{N\}$) is **coNP**-complete.

4.4 From the CSP to algebras and varieties

Most of the results presented in this section were first obtained in [14, 17, 16].

With any constraint language $\Gamma \subseteq R_D$ one can associate an algebra $\mathbb{A}_\Gamma = (D; \mathsf{Pol}(\Gamma))$. In this section we show that the complexity of the problem CSP(Γ) is completely determined by certain properties of $\mathbb{A}_\Gamma$. (We refer the reader to [62] for a general background in universal algebra.)

Recall that algebras are said to be *term equivalent* if they have the same set of term operations. Since, the term operations of $\mathbb{A}_\Gamma$ are precisely the operations in $\mathsf{Pol}(\Gamma)$, Theorem 4.9 implies that term equivalent algebras give rise to problem classes of the same complexity.

4.16 Proposition *Let $\Gamma_1, \Gamma_2 \subseteq R_D$, where D is finite. If $\mathbb{A}_{\Gamma_1}$ and $\mathbb{A}_{\Gamma_2}$ are term equivalent then Γ_1 and Γ_2 are tractable or* **NP***-complete simultaneously.*

This allows us to introduce the notion of a tractable algebra.

4.17 Definition An algebra $\mathbb{A} = (D; F)$ is said to be *tractable* if the constraint language $\mathsf{Inv}(F)$ is tractable. It is said to be **NP***-complete* if $\mathsf{Inv}(F)$ is **NP**-complete.

Thus, the complexity classification problem for constraint languages reduces to the complexity classification problem for finite algebras. Furthermore, the next results show that it is possible to significantly restrict the class of algebras which need to be classified.

Let $\mathbb{A} = (D; F)$ be an algebra, and $U \subseteq D$. Let $\mathbb{A}|_U$ denote the algebra $\mathbb{A}|_U = (U, F')$, where F' consists of all operations of the form $f|_U$ (the restriction of f to U), for each term operation f of $\mathbb{A}$ such that $f \in \mathsf{Pol}(U)$.

4.18 Proposition *Let $\mathbb{A}$ be a finite algebra, f a unary term operation such that $f(f(x)) = f(x)$ and $U = f(D)$. Then $\mathbb{A}$ is tractable if and only if $\mathbb{A}|_U$ is tractable.*

Hence, by choosing a unary term operation with a minimal range, we may restrict ourselves to considering only *surjective* algebras, that is, algebras all of whose term operations are surjective.

Recall that an operation f is called *idempotent* if it satisfies the identity $f(x, \dots, x) = x$, and the *full idempotent reduct* of an algebra $\mathbb{A} = (D; F)$ is the algebra $\mathsf{Id}(\mathbb{A}) = (D, F')$ where F' consists of all idempotent term operations of $\mathbb{A}$.

4.19 Proposition *A surjective finite algebra $\mathbb{A}$ is tractable if and only if its full idempotent reduct is tractable.*

It follows that to classify the complexity of arbitrary finite algebras it is sufficient to consider only idempotent algebras, that is, algebras whose operations are all idempotent.

Next, we show that the standard algebraic constructions preserve the tractability of an algebra.

4.20 Theorem *Let $\mathbb{A}$ be a finite algebra. If $\mathbb{A}$ is tractable, then all of its subalgebras, homomorphic images and finite direct powers are also tractable. Conversely, if $\mathbb{A}$ has an* **NP***-complete subalgebra, homomorphic image, or finite direct power, then it is* **NP***-complete itself.*

For an algebra $\mathbb{A}$, we denote the *pseudo-variety* and the *variety* generated by $\mathbb{A}$ by $\mathsf{pvar}(\mathbb{A})$ and $\mathsf{var}(\mathbb{A})$, respectively.

4.21 Corollary *A finite algebra $\mathbb{A}$ is tractable if and only if every algebra from $\mathsf{pvar}(\mathbb{A})$ is tractable.*

As is well known, if $\mathfrak{A}$ is a finite class of finite algebras, then the pseudo-variety generated by $\mathfrak{A}$ equals the class of finite algebras from the variety generated by $\mathfrak{A}$.

4.22 Corollary *A finite algebra $\mathbb{A}$ is tractable if and only if every finite algebra from $\mathsf{var}(\mathbb{A})$ is tractable.*

4.23 Corollary *If $\mathbb{A}$ is a finite algebra, and $\mathsf{var}(\mathbb{A})$ contains a finite* **NP***-complete algebra, then $\mathbb{A}$ is* **NP***-complete.*

Thus, the tractability of an algebra is a property which can be determined by identities.

We call an algebra a *set* if it contains more than one element and all of its operations are projections. By combining Proposition 4.10 with Corollary 4.23, we get the following result.

4.24 Corollary *If the pseudovariety generated by a finite idempotent algebra $\mathbb{A}$ contains a set then $\mathbb{A}$ is* **NP***-complete.*

A homomorphic image of a subalgebra of an algebra $\mathbb{A}$ is called a *factor* of $\mathbb{A}$.

4.25 Proposition *If $\mathbb{A}$ is an idempotent algebra and $\mathsf{pvar}(\mathbb{A})$ contains a set then some factor of $\mathbb{A}$ is a set.*

Remarkably, the presence of a set as a factor is the only known reason for an idempotent algebra to be **NP**-complete. This prompts us to suggest the following conjecture.

4.26 Conjecture A finite idempotent algebra $\mathbb{A}$ is tractable if and only if

$$\text{none of the factors of } \mathbb{A} \text{ is a set;} \qquad \text{(No-Set)}$$

otherwise it is **NP**-complete.

It was shown in [17] that if one strengthens Conjecture 4.26 by removing the condition of idempotency, or replacing "factor" by either "subalgebra" or "homomorphic image", then the resulting conjecture is false.

It was proved in [78] that a variety $\mathcal{V}$ generated by a finite idempotent algebra contains no set if and only if there is an n-ary term f (called a *Taylor term*) in $\mathcal{V}$ such that $\mathcal{V}$ satisfies n identities of the form

$$f(x_{i1}, \ldots, x_{in}) = f(y_{i1}, \ldots, y_{in}), \quad i = 1, \ldots, n,$$

where $x_{ij}, y_{ij} \in \{x, y\}$ and $x_{ii} \neq y_{ii}$ for all i, j. Therefore, Corollary 4.24 can be restated as follows.

4.27 Corollary *If a finite idempotent algebra has no Taylor term, then it is* **NP***-complete.*

This corollary was used in [57] to study systems of polynomial equations over finite algebras, where it was proved, in particular, that solving systems of equations over a non-trivial algebra from a congruence-distributive variety is **NP**-complete, and, furthermore, solving systems of equations over a Mal'tsev algebra is tractable if this algebra is polynomially equivalent to a module, otherwise it is **NP**-complete.

Using the result from [78] mentioned above, Conjecture 4.26 can be restated in terms of identities.

4.28 Conjecture A finite idempotent algebra $\mathbb{A}$ is tractable if it has a Taylor term; otherwise it is **NP**-complete.

Finally, the condition (No-Set) from Conjecture 4.26 can be expressed in terms of tame congruence theory [40]: a finite idempotent algebra satisfies this condition if and only if the variety it generates "omits type **1**" [14].

4.5 Tractable algebras, classification results and tractability tests

4.5.1 Tractable algebras

During the last decade several particular identities (particular forms of the Taylor term) have been identified that guarantee the tractability of algebras satisfying one of these identities (that is, having a Taylor term of one of these special forms) [12, 10, 26, 42, 43, 44].

Recall that a binary operation $\cdot$ is called a 2-*semilattice* operation if it satisfies the identities $x \cdot x = x$, $x \cdot y = y \cdot x$ and $(x \cdot x) \cdot y = x \cdot (x \cdot y)$. Note that a semilattice operation is a particular case of a 2-semilattice operation. A ternary operation f satisfying the identities

$f(x,y,y) = f(y,y,x) = x$ is called a *Mal'tsev* operation, and an n-ary operation g is called a *near-unanimity* operation if it satisfies the identities

$$f(y,x,\dots,x) = f(x,y,x,\dots,x) = \cdots = f(x,\dots,x,y) = x.$$

An n-ary operation is called *totally symmetric* if, for all $x_1,\dots,x_n$ and $y_1,\dots,y_n$ such that $\{x_1,\dots,x_n\} = \{y_1,\dots,y_n\}$, it satisfies the identities

$$f(x_1,\dots,x_n) = f(y_1,\dots,y_n).$$

(Note that, in [26], a family $(f_n)_{n\geq 2}$ of totally symmetric operations, where f_n is n-ary, was called a *set function*).

4.29 Theorem [12, 10, 43, 26] *A finite algebra is tractable if it has (at least) one the following:*

- *a 2-semilattice term operation;*
- *a Mal'tsev term operation;*
- *a near-unanimity term operation;*
- *n-ary totally symmetric term operations for all $n \geq 2$.*

Another class of algebras which has been shown to be tractable [24] is the class of *para-primal* algebras, which are defined as follows. Let ϱ an n-ary relation on D, and $I = \{i_1,\dots,i_k\} \subseteq \{1,\dots,n\}$ with $i_1 < \cdots < i_k$. By the *projection* of ϱ onto I we mean the relation $\varrho|_I = \{(a_{i_1},\dots,a_{i_k}) \mid (a_1,\dots,a_n) \in \varrho\}$. The set I is said to be *ϱ-reduced* if it is minimal with the property that the natural mapping $\varrho \mapsto \varrho|_I$ is one-to-one. A finite algebra $\mathbb{A}$ is called *para-primal* if, for every $n \in \mathbb{N}$, every subuniverse ϱ of $\mathbb{A}^n$, and every ϱ-reduced set I, we have $\varrho|_I = \prod_{i\in I} \varrho|_{\{i\}}$. However, it is known that every para-primal algebra has a Mal'tsev term operation (see [76, Theorem 4.7]), and, hence, tractability of para-primal algebras follows from Theorem 4.29.

4.5.2 Classification results

Algebras of several special types have been completely classified with respect to the complexity of the corresponding constraint satisfaction problems.

Strictly simple algebras A finite algebra is said to be *strictly simple* if it is simple and has no subalgebras with more than one element. Strictly simple algebras are completely described in [77].

4.30 Proposition [17, 16] *A finite idempotent strictly simple algebra is tractable if it is not a set; otherwise it is* **NP***-complete.*

Homogeneous algebras An algebra is called *homogeneous* if every permutation on its base set is an automorphism of the algebra. Finite homogeneous algebras are completely described in [60].

4.31 Proposition [24] *A finite homogeneous algebra is tractable if it satisfies the condition (*No-Set*); otherwise it is* **NP***-complete.*

Finite semigroups A semigroup is called a *left-* [*right-*]*zero semigroup* if $x \cdot y = x$ [$x \cdot y = y$] for all x, y. It is called a *block-group* if none of its subsemigroups is a left- or right-zero semigroup. As is easily seen, block-groups are exactly those semigroups that have no factor which is a set.

4.32 Proposition [15] *A finite semigroup is tractable if it is a block-group; otherwise it is* **NP***-complete.*

Small algebras Conjecture 4.26 has been proved for 2- and 3-element algebras.

4.33 Theorem [74, 9]

(1) *An idempotent two-element algebra is tractable if it is not a set; otherwise it is* **NP***-complete.*

(2) *An idempotent three-element algebra is tractable if it satisfies the condition* (No-Set)*; otherwise it is* **NP***-complete.*

Conservative algebras An algebra is said to be *conservative* if every subset of its universe is a subalgebra, or, equivalently, if $f(x_1, \ldots, x_n) \in \{x_1, \ldots, x_n\}$ for every term operation f and all $x_1, \ldots, x_n$.

4.34 Theorem [11] *A conservative algebra $\mathbb{A}$ is tractable if every 2-element subalgebra $\mathbb{B}$ has a term operation of one of the following types: a semilattice operation, a ternary near-unanimity operation (that is, a majority operation), or a Mal'tsev operation; otherwise $\mathbb{A}$ is* **NP***-complete.*

It is not hard to check that the conditions stated in Theorem 4.34 are equivalent to (No-Set).

4.5.3 Testing tractability

We now consider the problem of deciding whether a given constraint language or idempotent algebra is tractable. Following [20], we call such a problem a *meta-problem*. A priori, there is no upper complexity bound for this problem; it may even be undecidable. However, if Conjecture 4.26 is true, then, given the basic operations of an idempotent algebra, one can straightforwardly check whether any factor of the algebra is a set. If we are given a finite constraint language Γ on a finite set A, then the presence of a factor which is a set can be detected by examining all polymorphisms of Γ of arity at most $|A|$. Thus, the meta-problem is decidable, assuming Conjecture 4.26 holds. In this section we study its complexity.

For a constraint language Γ, let $f \in \mathsf{Pol}(\Gamma)$ be a unary operation with minimal range U, and let $f(\Gamma) = \{f(\varrho) \mid \varrho \in \Gamma\}$ where $f(\varrho) = \{f(\vec{a}) \mid \vec{a} \in \varrho\}$. Denote by $\overline{\Gamma}$ the constraint language $f(\Gamma) \cup \{\{a\} \mid a \in U\}$. It follows from Propositions 4.18 and 4.19 that Γ is tractable if and only if $\overline{\Gamma}$ is tractable. Moreover, the algebra $\mathbb{A}_{\overline{\Gamma}} = (U, \mathsf{Pol}(\overline{\Gamma}))$ is idempotent.

We consider three combinatorial decision problems related to the condition (No-Set).

CSP-Tractability-of-algebra

Instance: A finite set A and operation tables of idempotent operations $f_1, \ldots, f_n$ on A.
Question: Does the algebra $\mathbb{A} = (A; \{f_1, \ldots, f_n\})$ satisfy (No-set)?

CSP-Tractability

Instance: A finite set A and a finite constraint language Γ on A.
Question: Does the algebra $\mathbb{A} = (U, \mathsf{Pol}(\overline{\Gamma}))$ satisfy (No-set)?

CSP-Tractability(k)

Instance: A finite set A, $|A| \leq k$, and a finite constraint language Γ on A.
Question: Does the algebra $\mathbb{A} = (U, \mathsf{Pol}(\overline{\Gamma}))$ satisfy (No-set)?

4.35 Theorem [14]

(1) *The problem* CSP-Tractability-of-algebra *is tractable.*

(2) *The problem* CSP-Tractability(k) *is tractable.*

(3) *The problem* CSP-Tractability *is* **NP***-complete.*

The algorithm solving CSP-Tractability-of-algebra is, in fact, an adapted version of the algorithm presented in [3] for finding the type set of a finite algebra. Since this algorithm uses operations of arity bounded by the size of the algebra, it can be further transformed to an algorithm for solving CSP-Tractability(k). Finally, it is possible to show that CSP-Tractability is in **NP** and to reduce the **NP**-complete problem Not-All-Equal-Satisfiability (see Example 3.4) to CSP-Tractability in polynomial time.

4.6 The counting CSP

In this section we discuss Problem 3.6 for the counting constraint satisfaction problem (#CSP), which is the problem of counting solutions to an instance of CSP. Using the logical and the homomorphism forms of the CSP (see Section 1) this problem can also be formulated as the problem of counting satisfying assignments to a conjunctive formula, that is, a formula of the form $\varrho_1 \wedge \cdots \wedge \varrho_n$, where each ϱ_i is an atomic formula, in a given interpretation, or alternatively as the problem of finding the number of homomorphisms between two finite relational structures. For any constraint language Γ, giving rise to the decision constraint satisfaction problem CSP(Γ), we also define the corresponding class #CSP(Γ) of counting problems.

The results of this section were first obtained in [13].

4.36 Example An instance of the #3-SAT problem [19, 20, 80, 81] is specified by giving an instance of the 3-satisfiability problem (see Example 1.3) and asking how many assignments satisfy it. Therefore, #3-SAT is equivalent to #CSP(Γ) where Γ is the set of ternary Boolean relations which are expressible by clauses.

4.37 Example In the problem Antichain [72], we are given a finite poset $(P; \leq)$, and we aim to compute the number of antichains in P. This problem can be expressed in the #CSP-form as follows. Let $\varrho_{\leq}$ be the predicate of the natural order on $\{0, 1\}$. We assign a variable x_a to each element $a \in P$. Then the #CSP($\{\varrho_{\leq}\}$) instance $\Phi = \bigwedge_{a \leq b} \varrho_{\leq}(x_a, x_b)$ can be shown to be equivalent to the original Antichain instance.

To show this, notice that every model φ to Φ satisfies the following condition: if $\varphi(x_a) = 1$ and $a \leq b$ then $\varphi(x_b) = 1$. This means that the set $F_\varphi = \{a \in P \mid \varphi(x_a) = 1\}$ is a filter of P. Hence the models of Φ correspond one-to-one to the filters of P, and consequently, to the antichains of P.

On the other hand, any #CSP($\{\varrho_{\leq}\}$) instance is reducible to an ANTICHAIN instance, though not so straightforwardly (see [13]). Thus ANTICHAIN is equivalent to #CSP($\{\varrho_{\leq}\}$).

The general #CSP is known to be #**P**-complete, as follows from Theorem 4.14, or the results of [81] and the examples above. We call a constraint language Γ *#-tractable* if, for every finite $\Gamma_0 \subseteq \Gamma$, the problem #CSP($\Gamma_0$) is polynomial time solvable. The language Γ is said to be #**P**-*complete* if #CSP(Γ_0) is #**P**-complete for a certain finite $\Gamma_0 \subseteq \Gamma$.

The expressive power and the polymorphisms of a constraint language again play crucial roles in determining the complexity of #CSP(Γ).

4.38 Proposition *For any constraint languages, Γ, Γ_0, on a finite set D, with Γ_0 finite, if $\Gamma_0 \subseteq \langle\Gamma\rangle$, then* #CSP($\Gamma_0$) *is reducible to* #CSP(Γ) *in polynomial time.*

4.39 Proposition *For any constraint languages Γ, Γ_0, on a finite set D, with Γ_0 finite, if $\mathsf{Pol}(\Gamma) \subseteq \mathsf{Pol}(\Gamma_0)$, then* #CSP($\Gamma_0$) *is reducible to* #CSP(Γ) *in polynomial time.*

Proposition 4.39 implies that, as with the decision CSP, the algebra $\mathbb{A}_\Gamma$ fully determines the counting complexity of a constraint language Γ. We will say that a finite algebra $\mathbb{A} = (D; F)$ is *#-tractable* [#**P**-*complete*] if so is the constraint language $\mathsf{Inv}(F)$.

The next result shows that, once again, standard constructions preserve tractability.

4.40 Theorem *Let $\mathbb{A}$ be a finite algebra. If $\mathbb{A}$ is #-tractable, then all of its subalgebras, homomorphic images and finite direct powers are also #-tractable. Conversely, if $\mathbb{A}$ has a* #**P**-*complete subalgebra, homomorphic image, or finite direct power, then $\mathbb{A}$ is* #**P**-*complete itself.*

4.41 Theorem *A finite algebra is #-tractable (*#**P**-*complete) if and only if its full idempotent reduct is #-tractable (*#**P**-*complete).*

The benchmark hard counting problems arise from binary reflexive non-symmetric relations.

4.42 Proposition *If ϱ is a binary reflexive non-symmetric relation on a finite set, then* #CSP($\{\varrho\}$) *is* #**P**-*complete.*

Theorem 4.40, Proposition 4.42, and the results from [40], provide a link between the complexity of #CSP and Mal'tsev operations, which we will now investigate. The next statement follows from [40, Theorem 9.13].

4.43 Theorem *For a finite algebra $\mathbb{A}$ the following conditions are equivalent:*

(1) *$\mathbb{A}$ does not have a Mal'tsev term operation.*

(2) *There is a finite algebra $\mathbb{B} = (B; F) \in \mathsf{var}(\mathbb{A})$, such that $\mathsf{Inv}(F)$ contains a binary reflexive non-symmetric relation.*

By Proposition 4.42, the algebra $\mathbb{B}$ from Theorem 4.43 (2) is #**P**-complete. Furthermore, Theorem 4.40 implies that $\mathbb{A}$ is also #**P**-complete.

4.44 Corollary *Every finite algebra having no Mal'tsev term operation is* #**P***-complete.*

By making use of Corollary 4.44 we can obtain a very easy proof of the dichotomy theorem for the Boolean #CSP ([19], see Theorem 4.14). First, it follows from the results of [71], that any Boolean relation which is invariant under some Mal'tsev operation on $\{0,1\}$ is also invariant under the unique affine operation on $\{0,1\}$, $x-y+z$. Hence, by Corollary 4.44, any Boolean constraint language is either #**P**-complete, or else a subset of $\mathsf{Inv}(\{x-y+z\})$. Any relation belonging to $\mathsf{Inv}(\{x-y+z\})$ is the solution space of a system of linear equations over the 2-element field, so it is possible to find a basis for this set in polynomial time. Furthermore, the number of solutions in this set equals 2^n, where n is the number of vectors in the basis.

4.45 Example The #H-COLORING problem is the counting version of the GRAPH H-COLORING problem (see Example 3.5). In this problem, the goal is to find the number of homomorphisms from a given graph G to the fixed graph H.

If the #H-COLORING problem is restricted to undirected graphs then, as proved in [28], the problem is tractable if every connected component of H is either an isolated vertex, or a complete graph with all loops, or a complete unlooped bipartite graph; otherwise the problem is #**P**-complete. The tractability part of this result is easy, and the hardness part can be easily derived from Corollary 4.44, since symmetric relations (or graphs) invariant under a Mal'tsev operation must be of the form specified above.

We will now describe a sufficient condition for Mal'tsev algebras to be #-tractable. An algebra is said to be *uniform* if, for any subalgebra $\mathbb{B}$, the blocks of every congruence of $\mathbb{B}$ are of the same size. Clearly, all two-element algebras, groups and quasi-groups are uniform.

4.46 Theorem *Every uniform Mal'tsev algebra is #-tractable.*

4.7 The Quantified CSP

The standard constraint satisfaction problem over an arbitrary finite domain can be expressed as follows: given a first-order sentence of the form $\exists x_1 \dots \exists x_l(\varrho_1 \wedge \dots \varrho_q)$, where each ϱ_i is an atomic formula, and $x_1, \dots, x_l$ are the variables appearing in the ϱ_i, determine whether the sentence is true (see Section 1). In this subsection we consider a more general framework which allows arbitrary quantifiers over constrained variables, rather than just existential quantifiers. This form of the CSP is called the *quantified* CSP, or QCSP for short. The Boolean QCSP (also known as QSAT or QBF), and some of its restrictions (such as Q3SAT), have always been standard examples of **PSPACE**-complete problems [32, 66, 74].

All the results presented in this section were first obtained in [8, 18].

4.47 Definition For a constraint language $\Gamma \subseteq R_D$, an *instance* of QCSP(Γ) is a first-order sentence $\mathcal{Q}_1 x_1 \dots \mathcal{Q}_l x_l$ $(\varrho_1 \wedge \dots \wedge \varrho_q)$, where each ϱ_i is an atomic formula involving a predicate from Γ, $x_1, \dots, x_l$ are the variables appearing in the ϱ_i, and $\mathcal{Q}_1, \dots, \mathcal{Q}_l$ are arbitrary quantifiers. The *question* is whether the sentence is true.

Clearly, an instance of CSP(Γ) corresponds to an instance of QCSP(Γ) in which all the quantifiers happen to be existential.

We note that in the Boolean case, the complexity of QCSP(Γ) has been completely classified (see Theorem 4.11). For problems over larger domains no complete classification has yet been obtained, but there are a number of known results concerning the complexity of special cases.

4.48 Example Consider the following Coloring Construction Game played by two players, Player 1 and Player 2: given an undirected graph $G = (V, E)$, a linear ordering on V (i.e., a bijection $f : V \to \{1, \ldots, |V|\}$), an ownership function $w : V \to \{1, 2\}$ (that is, each vertex v is "owned" by Player $w(v)$), and a finite set of colours D with $|D| \geq 3$. In the i'th move, the player who owns vertex $f^{-1}(i)$ (that is, Player $w(f^{-1}(i))$) colours it in one of $|D|$ available colours. Player 1 wins if all vertices are coloured at the end of the game.

Deciding whether Player 1 has a winning strategy in an instance of this game can be translated into an instance of the quantified version of the Graph $|D|$-colorability problem, QCSP($\{\neq_D\}$). To make this translation we view elements from V as variables, elements of E as constraint scopes, the relation $\neq_D$ as the only available constraint relation, the variables from $w^{-1}(1)$ as existentially quantified, the variables from $w^{-1}(2)$ as universally quantified, and the order of quantification as specified by the function f.

The problem QCSP($\{\neq_D\}$) was shown to be **PSPACE**-complete [8].

It can be shown that, for quantified constraint satisfaction problems, *surjective* polymorphisms play a similar role to that played by arbitrary polymorphisms for ordinary CSPs (cf. Theorem 4.9). Let $\mathsf{s\text{-}Pol}(\Gamma)$ denote the set of all surjective operations from $\mathsf{Pol}(\Gamma)$.

4.49 Theorem *For any constraint languages $\Gamma, \Gamma_0 \subseteq R_D$, with Γ_0 finite, if $\mathsf{s\text{-}Pol}(\Gamma) \subseteq \mathsf{s\text{-}Pol}(\Gamma_0)$, then* QCSP($\Gamma_0$) *is reducible to* QCSP(Γ) *in polynomial time.*

This theorem follows immediately from the next two propositions.

4.50 Definition For any set $\Gamma \subseteq R_D$, the set $[\Gamma]$ consists of all predicates that can be expressed using:

(1) predicates from Γ, together with the binary equality predicate $=_D$ on D,

(2) conjunction,

(3) existential quantification,

(4) universal quantification.

4.51 Proposition *For any constraint languages $\Gamma, \Gamma_0 \subseteq R_D$, with Γ_0 finite, if $[\Gamma_0] \subseteq [\Gamma]$, then* QCSP($\Gamma_0$) *is reducible to* QCSP(Γ) *in polynomial time.*

4.52 Proposition *For any constraint language Γ over a finite set, $[\Gamma] = \mathsf{Inv}(\mathsf{s\text{-}Pol}(\Gamma))$.*

Note that Proposition 4.52 intuitively means that the expressive power of constraints in the QCSP is determined by their surjective polymorphisms. Hence, in order to show that some relation ϱ belongs to $[\Gamma]$, one does not have give an explicit construction, but instead one can show that ϱ is invariant under all surjective polymorphisms of Γ, which often turns out to be significantly easier.

We remark that the operators $\mathsf{Inv}()$ and $\mathsf{s\text{-}Pol}()$ used in Proposition 4.52 form a Galois connection between R_D and the set of all surjective members of O_D which has not previously been investigated (see, e.g., survey [69]).

Using Theorem 4.49, together with Example 4.48, we can obtain a sufficient condition for **PSPACE**-completeness of QCSP(Γ), in terms of the surjective polymorphisms of Γ.

4.53 Theorem *For any finite set D with $|D| \geq 3$, and any $\Gamma \subseteq R_D$, if every $f \in \mathsf{s\text{-}Pol}(\Gamma)$ is of the form $f(x_1, \ldots, x_n) = g(x_i)$ for some $1 \leq i \leq n$ and some permutation g on D, then* QCSP(Γ) *is* **PSPACE***-complete.*

The next example uses this result to show that even predicates that give rise to trivial constraint satisfaction problems can give rise to intractable quantified constraint satisfaction problems. This can happen because non-surjective polymorphisms, which may guarantee the tractability of the CSP, do not affect the complexity of the QCSP.

4.54 Example Let τ_s be the s-ary "not-all-distinct" predicate holding on a tuple $(a_1, \ldots, a_s)$ if and only if $|\{a_1, \ldots, a_s\}| < s$. Note that $\tau_s \supseteq \{(a, \ldots, a) \mid a \in D\}$, so every instance of CSP($\{\tau_s\}$) is trivially satisfiable by assigning the same value to all variables.

However, by [70, Lemma 2.2.4], the set $\mathsf{Pol}(\{\tau_{|D|}\})$ consists of all non-surjective operations on D, together with all operations of the form given in Theorem 4.53. Hence, $\{\tau_{|D|}\}$ satisfies the conditions of Theorem 4.53, and QCSP($\{\tau_{|D|}\}$) is **PSPACE**-complete. Similar arguments can be used to show that QCSP($\{\tau_s\}$) is **PSPACE**-complete, for any s in the range $3 \leq s \leq |D|$.

On the tractability side, we have the following result. We call a semilattice operation *bounded* if the corresponding partial order is bounded (that is, it is a lattice order). Recall that the dual discriminator operation is defined by the rule

$$d(x, y, z) = \begin{cases} y & \text{if } y = z, \\ x & \text{otherwise.} \end{cases}$$

Note that the dual discriminator is a special type of near-unanimity operation.

4.55 Theorem *For any constraint language Γ over a finite set:*

(1) *if* $\mathsf{Pol}(\Gamma)$ *contains a Mal'tsev operation, or a near-unanimity operation, or a bounded semilattice operation, then* QCSP(Γ) *is tractable;*

(2) *if* $\mathsf{Pol}(\Gamma)$ *contains the dual discriminator operation, then* QCSP(Γ) *is in* **NL**.

Recall that the graph of a permutation π is the binary relation $\{(x, y) \mid y = \pi(x)\}$ (or the binary predicate $\pi(x) = y$), For the special case when Γ contains the set Δ of all graphs of permutations, there is a trichotomy result which says that such problems are either tractable, or **NP**-complete, or **PSPACE**-complete. (We remark that the complexity of the standard CSP(Γ) for such sets Γ was completely classified in [24].)

To state this trichotomy result we need to define two additional surjective operations:

- The k-ary *near projection* operation,

$$l_k(x_1, \dots, x_k) = \begin{cases} x_1 & \text{if } x_1, \dots, x_k \text{ are all different,} \\ x_k & \text{otherwise.} \end{cases}$$

- The ternary *switching* operation,

$$s(x, y, z) = \begin{cases} x & \text{if } y = z, \\ y & \text{if } x = z, \\ z & \text{otherwise.} \end{cases}$$

4.56 Theorem *Let $\Delta \subseteq \Gamma \subseteq R_D$, and $|D| \geq 3$.*

- *If* s-Pol(Γ) *contains the dual discriminator d, or the switching operation s, or (when $|D| \in \{3, 4\}$) an affine operation, then* QCSP(Γ) *is in* **PTIME***;*
- *otherwise, if* s-Pol(Γ) *contains $l_{|D|}$, then* QCSP(Γ) *is* **NP***-complete;*
- *otherwise* QCSP(Γ) *is* **PSPACE***-complete.*

5 The infinite-valued CSP

There are many computational problems which can be represented as constraint satisfaction problems, but require an infinite set of values. In order to avoid representation problems for infinite objects, we will consider CSPs with infinite sets of values in the following form: fix an infinite relational structure $\mathcal{B}$ of finite signature; the input then is a finite structure $\mathcal{A}$ of the same signature, and the question is whether there is a homomorphism from $\mathcal{A}$ to $\mathcal{B}$.

Here are two well-known examples of problems with an infinite set of possible values.

5.1 Example An instance of the Acyclic Digraph problem is a directed graph G, and the question is whether G is acyclic, that is, contains no directed cycles. It is easy to see that this problem is equivalent to Hom($\mathcal{B}$) where $\mathcal{B} = (\mathbb{N}; <)$, since a directed graph is acyclic if and only if its vertices can be numbered in such a way that every arc leads from a vertex with smaller number to a vertex with a greater one. This problem is tractable.

5.2 Example An instance of the Betweenness problem is a pair (A, T) where A is a finite set and $T \subseteq A^3$; the question is whether there is a function $f : A \to \{1, \dots, |A|\}$ such that, for every triple $(a, b, c) \in T$, we have either $f(a) < f(b) < f(c)$ or $f(a) > f(b) > f(c)$. This problem is equivalent to Hom($\mathcal{B}$) with $\mathcal{B} = (\mathbb{N}, R)$ where

$$R = \{(x, y, z) \in \mathbb{N}^3 \mid x < y < z \text{ or } x > y > z\}.$$

This problem is **NP**-complete [32].

It was shown in Proposition 3.7 of [4] that the former problem cannot be represented as CSP(Γ) for any constraint language Γ over a finite set D (in fact, the above mentioned proposition is an even stronger claim); for the latter problem, the proof is similar.

5.1 Applicability of polymorphisms

For a family Γ of relations over an infinite set, let $\langle\Gamma\rangle$ be defined exactly as in the finite case (see Definition 4.3). In order to investigate the applicability of the algebraic approach, described in previous sections, to the infinite-valued CSP, the first question to be asked is whether the complexity is determined by the polymorphisms of the constraint relations; that is, whether $\langle\Gamma\rangle = \mathsf{Inv}(\mathsf{Pol}(\Gamma))$ when Γ is a finite constraint language over an infinite domain. It is not hard to see that the inclusion $\langle\Gamma\rangle \subseteq \mathsf{Inv}(\mathsf{Pol}(\Gamma))$ always holds. However, this inclusion can be strict, as the next example shows.

5.3 Example Consider $\Gamma = \{R_1, R_2, R_3\}$ on $\mathbb{N}$, where $R_1 = \{(a,b,c,d) \mid a = b \text{ or } c = d\}$, $R_2 = \{(1)\}$, and $R_3 = \{(a, a+1) \mid a \in \mathbb{N}\}$. It is not difficult to show that every polymorphism of Γ is a projection, and hence $\mathsf{Inv}(\mathsf{Pol}(\Gamma))$ is the set of all relations on $\mathbb{N}$. However, one can check that, for example, the unary relation consisting of all even numbers does not belong to $\langle\Gamma\rangle$.

However, for some countable structures $\mathcal{B}$, the required equality does hold, as the next result indicates.

A countable structure $\mathcal{B}$ (of finite signature) is called *homogeneous* if every isomorphism between any pair of substructures is induced by an automorphism of $\mathcal{B}$. A countable structure is called *ω-categorical* if it is determined (up to isomorphism) by its first-order theory. It is known that every countable homogeneous structure is ω-categorical, and that a countable structure is ω-categorical if and only if its automorphism group, when acting on the set of all n-tuples (for any n) of elements from the structure, has only finitely many orbits (see, e.g., [41]).

5.4 Theorem [4, 5] *If $\mathcal{B}_\Gamma$ is a countable ω-categorical structure then $\langle\Gamma\rangle = \mathsf{Inv}(\mathsf{Pol}(\Gamma))$.*

Many examples of countable homogeneous structures, as well as remarks on the complexity of the corresponding constraint satisfaction problems, can be found in [4, 5].

5.2 The interval-valued CSP

One form of infinite-valued CSP which has been widely studied in artificial intelligence is the case where the values taken by the variables are intervals on the real line. This setting is used to model temporal behaviour of systems, where the intervals represent time intervals during which events occur. The most popular such formalism is Allen's interval algebra (AIA for short), introduced in [1], which concerns binary qualitative relations between intervals. This algebra contains 13 basic relations (see Table 1), corresponding to the 13 distinct ways in which two given intervals can be related. The complete set of relations in AIA consists of the $2^{13} = 8192$ possible unions of the basic relations.

Let Γ be a constraint language over the set of intervals on the real line, whose elements are members of Allen's interval algebra, and let $\mathcal{B}_\Gamma$ be the corresponding relational structure. It is not hard to see that every instance of CSP(Γ) can also be (more graphically) viewed as a directed graph whose vertices represent the variables and whose arcs are each labelled with a relation from Γ. The question would then be whether one can assign intervals to the vertices so that all constraints on the arcs are satisfied.

Basic relation		Example	Endpoints
I precedes J	p	III	$I^+ < J^-$
J preceded by I	p^{-1}	JJJ	
I meets J	m	IIII	$I^+ = J^-$
J met by I	m^{-1}	JJJJ	
I overlaps J	o	IIII	$I^- < J^- < I^+$,
J overl. by I	o^{-1}	JJJJ	$I^+ < J^+$
I during J	d	III	$I^- > J^-$,
J includes I	d^{-1}	JJJJJJJ	$I^+ < J^+$
I starts J	s	III	$I^- = J^-$,
J started by I	s^{-1}	JJJJJJJ	$I^+ < J^+$
I finishes J	f	III	$I^+ = J^+$,
J finished by I	f^{-1}	JJJJJJJ	$I^- > J^-$
I equals J	$\equiv$	IIII	$I^- = J^-$,
		JJJJ	$I^+ = J^+$

Table 1: The 13 basic relations in Allen's interval algebra.

Some well-known combinatorial problems can be represented as CSP(Γ) for a suitable subset Γ of AIA, as the next example indicates.

5.5 Example An undirected graph is called an interval graph if it possible to assign (open) intervals to its nodes so that two intervals intersect if and and only if the corresponding nodes are adjacent. An instance of the Interval Graph Sandwich problem [34] consists of two (undirected) graphs $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ such that $E_1 \subseteq E_2$. The question is whether there is E such that $E_1 \subseteq E \subseteq E_2$ and $G = (V, E)$ is an interval graph. This problem is known to be **NP**-complete [34].

This problem can be represented as CSP(Γ) where Γ consists of two relations: "disjoint" (given by $\mathsf{p} \cup \mathsf{p}^{-1} \cup \mathsf{m} \cup \mathsf{m}^{-1}$) and its complement, "intersect" (the union of the other nine basic relations). Indeed, let V be the set of variables; then, to any edge $e \in E_1$ assign the constraint "intersect", to any edge $e \notin E_2$ assign the constraint "disjoint", and leave all other pairs of variables unrelated. Solutions of this CSP precisely correspond to interval graph sandwiches.

Note that the case when $G_1 = G_2$ is known as Interval Graph Recognition problem, which is tractable, but this problem is not of the form CSP(Γ) because we cannot leave variables unrelated.

Choosing other pairs of complementary relations, one can obtain other graph sandwich problems, such as the Overlap (or Circle) Graph Sandwich problem [34, 55]

The general CSP problem for AIA is **NP**-complete, as follows from the above example. The problem of classifying subsets of AIA with respect to the complexity of the corresponding CSP has attracted much attention in artificial intelligence (see, for example, [75]).

Allen's interval algebra has three operations on relations: composition, intersection, and inversion. Note that these three operations can each be represented by using conjunction and existential quantification, so, for any subset Γ of AIA, the subalgebra Γ' of AIA generated by Γ has the property that $\Gamma' \subseteq \langle\Gamma\rangle$. It follows from Lemma 3.3 of [4] (which is Corollary 4.6

$$\mathcal{S}_{\mathsf{p}} = \{r \mid r \cap (\mathsf{pmod}^{-1}\mathsf{f}^{-1})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{p})^{\pm 1} \subseteq r\}$$
$$\mathcal{S}_{\mathsf{d}} = \{r \mid r \cap (\mathsf{pmod}^{-1}\mathsf{f}^{-1})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{d}^{-1})^{\pm 1} \subseteq r\}$$
$$\mathcal{S}_{\mathsf{o}} = \{r \mid r \cap (\mathsf{pmod}^{-1}\mathsf{f}^{-1})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{o})^{\pm 1} \subseteq r\}$$
$$\mathcal{A}_1 = \{r \mid r \cap (\mathsf{pmod}^{-1}\mathsf{f}^{-1})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{s}^{-1})^{\pm 1} \subseteq r\}$$
$$\mathcal{A}_2 = \{r \mid r \cap (\mathsf{pmod}^{-1}\mathsf{f}^{-1})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{s})^{\pm 1} \subseteq r\}$$
$$\mathcal{A}_3 = \{r \mid r \cap (\mathsf{pmodf})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{s})^{\pm 1} \subseteq r\}$$
$$\mathcal{A}_4 = \{r \mid r \cap (\mathsf{pmodf}^{-1})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{s})^{\pm 1} \subseteq r\}$$

$$\mathcal{E}_{\mathsf{p}} = \{r \mid r \cap (\mathsf{pmods})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{p})^{\pm 1} \subseteq r\}$$
$$\mathcal{E}_{\mathsf{d}} = \{r \mid r \cap (\mathsf{pmods})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{d})^{\pm 1} \subseteq r\}$$
$$\mathcal{E}_{\mathsf{o}} = \{r \mid r \cap (\mathsf{pmods})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{o})^{\pm 1} \subseteq r\}$$
$$\mathcal{B}_1 = \{r \mid r \cap (\mathsf{pmods})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{f}^{-1})^{\pm 1} \subseteq r\}$$
$$\mathcal{B}_2 = \{r \mid r \cap (\mathsf{pmods})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{f})^{\pm 1} \subseteq r\}$$
$$\mathcal{B}_3 = \{r \mid r \cap (\mathsf{pmod}^{-1}\mathsf{s}^{-1})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{f}^{-1})^{\pm 1} \subseteq r\}$$
$$\mathcal{B}_4 = \{r \mid r \cap (\mathsf{pmod}^{-1}\mathsf{s})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{f}^{-1})^{\pm 1} \subseteq r\}$$

$$\mathcal{E}^* = \left\{ r \;\middle|\; \begin{array}{l} 1)\ r \cap (\mathsf{pmod})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{s})^{\pm 1} \subseteq r, \text{ and} \\ 2)\ r \cap (\mathsf{ff}^{-1}) \neq \emptyset \Rightarrow (\equiv) \subseteq r \end{array} \right\}$$

$$\mathcal{S}^* = \left\{ r \;\middle|\; \begin{array}{l} 1)\ r \cap (\mathsf{pmod}^{-1})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{f}^{-1})^{\pm 1} \subseteq r, \text{ and} \\ 2)\ r \cap (\mathsf{ss}^{-1}) \neq \emptyset \Rightarrow (\equiv) \subseteq r \end{array} \right\}$$

$$\mathcal{H} = \left\{ r \;\middle|\; \begin{array}{l} 1)\ r \cap (\mathsf{os})^{\pm 1} \neq \emptyset \;\&\; r \cap (\mathsf{o}^{-1}\mathsf{f})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{d})^{\pm 1} \subseteq r, \text{ and} \\ 2)\ r \cap (\mathsf{ds})^{\pm 1} \neq \emptyset \;\&\; r \cap (\mathsf{d}^{-1}\mathsf{f}^{-1})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{o})^{\pm 1} \subseteq r, \text{ and} \\ 3)\ r \cap (\mathsf{pm})^{\pm 1} \neq \emptyset \;\&\; r \not\subseteq (\mathsf{pm})^{\pm 1} \Rightarrow (\mathsf{o})^{\pm 1} \subseteq r \end{array} \right\}$$

$$\mathcal{A}_{\equiv} = \{r \mid r \neq \emptyset \Rightarrow (\equiv) \subseteq r\}$$

Table 2: The 18 maximal tractable subalgebras of Allen's algebra.

for the infinite case) that CSP(Γ) and CSP(Γ') are polynomial-time equivalent. Hence it is sufficient to classify all *subalgebras* of AIA.

Using computations in subalgebras of AIA, manipulations with primitive positive formulas (called *derivations* in [55]) and a number of new **NP**-completeness results, a complete classification of the complexity of all subsets of AIA was accomplished in [55], where the following result was obtained.

5.6 Theorem *Let Γ be a subset of Allen's interval algebra. If Γ is contained in one of the eighteen subalgebras listed in Table* 2*, then* CSP(Γ) *is tractable; otherwise it is* **NP***-complete.*

In Table 2, for the sake of brevity, relations between intervals are written as collections of basic relations. So, for instance, we write (pmod) instead of $\mathsf{p} \cup \mathsf{m} \cup \mathsf{o} \cup \mathsf{d}$. We also use the symbol $\pm$, which should be interpreted as follows: a condition involving $\pm$ means the conjunction of two conditions, one corresponding to $+$ and one corresponding to $-$. For example, the condition $(\mathsf{o})^{\pm 1} \subseteq r \iff (\mathsf{d})^{\pm 1} \subseteq r$ means that both $(\mathsf{o}) \subseteq r \iff (\mathsf{d}) \subseteq r$ and $(\mathsf{o}^{-1}) \subseteq r \iff (\mathsf{d}^{-1}) \subseteq r$ hold.

It follows from Theorem 5.6 that CSP($\{r\}$), where r is a single relation in AIA, is **NP**-complete if and only if r either satisfies $r \cap r^{-1} = (\mathsf{mm}^{-1})$ or is a relation with $r \cap r^{-1} = \emptyset$ and such that neither r nor r^{-1} is contained in one of $(\mathsf{pmod}^{-1}\mathsf{sf}^{-1})$, $(\mathsf{pmod}^{-1}\mathsf{s}^{-1}\mathsf{f}^{-1})$, (pmodsf) and (pmodsf^{-1}).

It was noted in [4, 5] that AIA (without its operations) is in fact a homogeneous relational structure. Since we may assume, without loss of generality, that all intervals under consideration have rational endpoints, we obtain a countable homogeneous structure of finite signature. Therefore, by Theorem 5.4, the complexity classification problem for subsets of AIA can be tackled using polymorphisms. Such an approach may provide a route to simplifying the involved classification proof given in [55].

References

[1] J. F. Allen, Maintaining knowledge about temporal intervals, *Comm. ACM* **26** (1983), 832–843.

[2] J. F. Allen, *Natural Language Understanding*, Benjamin Cummings, 1994.

[3] J. Berman, E. Kiss, P. Pröhle, and Á. Szendrei, The set of types of a finitely generated variety, *Discrete Math.* **112** (1–3) (1993), 1–20.

[4] M. Bodirsky, Constraint satisfaction with infinite domains, Ph.D. Thesis, Humboldt University, Berlin, 2004.

[5] M. Bodirsky and J. Nešetřil, Constraint satisfaction problems with countable homogeneous structures, in: *Computer Science Logic (Vienna, 2003)*, Lecture Notes in Comput. Sci. **2803**, Springer, Berlin, 2003, 44–57.

[6] E. Böhler, E. Hemaspaandra, S. Reith, and H. Vollmer, Equivalence and isomorphism for Boolean constraint satisfaction, in: *Computer Science Logic (Edinburgh, 2002)*, Lecture Notes in Comput. Sci. **2471**, Springer, Berlin, 2002, 412–426.

[7] E. Böhler, E. Hemaspaandra, S. Reith, and H. Vollmer, The complexity of Boolean constraint isomorphism, in: *STACS 2004 (Montpellier, 2004)*, Lecture Notes in Comput. Sci. **2996**, Springer, Berlin, 2004, 164–175.

[8] F. Börner, A. Bulatov, P. Jeavons, and A. Krokhin, Quantified constraints: Algorithms and complexity, in: *Computer Science Logic (Vienna, 2003)*, Lecture Notes in Comput. Sci. **2803**, Springer, Berlin, 2003, 58–70.

[9] A. Bulatov, A dichotomy theorem for constraints on a three-element set, in: *Foundations of Computer Science (Vancouver, BC, 2002)*, IEEE Comput. Soc., 2002, 649–658.

[10] A. Bulatov, Mal'tsev constraints are tractable, Technical Report PRG-RR-02-05, Computing Laboratory, University of Oxford, UK, 2002.

[11] A. Bulatov, Tractable conservative constraint satisfaction problems, in: *Logic in Computer Science (Ottawa, ON, 2003)*, IEEE Comput. Soc., 2003, 321–330.

[12] A. Bulatov, Combinatorial problems raised from 2-semilattices, *J. Algebra*, accepted for publication.

[13] A. Bulatov and V. Dalmau, Towards a dichotomy theorem for the counting constraint satisfaction problem, in: *Foundations of Computer Science (Boston, MA, 2003)*, IEEE Comput. Soc., 2003, 562–571.

[14] A. Bulatov and P. Jeavons, Algebraic structures in combinatorial problems, Technical Report MATH-AL-4-2001, Technische Universität Dresden, Germany, 2001.

[15] A. Bulatov, P. Jeavons, and M. Volkov, Finite semigroups imposing tractable constraints, in: *Semigroups, Algorithms, Automata and Languages* (Gracinda M.S.Gomes, Jean-Eric Pin, Pedro V.Silva, eds), World Scientific, Singapore, 2002, 313–329.

[16] A. Bulatov, A. Krokhin, and P. Jeavons, Constraint satisfaction problems and finite algebras, in: *Automata, Languages and Programming (Geneva, 2000)*, Lecture Notes in Comput. Sci. **1853**, Springer, Berlin, 2000, 272–282.

[17] A. Bulatov, A. Krokhin, and P. Jeavons, Classifying complexity of constraints using finite algebras, *SIAM J. Comput.*, accepted for publication.

[18] H. Chen, Quantified constraint satisfaction problems: Closure properties, complexity, and algorithms, manuscript, 2003.

[19] N. Creignou and M. Hermann, Complexity of generalized satisfiability counting problems, *Inform. and Comput.* **125** (1) (1996), 1–12.

[20] N. Creignou, S. Khanna, and M. Sudan. *Complexity Classifications of Boolean Constraint Satisfaction Problems*, SIAM Monographs on Discrete Mathematics and Applications **7**, SIAM, Philadelphia, 2001.

[21] P. Crescenzi and G. Rossi, On the Hamming distance of constraint satisfaction problems, *Theoret. Comput. Sci.* **288** (1) (2002), 85–100.

[22] V. Dalmau, Some dichotomy theorems on constant-free quantified Boolean formulas, Technical Report TR LSI-97-43-R, Department LSI, Universitat Politecnica de Catalunya, 1997.

[23] V. Dalmau, Constraint satisfaction problems in non-deterministic logarithmic space, in: *Automata, Languages and Programming (Malaga, 2002)*, Lecture Notes in Comput. Sci. **2380**, Springer, Berlin, 2002, 414–425.

[24] V. Dalmau, A new tractable class of constraint satisfaction problems, *Ann. Math. Artif. Intell.*, to appear.

[25] V. Dalmau, Ph. G. Kolaitis, and M. Y. Vardi, Constraint satisfaction, bounded treewidth, and finite-variable logics, in: *Principles and Practice of Constraint Programming (Ithaca, NY, 2002)*, Lecture Notes in Comput. Sci. **2470**, Springer, Berlin, 2002, 310–326.

[26] V. Dalmau and J. Pearson, Set functions and width 1 problems, in: *Principles and Practice of Constraint Programming (Alexandria, VA, 1999)*, Lecture Notes in Comput. Sci. **1713**, Springer, Berlin, 1999, 159–173.

[27] N. W. Dunkin, J. E. Bater, P. G. Jeavons, and D. A. Cohen, Towards high order constraint represenations for the frequency assignment problem, Technical Report CSD-TR-98-05, Department of Computer Science, Royal Holloway, University of London, Egham, Surrey, UK, 1998.

[28] M. Dyer and C. Greenhill, The complexity of counting graph homomorphisms, *Random Structures Algorithms* **17** (2000), 260–289.

[29] H.-D. Ebbinghaus and J. Flum, *Finite Model Theory*, 2nd ed., Perspectives in Mathematical Logic, Springer, Berlin, 1999.

[30] T. Feder and M. Y. Vardi, The computational structure of monotone monadic SNP and constraint satisfaction: A study through Datalog and group theory, *SIAM J. Comput.* **28** (1998), 57–104.

[31] E. C. Freuder and R. Wallace, Partial constraint satisfaction, *Artificial Intelligence* **58** (1992), 21–70.

[32] M. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, Freeman, San Francisco, CA, 1979.

[33] M. Goldmann and A. Russell, The complexity of solving equations over finite groups, *Inform. and Comput.* **178** (1) (2002), 253–262.

[34] M.C. Golumbic, H. Kaplan, and R. Shamir, Graph sandwich problems, *J. Algorithms* **19** (3) (1995), 449–473.

[35] G. Gottlob, L. Leone, and F. Scarcello, Hypertree decomposition and tractable queries, *J. Comput. System Sci.* **64** (3) (2002), 579–627.

[36] M. Grohe, The complexity of homomorphism and constraint satisfaction problems seen from the other side, in: *Foundations of Computer Science (Boston, MA, 2003)*, IEEE Comput. Soc., 2003, 552–561.

[37] M. Gyssens, P. G. Jeavons, and D. A. Cohen, Decomposing constraint satisfaction problems using database techniques, *Artificial Intelligence* **66** (1) (1994) 57–89.

[38] P. Hell, Algorithmic aspects of graph homomorphisms, in: *Surveys in Combinatorics 2003* (C. Wensley, ed.), LMS Lecture Note Series **307**, Cambridge University Press, 2003, 239–276.

[39] P. Hell and J. Nešetřil, On the complexity of H-coloring, *J. Combin. Theory Ser. B* **48** (1990), 92–110.

[40] D. Hobby and R. N. McKenzie, *The Structure of Finite Algebras*, Contemporary Mathematics **76**, Amer. Math. Soc., Providence, R.I., 1988.

[41] W. Hodges, *A Shorter Model Theory*, Cambridge University Press, 1997.

[42] P. G. Jeavons, On the algebraic structure of combinatorial problems, *Theoret. Comput. Sci.* **200** (1998), 185–204.

[43] P. G. Jeavons, D. A. Cohen, and M. C. Cooper, Constraints, consistency and closure, *Artificial Intelligence* **101** (1–2) (1998), 251–265.

[44] P. G. Jeavons, D. A. Cohen, and M. Gyssens, Closure properties of constraints, *J. ACM* **44** (1997), 527–548.

[45] P. G. Jeavons, D. A. Cohen, and M. Gyssens, How to determine the expressive power of constraints, *Constraints* **4** (1999), 113–131.

[46] P. Jonsson and A. Krokhin, Recognizing frozen variables in constraint satisfaction problems, *Theoret. Comput. Sci.* **329** (1–3) (2004), 93–113.

[47] L. Juban, Dichotomy theorem for generalized unique satisfiability problem, in: *Fundamentals of Computation Theory (Iasi, 1999)*, Lecture Notes in Comput. Sci. **1684**, Springer, Berlin, 1999, 327–337.

[48] H. Kautz and B. Selman, Planning as satisfiability, in: *Tenth European Conference on Artificial Intelligence (Vienna, 1992)*, John Wiley and Sons, Chichester, 1992, 359–363.

[49] D. Kavvadias and M. Sireni, The inverse satisfiability problem, *SIAM J. Comput.* **28** (1) (1998) 152–163.

[50] S. Khanna, M. Sudan, L. Trevisan, and D. Williamson, The approximability of constraint satisfaction problems, *SIAM J. Comput.* **30** (6) (2001), 1863–1920.

[51] L. Kirousis and Ph. Kolaitis, A dichotomy in the complexity of propositional circumscription, in: *Logic in Computer Science (Boston, MA, 2001)*, IEEE Comput. Soc., 2001, 71–80.

[52] L. Kirousis and Ph. Kolaitis, On the complexity of model checking and inference in minimal models, in: *Logic Programming and Nonmonotonic Reasoning (Vienna, 2001)*, Lecture Notes in Comput. Sci. **2173**, Springer, Berlin, 2001, 42–53.

[53] L. Kirousis and Ph. Kolaitis, The complexity of minimal satisfiability problems, *Inform. and Comput.* **187** (2003), 20–39.

[54] Ph. G. Kolaitis and M. Y. Vardi, Conjunctive-query containment and constraint satisfaction, *J. Comput. System Sci.* **61** (2000), 302–332.

[55] A. Krokhin, P. Jeavons, and P. Jonsson, Reasoning about temporal relations: The maximal tractable subalgebras of Allen's interval algebra, *J. ACM* **50** (5) (2003), 591–640.

[56] R. E. Ladner, On the structure of polynomial time reducibility, *J. ACM* **22** (1975), 155–171.

[57] B. Larose and L. Zádori, Taylor terms, constraint satisfaction and the complexity of polynomial equations over finite algebras, manuscript, 2003.

[58] T. Łuczak and J. Nešetřil, A probabilistic approach to the dichotomy problem, Technical report 2003-640, KAM-DIMATIA Series, 2003.

[59] A. K. Mackworth, Constraint satisfaction, in: *Encyclopedia of Artificial Intelligence, vol. 1* (S. C. Shapiro, ed.), Wiley Interscience, 1992, 285–293.

[60] S. S. Marchenkov, Homogeneous algebras, *Problemy Kibernet.* **39** (1982), 85–106, (in Russian).

[61] K. Marriott and P. J. Stuckey, *Programming with Constraints: an Introduction*, MIT Press, 1998.

[62] R. N. McKenzie, G. F. McNulty, and W. F. Taylor, *Algebras, Lattices and Varieties*, vol. I, Wadsworth and Brooks, CA, 1987.

[63] U. Montanari, Networks of constraints: Fundamental properties and applications to picture processing, *Inform. Sci.* **7** (1974), 95–132.

[64] C. Moore, P. Tesson, and D. Therien, Satisfiability of systems of equations over finite monoids. in: *Mathematical Foundations of Computer Science (Marianske Lazne, 2001)*, Lecture Notes in Comput. Sci. **2136**, Springer, Berlin, 2001, 537–547.

[65] B. A. Nadel, Constraint satisfaction in Prolog: complexity and theory-based heuristics, *Inform. Sci.* **83** (3–4) (1995), 113–131.

[66] C. H. Papadimitriou, *Computational Complexity*, Addison-Wesley, 1994.

[67] J. K. Pearson and P. G. Jeavons, A survey of tractable constraint satisfaction problems, Technical Report CSD-TR-97-15, Royal Holloway, University of London, July 1997.

[68] N. Pippenger, *Theories of Computability*, Cambridge University Press, Cambridge, 1997.

[69] R. Pöschel, Galois connections for operations and relations, Technical Report MATH-AL-8-2001, Technische Universität Dresden, Germany, 2001.

[70] R. Pöschel and L. A. Kalužnin, *Funktionen- und Relationenalgebren*, DVW, Berlin, 1979.

[71] E. L. Post, The two-valued iterative systems of mathematical logic, in: *Annals of Mathematical Studies* **5**, Princeton University Press, 1941.

[72] J. S. Provan and M. O. Ball, The complexity of counting cuts and of computing the probability that a graph is connected, *SIAM J. Comput.* **12** (4) (1983), 777–788.

[73] S. Reith and H. Vollmer, Optimal satisfiability for propositional calculi and constraint satisfaction problems, *Inform. and Comput.* **186** (1) (2003), 1–19.

[74] T. J. Schaefer, The complexity of satisfiability problems, in: *Tenth ACM Symposium on Theory of Computing (San Diego, CA, 1978)*, ACM Press, New York, 1978, 216–226.

[75] E. Schwalb and L. Vila, Temporal constraints: a survey, *Constraints* **3** (2–3) (1998), 129–149.

[76] A. Szendrei, *Clones in Universal Algebra*, Université de Montréal, 1986.

[77] A. Szendrei, Simple surjective algebras having no proper subalgebras, *J. Austral. Math. Soc. Ser. A* **48** (1990), 434–454.

[78] W. Taylor, Varieties obeying homotopy laws, *Canad. J. Math.* **29** (1977) 498–527.

[79] E. Tsang, *Foundations of Constraint Satisfaction*, Academic Press, London, 1993.

[80] L. Valiant, The complexity of computing the permanent, *Theoret. Comput. Sci.* **8** (1979), 189–201.

[81] L. Valiant, The complexity of enumeration and reliability problems, *SIAM J. Comput.* **8** (3) (1979), 410–421.

On the automata functional systems

V. B. KUDRYAVTSEV

MaTIS Dept.
M. V. Lomonosov Moscow State University
Vorobjevy Gory, GZ MGU, 119899 Moscow
Russia

Abstract

The paper gives the main results on the problems of expressibility and completeness for the automata functional systems. These results were obtained over the past 30 years, that is, since the appearance and during the years of formation of automata theory. The description of the properties of the automata functional systems is done for model systems in order of their increasing complexity. The first to be considered are automata without memory, i.e., the functions of k-valued logic; then we consider automata with limited memory, i.e., the above-mentioned functions with delays, and finally, finite automata, i.e., automata functions.

1 Introduction

The notion of automaton is one of the most important notions in mathematics. It appeared at the juncture of its different branches, as well as in engineering, biology and other fields of science. From the content point of view an automaton is a system with input and output channels. Its input channels receive sequential information which the automaton processes with regard to the structure of the sequence and emits it through its output channels. These systems permit the interconnection of channels. The mapping of input into output sequences is called an automaton function, and the possibility of creating such new mappings by means of combining automata leads to the algebra of automata functions.

The automata and their algebras were first investigated in the 1930's, and most intensively since the 1950's.

The foundations of the science were laid out in the works of A. Turing, C. Shannon, E. Moore, S. Kleene, and other authors of the famous collection of papers "Automata Studies" [22]. The subsequent investigations of automata algebra were carried out under the great influence of the well-known article by Yablonskii on the theory of functions of k-valued logic [29]. The set of such functions can be considered as a set of automata without memory with superposition operation on it. A number of problems were formulated for these functions such as problems of expressibility, completeness, bases, problems of a closed class lattice, and others. A well-developed technique of preserving predicates appeared as the key to solving these problems. All this turned out to be very effective for automata algebras which, in what follows, are referred to as functional systems. Expressibility is understood as the possibility of obtaining functions of one set from the functions of another set with the

V. B. Kudryavtsev and I. G. Rosenberg (eds.),
Structural Theory of Automata, Semigroups, and Universal Algebra, 215–240.

help of prescribed operations, and completeness means the expressibility of all functions in terms of the prescribed ones.

In the review the study of functional systems is carried out on a number of model objects, beginning with automata without memory, i.e., the functions of k-valued logic; then come automata with limited memory, i.e., such functions with delays, and finally we consider finite automata, i.e., automata functions of the general form. Superpositions are used as operations, and in the last case feedback is also used.

The fundamental results of Post concerning the structure of the lattice of closed classes of Boolean functions are given for automata without memory. (It is not easy to become acquainted with these results today as the books which contain them [24, 30] have become a rarity.) Then the most essential results for the functions of k-valued logic are given. Their basis is formed by the approach developed by A. V. Kuznetsov and S. V. Yablonskii. The central idea of this approach is the concept of a precomplete class. For finitely-generated systems of such functions the family of precomplete classes forms a criterial system; in other words, an arbitrary set is complete exactly when it is not a subset of a precomplete class. The set of these precomplete classes turned to be finite, and from their characterization follows the algorithmic solvability of the completeness problem. Proceeding in this direction and describing explicitly all precomplete classes S. V. Yablonskii solved the completeness problem for functions of three-valued logic, and, together with A. V. Kuznetsov, found certain families of precomplete classes for an arbitrary finite-valued logic. Then by the efforts of many researchers new families of this kind were discovered in succession, and the conclusive constructions were obtained by Rosenberg [26]. A summary of these constructions is also given here.

For automata with limited memory, solutions of completeness and expressibility problems are given as well as problems of weak versions of these assertions. By an automaton of this kind is meant a pair (f, t), where f is a function of k-valued logic and t is its calculation time. Weak completeness means the possibility of getting any function with some kind of delay from initial pairs with the help of superpositions. The case of functions of two-valued logic is considered in full detail. Here precomplete classes are also used as solving tools. In contrast to automata without memory, here the family of precomplete classes turned out to be countable. At the same time the weak completeness problem remains algorithmically decidable.

Another generalization of automata without memory is the class of linear automata with superposition and feedback operations. Here the situation is similar to the case of automata with limited memory. The description of the set of all precomplete classes is also possible; this set is countable. From this description an algorithm deciding the completeness of finite automata systems is deduced [9].

The transition to the general case of automata affords us an opportunity to discover the continuum set of precomplete classes [17] and the algorithmic undecidability of the completeness problem [15]. Therefore it is of current importance to find ways of both weakening the completeness properties and, conversely, of enriching this notion.

The first direction is realized by considering problems of r-completeness and A-completeness, which consist, respectively, in checking the possibility of generating all mappings on words of length r and also such mappings for any fixed r. The main results here are an explicit description of all r- and A-precomplete classes and the algorithmic undecidability of the A-completeness problem [8].

Another realization in this direction is considering automata with Baire metrics with the given precision ε [23]. The problem of ε-completeness also turned out to be undecidable, though there exists a constructive automata kernel (consisting of strongly connected automata) with ε-completeness property.

There exists a dual view to the sequential view of automata—the acceptor view proposed by Kleene. In this case automata are factored by the property of representing identical events. The completeness property in this case is called Kleene-completeness. This problem is also undecidable [13].

We can also consider Kleene algebra over regular events. In this case the completeness problem is still undecidable, yet there exist interesting subalgebras for which completeness is decidable [14].

The second direction is realized by stratifying all finite automata systems from the point of view of completeness and of types. Each type is formed by all systems that contain a given Post class of automata without memory. The main result here is an explicit indication of the separability level of the algorithmically decidable cases of the automata systems on the Post diagram. This level turns out to be correct for the A-completeness case, too [6].

Together with the achievements in solving the main problems connected with expressibility and completeness, the review also outlines those directions which are insufficiently or poorly developed. Only model cases of automata functional systems are dealt with here. The general constructions implemented by the author in [18] are not touched upon here.

2 The main notions and problems

Let $N = \{1, 2, \dots\}$, $N_0 = N \cup \{0\}$, $N_1 = N \setminus \{1\}$, and for $h \in N$ let ${N_1}^h = \{1, 2, \dots, h\}$. Let us consider a set M and a mapping $\omega : M^n \to M$, where M^n is the n-th Cartesian power of the set M and $n \in N$. Let P_M be the set of all such mappings ω for any n, and $Q \subseteq P_M$.

We consider the universal algebra (u.a.) $\mathcal{M} = (M, \Omega)$ where M is called a *carrier* and Ω is a class of operations. We associate a sequence of sets $\overline{M}^{(i)}, i \in N$, with a subset $\overline{M} \subseteq M$ in the following way.

Let $\overline{M}^{(1)} = \overline{M}$. Then, the set $\overline{M}^{(i+1)}$ consists of all elements m from M for which there exist ω from Ω and $m_1, m_2, \dots, m_n$ from $\overline{M}^{(i)}$ such that $m = \omega(m_1, m_2, \dots, m_n)$. We denote

$$I_\Omega(\overline{M}) = \bigcup_{i=1}^{\infty} \overline{M}^{(i)}.$$

It is not difficult to see that I_Ω is a closure operator on the set $\mathcal{B}(M)$ formed by all subsets of the set M. Thus, for I_Ω the conditions $I_\Omega(\overline{M}) \supseteq \overline{M}$, $I_\Omega(I_\Omega(\overline{M})) = I_\Omega(\overline{M})$ are always fulfilled, and if $\overline{M} \supseteq M'$, then $I_\Omega(\overline{M}) \supseteq I_\Omega(M')$. The set $I_\Omega(\overline{M})$ is called the *closure* of the set $\overline{M}$, and $\overline{M}$ itself is a generating set for $I_\Omega(\overline{M})$. The set $\overline{M}$ is called *closed* if $\overline{M} = I_\Omega(\overline{M})$.

Let $\Sigma(\mathcal{M})$ be the set of all closed subsets in $\mathcal{M}$. It is said that $\overline{M}$ is *expressible in terms of* M' if $\overline{M} \subseteq I_\Omega(M')$. The set $\overline{M}$ is called *complete* if $I_\Omega(\overline{M}) = M$. A complete set is called a *basis* if none of its proper subsets is complete.

The main problems for M which will interest us are those of expressibility and completeness as well as the problems on bases, on the lattice of closed classes, on their modification and some other related questions.

By the problem of expressibility is meant the indication of all pairs $\overline{M}$, M' such that $\overline{M}$ is expressible in terms of M'; by the problem of completeness is meant the indication of all complete subsets; the problem of bases is the description of all bases if they exist; the problem of the lattice is the construction of the lattice of all closed classes and the determination of its properties.

The knowledge of the lattice $\Sigma(\mathcal{M})$ gives the solution of the problems of expressibility and completeness. Thus, expressibility of $\overline{M}$ in terms of M' means verification of the condition $I_\Omega(\overline{M}) \subseteq I_\Omega(M')$. For the solution of the completeness problem the following scheme is used. A system $\Sigma' \subseteq \Sigma(\mathcal{M})$ is called *criterial* (*c-system*) if for any set $S \subseteq M$, S is complete if and only if S is not a subset of any set M', $M' \in \Sigma'$. It is obvious that if $\Sigma(\mathcal{M}) \setminus \{M\} \neq \emptyset$, which is later assumed, then $\Sigma(\mathcal{M}) \setminus \{M\} \neq \emptyset$ is a c-system. It is not difficult to see that dual atoms of the lattice $\Sigma(\mathcal{M})$, which are also called *precomplete classes*, belong to any c-system. Let $\Sigma_\pi(\mathcal{M})$ be the set of all precomplete classes and $\Sigma_{\overline{\pi}}(\mathcal{M})$ be the set of all classes from $\Sigma(\mathcal{M})$ which are not a subsets of any precomplete class from $\Sigma_\pi(\mathcal{M})$. It is not difficult to verify that the following assertion holds.

2.1 Proposition *The set $\Sigma_\pi(\mathcal{M}) \cup \Sigma_{\overline{\pi}}(\mathcal{M})$ forms a c-system in u.a. $\mathcal{M}$.*

Of special interest is the situation where $\Sigma_{\overline{\pi}}(\mathcal{M})$ is the empty set, since in this case the system $\Sigma_\pi(\mathcal{M})$ forms a c-system, which means that the completeness problem is reduced to the description of all precomplete classes. Let us mention an important case of this kind. A u.a. $\mathcal{M}$ is called *finitely-generated* if there exists a finite subset $M' \subseteq M$ which is complete. The following assertion is known [12].

2.2 Proposition *If a u.a. $\mathcal{M}$ is finitely-generated, then $\Sigma_\pi(\mathcal{M})$ forms a c-system.*

We note that in the general case the converse is false. Let us now consider the case where the set M consists of functions. It is the main case for us. In this case the u.a. $\mathcal{M}$ is called a *functional system* (f.s.).

Let E be a set, and let a function f have the form $f : E^n \to E$, where $n \in N$. Let $U = \{u_1, u_2, \dots\}$ be the alphabet of variables with values in E, $i \in N$. To write a function f, we use the expression $f = (u_{i_1}, u_{i_2}, \dots, u_{i_n})$. Denote the class of all such functions by P_E. In order to avoid complex indices in the variables u_i we use metasymbols x, y, z for them, possibly with indices.

Following Maltsev [20], we introduce in P_E unary operations $\eta, \tau, \triangle, \nabla$, which are defined in the following way:

$$\begin{aligned}
(\eta f)(x_1, x_2, \dots, x_n) &= f(x_2, x_3, \dots, x_n, x_1);\\
(\tau f)(x_1, x_2, \dots, x_n) &= f(x_2, x_1, x_3, \dots, x_n);\\
(\Delta f)(x_1, x_2, \dots, x_{n-1}) &= f(x_1, x_1, x_2, \dots, x_{n-1}) \quad \text{if} \quad n > 1;\\
(\eta f) = (\tau f) = (\Delta f) &= f \quad \text{if} \quad n = 1;\\
(\nabla f)(x_1, x_2, \dots, x_{n+1}) &= f(x_2, x_3, \dots, x_{n+1}).
\end{aligned}$$

The form of these operations refines the operations from [17].

Let us introduce in P_E a binary operation $*$ in the following way. For the functions $f(x_1, x_2, \dots, x_n)$ and $g(x_{n+1}, x_{n+2}, \dots, x_{n+m})$ we assume that

$$(f * g)(x_2, x_3, \dots, x_n, x_{n+1}, x_{n+2}, \dots, x_{n+m}) = (f(g(x_{n+1}, x_{n+2}, \dots, x_{n+m}), x_2, \dots, x_n).$$

The described operations are called, respectively, *shift*, *transposition*, *identification*, *extension*, and *substitution*, and, in total, the *operations of superposition*. The set of these operations is indicated by Ω_s. Let $M \subseteq P_E$ and $I_{\Omega_s}(M) = M$, then f.s. $\mathcal{M} = (M, \Omega_s)$ is called the *Post iterative f.s.*

3 Functions of l-valued logic

The functions from P_E are called *functions of l-valued logic* if

$$E = E_l := \{0, 1, 2, \ldots, l-1\}, \quad l \geq 2.$$

In this case the symbol P_l is used instead of P_E. The f.s. $\mathcal{P}_l = (P_l, \Omega_s)$ is considered to be one of the main models of the Post iterative f.s. (for brevity's sake, P.f.s.), the study of which served as the basis for the formulation of the problems and methods of f.s. theory. If $\mathcal{M}_l = (M, \Omega_s)$ and $M \subseteq P_l$ then $\mathcal{M}$ is called *P.f.s. of kind l*. Let us briefly sum up the main results in the study of $\mathcal{P}_l$ which will be important for the consideration of the f.s. of automaton functions.

Post [24] gave a complete solution of the indicated problems on completeness, expressibility, bases and closed classes lattice, for P_l. Let us describe this lattice preserving his notation. Consider the set $\mathcal{Q}$ of the classes

$$C_i,\ A_i,\ D_j,\ L_r,\ O_s,\ S_t,\ P_t,\ F_\nu^m,\ F_\nu^\infty$$

where $i = 1, 2, 3, 4$; $j = 1, 2, 3$; $r = 1, 2, 3, 4, 5$; $s = 1, 2, \ldots, 9$; $t = 1, 3, 5, 6$; $\nu = 1, 2, \ldots, 8$; $m = 1, 2, \ldots$.

The functions from P_2 are called *Boolean functions* (B.f.). Post denotes the class P_2 by C_1. The class C_2 contains all B.f. f such that $f(1, 1, \ldots, 1) = 1$; C_3 is the class of all B.f. such that $f(0, 0, \ldots, 0) = 0$; $C_4 = C_2 \cup C_3$. It is said that a B.f. f is monotonic if the inequalities $a_i \leq b_i$ for all $i = 1, 2, \ldots, n$ imply that $f(a_1, \ldots, a_n) \leq f(b_1, \ldots, b_n)$. The class A_1 consists of all monotonic B.f.; $A_2 = C_2 \cap A_1$; $A_3 = C_3 \cap A_1$; $A_4 = A_2 \cap A_3$. The class D_3 consists of all B.f. f such that $f(x_1, \ldots, x_n) = \overline{f}(\overline{x}_1, \ldots, \overline{x}_n)$ where B.f. $\overline{x}$ is called the negation and $\overline{0} = 1, \overline{1} = 0; D_1 = C_4 \cup D_3; D_2 = A_1 \cup D_3$. The class L_1 consists of all B.f. $f(x_1, \ldots, x_n) = x_1 + x_2 + \cdots + x_n + \alpha \pmod 2$, $\alpha \in E_2$; $L_2 = C_2 \cap L_1$; $L_3 = C_3 \cap L_1$; $L_4 = L_2 \cap L_3$; $L_5 = D_3 \cap L_1$. The class O_9 consists of all B.f. depending essentially on no more than one variable; $O_8 = A_1 \cap O_9$; $O_4 = D_3 \cap O_9$; $O_5 = C_2 \cap O_9$; $O_6 = C_3 \cap O_9$; $O_1 = O_5 \cap O_6$; $O_7 = \{0, 1\}$; $O_2 = O_5 \cap O_7$; $O_3 = O_6 \cap O_7$. The class S_6 consists of all B.f. of the form $x_1 \vee x_2 \vee \cdots \vee x_n$ and of constants; $S_3 = C_2 \cap S_6$; $S_5 = C_3 \cap S_6$; $S_1 = S_3 \cap S_5$. The class P_6 consists of all B.f. of the formal $x_1 \wedge x_2 \wedge \cdots \wedge x_n$ and constants; $P_5 = C_2 \cap P_6$; $P_3 = C_3 \cap P_6$; $P_1 = P_3 \cap P_5$. It is said that a B.f. satisfy the condition a_μ, $\mu \in N_1$ if all μ collections at which it equals 0 have a common coordinate 0. The property A_μ is determined similarly with the change of 0 to 1. The class F_4^μ consists of all B.f. with the property a^μ; $F_1^\mu = C_4 \cap F_4^\mu$; $F_3^\mu = A_1 \cap F_4^\mu$; $F_2^\mu = F_1^\mu \cap F_3^\mu$. The class F_8^μ consists of all B.f. with the property A^μ; $F_5^\mu = C_4 \cap F_8^\mu$; $F_7^\mu = A_3 \cap F_8^\mu$; $F_6^\mu = F_5^\mu \cap F_7^\mu$. A B.f. satisfies the condition a^∞ if all the collections at which it equals 0 have a common coordinate 0. The property A^∞ is introduced by analogy with the change of 0 to 1.

The class F_4^∞ consists of all B.f. with the property a^∞; $F_1^\infty = C_4 \cap F_4^\infty$; $F_3^\infty = A_1 \cap F_4^\infty$; $F_2^\infty = F_1^\infty \cap F_3^\infty$. The class F_8^∞ consists of all B.f. with the property A^∞; $F_5^\infty = C_4 \cap F_8^\infty$; $F_7^\infty = A_3 \cap F_8^\infty$; $F_6^\infty = F_5^\infty \cap F_7^\infty$.

3.1 Theorem [24] *For P.f.s. $\mathcal{P}_2$ the following assertions are true:*

(1) *The set of all closed classes in $\mathcal{P}_2$ is countable and coincides with the set $\mathcal{Q}$.*

(2) *The classes from $\mathcal{Q}$ form the inclusion lattice given in Fig. 1.*

(3) *Every closed class in $\mathcal{P}_2$ has a basis, and its cardinality is always no more than 4.*

(4) *The problems of completeness and expressibility for P.f.s. of kind 2 applied to finite sets of B.f. are algorithmically decidable.*

The properties of P.f.s. of kind l, $l > 2$, turned out to be more complex, which follows from the assertions given below.

Denote by $P_l^{(n)}$ the set of all functions from P_l depending on no more than n variables $u_1, u_2, \ldots, u_n$.

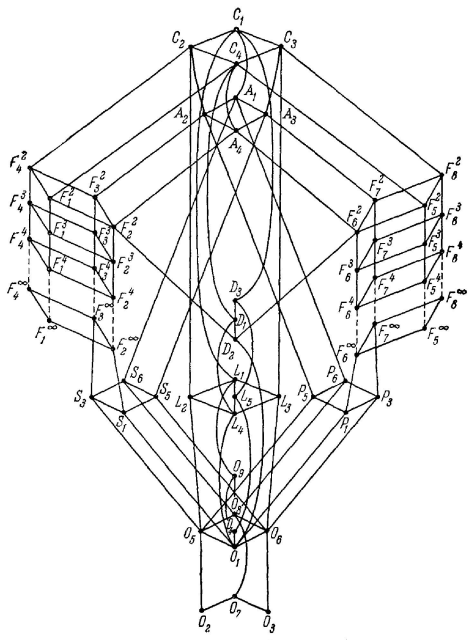


Figure 1

It is clear that the number of functions in $P_l^{(n)}$ is equal to l^{l^n}. Let S_l^n be the set of all functions from $P_l^{(n)}$ each of which is equal to u_i for some $i, i = 1, 2, \ldots, n$. If $M \subseteq P_i$ and M is finite, then $\alpha(M)$ indicates the greatest number of variables of the functions from M.

For a finitely-generated P.f.s. $\mathcal{M}$ of kind l let $\alpha(\mathcal{M})$ be the smallest number α' such that $I_{\Omega_s}(M') = M$ and $\alpha(M') = \alpha'$ for some $M' \subseteq M$. A non-empty set $M' \subseteq P_l^{\alpha(\mathcal{M})} \cap M$ is called an *R-set* in $\mathcal{M}$ if $I_{\Omega_s}(M') \cap P_l^{\alpha(\mathcal{M})} = M'$ and $M' \neq P_l^{\alpha(\mathcal{M})}$, and is denoted by $R^{\alpha(\mathcal{M})}$. Let $\mathcal{R}^{\alpha(\mathcal{M})} = R^{\alpha(\mathcal{M})} \cup S_l^{\alpha(\mathcal{M})}$. We say that a function $f(x_1, x_2, \ldots, x_n)$ from P_i preserves $\mathcal{R}^{\alpha(\mathcal{M})}$ if for any collection of functions $g_1, g_2, \ldots, g_n$ from $\mathcal{R}^{\alpha(\mathcal{M})}$

$$f(g_1, g_2, \ldots, g_n) \in R^{\alpha(\mathcal{M})}.$$

The class of all functions from M preserving $\mathcal{R}^{\alpha(\mathcal{M})}$ is denoted by $U(\mathcal{R}^{\alpha(\mathcal{M})})$. An R-set $\mathcal{R}^{\alpha(\mathcal{M})}$ is called *maximal* if there is no R-set $\mathcal{R}_1^{\alpha(\mathcal{M})}$ such that $U(\mathcal{R}_1^{\alpha(\mathcal{M})}) \subset U(\mathcal{R}^{\alpha(\mathcal{M})})$. Let $\mathbf{R}(\mathcal{M})$ be the set of all maximal R-sets, and $U(\mathbf{R}(\mathcal{M}))$ be the set of all classes of preservation of elements from $\mathbf{R}(\mathcal{M})$. A P.f.s. $\mathcal{M}$ is called trivial if $M = I_{\Omega_s}(S_k^1)$ or $M = I_{\Omega_s}(\{c(x)\})$, where $c(x) = c$, $c \in E_l$. The cardinality of a set A is denoted by $|A|$.

3.2 Theorem [18] *If a P.f.s. $\mathcal{M} = (M, \Omega_s)$ of kind l is non-trivial and finitely-generated, then:*

(1) $U(\mathbf{R}(\mathcal{M})) = \Sigma_\pi(\mathcal{M})$*;*

(2) $|U(\mathbf{R}(\mathcal{M}))| \leq 2^{|P_l^{\alpha(\mathcal{M})}|}$*;*

(3) $\mathbf{R}(\mathcal{M})$ *is constructed effectively.*

This theorem is a development of an assertion from [29].

3.3 Corollary *The problem of completeness for finitely generated P.f.s. of kind l is algorithmically decidable for any l.*

3.4 Theorem [18] *The problem of expressibility for finite sets of finitely-generated P.f.s. of kind l is algorithmically decidable for any l.*

3.5 Theorem [31] *For every $l \geq 3$ there exists P.f.s. of kind l such that:*

(1) *the P.f.s. have a countable basis;*

(2) *the P.f.s. have a basis of an assigned finite power;*

(3) *the P.f.s. have no basis.*

3.6 Corollary [31] *For every $l \geq 3$ the lattice of closed classes in the P.f.s. $\mathcal{P}_l$ is a continuum set.*

As was established in [29], the P.f.s. $\mathcal{P}_l$ is finitely generated, therefore Theorems 2 and 3 are valid, and for the set $U(\mathbf{R}(\mathcal{P}_l))$ its explicit description is found. It is more convenient for us to present it in a predicate form.

Let $\rho(y_1, y_2, \ldots, y_h)$ be an h-ary predicate, the arguments of which take values from E_l. If $\rho(a_1, a_2, \ldots, a_h) = a$, then for $a = 1$ the collection is called true, and for $a = 0$ it is false. The set of all true collections is denoted by ρ_1, and the set of false ones by ρ_0. We say that a function $f(x_1, x_2, \ldots, x_n)$ from P_l preserves ρ, if the truth of every element

$\rho(a_1^1, a_2^1, \ldots, a_h^1), \rho(a_1^2, a_2^2, \ldots, a_h^2), \ldots, \rho(a_1^n, {}_2^n, \ldots, {}_h^n)$ implies the truth of $\rho(f(a_1^1, a_2^1, \ldots, a_h^1), f(a_1^2, a_2^2, \ldots, a_h^2), \ldots, f(a_1^n, {}_2^n, \ldots, {}_h^n))$.

A set M of functions from P_l preserves ρ if every function from M preserves ρ. The class of all functions preserving ρ is denoted by $U(\rho)$. Let us describe six special families of predicates.

Family P Let the truth region of a predicate $\rho(y_1, y_2)$ be the diagram of a permutation $\sigma(x)$ from P_l, which is decomposed into the product of cycles of equal prime length p, $p \geq 2$. The family P consists exactly of all such predicates.

Family E Let $E_l = E^1 \cup E^2 \cup \cdots \cup E^t$, where $1 \leq t \leq l$, $E^i \cap E^j = \emptyset$ for $i \neq j$. We consider the predicate $\rho(y_1, y_2)$ that is true precisely on those collections for which $a, b \in E^i$ for some i. The family E consists exactly of all such predicates for the decompositions indicated.

Family $\mathcal{L}$ Let $l = p^m$, where p is prime, and let $G = \langle E_l, + \rangle$ be an Abelian group whose every non-zero element has the order p, i.e., G is an elementary p-group.

For $p = 2$ we consider the predicate $\rho_G(y_1, y_2, y_3, y_4)$ which is true exactly on those collections for which $y_1 + y_2 = y_3 + y_4$. In this case $\mathcal{L}$ consists exactly of those predicates ρ_G which correspond to all indicated elementary 2-groups G.

For $p \neq 2$ we consider the predicate $\rho_G(y_1, y_2, y_3)$ which is true exactly on those collections for which $y_3 = 2^{-1}(y_1 + y_2)$, where 2^{-1} denotes the number from E_p for which $2 \cdot 2^{-1} \equiv 1 \pmod{p}$. In this case $\mathcal{L}$ consists exactly of those predicates ρ_G which correspond to all indicated elementary p-groups.

Family B Let h, m be natural numbers such that $h \geq 3$, $m \geq 1$, $h^m \leq l$, and let $\varphi(x)$ be an arbitrary mapping of E_l into E_{h^m}. Let $a \in E_{h^m}$; denote by $[a]_l$ the l-th coefficient in the decomposition

$$a = \sum_{l=0}^{m-1} [a]_l h^l, \quad [a]_l \in E_{h^m}.$$

We consider the predicate $\rho(y_1, y_2, \ldots, y_h)$ which is true exactly on those collections for which the collection $([\varphi(a_1)]_l, [\varphi(a_2)]_l, \ldots, [\varphi(a_h)]_l)$ is nondifferent-valued for any l from E_m, i.e., ρ is determined by the triple (h, m, φ); in addition, if $\varphi(x) = x$, then ρ is called *elementary*. The family B consists exactly of the predicates ρ determined by all the triples (h, m, φ) indicated.

Family Z A predicate $\rho(y_1, y_2, \ldots, y_h)$, $h \geq 1$, is reflexive if its truth follows from the nondifferent-valuedness of the collection of values of the variables, and it is symmetrical if for any substitution $t(u)$ of the numbers $1, 2, \ldots, h$,

$$\rho(y_1, y_2, \ldots, y_h) = \rho(y_{t(1)}, y_{t(2)}, \ldots, y_{t(h)}).$$

The non-empty set of all elements c from E_l such that for all values of the variables $\rho(y_1, y_2, \ldots, y_{h-1}, c) = 1$ is called the *centre* of the symmetrical predicate ρ. A predicate ρ is called *central* if it is symmetrical, reflexive and has a centre C such that $C \subset E_l$. The family Z consists exactly of all the central predicates $\rho(y_1, y_2, \ldots, y_h)$ such that $1 \leq h \leq l - 1$.

Family M A partial order on E_l with one the greatest and one the smallest elements can be assigned by a binary predicate $\rho(y_1, y_2)$. The family M consists exactly of all such predicates.

Let $W = P \cup E \cup L \cup B \cup Z \cup M$ and let $U(W)$ consist of all classes of functions preserving predicates from W.

3.7 Theorem *The following relations are true:*

(1) $U(W) = \Sigma(\mathcal{P}_i)$;

(2) $|U(W)| \sim l \cdot (l - 2 \cdot [\frac{l}{2}] + 1) \cdot 2^{\binom{l-1}{[(l-1)/2]}}$.

Part (1) of this assertion was gradually established by the authors of the papers [10,11,21, 24,26,30,32,33], and the concluding study was carried out in [26]. Part (2) was established in [7].

The sharp difference in the properties of $\mathcal{P}$ for $l = 2$ and $l > 2$ led to the study of different variations of the main problems for P.f.s. such as testing completeness of a system of functions with assigned properties, like, for example, the Slupecki system which contains all one-place functions, or examining the structure of fragments of lattice of closed classes, etc. Besides that, the objects of scientific investigation were generalizations of $\mathcal{P}_l$ in the form of P.f.s. of non-homogeneous functions, i.e., of functions depending on groups of variables whose domains of definition are different [18], and also functions whose variables, like the functions themselves, take a countable number of values.

The study of the Cartesian products of such generalizations and of other cases has begun (see [25]). We do not discuss these directions here. We consider only the generalizations of functions from P_k which appear if we take into account the time which is necessary for their calculations.

4 Functions with delays

Let $f(x_1, x_2, \ldots, x_n) \in P_l$ and $t \in N_0$. A pair $(f(x_1, x_2, \ldots, x_n), t)$ is called a *function* f *with delay* t and the set of all such pairs is denoted by $\tilde{P}_l$. We extend the operations η, τ, Δ and $\bigtriangledown$ to $\tilde{P}_l$ assuming that if μ is any of them, then $\mu((f,t)) = (\mu(f), t)$. We introduce one more operation $*_s$ that is called a *synchronous substitution*, assuming that for the pairs $(f,t), (f_1,t'), (f_2,t'), \ldots, (f_n,t')$ in which the sets of variables of the functions $f_1, f_2, \ldots, f_n$ are mutually disjoint the relation

$$(f,t) *_s ((f_1,t'), (f_2,t'), \ldots, (f_n,t')) = (f(f_1, f_2, \ldots, f_n), t + t')$$

holds.

The set of operations $\eta, \tau, \Delta, \bigtriangledown, *s$ is denoted by Ω_{ss} and is called the *operations of synchronous superposition*. Let $M \subseteq \tilde{P}_l$ and $J_{\Omega_{ss}}(M) = M$, then the f.s. $\mathcal{M} = (M, \Omega_{ss})$ is called an *iterative P.f.s. of functions with delays of kind* l (*P.f.s.f.d.*).

Let us briefly sum up the main results in the study of these f.s. [16,18].

4.1 Theorem *For a finitely-generated P.f.s.f.d.* $\mathcal{M}$ *of kind* l *the set* $\Sigma_\pi(\mathcal{M})$ *is finite and is effectively constructed for any* l.

4.2 Theorem *For a finitely-generated P.f.s.f.d. of kind l the problems of completeness and expressibility are algorithmically decidable for any l.*

4.3 Theorem *For every l from N_1 there exists a P.f.s.f.d. of kind l such that:*

(1) *there is a countable basis;*

(2) *there is a finite basis of an assigned power;*

(3) *there is no basis.*

4.4 Corollary *The lattice of closed classes in $\tilde{\mathcal{P}}_l$ has cardinality of continuum for any l.*

An example of a finitely-generated P.f.s.f.d. is $\tilde{\mathcal{P}}_l = (\tilde{\mathcal{P}}_l, \Omega_{ss})$.

For $\tilde{\mathcal{P}}_l$ Theorem 4.1 can be refined as follows. Denote by $M^{(1)}$ the set of all functions f such that $(f,t) \in M$ for some t.

4.5 Theorem *A set $M \subseteq \tilde{P}_l$ is complete in $\tilde{\mathcal{P}}_l$, if and only if $J_{\Omega_s}(M^{(1)}) = \tilde{P}_l$, and $M \supseteq \{(f,1)\}$, where $f \notin E_l$.*

Let $\mathcal{M} = (M, \Omega_{ss})$ and $M' \subset M$. We say that M' is *weakly complete* in $\mathcal{M}$ if for any function f from $M^{(1)}$ there is t such that $(f,t) \in J_{\Omega_{ss}}(M')$. The set M' is called *weakly precomplete* if it is not weakly complete but turns into such by adding any pair from $M \setminus M'$. The class of all such sets is denoted by $\Sigma_{W\pi}(\mathcal{M})$. A class Σ is called *weakly criterial* if the set M' is weakly complete if and only if $M' \nsubseteq T$ for any T from Σ. It is obvious that for any weakly criterial system K the inclusion $K \supseteq \Sigma(\mathcal{M})$ is true.

4.6 Theorem *For a finitely-generated P.f.s.f.d. $\mathcal{M} = (M, \Omega_{ss})$ of kind l the following assertions are true.*

(1) *The set $\Sigma_{W\pi}(\mathcal{M})$ is finite or countable.*

(2) *The set $\Sigma_{W\pi}(\mathcal{M})$ is weakly criterial.*

(3) *The set $\Sigma_{W\pi}(\mathcal{M})$ is constructed effectively.*

4.7 Theorem *For a finitely-generated P.f.s.f.d. of type l the problem of weak completeness is algorithmically decidable for any l.*

An explicit description of the set $\Sigma_{W\pi}(\mathcal{M})$ is obtained only for $l = 2, 3$ [16, 25]. Let us give here the description of the case $l = 2$.

Let $M \subseteq P_2$ and $M \in \{C_2, C_3, D_3, A_1, L_1\}$; we denote by $\widetilde{M}$ the set of all pairs (f,t) such that $f \in M$ and $t \in N_0$.

A function $f(x_1, x_2, \ldots, x_n)$ from P_2 is called an α-, β-, γ- or δ-function if $f(x, x, \ldots, x)$ is equal to $x, 1, 0$ or $\bar{x}$ respectively. Let A, B, Γ, D be the classes of all α-, β-, γ- or δ-functions respectively.

Denote by $\widetilde{C}$ the set of all pairs of the form

$$(f, t+1),\ (\varphi, t+1),\ (\psi, 0),$$

where $f \in B$, $\varphi \in \Gamma$, $\psi \in A$, and $t \in N_0$.

Let $i \in \{0,1\}$; we denote by $\widetilde{E}_i$ the set of all pairs of the form

$$(f,0),\ (\bar{\imath},t+1),\ (i,t),$$

where $f \in C_{i+2}$ and $t \in N_0$.

A function f is called *even* if $f(x_1,x_2,\dots,x_n) = f(\bar{x_1},\bar{x_2},\dots,\bar{x_n})$.

Let Y be the set of all even functions. Let $\widetilde{H}$ be the set of all pairs of the form

$$(f,0),\ (\varphi,t+1),$$

where $\varphi \in Y$, $f \in D_3$, and $t \in N_0$.

For every r from N_0 we denote by $\widetilde{Z_r}$ the set of pairs of the form

$$(f,(2t+1)2^r),\ (\varphi,(2t+1)2^s),\ (\psi,0),$$

where $f \in D$, $\varphi \in A$, $\psi \in A$, $t \in N_0$, and $s \in N_0 \setminus \{0,1,2,\dots,r\}$.

For every r from N_0 we denote by $\widetilde{W}_r$ the set of all pairs of the form

$$(f,(2t+1)2^r),\ (0,t),\ (1,t),\ (\varphi,(2t+1)2^s),\ (\psi,0),$$

where $\bar{f} \in M$, $\varphi \in M$, $\psi \in M$, $t \in N_0$, and $s \in N_0 \setminus \{0,1,2,\dots,r\}$.

Let $\widetilde{W} = \{\widetilde{C_2},\widetilde{C_3},\widetilde{D_3},\widetilde{A_1},\widetilde{L_1},\widetilde{C},\widetilde{E_0},\widetilde{E_1},\widetilde{H},\widetilde{Z_0},\widetilde{Z_1},\dots,\widetilde{W_0},\widetilde{W_1},\dots\}$.

4.8 Theorem [16] *The equality* $\Sigma_{W\pi}(\widetilde{\mathcal{P}_2}) = \widetilde{W}$ *holds.*

We note that an explicit description of $\Sigma_{W\pi}(\widetilde{\mathcal{P}_1}) = \widetilde{W}$ for any l exists only in the form of separate families of weakly precomplete classes [2,18].

From the point of view of content the modifications of the problem of completeness for P.f.s.f.d. considered in [18] are of interest.

5 Automata functions

The extension of the functions of l-valued logic to functions with delays (considered in [24]) is intermediate in the transition to the class of automaton functions, whose properties logically look essentially more complex than those of functions with delays. In order to introduce the notion of automaton function we shall need auxiliary notations and definitions.

Let C be a finite or countable set which is called an *alphabet*. A sequence of letters from C is called a *word* if it is finite, and a *superword* if it is infinite. The class of all such words is denoted by C^*, and of all superwords, by C^ω. Let $\overline{C} = C^* \cup C^\omega$ and $\xi \in \overline{C}$. The word formed by the first r letters from ξ is called a *prefix for* ξ and is labeled by $\xi]_r$. Let A and B be alphabets and $f : A^* \to B^*$. If $\xi = c(1)c(2)\dots c(r)$, then r is called the *length of the word* ξ and is denoted by $\|\xi\|$. Let A and B be alphabets and $f : A^* \to B^*$. The function f is called *determined* (*d. function*) if for any ξ from A^* the equality $\|f(\xi)\| = \|\xi\|$ is valid, and for any ξ_1 and ξ_2 from A^* and any i such that $1 \le i \le \min(|\xi_1|,|\xi_2|)$ if $\xi_1]_i = \xi_2]_i$, then $f(\xi_1)]_i = f(\xi_2)]_i$. It is known that a d. function f can be recurrently represented by the so-called canonical equations of the form

$$\begin{aligned} q(1) &= q_0, \\ q(t+1) &= \varphi(q(t),a(t)), \\ b(t) &= \psi(q(t),a(t)), \quad t = 1,2,\dots, \end{aligned} \tag{5.1}$$

where the parameter q is called a *state of* f and takes values in the alphabet Q.

This recurrence is determined in the following way. If $\alpha \in A^*$, $\beta \in B^*$, $\kappa \in Q$, and

$$\alpha = a(1)a(2)\dots a(r), \qquad \beta = b(1)b(2)\dots b(r), \qquad \kappa = q(1)q(2)\dots q(r),$$

then for $f(\alpha) = \beta$ the word β is inductively calculated by α from the following way:

(1) $b(1) = \psi(q(1), a(1))$, where $q(1) = q_0$;

(2) if $q(t)$ is calculated for $1 \le t \le r-1$, then $q(t+1) = \varphi(q(t), a(t))$ and $b(t) = \psi(q(t), a(t))$.

It is often assumed that the alphabets A and B are Cartesian degrees of E_l, i.e., $A = (E_l)^n$ and $B = (E_l)^m$, where $n, m, p \in N$. In that case it is convenient to turn from a one-place d. function $f(x)$ of the form $f : ((E_l)^n)^* \to ((E_l)^m)^*$ to an n-ary d. function $f(x_1, x_2, \dots, x_n)$ of the form $f' : ((E_l)^*)^n \to ((E_l)^*)^m$ in the following way. Let $\zeta^n \in (E_k^*)$ and $f(\zeta^n) = \zeta'^m$, where

$$\zeta^n = c^n(1)c^n(2)\dots c^n(r), \qquad \zeta'^m = c'^m(1)c'^m(2)\dots c'^m(r),$$

in addition,

$$c^n(t) = (e_1(t), e_2(t), \dots, e_n(t)), \qquad c'^m(t) = (e'_1(t), e'_2(t), \dots, e'_m(t)),$$

where $1 \le t \le r$. Let

$$\zeta_i = e_i(1)e_i(2)\dots e_i(r) \quad \text{and} \quad \zeta'_j = e'_j(1)e'_j(2)\dots e'_j(r),$$

where $1 \le i \le n$ and $1 \le j \le m$.

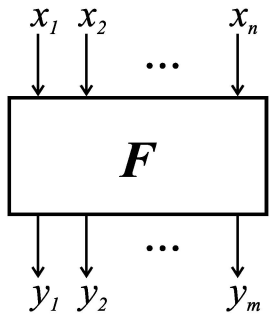


Figure 2

Now we assume that

$$f'(\zeta_1, \zeta_2, \dots, \zeta_n) = (\zeta'_1, \zeta'_2, \dots, \zeta'_n).$$

The function f' is actually obtained from f only at the cost of presenting the matrices formed by vector-letters (rows) of words ζ^n and, respectively, ζ'^m as matrices formed by the columns.

Canonical equations (5.2) for f' are obtained from (5.1) by replacing all the parameters there with the corresponding vector values, i.e.,

$$\begin{aligned} q(1) &= q_0, \\ q(t+1) &= \varphi(q(t), e_1(t), \dots, e_n(t)), \\ b_j(t) &= \psi_j(q(t), e_1(t), \dots, e_n(t)), \quad t = 1, 2, \dots, \quad j = 1, 2, \dots, m. \end{aligned} \tag{5.2}$$

The function f' is considered to be an interpretation of f and is called an *automaton function* (*a. function*). The parameters n and m are called, respectively, the *locality* and the *dimensionality* of the a. function, and the cardinality of the set of values of the parameter q is called the *number of states*. The significant interpretation of the a. function

$$f'(x_1, x_2, \dots, x_n) = (y_1, y_2, \dots, y_m)$$

is the functioning of the technical device F shown in Fig. 2. Here the input arrows are labelled by the letters $x_i, i = 1, \dots, n$, and the output ones, by the letters y_j, $j = 1, 2, \dots, m$. It is assumed that F functions at the discrete moments of time $t = 1, 2, \dots$. At these moments every input x_i and output y_i can take values in E_l; the device itself may be in states that are coded by values from Q, these states are also called *memory* for F. The values of outputs and the state at the moment $t + 1$ are determined by the collection of values of inputs and by the state of the device F at the moment t according to rules (5.2).

Denote by $P_{a,l}^{n,m}$ the class of all a. functions with the given parameters n and m from N. Let

$$P_{a,l} = \bigcup_{n,m \geq 1} P_{a,l}^{n,m}.$$

We extend the operations $\eta, \tau, \Delta, \nabla$ to $P_{a,l}$, and also introduce some other operations.

Let $f'(x_1, x_2, \dots, x_n) = (y_1, \dots, y_j, \dots, y_m)$, then

$$(\pi_j f')(x_1, x_2, \dots, x_n) = f'_j(x_1, x_2, \dots, x_n),$$

where $f'_j(x_1, x_2, \dots, x_n)$ coincides with the value y_j in $f'(x_1, x_2, \dots, x_n)$.

Let $f''(x_{n+1}, x_{n+2}, \dots, x_{n+v}) = (y_{m+1}, \dots, y_{m+w})$, then

$$\begin{aligned} &(f' \sigma f'')(x_1, x_2, \dots, x_n, x_{n+1}, x_{n+2}, \dots, x_{n+v}) = (f'(x_1, x_2, \dots, x_n), \\ &f''(x_{n+1}, x_{n+2}, \dots, x_{n+v})) = (y_1, y_2, \dots, y_m, y_{m+1}, y_{m+2}, \dots, y_{m+w}). \end{aligned}$$

Let u be such that $m + 1 \leq u \leq m + u$, then

$$(f' *_u f'')(x_2, x_3, \dots, x_n, x_{n+1}, x_{n+2}, \dots, x_{n+v}) = (z_1, z_2, \dots, z_{m+w-1}),$$

where $z_j = f'_j * f''_u$ for $j = 1, 2, \dots, m$ and $z_{j'} = f''_{j'}(x_{n+1}, x_{n+2}, \dots, x_{n+v})$ for $j' = m + 2, m + 3, \dots, m + t$.

The operations π and σ are called, respectively, *projection* and *union*, and the operation of the form $*_u$ is the extension of the operation $*_s$ from one-dimensional functions to vector functions and, as before, is called *substitution*.

In total the operations $\eta, \tau, \Delta, \nabla, \pi, \sigma$ and $*_u$ are called *operations of extended superposition* and is denoted by Σ_{es}.

Let us introduce one more operation on a. functions, which is called *feedback* and is denoted by **F**.

It is said that an a. function f' is of type $i-j$ if for f' there is a system of kind (5.2) such that the function $\psi_j(q, e_1, e_2, \dots, e_n)$ fictitiously depends on e_i. Let f' be such an a. function; we consider a function of the form

$$(\mathbf{F}_j^i f')(x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = (y_1, y_2, \dots, y_{j-1}, y_{j+1}, \dots, y_m),$$

which is determined in the following way. Let words of the form

$$\xi_l = e_l(1)e_l(2)\dots e_l(r),$$

where $l = 1, 2, \dots, i-1, i+1, \dots, n$ be given. Then with the help of (5.2), using the collection $(e_1(1), e_2(1), \dots, e_{i-1}(1), e_{i+1}(1), \dots, e_n(1))$, it is possible to calculate the value $b_j(1)$. Now in (5.2) we substitute $b_j(1)$ for $e_i(1)$; after that it is possible to calculate the collections $(q(2))$ and $(b_1(1), b_2(1), \dots, b_m(1))$. Then it is also possible to calculate the value $b_j(2)$, using the collection

$$(e_1(2), e_2(2), \dots, e_{i-1}(2), e_{i+1}(2), \dots, e_n(2)).$$

Again, the substitution of $e_i(2)$ for $b_j(2)$ in (5.2) makes it possible to calculate the collections $(q(3))$ and $(b_1(2), b_2(2), \dots, b_m(2))$ and so on.

Now we assume that

$$(F_j^i f')(\zeta_1, \zeta_2, \dots, \zeta_{i-1}, \zeta_{i+1}, \dots, \zeta_n) = (\zeta'_1, \zeta'_2, \dots, \zeta'_{i-1}, \zeta'_{i+2}, \dots, \zeta'_m),$$

where $\zeta'_{l'} = b_{l'}(1)b_{l'}(2)\dots b_{l'}(r)$ and $l' = 1, 2, \dots, j-1, j+1, \dots, m$. Suppose that $\Omega_{\mathrm{es},F} = \Omega_{\mathrm{es}} \cup \{F\}$. The class of operations $\Omega_{\mathrm{es},F}$ is called *composition*. It is said that an a. function f' from $P_{a,l}$ is a finite automata function (*f.a. function*) if the alphabet Q in system (5.2) determining this function is finite. Denote by $P_{a,l,k}^{n,m}$ the class of all f.a. functions with parameters n and m. Suppose that

$$P_{a,l,k} = \bigcup_{n,m \geq 1} P_{a,l,k}^{n,m}.$$

It is said that an a. function is true (*t.a. function*) if the alphabet Q in system (5.2) determining this function is a one-element set. We denote by $P_{a,l,t}^{n,m}$ and $P_{a,l,k,t}^{n,m}$ the classes of all true a. functions and of true f.a. functions (*t.f.a. functions*), respectively, with parameters n and m. Suppose that

$$P_{a,l,t} = \bigcup_{n,m \geq 1} P_{a,l,t}^{n,m}, \qquad P_{a,l,k,t} = \bigcup_{n,m \geq 1} P_{a,l,k,t}^{n,m}.$$

From the content point of view, true a. functions are interpreted as the functioning of a device F without memory, on the one hand, and, on the other hand, they can be considered to be realizations of the functions from P_l with regard to time t that runs over the values $1, 2, \dots$. At each moment of time the dependence of the value of the function on the values of the variables is the same.

Thus, a P.f.s. (P_l, Ω_s) actually leads to $\mathcal{P}_{a,l,t} = (P_{a,l,t}, \Omega_{\mathrm{es}})$, if we extend the object P_l to the set of vector functions of l-valued logic and, respectively, extend the operations to Ω_{es}.

We also note that a function with delay is interpreted as the functioning of such device F with n inputs and one output whose value $b(t)$ for some τ from N_0 and $f(x_1, x_2, \ldots, x_n)$ from P_l is determined at any moment $t \geq \tau + 1$ by the relation

$$b(t) = f(a_1(t-\tau), a_2(t-\tau), \ldots, a_n(t-\tau))$$

that, evidently, can be described by a system of form (5.2).

Let $M \subseteq P_{a,l}$; then for $J_{\Omega_{\text{es}}}(M) = M$ the f.s. $\mathcal{M} = (M, \Omega_{\text{es}})$, and for $J_{\Omega_{\text{es},F}}(M) = M$ the f.s. $M = (\mathcal{M}, \Omega_{\text{es},F})$ are called iterative P.f.s. of automaton functions (P.f.s.a.f.). Examples of such f.s. are f.s. of the form $\mathcal{P}_{a,l,t}$. For a given $l \in N_1$, the main P.f.s.a.f. are

$$\begin{aligned} \mathcal{P}_{a,l,k} &= (P_{a,l,k}, \Omega_{\text{es}}), \qquad & \mathcal{P}^*_{a,l,k} &= (P_{a,l,k}, \Omega_{es,F}), \\ \mathcal{P}_{a,l} &= (P_{a,l}, \Omega_{\text{es}}), & \mathcal{P}^*_{a,l} &= (P_{a,l}, \Omega_{es,F}). \end{aligned} \tag{5.3}$$

Let us give the main results concerning our problems for P.f.s.a.f.

5.1 Theorem [19] *For any l from N_1 among the main P.f.s.a.f. only f.s. $\mathcal{P}^*_{a,l,k}$ is finitely generated.*

5.2 Theorem [17] *For any l from N and m from N_1 in $\mathcal{P}^*_{a,l,k}$ there exists a countable set of bases of power m.*

An a. function forming a complete system in $\mathcal{M}$ is called a *Sheffer function*. Further simplification of the basis is achieved by minimization of the number of variables, dimensionality and the number of states of a Sheffer a. function. The next assertion gives the final answer concerning the form of the simplest, in this sense, Sheffer a. functions.

5.3 Theorem [28] *For every l from N_1 in $\mathcal{P}^*_{a,l,k}$ there exist Sheffer one-dimensional a. functions of two variables and with two states.*

For the other two main P.f.s.a.f. the following assertion gives the answer to the basis problem.

5.4 Theorem [19] *For any l from N_1 in $\mathcal{P}_{a,l}$ and $\mathcal{P}^*_{a,l}$, there are no bases, in $\mathcal{P}_{a,l,k}$ there is a countable basis as well as a complete system containing no basis.*

As for the completeness problem, the situation is described by the following assertions.

5.5 Theorem [3, 19] *For the main P.f.s.a.f. the set $\Sigma_\pi(\mathcal{M})$ forms a c-system only for $\mathcal{M} = \mathcal{P}^*_{a,l,k}$ for any l from N_1.*

5.6 Theorem [17] *For any l from N_1 the set $\Sigma_\pi(\mathcal{M})$ has cardinality of continuum for $\mathcal{M} \in \{\mathcal{P}_{a,l,k}, \mathcal{P}^*_{a,l,k}\}$ and hypercontinuum for $\mathcal{M} \in \{\mathcal{P}_{a,l}, \mathcal{P}^*_{a,l}\}$.*

As a corollary we obtain that the corresponding lattices of closed classes in the main P.f.s.a.f. have the cardinalities mentioned in Theorem 5.6.

A system $\Sigma' \subseteq \Sigma(\mathcal{P}^*_{a,l,k})$ is called *c-criterial* if any finite set $M \subseteq \mathcal{P}_{a,l,k}$ is complete if and only if for any set $K \in \Sigma'$ the condition $M \not\subseteq K$ is valid.

5.7 Theorem [17] *In $\mathcal{P}^*_{a,l,k}$ there exist countable c-criterial systems of the form $\Sigma' \subseteq \Sigma_\pi(\mathcal{P}^*_{a,l,k})$ for any l from N_1.*

It should be pointed out that in the general case assigning a. functions from $\mathcal{P}_{a,l}$ is not efficient; therefore the problem of expressibility and completeness can be posed only for effectively assigned systems.

5.8 Theorem [15] *The problem of expressibility for effectively assigned finite systems of a. functions in the main P.f.s.a.f. and the problem of completeness in $\mathcal{P}_{a,l}$ are not algorithmically decidable for any l from N_1.*

Thus the extension of the functional possibilities of a. functions in comparison to the functions of l-valued logic and functions with delays considerably complicates the solution of the problems we are interested in for algebraic a. functions. The study of the nature of this complexity was carried out in different directions.

Here we dwell on the problem of approximate completeness and on the problem of completeness of specially enriched systems of a. functions.

The first of these problems has two modifications: the problem of r-completeness, $r \in N$, and the problem of approximate completeness (A-completeness), to which the next section is devoted. We call attention to special f.s.'s $\mathcal{P}_1$, $\mathcal{P}_2$, $\mathcal{P}_3$ and $\mathcal{P}_4$ that are subalgebras of the corresponding main P.f.s.a.f. from (5.3). Each of them consists of all one-place and one-dimensional a. functions from the main P.f.s., and operations employed in them are the same as in the corresponding P.f.s.a.f., except the operations σ and π. As has been established in [1, 19], they have no bases. Besides that $\mathcal{P}_1$ contains a subalgebra $\mathcal{P}'_1$, of all one-to-one mappings, which is a group with the operation of substitution and which models by one of its parts a group of Burnside type [2], i.e., a finitely-generated group in which the orders of the elements are finite, but are not bounded in totality. Still open are questions of the existence of bases in $\mathcal{P}'_1$, as well as the problem of algorithmic decidability of the property of finiteness of the order of its elements and expressibility of these elements by other elements. In conclusion we note that Theorems 5.1, 5.2, 5.5 and 5.6, and the facts mentioned here concerning one-place algebras also remain valid for the case where the value l is extended to countable in the f.s. of functions; in this way we generalize a. functions.

6 Conditions of r- and A-completeness for a. functions

A. functions f and g are *r-equivalent* if they coincide on all input words of length r (in this case we write $f\ r\ g$) and *A-equivalent* if $f\ r\ g$ for any r in N.

On the set $\mathcal{B}(P_{a,l})$ we introduce a relation Δ_r such that $M\ \Delta_r\ M'$ for $M, M' \subseteq P_{a,l}$, if for every function f from M there is g from M' such that $f\ r\ g$. It is clear that this relation forms a preorder and, consequently, can be represented as a relation of partial order on equivalence classes including all elements M and M' for which the relations $M\ \Delta_r\ M'$ and $M'\ \Delta_r\ M$ are valid; in this case we write $M\ r\ M'$ and the elements themselves are called r-equivalent.

On $\mathcal{B}(P_{a,l})$ we introduce one more relation, assuming that for $M, M' \subseteq P_{a,l}$ the relation $M\ \Delta_A\ M'$ is fulfilled if $M\ r\ M'$ for any r from N. This relation is also a preorder for representatives M and M' of its equivalence class, and when $M\ \Delta_A\ M'$ and $M'\ \Delta_A\ M$ we write $M\ A\ M'$ and the representatives themselves are called A-equivalent.

6.1 Theorem [8] *For any $M \subseteq P_{a,l}$ and $r \in N$*

$$J_{\Omega_{\text{es}}}(M)\ r\ J_{\Omega_{\text{es},F}}(M), \qquad J_{\Omega_{\text{es}}}(M)\ A\ J_{\Omega_{\text{es},F}}(M).$$

Thus the actions of the operators $J_{\Omega_{\text{es}}}(M)$ and $J_{\Omega_{\text{es},F}}(M)$ coincide up to r- and A-equivalence, and thereby in this sense the operation of feedback **F** turns out to be modeled by operations of extended superpositions, which we are going to use later.

Let $M, M' \subseteq P_{a,l}$. It is said that M is *r-expressible* by M' if $M\ \Delta_r\ J_{\Omega_{\text{es}}}(M')$, and *$A$-expressible* by M'' if $M\ \Delta_A\ J_{\Omega_{\text{es}}}(M'')$.

6.2 Theorem *For effectively assigned finite sets $M, M' \subseteq P_{a,l}$, the relation*

$$M\ \Delta_r\ J_{\Omega_{\text{es}}}(M')$$

is algorithmically decidable for any r in N.

6.3 Theorem [7] *For finite sets $M, M' \subseteq P_{a,l,k}$, the relation $M\ \Delta_A\ J_{\Omega_{\text{es}}}(M')$ is algorithmically undecidable.*

Let $M \subseteq P_{A,l}$ and $M\ A\ J_{\Omega_{\text{es}}}(M')$. The set $M' \subseteq M$ is called *r-complete* in M if $J_{\Omega_{\text{es}}}(M')\ r\ M$ and *A-complete* if $J_{\Omega_{\text{es}}}(M')\ A\ M$.

6.4 Theorem *If $M \subseteq P_{a,l}$, $M\ A\ J_{\Omega_{\text{es}}}(M)$, M is decidable, there is a finite A-complete subset of M and $r \in N$, then there exists an algorithm determining whether any finite decidable subset $M' \subseteq M$ is r-complete in M.*

In fact this theorem follows from Theorem 3.2; we verify this fact in the following way. Let $f \in P_{a,l}$ and $r \in N$. Consider the set E_l^r assuming that its elements are coded by the words of length r in the alphabet E_l. Then, examining the function f from $P_{a,l}$ only on words of length r, we can consider it to belong to P_l^r. Thus, from examining a. functions we passed to functions of l^r-valued logic. It remains to note that the operations of extended superposition with respect to expressibility and completeness are actually reduced to the operations of superposition. The following assertion is a corollary of Theorem 6.1 and the relation $P_{a,l,k}\ A\ P_{a,l}$.

6.5 Theorem *The conditions of r-completeness and of A-completeness, respectively, coincide for all main P.f.s.a.f.'s.*

We point out an essential difference between the notions of r-completeness and A-completeness which is given in the following assertion.

6.6 Theorem *In each of the main P.f.s.a.f. there exists finite A-complete systems and countable A-complete systems containing no finite A-complete subsystems.*

The difference in the notions of r-completeness and A-completeness appears in the following assertion.

6.7 Theorem [7] *If $M \subseteq P_{a,l,k}$ and M is finite, then there is no algorithm establishing whether M is A-complete in $P_{a,l,k}$.*

At the same time there is some similarity of the notions of r- and A-completeness and it reveals itself in the approach to solving problems of r- and A-completeness in terms of precomplete classes.

If $M \subseteq P_{a,l}$, M A $J_{\Omega_{\text{es}}}(M)$ and $M' \subseteq M$, then M' is called *r-precomplete* in M if it is not r-complete in M , but for any function f from $M \setminus M'$ the set $M' \cup \{f\}$ is r-complete in M. The notion of an *A-precomplete* is introduced analogously. Let $\sum_{\pi,r}(M)$ and $\sum_{\pi,A}(M)$ be the sets of all r-precomplete and A-precomplete sets in M respectively.

6.8 Theorem *If $M \subseteq P_{a,l}$ and M A $J_{\Omega_{\text{es}}}(M)$, then*

$$\sum_{\pi,A}(M) = \bigcup_{r\geq 1} \sum_{\pi,r}(M).$$

6.9 Theorem *If $\lambda \in \{r, A\}$, $r \in N$, $M \subseteq P_{a,l}$, M A $J_{\Omega_{es}}(M)$ in M where it is a finite λ-complete subset, and $M' \subseteq M$, then M' is λ-complete in M if and only if $M' \subset K$ for any $K \in \sum_{\pi,\lambda} M$.*

This assertion with regard to Theorems 6.6 and 6.8 reduces the solution of r- and A-completeness problems in the main P.f.s.a.f. to the description of the set $\sum_{\pi,r} P_{a,l}$ (this fact was obtained in [8]). Let us give this description.

Let $t \in N$; denote by E_l^t the set of all words $\epsilon = e(1)e(2)\dots e(t)$ of length t in the alphabet E_l. For $h \in N$ and $T = (t_1, t_2, \dots t_h)$, where $t_i \in N$, $i = 1, \dots h$, we put $E_l^T = E_l^{t_1} \times E_l^{t_2} \times \dots \times E_l^{t_h}$, $T \in N^h$. Let $\rho(y_1, y_2, \dots y_h)$ be an h-ary predicate whose arguments y_i take values in $E_l^{t_i}$, $i = 1, \dots h$. As previously, let ρ_1 and ρ_0 be, respectively, the sets of true and false collections of values of the variables for ρ. We say that a function $f(x_1, x_2, \dots x_n)$ from $P_{a,l}$ preserves ρ if the truth of the expression

$$\rho(f(\alpha_1^1, \alpha_1^2, \dots, \alpha_1^n), f(\alpha_2^1, \alpha_2^2, \dots, \alpha_2^n), \dots, f(\alpha_h^1, \alpha_h^2, \dots, \alpha_h^n))$$

follows from the truth of each element of the row

$$\rho(\alpha_1^1, \alpha_2^1, \dots, \alpha_h^1), \rho(\alpha_1^2, \alpha_2^2, \dots, \alpha_h^2), \dots, \rho(\alpha_1^n, \alpha_2^n, \dots, \alpha_h^n).$$

We denote the class of all such functions by $U_a(\rho)$.

We introduce the function $\nu : E_l^* \times E_l^* \to N_0$, putting for $\epsilon_1 = E_l^{t_1}$, $\epsilon_2 = E_l^{t_2}$, and $t = \min(t_1, t_2)$,

$$\nu(\epsilon_1, \epsilon_2) = \begin{cases} 0, & \text{if } e_1(1) = e_2(1), \dots, e_1(t) = e_2(t), \\ i, & \text{if } 1 \leq i \leq t-1 \text{ and } e_1(1) = e_2(1), \dots, e_1(t-i) = e_2(t-i), \\ & \text{but } e_1(t-i+1) \neq e_2(t-i+1), \\ t, & \text{if } e_1(1) \neq e_2(1). \end{cases}$$

We define the relation of preorder $\leq$ on the set E_l^T.

Let $A = (\alpha_1, \alpha_2, \dots, \alpha_h)$ and $A' = (\alpha_1', \alpha_2', \dots, \alpha_h')$ be elements from E_l^T. We write $A' \leq A$, if $\nu(\alpha_i', \alpha_j') \leq \nu(\alpha_i, \alpha_j)$ for any $i, j \in N_1^h$.

Let $t' = \max\{t_1, t_2, \dots, t_h\}$, $h \leq l^r$, and $t' \geq 2$; if $l = 2$, then we put $h = 2$. For $h \geq 2$, let

$$\nu(\alpha_i, \alpha_j) \neq 0, \quad i \neq j, \quad \text{in } A.$$

The set of all A' such that $A' \leq A$ is called the *ν-set determined by the element A*. We denote this set by ξ. ξ is divided into two subsets: $\xi^{(m)}$ consisting of all maximal elements with respect to the relation $\leq$ and $\xi^{(\underline{m})}$ containing the rest of the elements in ξ. Thus, for $h = 1$ we have $\xi^{(\underline{m})} = \emptyset$. It is clear that the value $\nu = (\alpha_i, \alpha_j)$ does not depend on the choice of A from $\xi^{(m)}$; therefore instead of $\nu = (\alpha_i, \alpha_j)$ we write $\nu = (i, j)$.

For $l \geq 2$ and $t \geq 1$ we indicate seven families of predicates.

Let $h \geq 1$, $T = (t_1, t_2, \ldots, t_h)$ and $\xi \subseteq E_l^T$. A substitution γ of the numbers $1, 2, \ldots, h$ is called a *ξ-substitution* if $(\alpha_{\gamma(1)}, \alpha_{\gamma(2)}, \ldots, \alpha_{\gamma(h)}) \in \xi$ for any $(\alpha_1, \alpha_2, \ldots, \alpha_h)$ in ξ.

Let $s \geq 1$; we say that the set

$$\{(\alpha_1^1, \alpha_2^1, \ldots, \alpha_h^1),\ (\alpha_1^2, \alpha_2^2, \ldots, \alpha_h^2), \ldots, (\alpha_1^s, \alpha_2^s, \ldots, \alpha_h^s)\}$$

of elements from E_l^T is *ξ-consistent* if there exists a collection of ξ-substitutions $\gamma_1, \gamma_2, \ldots, \gamma_s$ such that $\nu(i, j) \leq \nu(\alpha^q_{\gamma_q}(i), \alpha^p_{\gamma_p}(j))$ for any q and p in N_1^s, and any i and j in N_1^h.

Let $\rho(y_1, y_2, \ldots, y_h)$ be a predicate for which $\rho_1 \subseteq \xi$. We say that ρ is *ξ-reflexive* if $\xi^{(m)} \in \rho_1$, and *ξ-symmetrical* if $(\alpha_{\gamma(1)}, \alpha_{\gamma(2)}, \ldots, \alpha_{\gamma(h)}) \in \rho_1$ for any $(\alpha_1, \alpha_2, \ldots, \alpha_h)$ in ρ_1 and any ξ-substitution γ.

A ξ-reflexive and ξ-symmetrical predicate ρ is called *ξ-elementary* if the set $\xi \setminus \rho_1$ is ξ-consistent. For such ρ, $A \in \xi \setminus \rho_1$ and $i \in N_1^h$ we determine subsets $C_\rho^i(A)$, $Q_\rho^i(A)$ and $\epsilon_\rho^i(A)$ of the set E_l in the following way:

(1) $a \in C_\rho^i(A)$ if and only if there exists $\alpha_i' \in E_l^{t_i}$ such that $\nu(\alpha_i, \alpha_i') \leq 1$, $\alpha_i'(t_i) = a$, and any element $(\alpha_1', \alpha_2', \ldots, \alpha_h')$ from ξ is contained in ρ_1;

(2) $b \in Q_\rho^i(A)$ if and only if there exists an element

$$(\alpha_1, \alpha_2, \ldots, \alpha_{i-1}, \alpha_i'', \ldots, \alpha_h) \quad \text{in } \xi \setminus \rho_1$$

such that $\nu(\alpha_i, \alpha_i'') \leq 1$, $\alpha_i''(t_i) = b$;

(3) The set $\epsilon_\rho^i(A)$ coincides with $C_\rho^i(A)$ if $C_\rho^i(A) \neq \emptyset$ and with $Q_\rho^i(A)$ otherwise.

Let $n \geq 1$ and let $R = \{\rho^1, \rho^2, \ldots, \rho^n\}$ be an arbitrary system of ξ-elementary predicates. We say that R is *T-consistent* if $Q_\rho^i(A) \neq E_l$ for any $\sigma \in N_1^n$, $i \in N_1^h$, and $A \in \xi \setminus \rho_1^\sigma$. We say that R is *W-consistent* if for all σ, $\sigma \in N_1^n$, $i \in N_1^h$, $A \in \xi \setminus \rho_1^\sigma$, $A' \in \xi \setminus \rho_1^\sigma$ the sets $C^i_{\rho^\sigma}(A)$ and $C^i_{\rho^\sigma}(A')$ are simultaneously either empty or not empty, and also if $C^i_{\rho^\sigma}(A) = \emptyset$, then for any $b \in E_l$ there exists $j \in N_1^h$ such that $t_j = t_i$, $\nu(i, j) \leq 1$, $b \in Q^j_{\rho^\sigma}(A)$.

Let $n \geq 1$, $\sigma_i \in N_1^q$, $A^i = (\alpha_1^i, \alpha_2^i, \ldots, \alpha_n^i)$, $A^i \in \xi \setminus \rho_1^{\sigma_i}$ and $j_i \in N_1^h$ for $i \in N_1^n$. Then if $\sigma_i \neq \sigma_{i'}$ for $i \neq i'$ and $t_{j_i} = t_{j_{i'}}$, $\nu(\alpha^1_{j_1}, \alpha^2_{j_{i''}}) \leq 1$, for all $i, i' \in N_1^q$ and $i'' \in N_1^n \setminus \{1\}$, and

$$\bigcap_{i=1}^{n} \epsilon^{j_l}_{\rho_1^{\sigma_i}}(A^i) = \emptyset,$$

then the system R is called *Q-consistent*.

Let $n \geq 2$, $i \in N_1^n$, $\sigma_j \in N_1^q$ and $A^i \in \xi \setminus \rho_1^{\sigma_i}$. We assume that $\nu(\alpha_1^j, \alpha_1^{j'}) \neq 1$ for all $j, j' \in N_1^n$, $j \neq j'$. Let $t_{i_j} = t_{i_{j'}}$ and for all $i_j, i_{j'} \in N_1^h$, $\nu(\alpha^1_{i_1}, \alpha_{i_j}) \leq 1$ for $j > 1$, and

$$\bigcap_{j=1}^{n} \epsilon^{i_j}_{\rho^{\sigma_j}}(A^j) = \emptyset.$$

We suppose that under these conditions there exists $l_\nu \in N_1^n$, $\nu \in N_1^q$ such that $i_{l_\nu} \in N_1^2$, the sets A^{l_ν} are ξ-consistent and

$$\bigcap_{\nu=1}^{q} \epsilon_{\rho^{\sigma_{l_\nu}}}^{i_{l_\nu}}(A^{l_\nu}) = \emptyset.$$

Then we say that the system R is *D-consistent*.

The family of predicates $Z_l(r)$ This family is not empty for $l \geq 2$ and $r \geq 1$. A predicate ρ belongs to $Z_l(r)$ if and only if $\rho_1 \subset \xi$, $\xi \subseteq E_l^T$, $T = (t_1, t_2, \ldots, t_h)$, $h \geq 1$, $\max\{t_1, t_2, \ldots, t_h\} \leq r$ and for some $m \geq 1$

$$\rho_1 = \bigcap_{i=1}^{m} \rho_1^i,$$

where ρ^i is ξ-elementary predicate, and the ρ^i themselves form a T-, W- and Q-consistent system, and for any $j \in N_1^h$ and $A \in \xi \setminus \rho_1$ the set $C_\rho^i(A)$ is non-empty.

The family of predicates $J_l(r)$ This family is not empty for $r \geq 1$ and $l > 2$, and also for $r \geq 2$ and $l = 2$. A predicate ρ belongs to $J_l(r)$ if and only if $\rho_1 \subseteq \xi \subseteq E_l^T$, $T = (t_1, t_2, \ldots, t_h)$, $h \geq 3$, $\max\{t_1, t_2, \ldots, t_h\} \leq r$, for some $m \geq 1$

$$\rho_1 = \bigcap_{i=1}^{m} \rho_1^i,$$

where ρ^i is ξ-elementary, and the ρ^i themselves form a T-, W- and Q-consistent system; there are numbers $i, j, l \in N_1^h$, $i \neq j$, $i \neq l$, $j \neq l$, such that for A in $\xi \setminus \rho_1^1$ the set $C_\rho^u(A)$ is empty for $u \in \{i, j, l\}$.

The family of predicates $D_l(r)$ This family is not empty for $r \geq 1$ if $l > 2$ and for $r \geq 2$ and $l = 2$. A predicate ρ belongs to $J_l(r)$ if and only if $\rho_1 \subset \xi \subseteq E_l^T$, $T = (t_1, t_2, \ldots, t_h)$, $h \geq 2$, $\max\{t_1, t_2, \ldots, t_h\} \leq r$, for some $m \geq 1$

$$\rho_1 = \bigcap_{i=1}^{m} \rho_1^i,$$

where ρ^i is ξ-elementary, and the ρ^i themselves form a T-, W- and Q-consistent system. For any $A \in \xi \setminus \rho$, the sets $C_\rho^1(A)$ and $C_\rho^2(A)$ are empty. If $h \geq 3$, then $C_\rho^i(A) \neq \emptyset$ for any i from N_1^h; if $h = 2$, then $\rho_1 \cap \xi^{(m)} \neq \emptyset$.

Further $l \geq 2$, $t \geq 1$, $T = (t, t)$, $\xi_t \subseteq E_l^T$, ξ_t be a ν-subset and let $\nu_{\xi_t}(1, 2) = 1$.

The family of predicates $M_l(r)$ This family is not empty for $l \geq 2$ and $r \geq 1$. A predicate ρ belongs to $M_l(r)$ if and only if $\rho_1 \subset \xi_t$, $t \leq r$, and ρ_1 coincides with the relation of partial order determined on E_l^T and having exactly l^{t-1} minimal and l^{t-1} maximal elements.

The family of predicates $S_l(r)$ This family is not empty for $l \geq 2$ and $r \geq 1$. A predicate ρ belongs to $S_l(r)$ if and only if $\rho_1 \subset \xi_t$, $t \leq r$, and there is a substitution σ_ρ on E_l^t decomposing into a product of cycles of equal prime length $p \geq 2$ whose graph coincides with ρ_1, i.e., if $a \in E_l^T$, then $(a, \sigma_\rho(a)) \in \rho_1$, and if $(a_1, a_2) \in \rho_1$, then $a_2 = \sigma_\rho(a_1)$.

Let $\tilde{\Phi}_t$, $t \geq 1$, be the class of all mappings ϕ of the set E_l^T into the set of substitutions on E_l. The value ϕ at a is denoted by ϕ_a. Let $\Phi_t \subseteq \tilde{\Phi}_t$, and let Φ_t, consist of all ϕ such that $\phi_a = \phi_{a'}$ for $v(a, a') \leq 1$. Let $h \in \{3, 4\}$, $T_h = \{t, t, \ldots, t\}$, $K_t^h \subseteq E_l^{T_h}$ and let K_t^h consist of all elements $(a_1, a_2, \ldots, a_h)$ such that $v(a_i, a_j) \leq 1$ for $i, j \in N_1^h$. Let $l = p^m$, where p is prime, $m \geq 1$, and let $G = \langle E_l, + \rangle$ be an Abelian group whose every non-zero element has a prime order p. For $p \neq 2$ let $l_p \in N_1^{p-1}$ and $2l_p = 1 \pmod p$.

The family of predicates $L_l(r)$ This family is not empty only for $l = p^m$, where p is prime, $m \geq 1$, $t \leq r$. A predicate ρ belongs to $L_l(r)$ if and only if for some ϕ from Φ_t, $t \leq r$, the following assertions are true:

(1) Let $k = p^m$ and $p > 2$, then $\rho_1 \subset K_t^3$ and (a_1, a_2, a_3) from K_t^3 belongs to ρ_1 if and only if $\phi_{a_1}(a_3(t)) = l_p(\phi_{a_1}(a_1(t)) + \phi_{a_1}(a_2(t)))$.

(2) Let $k = p^m$, then $\rho_1 \subset K_t^h$ and (a_1, a_2, a_3, a_4) from K_t^4 belongs to ρ_1 if and only if $\phi_{a_1}(a_1(t)) + \phi_{a_1}(a_2(t)) = \phi_{a_1}(a_3(t)) + \phi_{a_1}(a_4(t))$.

We note that the families indicated for $r = 1$ coincide, respectively, with the known families for P_l, from Section 3. Let $t > 2$, $T = (t, t)$, and let $\tilde{\xi}_t$, be a μ-subset E_l^T such that $\mu_{\tilde{\xi}_t}(1, 1) = 2$.

The family of predicates $V_l(r)$ This family is not empty for $l \geq 2$ and $r \geq 2$. A predicate ρ belongs to $V_l(r)$ if and only if $\rho_1 \subset \tilde{\xi}_t$, $t \leq r$ and $(a_1, a_2) \in \tilde{\xi}_t^{(m)} \cap \rho_1$, if and only if $a_1(t) = a_2(t)$; there is ϕ from Φ_t, such that the inclusion $(a_1, a_2) \in \tilde{\xi}_t^{(m)} \cap \rho_1$ is equivalent to the existence of α in E_l such that $a_1(t) = \phi_{a_1}(\alpha)$ and $a_2(t) = \phi_{a_2}(t)$.

Let $W_l(r) = Z_l(r) \cup J_l(r) \cup D_l(r) \cup M_l(r) \cup S_l(r) \cup L_l(r) \cup V_l(r)$, and let $U(W_l(r))$ be the set of classes of a. functions preserving predicates from $W_l(r)$.

6.10 Theorem [8] *The equality* $\Sigma_{\pi,r}(P_{a,l}) = U(W_l(r))$ *is true.*

7 Conditions of completeness for special systems of automaton functions

Another way of studying the completeness properties of a. functions systems consists in enriching the operations performed both on a. functions and on a. function systems which are tested for completeness. We consider the second situation. Here two approaches are of interest. The first approach consists in considering Slupecki's problem for f.s. $\mathcal{P}_{a,l,k}$ and the second one concerns such systems of a. functions in $\mathcal{P}^*_{a,l,k}$ that contain some closed class of truth a. functions. Following [27], a finite system T of f.a. functions in $\mathcal{P}_{a,l,k}$ is called a *Slupecki system* if $J_{\Omega_{\text{es}}}(P^{1,1}_{a,l,k} \cup T) = P_{a,l,k}$.

7.1 Theorem [25] *For any* l *from* N_1 *there exist Slupecki systems of f.a. functions from* $P_{a,l,k}$.

V. A. Buyevich proved the following assertion.

7.2 Theorem *For any l from N_1 there exists an algorithm which establishes for a finite set of f.a. functions in $\mathcal{P}_{a,l,k}$ whether it is a Slupecki system.*

A precomplete class of the main P.s.a.f. is called a *Slupecki class* if it contains all homogeneous functions from the carrier of P.s.a.f. It is interesting to compare Theorems 7.1 and 7.2 with the following assertion.

7.3 Theorem *The cardinality of the set of Slupecki classes is equal to the continuum in the f.s.'s $\mathcal{P}_{a,l,k}$ and $\mathcal{P}^*_{a,l,k}$, and is equal to the hypercontinuum in the f.s.'s $\mathcal{P}_{a,l}$ and $\mathcal{P}^*_{a,l}$, for any l from N_1.*

This assertion indicates a very important difference in completeness conditions in the general case and in the case of Slupecki systems. Note that the hypercontinuum property in this theorem remains valid also for f.s. functions which are obtained from a. functions by extending the value of the parameter l to countable. The second approach was first realized in the form of the following assertion.

7.4 Theorem [5] *For any l in N_1, there exists an algorithm which establishes for any finite set $M \subseteq P_{a,l,k}$ whether the set $M \cup P_{a,l,k,t}$ is complete in $\mathcal{P}^*_{a,l,k}$.*

Later on this approach was developed for $l = 2$ in the following way. Let Q be a closed Post class of Boolean functions. We interpret these functions as truth f.a. functions. Denote by K_Q the class of all finite sets M of f.a. functions from $P_{a,2,k}$ such that $M \supseteq Q$. Our primary interest here is the existence of an algorithm which establishes for any M from K_Q whether the set is complete in $\mathcal{P}^*_{a,2,k}$ (to be more exact, for which Q this algorithm exists). It is clear that if such an algorithm exists for some Q, then it also exists for any Post class Q' such that $Q' \supseteq Q$. On the contrary, if such an algorithm does not exist, then it does not exist for any Post class $Q'' \subseteq Q$. Thus, the separation of algorithmically decidable cases from algorithmically undecidable ones is achieved by indicating such a family Φ of the Post classes Q for which the completeness property is not algorithmically decidable. However for any Post class such that $Q' \supset Q$ it is decidable.

7.5 Theorem [6] *The family Φ is equal to*

$$\Phi = \{F_1^\infty, F_2^\infty, F_3^\infty, F_4^\infty, F_5^\infty, F_6^\infty, F_7^\infty, F_8^\infty, L_2, L_3, L_5, S_6, P_6, O_9\}.$$

A special case of this assertion is Theorem 7.4 for $l = 2$. As is shown in [6], the assertion of Theorem 7.5 holds also for the case of A-completeness with the proper interpretation of the set Φ.

8 Linear a. functions

An a. function from $P_{a,l,k}$ is *linear* (*l.a.f.*) if in its system (5.1) the vector-functions ϕ and ψ are linear modulo l with respect to their variables. Denote by $L_{a,l,k}$ the class of all linear a. functions. It is not difficult to see that $I_{\Omega_{\mathrm{es},F}}(L_{a,l,k}) = L_{a,l,k}$. Let us consider the f.s. $\mathcal{L}_{a,l,k} = (L_{a,l,k}, \Omega_{\mathrm{es}})$ and $\mathcal{L}^*_{a,l,k} = (L_{a,l,k}, \Omega_{\mathrm{es},F})$.

8.1 Theorem *Among the f.s.* $\mathcal{L}_{a,l,k}$ *and* $\mathcal{L}^*_{a,l,k}$, *only* $\mathcal{L}^*_{a,l,k}$ *is finitely-generated.*

A solution of the completeness problem for $\mathcal{L}^*_{a,l,k}$ is obtained for $l = 2$, and we present it [9] using the notation L_a instead of $L_{a,l,k}$ and the notation $\mathcal{L}_a$ instead of $\mathcal{L}_{a,l,k}$. Let Γ_a, $a \in E_2$, be the class of all B.f. $f(x_1, x_2, \ldots, x_n) = (y_1, y_2, \ldots, y_m)$ such that $f(a, a, \ldots, a) = (a, a, \ldots, a)$. It is said that a B.f. $f(x_1, x_2, \ldots, x_n)$ depends on x_i with a shift, if in its system (5.1), x_i is absent in the equations for ψ; otherwise x_i, is called an *immediate variable*. Let V_1 be the class of all B.f. having no more than one immediate variable. Let V_2 be the class of all B.f. having an odd number of immediate variables. Denote by R_C the class of all B.f. having exactly one essential variable, and by R_H, the class of all B.f. with exactly one immediate variable. The class of all B.f. without immediate variables is denoted by C. Let $\alpha^s[0]$ be the word of length s of the form $00\ldots0$. Denote by L^0 the class of all B.f. $f(x_1, x_2, \ldots, x_n)$ such that

$$f(\alpha^s[0], \alpha^s[0], \ldots, \alpha^s[0]) = (\alpha^s[0], \alpha^s[0], \ldots, \alpha^s[0])$$

for any s in N. Let L_1 be the class of all B.f.'s of one variable and $L^0_1 = L^0 \cap L_1$. Denote by $E_2[z]$ the ring of polynomials in z over the field E_2 with the usual operations of addition and multiplication of polynomials. For $\{u, v, u', v'\} \subset E_2[z]$ we consider the fractions u/v, u'/v' which we assume equal if $uu_1 = u'u_2$ and $vu_1 = v'u_2$ for some u_1, u_2 in $E_2[z] \setminus \{0\}$. The degree of a polynomial u is denoted by $\deg u$. Let $Q_2(z)$ be the set of all fractions u/v, u/v is incontractible, $v \in E_2[z] \setminus \{0\}$, then v is not divided by z. It is possible to show that L^0_1 and $Q_2(z)$ are isomorphic (we write $\sim$) and retain the operations of addition and multiplication; therefore they are identical in a sense.

Let $u/v \in Q_2(z)$, $f \in L_a$ and $u/v \sim f$. It is said that f possesses the *O-property* if either $u/v \in R_H$ and $\deg u = \deg v$, or $u/v \notin R_H$ and $\deg u < \deg v$. We consider polynomials p_i from $Q_2(z)$ ordered according to ascending degree, i.e., $\deg p_i \leq \deg p_{i+i}$ for $i \in N$, and let $p_1 = \xi$. If $u + v$ or u is divisible by ξp_i, then we say that f possesses the *i-property*. If $\deg u < \deg v$, then f possesses the *O′-property*; if $\deg u \leq \deg v$, then f possesses the *O″-property*. If u is divided by p_i then f possesses *i′-property*; if v is not divisible by p_i, then it possesses *i″-property*.

Let $M^{(1)}_i$ consist of all B.f. f having i-property, $i \in N_0$, and let $R^{(1)}_i$ consist of all B.f. f with the i'-property.

For a B.f. $f(x_1, x_2, \ldots, x_n)$ there exist functions $f_1(x_1), f_2(x_2), \ldots, f_n(x_n)$ from L_1 and γ from C such that

$$f(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} f_i(x_i) + \gamma \pmod 2.$$

Denote the set of functions $f_1, f_2, \ldots, f_n$ by $\mu(f)$. Let M_i consist of all B.f.'s f such that $\mu(F) \subset M^{(1)}_i$. Let a B.f. f satisfy the following property: if x_j is the only essential variable, then f_j in $\mu(f)$ possesses the i''-property; and if the indicated property is absent in x_j, then f_j possesses the i'-property. The class of such f is denoted by R^C_j.

Let a B.f. f satisfy the following property: if x_j is the only immediate variable, then f_j in $\mu(f)$ possesses the i''-property; otherwise f possesses the i'-property. The class of such f is denoted by R^H_j.

Denote by J the family consisting of all classes $\Gamma_0, \Gamma_1, V_1, V_2, M_i, R^C_j, R^H_j$ for $i \in N_1$, $j \in N$.

8.2 Theorem [9] *The relation* $\Sigma_\pi(\mathcal{L}_a^*) = J$ *holds.*

Thus in the f.s. $\mathcal{L}_a^*$ there is only a countable set of precomplete classes and by virtue of the fact that $\mathcal{L}_a^*$ is finitely generated they form a criterial system. By analogy with the case of functions with delays the following assertion holds.

8.3 Theorem [9] *There is an algorithm establishing whether any finite system of B.f.'s in* $\mathcal{L}_a^*$ *is complete.*

References

[1] S. V. Aleshin, On the absence of bases in some classes of initial automata, *Probl. Cybern.* **22** (1970), 67–75 (Russian).

[2] S. V. Aleshin, Finite automata and the Burnside problem for torsion groups, *Math. Notes* **29** (1972), 319–328 (Russian).

[3] D. N. Babin, On superpositions of bounded-determined functions, *Math. Notes* **47** (1990), 3–10 (Russian).

[4] D. N. Babin, On completeness of the binary boundedly determined functions with respect to superposition, *Discrete Math. Appl.* **1** (1991), 423–431.

[5] D. N. Babin, A decidable case of the completeness problem for automata functions, *Diskretnaya Matematika* **4** (4) (1992), 41–55 (Russian).

[6] D. N. Babin, Undecidability of the completeness and A-completeness problems for some systems of automaton functions, *Discrete Math. Appl.* **5** (1995), 31–42.

[7] V. A. Buyevich, On algorithmic unsolvability of A-completeness recognition for bounded-determined functions, *Math. Notes* **29** (1972), 687–697 (Russian).

[8] V. A. Buyevich, On r-completeness in the class of determined functions, *Soviet Acad. Sci. Dokl.* **326** (1992), 399–404 (Russian).

[9] A. A. Chasovskikh, On completeness in the linear automata class, *Math. Probl. Cybern.* **3** (1995), 140–166 (Russian).

[10] L. Chzhu-Kai, Precomplete classes determined by the k-nary relations in k-valued logic, *Acta Sci. Natur. Univ. Jilinensis* **3** (1964).

[11] L. Chzhu-Kai and Lu.Sui-Khua, Precomplete classes determined by the binary relations in many-valued logic, *Acta Sci. Natur. Univ. Jilinensis* **4** (1964).

[12] P. Cohn, *Universal Algebra*, Wiley, New York, 1965.

[13] J. Dassow, Ein modifizierter Vollständigkeitsbergiff in einen Algebra von Automatenabbildungen, Dissertation fur Doctor B, Uni Rostock, 1978.

[14] K. B. Kolyada, On completeness of regular transformations, Ph.D. thesis, Moscow State University, 1987 (Russian).

[15] M. I. Kratko, Algoritmic undecidability of the completeness recognition problem for finite automata, *Soviet Acad. Sci. Dokl.* **155** (1964), 35–37 (Russian).

[16] V. B. Kudryavtsev, A completeness theorem for one class of automata without feedback, *Probl. Cybern.* **8** (1962), 91–115 (Russian).

[17] V. B. Kudryavtsev, On the cardinality of the sets of precomplete sets of some functional systems connected with automata, *Probl. Cybern.* **13** (1965), 45–74 (Russian).

[18] V. B. Kudryavtsev, *Functional Systems*, Moscow Univ. Press, Moscow, 1982 (Russian).

[19] V. B. Kudryavtsev, S. V. Aleshin, and A. S. Podkolzin, *An Introduction to Automata Theory*, Nauka, Moscow, 1985 (Russian).

[20] A. I. Maltsev, Iterative algebras and the Post's varieties, *Algebra and Logic* **5** (2) (1966), 5–24 (Russian).

[21] V. V. Martynyuk, An investigation of some classes of functions in many-valued logic, *Probl. Cybern.* **3** (1960), 49–60 (Russian).

[22] J. McCarthy and C. A. Shannon, eds., *Automata Studies*, Princeton Univ. Press, Princeton, 1956.

[23] A. S. Strogalov, On ε-modelling of finite automata behavior, Ph.D. thesis, Moscow State University, 1985 (Russian).

[24] E. Post, *Two-valued Iterative Systems of Mathematical Logic*, Princeton Univ. Press, Princeton, 1941.

[25] I. G. Rosenberg and T. Hikita, Completeness for uniformly delayed circuits, in: *Proc. 13th Intern. Symposium on Multiple-Valued Logic*, Kyoto, 1983, 1–9.

[26] I. Rosenberg, La structure des functions de plusieurs variables sur un ensemble fini, *Comptes Rendus Acad. Sci. Paris* **260** (1965), 3817–3819.

[27] J. Słupecki, Kriterium pelnosci wielowartosciowych systemow logiki zdan, *Comptes rendus des scéances de la Société des Sciences et des Lettres de Varsovie* **32** (1939), 102–109.

[28] E. V. Vetrennikova, A simple example of the universal b.d. function, *Discrete Analysis* (1983), 5–11 (Russian).

[29] S. V. Yablonskii, Functional constructions in the k-valued logic, *Proc. Steklov Institute Math.*, **51** (1958), 5–142 (Russian).

[30] S. V. Yablonskii, O. P. Gavrilov, and V. B. Kudryavtsev, *Boolean Functions and the Post Classes*, Nauka, Moscow, 1966 (Russian).

[31] Yu. I. Yanov and A. A. Muchnik, On the existence of k-valued closed classes that do not have a basis, *Soviet Acad. Sci. Dokl.* **127** (1959), 144–146 (Russian).

[32] P. Yun-Tse, A method of discovering all precomplete classes in many-valued logic, *Acta Sci. Natur. Univ. Jilinensis* **2** (1964).

[33] E. Yu. Zakharova, Completeness criteria of a system of functions from P_k, *Probl. Cybern.* **18** (1967), 5–10 (Russian).

[34] E. Yu. Zakharova, V. B. Kudryavtsev, and S. V. Yablonskii, On precomplete classes in k-valued logics, *Soviet Acad. Sci. Dokl.* **186** (1969), 509–512 (Russian).

Algebra of behavior transformations and its applications

Alexander LETICHEVSKY

Glushkov Institute of Cybernetics
National Academy of Sciences of Ukraine
Ukraine

Abstract

The model of interaction of agents and environments is considered. Both agents and environments are characterized by their behaviors represented as the elements of continuous behavior algebra, a kind of the ACP with approximation relation, but in addition each environment is supplied by an insertion function, which takes the behavior of an agent and the behavior of an environment as arguments and return a new behavior of this environment. Each agent can be considered as a transformer of environment behaviors and a new kind of equivalence of agents weaker than bisimulation is defined in terms of the algebra of behavior transformations. Arbitrary continuous functions can be used as insertion functions and rewriting logic is used to define computable ones. The theory has applications for studying distributed computations, multi agent systems and semantics of specification languages.

1 Introduction

The topic of these lectures belongs to an intensively developing area of computer science: the mathematical theory of communication and interaction. The paradigm shift from computation to interaction and the wide-spread occurrence of distributed computations attract a great deal of interest among researchers in this area.

Concurrent processes or agents are the main objects of the theory of interaction. Agents are objects, which can be recognized as separate from the rest of a world or an environment. They exist in time and space, change their internal state, and can interact with other agents and environments, performing observable actions and changing their place among other agents (mobility). Agents can be objects in real life or models of components of an information environment in a computerized world. The notion of agent formalizes such objects as software components, programs, users, clients, servers, active components of distributed knowledge bases and so on. A more specific and rich notion of agent is also used in so called agent programming, an engineering discipline devoted to the design of intelligent interactive systems.

Theories of communication, interaction, and concurrency have a long history, that starts from the structural theory of automata (50-th years of the last century). Petri Nets is another very popular general model of concurrency. However Petri Nets is very specific, and structural automata theory requires too many details to represent interaction in a sufficiently

V. B. Kudryavtsev and I. G. Rosenberg (eds.),
Structural Theory of Automata, Semigroups, and Universal Algebra, 241–272.

abstract form. Theories of communication and interaction which appeared in 70-th captured the fundamental properties of the notion of interaction and constitute the basis of modern research in this field. They include CCS (Calculus of Communicated Processes) [19, 20] and the π-calculus [21] of R. Milner, CSP (Communicated Sequential Processes) of T. Hoar [13], ACP (Algebra of Communicated Processes) [5] and many branches of these basic theories. The current state-of-the-art is well represented in recently edited Handbook of Process Algebra [6]. A new look at the area appeared recently in connection with the coalgebraic approach to interaction [24, 4].

The main notion of the theory of interaction is the behavior of agents or processes interacting with each other within some environment. At the same time traditional theories of interaction do not formalize the notion of an environment where agents are interacting or consider very special cases of it. The usual point of view is that an environment for a given agent is the set of all other agents surrounding it. In the theory of interaction of agents and environments developed in [15] and presented in the lectures, the notion of environment is formalized as an agent supplied by an insertion function, which describes the change of behavior of an environment after the agent is inserted. After inserting one agent an environment is ready to accept another one and, considered as an agent, it can itself be inserted into another environment of higher level. Therefore multi-agent and multilevel environments can be created using insertion functions.

Both agents and environments are characterized by their behaviors represented as the elements of a continuous behavior algebra, a kind of ACP with approximation relation [14]. The insertion function takes the behavior of an agent and the behavior of environment as arguments and returns a new behavior of this environment. Each agent therefore can be considered as a transformer of environment behaviors and a new equivalence of agents is defined in terms of the algebra of behavior transformations. This idea comes from Glushkov discrete transformers (processors) [9, 10], an approach considering programs and microprograms as the elements of the algebra of state space transformations. In the theory of agents and environments the transformations of a behavior space is considered instead, so this transition is similar to the transition from point spaces to functional spaces in mathematical analysis.

Arbitrary continuous functions can be used as insertion functions and rewriting logic is applied to define computable ones. The theory has applications for the study of distributed computations and multi agent systems. It is used for the development specification languages and tools for the design of concurrent and distributed software systems. Three lectures correspond to the three main sections of the paper and contain the following.

The first section gives an introduction to the algebraic theory of processes considered as agent behaviors. Agents are represented by means of labeled transition systems with divergence and termination, and considered up to bisimilarity or other (weaker) equivalences. The theorem characterizing bisimilarity in terms of a complete behavior algebra (cpo with algebraic structure) is proved and the enrichment of a behavior algebra by sequential and parallel compositions is considered. The second section introduces algebras of behavior transformations. These algebras are classified by the properties of insertion functions and in many cases can be considered as behavior algebras as well. The enrichment of a transformation algebra by parallel and sequential composition can be done only in very special cases. Two aspects of studying transformation algebras can be distinguished.

Mathematical aspect The study of algebras of environments and behavior transformations as mathematical objects, classifying insertion functions and algebras of behavior transformations generated by them, developing specific methods of proving properties of systems represented as environments with inserted agents.

Application aspect The insertion programming, that is a programming based on agents and environments. This is an answer to the paradigm shift from computation to interaction and consists of developing a methodology of insertion programming (design environment and agents inserted in it) as well as the development of tools supporting insertion programming. The application of the insertion programming approach to the development of proof systems is considered in the last section.

2 Behavior algebras

2.1 Transition systems

Transition systems are used to describe the dynamics of systems. There are several kinds of transition systems which are obtained by the enrichment of an ordinary transition system with additional structures.

An *ordinary transition system* is defined as a couple

$$\langle S, T\rangle, \quad T \subseteq S^2$$

where S is the set of states and T is a *transition relation* denoted also as $s \to s'$. If there are no additional structures, perhaps the only useful construction is the transitive closure of the transition relation denoted as $s \xrightarrow{*} s'$ and expressing the reachability in the state space S.

A *labeled transition system* is defined as a triple

$$\langle S, A, T\rangle, \quad T \subseteq S \times A \times S$$

where S is again a set of states, A is a set of actions (alternative terminology: labels or events), and T is a set of labeled transitions. Belonging to a transition relation is denoted by $s \xrightarrow{a} s'$. This is the main notion in the theory of interaction. We can consider the external behavior of a system and its internal functioning using the notion of labeled transitions. As in automata theory two states are considered to be equivalent if we cannot distinguish them by observing only external behavior, that is actions produced by a system during its functioning. This equivalence is captured by the notion of bisimilarity discussed below. Both the notion of transition system and bisimilarity go back to R. Milner and, in its modern form, were introduced by D. Park [22] who studied infinite behavior of automata.

The *mixed* version

$$\langle S, A, T\rangle, \quad T \subseteq S \times A \times S \cup S^2$$

combines unlabeled transitions $s \to s'$ with labeled ones $s \xrightarrow{a} s'$. In this case we discuss unobservable or hidden and observable transitions. However as it will be demonstrated later, the mixed version can be reduced to labeled systems. Technically sometime it is easier to define a mixed system and then reduce it to labeled one.

The *attributed transition systems*:

$$\langle S, A, U, T, \varphi \rangle, \quad \varphi : S \to U.$$

This kind of transition system is used when not only transitions but also states should be labeled. The function φ is called a state label function. Usually a set of state labels is structured as $U = D^R$, where the set R is called the set of attributes and the set D is a set of attribute values. These sets can also be typed and in this case $U = (D_\xi^{R_\xi})_{\xi \in \Xi}$ (Ξ is the set of type symbols).

Adjusted transition systems are obtained distinguishing three kinds of subsets,

$$S_0, S_\Delta, S_\perp \subseteq S$$

in a set S of system states. They are *initial states*, *states of successful termination* and *undefined (divergent) states*, respectively. The supposed meaning of these adjustments is the following: from initial states a system can start in an initial state and terminate in a state of successful termination, undefined states are used to define an approximation relation on the set of states; the behavior of a system can be refined (extended) in undefined states. The states of successful termination must be distinguished from the *dead lock states*, that is the states from which there are no transitions but which are neither states of successful termination nor undefined states. The property of a state having no transitions is denoted as $s \not\to$.

Other important classes of transition systems are stochastic, fuzzy, and real time transition systems. All of them are obtained by introducing some additional numeric structure to different kinds of transition systems and will not be considered here. Attributed transition systems as well as mixed systems can be reduced to labeled ones, so the main kind of system will be labeled an adjusted transition system (usually with $S_0 = S$) and other kinds will be used only in examples.

Let us consider some useful examples (without details which the reader is encouraged to supply himself/herself).

Automata The set A of actions is identified with an input (output) alphabet or with the set of pairs input/output.

Programs The set A of actions is an instruction set or only input/output instructions according to what should be considered as observable actions. The set S is the set of states of a program (including memory states). If we want some variables to be observable, a system can be defined with a state label function mapping the variable symbols to their values in a given state.

Program schemata Symbolic (allowing multiple interpretations) instructions and states are considered. The set of actions are the same as in the model of a program.

Parallel or distributed programs and program schemata The set A of actions is a set of observable actions performed in parallel or sequentially (in this case parallelism is modelled by interleaving) in different components; communications are usually represented by hidden transitions (as in CCS). The states are composed with the states of components by parallel composition. This example will be considered below in more details.

Calculi States are formulas; actions are the names of inference rules.

Data and knowledge bases Actions are queries.

There are two kinds of non-determinism inherent in transition systems. The first one is the existence of two transitions $s \xrightarrow{a} s'$ and $s \xrightarrow{a} s''$ for some state s with $s' \neq s''$. This non-determinism means that, after performing an action a, a system can choose the next state non-deterministically. The second kind of non-determinism is the possibility of different adjustments of the same state, that is a state can be at the same time a state of successful termination as well as undefined or initial.

A labeled transition system (without hidden transitions) is called *deterministic* if for arbitrary transitions from $s \xrightarrow{a} s' \wedge s \xrightarrow{a} s''$ it follows that $s' = s''$ and $S_\Delta \cap S_\perp = \emptyset$.

2.2 Trace equivalence

A *history* of system performance is defined as a sequence of transitions starting from some initial state s_1 and continuing at each step by application of transition relation, to a state obtained at this step:

$$s_1 \xrightarrow{a_1} s_2 \xrightarrow{a_2} \cdots \xrightarrow{a_n} \cdots$$

A history can be finite or infinite. It is called *final* if it is infinite or cannot be continued. A *trace* corresponding to a given history is a sequence of actions performed along this history:

$$a_1 a_2 \ldots a_n \ldots$$

For an attributed transition system the trace includes the state labels:

$$\varphi(s_1) \xrightarrow{a_1} \varphi(s_2) \xrightarrow{a_2} \cdots \xrightarrow{a_n} \cdots$$

Different sets of traces can be used as invariants of system behavior. They are called *trace invariants*. Examples of trace invariants of a system S are the following sets: $L(S)$—the set of all traces of a system S; $L_S(s)$—the set of all traces starting at the state s; $L_\Delta(S)$—the set of all traces finishing at a terminal state, $L^0_\Delta(S)$—the set of all traces starting at an initial state and finishing at a terminal state, etc. All these invariants can be easily computed for finite state systems as regular languages.

We obtain the notion of *trace equivalence* considering $L^0_\Delta(S)$ as the main trace invariant: systems S and S' are trace equivalent ($S \sim_T S'$) if $L^0_\Delta(S) = L^0_\Delta(S')$. Unfortunately trace equivalence is too weak to capture the notion of transition system behavior. Consider the two systems presented in Fig. 1.

Both systems in the figure start their activity by performing an action a. But the first of the two systems has a choice at the second step. It can perform action b or c. At the same

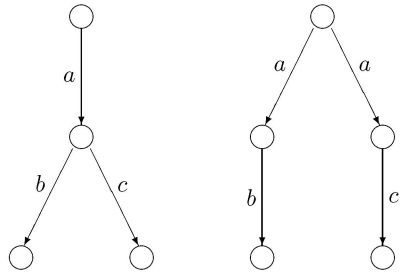


Figure 1: Trace equivalent systems with different behaviors

time the second system will only perform an action b and can never perform c or it can only perform c and never perform b, dependent on what decision was made at the first step. The equivalence, stronger than trace equivalence, that captures the difference between the two systems in Fig. 1 is bisimilarity. It is considered in the next section.

2.3 Bisimilarity

2.1 Definition A binary relation $R \subseteq S^2$ is called a *bisimulation* if:

(1) $(s,s') \in R \implies (s \in S_\Delta \iff s' \in S_\Delta,\ s \in S_\perp \iff s' \in S_\perp)$;

(2) $(s,s') \in R \wedge s \xrightarrow{a} t \implies \exists t'\ ((t,t') \in R \wedge s' \xrightarrow{a} t')$;

(3) $(s,s') \in R \wedge s' \xrightarrow{a} t' \implies \exists t\ ((t,t') \in R \wedge s \xrightarrow{a} t)$.

States s and s' are called *bisimilar* ($s \sim_B s'$) if there exists a bisimulation R such that $(s,s') \in R$. For attributed transition systems an additional requirement is: $(s,s') \in R \implies \varphi(s) = \varphi(s')$. We can also extend this definition to mixed transition systems if $\exists s'\ (s \xrightarrow{*} s' \xrightarrow{a} t)$ will be used instead of $s \xrightarrow{a} t$ and use $\exists s'\ (s \xrightarrow{*} s' \wedge s' \in S_\Delta(S_\perp))$ instead of $s \in S_\Delta(S_\perp)$.

2.2 Proposition *Bisimilarity is an equivalence relation.*

Proof Note that $\{(s,s) \mid s \in S\}$ is a bisimulation. If R is a bisimulation then R^{-1} is a bisimulation and if R and R' are bisimulations then $R \circ R'$ is also a bisimulation. □

2.3 Proposition *Bisimilarity is a maximal bisimulation on S.*

Proof An arbitrary union of bisimulations is again a bisimulation; therefore a bisimilarity is a union of all bisimulations on S. □

Bisimilarity of two states can be extended to the case when they are the states of different systems in a usual way (consider the disjoint union of the two systems). The bisimilarity of two systems can also be defined so that each state of one of them must be bisimilar to some state in the other.

Reduction of mixed transition systems Let S be a mixed transition system. Add new rules to define new labeled transitions and extend termination states in the following way.

$$\frac{s \xrightarrow{*} s',\ s' \xrightarrow{a} s''}{s \xrightarrow{a} s'}$$

$$s \xrightarrow{*} s',\ s' \in S_\Delta(S_\perp) \implies s \in S_\Delta(S_\perp).$$

Now delete unlabeled transitions. The new labeled system is called a reduction of the system S.

2.4 Proposition *A mixed transition system and its reduction are bisimilar.*

Proof The relation $s' \xrightarrow{*} s$ between s, considered as a state of a reduced system, and s', considered as a state of a mixed system, is a bisimulation. □

For a deterministic system the difference between trace equivalence and bisimilarity disappears.

2.5 Proposition *For deterministic systems* $s \sim_T s' \implies s \sim_B s'$.

Th spectrum of different equivalences, from trace equivalence to bisimilarity, can be found in the paper of Glabbeek [8]. Bisimilarity is the strongest; trace equivalence is the weakest.

To define an approximation relation on the set of states of a transition system, the notion of partial bisimulation will be introduced.

2.6 Definition The binary relation $R \subseteq S^2$ is called a *partial bisimulation* if:

(1) $(s,s') \in R \implies (s \in S_\Delta \implies s' \in S_\Delta,\ s \notin S_\perp \implies s' \notin S_\perp) \wedge (s \notin S_\perp \wedge s' \in S_\Delta \implies s \in S_\Delta)$;

(2) $(s,s') \in R \wedge s \xrightarrow{a} t \implies \exists\, t'\ ((t,t') \in R \wedge s' \xrightarrow{a} t')$ (the same as for bisimilarity);

(3) $(s,s') \in R \wedge s \notin S_\perp \wedge s' \xrightarrow{a} t' \implies \exists\, t\ ((t,t') \in R \wedge s \xrightarrow{a} t)$ (the same as for bisimilarity with the additional restriction $s \notin S_\perp$).

We say that s is less defined then s' or s *approximates* s' ($s \sqsubseteq_B s'$), if there exists a partial bisimulation such that $(s,s') \in E$. A partial bisimulation is a preorder and from the definitions it follows that:

2.7 Proposition $s \sim_B s' \iff s \sqsubseteq_B s' \sqsubseteq_B s$.

2.4 Behavior algebras

The invariant of a trace equivalence is a language. What is the invariant of a bisimilarity? To answer this question one should define the notion of behavior of a transition system (in a given state). Intuitively it is a node of a diagram of a transition system unfolded into a (finite or infinite) labeled tree (synchronization tree), with some nodes of this tree being identified. More precisely, two transitions from the same node labeled by the same action should be identified if they lead to bisimilar subtrees. Different approaches are known for studying bisimilarity. Among them are Hennessy-Milner logic [12], the domain approach of

S. Abramsky [1], and the final coalgebra approach of Aczel and Mendler [3]. A comparative study of different approaches to characterize bisimilarity can be found in [23]. Here we shall give the solution based on continuous algebras [11] or algebras with an approximation [14]. The variety of algebras with approximation relation will be defined and a minimal complete algebra $F(A)$ over a set of actions A will be constructed and used for the characterization of bisimilarity. It is not the most general setting, but the details of direct constructions are important for the next steps in developing the algebra of transformations.

Behavior algebra $\langle U, A\rangle$ is a two sorted algebra. The elements of sort U are called *behaviors*, the elements of A are called *actions*. The signature and identities of a behavior algebra are the following.

Signature *Prefixing* $a.u$, $a \in A$, $u \in U$, *non-deterministic choice* $u + v$, $u, v \in U$, *termination constants* Δ, $\bot$, 0, called *successful termination*, *divergence* and *dead lock* correspondingly, and *approximation relation* $u \sqsubseteq v$ (u approximates v), $u, v, \in U$.

Identities Non-deterministic choice is an associative, commutative, and idempotent operation with 0 as a neutral element ($u+0 = u$). Approximation relation $\sqsubseteq$ is a partial order with minimal element $\bot$. Both operations (prefixing and non-deterministic choice) are monotonic with respect to the approximation relation:

$$\begin{aligned} &\bot \sqsubseteq u,\\ &u \sqsubseteq v \implies u + w \sqsubseteq v + w,\\ &u \sqsubseteq v \implies a.u \sqsubseteq a.v. \end{aligned}$$

Continuity Prefixing and non-deterministic choice are continuous with respect to approximation, that is they preserve least upper bounds of directed sets of behaviors if they exist.

More precisely, let $D \subseteq U$ be a directed set of behaviors, that is for any two elements $d', d'' \in D$ there exists $d \in D$ such that $d' \sqsubseteq d$, $d'' \sqsubseteq d$. The least upper bound of the set D if it exists will be denoted as $\bigsqcup D$ or $\bigsqcup_{d\in D} d$. The continuity condition for U means that

$$\begin{aligned} a.\bigsqcup D &= \bigsqcup_{d\in D} a.d,\\ \bigsqcup D + u &= \bigsqcup_{d\in D} (d + u). \end{aligned}$$

Note that monotonicity follows from continuity.

Some additional structures can be defined on the components of a behavior algebra.

Actions A combination $a \times b$ of actions can be introduced as a binary associative and commutative (but in general case not idempotent) operation to describe communication or simultaneous (parallel) performance of actions. In this case an impossible action $\emptyset$ is introduced as unnulator for combination and unit action δ with identities

$$a \times \emptyset = \emptyset,$$
$$a \times \delta = a,$$
$$\emptyset.u = 0.$$

In CCS each action a has a dual action $\overline{a}$ ($\overline{\overline{a}} = a$) and the combination is defined as $a \times \overline{a} = \delta$ and $a \times b = \emptyset$ for non-dual actions (the symbol τ is used in CCS instead of δ; it denotes the observation of hidden transitions and two states are defined as weakly bisimilar if they are bisimilar after changing τ transitions to hidden ones). In CSP another combination is used: $a \times a = a$, $a \times b = \emptyset$ for $a \neq b$.

Attributes A function defined on behaviors and taking values in an attribute domain can be introduced to define behaviors for attributed transition systems.

To characterize bisimilarity we shall construct a complete behavior algebra $F(A)$. Completeness means that all directed sets have least upper bounds. We start from the algebra $F_{\text{fin}}(A)$ of finite behaviors. This is a free algebra generated by termination constants (an initial object in the variety of behavior algebras). Then this algebra is extended to a complete one adding the limits of directed sets of finite behaviors. To obtain infinite, convergent (definition see below), non-deterministic sums this extension must be done through the intermediate extension F_{fin}^{∞} of the algebra of finite depth elements.

Algebra of finite behaviors $F_{\text{fin}}(A)$ is the algebra of behavior terms generated by termination constants considered up to identities for non-deterministic choice, and with the approximation relation defined in the following way.

2.8 Definition (Approximation in $F_{\text{fin}}(A)$) $u \sqsubseteq v$ if and only if there exists a term $\varphi(x_1, \ldots, x_n)$ generated by termination constants and variables $x_1, \ldots, x_n$ and terms $v_1, \ldots, v_n$ such that $u = \varphi(\bot, \ldots, \bot)$ and $v = \varphi(v_1, \ldots, v_n)$.

2.9 Proposition *Each element of $F_{\text{fin}}(A)$ can be represented in the form*

$$u = \sum_{i \in I} a_i.u_i + \varepsilon_u$$

where I is a finite set of indices and ε is a termination constant. If the $a_i.u_i$ are all different and all u_i are represented in the same form, this representation is unique up to commutativity of non-deterministic choice.

Proof The proof is by induction on the height $h(u)$ of a term u defined in the following way: $h(\varepsilon) = 0$ for the termination constant ε, $h(a.u) = h(u) + 1$, $h(u + v) = \max\{h(u), h(v)\}$. □

A termination constant ε_u can possess the following values: 0, Δ, $\bot$, $\bot + \Delta$. Behavior u is called *divergent* if $\varepsilon_u = \bot, \bot + \Delta$, otherwise it is called *convergent*. For *terminal* behaviors $\varepsilon_u = \Delta, \bot + \Delta$ and behavior u is *guarded* if $\varepsilon_u = 0$.

2.10 Proposition *$u \sqsubseteq v$ if and only if*

(1) $\varepsilon_u \sqsubseteq \varepsilon_v$;

(2) $u = a.u' + u'' \implies v = a.v' + v'',\ u' \sqsubseteq v'$;

(3) $v = a.v' + v''$ *and* u *is convergent* $\implies u = a.u' + u'',\ u' \sqsubseteq v'$.

2.11 Proposition *The algebra* $F_{\text{fin}}(A)$ *is a free behavior algebra.*

Proof Only properties of approximation need proof. To prove that approximation is a partial order and that prefixing and nondeterministic choice are monotoniv is an easy exercise (to prove antisymmety use Proposition 2.10). To prove that the operations are continuous note that each finite behavior has only a finite number of approximations and therefore only finite directed sets have least upper bounds. The property $\varphi(\bot,\dots,\bot) \sqsubseteq \varphi(v_1,\dots,v_n)$ is true in an arbitrary behavior algebra (induction); therefore the approximation in $F_{\text{fin}}(A)$ is a minimal one. □

Note that in $F_{\text{fin}}(A)$

$$x = y \iff x \sqsubseteq y \sqsubseteq x.$$

Algebra of finite height behaviors $F^{\infty}_{\text{fin}}(A)$ is defined in the following way. Let

$$\begin{aligned} F^{(\infty)}_{\text{fin}}(A) &= \bigcup_{n=0}^{\infty} F^{(n)}, \\ F^{(0)}(A) &= \{\Delta, \bot, \Delta + \bot, 0\}, \\ F^{(n+1)}(A) &= \{\textstyle\sum_{i\in I} a_i.u_i + v \mid u_i, v \in F^{(n)}\}, \end{aligned}$$

where I is an arbitrary set of indices, but expressions $\sum_{i\in I} a_i.u_i$ and $\sum_{j\in J} b_j.v_j$ are identified if $\{a_i.u_i \mid i \in I\} = \{b_j.v_j \mid j \in J\}$. Therefore one can restrict the cardinality of infinite I to be no more then $2^{|A|}$ for $F^{(1)}(A)$ and no more then $2^{|F^{(n)}|}$ for $F^{(n+1)}(A)$.

Take Proposition 2.10 as the definition of an approximation relation on the set $F^{\infty}_{\text{fin}}(A)$. Taking into account the identification of infinite sums we have again that $x = y \iff x \sqsubseteq y \sqsubseteq x$. Define prefixing as $a.u$ for $u \in F^{(n)}(A)$ and define $\sum_{i\in I} a_i.u_i + \sum_{j\in J} b_j.v_j = \sum_{k\in I\cup J} c_k.w_k$ where $I \cap J = \emptyset$ and $c_k.w_k = a_k.u_k$ for $k \in I$ and $c_k.w_k = b_k.v_k$ for $k \in J$ (disjoint union).

2.12 Proposition *The algebra* $F^{\infty}_{\text{fin}}(A)$ *is a behavior algebra.*

Proof Use induction on the height. □

However the algebra $F^{\infty}_{\text{fin}}(A)$ has the same identities as $F_{\text{fin}}(A)$. It is not free because it has no free generators and the equality

$$\sum_{i\in I} a_i.u_i + \sum_{j\in J} b_j.v_j = \sum_{k\in K} c_k.w_k$$

for $\{a_i.u_i \mid i \in I\} \cup \{b_j.v_j \mid j \in J\} = \{c_k.w_k \mid k \in K\}$ does not follow from the identities when at least one of I or J is infinite. But they are the only equalities except for identities in $F^{\infty}_{\text{fin}}(A)$ (infinite associativity).

In the algebra $F^{\infty}_{\text{fin}}(A)$ the canonical representation of Proposition 2.9 is still valid for infinite sets of indices.

Let X be a set of variables. Define the set $F^{\infty}_{\text{fin}}(A,X)$ in the same way as $F^{\infty}_{\text{fin}}(A)$, but redefine $F^{(0)}(A)$ as $F^{(0)}(A,X) = \{\sum_{i\in I}\varepsilon_i \mid \varepsilon_i \in F^{(0)}(A) \cup X,\ i \in I\}$ so that, besides the set of termination constants, it also includes the sums of variables. The set $F^{\infty}_{\text{fin}}(A,X)$ is a behavior algebra with operations and approximation defined in the same way as for $F^{\infty}_{\text{fin}}(A)$.

Define substitution $\sigma = \{x_i := v_i \mid i \in I\}$ as a homomorphism $u \mapsto u\sigma$ such that $x_i\sigma = v_i$, $\varepsilon\sigma = \varepsilon$ for termination constants, and $(\sum u_i)\sigma = \sum u_i\sigma$. If $u, v \in F^{\infty}_{\text{fin}}(A,X)$ then $u(v)$ denotes $u\{x := v,\ x \in X\}$.

2.13 Proposition *For elements $u, v \in F^{\infty}_{\text{fin}}(A,X)$ the approximation relation satisfies the following statement: $u \sqsubseteq v$ if and only if there exists $\varphi \in F^{\infty}_{\text{fin}}(A,X)$ and substitution $\sigma = \{x_i := v_i \mid i \in I, x_i \in X\}$ such that $\varphi(\bot) = u$, and $\varphi\sigma = v$.*

Proof By induction on the height of u. □

Complete behavior algebra $F(A)$

The elements of $F(A)$ are directed sets of $F^{\infty}_{\text{fin}}(A)$ considered up to the following equivalence.

2.14 Definition Directed sets U and V in $F^{\infty}_{\text{fin}}(A)$ are called equivalent ($U \sim V$) if for each $u \in U$ there exists $v \in V$ such that $u \sqsubseteq v$ and for each $v \in V$ there exists $u \in U$ such that $v \sqsubseteq u$.

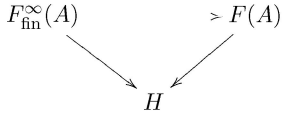
Define operations and approximation on directed sets in the following way.

- Prefixing: $a.U = \{a.u \mid u \in U\}$;
- Non-deterministic choice: $U + V = \{u + v \mid u \in U, V \in V\}$;
- Approximation: $U \sqsubseteq V \iff \forall(u \in U)\ \exists(v \in V)\ (u \sqsubseteq v)$.

These operations preserve equivalence and therefore can be extended to classes of equivalent directed sets.

2.15 Proposition *The algebra $F(A)$ is a behavior algebra. It is a minimal complete conservative extension of the algebra $F^{\infty}_{\text{fin}}(A)$.*

Proof The least upper bound of a directed set of elements of $F(A)$ is a (set theoretical) union of these elements. Also the algebra $F^{\infty}_{\text{fin}}(A)$ can be isomorphically embedded into $F(A)$ by the mapping of $u \in F^{\infty}_{\text{fin}}(A)$ to $\{v \mid v \sqsubseteq u\}$. Minimality means that if H is another complete conservative extension of $F^{\infty}_{\text{fin}}(A)$ then there exists a continuous homomorphism from $F(A)$ to H such that the following diagram is commutative:



For details see [14]. □

Let us define $\sum_{i\in I} u_i$ for $u_i \in F(A)$ and infinite I as $\{\sum_{i\in I} v_i \mid v_i \in u_i\}$. Note that $F(A,X)$ can be defined as a free complete extension of $F(A)$ and Proposition 2.13 can be proved for $F(A,X)$

2.16 Proposition *Each $u \in F(A)$ can be represented in the form $u = \sum_{i\in I} a_i.u_i + \varepsilon_u$ and this representation is unique if all $a_i.u_i$ are different.*

Proof Let $M(a)$ be the set of all solutions of the equation $a.x + y = u$ with unknowns $x, y \in F(A)$ and $S(a)$ the set of all x such that, for some y, $(x,y) \in M(a)$. Let $I = \{(a,u) \mid a \in A,\ u \in S(a)\}$ and $a_{(a,u)} = a$, $u_{(a,u)} = u$. Then $u = \sum_{i\in I} a_i.u_i + \varepsilon_u$ and uniqueness is obvious. □

Another standard representation of behaviors is through the definition of a minimal solution of the system of equations

$$x_i = F_i(X), \quad i \in I$$

where $F_i(X) \in F^{\infty}_{\mathrm{fin}}(A,X)$ and $x_i \in X$. As usually, this minimal solution is defined as $x_i = \bigsqcup_{i=0}^{(\infty)} x_i^{(n)}$ where $x_i^{(0)} = \bot$, $x_i^{(n+1)} = (F_i(X))\sigma_n$, $\sigma_{(n+1)} = \{x_i := x_i^{(n)},\ i \in I\}$. Note that the first representation is used in the co-algebraic approach and the second is a slight generalisation of the traditional fixed point approach.

2.5 Behaviors of transition systems

Let S be a labeled transition system over A. For each state $s \in S$, define the behavior $\mathtt{beh}(s) = u_s$ of a system S in a state s as a minimal solution of the system

$$u_s = \sum_{s \xrightarrow{a} t} a.u_t + \varepsilon_s$$

where ε_s is defined in the following way:

$$\begin{aligned} s \notin S_\Delta \cup S_\bot &\implies \varepsilon_s = 0,\\ s \in S_\Delta \backslash S_\bot &\implies \varepsilon_s = \Delta,\\ s \in S_\bot \backslash S_\Delta &\implies \varepsilon_s = \bot,\\ s \in S_\Delta \cap S_\bot &\implies \varepsilon_s = \Delta + \bot. \end{aligned}$$

Behaviors as states

A set of behaviors $U \subseteq F(A)$ is called *transition closed* if

$$a.u + v \in U \implies u \in U.$$

In this case U can be considered as a transition system if transitions and adjustment are defined in the following way:

$$\begin{aligned} a.u + v &\xrightarrow{a} u,\\ U_\Delta &= \{u \mid u = u + \Delta\},\\ U_\bot &= \{u \mid u = u + \bot\}. \end{aligned}$$

2.17 Theorem *Let s and t be states of a transition system, u and v behaviors. Then*

(1) $s \sqsubseteq_B t \iff u_s \sqsubseteq u_t$;

(2) $s \sim_B t \iff u_s = u_t$;

(3) $u \sqsubseteq v \iff u \sqsubseteq_B v$;

(4) $u = v \iff u \sqsubseteq_B v$.

Proof The first follows from the bisimilarity of s and u_s considered as a state. (1) $\Rightarrow$ (2) because $\sqsubseteq$ is a partial order, and (2) $\Rightarrow$ (3) because $\mathtt{beh}(u) = u$. □

An *agent* is an adjusted labeled transition system. An *abstract agent* is an agent with states considered up to bisimilarity. Identifying the states with behaviors we can consider an abstract agent as a transition closed set of behaviors. Conversely, considering behaviors as states we obtain a standard representation of an agent as a transition system. This representation is defined uniquely up to bisimilarity. We should distinguish an agent as a set of states or behaviors from an agent in a given state. In the latter case we consider each individual state or behavior of an agent as the same agent in a given state adjusted in such a way that it has the unique initial state. Usually this distinction is understood from the context.

2.6 Sequential and parallel compositions

There are many compositions enriching the base process algebra or the algebra of behaviors. Most of them are defined independently on the representation of an agent as a transition system. These operations preserve bisimilarity and can be considered as operations on behaviors. Another useful property of these operations is continuity. The use of definitions in the style of SOS semantics [2] or the use of conditional rewriting logic [18] always produces continuous functions if these definitions are expressed in terms of behavior algebras. The mostly popular operations are *sequential* and *parallel* compositions.

Sequential composition is defined by means of the following inference rules and equations:

$$\frac{u \xrightarrow{a} u'}{(u;v) \xrightarrow{a} (u';v)},$$
$$((u+\Delta);v) = (u;v) + v,$$
$$((u+\bot);v) = (u;v) + \bot,$$
$$(u;0) = 0.$$

These definitions should be understood in the following way. First we extend the signature of the behavior algebra adding new binary operation $(( );( ))$. Then add identities for this operation and convince yourself that no new equation appears in the original signature (extension is conservative). Then a transition relation is defined on the set of equivalence classes of extended behavior expressions (independence of the choice of representative must be shown). These classes now become the states of a transition system, and the value of the expression is defined as its behavior. In the sequel we shall use the notation uv instead of $(u;v)$.

2.18 Exercise Prove identities $\Delta u = u\Delta = u$, $\bot u = \bot$, $(uv)w = u(vw)$, $(u+v)w = uw+vw$. Hint: Define bisimilarity (for non-trivial cases).

Sequential composition can be also defined explicitly by the following recursive definition:

$$uv = \sum_{u \xrightarrow{a} u'} a(u'v) + \sum_{u=u+\varepsilon} \varepsilon v,$$
$$0v = 0, \qquad \Delta v = v, \qquad \bot v = \bot.$$

If an action a is identified with the agent $a.\Delta$, then we have $a = a.\Delta = (a;\Delta) = a\Delta$.

Parallel composition of behaviors assumes that a combination of actions is defined. It is considered as a associative and commutative operation $a \times b$ with annulator $\emptyset$. Rules and identities for the parallel composition are

$$\frac{u \xrightarrow{a} u',\ v \xrightarrow{b} v',\ a \times b \neq \emptyset}{u\|v \xrightarrow{a\times b} u'\|v'},$$

$$\frac{u \xrightarrow{a} u',\ v \xrightarrow{b} v'}{u\|v \xrightarrow{a} u'\|v,\ u\|v \xrightarrow{b} u\|v',\ u\|(v+\Delta) \xrightarrow{a} u',\ (u+\Delta)\|v \xrightarrow{b} v'},$$

$$(u+\Delta)\|(v+\Delta) = (u+\Delta)\|(v+\Delta) + \Delta,$$
$$(u+\bot)\|v = (u+\bot)\|v + \bot,$$
$$u\|(v+\bot) = u\|(v+\bot) + \bot.$$

2.19 Exercise Prove associativity and commutativity of parallel composition.

An explicit definition of parallel composition is

$$u\|v = \sum_{u \xrightarrow{a} u', v \xrightarrow{b} v'} (a \times b)(u'\|v') + \sum_{u \xrightarrow{a} u'} a(u'\|v) + \sum_{v \xrightarrow{b} v'} b(u\|v') + \varepsilon_u\|\varepsilon_v$$

where $\varepsilon_u(\varepsilon_v)$ is a termination constant in the representation $u = \sum a_i u_i + \varepsilon_u$ of a behavior u.

3 Algebra of behavior transformations

3.1 Environments and insertion functions

An *environment* is an abstract agent E over the set C of environment actions together with a continuous *insertion function* $\mathtt{Ins}\colon E \times F(A) \to E$. All states of E are considered as possible initial states. Therefore an environment is a tuple $\langle E, C, A, \mathtt{Ins}\rangle$. In the sequel C, A, and $\mathtt{Ins}$ will be used implicitly and $\mathtt{Ins}(e,u)$ will be denoted as $e[u]$. After inserting an agent u (in a given state u), the new environment is ready for new agents to be inserted and the insertion of several agents is something that we will often wish to describe. Therefore the notation

$$e[u_1, \ldots, u_n] = e[u_1] \ldots [u_n]$$

will be used to describe this insertion.

Each agent (behavior) u defines a transformation $[u]$ of environment (behavior transformation); $[u]\colon E \to E$ is such that $[u](e) = e[u]$. The set of all behavior transformations of a type $[u]$ of environment E is denoted by $T(E) = \{[u] \mid u \in F(A)\}$. This is a subset of the set $\Phi(E)$ of all continuous transformations of E. A semigroup multiplication $[u] * [v]$ of two transformations $[u]$ and $[v]$ can be defined as follows:

$$([u] * [v])(e) = (e[u])[v] = e[u, v]$$

The semigroup generated by $T(E)$ is denoted as $T^*(E)$ and (for a given insertion function) we have:

$$T(E) \subseteq T^*(E) \subseteq \Phi(E)$$

An insertion function is called a *semigroup insertion* if $T(E) = T^*(E)$. It is possible if and only if for all $u, v \in F(A)$ there exists $w \in F(A)$ such that for all $e \in E$, $e[u, v] = e[w]$.

It is interesting also to fix the cases when $T(E) = \Phi(E)$. Such an insertion is called *universal*. A trivial universal insertion exists if the cardinality of A is not less then the cardinality of $\Phi(E)$. In this case all functions can be enumerated by actions with the mapping $\varphi \mapsto a_\varphi$ and insertion function can be defined so that $e[a_\varphi] = \varphi(e)$.

The kernel of the mapping $u \mapsto [u]$ of $F(A)$ to $T(E)$ defines an equivalence relation on the set F(A) of agents (behaviors). This equivalence is called an *insertion equivalence*:

$$u \sim_E v \iff \forall (e \in E)\ (e[u] = e[v])$$

Generally speaking, insertion equivalence is not a congruence.

Let $\sim_E$ is a congruence. In this case the operations of the behavior algebra $F(A)$ can be transferred to $T(E)$ so that

$$e([u] + [v]) = e[u + v]$$
$$e(a.[u]) = e[a.u]$$

Define an approximation relation on $T(E)$ so that

$$[u] \sqsubseteq [v] \iff \forall (e \in E)\ (e[u] \sqsubseteq e[v])$$

3.1 Theorem *If $\sim_E$ is a congruence, $T(E)$ is a behavior algebra and the mapping $u \mapsto [u]$ is a continuous homomorphism of $F(A)$ on $T(E)$.*

An environment can be also defined as a two sorted algebra $\langle E, T(E)\rangle$ with the insertion function considered as an external operation on E.

Let us consider some simple examples.

Parallel insertion An insertion function is called a *parallel insertion* if it satisfies the following condition:

$$e[u, v] = e[u \| v].$$

A parallel insertion is a semigroup insertion with composition defined as $[u] * [v] = [u \| v]$. An example of parallel insertion is the insertion function $e[u] = e \| u$ (for $A = C$). This function is called a *strong parallel insertion*. In this case $\sim_E$ is a congruence and if $\Delta \in E$ it coincides with a bisimilarity. Strong parallel insertion models the situation when an environment for a given agent is a parallel composition of all other agents interacting with it.

Sequential insertion An insertion function is called a *sequential insertion* if it satisfies the following condition:

$$e[u, v] = e[uv].$$

A sequential insertion is also a semigroup insertion with $[u]*[v] = [uv]$ and a *strong sequential insertion* defined by the equation $e[u] = eu$ $(A = C)$ is a congruence and in this case the insertion equivalence is a bisimilarity if $\Delta \in E$.

Trace environment A *trace environment* is generated by one state: $E = \{e_0[u] \mid u \in F(A)\}$. An insertion function is defined by the equations $e_0[u, v] = e_0[uv]$, $e_0[\Delta] = e_0$, $e_0[\bot] = \bot$, $e_0[0] = 0$, and

$$e_0\left[\sum_{i\in I} a_i u_i + \varepsilon\right] = \sum_{a\in A} a.e_0\left[\sum_{i\in I,\ a_i=a} u_i\right] + e_0[\varepsilon].$$

It is easy to prove the following.

3.2 Theorem *For a trace environment E, $u \sim_E v$ if and only if $u \sim_T v$.*

In a trace environment we also have a distributive law $[x] * ([y] + [z]) = [x] * [y] + [x] * [z]$ and a Klinee like algebra can be defined by introducing an iteration $[u]^* = \sum_{n=0}^{\infty} [u]^n$. But this algebra contains not only finite but also infinite behaviors and there are equalities like $uv = u$ if u has no termination constant Δ at the end of some history.

Problem Find environments for all the equivalences between trace and bisimulation defined in [8].

3.2 Classification of insertion functions

In this and the following sections assume that the set E of environment behaviors is not only transition closed but, for each behavior e, also contains all of its approximations. That is from $e \in E$ and $e' \sqsubseteq e$ it follows that $e' \in E$.

Every insertion function is a continuous one, therefore it can be represented in the form

$$e[u] = \bigsqcup_{e'\sqsubseteq e,\ u'\sqsubseteq u} e'[u']$$

where $e' \in F_m^\infty(C)$, $u' \in F_m^\infty(A)$. Using this representation the following proposition can be proved.

3.3 Proposition

(1) $e[u] \xrightarrow{c} f \implies$ *there exist, for some m, $e' \in F_m^\infty(C)$, $u' \in F_m^\infty(A), f' \in F(C)$ such that $e'[u'] \xrightarrow{c} f'$, $e' \sqsubseteq e$, $u' \sqsubseteq u$, $f' \sqsubseteq f$;*

(2) $e'[u'] \xrightarrow{c} f'$, $e' \in F_m^\infty(C)$, $u' \in F_m^\infty(A), f' \in F(C) \implies$ *there exist $e \in E$, u, f such that $e[u] \xrightarrow{c} f$ and $e' \sqsubseteq e$, $u' \sqsubseteq u$, $f' \sqsubseteq f$.*

From this proposition and Proposition 2.13 it follows that for each transition $e'[u'] \xrightarrow{c} f'$ there must be a "transition equation"

$$e(X)[u(Y)] \xrightarrow{c} f(X \cup Z)\sigma \tag{3.1}$$

such that $e(\perp) = e'$, $u(\perp) = u'$, $f(\perp) = f'$ and σ substitutes continuous functions of X and Y to Z. To cover more instances the intersection of X and Y must be empty and all occurrences of variables in the left hand side must be different (left linearity). These equations belong to some extended transformation algebra enriched by corresponding functions. More precisely the meaning of (3.1) can be described by the following equational formula:

$$\forall\, \sigma_1,\ \forall\, \sigma_2\ \exists\, g((e(X)\sigma_1)[u(Y)\sigma_2] = c.(f(X \cup Z)\sigma_1)\sigma' + g)$$

where $\sigma_1 : X \to F(C)$, $\sigma_2 : Y \to F(A)$, $g \in F(C)$, $\sigma' = \sigma(\sigma_1 + \sigma_2)$ (disjoint union of two substitutions in the right hand sides of substitution σ). The set of termination equations

$$e(X)[u(Y)] = e(X)[u(Y)] + \varepsilon \tag{3.2}$$

must be considered together with the transition ones. The set of all transition and termination equations uniquely defines an insertion function and can be used for its computation.

The previous discussion results in trying to apply rewriting logic [18] for recursive computation of insertion functions and modeling the behavior of an environment with inserted agents. For this purpose one should restrict consideration to only those substitutions expressed in the form $\sigma = \{z_i := f_i[u_i] \mid z_i \in Z,\ f_i \in F^{\infty}_{\text{fin}}(C, X),\ u_i \in F^{\infty}_{\text{fin}}(A, Y)\}$. This restriction means that the insertion function is defined by means of identities of a two-sorted algebra of environment transformations (with insertion as the operation). Insertions (environments) defined in this way will be called *equationally defined* insertions (environments).

Equationally defined environments can be classified by additional restrictions on the form of rewriting rules for insertion functions. The first classification is on the height of e and u in the left hand side of (3.1):

- One-step insertion: $(1, 1)$; both environment and agent terms have the height 1.
- Head insertion: $(m, 1)$; the environment term is of arbitrary height and the agent term of height 1. This case is reduced to one-step insertion.
- Look-ahead insertion: (m, m); the general case, reduced to the case $(1, m)$.

In addition we restrict the insertion function by the additivity conditions:

$$\left(\sum e_i\right)[u] = \sum (e_i[u]) \tag{3.3}$$

$$e\left[\sum u_i\right] = \sum (e[u_i]) \tag{3.4}$$

Both conditions will be used for a one-step insertion and the second one—for a head insertion. Restictions for termination equations will not be considered. They are assumed to be in a general form.

3.3 One-step insertion

First we shall consider some special cases of one-step insertion and later it will be shown that the general case can be reduced to this one. From the additivity conditions it follows that the transitions for $c.e[a.u]$, $\varepsilon[a.u]$, $(c.e)[\varepsilon]$, and $\varepsilon[\varepsilon']$ should be defined to dependend only on a, c, and $\varepsilon, \varepsilon' \in \mathrm{E} = \{\bot, \Delta, 0\}$. Other restrictions follow from the rules below. To define insertion assume that two functions $D_1 : A \times C \to 2^C$ and $D_2 : C \to 2^C$ are given. The rules for insertion are defined in the following way.

$$\frac{u \xrightarrow{a} u',\ e \xrightarrow{c} e',\ d \in D_1(a,c)}{e[u] \xrightarrow{d} e'[u']} \qquad \text{(interaction)},$$

$$\frac{e \xrightarrow{c} e',\ d \in D_2(c)}{e[u] \xrightarrow{d} e'[u]} \qquad \text{(environment move)}.$$

In addition we must define a continuous function $\varphi_\varepsilon(u) = \varepsilon[u]$ for each $\varepsilon \in E$. This function must satisfy the following conditions. For all $e \in E$ and $u \in F(A)$

$$\bot[u] \sqsubseteq e[u], \qquad e[u] + 0[u] = e[u]. \tag{3.5}$$

The simplest way to meet these conditions is to define $\bot[u] = \bot$ and $0[u] = 0$. There are no specific assumptions for $\Delta[u]$, but usually neither Δ nor 0 belongs to E. Note that in the case when $\Delta \in E$ and $\Delta[u] = u$ the insertion equivalence is a bisimulation.

3.4 Theorem *For a one-step insertion the equivalence on $F(A)$ is a congruence and $T(A)$ is an environment algebra isomorphic to the quotient algebra of $F(A)$ by insertion equivalence.*

First let us prove the following statement.

3.5 Proposition *For a one-step insertion there exists a continuous function $F : A \times T(E) \to T(E)$ such that $[a.u] = F(a, [u])$,*

Proof Let

$$e = \sum_{i \in I} c_i e_i + \varepsilon_e. \tag{3.6}$$

Then

$$e[a.u] = \sum_{d \in D_1(a,c_i)} d.e_i[u] + \sum_{d \in D_2(c)} d.e_i[a.u] + \varepsilon_e[a.u].$$

Therefore for an arbitrary e the value $e[a.u]$ can be found from the minimal solution of the system defined by these equations with unknowns $e[u]$ and $e[a.u]$ (for arbitrary e and u). □

Proof of Theorem 3.4 Define $a.[u] = F(a, [u])$, $[u] + [v] = \lambda e.(e[u] + e[v])$, $[u] \sqsubseteq [v] \iff \forall\, e.(e[u] \sqsubseteq e[v])$, and $[\varepsilon] = \lambda e.\psi_\varepsilon(e)$, where $\psi_\varepsilon(e) = e[\varepsilon]$ for e defined by (3.6) is found from the system of equations

$$\psi_\varepsilon(e) = \sum_{i \in I} \sum_{d \in D_2(c_i)} d.\psi_\varepsilon(e_i) + \varphi_\varepsilon(\varepsilon_e).$$

Now $T(E)$ is a behavior algebra and $u \mapsto [u]$ is a continuous homomorphism. □

3.6 Example Let $A \subseteq C$. Define combinations $c \times c'$ of actions on the set C as arbitrary ac-operation with identities $c \times \delta = c,\ c \times \emptyset = \emptyset$. Now define the functions for a one-step insertion as follows: $D_1(a,c) = \{d \mid c = a \times d\},\ D_2(c) = \{c\}$. It is easy to prove that this is a parallel insertion: $e[u,v] = e[u \| v]$.

Parallel computation over shared and distributed memory The insertion of the previous example can be used to model parallel computation over shared memory. In this case

$$E = \{e[u_1, u_2, \dots] \mid e : R \to D\}.$$

Here R is a set of names, D is a data domain and the environment is called a shared memory over R. Actions $c \in C$ correspond to statements about memory such as assignements or conditions. The combination $c \times c' \neq \emptyset$ if and only if c and c' are consistent. The notion of consistency depends on the nature of actions and intuitively means that they can be performed simultaneously. The transition rules are:

$$\frac{e \xrightarrow{a \times d} e',\ u \xrightarrow{a} u'}{e[u] \xrightarrow{d} e'[u']}.$$

As a consequence

$$\frac{e \xrightarrow{a_1 \times a_2 \times \cdots \times d} e',\ u_1 \xrightarrow{a_1} u_1', \dots}{e[u_1 \| u_2 \| \dots] \xrightarrow{d} e'[u_1',\ u_2', \dots]}.$$

The residual action d in the transition $e[u_1 \| u_2 \| \dots] \xrightarrow{d} e'[u_1',\ u_2', \dots]$ is intended to be used by external agents inserted later, but it can be a convenient restricted interaction only with a given set of agents already inserted. For this purpose a shared memory environment can be inserted into a higher level closure environment with the insertion function defined by the equation $g[e[u]][v] = g[e[u \| v]]$ where g is a state of this environment, e is a state of a shared memory environment, and the only rule used is for the transition: $u \xrightarrow{\delta} u' \vdash g[u] \xrightarrow{\delta} g[u']$.

The idea of a two-level insertion can be used to model distributed and shared memory in the following way. Let $R = R_1 \cup R_2$ be divided into two non-intersecting parts (external and internal memories correspondingly). Let C_1 be the set of actions that change only the values of R_1 (but can use the values of R_2). Let C_2 be the set of statements and conditions that change and use only R_2. Generalize the closure environment in the following way:

$$\frac{e[u] \xrightarrow{d} e'[u'],\ d \in C_1}{g[e[u]] \xrightarrow{d'} g[e'[u']]}$$

where d' is the result of substituting the values of R_2 into d. Now closed environments over R_2 can be inserted into the shared memory environment over R_1:

$$e[g[u_1] \| g[u_2] \| \dots]$$

and we obtain a two-level system with shared memory R_1 and distributed memory R_2. This construction can be iterated to obtain multilevel systems and enriched by message passing.

The importance of these constructions for applications is that most problems of proving equivalence, equivalent transformations and proving properties of distributed programs are reduced to the corresponding problems for behavior transformations that have the structure of behavior algebra.

3.4 Head insertion

Again we start with the special case of head insertion. It is defined by two sets of transition equations:

$$G_{i,a}(X)[a.y] \xrightarrow{d} G'_{i,a}(X)[y], \quad i \in I(a), \quad d \in D_{i,a} \subseteq C \qquad \text{(interaction)},$$
$$H_j(X)[y] \xrightarrow{c} H'_j(X)[y], \quad j \in J, \quad c \in C_j \subseteq C \qquad \text{(environment move)},$$

and the function $\varphi_\varepsilon(u) = \varepsilon[u]$ satisfying conditions (3.5). Here $G_i(X)$, $G'_{i,a}(X)$, $H_j(X)$, $H'_j(X) \subseteq F^\infty_{\text{fin}}(C, X)$ and y is a variable running over the agent behaviors. It is easy to prove that one-step insertion is a special case of head insertion. Note that transition rules for head insertion are not independent. An insertion function defined by these rules must be continuous. Corresponding conditions can be derived from the basic definitions.

Reduction of general case

Two environment states e and e' are called *insertion equivalent* if for all $u \in F(A)$ $e[u] = e'[u]$. This definition is also valid if e and e' are the states of two different environments, E and E', over the same set of agent actions. Environments E and E' are called *insertion equivalent* if for all $e \in E$ there exists $e' \in E'$ such that e and e' are equivalent and vice versa.

3.7 Theorem *For each head insertion environment of the general case there exists an equivalent head insertion environment of the special case.*

Proof The transitions of the general case have the form

$$G(X)[a.y] \xrightarrow{d} G'(X \cup Z)\sigma \tag{3.7}$$

where $\sigma = \{z_i := f_i(X)[g_i(y)] \mid z_i \in Z,\ i \in I\}$, $f_i(X) \in F^\infty_{\text{fin}}(C, X)$, $g_i(y) \in F^\infty_{\text{fin}}(A, \{y\})$, or

$$G(X)[y] \xrightarrow{d} G'(X \cup Z)\sigma \tag{3.8}$$

with the same description of σ. Introduce a new environment E' in the following way. The states of this environment (represented as a transition system) are the states of E and the insertion expressions are $\overline{e}[u]$ where $e = f\sigma$, $f \in F^\infty_{\text{fin}}(A, Z)$, $\sigma = \{z_i := f_i[g_i] \mid z_i \in Z,\ f_i \in F(C),\ g_i \in F^\infty_{\text{fin}}(A, \{\overline{y}\}),\ i \in I\}$. The symbol $\overline{y}$ is called suspended agent behavior and the insertion expressions are considered up to identity $\overline{e[\overline{y}]}[u] = e[u]$.

Define a new insertion function so that if there is a transition rule (3.7) in E, then there is a rule $G(X)[a.y] \xrightarrow{d} \overline{G'(X \cup Z)\sigma'}[y]$ in E' where $\sigma' = \sigma\{y := \overline{y}\}$ and if there is a transition rule (3.8) in E then there is a rule $G(X)[y] \xrightarrow{d} \overline{G'(X \cup Z)\sigma'}[y]$ in E' with the same σ'. Add also the rule: if $e \xrightarrow{c} e'$ in E then $\overline{e}[u] \xrightarrow{c} \overline{e'}[u]$ in E'. The termination insertion function φ_ε is derived from those transition rules that have ε as the left hand side. The equivalence of the two environments follows from their bisimilarity as transition systems. □

Reduction to one-step insertion

3.8 Theorem *For each head insertion environment there exists a one-step insertion environment equivalent to it.*

Proof Let E be a special case of a head insertion environment with the set X common to all insertion identities (both assumptions do not influence generality). Define E' as an environment with the set of actions $C' = F_m^\infty(C, X)$. Define mapping $\gamma : E \to F(C')$ so that

$$\begin{aligned}\gamma(e) &= \gamma_1(e) + \gamma_2(e),\\ \gamma_1(e) &= \sum_{e=G_{i,a}(X)\sigma+e'} G_{i,a}(X).\gamma(G'_{i,a}(X)\sigma),\\ \gamma_2(e) &= \sum_{e=H_i(X)\sigma+e'} H_i(X).\gamma(H'_i(X)\sigma).\end{aligned}$$

Define E' as the image of γ, the insertion function, by means of equations

$$\begin{aligned}D_1(a, G_{i,a}(X)) &= \{d | G_{i,a}(X)[a.y] \xrightarrow{d} G'_{i,a}(X)[y]\}\\ D_2(H_i(X)) &= \{d | H_i(X)[y] \xrightarrow{d} H'_i(X)[y]\}\end{aligned}$$

for a one-step insertion and a termination function derived from environment moves for E. Equivalence of two environments follows from the statement that $\{(e[u], \gamma(e)[u]) \mid e \in E\}$ is a bisimulation. □

3.5 Look-ahead insertion

A special case of look-ahead insertion is defined by a set of transition equations of the following type:

$$G(X)[H(Y)] \xrightarrow{d} G'(X)[H'(Y)].$$

Reduction of a general case to special one can be done in the same way as for head insertion. Constructions for head insertion reduction to one-step insertion can be used to reduce look-ahead insertion to $(1, m)$ insertion as well. Further reductions have not been considered yet.

If $A = C$, look-ahead can be generalized:

$$G(X)[H(Y)] \xrightarrow{d} G'(X, Y)[H'(X, Y)].$$

3.6 Enrichment transformation algebra by sequential and parallel composition

In this section a transformation algebra of a one-step environment is considered. Some one-step environments allow introducing sequential and parallel compositions of behavior transformations so that they are gomomorphically transferred from corresponding behavior algebras. The following conditional equation easily follows from the definition of one-step insertion:

$$\forall(i \in I)\ ([u_i] = [u]) \implies \left[\sum_{i\in I} a_i u_i\right] = \left[\left(\sum_{i\in I} a_i\right) u\right].$$

Behavior $\sum_{i\in I} a_i$ is called a one-step behavior. Therefore the following normal form can be proved for one-step insertion:

$$[u] = \sum_{i\in I}[p_i u_i] + [\varepsilon], \tag{3.9}$$

$[u_i] \neq [u_j]$ if $i \neq j$, $[p_i u_i] \neq [0]$, p_i are one-step behaviors.

For one-step behavior $p = \sum_{i\in I} a_i$ define $h(p) = \bigcup_{i\in I}\bigcup_{c\in C} D_1(c, a_i)$.

A one-step environment E is called *regular* if:

(1) For all $(e \in E,\ c \in C)$ $(c.e \in E)$;

(2) For all $(a \in A,\ c \in C)$ $(D_1(c, a) \cap D_2(c) = \emptyset)$;

(3) The function $\varphi_\varepsilon(u)$ does not depend on u and all termination equations are consequences of the definition of this function.

3.9 Proposition *For one-step behaviors p and q and a regular environment, $[p] = [q]$ if and only if $h(p) = h(q)$.*

Proof $c.e[p] = \sum_{d\in h(p)} d.e[\Delta] + \sum_{d\in D_2(c)} d.e[p]$. □

3.10 Proposition *For nonempty $h(p)$ and a regular environment there is only one transition $(c.e)[pu] \xrightarrow{d} e[u]$ labeled by $d \in h(p)$.*

Proof It follows from the definition of a regular environment. □

3.11 Proposition *When $h(p) = \emptyset$ and the environment is regular then $e[pu] = e[0]$.*

Proof If $h(p) = \emptyset$ then $(c.e)[pu] = \sum_{d\in D_2(c)} d.e[pu] = (c.e)[0]$, $\varepsilon[pu] = \varphi_\varepsilon(pu) = \varphi_\varepsilon(0)$. □

Using this proposition we can strengthen the normal form excluding in (3.9) one-step behavior coefficients p with $h(p) = \emptyset$ and termination constant $[\varepsilon]$ if $I \neq \emptyset$ (in the case $I = \emptyset$ a constant $[0]$ can be chosen as a termination constant).

3.12 Theorem *For a regular environment, normal form is defined uniquely up to commutativity of non-deterministic choice and equivalence of one-step behavior coefficients.*

Proof First prove that if $h(p) \neq \emptyset$ then $[pu] = [qv] \iff [p] = [q]$ and $[u] = [v]$. If $d \in h(p)$ then for some $a \in A$ and $c \in C$, $d \in D_1(c, a)$. Take arbitrary behavior $e \in E$. Since E is regular, $c.e \in E$ and $d \notin D_2(c)$. Therefore $(c.e)[pu] \xrightarrow{d} e[u]$. From the equivalence of pu and qv it follows that $(c.e)[qv] \xrightarrow{d} e[v]$ and this is the only transition from $c.e[qv]$ labeled by d. Symmetric reasoning gives $d \in h(q)$ implies $d \in h(p)$ and $[p] = [q]$. Therefore $[pu] = [qv] \to [p] = [q]$ and $[u] = [v]$. The inverse is evident.

Now let $[u] = [v]$ and $[u] = \sum_{i\in I}[p_i u_i]$, $[v] = \sum_{j\in J}[q_j v_j]$ be their normal forms (exclude trivial case when $I = \emptyset$). For each transition $(c.e)[u] \xrightarrow{d} e'$ there exists a transition $(c.e)[v] \xrightarrow{d} e''$ such that $e' = e''$. Select arbitrary $d \in h(p_i)$, c and e in the same way as in the previous part of the proof. Therefore $e' = e[u_i]$ and there exists only one j such that $e'' = e[v_j] = e[u_i]$ and $d \in h(q_j)$. From symmetry and the arbitrariness of e we have $[u_i] = [v_j]$, $[p_i] = [q_j]$ and $[p_i u_i] = [q_j v_j]$. □

3.13 Theorem *For regular environment,*

$$[u] = [u'],\ [v] = [v'] \implies [uv] = [u'v'].$$

Proof Simply prove that relation $\{(e[uv], e[u'v']) =|\ [u] = [u'],\ [v] = [v']\}$, defined on insertion expressions as states, is a bisimilarity. Use normal forms to compute transitions. □

Parallel composition does not in general have a congruence property. To find the condition when it does, let us extend the combination of actions to one-step behaviors assuming that

$$p \times q = \sum_{p=a+p',\ q=b+q'} a \times b.$$

The equivalence of one-step behaviors is a congruence if $h(p) = h(q) \implies h(p \times r) = h(q \times r)$.

3.14 Theorem *Let E be a regular one-step environment and the equivalence of one-step behaviors be a congruence. Then $[u] = [u'] \wedge [v] = [v']$ Longrightarrow $[u \| v] = [u' \| v']$.*

Proof As in the previous theorem we prove that the relation $\{(e[u\|v],\ e[u'\|v'])|[u] = [u'],\ [v] = [v']\}$ defined on the set of insertion expressions is a bisimilarity. To compute transitions, normal forms for the representation of behavior transformations must be used as well as the algebraic representation of parallel composition:

$$u \| v = u \times v + u \lfloor\!\lfloor v + v \lfloor\!\lfloor u.$$

□

4 Application to automatic theorem proving

The theory of interaction of agents and environments can be used as a theoretical foundation for a new programming paradigm called *insertion programming* [17]. The methodology of this paradigm includes the development of an environment with an insertion function as a basis for subject domain formalization and writing insertion programs as agents to be inserted into this environment. The semantics of behavior transformations is a theoretical basis for understanding insertion programs, their verification and transformations. In this section an example of the development of an insertion program for interactive theorem proving is considered. The program is based on the evidence algorithm of V. M. Glushkov that has a long history [7] and recently has been redesigned in the scope of insertion programming. According to the present time classification evidence algorithm is related to some sort of tableau method with sequent calculus and is oriented to the formalization of natural mathematical reasoning. We restrict ourselves to consider only first order predicate calculus. However in the implementation it is possibles to integrate the predicate calculus with applied theories and higher order functionals.

4.1 Calculus for interactive evidence algorithm

The development of an insertion program for the evidence algorithm starts with its specification by two calculi: the calculus of conditional sequents and the calculus of auxiliary goals. The formulas of the first one are

$$(X, s, w, (u_1 \Rightarrow v_1) \wedge (u_2 \Rightarrow v_2) \wedge \dots)$$

where, $u_1, v_1, u_2, v_2, \ldots$ are first order formulas; other symbols will be explained later.

All free variables occurring in the formulas are of two classes: fixed and unknown. The first ones are obtained by deleting universal quantifiers, the second ones by deleting the existential quantifiers. The expressions of a type $(u_i \Rightarrow v_i)$ are called ordinary sequents (u_i are called assumptions, v_i goals). Symbol w denotes a conjunction of literals, used as an assumption common to all sequents. Symbol s represents a partially ordered set of all free variables occurring in the formula, where partial order corresponds to the order of quantifier deletion (when quantifiers are deleted from different independent formulas new variables are not ordered). It is used to define dependencies between variables. The values of unknowns can depend on variables which only appear before them. A symbol X denotes substitution, partially defined function from unknowns to their values (terms). All logical formulas and terms with interpreted functional symbols and conditional sequents are considered up to some equivalence (associativity and commutativity of logical connectives, deMorgan identities and other Boolean identities excluding distributivity).

The formulas of the calculus of auxiliary goals are:

$$\mathrm{aux}(s, v, u \Rightarrow z, Q)$$

where s is a partial order on variables, v and u are logical formulas, and Q is a conjunction of sequents. The inference rules of the calculus of conditional sequents define the backward inference: from goal to axioms. They reduce the proof of the conjunction of sequents to the proof of each of them and the proof of an ordinary sequent to the proof of an ordinary sequent with literal as a goal. If the reduced sequent has a form $(X, s, w, u \Rightarrow z)$, where z is a literal then the rule of auxiliary goal is used at the next step:

$$\frac{\mathrm{aux}(s, 1, w \wedge u \Rightarrow z, 1) \vdash \mathrm{aux}(t, v, x \wedge y \Rightarrow z, P)}{(X, s, w, u \Rightarrow z) \vdash (Y, t, w \wedge \neg z, P)}.$$

In this rule z and x are unifiable literals, Y is the most general unifier extending X, and P is a conjunction of ordinary sequents obtained as an auxiliary goal in the calculus of auxiliary goals. Proving this conjunction is sufficient to prove $(X, s, w, u \Rightarrow z)$. The rule is applicable only if the substitution Y is consistent with the partial order t.

The axioms of the calculus of conditional sequents are:

$$(X, s, w, u \Rightarrow 1);$$
$$(X, s, w, 0 \Rightarrow u);$$
$$(X, s, 0, Q);$$
$$(X, s, w, 1).$$

The inference rules are:

$$\frac{(X, s, w, F) \vdash (X', s', w', F')}{(X, s, w, F \wedge H) \vdash (X', s', w, H)},$$

applied only when (X', s', w', F') is an axiom. This rule is called the sequent conjunction rule.

$$(X, s, w, u \Rightarrow 0) \vdash (X, s, w, 1 \Rightarrow \neg u)$$
$$(X, s, w, u \Rightarrow x \wedge y) \vdash (X, s, w, (u \Rightarrow x) \wedge (u \Rightarrow y))$$
$$(X, s, w, u \Rightarrow x \vee y) \vdash (X, s, w, \neg x \wedge u \Rightarrow y)$$

This rule can be applied in two ways permuting x and y because of commutativity of disjunction.

$$(X, s, w, u \Rightarrow \exists\, xp) \vdash (X, \mathrm{addv}(s, y), w, \neg(\exists\, xp) \wedge u \Rightarrow \mathrm{lsub}(p, x := y))$$
$$(X, s, w, u \Rightarrow \exists\, x(z)p) \vdash (X, \mathrm{addv}(s, (z, y)), w, \neg(\exists\, xp) \wedge u \Rightarrow \mathrm{lsub}(p, x := y))$$
$$(X, s, w, u \Rightarrow \forall\, xp) \vdash (X, \mathrm{addv}(s, a), w, u \Rightarrow \mathrm{lsub}(p, x := a))$$
$$(X, s, w, u \Rightarrow \forall\, x(z)p) \vdash (X, \mathrm{addv}(s, (z, a)), w, u \Rightarrow \mathrm{lsub}(p, x := a))$$

In these rules y is a new unknown and a a new fixed variable. The function $\mathrm{lsub}(p, x := z)$ substitutes z into p instead of all free occurrences of x. At the same time it joins z to all variables in the outermost occurrences of quantifiers. Therefore a formula $\exists\, x(z)p$, for instance, appears just after deleting some quantifier and introducing a new variable z. The function $\mathrm{addv}(s, y)$ adds a new element, y, to s without ordering it with other elements of s, and $\mathrm{addv}(s, (z, y))$ adds y, ordering it after z.

The rules of the calculus of auxiliary goals are the following:

$$\mathrm{aux}(s, v, x \wedge y \Rightarrow z, P) \vdash \mathrm{aux}(s, v \wedge y, x \Rightarrow z, P);$$
$$\mathrm{aux}(s, v, x \vee y \Rightarrow z, P) \vdash \mathrm{aux}(s, v, x \Rightarrow z, (v \Rightarrow \neg y) \wedge P);$$
$$\mathrm{aux}(s, v, \exists\, xp \Rightarrow z, P) \vdash \mathrm{aux}(\mathrm{addv}(s, a), v, \mathrm{lsub}(p, x := a) \Rightarrow z, P);$$
$$\mathrm{aux}(s, v, \exists\, x(y)p \Rightarrow z, P) \vdash \mathrm{aux}(\mathrm{addv}(s, (y, a)), v, \mathrm{lsub}(p, x := a) \Rightarrow z, P);$$
$$\mathrm{aux}(s, v, \forall\, xp \Rightarrow z, P) \vdash \mathrm{aux}(\mathrm{addv}(s, u), v \wedge \forall xp, \mathrm{lsub}(p, x := u) \Rightarrow z, P);$$
$$\mathrm{aux}(s, v, \forall\, x(y)p \Rightarrow z, P) \vdash \mathrm{aux}(\mathrm{addv}(s, (y, u)), v \wedge \forall\, x(y)p, \mathrm{lsub}(p, x := u) \Rightarrow z, P).$$

Here a is a new fixed varaible and u is a new unknown as in the calculus of conditional sequents.

4.2 A transition system for the calculus

Each of two calculi can be considered as a non-deterministic transition system. For this purpose the sequent conjunction rule must be split into ordinary rules. First we introduce the extended conditional sequent as a sequence $C_1; C_2; \ldots$ of conditional sequents and the following new rules:

$$(X, s, w, H_1 \wedge H_2) \vdash ((X, s, w, H_1); (X, s, w, H_2));$$
$$((X, s, w, H_1 \wedge H_2); P) \vdash ((X, s, w, H_1); (X, s, w, H_2); P);$$
$$\frac{C \vdash C'}{(C; P) \vdash (C'; P)}.$$

For axioms (X', s', w', F) the ruless are:

$$((X', s', w', F); (X, s, w, H)) \vdash (X', s', w, H);$$
$$((X', s', w', F); (X, s, w, H); P) \vdash ((X', s', w, H); P).$$

Futhermore the calculus of auxiliary goals always terminates (each inference is finite). Therefore the rule of auxiliary goals can be considered as a one-step rule of the calculus of

conditional sequents. Now the transitions of a transition system are transitions of this calculus. Each transition corresponds to some inference rule. The states of successful termination are axioms and a formula P is valid if aqnd only if one of the successful termination states is reachable from the initial state $(X_0, s_0, 1, 1 \Rightarrow P)$.

Let us label the transitions by actions. Each action is a message on which a rule has been applied in the corresponding transition. For example a transition corresponding to the rule of deleting a universal quantifier is:

$$(X, s, w, u \Rightarrow \forall\, xp) \xrightarrow{\text{mes}} (\text{addv}(s, a), w, u \Rightarrow \text{lsub}(p, x := a))$$

where a message "To prove a statement $\forall\, xp$ let us consider an arbitrary element a and prove q" is used as an action mes where $q = \text{lsub}(p, x := a)$. Other rules can be labeled in a similar way. Proving a statement T is therefore reduced to finding the trace which labels the transition of a system from the initial state $(\emptyset, \emptyset, 1, 1 \Rightarrow T)$ to the state corresponding to one of the axioms. If the axioms are defined as the states of successful termination, the problem is to find the trace from the initial state to one of the successfully terminated states. The sequence of messages corresponding to this trace is a text of a proof of a statement T.

This form of an evidence algorithm representation can be implemented in the system of insertion programming using a trivial environment which allows arbitrary behavior of the inserted agent. In automatic mode a system is looking for a trace with successful termination, if possible, and prints the proof when the trace is found. In interactive mode a system addresses to a user each time when it is necessary to make a non-deterministic choice. A user is offered several options to choose an inference rule and a user makes this choice. It is possible to go back and jump to other brunches. An environment which provides such possibilities is a proof system based on an evidence algorithm.

4.3 Decomposition of the transition system

A more interesting implementation of the evidence algorithm can be obtained if a state of a calculus is split into an environment and an agent inserted into this environment. This splitting is very natural if a substitution, partial order, conjunction of literals, or assumption of a current sequent is considered as a state of the environment and the goal of the current sequent, as well as all other sequents, is considered as an agent. Moreover it is possible to change the conjunction of all other sequents to a sequential composition. Call this agent a formula agent. A formula agent is considered as a state of the transition system used for the representation of an agent. To compute the behavior of agents, a recursive unfolding function must be defined on the set of agent expressions. It is easy to extract the unfolding function and the transition relation for a formula agent from the inference rules. Actions produced by formula agent define necessary changes in the environment and are computed by the insertion function. For example a formula agent $(\texttt{prove}\ \forall\, xp; P)$ is unfolded according to its recursive definition to an agent expression $Q = \texttt{fresh}\ C(\forall\, xp).P$, where $\texttt{fresh}\ C(\forall\, xp)$ is an action which substitutes a new fixed variable for p. This substitution is performed by the insertion function with transition

$$e[Q] \xrightarrow{\text{mes}} e'[\texttt{prove}\ \text{lsub}(p, x := a).P],$$

where mes is a message considered before and the transition from e to e' corresponds to the generation of a new fixed a.

In a general extended conditional sequent

$$((X, s, w, u \Rightarrow v); P)$$

is decomposed into an environment state and a sequential composition of agents. The environment state is

$$(X, s, w, u).$$

Initially $(\emptyset, \emptyset, 1, 1)$. The main types of agent expressions are:

- `prove` v; v is a simple sequent or formula,
- `prove` $(H_1 \wedge H_2 \wedge \dots)$; H_i are simple sequents,
- `block` P; P is an agent,
- `end_block` w; w is a conjunction of literals.

The following equations define unfolding and insertion function for blocks.

- `prove` $(H_1 \wedge H_2 \wedge \dots) =$ (`prove` H_1; `prove` $H_2 \wedge \dots$);
- `prove` $(u \Rightarrow v) =$ `block` (`Let` u.(`ask` $0.m_1 +$ `ask` $1.m_2$.`prove` v).m_3, where
  - $m_1 =$ `mes` ($\neg u$ `is evident by contradiction`);
  - $m_2 =$ `mes` (`prove` $u \Rightarrow v$);
  - $m_3 =$ `mes` (`sequent proved`);
- e[`block` $P.Q$] = `mes`(`begin`).e[P; `end_block` $w; Q$], where $e = (X, s, w, F)$;
- e[`end_block` w'] = `mes`(`end`).$e'[\Delta]$, where $e' = (X, s, w', 1)$ if $e = (X, s, w, u)$.

Actions are easily recognized by dots following them. The meaning of actions `mes` x and `block` P is clear from these definitions. Other actions will be explained later.

4.3.1 Unfolding conjunction and disjunction

In this section unfolding rules for conjunction and disjunction are explained as well as some auxiliary rules.

```
prove (u ⇒ 0) = prove (¬u);
prove (x ∧ y) = (prove (x); prove (y));
prove (x ∨ y) = block(
   Let (¬x).(
      ask 0.mes (x is evident by contradiction)
       +
      ask 1.mes(To prove x ∨ y let ¬x, prove y).
      prove (y)
   ); mes (disjunction proved)
) + (
   Let ¬y.(
```

```
        ask 0.mes(y is evident by contradiction)
        +
        ask 1.mes(To prove x ∨ y let ¬y, prove x).
        prove x
    );mes(disjunction proved)
)
```

4.3.2 Unfolding quantifiers

In these definitions e' is a state of an environment after generating a new unknown (`fresh` V) or fixed (`fresh` C) variable y.

```
prove ∃xp = (fresh V prove ∃xp).Δ;
prove ∀xp = (fresh C prove ∀xp).Δ;
e[fresh V q] = e'[get_fresh y q];
e[fresh C q] = e'[get_fresh y q];
get_fresh y prove ∃xp = mes(
   To prove ∃xp find x = y such that p
).block(
   Let ¬∃xp.
   prove lsub(p, x := y);
   mes(existence proved)
);
get_fresh y prove ∀xp = mes(
   Prove ∀xp. Let y is arbitrary constant.
).(
   prove lsub(p, x := y);
   mes(forall proved)
)
```

4.3.3 An auxiliary goal

The calculus of the auxiliary goal is implemented on the level of the environment. It is hidden from an external observer who can only see the result, that is the auxiliary goal represented by a corresponding message.

```
prove z = mes(To prove z find auxiliary goal).start_aux z;
e[start_aux z] = e'[prove_aux(z, Q)];
prove_aux(z, 1) = mes(z is evident);
prove_aux(z, Q) = mes(auxiliary goal is Q).prove Q.
```

Note that the rules for insertion of formula agents can be interpreted as one step insertions except for the rules for `start_aux` and `fresh`. Both can be interpreted as an instantiation of the rule:

$$e\frac{\xrightarrow{a} e'[v],\ u \xrightarrow{a} [u']}{e[u] \xrightarrow{a} e'[v; u']}$$

In the case of sequential insertion this rule can also be considered as a one-step insertion rule because in this case $e'[v; u'] = (e'[v])[u]$.

The last step in the development of an insertion program is verification. The correctness of the specification calculi is proved as a sound and completeness theorem comparing it with algorithms based on advanced tableau methods. After decomposition we should prove the bisimilarity of the corresponding initial states of two systems. We can also improve the insertion program. In this case we should do optimization preserving insertion equivalence of agents.

4.4 A proving machine

Formula agent actions can be considered as instructions of a proving machine representing an environment for such kind of agents. Here are some of the main instructions of the proving machine used for the development of the evidence algorithm kernel.

- `Let` <formula> adds the formula to assumptions in the environment. It is used for example for the unfolding sequent: `prove` $(u \Rightarrow v) =$ `Let` u.`prove` v or for the unfolding disjunction

  $$\texttt{prove}\ (u \vee v) = \texttt{Let}\ \neg v.\texttt{prove}\ v + \texttt{Let}\ \neg v.\texttt{prove}\ u.$$

  An essential reconstruction of the environment is performed each time a new assumption is added to the environment. Formulas are simplified, conjunctive literals are distinguished, substitution is applied etc. Moreover assumptions or a literal conjunction can be simplified up to 0 (false).

- `tell` <literal> adds a literal to conjunction of literals without reconstruction of the environment.

- `ask` 0 checks inconsistency of the environment state (0 in assumptions or in literals).

- `ask` 1 checks that there is no explicit inconsistency.

- `start_aux` <literal> starts the calculus of auxiliary goals:

  $$e[\texttt{start\_aux}\ p.Q] = \texttt{Let}\ \neg p.(\texttt{prove}\ P_1 + \texttt{prove}\ P_2 + \cdots)$$

  where $P_1, P_2, \ldots$ are auxiliary goals (conjunctions of sequents) extracted from assumptions according to the calculus of auxiliary goals (actually the real relations are slightly more complex, because they include messages about the proof development and they anticipate the case when there are no auxiliary goals at all).

- `fresh V` <formula> substitutes a new unknown into the formula.

- `fresh C` <formula> substitutes a new fixed variable into the formula.

- `solve` <equation> is used for solving equations in the case when equality is used.

- `block` <program> localizes all assumptions within the block. It is used for example when conjunction is proved.

- `Mesg` <text> inserts the printing of messages at the different stages of the proof search.

- `start_lkb` <literal>. It is used to address the local knowledge base for extracting auxiliary goals from assumptions presented in this base. The local knowledge base is prepared in advance according to requirements of the subject domain where the proof is searched for.

The advantages of a proving machine (or more generally an insertion machine for other environment structures) is that it is possible to change in a wide area the algorithms of a proof search without changing the structure of the environment by varying the recursive unfolding rules of formula agents. Moreover the environment itself can be extended by introducing new instructions into the instruction set and adding new components to an environment state.

To run a program on a proving machine some higher level environment should be used to implement back-tracking. Such an environment can work in two modes: interactive and automatic. The following equations demonstrate the approximate meaning of these two modes.

The interactive mode:

$$e[a_1.u_1 + a_2.u_2 + \dots] = \texttt{mes}(\texttt{select}\ a_1, a_2, \dots).(a_1.e'[u_1] + a_2.e'[u_2] + \dots) + \texttt{back}.e''[u] + \Delta.$$

The automatic mode:

$$e[a_1.u_1 + a_2.u_2 + \dots] = e'[(a_1.u_1); (\texttt{return if fail or stop}); (a_2.u_2 + \dots)].$$

This is a depth first search. It works only with restrictions on the admissible depth. A breadth first search is more complex.

5 Conclusions

A model of interaction of agents and environments has been introduced and studied. The first two sections extend and generalize results previously obtained in [15]. The algebra of behavior transformations is a good mathematical basis for the description and explanation of agent behavior restricted by the environment in which it is inserted. System behavior has two dimensions. The first one is a branching time which defines the height of a behavior tree and can be infinite (but no more than countable). The second one is a non-deterministic branching at a given point. Our construction of a complete behavior algebra used for the characterization of bisimilarity allows for branching of arbitrary cardinality.

The restriction of the insertion function to be continuous is too broad and we consider a more restricted classes of equationally defined insertion functions and one-step insertions to which more general head insertion is reduced. The question about reducing or restricting look-ahead insertion is open at this moment. At the same time the algebra of behavior transformations based on regular one-step insertion can be enriched by sequential and parallel compositions.

The algebra of behavior transformations is the mathematical foundation of a new programming paradigm: insertion programming. It has been successfully applied for automatic theorem proving and verification of distributed software systems. In [16] operational semantics of timed MSC (specification language of Message Sequencing Charts) has been defined on the basis of behavior transformations. This semantics has been used for the development of tools for verification of distributed systems.

References

[1] S. Abramsky, A domain equation for bisimulation, *Information and Computation* **92** (2) (1991), 161–218.

[2] L. Aceto, W. Fokking, and C. Verhoef, Structural operational semantics, in: *Handbook of Process Algebra* (J. A. Bergstra, A. Ponce, and S. A. Smolka, eds.), North-Holland, 2001.

[3] P. Aszel and N. Mendler, A final coalgebra theorem, in: *LNCS 389*, Springer-Verlag, 1989.

[4] P. Azcel, J. Adamek, S. Milius, and J. Velebil, Infinite trees and comletely iterative theories: a coalgebraic view, *TCS* **300** (1–3) (2003), 1–45.

[5] J. A. Bergstra and J. W. Klop, Process algebra for synchronous communications, *Information and Control* **60** (1/3) (1984), 109–137.

[6] J. A. Bergstra, A. Ponce, and S. A. Smolka, eds., *Handbook of Process Algebra*, North-Holland, 2001.

[7] A. Degtyarev, J. Kapitonova, A. Letichevsky, A. Lyaletsky, and M. Morokhovets, Evidence algorithm and problems of representation and processing of computer mathematical knowledge, *Kibernetika and System Analysis* (6) (1999), 9–17.

[8] R. J. Glabbeek, The linear time—branching time spectrum i. the semantics of concrete, sequential processes, in: *Handbook of Process Algebra* (J. A. Bergstra, A. Ponce, and S. A. Smolka, eds.), North-Holland, 2001.

[9] V. M. Glushkov, Automata theory and formal transformations of microprograms, *Kibernetika* (5).

[10] V. M. Glushkov and A. A. Letichevsky, Theory of algorithms and descrete processors, in: *Advances in Information Systems Science* (J. T. Tou, ed.), vol. 1, Plenum Press, 1969.

[11] J. A. Goguen, S. W. Thetcher, E. G. Wagner, and J. B. Write, Initial algebra semantics and continuous algebras, *J. ACM* (24) (1977), 68–95.

[12] M. Hennessy and R. Milner, On observing nondeterminism and concurrency, in: *Proc. ICALP'80, LNCS 85*, Springer-Verlag, 1980.

[13] C. A. R. Hoare, *Communicating Sequential Processes*, Prentice Hall, 1985.

[14] A. Letichevsky, Algebras with approximation and recursive data structures, *Kibernetika and System Analysis* (5) (1987), 32–37.

[15] A. Letichevsky and D. Gilbert, Interaction of agents and environments, in: *Resent trends in Algebraic Development technique, LNCS 1827* (D. Bert and C. Choppy, eds.), Springer-Verlag, 1999.

[16] A. Letichevsky, J. Kapitonova, V. Kotlyarov, A. Letichevsky, Jr., and V. Volkov, Semantics of timed msc language, *Kibernetika and System Analysis* (4).

[17] A. Letichevsky, J. Kapitonova, V. Volkov, V. Vyshemirskii, and A. Letichevsky, Jr., Insertion programming, *Kibernetika and System Analysis* (1) (2003), 19–32.

[18] J. Meseguer, Conditional rewriting logic as a unified model of concurrency, *Theoretical Computer Science* **96** (1992), 73–155.

[19] R. Milner, *A Calculus of Communicating Systems*, vol. 92 of *Lecture Notes in Computer Science*, Springer-Verlag, 1980.

[20] R. Milner, *Communication and Concurrency*, Prentice Hall, 1989.

[21] R. Milner, The polyadic π-calculus: a tutorial, Tech. Rep. ECS–LFCS–91–180, Laboratory for Foundations of Computer Science, Department of Computer Science, University of Edinburgh, UK (1991).

[22] D. Park, Concurrency and automata on infinite sequences, in: *LNCS 104*, Springer-Verlag, 1981.

[23] M. Roggenbach and M. Majster-Cederbaum, Towards a unified view of bisimulation: a comparative study, *TCS* **238** (2000), 81–130.

[24] J. Rutten, Coalgebras and systems, *TCS* **249**.

Congruence modular varieties: commutator theory and its uses

Ralph MCKENZIE

Department of Mathematics
Vanderbilt University
Nashville, TN 37240
USA

John SNOW

Department of Mathematics
Concordia University
Seward, NE 68434
USA

Abstract

We present the basic theory of commutators of congruences in congruence modular varieties (or equationally defined classes) of algebras. The theory we present was first introduced to the mathematical world in a 1976 monograph of J. D. H. Smith, devoted to varieties with permuting congruences. It was extended to congruence modular varieties in a 1979 paper by J. Hagemann and C. Herrmann, and has since been elaborated into an impressive machinery for attacking diverse problems in the domain of congruence modular varieties. Three notable applications of this commutator theory are presented in detail, and others are described.

1 Introduction

The commutator in group theory is a natural operation defined on the lattice of normal subgroups of any group which plays a basic role in the definition and study of solvable and Abelian groups. This commutator has a companion operation in the theory of rings, defined on any lattice of ideals. These two operations share many common properties, including the ability to capture the notion of Abelian-ness.

In [43], J. D. H. Smith used category theory to extend structural properties of groups and rings to varieties with permuting congruences. In doing so, he laid the framework for generalizing the commutator from groups and rings to an operation on the congruence lattices of algebras in congruence permutable varieties.

J. Hagemann and C. Herrmann in [19] extended some of Smith's results to congruence modular varieties. Their techniques include subtle and difficult calculations in $\operatorname{Con}\mathbf{A}$, $\operatorname{Con}\mathbf{A}^2$, and $\operatorname{Con}\mathbf{A}^3$ using modular arithmetic. In their work, they mentioned the term

V. B. Kudryavtsev and I. G. Rosenberg (eds.),
Structural Theory of Automata, Semigroups, and Universal Algebra, 273–329.

condition which would later become the basis for what seems to be the most useful definition of the commutator in congruence modular varieties. H.-P. Gumm [17] further extended these structural results for congruence modular varieties by viewing the structure imposed on algebras by congruence relations geometrically. R. Freese and R. McKenzie [12] developed the commutator for congruence modular varieties based on the term condition mentioned by Hagemann and Herrmann.

In this manuscript, we give a gentle introduction to the commutator theory presented in [12]. We then present several applications of the commutator along with some open problems which may involve commutator theory. In Section 2 we review the classical commutator in groups and demonstrate how the term condition arises naturally in this environment. In Section 3 we lay out some basic notation which will be pervasive throughout the manuscript. In Section 4 we use the notion of centrality to define the commutator and prove a few properties of the commutator which hold in any environment. In Section 5 we give examples of the commutator in some familiar environments including rings, lattices, and modules. In Section 6 we give the classical Maltsev type characterizations of congruence permutability, distributivity, and modularity due to Maltsev, Jónsson, and Day. These characterizations are exploited heavily in the development of commutator theory for congruence modular varieties. In Section 7 we use Jónsson's characterization of congruence distributivity to prove that in a congruence distributive variety the commutator is nothing other than congruence intersection. In Section 8 we extend all of the properties of the group commutator mentioned in Section 2 to the commutator in congruence modular varieties. In Section 9 we prove the Fundamental Theorem of Abelian Algebras that every Abelian algebra (in a congruence modular variety) is affine (polynomially equivalent to a module). In Section 10 we extend the ideas of solvability and nilpotence using the commutator. We prove that every nilpotent or solvable algebra in a congruence modular variety has a Maltsev term and use this to give some structural results about nilpotent algebras. In Section 11 on applications, we briefly discuss seven outstanding instances of basic problems that have been solved, for modular varieties, with the aid of commutator theory. In the following sections, we present three of these applications in detail: In Section 12 we prove that every finitely generated, residually small, congruence modular variety has a finite residual bound, and that such varieties are characterized by a commutator equation. In Section 13 we characterize directly representable varieties—i.e., those finitely generated varieties that possess only a finite number of non-isomorphic finite, directly indecomposable, algebras—and we characterize the larger family of finitely generated varieties whose spectrum is contained in a finitely generated monoid of positive integers. All of these varieties are shown to be congruence modular. In Section 14 we characterize the locally finite congruence modular varieties for which the function giving the number of non-isomorphic n-generated algebras is dominated by a polynomial in n. They are precisely the directly representable Abelian varieties. In Section 15 we survey some problems which either involve the commutator or for which there is evidence that the commutator might prove useful. We note that the results herein are not original. Excepting the results of Sections 13–14, almost all of them appear with proofs in [12].

2 The commutator in groups

In this section, we discuss the group commutator and some of its most basic properties, and we illustrate how the term condition arises naturally in this environment.

2.1 Definition Suppose that $\mathbf{G}$ is a group and M and N are normal subgroups of $\mathbf{G}$. The *group commutator* of M and N is defined as

$$[M, N] = \mathrm{Sg}_{\mathrm{G}}(\{m^{-1}n^{-1}mn : m \in M \text{ and } n \in N\}), \tag{2.1}$$

Suppose that $\mathbf{G}$ is a group, that M, N, and $\{N_i : i \in I\}$ are normal subgroups of $\mathbf{G}$, and that $f : \mathbf{G} \to \mathbf{H}$ is a surjective group homomorphism. Then the following properties of the group commutator are easy exercises in any first class on group theory.

(1) $[M, N] \subseteq M \cap N$.

(2) $f([M, N]) = [f(M), f(N)]$.

(3) $[M, N] = [N, M]$.

(4) $[M, \bigvee_{i\in I} N_i] = \bigvee_{i\in I}[M, N_i]$.

(5) For any normal subgroup K of $\mathbf{G}$ included in $M \cap N$, the elements of M/K commute with the elements of N/K if and only if $K \supseteq [M, N]$.

(6) $\mathbf{G}$ is Abelian if and only if $[G, G] = \{1\}$ (where 1 is the identity element of $\mathbf{G}$).

When we generalize the commutator later to congruences in a congruence modular variety, we will want an operation which will share these properties. Our operation will be defined from a generalization of the condition (5). From this definition, we will get (1), (3), and (4) directly and a slight modification of (2). We will take (6) to be our definition of Abelian.

The group commutator has one more property which is less transparent but which will also carry over to the generalization (in fact, it also could be used to define the modular commutator). This is

(7) The commutator operation is the largest binary operation defined across all normal subgroup lattices of all groups which satisfies conditions (1) and (5).

Suppose that $C(x, y)$ is another binary operation defined on the normal subgroup lattice of every group which satisfies (1) and (2). Let M and N be normal subgroups of a group $\mathbf{G}$. We will prove that $C(M, N) \subseteq [M, N]$. To do so, we need to define four subgroups of $\mathbf{G} \times \mathbf{G}$.

$$\mathbf{G}(M) = \{\langle x, y\rangle : x, y \in \mathbf{G} \text{ and } x^{-1}y \in M\} \tag{2.2}$$

$$\Delta = \{\langle x, y\rangle : x \in N, y \in G, \text{ and } x^{-1}y \in [M, N]\} \tag{2.3}$$

$$B = \{\langle x, 1\rangle : x \in [M, N]\} \tag{2.4}$$

$$M_1 = \{\langle x, 1\rangle : x \in M\}. \tag{2.5}$$

The subgroups Δ, B, and M_1 are normal subgroups of $\mathbf{G}(M)$. Let π be the projection of $\mathbf{G}(M)$ onto the first coordinate. Then the reader can check that

$$\pi(\mathbf{G}(M)) = \mathbf{G} \tag{2.6}$$

$$\pi(\Delta) = N \tag{2.7}$$

$$\pi(B) = [M, N] \tag{2.8}$$

$$\pi(M_1) = M. \tag{2.9}$$

From property (1) we see that $C(M_1, \Delta) \subseteq M_1 \cap \Delta \subseteq B$ and by (2)

$$C(M, N) = \pi(C(M_1, \Delta)) \subseteq \pi(B) = [M, N]. \tag{2.10}$$

We would now like to state a condition equivalent to (5) which will be the basis for our generalization of the commutator. Let $K = [M, N]$. Suppose that t is an $(n + m)$-ary group term. We will address here how t behaves in $\mathbf{G}/K$ when evaluated on elements of M/K and N/K. For convenience in the following calculations, we will write $\bar{g}$ for elements gK of $\mathbf{G}/K$. Suppose that $a_1, \ldots, a_n, b_1, \ldots, b_n \in M$ and $x_1, \ldots, x_m, y_1, \ldots, y_m \in N$ and that

$$t(\bar{a}_1, \ldots, \bar{a}_n, \bar{x}_1, \ldots, \bar{x}_m) = t(\bar{a}_1, \ldots, \bar{a}_n, \bar{y}_1, \ldots, \bar{y}_m). \tag{2.11}$$

Since $K = [M, N]$, we can permute some of the elements in this equation so that we have

$$t(\bar{a}_1, \ldots, \bar{a}_n, \bar{x}_1, \ldots, \bar{x}_m) = \bar{a}_{j_1}^{e_1} \bar{a}_{j_2}^{e_2} \cdots \bar{a}_{j_u}^{e_u} \bar{x}_{l_1}^{d_1} \bar{x}_{l_2}^{d_2} \cdots \bar{x}_{l_v}^{d_v} \tag{2.12}$$

where each e_i and each d_i is either 1 or -1. Similarly

$$t(\bar{a}_1, \ldots, \bar{a}_n, \bar{y}_1, \ldots, \bar{y}_m) = \bar{a}_{j_1}^{e_1} \bar{a}_{j_2}^{e_2} \cdots \bar{a}_{j_u}^{e_u} \bar{y}_{l_1}^{d_1} \bar{y}_{l_2}^{d_2} \cdots \bar{y}_{l_v}^{d_v}. \tag{2.13}$$

Combining these, we have

$$\bar{a}_{j_1}^{e_1} \bar{a}_{j_2}^{e_2} \cdots \bar{a}_{j_u}^{e_u} \bar{x}_{l_1}^{d_1} \bar{x}_{l_2}^{d_2} \cdots \bar{x}_{l_v}^{d_v} = \bar{a}_{j_1}^{e_1} \bar{a}_{j_2}^{e_2} \cdots \bar{a}_{j_u}^{e_u} \bar{y}_{l_1}^{d_1} \bar{y}_{l_2}^{d_2} \cdots \bar{y}_{l_v}^{d_v}. \tag{2.14}$$

Suitable cancellation and multiplication by b's now gives

$$\bar{b}_{j_1}^{e_1} \bar{b}_{j_2}^{e_2} \cdots \bar{b}_{j_u}^{e_u} \bar{x}_{l_1}^{d_1} \bar{x}_{l_2}^{d_2} \cdots \bar{x}_{l_v}^{d_v} = \bar{b}_{j_1}^{e_1} \bar{b}_{j_2}^{e_2} \cdots \bar{b}_{j_u}^{e_u} \bar{y}_{l_1}^{d_1} \bar{y}_{l_2}^{d_2} \cdots \bar{y}_{l_v}^{d_v}. \tag{2.15}$$

After commuting as before, we end up with the equality

$$t(\bar{b}_1, \ldots, \bar{b}_n, \bar{x}_1, \ldots, \bar{x}_m) = t(\bar{b}_1, \ldots, \bar{b}_n, \bar{y}_1, \ldots, \bar{y}_m). \tag{2.16}$$

We have proven the following property which will replace property (5) above.

(5') Suppose that t is an $(n + m)$-ary group term and that $a_1, \ldots, a_n, b_1, \ldots, b_n \in M$ and $x_1, \ldots, x_m, y_1, \ldots, y_m \in N$. Let $K = [M, N]$. If

$$t(a_1K, \ldots, a_nK, x_1K, \ldots, x_mK) = t(a_1K, \ldots, a_1K, y_1K, \ldots, y_mK) \tag{2.17}$$

then also

$$t(b_1K, \ldots, b_nK, x_1K, \ldots, x_mK) = t(b_1K, \ldots, b_1K, y_1K, \ldots, y_mK). \tag{2.18}$$

We will describe this situation by saying that $\mathbf{G}$ satisfies the M, N term condition modulo K or that M centralizes N modulo K. This term condition will be the basis of the modular commutator.

3 Notation

We assume that the reader is familiar with the basics of universal algebra, and we will usually use notation consistent with [37]. In this section, we emphasize a few key ideas and pieces of notation.

Generally, we will use plain text capital letters to refer to sets. We will use bold faced text to refer to algebras. Usually (but not always), the same letter will be used for the set and the algebra. For example, an algebra on a set A will almost always be called $\mathbf{A}$. We use script letters (such as $\mathcal{V}$) to refer to varieties, classes of varieties, and classes of algebras. For any algebra $\mathbf{A}$ in a variety $\mathcal{V}$ and any term $t(x_0, \ldots, x_n)$ of $\mathcal{V}$, it is customary to use a superscript to denote the term operation of $\mathbf{A}$ induced by t (that is, $t^{\mathbf{A}}(x_0, \ldots, x_n)$). In most of our proofs, the algebra will be understood, so we will often (usually) leave off the superscript to allow for cleaner notation. If $\mathbf{A}$ is an algebra, we will use bold faced lowercase letters to represent elements of direct powers of $\mathbf{A}$. For example, an element $\mathbf{a} \in A^n$ is a vector $\langle a_0, \ldots, a_{n-1} \rangle$. Notice that with this notation, we will always assume our subscripts begin at 0 and go to $n-1$. When applying an $(n+m)$-ary term t to a vector $\langle x_0, \ldots, x_{n-1}, y_0, \ldots, y_{m-1} \rangle$, it is often more convenient (and notationally cleaner) to write $t(\mathbf{x}, \mathbf{y})$.

For the subalgebra of $\mathbf{A}$ generated by a subset $X \subseteq \mathbf{A}$, we will write $\mathrm{Sg}_{\mathbf{A}}(X)$. For the subalgebra generated by elements $a_1, \ldots, a_n$, we may abuse notation and write $\mathrm{Sg}_{\mathbf{A}}(a_1, \ldots, a_n)$. Similarly, we use $\mathrm{Cg}_{\mathbf{A}}(X)$ for the congruence on $\mathbf{A}$ generated by a subset $X \subseteq A^2$. For the principal congruence generated by identifying elements a and b, we will write $\mathrm{Cg}_{\mathbf{A}}(a, b)$. In all cases, we may omit the subscripted $\mathbf{A}$ if the underlying algebra is understood. We will use $\operatorname{End} \mathbf{A}$ for the endomorphism monoid of $\mathbf{A}$.

Depending on context, there are three notations we may use to assert that two elements x and y are related by a binary relation α. These are

$$x \alpha y,$$
$$\langle x, y \rangle \in \alpha, \text{ and}$$
$$x \equiv y \pmod{\alpha}.$$

By a tolerance on an algebra $\mathbf{A}$, we mean a subalgebra of $\mathbf{A}^2$ which is reflexive and symmetric (but not necessarily transitive). We will use $\operatorname{Con} \mathbf{A}$ to represent the congruence lattice of $\mathbf{A}$ and $\operatorname{Tol} \mathbf{A}$ to represent the tolerance lattice of $\mathbf{A}$. If α is any binary relation then $\mathrm{Tr}(\alpha)$ will be the transitive closure of α. The universal relation on a set A will be denoted 1_A, and the identity relation will be denoted by 0_A.

If $\mathcal{V}$ is any variety, we will use the notation $\mathcal{V} \models_{\mathrm{Con}} \ldots$ to indicate that all congruences of all algebras in $\mathcal{V}$ satisfy the property $\ldots$. For example, $\mathcal{V} \models_{\mathrm{Con}} (\alpha \cap \beta \approx [\alpha, \beta])$ means that for every algebra $\mathbf{A} \in \mathcal{V}$ and for all $\alpha, \beta \in \operatorname{Con} \mathbf{A}$ the equality $\alpha \cap \beta = [\alpha, \beta]$ holds. Usually, $\approx$ will be used to represent the equality symbol of a first-order language, and $=$ will be used for a specific instance of equality.

Much of this manuscript will deal with congruence lattices which are modular or distributive. Therefore, we remind ourselves of the definitions of these properties and state some basic facts about them. The realization of the concept of a lattice as an independent algebraic object of interest and the formulation of the modular law dates back to Richard Dedekind [9].

3.1 Definition Let $\mathbf{L} = \langle L, \wedge, \vee \rangle$ be a lattice. $\mathbf{L}$ is *modular* if for all elements $a, b, c \in \mathbf{L}$

with $c \leq a$ the equality $a \wedge (b \vee c) = (a \wedge b) \vee c$ holds. $\mathbf{L}$ is *distributive* if for all elements $a, b, c \in L$ the equality $a \wedge (b \vee c) = (a \wedge b) \vee (a \wedge c)$ holds.

These characterizations of modularity and distributivity should be familiar.

3.2 Theorem *For any lattice* $\mathbf{L}$*, the following are equivalent.*

(1) $\mathbf{L}$ *is distributive.*

(2) $a \wedge (b \vee c) \leq (a \wedge b) \vee (a \wedge c)$ *for all* $a, b, c \in \mathbf{L}$.

(3) $a \vee (b \wedge c) = (a \vee b) \wedge (a \vee c)$ *for all* $a, b, c \in \mathbf{L}$.

(4) $\mathbf{L}$ *has no sublattice isomorphic to either* $\mathbf{N}_5$ *or* $\mathbf{M}_3$ (*see Fig.* 1).

3.3 Theorem *The following are equivalent for any lattice* $\mathbf{L}$.

(1) $\mathbf{L}$ *is modular.*

(2) *For any* $a, b, c \in L$ *if* $c \leq a$ *then* $a \wedge (b \vee c) \leq (a \wedge b) \vee c$.

(3) $((a \wedge c) \vee b) \wedge c = (a \wedge c) \vee (b \wedge c)$ *for all* $a, b, c \in L$.

(4) $\mathbf{L}$ *has no sublattice isomorphic to* $\mathbf{N}_5$ (*see Fig.* 1).

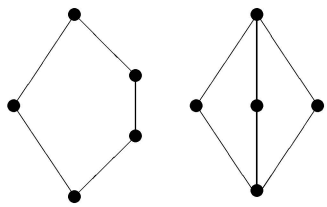


Figure 1: The lattices $\mathbf{N}_5$ (left) and $\mathbf{M}_3$ (right).

4 Centrality and the term condition commutator

4.1 Definition Suppose that α, β, and δ are congruences on an algebra $\mathbf{A}$. Then α *centralizes* β *modulo* δ (in symbols $C(\alpha, \beta; \delta)$) if for any $(m+n)$-ary term operation T of $\mathbf{A}$, for any $\mathbf{a}, \mathbf{b} \in A^m$ with $a_i \alpha b_i$ for all i, and for any $\mathbf{c}, \mathbf{d} \in A^n$ with $c_i \beta d_i$ for all i, the relation $T(\mathbf{a}, \mathbf{c}) \delta T(\mathbf{a}, \mathbf{d})$ holds if and only if $T(\mathbf{b}, \mathbf{c}) \delta T(\mathbf{b}, \mathbf{d})$. When $C(\alpha, \beta; \delta)$ holds, we will also say that $\mathbf{A}$ *satisfies the* α, β *term condition modulo* δ.

It will be convenient for us at times to view the elements of $\mathbf{A}^4$ as 2×2 matrices so that the 4-tuple $\langle x_0, x_1, x_2, x_3 \rangle$ corresponds to the matrix

$$\begin{pmatrix} x_0 & x_1 \\ x_2 & x_3 \end{pmatrix}.$$

4.2 Definition Suppose that α and β are congruences on an algebra $\mathbf{A}$. Define $M(\alpha, \beta)$ to be the subalgebra of $\mathbf{A}^4$ generated by all matrices either of the form

$$\begin{pmatrix} a & a \\ b & b \end{pmatrix}$$

where $a\alpha b$ or of the form

$$\begin{pmatrix} c & d \\ c & d \end{pmatrix}$$

where $c\beta d$.

It follows immediately from the definitions that

4.3 Lemma *For any congruence α, β, and δ on an algebra $\mathbf{A}$, $C(\alpha, \beta; \delta)$ holds if and only if for all $\left(\begin{smallmatrix} x & y \\ u & v \end{smallmatrix}\right) \in M(\alpha, \beta)$, $x\delta y \leftrightarrow u\delta v$.*

These basic properties of centrality hold without any additional assumptions about the underlying variety or algebra.

4.4 Lemma *Suppose that α, β, δ, $\{\alpha_i : i \in I\}$, and $\{\delta_j : j \in J\}$ are congruences on an algebra $\mathbf{A}$.*

(1) *If $C(\alpha_i, \beta; \delta)$ for all $i \in I$, then $C(\bigvee_{i\in I} \alpha_i, \beta; \delta)$.*

(2) *If $C(\alpha, \beta; \delta_j)$ for all $j \in J$, then $C(\alpha_j, \beta; \bigcap_{j\in J} \delta_j)$*

(3) *$C(\alpha, \beta; \alpha \cap \beta)$.*

Proof (1) Let $\gamma = \bigvee_{i\in I} \alpha_i$. Suppose that T is an $(n+m)$-ary term of $\mathbf{A}$ and that $\mathbf{a}, \mathbf{b} \in A^n$ and $\mathbf{x}, \mathbf{y} \in A^m$. Assume also that $a_t \gamma b_t$ for all t and $x_t \beta y_t$ for all t. There exist vectors $\mathbf{u}^1, \ldots, \mathbf{u}^l$ so that for each t

$$a_t = u_t^1 \alpha_{j_1} u_t^2 \alpha_{j_2} u_t^3 \ldots u_t^l = b_t$$

Suppose that $T(\mathbf{a}, \mathbf{x})\delta T(\mathbf{a}, \mathbf{y})$. We can prove by induction that for each $i = 1, \ldots, l$ the relation $T(\mathbf{u}^i, \mathbf{x})\delta T(\mathbf{u}^i, \mathbf{y})$ holds using $C(\alpha_{j_i}, \beta; \delta)$. It follows then that $T(\mathbf{b}, \mathbf{x})\delta T(\mathbf{b}, \mathbf{y})$.

(2) Suppose that $C(\alpha, \beta; \delta_j)$ for all $j \in J$. If $\left(\begin{smallmatrix} x & y \\ u & v \end{smallmatrix}\right) \in M(\alpha, \beta)$ is any matrix in $M(\alpha, \beta)$ so that $\langle x, y\rangle$ is in $\bigcap_{j\in J} \delta_j$, then $x\delta_j y$ for all j, so by centrality $u\delta_j v$ for all j. Hence $\langle u, v\rangle \in \bigcap_{j\in J} \delta_j$.

(3) Suppose that $\left(\begin{smallmatrix} x & y \\ u & v \end{smallmatrix}\right) \in M(\alpha, \beta)$ is any matrix in $M(\alpha, \beta)$. If $x(\alpha \cap \beta)y$, then $u\alpha x(\alpha \cap \beta)y\alpha v$ so $u\alpha v$. Since $u\beta v$ by assumption, it follows that $u(\alpha \cap \beta)v$. □

This lemma makes possible the following definition.

4.5 Definition Suppose that $\mathbf{A}$ is any algebra and $\alpha, \beta \in \operatorname{Con}\mathbf{A}$. The *commutator* of α and β is defined as $[\alpha, \beta] = \bigcap\{\delta \in \operatorname{Con}\mathbf{A} : C(\alpha, \beta; \delta)\}$.

This lemma is an immediate consequence of the fact that if $\alpha' \subseteq \alpha$ and $\beta' \subseteq \beta$ then $M(\alpha', \beta) \subseteq M(\alpha, \beta)$ and $M(\alpha, \beta') \subseteq M(\alpha, \beta)$.

4.6 Lemma [,] *is monotone in both variables.*

This lemma follows immediately from (3) of Lemma 4.4.

4.7 Lemma *Suppose that α and β are congruences on an algebra $\mathbf{A}$. Then $[\alpha, \beta] \leq \alpha \cap \beta$.*

We take the opportunity here to state our definition of what it means for an algebra, or a congruence, to be Abelian.

4.8 Definition An algebra $\mathbf{A}$ is *Abelian* if $\mathbf{A} \models [1_A, 1_A] = 0_A$. A congruence α on $\mathbf{A}$ is Abelian if $\mathbf{A} \models [\alpha, \alpha] = 0_A$.

By Lemma 4.4 (1), the following definition makes sense.

4.9 Definition Suppose that $\mathbf{A}$ is any algebra. Define the *center* of $\mathbf{A}$ to be the largest congruence $\zeta \in \operatorname{Con} \mathbf{A}$ so that $[\zeta, 1_A] = 0_A$. Denote the center of $\mathbf{A}$ as $\zeta_{\mathbf{A}}$.

From the definitions, it is clear that

4.10 Lemma *An algebra $\mathbf{A}$ is Abelian if and only if $\zeta_{\mathbf{A}} = 1_A$.*

We also have this useful universal substitution property of the center.

4.11 Lemma *Suppose $\mathbf{A}$ is any algebra. Then $\zeta_{\mathbf{A}}$ is the set of all $\langle x, y \rangle$ so that for all positive integers n, for all $(n+1)$-ary terms t of $\mathbf{A}$, and for all $\mathbf{a}, \mathbf{b} \in A^n$*

$$t(x, \mathbf{a}) = t(x, \mathbf{b}) \leftrightarrow t(y, \mathbf{a}) = t(y, \mathbf{b}).$$

Proof Let θ be the set of all $\langle x, y \rangle$ satisfying the conditions of the lemma. We will prove that $\theta = \zeta_{\mathbf{A}}$. Since $[\zeta_{\mathbf{A}}, 1_A] = 0_A$, it should be clear that $\zeta_{\mathbf{A}} \subseteq \theta$. We need only establish the reverse inclusion. To do so, we need to know that θ is a congruence on $\mathbf{A}$ and that $[\theta, 1_A] = 0_A$. It is easy to see that θ is an equivalence relation. To prove that it is a congruence, we prove that θ is closed under all unary polynomials of $\mathbf{A}$. Let p be any unary polynomial of $\mathbf{A}$. This means that for some k there is an $(k+1)$-ary term s of $\mathbf{A}$ and constants $\mathbf{c} \in \mathbf{A}^k$ so that $p(x) = s(x, \mathbf{c})$. Let $\langle x, y \rangle \in \theta$. We show that $\langle p(x), p(y) \rangle \in \theta$. Suppose that t is an $(n+1)$-ary term of $\mathbf{A}$ and that $\mathbf{a}, \mathbf{b} \in \mathbf{A}^n$. Then $t(p(x), \mathbf{a}) = t(p(x), \mathbf{b})$ if and only if $t(s(x, \mathbf{c}), \mathbf{a}) = t(s(x, \mathbf{c}), \mathbf{b})$. Since $x \theta y$, this happens if and only if $t(s(y, \mathbf{c}), \mathbf{a}) = t(s(y, \mathbf{c}), \mathbf{b})$, which happens if and only if $t(p(y), \mathbf{a}) = t(p(y), \mathbf{b})$. Thus $p(x) \theta p(y)$. The equivalence relation θ is closed under all unary polynomials of $\mathbf{A}$, so $\theta \in \operatorname{Con} \mathbf{A}$ as desired.

Now we only have left to prove that $[\theta, 1_A] = 0_A$. To do so, we show that $C(\theta, 1_A; 0_A)$. Suppose that t is an $(n+m)$-ary term operation of $\mathbf{A}$, that $\mathbf{x}, \mathbf{y} \in \mathbf{A}^m$, that $\mathbf{a}, \mathbf{b} \in \mathbf{A}^n$ with $x_i \theta y_i$ for all i. Suppose that $t(\mathbf{x}, \mathbf{a}) = t(\mathbf{x}, \mathbf{b})$. We must establish that $t(\mathbf{y}, \mathbf{a}) = t(\mathbf{y}, \mathbf{b})$. We will prove by induction on $i = 0, 1, \ldots, (m-1)$ that

$$t(y_0, \ldots, y_i, x_{i+1}, \ldots, x_{m-1}, \mathbf{a}) = t(y_0, \ldots, y_i, x_{i+1}, \ldots, x_{m-1}, \mathbf{b}).$$

For $i = 0$, since $x_0 \theta y_0$ and $t(\mathbf{x}, \mathbf{a}) = t(\mathbf{x}, \mathbf{b})$, it follows immediately that

$$t(y_0, x_1, \ldots, x_{m-1}, \mathbf{a}) = t(y_0, x_1, \ldots, x_{m-1}, \mathbf{b}).$$

Suppose then that $0 \leq i < m-1$ and that

$$t(y_0, \ldots, y_i, x_{i+1}, \ldots, x_{m-1}, \mathbf{a}) = t(y_0, \ldots, y_i, x_{i+1}, \ldots, x_{m-1}, \mathbf{b}).$$

That

$$t(y_0, \ldots, y_i, y_{i+1}, x_{i+2}, \ldots, x_{m-1}, \mathbf{a}) = t(y_0, \ldots, y_i, y_{i+1}, x_{i+2}, \ldots, x_{m-1}, \mathbf{b})$$

follows now from $x_{i+1}\theta y_{i+1}$. This completes the induction argument. Taking $i = m-1$ now yields $t(\mathbf{y}, \mathbf{a}) = t(\mathbf{y}, \mathbf{b})$ as desired. □

5 Examples

In this section we give a few examples of centrality and the commutator in some familiar varieties. Hopefully, these examples will motivate some of the results we will prove later and the techniques necessary to prove them. We first consider the commutator in rings.

Suppose that $\mathbf{R}$ is a ring and let I, J, and K be ideals of $\mathbf{R}$. Denote the congruence relations corresponding to I, J, and K by α, β, and δ. Then, for example, $x\alpha y$ if and only if $x - y \in I$. Suppose further that $C(\alpha, \beta; \delta)$. Taking $x \in I$ and $y \in J$, we will use centrality to prove that $xy \in K$. We have the relations $x\alpha 0$ and $y\beta 0$, so the matrix

$$\begin{pmatrix} 0y & 00 \\ xy & x0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ xy & 0 \end{pmatrix}$$

is in $M(\alpha, \beta)$. Since obviously the top row of this matrix in in δ, we have that $xy\delta 0$. This means that $xy = xy - 0 \in K$. A symmetric argument will show that $yx \in K$ also. This implies that K must contain the ideal $IJ + JI$. Suppose on the other hand that K is an ideal containing $IJ + JI$. We will use this assumption to prove that $C(\alpha, \beta; \delta)$. First of all, notice that in $\mathbf{R}/K$, the product of any element from $\mathbf{I}/K$ and an element from J/K is 0. Suppose now that t is an $(n+m)$-ary ring term, $\mathbf{a}, \mathbf{b} \in R^n$ with $a_i \alpha b_i$ for all i, $\mathbf{x}, \mathbf{y} \in R^m$ with $x_i \beta y_i$ for all i, and that $t(\mathbf{a}, \mathbf{x})\delta t(\mathbf{a}, \mathbf{y})$. We need to show that $t(\mathbf{b}, \mathbf{x})\delta t(\mathbf{b}, \mathbf{y})$. To simplify notation in the next calculations, we will write $\bar{r}$ for any coset $r + K$ in $\mathbf{R}/K$. Since $t(\mathbf{a}, \mathbf{x})\delta t(\mathbf{a}, \mathbf{y})$, in $\mathbf{R}/K$ we have $t(\bar{\mathbf{a}}, \bar{\mathbf{x}}) = t(\bar{\mathbf{a}}, \bar{\mathbf{y}})$. Through distribution we can find ring terms q, r, s so that

$$t(\mathbf{u}, \mathbf{v}) = q(\mathbf{u}) + r(\mathbf{v}) + s(\mathbf{u}, \mathbf{v}) \tag{5.1}$$

where each of q, r, and s is a sum of products of variables and negated variables and so that in each product of s at least one u_i and at least one v_i occurs. Notice that by our assumptions

$$s(\bar{\mathbf{a}}, \bar{\mathbf{x}}) = s(\bar{\mathbf{a}}, \bar{\mathbf{y}}) = s(\bar{\mathbf{b}}, \bar{\mathbf{x}}) = s(\bar{\mathbf{b}}, \bar{\mathbf{y}}) = 0. \tag{5.2}$$

From $t(\bar{\mathbf{a}}, \bar{\mathbf{x}}) = t(\bar{\mathbf{a}}, \bar{\mathbf{y}})$ it follows that

$$q(\bar{\mathbf{a}}) + r(\bar{\mathbf{x}}) + s(\bar{\mathbf{a}}, \bar{\mathbf{x}}) = q(\bar{\mathbf{a}}) + r(\bar{\mathbf{y}}) + s(\bar{\mathbf{a}}, \bar{\mathbf{y}}) \tag{5.3}$$

and hence that

$$q(\bar{\mathbf{a}}) + r(\bar{\mathbf{x}}) + 0 = q(\bar{\mathbf{a}}) + r(\bar{\mathbf{y}}) + 0. \tag{5.4}$$

Appropriate cancellation and addition of $q(\bar{\mathbf{b}})$ now gives

$$q(\bar{\mathbf{b}}) + r(\bar{\mathbf{x}}) + 0 = q(\bar{\mathbf{b}}) + r(\bar{\mathbf{y}}) + 0 \tag{5.5}$$

and hence

$$q(\bar{\mathbf{b}}) + r(\bar{\mathbf{x}}) + s(\bar{\mathbf{b}}, \bar{\mathbf{x}}) = q(\bar{\mathbf{b}}) + r(\bar{\mathbf{y}}) + s(\bar{\mathbf{b}}, \bar{\mathbf{y}})\,. \tag{5.6}$$

This gives $t(\bar{\mathbf{b}}, \bar{\mathbf{x}}) = t(\bar{\mathbf{b}}, \bar{\mathbf{y}})$ or $t(\mathbf{b}, \mathbf{x})\delta t(\mathbf{b}, \mathbf{y})$ as desired. We have proven that $C(\alpha, \beta; \delta)$ holds if and only if K contains $IJ + JI$. This gives us

5.1 Fact *Let* $\mathbf{R}$ *be a ring and let* I *and* J *be ideals of* $\mathbf{R}$*. Suppose that* α *and* β *are the congruences associated with* I *and* J*. Then* $[\alpha, \beta]$ *is the congruence associated with the ideal* $IJ + JI$*.*

This fact tells us what Abelian rings look like. Suppose that $\mathbf{R}$ is an Abelian ring. This means that $[1_R, 1_R] = 0_R$, which by the last fact tells us that $R{\cdot}R + R{\cdot}R = \{0\}$. This happens exactly when multiplication in $\mathbf{R}$ is trivial in the sense that all products are 0. Hence

5.2 Fact *A ring* $\mathbf{R}$ *is Abelian if and only if all products in* $\mathbf{R}$ *are* 0*.*

Since we know what ring commutators look like, it is easy to see what the center of a ring is.

5.3 Fact *If* $\mathbf{R}$ *is any ring, then* $\zeta_{\mathbf{R}}$ *is the annihilator of* $\mathbf{R}$*—the set of all* $x \in R$ *so that* $xr = rx = 0$ *for all* $r \in R$*.*

We next turn our attention to the commutator in lattices. Suppose that $\mathbf{L}$ is a lattice and that α and β are congruences on $\mathbf{L}$. We will prove that $[\alpha, \beta] = \alpha \cap \beta$. That $[\alpha, \beta] \subseteq \alpha \cap \beta$ is always true. We just need to prove the reverse inclusion. Suppose that $\langle a, b\rangle \in \alpha \cap \beta$. We will use the α, β term condition modulo $[\alpha, \beta]$ to show that $\langle a, b\rangle \in [\alpha, \beta]$. Consider the lattice term $t(x, y, z) = (x \wedge y) \vee (x \wedge z) \vee (y \wedge z)$. This term satisfies

$$t(x, x, x) \approx t(x, x, y) \approx t(x, y, x) \approx t(y, x, x) \approx x. \tag{5.7}$$

Any ternary term which satisfies these equations is called a majority term. It is well known that any variety with a majority term is congruence distributive. The matrix

$$\begin{pmatrix} a & a \\ a & b \end{pmatrix} = \begin{pmatrix} t(a,a,a) & t(a,a,b) \\ t(a,b,a) & t(a,b,b) \end{pmatrix}$$

is in $M(\alpha, \beta)$. Since we have equality in the first row—and hence a relation via $[\alpha, \beta]$, the α, β term condition modulo $[\alpha, \beta]$ tells us the second row is in $[\alpha, \beta]$. We have proven

5.4 Fact *The commutator operation in the variety of lattices is congruence intersection.*

Actually, our argument proves

5.5 Fact *Suppose that* $\mathcal{V}$ *is any variety with a majority operation. Then the commutator operation in* $\mathcal{V}$ *is congruence intersection.*

We will extend this fact in Section 7 to all congruence distributive varieties. For now, we will use it to see what Abelian lattices look like. A lattice $\mathbf{L}$ is Abelian if and only if $0_L = [1_L, 1_L] = 1_L \cap 1_L = 1_L$. This happens if and only if $\mathbf{L}$ is a one element lattice. Thus

5.6 Fact *The only Abelian lattice is the one element lattice.*

This could also be proven quickly by considering the term $t(x, y, z) = x \wedge y \wedge z$ and the term condition. Since such an argument would only involve the one lattice operation, it would also establish the same result for semilattices. Also notice that since the commutator in lattices is intersection, the center of a lattice is trivial (the identity relation).

As a final example in this section, we will consider the commutator in modules. We will prove that the commutator in any module $\mathbf{M}$ over a ring $\mathbf{R}$ is constantly 0_M. To do this, it suffices to show that $[1_M, 1_M] = 0_M$. In particular, we will see that every module is Abelian. Suppose that t is any $(n+m)$-ary term of $\mathbf{M}$. We can assume that $\mathbf{R}$ has a unity. The term t can be expressed as

$$t(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^{l} r_i u_{j_i} + \sum_{i=1}^{l'} s_i v_{j_i} \tag{5.8}$$

where each r_i and each s_i is in $\mathbf{R}$. Let $\mathbf{a}, \mathbf{b} \in M^n$ and $\mathbf{x}, \mathbf{y} \in M^m$ and suppose that $t(\mathbf{a}, \mathbf{x}) = t(\mathbf{a}, \mathbf{y})$. This means that

$$\sum_{i=1}^{l} r_i a_{j_i} + \sum_{i=1}^{l'} s_i x_{j_i} = \sum_{i=1}^{l} r_i a_{j_i} + \sum_{i=1}^{l'} s_i y_{j_i}. \tag{5.9}$$

Appropriate cancellation and addition now gives

$$\sum_{i=1}^{l} r_i b_{j_i} + \sum_{i=1}^{l'} s_i x_{j_i} = \sum_{i=1}^{l} r_i b_{j_i} + \sum_{i=1}^{l'} s_i y_{j_i} \tag{5.10}$$

and hence that $t(\mathbf{b}, \mathbf{x}) = t(\mathbf{b}, \mathbf{y})$. Thus we have established that $C(1_M, 1_M; 0_M)$ so $[1_M, 1_M] = 0_M$. This means

5.7 Fact *Every module over a ring is Abelian.*

It follows, of course, that the center of any module is the universal relation. We will see in Section 9 that every Abelian algebra in a congruence modular variety is (polynomially equivalent to) a module over a ring.

6 Maltsev conditions

A variety $\mathcal{W}$ is said to *interpret* a variety $\mathcal{V}$ if for every basic operation t of $\mathcal{V}$ there is a $\mathcal{W}$-term s_t so that for every algebra $\mathbf{A} \in \mathcal{W}$ the algebra $\langle A, \{s_t^{\mathbf{A}}\}\rangle$ is a member of $\mathcal{V}$. We denote this situation by $\mathcal{V} \leq \mathcal{W}$. We say that a class $\mathcal{K}$ of varieties is a *strong Maltsev class* (or that $\mathcal{K}$ is defined by a *strong Maltsev condition*) if and only if there is a finitely presented variety $\mathcal{V}$ so that $\mathcal{K}$ is precisely the class of all varieties $\mathcal{W}$ for which $\mathcal{V} \leq \mathcal{W}$. If there are finitely presented varieties $\cdots \leq \mathcal{V}_3 \leq \mathcal{V}_2 \leq \mathcal{V}_1$ so that $\mathcal{K}$ is the class of all varieties $\mathcal{W}$ so that $\mathcal{V}_i \leq \mathcal{W}$ for some i, then $\mathcal{K}$ is a *Maltsev class* (or is defined by a *Maltsev condition*). Finally, if $\mathcal{K}$ is the intersection of countably many Maltsev classes, then $\mathcal{K}$ is a *weak Maltsev class* (or is defined by a *weak Maltsev condition*).

Most Maltsev conditions in practice take the form of an assertion that a variety has a set of terms satisfying one of a sequence of sets of weaker and weaker equations. The first example of a strong Maltsev class was the class of all varieties with permuting congruences.

6.1 Theorem (A. I. Maltsev [31]) *A variety $\mathcal{V}$ has permuting congruences if and only if there is a term p of the variety so that $\mathcal{V}$ models the equations:*

$$p(x,z,z) \approx x \quad \textit{and} \quad p(z,z,x) \approx x.$$

Proof Suppose that a variety $\mathcal{V}$ has such a term p. Let $\mathbf{A} \in \mathcal{V}$ and let θ and ϕ be congruences of $\mathbf{A}$. Suppose that $x, z \in A$ and $\langle x, z\rangle \in \theta \circ \phi$. Then there is a $y \in A$ so that $x\theta y$ and $y\phi z$, and we have

$$x = p^{\mathbf{A}}(x,z,z)\phi p^{\mathbf{A}}(x,y,z)\theta p^{\mathbf{A}}(x,x,z) = z.$$

Thus $\theta \circ \phi \subseteq \phi \circ \theta$. Also

$$\begin{aligned} \phi \circ \theta &= \phi^{\cup} \circ \theta^{\cup} \\ &= (\theta \circ \phi)^{\cup} \\ &\subseteq (\phi \circ \theta)^{\cup} \\ &= \theta^{\cup} \circ \phi^{\cup} \\ &= \theta \circ \phi. \end{aligned} \tag{6.1}$$

So $\theta \circ \phi = \phi \circ \theta$. It follows that $\mathcal{V}$ has permuting congruences.

Suppose now that $\mathcal{V}$ has permuting congruences. Let $\mathbf{F}$ be the algebra in $\mathcal{V}$ free on $\{x, y, z\}$. Let $f : \mathbf{F} \to \mathbf{F}$ be the homomorphism which maps x and y to x and z to z and let $\theta = \ker f$. Let $g : \mathbf{F} \to \mathbf{F}$ be the homomorphism mapping x to x and y and z to z and let $\phi = \ker g$. Clearly, we have $\langle x, z\rangle \in \theta \circ \phi$. Since we are assuming congruences permute, $\langle x, z\rangle \in \phi \circ \theta$, so there must be a $w \in F$ with $x\phi w\theta z$. Since $\mathbf{F}$ is generated by $\{x, y, z\}$, there is a term p of $\mathcal{V}$ so that $p^{\mathbf{F}}(x, y, z) = w$. Observe:

$$x = g(x) = g(w) = g(p^{\mathbf{F}}(x,y,z)) = p^{\mathbf{F}}(g(x), g(y), g(z)) = p^{\mathbf{F}}(x,z,z).$$

Using f, we can similarly show that $p^{\mathbf{F}}(z,z,x) = x$. Since $\mathbf{F}$ is freely generated by $\{x, y, z\}$, it follows that these equalities hold throughout $\mathcal{V}$. □

The second example of a strong Maltsev class was found in 1963 by A. F. Pixley. It is the class of all varieties in which congruences permute and in which congruence lattices are distributive. Such varieties are called *arithmetical*. The fact that this is a Maltsev class is a consequence of the following theorem whose proof is similar to the proof of Maltsev's theorem:

6.2 Theorem (A. F. Pixley [38]) *A variety $\mathcal{V}$ is arithmetical if and only if it has a term t so that $\mathcal{V}$ models*

$$t(x,y,y) \approx t(y,y,x) \approx t(x,y,x) \approx x.$$

The first class of varieties which was shown to be a Maltsev class but not a strong Maltsev class was the class of all varieties in which all congruence lattices are distributive. Such a variety is said to be congruence distributive.

6.3 Theorem (B. Jónsson [25]) *A variety $\mathcal{V}$ is congruence distributive if and only if for some positive integer n, $\mathcal{V}$ has ternary terms $d_0, \ldots, d_n$ so that $\mathcal{V}$ models the following equations:*

$$\begin{aligned} x &\approx d_0(x,y,z) \\ x &\approx d_i(x,y,x) \quad \text{for } 0 \leq i \leq n \\ d_i(x,x,y) &\approx d_{i+1}(x,x,y) \quad \text{for even } i < n \\ d_i(x,y,y) &\approx d_{i+1}(x,y,y) \quad \text{for odd } i < n \\ d_n(x,y,z) &\approx z. \end{aligned}$$

Proof Suppose first that $\mathcal{V}$ has ternary terms as described. Let $\mathbf{A} \in \mathcal{V}$ and let θ, ϕ and ψ be congruences of $\mathbf{A}$. Suppose that $\langle a, c\rangle \in \theta \cap (\phi \vee \psi)$ and let $\alpha = (\theta \cap \phi) \vee (\theta \cap \psi)$. We show $\langle a, c\rangle \in \alpha$. There must be $a = x_0, x_1, \ldots, x_k = c$ in A with $\langle x_i, x_{i+1}\rangle \in \phi \cup \psi$ for $i < k$. For any $j \leq n$ and for any $i < k$, we have that $\langle d_j^{\mathbf{A}}(a, x_i, c), d_j^{\mathbf{A}}(a, x_{i+1}, c)\rangle \in \phi \cup \psi$. Also,

$$d_j^{\mathbf{A}}(a, x_i, c)\theta d_j^{\mathbf{A}}(a, x_i, a) = a = d_j^{\mathbf{A}}(a, x_{i+1}, a)\theta d_j^{\mathbf{A}}(a, x_{i+1}, c).$$

Hence, $\langle d_j^{\mathbf{A}}(a, x_i, c), d_j^{\mathbf{A}}(a, x_{i+1}, c)\rangle \in \alpha$. By transitivity, for all $0 \leq j \leq n$:

$$d_j^{\mathbf{A}}(a, a, c) = d_j^{\mathbf{A}}(a, x_0, c)\alpha d_j^{\mathbf{A}}(a, x_k, c) = d_j^{\mathbf{A}}(a, c, c).$$

By the third and fourth equations above, this yields $d_j^{\mathbf{A}}(a, c, c)\alpha d_{j+1}^{\mathbf{A}}(a, c, c)$ for all $j \leq n$. Hence, $a = d_0^{\mathbf{A}}(a, c, c)\alpha d_n^{\mathbf{A}}(a, c, c) = c$. Thus, $\theta \cap (\phi \vee \psi) \subseteq (\theta \cap \phi) \vee (\theta \cap \psi)$. The reverse inclusion always holds, so we have established that $\operatorname{Con} \mathbf{A}$ is distributive.

Now assume that $\mathcal{V}$ is congruence distributive and let $\mathbf{F}$ be the free algebra in $\mathcal{V}$ generated by $\{x, y, z\}$. Let f, g, and h be homomorphisms from $\mathbf{F}$ to $\mathbf{F}$ given by:

$$\begin{aligned} f(x) &= f(y) = x, \\ f(z) &= z, \\ g(x) &= x, \\ g(y) &= g(z) = y, \\ h(x) &= h(z) = x, \text{ and} \\ h(y) &= y. \end{aligned}$$

Let $\phi = \ker f$, $\psi = \ker g$, and $\theta = \ker h$. Since $\langle x, z\rangle \in \theta \cap (\phi \vee \psi) \leq (\theta \cap \phi) \vee (\theta \cap \psi)$, there must be elements $w_0 = x, x_1, \ldots, w_n = z$ in F so that

$$\begin{aligned} w_i \theta x &\text{ for all } i \leq n, \\ w_i \psi w_{i+1} &\text{ for all even } i < n, \text{ and} \\ w_i \phi w_{i+1} &\text{ for all odd } i < n. \end{aligned}$$

Since $\mathbf{F}$ is generated by $\{x, y, z\}$, there must be ternary terms $d_0, \ldots d_n$ so that $d_i^{\mathbf{F}}(x, y, z) = w_i$ for $i = 0, \ldots, n$. That these terms satisfy the desired equations follows as in the proof of Maltsev's theorem above. □

The following Maltsev characterization of congruence modularity is critical for work with the modular commutator.

6.4 Theorem (A. Day [7]) *A variety $\mathcal{V}$ is congruence modular if and only if $\mathcal{V}$ has 4-ary terms $m_0, \dots, m_n$ for which $\mathcal{V}$ satisfies the equations*

$$\begin{aligned} x &\approx m_0(x,y,z,u) \\ x &\approx m_i(x,y,y,x) \quad \textit{for all } i \\ m_i(x,x,z,z) &\approx m_{i+1}(x,x,z,z) \quad \textit{for } i \textit{ even} \\ m_i(x,y,y,u) &\approx m_{i+1}(x,y,y,u) \quad \textit{for } i \textit{ odd} \\ m_n(x,y,z,u) &\approx u. \end{aligned}$$

The terms in Theorem 6.4 are called Day terms. To prove Day's theorem, we need the following lemmas.

6.5 Lemma *Let $m_0, \dots, m_n$ be Day terms for a variety $\mathcal{V}$. Let $\mathbf{A} \in \mathcal{V}$ and $a, b, c, d \in A$. Let $\gamma \in \operatorname{Con}\mathbf{A}$ with $b\gamma d$. Then $a\gamma c$ if and only if $m_i(a,a,c,c)\gamma m_i(a,b,d,c)$ for each $i = 0, \dots, n$.*

Proof First suppose that $a\gamma c$. Then

$$m_i(a,a,c,c)\gamma m_i(a,a,a,a) = a = m_i(a,b,b,a)\gamma m_i(a,b,d,c). \tag{6.2}$$

Next, suppose that $m_i(a,a,c,c)\gamma m_i(a,b,d,c)$ for all i. We will prove by induction that $m_i(a,b,d,c)\gamma a$ for all i. This is trivial for $i = 0$ since $m_0(a,b,d,c) = a$. Assume that $0 \le i < n$ and that $m_i(a,b,d,c)\gamma a$. If i is odd, then

$$m_{i+1}(a,b,d,c)\gamma m_{i+1}(a,b,b,c) = m_i(a,b,b,c)\gamma m_i(a,b,d,c)\gamma a. \tag{6.3}$$

If i is even, then

$$m_{i+1}(a,b,d,c)\gamma m_{i+1}(a,a,c,c) = m_i(a,a,c,c)\gamma m_i(a,b,d,c)\gamma a. \tag{6.4}$$

This finishes the proof that $m_i(a,b,d,c)\gamma a$ for all i. In particular, we now know that $a\gamma m_n(a,b,d,c) = c$. □

6.6 Lemma (Shifting Lemma—Gumm [17]) *Suppose that $\mathbf{A}$ is an algebra in a variety $\mathcal{V}$ with Day terms $m_0, \dots, m_n$. Let $\alpha, \gamma \in \operatorname{Con}\mathbf{A}$ and let β be a compatible reflexive binary relation on $\mathbf{A}$. Suppose $\alpha \cap \beta \subseteq \gamma$. If $a\beta b$, $c\beta d$, $a\alpha c$, and $b(\alpha \cap \gamma)d$, then $a\gamma c$ (see Fig. 2).*

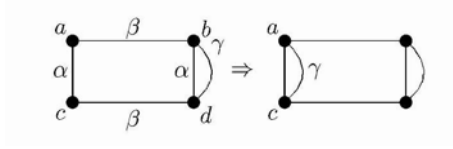


Figure 2: The Shifting Lemma.

Proof We will employ Lemma 6.5, so we need $m_i(a,a,c,c)\gamma m_i(a,b,d,c)$ for all i. First note that for all i our assumptions imply $m_i(a,a,c,c)\beta m_i(a,b,d,c)$. Next, note that for all i

$$m_i(a,a,c,c)\alpha m_i(a,a,a,a) = a = m_i(a,b,b,a)\alpha m_i(a,b,d,c)\,. \tag{6.5}$$

Thus, we have

$$\langle m_i(a,a,c,c,), m_i(a,b,d,c)\rangle \in \alpha \cap \beta \subseteq \gamma. \tag{6.6}$$

By Lemma 6.5, $a\gamma c$. □

Proof of Theorem 6.4 Suppose first that $\mathcal{V}$ is a variety with terms $m_0, \ldots, m_n$ satisfying Day's equations. Let α, β, and γ be congruences on an algebra $\mathbf{A}$ in $\mathcal{V}$ with $\alpha \geq \gamma$. We must show that $\alpha \cap (\beta \vee \gamma) = (\alpha \cap \beta) \vee \gamma$. The inclusion $\supseteq$ is always true. We establish the forward inclusion. Define compatible reflexive binary relations $\Gamma_0, \Gamma_1, \ldots$ on $\mathbf{A}$ recursively by $\Gamma_0 = \beta$ and $\Gamma_{n+1} = \Gamma_n \circ \gamma \circ \Gamma_n$. Then $\beta \vee \gamma = \bigcup_{n=0}^{\infty} \Gamma_n$. We will prove by induction that $\alpha \cap \Gamma_n \subseteq (\alpha \cap \beta) \vee \gamma$ for all n. First, $\alpha \cap \Gamma_0 = \alpha \cap \beta$ which is clearly contained in $(\alpha \cap \beta) \vee \gamma$. Assume that $n \geq 0$ and $\alpha \cap \Gamma_n \subseteq (\alpha \cap \beta) \vee \gamma$. Let $\langle a, c\rangle \in \alpha \cap \Gamma_{n+1}$. Since $\gamma \leq \alpha$ and $\gamma \leq (\alpha \cap \beta) \vee \gamma$, we have the relations in Fig. 3 for some b and d. The Shifting

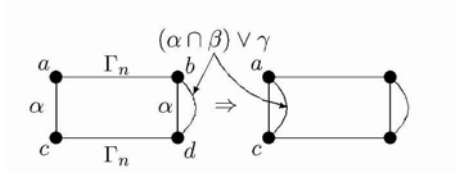


Figure 3: The inductive step for the first half of the proof of Theorem 6.4.

Lemma now gives that $\langle a, c\rangle \in (\alpha \cap \beta) \vee \gamma$. By induction, $\alpha \cap \Gamma_n \subseteq (\alpha \cap \beta) \vee \gamma$ for all n, so $\alpha \cap (\beta \vee \gamma) \subseteq (\alpha \cap \beta) \vee \gamma$. This inclusion gives equality and completes the proof that $\operatorname{Con}\mathbf{A}$ is modular.

Next, suppose that $\mathcal{V}$ is a congruence modular variety. Let $\mathbf{A}$ be the free algebra in $\mathcal{V}$ on the generators $\{x, y, z, u\}$. Let $\alpha = \operatorname{Cg}_{\mathbf{A}}(x,u) \vee \operatorname{Cg}_{\mathbf{A}}(y,z)$, $\beta = \operatorname{Cg}_{\mathbf{A}}(x,y) \vee \operatorname{Cg}_{\mathbf{A}}(u,z)$, and $\gamma = \operatorname{Cg}_{\mathbf{A}}(y,z)$. Then $\langle x, u\rangle$ is in $\alpha \cap (\beta \vee \gamma)$ which by modularity is equal to $(\alpha \cap \beta) \vee \gamma$. This means there are elements $u_0, \ldots, u_n \in A$ so that $x = u_0$, $u = u_n$, $u_i \alpha \cap \beta u_{i+1}$ for i even, and $u_i \gamma u_{i+1}$ for i odd. Let $m_0, \ldots, m_n$ be 4-ary terms so that $m_i(x,y,z,u) = u_i$. Immediately, then, we have $m_0(x,y,z,u) = x$ and $m_n(x,y,z,u) = u$. Since $\gamma \leq \alpha$, all of the u_i's are α related. This means that $x \alpha m_i(x,y,z,u) \alpha m_i(x,y,y,x)$. However, by our definitions, $(x/\alpha) \cap \operatorname{Sg}_{\mathbf{A}}(x,y) = \{x\}$, so we have $m_i(x,y,y,x) = x$ for all i. Let $g : \mathbf{A} \to \mathbf{A}$ be the unique homomorphism defined by $g(x) = g(y) = x$ and $g(u) = g(z) = z$. Then $\ker g = \beta$. Suppose that i is even. Since $m_i(x,y,z,u)(\alpha \cap \beta) m_{i+1}(x,y,z,u)$, we have

$$\begin{aligned} m_i(x,x,z,z) &= m_i(g(x), g(y), g(z), g(u)) \\ &= g(m_i(x,y,z,u)) \\ &= g(m_{i+1}(x,y,z,u)) \\ &= m_{i+1}(g(x), g(y), g(z), g(u)) \\ &= m_{i+1}(x,x,z,z). \end{aligned} \tag{6.7}$$

Let $f : \mathbf{A} \to \mathbf{A}$ be the unique homomorphism defined by $f(x) = x$, $f(y) = f(z) = y$, and

$f(u) = u$. Then $\ker f = \gamma$. Suppose that i is odd. Since $m_i(x, y, z, u)\gamma m_{i+1}(x, y, z, u)$ we see

$$\begin{aligned} m_i(x, y, y, u) &= m_i(f(x), f(y), f(z), f(u)) \\ &= f(m_i(x, y, z, u)) \\ &= f(m_{i+1}(x, y, z, u)) \\ &= m_{i+1}(f(x), f(y), f(z), f(u)) \\ &= m_{i+1}(x, y, y, u). \end{aligned} \tag{6.8}$$

We have established these equalities in $\mathbf{A}$:

$$\begin{aligned} x &= m_0(x, y, z, u) \\ x &= m_i(x, y, y, x) \quad \text{for all } i \\ m_i(x, x, z, z) &= m_{i+1}(x, x, z, z) \quad \text{for } i \text{ even} \\ m_i(x, y, y, u) &= m_{i+1}(x, y, y, u) \quad \text{for } i \text{ odd} \\ m_n(x, y, z, u) &= u. \end{aligned}$$

Since $\mathbf{A}$ is freely generated in $\mathcal{V}$ by $\{x, y, z, u\}$, it follows that these hold as equations in all of $\mathcal{V}$. □

The lemma usually referred to as the shifting lemma assumes the underlying variety is modular and that β is a congruence. The version we have stated happens to be equivalent and is what we need to prove Day's theorem directly. We actually proved that the existence of Day terms implies the shifting lemma, that the shifting lemma implies congruence modularity, and that modularity implies the existence of Day terms. Thus, these three conditions are equivalent.

It has long been known that any lattice identity interpreted as a congruence equation is equivalent to a weak Maltsev condition [46, 39], but it is still an open problem as to which lattice equations are equivalent to Maltsev conditions. There are lattice equations which do not imply modularity among lattices but which, when satisfied by the congruence lattice of every algebra in a variety, do imply congruence modularity [8]. It was all but conjectured on p. 155 of [12] that all of these equations are Maltsev conditions for congruences. This has recently been proven to be true.

6.7 Theorem ([6]) *A variety $\mathcal{V}$ is congruence modular if and only if for all tolerances α and β on any algebra in $\mathcal{V}$ it is the case that* $\mathrm{Tr}(\alpha) \cap \mathrm{Tr}(\beta) = \mathrm{Tr}(\alpha \cap \beta)$.

Using this theorem, it is easy to show that

6.8 Theorem ([5]) *Suppose that ϵ is a lattice equation so that for any variety $\mathcal{V}$ if $\mathcal{V} \models_{\mathrm{Con}} \epsilon$, then $\mathcal{V}$ is congruence modular. Then the class of all varieties $\mathcal{V}$ with $\mathcal{V} \models_{\mathrm{Con}} \epsilon$ is a Maltsev class.*

7 Congruence distributive varieties

The commutator in congruence distributive varieties reduces to congruence intersection. This can be proved later after we have derived some of the properties of the commutator in

congruence modular varieties. However, we offer a proof here based on the Jónsson terms. This will illustrate in an isolated setting the intimate relationship between the commutator and equations for Maltsev conditions. The proof we are about to see should be reminiscent of Fact 5.5.

7.1 Theorem *A variety $\mathcal{V}$ is congruence distributive if and only*

$$\mathcal{V}\models_{\mathrm{Con}}[\alpha\vee\gamma,\beta]\approx[\alpha,\beta]\vee[\gamma,\beta]\quad \textit{and}\quad [\alpha,\beta]=\alpha\cap\beta. \tag{7.1}$$

Proof Half of this is almost obvious. Suppose that congruences in $\mathcal{V}$ satisfy the stated commutator equations. Let α,β,γ be congruences on an algebra $\mathbf{A}\in\mathcal{V}$. Then

$$\begin{aligned}(\alpha\cap\beta)\vee(\gamma\cap\beta)&=[\alpha,\beta]\vee[\gamma,\beta]\\&=[\alpha\vee\gamma,\beta]\\&=(\alpha\vee\gamma)\cap\beta.\end{aligned} \tag{7.2}$$

Thus $\operatorname{Con}\mathbf{A}$ is distributive.

For the reverse direction, we will need to use Jónsson's terms. Suppose that $\mathcal{V}$ is a congruence distributive variety and let $d_0,\dots,d_n$ be Jónsson terms for $\mathcal{V}$. Let $\mathbf{A}\in\mathcal{V}$ and $\alpha,\beta\in\operatorname{Con}\mathbf{A}$. We will prove that $[\alpha,\beta]=\alpha\cap\beta$. That $[\alpha,\beta]\subseteq\alpha\cap\beta$ is always true. We will prove the reverse inclusion. Let $\delta=[\alpha,\beta]$ and let $\langle x,y\rangle\in\alpha\cap\beta$. We will prove by induction that $d_i(x,y,x)\delta d_i(x,y,y)$ for all $i=0,\dots,n$. This is trivially true for d_0 since $x=d_0(x,y,x)=d_0(x,y,y)$. Suppose that $0\leq i<n$ and $d_i(x,y,x)\delta d_i(x,y,y)$. There are two cases—either i is even or it is odd. Supposing that i is odd, then

$$d_{i+1}(x,y,x)=d_i(x,y,x)\delta d_i(x,y,y)=d_{i+1}(x,y,y). \tag{7.3}$$

Suppose next that i is even. Since $\left(\begin{smallmatrix}d_i(x,y,x) & d_i(x,y,y)\\ d_i(x,x,x) & d_i(x,x,y)\end{smallmatrix}\right)\in M(\alpha,\beta)$, from $C(\alpha,\beta;\delta)$ we can conclude that $d_i(x,x,x)\delta d_i(x,x,y)$. It follows that

$$d_{i+1}(x,x,x)=d_i(x,x,x)\delta d_i(x,x,y)=d_{i+1}(x,x,y). \tag{7.4}$$

Now

$$\begin{pmatrix}d_{i+1}(x,x,x) & d_{i+1}(x,x,y)\\ d_{i+1}(x,y,x) & d_{i+1}(x,y,y)\end{pmatrix}\in M(\alpha,\beta),$$

so centrality now gives $d_{i+1}(x,y,x)\delta d_{i+1}(x,y,y)$. This completes the proof that for all $i\in\{0,\dots,n\}$, $d_i(x,y,x)\delta d_i(x,y,y)$. The particular case we care about is $i=n$, which gives

$$x=d_n(x,y,x)\delta d_n(x,y,y)=y. \tag{7.5}$$

Thus we have $\alpha\cap\beta\subseteq[\alpha,\beta]$, so these congruences are actually equal. We have that $\mathcal{V}\models_{\mathrm{Con}}[\alpha,\beta]=\alpha\cap\beta$. The other equation now follows immediately from distributivity. Suppose that α,β,γ are congruences on an algebra in $\mathcal{V}$. Then

$$\begin{aligned}[\alpha\vee\gamma,\beta]&=(\alpha\vee\gamma)\cap\beta\\&=(\alpha\cap\beta)\vee(\gamma\cap\beta)\\&=[\alpha,\beta]\vee[\gamma,\beta].\end{aligned} \tag{7.6}$$

□

Of course, this gives

7.2 Corollary *The only Abelian algebras in a congruence distributive variety are trivial.*

8 Congruence modular varieties

The commutator is particularly well behaved in congruence modular varieties. In fact, we can extend all of the properties listed for the group commutator in Section 2 to the commutator in congruence modular varieties. Our primary tool for doing so will be this next characterization of centrality in congruence modular varieties.

8.1 Definition Suppose that α and β are congruences on an algebra $\mathbf{A}$ in a variety with Day terms $m_0, \dots, m_n$. Define $\chi(\alpha, \beta)$ to be the set of all pairs $\langle m_i(x, x, u, u), m_i(x, y, z, u)\rangle$ for which $\left(\begin{smallmatrix} x & y \\ u & z \end{smallmatrix}\right) \in M(\alpha, \beta)$ and m_i is a Day term.

8.2 Theorem (R. Freese and R. McKenzie [12]) *Suppose that α, β, and γ are congruences on an algebra $\mathbf{A}$ in a congruence modular variety. The following are equivalent.*

(1) $C(\alpha, \beta; \gamma)$;

(2) $\chi(\alpha, \beta) \subseteq \gamma$;

(3) $C(\beta, \alpha; \gamma)$;

(4) $\chi(\beta, \alpha) \subseteq \gamma$.

Proof We will prove that (1)→(2)→(3). Then exchanging α and β in these implications will show that all four conditions are equivalent.

(1)→(2): Suppose that $C(\alpha, \beta; \gamma)$. Let t be an $(n+m)$-ary term of $\mathbf{A}$. Let $\mathbf{a}, \mathbf{b} \in A^n$ and $\mathbf{x}, \mathbf{y} \in A^m$ with $a_i \alpha b_i$ for all i and $x_i \beta y_i$ for all i. This makes $\left(\begin{smallmatrix} t(\mathbf{a},\mathbf{x}) & t(\mathbf{a},\mathbf{y}) \\ t(\mathbf{b},\mathbf{x}) & t(\mathbf{b},\mathbf{y}) \end{smallmatrix}\right)$ a generic element of $M(\alpha, \beta)$. To establish the implication, we need to prove that

$$m_i(t(\mathbf{a},\mathbf{x}), t(\mathbf{a},\mathbf{x}), t(\mathbf{b},\mathbf{x}), t(\mathbf{b},\mathbf{x}))\gamma m_i(t(\mathbf{a},\mathbf{x}), t(\mathbf{a},\mathbf{y}), t(\mathbf{b},\mathbf{y}), t(\mathbf{b},\mathbf{x}))\,. \tag{8.1}$$

The matrix

$$\begin{pmatrix} m_i(t(\mathbf{a},\mathbf{x}), t(\mathbf{b},\mathbf{x}), t(\mathbf{b},\mathbf{x}), t(\mathbf{a},\mathbf{x})) & m_i(t(\mathbf{a},\mathbf{x}), t(\mathbf{b},\mathbf{y}), t(\mathbf{b},\mathbf{y}), t(\mathbf{a},\mathbf{x})) \\ m_i(t(\mathbf{a},\mathbf{x}), t(\mathbf{a},\mathbf{x}), t(\mathbf{b},\mathbf{x}), t(\mathbf{b},\mathbf{x})) & m_i(t(\mathbf{a},\mathbf{x}), t(\mathbf{a},\mathbf{y}), t(\mathbf{b},\mathbf{y}), t(\mathbf{b},\mathbf{x})) \end{pmatrix} \tag{8.2}$$

is in $M(\alpha, \beta)$. Notice that by the Day equations both elements of the top row of this matrix equal $t(\mathbf{a}, \mathbf{x})$. In particular, the top elements are γ related. It follows then that the bottom elements are also γ related as desired.

(2)→(3): Suppose now that $\chi(\alpha, \beta) \subseteq \gamma$. Suppose that $\left(\begin{smallmatrix} b & d \\ a & c \end{smallmatrix}\right) \in M(\beta, \alpha)$ and that $b\gamma d$. It follows that $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in M(\alpha, \beta)$ and hence that $\langle m_i(a, a, c, c), m_i(a, b, d, c)\rangle \in \gamma$ for all i. By Lemma 6.5 it follows that $a\gamma c$, so $C(\beta, \alpha; \gamma)$.

We now have (1)→(2)→(3). By trading α and β, we get (3)→(4)→(1), so the statements are equivalent. □

We can now easily prove the following corollaries.

8.3 Theorem *Suppose that α, β, and γ are congruences on an algebra $\mathbf{A}$ in a congruence modular variety.*

(1) $C(\alpha, \beta; \gamma)$ *if and only if* $[\alpha, \beta] \leq \gamma$.

(2) $[\alpha, \beta] = \mathrm{Cg}_{\mathbf{A}}(\chi(\alpha, \beta)) = \mathrm{Cg}_{\mathbf{A}}(\chi(\beta, \alpha))$.

(3) $C(\alpha, \beta; \gamma)$ *if and only if* $C(\beta, \alpha; \gamma)$.

(4) $[\alpha, \beta] = [\beta, \alpha]$.

(5) *If* $\{\alpha_t : t \in T\} \subseteq \mathrm{Con}\,\mathbf{A}$ *then* $[\bigvee_t \alpha_t, \beta] = \bigvee_t [\alpha_t, \beta]$.

(6) *For any surjective homomorphism* $f : \mathbf{A} \to \mathbf{B}$, *if* $\pi = \ker f$ *then*

$$[\alpha, \beta] \vee \pi = f^{-1}([f(\alpha \vee \pi), f(\beta \vee \pi)])$$

and

$$[f(\alpha \vee \pi), f(\beta \vee \pi)] = f([\alpha, \beta] \vee \pi).$$

(*See Fig.* 4)

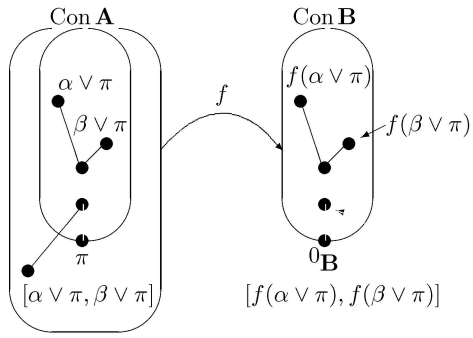


Figure 4: To calculate $[f(\alpha \vee \pi), f(\beta \vee \pi)]$, first pull back through f to $\alpha \vee \pi$ and $\beta \vee \pi$. Their commutator $[\alpha, \beta]$ might not lie above π, so join with π. The image of this congruence under f is $[f(\alpha \vee \pi), f(\beta \vee \pi)]$.

Proof (1)–(4) are immediate from the previous lemma and the definition of the commutator. We look first, then, at (5). That $\bigvee_t [\alpha_t, \beta] \subseteq [\bigvee_t \alpha_t, \beta]$ follows from monotonicity. To establish the reverse inclusion, it suffices to show that $C(\bigvee_t \alpha_t, \beta; \bigvee_t [\alpha_t, \beta])$. First, notice that by property (1), $C(\alpha_t, \beta; \bigvee_t [\alpha_t, \beta])$ holds for all t. Then Lemma 4.4 (1) gives the desired result.

For part (6), note that by part (5) $[\alpha, \beta] \vee \pi = [\alpha \vee \pi, \beta \vee \pi] \vee \pi$; and from this the two statements can easily be seen to be equivalent. Part (6) now follows from the fact that the function induced by f on $\mathbf{A}^4$ maps $M(\alpha \vee \pi, \beta \vee \pi)$ onto $M(f(\alpha \vee \pi), f(\beta \vee \pi))$ and maps $\chi(\alpha \vee \pi, \beta \vee \pi)$ onto $\chi(f(\alpha \vee \pi), f(\beta \vee \pi))$. □

The symmetry in the above theorem gives us this characterization of centrality in congruence modular varieties.

8.4 Corollary *Suppose that α, β, and δ are congruences on an algebra in a congruence modular variety. Then $C(\alpha, \beta; \delta)$ if and only if for all $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in M(\alpha, \beta)$*

$$a\delta b \leftrightarrow c\delta d \quad \text{and} \quad a\delta c \leftrightarrow b\delta d. \tag{8.3}$$

To make further arguments clearer, we will often write the elements of $\mathbf{A}^2$ as column vectors.

8.5 Definition Suppose that $\mathbf{A}$ is an algebra in a congruence modular variety and $\alpha, \beta \in \operatorname{Con}\mathbf{A}$. Let $\mathbf{A}(\alpha)$ be the subalgebra of $\mathbf{A}^2$ whose universe is α and define these three congruences on $\mathbf{A}(\alpha)$

$$[\alpha, \beta]_0 = \left\{ \left\langle \begin{pmatrix} x \\ u \end{pmatrix}, \begin{pmatrix} y \\ v \end{pmatrix} \right\rangle \in \mathbf{A}(\alpha)^2 : \langle x, y \rangle \in [\alpha, \beta] \right\} \tag{8.4}$$

$$[\alpha, \beta]_1 = \left\{ \left\langle \begin{pmatrix} u \\ x \end{pmatrix}, \begin{pmatrix} v \\ y \end{pmatrix} \right\rangle \in \mathbf{A}(\alpha)^2 : \langle x, y \rangle \in [\alpha, \beta] \right\} \tag{8.5}$$

$$\Delta_{\alpha,\beta} = \operatorname{Tr}\left(\left\{ \left\langle \begin{pmatrix} x \\ u \end{pmatrix}, \begin{pmatrix} y \\ z \end{pmatrix} \right\rangle : \begin{pmatrix} x & y \\ u & z \end{pmatrix} \in M(\alpha, \beta) \right\} \right). \tag{8.6}$$

8.6 Lemma *Suppose that α and β are congruences on an algebra $\mathbf{A}$ in a congruence modular variety. For $i \in \{0, 1\}$, let $\pi_i : \mathbf{A}^2 \to \mathbf{A}$ be the projection to the i^{th} coordinate and let $\eta_i = \ker \pi_i|_{\mathbf{A}(\alpha)}$. Then*

(1) $\eta_1 \cap \Delta_{\alpha,\beta} \subseteq [\alpha, \beta]_0$.

(2) $\eta_0 \cap \Delta_{\alpha,\beta} \subseteq [\alpha, \beta]_1$.

(3) $\Delta_{\alpha,\beta} \vee \eta_0 = \pi_0^{-1}(\beta)$.

(4) $\Delta_{\alpha,\beta} \vee \eta_1 = \pi_1^{-1}(\beta)$.

Proof For the proof, we will write Δ for $\Delta_{\alpha,\beta}$. Let $\langle \left(\begin{smallmatrix} x \\ u \end{smallmatrix}\right), \left(\begin{smallmatrix} y \\ u \end{smallmatrix}\right) \rangle \in \eta_1 \cap \Delta$. Then we have the arrangement in Fig. 5, so we can conclude that $\langle \left(\begin{smallmatrix} x \\ y \end{smallmatrix}\right), \left(\begin{smallmatrix} y \\ y \end{smallmatrix}\right) \rangle \in \eta_1 \cap \Delta$. This means that

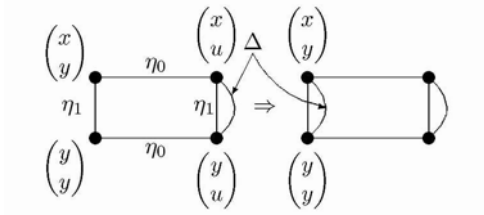


Figure 5: The shifting lemma for Lemma 8.6 (1).

there are $a_0, \ldots, a_n, b_0, \ldots, b_n \in A$ so that $a_0 = x$, $b_0 = a_n = b_n = y$ and for each $i < n$ $\left(\begin{smallmatrix} a_i & a_{i+1} \\ b_i & b_{i+1} \end{smallmatrix}\right) \in M(\alpha, \beta)$. Since $\langle a_n, b_n \rangle = \langle y, y \rangle \in [\alpha, \beta]$, we can use Corollary 8.4 to establish that $\langle a_i, b_i \rangle \in [\alpha, \beta]$ for all i. In particular, $\langle x, y \rangle \in [\alpha, \beta]$. This places $\langle \left(\begin{smallmatrix} x \\ u \end{smallmatrix}\right), \left(\begin{smallmatrix} y \\ u \end{smallmatrix}\right) \rangle \in [\alpha, \beta]_0$ as desired. We have established that $\eta_1 \cap \Delta \subseteq [\alpha, \beta]_0$. This proves (1). (2) now follows because

$$\left\langle \begin{pmatrix} x \\ u \end{pmatrix}, \begin{pmatrix} y \\ v \end{pmatrix} \right\rangle \in [\alpha, \beta]_0 \iff \left\langle \begin{pmatrix} u \\ x \end{pmatrix}, \begin{pmatrix} v \\ y \end{pmatrix} \right\rangle \in [\alpha, \beta]_1 \tag{8.7}$$

and

$$\left\langle \begin{pmatrix} x \\ u \end{pmatrix}, \begin{pmatrix} y \\ v \end{pmatrix} \right\rangle \in \Delta \iff \left\langle \begin{pmatrix} u \\ x \end{pmatrix}, \begin{pmatrix} v \\ y \end{pmatrix} \right\rangle \in \Delta. \tag{8.8}$$

For (3), suppose that $x\beta y$, $x\alpha u$ and $y\alpha v$. Then

$$\begin{pmatrix} x \\ u \end{pmatrix} \eta_0 \begin{pmatrix} x \\ x \end{pmatrix} \Delta \begin{pmatrix} y \\ y \end{pmatrix} \eta_0 \begin{pmatrix} y \\ v \end{pmatrix}. \tag{8.9}$$

This shows $\pi_0^{-1}(\beta) \subseteq \eta_0 \vee \Delta$. The other inclusion is trivial. (4) is similar. □

8.7 Theorem *Suppose that $\mathcal{V}$ is a congruence modular variety. The commutator is the greatest binary operation defined on the congruence lattice of every algebra in $\mathcal{V}$ so that for any $\mathbf{A}, \mathbf{B} \in \mathcal{V}$, for any $\alpha, \beta \in \operatorname{Con}\mathbf{A}$, and for any surjective homomorphism $f : \mathbf{A} \to \mathbf{B}$*

$$[\alpha, \beta] \leq \alpha \cap \beta \quad \textit{and} \tag{8.10}$$

$$[\alpha, \beta] \vee \pi = f^{-1}([f(\alpha \vee \pi), f(\beta \vee \pi)]). \tag{8.11}$$

Proof Let C be any other binary operation on the congruence lattices of algebras in $\mathcal{V}$ satisfying the stated properties. We will prove that the commutator always dominates C. The proof is essentially the same as that of this property for groups presented in Section 2. Let α and β be congruences on an algebra $\mathbf{A}$ in $\mathcal{V}$.

For $i \in \{0, 1\}$, let $\pi_i : \mathbf{A}(\alpha) \to \mathbf{A}$ be the projection to the i^{th} coordinate with $\eta_i = \ker \pi_i|_{\mathbf{A}(\alpha)}$. Let $\Delta = \Delta_{\alpha,\beta}$. In $\operatorname{Con}\mathbf{A}(\alpha)$ we have $C(\eta_1, \Delta) \subseteq \eta_1 \cap \Delta$ by our assumptions on C. Also, $\eta_1 \cap \Delta \subseteq [\alpha, \beta]_0$ by Lemma 8.6. Hence $C(\eta_1, \Delta) \subseteq [\alpha, \beta]_0$. Let $\alpha_0 = \pi_0^{-1}(\alpha)$ and $\beta_0 = \pi_0^{-1}(\beta)$. Then $\alpha_0 = \eta_0 \vee \eta_1$ and $\beta_0 = \eta_0 \vee \Delta$ so $\alpha = \pi_0(\eta_0 \vee \eta_1)$ and $\beta = \pi_0(\eta_0 \vee \Delta)$. It follows then by our assumptions on C that

$$\begin{aligned} C(\alpha, \beta) &= C(\pi_0(\eta_0 \vee \eta_1), \pi_0(\eta_0 \vee \Delta)) \\ &= \pi_0(C(\eta_1, \Delta) \vee \eta_0) \\ &\subseteq \pi_0([\alpha, \beta]_0) \\ &= [\alpha, \beta]. \end{aligned} \tag{8.12}$$

This concludes the proof that C is always dominated by the commutator. □

At this point in time, we have an extension of all seven of the properties of the group commutator listed in Section 2.

9 Abelian algebras and Abelian varieties

Recall that we defined an algebra $\mathbf{A}$ to be Abelian if $[1_A, 1_A] = 0_A$. In some respects, Abelian algebras and algebras generating congruence distributive varieties represent two extremes in congruence modular varieties. In Abelian algebras, the commutator is as small as possible. In algebras generating congruence distributive varieties, the commutator is as large as possible. This dichotomy is emphasized further in this theorem.

9.1 Theorem *Suppose that $\mathbf{A}$ is an algebra in a congruence modular variety. The following are equivalent.*

(1) *The projection congruences have a common complement in* $\operatorname{Con}\mathbf{A} \times \mathbf{A}$.

(2) $\mathbf{M}_3$ *is a* 0-1 *sublattice of* $\operatorname{Con}\mathbf{A}^2$.

(3) $\mathbf{M}_3$ *is a* 0-1 *sublattice of some subdirect product of two copies of* $\mathbf{A}$.

(4) $\mathbf{A}$ *is Abelian.*

Proof (1) → (2) and (2) → (3) are immediate. Suppose that $\mathbf{B}$ is a subalgebra of $\mathbf{A}^2$ which projects onto both coordinates and that $\mathbf{M}_3$ is a 0-1 sublattice of $\operatorname{Con}\mathbf{B}$. We claim that $\mathbf{B}$ is Abelian. Let α, β, and γ be the atoms of the copy of $\mathbf{M}_3$. Then

$$\begin{aligned}[1_B, 1_B] &= [\alpha \vee \beta, \alpha \vee \gamma] \\ &= [\alpha, \alpha] \vee [\alpha, \gamma] \vee [\beta, \alpha] \vee [\beta, \gamma] \\ &\subseteq \alpha \vee (\beta \cap \gamma) \\ &= \alpha \end{aligned} \tag{9.1}$$

so $[1_B, 1_B] \subseteq \alpha$. Similarly, $[1_B, 1_B]$ is below β and γ. Thus, $[1_B, 1_B] = 0_B$ and $\mathbf{B}$ is Abelian. Let $\pi : \mathbf{B} \to \mathbf{A}$ be either projection and $\eta = \ker \pi$. Now by Theorem 8.3 (6) we have

$$\begin{aligned}[1_A, 1_A] &= [\pi(1_B), \pi(1_B)] \\ &= \pi([1_B, 1_B] \vee \eta) \\ &= \pi(\eta) \\ &= 0_A. \end{aligned} \tag{9.2}$$

Thus $\mathbf{A}$ is Abelian, so (3) → (4).

Now assume that $\mathbf{A}$ is Abelian. For $i = 0, 1$, let π_i be the projection of $\mathbf{A}^2$ to the i^{th} coordinate and let $\eta_i = \ker \pi_i$. Let $\Delta = \Delta_{1_A, 1_A}$. It follows from Lemma 8.6 that $\Delta \vee \eta_i = 1_A$ for $i = 0, 1$. Also, $\Delta \cap \eta_i \subseteq [1_A, 1_A]_{1-i} = \eta_{1-i}$ for $i = 0, 1$, so $\Delta \cap \eta_i = 0_{A^2}$. Thus, Δ is a complement of both projection kernels. □

9.2 Definition Two algebras are *polynomially equivalent* if they have the same universe and the same polynomial operations. An algebra is *affine* if it is polynomially equivalent with a module over a ring.

J. D. H. Smith and R. McKenzie independently proved that any Abelian algebra in a congruence permutable variety is affine. C. Herrmann [20] proved that any Abelian algebra in a congruence modular variety is affine using a complex directed union construction which forced the existence of a Maltsev operation in the original algebra. This Maltsev operation is the key to the proof. H.-P. Gumm [13, 15] also constructed this term with his geometric arguments. Walter Taylor developed the following terms which we will be able to use to construct such a Maltsev term.

9.3 Lemma (W. Taylor [45]) *Suppose that* $\mathcal{V}$ *is a congruence modular variety and that* $m_0, \ldots m_n$ *are Day terms for the variety. Define ternary terms* $q_0, \ldots, q_n$ *recursively in the*

following manner

$$\begin{aligned} q_0(x,y,z) &= z \\ q_{i+1}(x,y,z) &= m_{i+1}(q_i(x,y,z),x,y,q_i(x,y,z)) \quad \textit{if } i \textit{ is even} \\ q_{i+1}(x,y,z) &= m_{i+1}(q_i(x,y,z),y,x,q_i(x,y,z)) \quad \textit{if } i \textit{ is odd.} \end{aligned}$$

Then

(1) $\mathcal{V} \models q_i(x,x,y) \approx y$ *for all* i.

(2) *For any congruence* β *on an algebra* $\mathbf{A} \in \mathcal{V}$ *and any* $\langle x,y\rangle \in \beta$, $\langle q_n(x,y,y),x\rangle \in [\beta,\beta]$.

Proof Part (1) we prove by induction on $i = 0,1,\ldots,n$. It is trivial for $i = 0$. Suppose that $i \geq 0$ and $\mathcal{V} \models q_i(x,x,y) \approx y$. Then

$$\begin{aligned} q_{i+1}(x,x,y) &\approx m_{i+1}(q_i(x,x,y),x,x,q_i(x,x,y)) \quad i \text{ even or odd} \\ &\approx m_{i+1}(y,x,x,y) \\ &\approx y. \end{aligned} \tag{9.3}$$

Let β be a congruence on an algebra $\mathbf{A}$ in $\mathcal{V}$ and let $x\beta y$. We will prove by induction that

$$q_i(x,y,y)[\beta,\beta]m_i(y,y,x,x) \text{for even } i \text{ and} \tag{9.4}$$

$$q_i(x,y,y)[\beta,\beta]m_i(y,y,y,x) \text{for odd } i. \tag{9.5}$$

This will be sufficient. The case of $i = 0$ is trivial. So suppose first that i is even and $q_i(x,y,y)[\beta,\beta]m_i(y,y,x,x)$. It follows that

$$\begin{aligned} q_{i+1}(x,y,y) = m_{i+1}(q_i(x,y,y),x,y,q_i(x,y,y)) \\ [\beta,\beta]m_{i+1}(m_i(y,y,x,x),x,y,m_i(y,y,x,x)). \end{aligned} \tag{9.6}$$

Also note that

$$\begin{aligned} m_{i+1}(m_i(y,y,x,x),x,\underline{x},m_i(y,y,x,x)) &= m_i(y,y,x,x) \\ &= m_{i+1}(y,y,x,x) \\ &= m_{i+1}(m_i(y,y,y,y),y,\underline{x},m_i(x,x,x,x)). \end{aligned} \tag{9.7}$$

By centrality, we can replace the underlined variable with the β-equivalent y and maintain equivalence modulo $[\beta,\beta]$. Thus

$$m_{i+1}(m_i(y,y,x,x),x,\underline{y},m_i(y,y,x,x))[\beta,\beta]m_{i+1}(m_i(y,y,y,y),y,\underline{y},m_i(x,x,x,x)). \tag{9.8}$$

Combining this with (9.6) gives

$$q_{i+1}(x,y,y)[\beta,\beta]m_{i+1}(y,y,y,x). \tag{9.9}$$

The case when i is odd is similar. □

Part (1) of the lemma tells us that q_n obeys half of Maltsev's equations. In the case when $\mathbf{A}$ is Abelian, we can take β in Part (2) to be 1_A and get the other half of the equations. Once

we have the Maltsev operation, we can prove that the Abelian algebra is affine. Any ternary term satisfying the two properties above for q_n is called a Gumm difference term. The Gumm difference term is all we need to conclude that any Abelian algebra in a congruence modular variety is congruence permutable. A weak difference term for a variety $\mathcal{V}$ is a term d so that whenever θ is a congruence on an algebra in $\mathcal{V}$ and $a\theta b$, then

$$d(b,b,a)[\theta,\theta]a[\theta,\theta]d(a,b,b).$$

The presence of just a weak difference term would be enough to conclude that all Abelian algebras are affine. In [28], K. Kearnes and A. Szendrei prove that having a weak difference term is equivalent to a Maltsev condition. In fact, any variety in which congruence lattices satisfy a nontrivial lattice equation has such a term.

9.4 Corollary (C. Herrmann [20]) *If* $\mathbf{A}$ *is an Abelian algebra in a congruence modular variety, then any Gumm difference term of* $\mathbf{A}$ *is a Maltsev operation. In particular, every Abelian algebra in a congruence modular variety has permuting congruences.*

A little more generally:

9.5 Corollary *If* β *is a congruence of an algebra* $\mathbf{A}$ *in a congruence modular variety and* $[\beta,\beta]=0_A$*, then every congruence of* $\mathbf{A}$ *permutes with* β*.*

Proof Suppose that β is a congruence on $\mathbf{A}$ satisfying $[\beta,\beta]=0_A$ and α is any congruence on $\mathbf{A}$. We will prove that $\alpha\circ\beta=\beta\circ\alpha$. Suppose that $\langle x,z\rangle\in\beta\circ\alpha$. There is some $y\in\mathbf{A}$ with $x\beta y\alpha z$. Denote the Gumm difference term of $\mathbf{A}$ by d. Then $d(y,y,z)=z$ and since $[\beta,\beta]=0_A$, $d(x,y,y)=x$. Therefore,

$$x=d(x,y,y)\alpha d(x,y,z)\beta d(y,y,z)=z.$$

We have shown that $\beta\circ\alpha\subseteq\alpha\circ\beta$. It follows that α and β commute as in the proof of Theorem 6.1. □

Before we prove that Abelian algebras in a congruence modular variety are affine, we need to know a few things about Abelian algebras with Maltsev terms.

9.6 Definition Suppose that $f(x_1,\dots,x_n)$ and $g(x_1,\dots,x_m)$ are operations on a set A. Then f and g commute if they satisfy the equation

$$f(g(x_1^1,\dots,x_m^1),g(x_1^2,\dots,x_m^2),\dots,g(x_1^n,\dots,x_m^n))$$
$$=g(f(x_1^1,\dots,x_1^n),f(x_2^1,\dots,x_2^n),\dots,f(x_m^1,\dots,x_m^n)).$$

Commutativity of operations can be viewed more easily using matrices. Consider an $m\times n$ matrix with entries from A:

$$\begin{pmatrix} x_1^1 & x_1^2 & \cdots & x_1^n \\ x_2^1 & x_2^2 & \cdots & x_2^n \\ & & \vdots & \\ x_m^1 & x_m^2 & \cdots & x_m^n \end{pmatrix}.$$

We could apply g to each column of this matrix and then evaluate f at the resulting vector, or we could evaluate f along each row and apply g to the result. If f and g commute, these will be the same.

9.7 Definition A ternary Abelian group is an algebra with a single ternary basic operation which satisfies Maltsev's equations and commutes with itself.

9.8 Theorem (H.-P. Gumm [13]) *Suppose that* $\mathbf{A} = \langle A, d\rangle$ *is an algebra with a single ternary basic operation. The following are equivalent.*

(1) $\mathbf{A}$ *is a ternary Abelian group.*

(2) *There is an Abelian group* $\langle A, +, -, 0\rangle$ *with universe* A *so that* $d(x, y, z) = x - y + z$.

Proof If there is such a group, it is easy to check that $d(x, y, z) = x - y + z$ is a Maltsev operation which commutes with itself, so $\mathbf{A}$ is a ternary Abelian group. On the other hand, suppose that $\mathbf{A}$ is a ternary Abelian group. Let $0 \in A$ be arbitrary and define $x+y = d(x, 0, y)$ and $-x = d(0, x, 0)$. It is routine to check that these operations make $\langle A, +, -, 0\rangle$ an Abelian group and that $d(x, y, z) = x - y + z$. □

9.9 Theorem *The following are equivalent for any algebra* $\mathbf{A}$.

(1) $\mathbf{A}$ *is affine.*

(2) $\mathbf{A}$ *has a Maltsev polynomial and satisfies* $C(1_A, 1_A; 0_A)$.

(3) $\mathbf{A}$ *has a Maltsev polynomial which commutes with every polynomial operation of* $\mathbf{A}$.

(4) $\mathbf{A}$ *has a Maltsev term which commutes with every term operation of* $\mathbf{A}$.

(5) $\mathbf{A}$ *has a Maltsev term and is Abelian.*

Proof Suppose that $\mathbf{A}$ is affine. Then $\mathbf{A}$ is polynomially equivalent to a module. The proof that $C(1_A, 1_A; 0_A)$ can be gleaned from the discussion on modules in Section 5. Thus (1) ⇒ (2).

Suppose now that $\mathbf{A}$ has a Maltsev polynomial m and that $C(1_A, 1_A; 0_A)$. We will prove that m commutes with every polynomial operation of $\mathbf{A}$. Suppose that t is an $(n + m)$-ary term of $\mathbf{A}$, $\mathbf{x}, \mathbf{y}, \mathbf{z} \in A^n$, and $\mathbf{c} \in A^m$. Then

$$m(t(\mathbf{y}, \mathbf{c}), t(\mathbf{y}, \mathbf{c}), t(\mathbf{z}, \mathbf{c})) = m(t(\mathbf{z}, \mathbf{c}), t(\mathbf{y}, \mathbf{c}), t(\mathbf{y}, \mathbf{c})).$$

Then

$$\begin{aligned} m(t(m(\underline{y_0}, y_0, y_0), \ldots, m(\underline{y_{n-1}}, y_{n-1}, y_{n-1}), \mathbf{c}), t(\mathbf{y}, \mathbf{c}), t(\mathbf{z}, \mathbf{c})) \\ = m(t(m(\underline{y_0}, y_0, z_0), \ldots, , m(\underline{y_{n-1}}, y_{n-1}, z_{n-1}), \mathbf{c}), t(\mathbf{y}, \mathbf{c}), t(\mathbf{y}, \mathbf{c})). \end{aligned}$$

Applying $C(1_A, 1_A; 0_A)$, we can replace the underlined variables with corresponding x's to get

$$\begin{aligned} m(t(m(\underline{x_0}, y_0, y_0), \ldots, m(\underline{x_{n-1}}, y_{n-1}, y_{n-1}), \mathbf{c}), t(\mathbf{y}, \mathbf{c}), t(\mathbf{z}, \mathbf{c})) \\ = m(t(m(\underline{x_0}, y_0, z_0), \ldots, m(\underline{x_{n-1}}, y_{n-1}, z_{n-1}), \mathbf{c}), t(\mathbf{y}, \mathbf{c}), t(\mathbf{y}, \mathbf{c})). \end{aligned}$$

Maltsev's equations now give

$$m(t(\mathbf{x}, \mathbf{c}), t(\mathbf{y}, \mathbf{c}), t(\mathbf{z}, \mathbf{c})) = t(m(x_0, y_0, z_0), \ldots, m(x_{n-1}, y_{n-1}, z_{n-1}), \mathbf{c}).$$

Thus m commutes with any polynomial, so (2) $\Rightarrow$ (3).

Now suppose that $\mathbf{A}$ has a Maltsev polynomial m which commutes with every polynomial of $\mathbf{A}$. We know immediately that m commutes with every term of $\mathbf{A}$ and with itself. We need only to prove that m is a term operation of $\mathbf{A}$. Since m is a polynomial of $\mathbf{A}$, there is a term operation S of $\mathbf{A}$ and elements $a_1, \ldots, a_n \in A$ so that for any $x, y, z \in A$

$$m(x, y, z) = S(x, y, z, a_1, \ldots, a_n).$$

Let $x, y, z \in A$. We will express $m(x, y, z)$ as a term evaluated only at x, y, and z. Let $0 \in A$ be arbitrary and define $\alpha = S(0, 0, 0, y, \ldots, y)$. We have:

$$\begin{aligned}
m(x,y,z) &= m(m(x,y,z), m(0,0,0), m(\alpha,\alpha,0)) \\
&= m(m(x,0,\alpha), m(y,0,\alpha), m(z,0,0)) \\
&= m(m(x,0,\alpha), m(y,0,\alpha), z) \\
&= m(m(x,0,\alpha), m(m(y,0,\alpha), m(y,0,\alpha), m(y,0,\alpha)), \\
&\quad m(z, m(y,0,\alpha), m(y,0,\alpha))) \\
&= m(m(x, m(y,0,\alpha), z), m(0, m(y,0,\alpha), m(y,0,\alpha)), m(\alpha, \\
&\quad m(y,0,\alpha), m(y,0,\alpha))) \\
&= m(m(x, m(y,0,\alpha), z), 0, \alpha) \\
&= m(m(x, m(m(y,y,y), m(0,0,0), \alpha), z), 0, \alpha) \\
&= m(m(x, m(S(y,y,y,a_1,\ldots,a_n), S(0,0,0,a_1,\ldots,a_n), \\
&\quad S(0,0,0,y,\ldots,y)), z), 0, \alpha) \\
&= m(m(x, S(m(y,0,0), m(y,0,0), m(y,0,0), m(a_1,a_1,y), \ldots, \\
&\quad m(a_n,a_n,y)), z), 0, \alpha) \\
&= m(m(x, S(y,y,y,y,\ldots,y), z), 0, \alpha) \\
&= m(S(x, S(y,y,y,y,\ldots,y), z, a_1, \ldots, a_n), S(0,0,0,a_1,\ldots,a_n), \\
&\quad S(0,0,0,y,\ldots,y) \\
&= S(m(x,0,0), m(S(y,y,y,y,\ldots,y),0,0), m(z,0,0), m(a_1,a_1,y), \ldots, \\
&\quad m(a_n,a_n,y)) \\
&= S(x, S(y,y,y,y,\ldots,y), z, y, \ldots, y).
\end{aligned} \tag{9.10}$$

Thus, m is actually a term and we have established that (3) $\Rightarrow$ (4).

Assume now that $\mathbf{A}$ has a Maltsev term m which commutes with every term operation of $\mathbf{A}$. We will prove that $\mathbf{A}$ is affine. Let $0 \in A$ be arbitrary and define $x + y = m(x, 0, x)$ and $-x = m(0, x, 0)$. Then by Theorem 9.8, $\hat{\mathbf{A}} = \langle A, +, -, 0\rangle$ is an Abelian group. We will define a ring $\mathbf{R}$ so that $\hat{\mathbf{A}}$ becomes an $\mathbf{R}$-module. Let R be the set of all unary polynomials of $\mathbf{A}$ which fix the element 0. R is nonempty since the unary projection operation and the constant 0 are both in R. Since m commutes with the terms of $\mathbf{A}$ and is idempotent, m also commutes with the polynomials of $\mathbf{A}$. Therefore, m commutes with each $r \in R$. Since each $r \in R$ fixes 0, it follows that each r is an endomorphism of $\hat{\mathbf{A}}$. Notice that R is closed under the operations of the ring $\operatorname{End}\hat{\mathbf{A}}$. This is because R contains the identity of $\operatorname{End}\hat{\mathbf{A}}$ (the unary projection) and the zero of $\operatorname{End}\hat{\mathbf{A}}$ (the constant 0), and because for each $r, s \in R$ the operations $r + s$, $-r$, and $rs = r \circ s$ are unary polynomials of $\mathbf{A}$ which fix 0. Thus R is the universe of a subring

$\mathbf{R}$ of End $\hat{\mathbf{A}}$. Since $\mathbf{R} \subseteq \operatorname{End} \hat{\mathbf{A}}$, we have the natural structure of $\hat{\mathbf{A}}$ as an $\mathbf{R}$ module. We will denote this module as ${}_{\mathbf{R}}\mathbf{A}$. We claim that $\mathbf{A}$ and ${}_{\mathbf{R}}\mathbf{A}$ are polynomially equivalent. We must show that $\operatorname{Pol}_{\mathbf{R}} \mathbf{A} = \operatorname{Pol} \mathbf{A}$. That $\operatorname{Pol}_{\mathbf{R}} \mathbf{A} \subseteq \operatorname{Pol} \mathbf{A}$ should be clear. We will prove inductively on rank that every polynomial $p(x_0, \dots, x_{n-1})$ of $\mathbf{A}$ is a polynomial of ${}_{\mathbf{R}}\mathbf{A}$. Suppose first that p is a unary polynomial of $\mathbf{A}$. Let $r(x) = p(x) - p(0)$. Then r is a unary polynomial of $\mathbf{A}$ which fixes 0 so $r \in \mathbf{R}$. Moreover, $p(x) = p(x) - p(0) + p(0) = r(x) + p(0)$. If $c = p(0)$, then $p(x) = rx + c$ is a polynomial of ${}_{\mathbf{R}}\mathbf{A}$ as desired. Next, assume that any n-ary polynomial of $\mathbf{A}$ is in $\operatorname{Pol}_{\mathbf{R}} \mathbf{A}$, and let p be an $(n+1)$-ary polynomial of $\mathbf{A}$. Then

$$\begin{aligned} p(x_0, \dots, x_n) &= p(m(x_0, 0, 0), \dots, m(x_{n-1}, 0, 0), m(0, 0, x_n)) \\ &= m(p(x_0, \dots, x_{n-1}, 0), p(0, \dots, 0), p(0, \dots, 0, x_n)) \\ &= p(x_0, \dots, x_{n-1}, 0) - p(0, \dots, 0) + p(0, \dots, 0, x_n). \end{aligned} \tag{9.11}$$

Now, $p(x_0, \dots, x_{n-1}, 0)$ is an n-ary polynomial of $\mathbf{A}$, and $p(0, \dots, 0, x_n)$ is a unary polynomial of $\mathbf{A}$. As such, each of these is a polynomial of ${}_{\mathbf{R}}\mathbf{A}$. Since $p(0, \dots, 0)$ is a constant, this makes p a polynomial of ${}_{\mathbf{R}}\mathbf{A}$. This proves that $\operatorname{Pol}_{\mathbf{R}} \mathbf{A} = \operatorname{Pol} \mathbf{A}$ and completes the proof that (4) $\Rightarrow$ (1).

We have proven that (1)–(4) are equivalent. It is easy to see that these combined are equivalent to (5). □

Suppose that $\mathbf{A}$ is an Abelian algebra in a congruence modular variety. Then $\mathbf{A}$ has a Maltsev term by Corollary 9.4. By the previous theorem, $\mathbf{A}$ is affine. On the other hand, any affine algebra is Abelian; so we have the *Fundamental Theorem of Abelian Algebras*:

9.10 Theorem (C. Herrmann [20]) *An algebra in a congruence modular variety is Abelian if and only if it is affine.*

This theorem has been extended by K. Kearnes and A. Szendrei [28] to the following.

9.11 Theorem *If a variety $\mathcal{V}$ satisfies a nontrivial lattice equation as a congruence equation, then the Abelian algebras in $\mathcal{V}$ are affine.*

A congruence modular variety in which every algebra is Abelian is termed an *Abelian variety* or an *affine variety*. In Sections 13 and 14, we will need some information about these varieties. We develop that information now without giving proofs (which in every case are easy and routine). For more detail on this topic, see R. Freese, R. McKenzie [12], Chapter IX.

9.12 Definition Two varieties $\mathcal{V}$ and $\mathcal{W}$ are said to be *polynomially equivalent* if every algebra in each of the varieties is polynomially equivalent with an algebra in the other.

Suppose that $\mathcal{A}$ is a congruence modular, Abelian variety. Let $d(x, y, z)$ be a Gumm term for $\mathcal{A}$ and let $\mathbf{F}$ be the free algebra on $\mathcal{A}$ freely generated by $\{x, y\}$. Let R be the set of all $t(x, y) \in F$ such that $\mathbf{A} \models t(x, x) \approx x$. For $r = r(x, y)$ and $s = s(x, y)$ in R, put $r \circ s = r(s(x, y), y)$, $r + s = d(r(x, y), y, s(x, y))$, $-r = d(y, r(x, y), y)$, $0 = y$, $1 = x$. Then $\mathbf{R} = \langle R, +, \circ, 0, 1 \rangle$ is a ring with unit and we have the *Fundamental Theorem of Affine Varieties*:

9.13 Theorem *If a variety $\mathcal{A}$ is congruence modular and Abelian, then $\mathcal{A}$ is polynomially equivalent with the variety ${}_{\mathbf{R}}\mathcal{M}$ of unitary left $\mathbf{R}$-modules where $\mathbf{R}$ is the ring of idempotent binary terms of $\mathbf{A}$ defined above.*

In fact, if $\mathbf{A} \in \mathcal{A}$, then the universe of $\mathbf{A}$ becomes an $\mathbf{R}$-module, denoted ${}_{\mathbf{R}}\mathbf{A}$, by choosing some element $a \in A$ and putting $0 = a$, $rb = r^{\mathbf{A}}(b, 0)$, $b + c = d^{\mathbf{A}}(b, 0, c)$, and $-b = d^{\mathbf{A}}(0, b, 0)$ for $\{b, c\} \subseteq A$ and $r \in R$. This module is, up to isomorphism, independent of the choice of 0 in A. The two algebras $\mathbf{A}$ and ${}_{\mathbf{R}}\mathbf{A}$ are polynomially equivalent. The Gumm term comes out to be $d^{\mathbf{A}}(x, y, z) = x - y + z$ (evaluated in the module). The passage from $\mathbf{A}$ to ${}_{\mathbf{R}}\mathbf{A}$ is generally many-to-one—i.e., ${}_{\mathbf{R}}\mathbf{A}_0 \cong {}_{\mathbf{R}}\mathbf{A}_1$ does not imply $\mathbf{A}_0 \cong \mathbf{A}_1$. But every $\mathbf{R}$-module occurs as ${}_{\mathbf{R}}\mathbf{A}$ for some $\mathbf{A} \in \mathcal{A}$.

Every algebra $\mathbf{A} \in \mathcal{A}$ is closely associated with an algebra $\mathbf{A}_\nabla$ with a one-element subalgebra (although $\mathbf{A}$ need not have any one-element subalgebra). Namely, $\mathbf{A}_\nabla = (\mathbf{A} \times \mathbf{A})/\Delta_{1,1}$ with $\Delta_{1,1}$ the congruence on $\mathbf{A} \times \mathbf{A}$ generated by $\{\langle\langle x, x\rangle, \langle y, y\rangle\rangle : \{x, y\} \subseteq A\}$. One verifies that if we choose $0 = a$ in ${}_{\mathbf{R}}\mathbf{A}$ and $0 = \langle a, a\rangle$ in ${}_{\mathbf{R}}(\mathbf{A} \times \mathbf{A})$, then ${}_{\mathbf{R}}(\mathbf{A} \times \mathbf{A}) = {}_{\mathbf{R}}\mathbf{A} \times {}_{\mathbf{R}}\mathbf{A}$. Since $\mathbf{A} \times \mathbf{A}$ and ${}_{\mathbf{R}}(\mathbf{A} \times \mathbf{A})$ have the same congruences (being polynomially equivalent), it is easily seen that $\Delta_{1,1} = \{\langle\langle x, y\rangle, \langle u, v\rangle\rangle \in A^2 \times A^2 : x - y = u - v\}$. Then ${}_{\mathbf{R}}\mathbf{A}_\nabla \cong {}_{\mathbf{R}}\mathbf{A}$ while $\mathbf{A}_\nabla$ has the one-element subalgebra $\nabla = \{\langle x, x\rangle : x \in A\}$.

10 Solvability and nilpotence

We can use the commutator to extend the ideas of solvability and nilpotence to congruence modular varieties. Before we tackle this topic, we will derive H.-P. Gumm's Maltsev characterization of congruence modularity—which is (relatively) easy to do thanks to the Gumm difference term supplied by W. Taylor's equations.

10.1 Theorem (H.-P. Gumm [16]) *A variety $\mathcal{V}$ is congruence modular if and only if $\mathcal{V}$ has ternary terms $d_0, \ldots, d_n$ and q satisfying*

$$\begin{aligned}
x &\approx d_0(x, y, z) \\
x &\approx d_i(x, y, x) \quad \textit{for all } i \\
d_i(x, x, z) &\approx d_{i+1}(x, x, z) \quad \textit{for even } i \\
d_i(x, z, z) &\approx d_{i+1}(x, z, z) \quad \textit{for odd } i \\
d_n(x, z, z) &\approx q(x, z, z) \\
q(x, x, z) &\approx z.
\end{aligned}$$

Proof Suppose that $\mathcal{V}$ is a congruence modular variety, and let q be a Gumm difference term for $\mathcal{V}$. Let $\mathbf{F}$ be the free algebra in $\mathcal{V}$ on the generators $\{x, y, z\}$. Let $\alpha = \mathrm{Cg}_{\mathbf{F}}(x, y)$, $\beta = \mathrm{Cg}_{\mathbf{F}}(y, z)$, and $\gamma = \mathrm{Cg}_{\mathbf{F}}(x, z)$. Then $\langle x, z\rangle \in \gamma \cap (\alpha \vee \beta)$. Now by the property of the difference term, $\langle x, q(x, z, z)\rangle \in [\gamma \cap (\alpha \vee \beta), \gamma \cap (\alpha \vee \beta)]$. However

$$\begin{aligned}
[\gamma \cap (\alpha \vee \beta), \gamma \cap (\alpha \vee \beta)] &\leq [\gamma, \alpha \vee \beta] \\
&= [\gamma, \alpha] \vee [\gamma, \beta] \\
&\leq (\gamma \cap \alpha) \vee (\gamma \cap \beta).
\end{aligned} \tag{10.1}$$

Thus $\langle x, d(x, z, z)\rangle \in (\gamma \cap \alpha) \vee (\gamma \cap \beta)$. This means there are $u_0, \ldots, u_n \in A$ with $u_0 = x$, $u_n = q(x, z, z)$, $u_i \alpha u_{i+1}$ if i is even, $u_i \beta u_{i+1}$ if i is odd, and $u_i \gamma u_{i+1}$ for all i. There are

ternary terms $d_0, \ldots, d_n$ so that $u_i = d_i(x, y, z)$ for each i. Now, $x = d_0(x, y, z)$ holds by design. Define $f, g, h : \mathbf{F} \to \mathbf{F}$ to be the unique homomorphisms defined by

$$f(x) = f(y) = x, \qquad f(z) = z \tag{10.2}$$
$$g(x) = x, \qquad g(y) = g(z) = z \tag{10.3}$$
$$h(x) = h(z) = x, \qquad h(y) = y. \tag{10.4}$$

Then $\ker f = \alpha$, $\ker g = \beta$, and $\ker h = \gamma$. For any i, notice that

$$\begin{aligned} x &= d_0(x, y, x) \\ &= d_0(h(x), h(y), h(z)) \\ &= h(d_0(x, y, z)) \\ &= h(d_i(x, y, z)) \\ &= d_i(h(x), h(y), h(z)) \\ &= d_i(x, y, x). \end{aligned} \tag{10.5}$$

Suppose now that $i < n$ is even. Then

$$\begin{aligned} d_i(x, x, z) &= d_i(f(x), f(y), f(z)) \\ &= f(d_i(x, y, z)) \\ &= f(d_{i+1}(x, y, z)) \\ &= d_{i+1}(f(x), f(y), f(z)) \\ &= d_{i+1}(x, x, z). \end{aligned} \tag{10.6}$$

If $i < n$ is odd then

$$\begin{aligned} d_i(x, z, z) &= d_i(g(x), g(y), g(z)) \\ &= g(d_i(x, y, z)) \\ &= g(d_{i+1}(x, y, z)) \\ &= d_{i+1}(g(x), g(y), g(z)) \\ &= d_{i+1}(x, z, z). \end{aligned} \tag{10.7}$$

Finally, $d_n(x, z, z) = q(x, z, z)$ by design, and $q(x, x, z) = z$ since q is a Gumm difference term. Since these terms satisfy these equations in $\mathbf{F}$, they satisfy the equations throughout $\mathcal{V}$.

Finally, assume that $\mathcal{V}$ has terms $d_0, \ldots, d_n, q$ satisfying the equations in this theorem. If n were even, then the equations would imply that $d_{n-1}(x, z, z) \approx q(x, z, z)$ also, so we can assume that n is odd. Define 4-ary terms $m_0, \ldots, m_{2n+2}$ in the following manner. $m_0(x, y, z, u) = x$ and

$$m_{2i-1}(x, y, z, u) = d_i(x, y, u) \quad \text{for } i \text{ odd} \tag{10.8}$$
$$m_{2i-1}(x, y, z, u) = d_i(x, z, u) \quad \text{for } i \text{ even} \tag{10.9}$$
$$m_{2i}(x, y, z, u) = d_i(x, z, u) \quad \text{for } i \text{ odd} \tag{10.10}$$
$$m_{2i}(x, y, z, u) = d_i(x, y, u) \quad \text{for } i \text{ even.} \tag{10.11}$$

Also, let $m_{2n+1} = q(y, z, u)$ and $m_{2n+2}(x, y, z, u) = u$. It is routine now to check that these are Day terms for $\mathcal{V}$. Hence $\mathcal{V}$ is congruence modular. □

H.-P. Gumm's Maltsev condition for congruence modularity may be viewed as a composition of B. Jónsson's condition for congruence distributivity and A. I. Maltsev's condition for congruence permutability. Notice that in the proof of Gumm's theorem, the term q was chosen to be the difference term of $\mathcal{V}$. Thus every difference term has Gumm terms associated with it. On the other had, every q arising from the Gumm terms is a difference term. Hence

10.2 Theorem *The following are equivalent for any ternary term q in a variety $\mathcal{V}$.*

(1) *$\mathcal{V}$ is congruence modular; $\mathcal{V} \models q(x, x, y) \approx y$; and for all $\mathbf{A} \in \mathcal{V}$, for all $\beta \in \operatorname{Con}\mathbf{A}$, for all $\langle a, b\rangle \in \beta$, $\langle a, q(a, b, b)\rangle$ is in $[\beta, \beta]$.*

(2) *For this particular q, there exist ternary terms $d_0, \ldots, d_n$ satisfying Gumm's equations for congruence modularity.*

Proof All we need to prove is that if $d_0, \ldots, d_n, q$ are Gumm terms for a variety $\mathcal{V}$, then q is a difference term. Suppose that β is a congruence on an algebra $\mathbf{A}$ in $\mathcal{V}$ and that $a\beta b$ in $\mathbf{A}$. We only need to establish that $\langle a, q(a, b, b)\rangle \in [\beta, \beta]$ since the other equation is part of Gumm's equations. We will first establish $\langle d_i(a, b, b), d_i(a, a, b)\rangle \in [\beta, \beta]$ for all i. This is true by centrality since

$$\begin{pmatrix} d_i(a,b,a) & d_i(a,a,a) \\ d_i(a,b,b) & d_i(a,a,b) \end{pmatrix} = \begin{pmatrix} a & a \\ d_i(a,b,b) & d_i(a,a,b) \end{pmatrix} \tag{10.12}$$

is in $M(\beta, \beta)$. Next, we establish $\langle d_i(a, b, b), d_{i+1}(a, b, b)\rangle \in [\beta, \beta]$ for all i. If i is odd, this is actually an equality, so there is nothing to show. If i is even then

$$d_i(a, b, b)[\beta, \beta]d_i(a, a, b) = d_{i+1}(a, a, b)[\beta, \beta]d_{i+1}(a, b, b). \tag{10.13}$$

Now it follows that

$$a = d_0(a, b, b)[\beta, \beta]d_n(a, b, b) = q(a, b, b). \tag{10.14}$$

□

10.3 Definition Suppose that β is a congruence on an algebra $\mathbf{A}$ in a congruence modular variety. Define $(\beta]^0, (\beta]^1, (\beta]^2, \ldots$ recursively as follows. First, $(\beta]^0 = \beta$. Next, if $(\beta]^n$ is defined, then $(\beta]^{n+1} = [\beta, (\beta]^n]$. If $(\beta]^n = 0$ for some n, then β is n-step nilpotent. If 1_A is n-step nilpotent, then $\mathbf{A}$ is also called n-step nilpotent. Also define the sequence $[\beta]^0$, $[\beta]^1$, $[\beta]^2, \ldots$ recursively by $[\beta]^0 = \beta$ and $[\beta]^{n+1} = [[\beta]^n, [\beta]^n]$. If $[\beta]^n = 0_A$ for some n, then β is n-step solvable. If 1_A is n-step solvable, then $\mathbf{A}$ is also called n-step solvable.

10.4 Definition Suppose that q is a Gumm difference term for a congruence modular variety $\mathcal{V}$. Define a sequence of ternary terms $q_0, q_1, q_2, \ldots$ recursively by $q_0 = q$ and $q_{n+1}(x, y, z) = q_0(x, q_n(x, y, y), q_n(x, y, z))$. We will call these terms *generalized Gumm terms*.

10.5 Theorem (H.-P. Gumm [17]) *Suppose that α and β are congruences on an algebra $\mathbf{A}$ in a congruence modular variety. Then $\alpha \circ \beta \subseteq [\alpha]^n \circ \beta \circ \alpha$ and $\alpha \circ \beta \subseteq (\alpha]^n \circ \beta \circ \alpha$ for all n.*

Proof We prove this by induction on n. For $n = 0$, $[\alpha]^n = \alpha$, so the inclusion is trivial. Suppose that the inclusion holds for $n \geq 0$ and suppose that $a\alpha b\beta c$. By our induction hypothesis, there are x and y so that $a[\alpha]^n x\beta y\alpha c$. Let q be the Gumm difference term of $\mathbf{A}$. Then

$$\langle a, q(a,x,x)\rangle \in [[\alpha]^n, [\alpha]^n] \subseteq [\alpha]^{n+1}\,. \tag{10.15}$$

Therefore

$$a[\alpha]^{n+1} q(a,x,x)\beta q(a,x,y)\alpha q(x,x,c) = c \tag{10.16}$$

so $\langle a, c\rangle \in [\alpha]^{n+1} \circ \beta \circ \alpha$. The other claim in the theorem is proved similarly. □

If α is n-step solvable, then $[\alpha]^n = 0$, so this theorem gives $\alpha \circ \beta \subseteq \beta \circ \alpha$. It follows that α and β permute (as in the proof of Theorem 6.1). Hence

10.6 Corollary *Suppose that α is an n-step solvable (or nilpotent) congruence of an algebra $\mathbf{A}$ in a congruence modular variety. Then α permutes with every congruence of $\mathbf{A}$.*

If $\mathbf{A}$ is n-step solvable, then every congruence on $\mathbf{A}$ is n-step solvable, so $\mathbf{A}$ has permuting congruences. Moreover, we can adapt the above proof using the generalized Gumm terms to manufacture a Maltsev term for $\mathbf{A}$ so that the entire variety generated by $\mathbf{A}$ has permuting congruences.

10.7 Theorem *Suppose that $\mathbf{A}$ is an algebra in a congruence modular variety with generalized Gumm terms $q_0, q_1, q_2, \ldots$. If $\mathbf{A}$ is $(n+1)$-step solvable (or nilpotent) for some n, then $\mathbf{A}$ has permuting congruences, and the term q_n is a Maltsev term for $\mathbf{A}$.*

Proof Let $a, b \in \mathbf{A}$. We will first prove that $\langle a, q_k(a,b,b)\rangle \in [1_A]^{k+1}$ for all k. We proceed by induction on k. Since q_0 is a difference term and $\langle a, b\rangle \in 1_A$, we clearly have $\langle a, q_0(a,b,b)\rangle \in [1_A, 1_A] = [1_A]^1$. Now assume that $k \geq 0$ and that $\langle a, q_k(a,b,b)\rangle \in [1_A]^{k+1}$. Since q_0 is a difference term, $\langle a, q_0(a, q_k(a,b,b), q_k(a,b,b))\rangle \in [[1_A]^{k+1}, [1_A]^{k+1}] = [1_A]^{k+2}$, as desired. We have proved that $\langle a, q_k(a,b,b)\rangle \in [1_A]^{k+1}$ for all k. Since $[1_A]^{n+1} = 0_A$, this means that $a = q_n(a,b,b)$.

Next, we will prove by induction that $q_k(a,a,b) = b$ for all k. This is true for $k = 0$ since q_0 is a difference term. Assume that $k \geq 0$ and that $q_k(a,a,b) = b$. Then $q_{k+1}(a,a,b) = q_0(a, q_k(a,a,a), q_k(a,a,b)) = q_0(a,a,b) = b$. We have shown that $q_k(a,a,b) = b$ for all k. In particular $q_n(a,a,b) = b$.

We have proven that for arbitrary $a, b \in A$, $q_n(a,b,b) = a$ and $q_n(a,a,b) = b$. Therefore, q_n is a Maltsev term for $\mathbf{A}$ and the result follows. The claim for nilpotence is proven in a similar manner. □

10.8 Theorem *The class of n-step solvable (nilpotent) algebras in a congruence modular variety $\mathcal{V}$ is a variety.*

Proof We prove the theorem for the case of nilpotence. Let $\mathcal{K}$ be the class of n-step nilpotent algebras in $\mathcal{V}$. We will prove that $\mathcal{K}$ is closed under homomorphic images, subalgebras, and products.

Suppose that $\mathbf{A} \in \mathcal{K}$ and that $f : \mathbf{A} \to \mathbf{B}$ is a surjective homomorphism with $\ker f = \pi$. We will prove by induction on k that $(1_B]^k = f((1_A]^k \vee \pi)$ for all k. This is trivial for $k = 0$,

so assume that $k \geq 0$ and that $(1_B]^k = f((1_A]^k \vee \pi)$. Then

$$\begin{aligned}(1_B]^{k+1} &= [1_B, (1_B]^k] \\ &= [f(1_A \vee \pi), f((1_A]^k \vee \pi)] \\ &= f([1_A, (1_A]^k] \vee \pi) \\ &= f((1_A]^{k+1} \vee \pi).\end{aligned} \tag{10.17}$$

It now follows that $(1_B]^n = f((1_A]^n \vee \pi) = f(\pi) = 0_B$ so $\mathbf{B}$ is n-step nilpotent and is in $\mathcal{K}$.

Next, suppose that $\mathbf{A} \in \mathcal{K}$ and that $\mathbf{B}$ is a subalgebra of $\mathbf{K}$. For any congruences $\alpha, \beta, \delta \in \operatorname{Con}\mathbf{A}$, it should be clear that $C(\alpha, \beta; \delta)$ (in $\mathbf{A}$) implies that $C(\alpha \cap B^2, \beta \cap B^2; \delta \cap B^2)$ (in $\mathbf{B}$). Therefore $[\alpha \cap B^2, \beta \cap B^2] \subseteq [\alpha, \beta] \cap B^2$. It follows that $(1_B]^n \subseteq (1_A]^n \cap B^2 = 0_B$, so $\mathbf{B}$ is n-step nilpotent and $\mathbf{B} \in \mathcal{K}$.

Finally, suppose that $\{\mathbf{A}_i : i \in I\} \subseteq \mathcal{K}$. Let $\mathbf{B} = \prod_{i \in I} \mathbf{A}_i$. Let $\pi_i : \mathbf{B} \to \mathbf{A}_i$ be the canonical projection and let $\eta_i = \ker \pi_i$. As we showed above, $\pi_i((1_B]^n \vee \eta_i) = (1_A]^n = 0_A$. Thus, $(1_B]^n \subseteq \eta_i$ for all i. Therefore, $(1_B]^n = 0_B$. Thus $\mathbf{B}$ is n-step nilpotent and is in $\mathcal{K}$. This finishes the proof that $\mathcal{K}$ is a variety. The proof of the theorem for solvable algebras is similar. □

To prove the next lemma, we need the following commutativity result.

10.9 Lemma *Suppose that α and δ are congruent on an algebra $\mathbf{A}$ with a Maltsev term m and that $C(\alpha, 1_A; \delta)$. If $\mathbf{a}, \mathbf{b}, \mathbf{c} \in A^3$ with $a_i \alpha b_i$ for all i then*

$$m(m(a_0, b_0, c_0), m(a_1, b_1, c_1), m(a_2, b_2, c_2)) \delta m(m(a_0, a_1, a_2), m(b_0, b_1, b_2), m(c_0, c_1, c_2)). \tag{10.18}$$

Proof Maltsev's equations give us

$$m(m(b_0, b_1, b_2), m(b_0, b_1, b_2), m(c_0, c_1, c_2)) = m(m(c_0, c_1, c_2), m(b_0, b_1, b_2), m(b_0, b_1, b_2)).$$

We expand the first subterm of each side of this equality to get

$$\begin{aligned}&m(m(m(\underline{b_0}, b_0, b_0), m(\underline{b_1}, b_1, b_1), m(\underline{b_2}, b_2, b_2)), m(b_0, b_1, b_2), m(c_0, c_1, c_2)) \\ &\qquad = m(m(m(\underline{b_0}, b_0, c_0), m(\underline{b_1}, b_1, c_1), m(\underline{b_2}, b_2, c_2)), m(b_0, b_1, b_2), m(b_0, b_1, b_2)).\end{aligned}$$

By centrality, we can replace the underlined b_i's with corresponding a_i's and maintain equivalence modulo δ so that

$$\begin{aligned}&m(m(m(\underline{a_0}, b_0, b_0), m(\underline{a_1}, b_1, b_1), m(\underline{a_2}, b_2, b_2)), m(b_0, b_1, b_2), m(c_0, c_1, c_2)) \\ &\qquad \delta m(m(m(\underline{a_0}, b_0, c_0), m(\underline{a_1}, b_1, c_1), m(\underline{a_2}, b_2, c_2)), m(b_0, b_1, b_2), m(b_0, b_1, b_2)).\end{aligned}$$

Maltsev's equations now give

$$m(m(a_0, a_1, a_2), m(b_0, b_1, b_2), m(c_0, c_1, c_2)) \delta m(m(a_0, b_0, c_0), m(a_1, b_1, c_1), m(a_2, b_2, c_2)).$$

□

10.10 Lemma *Suppose that* $\mathbf{A}$ *is an* n*-step nilpotent algebra with a Maltsev term* m*. There are ternary terms* l *and* r *so that for all* $b, c \in A$ *the function* $l(-, b, c)$ *is the inverse of* $m(-, b, c)$ *and the function* $r(-, b, c)$ *is the inverse of* $m(c, b, -)$*.*

Proof We will prove the existence of l. The existence of r is similar. We will prove this by induction on n. If $n = 0$, then $\mathbf{A}$ is trivial. If $n = 1$, then $\mathbf{A}$ is Abelian. For any $x \in \mathbf{A}$, Define $u +_x v = m(u, x, v)$ and $-_x u = m(x, u, x)$ for all $u, v \in \mathbf{A}$. Then by Theorem 9.8 these operations define an Abelian group on A with identity x so that $m(u, v, w) = u -_x v +_x w$. For any $b, c \in A$, the inverse of $m(x, b, c) = x -_x b +_x c$ is clearly $l(x, b, c) = x -_x c +_x b = m(x, c, b)$.

Next suppose that $n \geq 1$ and that the lemma holds for n-step nilpotent algebras. Suppose that $\mathbf{A}$ is $(n + 1)$-step nilpotent. Let $\theta = (1_A)^n$. Then $C(1_A, \theta; 0)$, and A/θ is n-step nilpotent. By our induction hypothesis, there is a term l' so that $l'(-, b/\theta, c/\theta)$ is the inverse of $m(-, b/\theta, c/\theta)$ for all $b, c, \in A$. Let $l(y, b, c) = m(m(y, m(l'(y, b, c), b, c), y), y, l'(y, b, c))$. We will show that l is the desired term. Let $y, b, c \in A$. Let $z = l'(y, b, c)$. By our choice of l', we know at least that $m(z, b, c)\theta y$. Then

$$\begin{aligned} m(l(y, b, c), b, c) &= m(m(m(y, m(z, b, c), y), y, z), m(y, y, b), m(y, y, c)) \\ &= m(m(m(y, m(z, b, c), y), y, y), m(y, y, y), m(z, b, c)) \\ &= m(m(y, m(z, b, c), y), y, m(z, b, c)) \end{aligned} \tag{10.19}$$

where the second equality follows from Lemma 10.9 since $m(z, b, c)\theta y$ and since $C(\theta, 1_A; 0)$. The set y/θ is closed under m since m is idempotent, and by Lemma 10.9 m commutes with itself on this θ block. Define $u +_y v = m(u, y, v)$ and $-_y u = m(y, u, y)$ on y/θ. By Theorem 9.8, these are Abelian group operations on y/θ with y as an identity. Then we can continue our calculations to see

$$\begin{aligned} m(l(y, b, c), b, c) &= m(m(y, m(z, b, c), y), y, m(z, b, c)) \\ &= -_y m(z, b, c) +_y m(z, b, c) \\ &= y. \end{aligned} \tag{10.20}$$

Now we look at the composition in the reverse order. Let $z = l'(m(y, b, c), b, c)$. Then $z\theta y$ and

$$\begin{aligned} l(m(y, b, c), b, c) &= m(m(m(y, b, c), m(z, b, c), m(y, b, c)), m(y, b, c), z) \\ &= m(m(m(y, z, y), m(b, b, b), m(c, c, c)), m(y, b, c), z) \\ &= m(m(m(y, z, y), b, c), m(y, b, c), z) \\ &= m(m(m(y, z, y), b, c), m(y, b, c), m(y, y, z)) \\ &= m(m(m(y, z, y), y, y), m(b, b, y), m(c, c, z)) \\ &= m(m(y, z, y), y, z) \\ &= -_y z +_y z \\ &= y. \end{aligned} \tag{10.21}$$

Note that the second and fifth equalities follow from Lemma 10.9 since $z\theta y$. □

The real mechanics of this proof are hidden from sight, but there is an elegant structure to these algebras. The Maltsev term m gives each block of θ the structure of a ternary Abelian group. For any $b, c \in A$, the functions $m(-, b, c) : b/\theta \to c/\theta$ and $m(-, c, b) : c/\theta \to b/\theta$ are

inverse isomorphisms of these group structures which exchange b and c, so all of the blocks are isomorphic to a ternary Abelian group $\mathbf{G}$. The algebra $\hat{\mathbf{A}} = \langle A, m\rangle$ is a sort of semi-direct product of $\hat{\mathbf{A}}/\theta$ and $\mathbf{G}$. To find the term l, we try to solve the equation $m(x,b,c) = y$ for x, assuming that there is a solution z modulo θ. To solve this equation, we use the isomorphisms $m(-,b,y)$, $m(-,c,y)$, and $m(-,z,y)$ to map everything into y/θ. Then we can treat y/θ as an Abelian group with identity y to solve the new equation. The nature of the semi-direct product is such that when we pull this solution back to z/θ using $m(-,y,z)$ we have a solution to the original equation.

More generally, if α is any congruence on $\mathbf{A}$ and $b, c \in A$ then the maps $m(-,b,c) : b/\alpha \to c/\alpha$ and $m(-,c,b) : c/\alpha \to b/\alpha$ may not be inverses, but they are both injective. Hence the congruence classes are the same size. Thus

10.11 Corollary *Any n-step nilpotent algebra in a congruence modular variety has uniform congruences.*

Suppose that a, b and c are elements of an n-step nilpotent algebra $\mathbf{A}$ in a congruence modular variety with Maltsev term m and that $\theta \in \operatorname{Con}\mathbf{A}$. If $a\theta b$, then $m(a,b,c)\theta m(b,b,c) = c$ so $m(a,b,c)\theta c$. On the other hand, if $m(a,b,c)\theta c$, then $a = l(m(a,b,c),b,c)\theta l(c,b,c)$. However, $m(l(c,b,c),b,c) = c = m(b,b,c)$, so by Lemma 10.10 $l(c,b,c) = b$. This means that $a\theta b$. We have that $a\theta b$ if and only if $m(a,b,c)\theta c$. This means that θ consists precisely of those pairs $\langle a, b\rangle$ for which $m(a,b,c) \in c/\theta$. Hence, θ is uniquely determined by any equivalence class c/θ. Thus

10.12 Corollary *Any n-step nilpotent algebra in a congruence modular variety has regular congruences.*

Suppose now that $\mathbf{A}$ is an n-step nilpotent algebra in a congruence modular variety. Let l and r be the terms guaranteed by the previous lemma and let $0 \in A$ be arbitrary. Now define these operations on A

$$\begin{aligned} u \cdot v &= m(u,0,v) \\ u/v &= l(u,0,v) \\ u\backslash v &= r(u,0,v). \end{aligned} \tag{10.22}$$

Then these are the multiplication and division operations of a loop on A. Hence

10.13 Theorem *Suppose that $\mathbf{A}$ is an n-step nilpotent algebra in the congruence modular variety $\mathcal{V}$. There is a loop operation in* Pol $\mathbf{A}$ *whose left and right division operations are also in* Pol $\mathbf{A}$.

11 Applications

We briefly describe here some of the important results that have been achieved through the application of commutator theory in the study of basic questions about varieties. We make no attempt at completeness.

In 1979 there appeared the paper J. Hagemann, C. Herrmann [19], which provided the first proofs of much of the basic commutator theory we have developed in the preceding

sections, and the same year appeared H.-P. Gumm, C. Herrmann [18] in which the new theory was applied to obtain new cancellation, refinement and uniqueness results for direct products of algebras in congruence modular varieties. H.-P Gumm and C. Herrmann proved, among other results, that if $\mathbf{A} \times \mathbf{B} \cong \mathbf{A} \times \mathbf{C}$ and if $\mathbf{A} \times \mathbf{B}$ belongs to a congruence modular variety, and if the congruence lattice of $\mathbf{A}$ has the ascending chain condition and the center of $\mathbf{A}$ is a congruence of "finite rank", then $\mathbf{B}$ is "affine-isotopic" to $\mathbf{C}$.

In 1981 appeared R. Freese, R. McKenzie [11] in which commutator theory was used to show that every residually small congruence modular variety $\mathcal{V}$ obeys a certain commutator law (C1), which must hold in the congruence lattice of every algebra of $\mathcal{V}$ augmented by the commutator operation. Conversely, if $\mathcal{V}$ is congruence modular and generated by a finite algebra $\mathbf{A}$ and if $\mathbf{A}$, along with all of its subalgebras, obeys (C1) then $\mathcal{V}$ is residually small, in fact has a finite residual bound.

Also in 1981 appeared a monograph by S. Burris and R. McKenzie [4], containing a proof that every locally finite congruence modular variety with decidable first-order theory must decompose as the varietal product of two subvarieties, a decidable affine variety and a decidable discriminator variety. Commutator theory was the essential tool for this work. The authors also provided an algorithm which can be used to reduce the question whether $\mathrm{HSP}(\mathbf{A})$ has decidable theory, where $\mathbf{A}$ is a given finite algebra, (the decidability problem) to the question whether the variety of $\mathbf{R}$-modules has decidable theory, where $\mathbf{R}$ is a certain finite ring correlated with $\mathbf{A}$, produced by the algorithm. The problem to characterize in some fashion those finite $\mathbf{A}$ for which the class of finite members of $\mathrm{HSP}(\mathbf{A})$ has decidable theory (the finite decidability problem) seems to be much more difficult than the decidability problem. In 1997, P. M. Idziak [22] provided a solution to the finite decidability problem for finite algebras in congruence modular varieties. His result is equally as satisfactory as the result of S. Burris and R. McKenzie, but is rather more difficult to state.

In 1982 appeared R. McKenzie [32] which used commutator theory to characterize the locally finite varieties having a finite bound on the cardinalities of their finite directly indecomposable members (the "directly representable" varieties). A breakthrough result achieved in the paper was the fact that every directly representable variety has permuting congruences.

The 1987 monograph R. Freese, R. McKenzie [12], which has been the principal source for the first ten sections of this text, contains the result (Chapter XIV) that any finite nilpotent algebra of finite type (i.e., which possesses only finitely many basic operations) which lies in a congruence modular variety and decomposes as the direct product of algebras of prime-power orders, has a finitely axiomatizable equational theory. R. McKenzie [33] proves that any finite algebra $\mathbf{F}$ belonging to a residually small congruence modular variety of finite type has a finitely axiomatizable equational theory. A main ingredient in this proof is the demonstration that $\mathbf{F}$ obeys finitely many equations which collectively imply that an algebra (which satisfies them) satisfies the commutator equation (C1) discovered by R. Freese and R. McKenzie.

In 1989 appeared K. A. Kearnes [26] which used commutator theory to prove that every residually small, congruence modular variety with the amalgamation property possesses the congruence extension property. Whether the words "congruence modular" can be removed from this result is unknown.

In 1996, P. M. Idziak and J. Berman began a study of the "generative complexity" of locally finite varieties. They define the generative complexity, or G-function, of a variety $\mathcal{V}$ to be the function $\mathrm{G}_{\mathcal{V}}$ defined for all positive integers n so that $\mathrm{G}_{\mathcal{V}}(n)$ is the number of

non-isomorphic n-generated algebras in $\mathcal{V}$. Perhaps their deepest result is a characterization of finitely generated congruence modular varieties $\mathcal{V}$ for which $\mathrm{G}_{\mathcal{V}}(n) \leq 2^{cn}$ (for all $n \geq 1$) for some constant c. The characterization is of the same order as P. M. Idziak's characterization of finite decidability for $\mathrm{HSP}(\mathbf{A})$ but even more complicated. A much easier result of P. M. Idziak and R. McKenzie [23] will be proved in Section 14 below; namely, a locally finite congruence modular variety $\mathcal{V}$ satisfies $\mathrm{G}_{\mathcal{V}}(n) \leq n^c$ (for all $n \geq 1$), for a constant c, if and only if $\mathcal{V}$ is Abelian and directly representable.

Perhaps this is the appropriate place to mention the tame congruence theory of D. Hobby, R. McKenzie [21]. With this theory, it became possible to extend most of the above-mentioned results that deal with finite algebras or locally finite varieties, either to all finite algebras and all locally finite varieties, or to a domain much broader than congruence modular varieties. Sometimes, tame congruence theory simply produces the result that every locally finite variety possessing a certain property must be congruence modular. For instance, it is proved in [21], Chapter 10 that every residually small locally finite variety that satisfies any non-trivial congruence equation must be congruence modular (i.e., must satisfy the modular law as a congruence equation).

In the book R. McKenzie, M. Valeriote [34] it is proved that every locally finite variety with decidable first-order theory decomposes as the varietal product of three decidable subvarieties: an affine variety, a discriminator variety and a combinatorial variety. This result contains the result proved five years earlier by S. Burris and R. McKenzie for congruence modular varieties. In the modular case, the combinatorial variety must consist just of one-element algebras.

P. M. Idziak, R. McKenzie and M. Valeriote [24] (unpublished) have extended the above-mentioned result of P. M. Idziak and R. McKenzie into a characterization of all locally finite varieties $\mathcal{V}$ with the property that $\mathrm{G}_{\mathcal{V}}(n) \leq n^c$ for all $n \geq 1$, for some constant c. Such a variety is a varietal product of an affine, directly representable, subvariety and a very special kind of combinatorial subvariety.

Successful applications of tame congruence theory, like those mentioned in the two previous paragraphs, have frequently begun with the idea to attempt an extension of results proved earlier for locally finite congruence modular varieties with the help of commutator theory. Tame congruence theory is a powerful tool for such efforts but, unlike modular commutator theory, its application appears to be essentially restricted to the realm of locally finite varieties.

12 Residual smallness

An algebra $\mathbf{A}$ is called subdirectly irreducible if it has a smallest non-zero congruence, (called the monolith of $\mathbf{A}$). According to a theorem of G. Birkhoff, every algebra can be embedded into a product, $\prod_{t \in T} \mathbf{S}_t$, of subdirectly irreducible algebras $\mathbf{S}_t$ in such a way that it projects onto each factor $\mathbf{S}_t$ (subdirect embedding).

12.1 Definition A variety $\mathcal{V}$ is *residually small* if there is a cardinal bound on the size of subdirectly irreducible algebras in $\mathcal{V}$. If $\mathcal{V}$ is residually small, then we will write $\mathrm{resb}(\mathcal{V})$ for the least cardinal κ such that every subdirectly irreducible algebra in $\mathcal{V}$ has cardinality less than κ. If the cardinalities of subdirectly irreducible algebras in $\mathcal{V}$ have no cardinal upper bound, then we write $\mathrm{resb}(\mathcal{V}) = \infty$, and say that $\mathcal{V}$ is *residually large*. For an algebra $\mathbf{A}$, we

put $\mathrm{resb}(\mathbf{A}) = \mathrm{resb}(\mathrm{HSP}(\mathbf{A}))$. A variety $\mathcal{V}$ will be called *residually finite* if $\mathrm{resb}(\mathcal{V}) \leq \aleph_0$. A *residual bound* for $\mathcal{V}$ is any cardinal $\kappa \geq \mathrm{resb}(\mathcal{V})$.

R. W. Quackenbush [42] proved that if a locally finite variety has an infinite subdirectly irreducible algebra, then it has unboundedly large finite subdirectly irreducible algebras. He posed the question, "Does every finitely generated residually finite variety have a finite residual bound?" While in general the answer to this question is no (R. McKenzie [35]), the answer is affirmative for finite algebras in congruence modular varieties (R. Freese and R. McKenzie [11]). Moreover, the class of finite algebras $\mathbf{A}$ in congruence modular varieties for which $\mathrm{HSP}(\mathbf{A})$ is residually finite is defined by a commutator equation. This equation, which takes the form $x \wedge [y,y] \leq [x,y]$, is called (C1). Notice that (C1) is equivalent to $x \wedge [y,y] = [x \wedge y, y]$, as can be proved by substituting $x \wedge y$ for x.

12.2 Theorem *Suppose that* $\mathbf{A}$ *is a finite algebra and that* $\mathrm{HSP}(\mathbf{A})$ *is congruence modular. The following are equivalent.*

(1) $\mathrm{resb}(\mathbf{A}) < \infty$.

(2) $\mathrm{resb}(\mathbf{A})$ *is a positive integer.*

(3) $\mathrm{HSP}(\mathbf{A}) \models_{\mathrm{Con}} (\alpha \wedge [\beta,\beta] \leq [\alpha,\beta])$.

(4) $\mathrm{S}(\mathbf{A}) \models_{\mathrm{Con}} (\alpha \wedge [\beta,\beta] \leq [\alpha,\beta])$.

Proof The implications (3) $\Rightarrow$ (4) and (2) $\Rightarrow$ (1) are trivial. We will prove that (4) implies the validity of (C1) in finite algebras of $\mathrm{HSP}(\mathbf{A})$, and that this in turn implies (2). Finally, we shall prove that (1) $\Rightarrow$ (3).

To begin, suppose that $\mathrm{S}(\mathbf{A}) \models$ (C1). Let $\mathbf{B}$ be any finite algebra in $\mathrm{HSP}(\mathbf{A})$. There is a positive integer n, a subalgebra $\mathbf{D}$ of $\mathbf{A}^n$, and a congruence θ on $\mathbf{D}$ such that $\mathbf{B}$ is isomorphic to $\mathbf{D}/\theta$. First, we show that $\mathbf{D} \models$ (C1); then using that, we show that $\mathbf{B} \models$ (C1). So let $\{\alpha, \beta\} \subseteq \mathrm{Con}\,\mathbf{D}$. We write η_i for the kernel of the ith projection homomorphism of $\mathbf{D}$ into $\mathbf{A}$, so that $\mathbf{D}/\eta_i \in \mathrm{S}(\mathbf{A})$.

To get a contradiction, we assume that $\alpha \wedge [\beta,\beta] \neq [\alpha \wedge \beta, \beta]$. This means that $\alpha \wedge [\beta,\beta] > [\alpha \wedge \beta, \beta]$. Since $\mathbf{D}/\eta_0 \models$ (C1), using Statement (6) of Theorem 8.3, we have that

$$[\alpha \wedge \beta, \beta] \vee \eta_0 = [(\alpha \wedge \beta) \vee \eta_0, \beta \vee \eta_0] \vee \eta_0 = ((\alpha \wedge \beta) \vee \eta_0) \wedge ([\beta \vee \eta_0, \beta \vee \eta_0] \vee \eta_0);$$

which gives that

$$[\alpha \wedge \beta, \beta] \vee \eta_0 \geq \alpha \wedge \beta \wedge [\beta,\beta] = \alpha \wedge [\beta,\beta]. \tag{12.1}$$

Modularity of $\mathrm{Con}\,\mathbf{D}$ thus implies that

$$\eta_0 \wedge \alpha \wedge [\beta,\beta] > \eta_0 \wedge [\alpha \wedge \beta, \beta] \geq [\eta_0 \wedge \alpha \wedge \beta, \beta]\,.$$

Replacing $i = 0$ by $i = 1$ and α by $\eta_0 \wedge \alpha$ in this argument, leads to

$$\eta_1 \wedge \eta_0 \wedge \alpha \wedge [\beta,\beta] > [\eta_1 \wedge \eta_0 \wedge \alpha \wedge \beta, \beta]. \tag{12.2}$$

Continuing in this fashion, we eventually reach the conclusion that

$$\bigwedge_{i<n} \eta_i \wedge \alpha \wedge [\beta,\beta] > \left[\bigwedge_{i<n} \eta_i \wedge \alpha \wedge \beta, \beta\right],$$

which is absurd since $\bigwedge_{i<n} \eta_i = 0_D$. This contradiction establishes that $\alpha \wedge [\beta,\beta] = [\alpha \wedge \beta, \beta]$. Thus $\mathbf{D} \models$ (C1).

Now to see that $\mathbf{B} \cong \mathbf{D}/\theta$ satisfies (C1), let $\{\alpha, \beta\} \subseteq \operatorname{Con}\mathbf{D}$ with $\alpha \wedge \beta \geq \theta$. By Theorem 8.3(6), what we need to show is that $\alpha \wedge ([\beta,\beta] \vee \theta) = [\alpha \wedge \beta, \beta] \vee \theta$. Since $\mathbf{D} \models$ (C1), and $\alpha \geq \theta$, we have that

$$\alpha \wedge ([\beta,\beta] \vee \theta) = (\alpha \wedge [\beta,\beta]) \vee \theta = [\alpha \wedge \beta, \beta] \vee \theta,$$

as required.

Next, suppose that all the finite algebras in HSP($\mathbf{A}$) satisfy (C1). If HSP($\mathbf{A}$) had an infinite subdirectly irreducible algebra, then the variety would contain arbitrarily large finite subdirectly irreducible algebras by [42]. Therefore, we need only find a finite bound on the size of the finite subdirectly irreducible algebras in the variety generated by $\mathbf{A}$. Let $\mathbf{B}$ be a finite subdirectly irreducible algebra in the variety generated by $\mathbf{A}$. Choose a positive integer n, a subalgebra $\mathbf{D}$ of $\mathbf{A}^n$, and a congruence θ on $\mathbf{D}$ with $\mathbf{B}$ isomorphic to $\mathbf{D}/\theta$. Let β be the monolith of $\mathbf{B}$.

By additivity of the commutator, $\mathbf{B}$ has a largest congruence ζ such that $[\zeta,\beta] = 0_B$. We will prove that $\mathbf{B}/\zeta \in \mathrm{HS}(\mathbf{A})$. Let $\alpha \in \operatorname{Con}\mathbf{D}$ be the congruence of $\mathbf{D}$ corresponding to ζ via the isomorphism of $\mathbf{B}$ with $\mathbf{D}/\theta$, and θ' be the congruence of $\mathbf{D}$ corresponding to β. (So that θ' is the unique cover of θ.) By our choice of ζ, α is the largest congruence in $\operatorname{Con}\mathbf{D}$ with $[\alpha,\theta'] \leq \theta$. For each $i = 0,\ldots,n-1$, let η_i be the kernel of the projection of $\mathbf{D}$ to the i^{th} coordinate. We will prove that there is some i so that $\eta_i \leq \alpha$. This will show that $\mathbf{B}/\zeta \cong \mathbf{D}/\alpha \in \mathrm{HS}(\mathbf{A})$. We do so by contradiction. Suppose that $\eta_i \not\leq \alpha$ for all i. By our choice of α, this means that $[\theta',\eta_i] \not\leq \theta$ for all i. It follows that for all i, $[\theta',\eta_i] \vee \theta$ is strictly larger than θ but contained in θ'. Hence, $[\theta',\eta_i] \vee \theta = \theta'$. If we substitute $[\theta',\eta_0] \vee \theta$ for θ' in $\theta' = [\theta',\eta_1] \vee \theta$, we get

$$\begin{aligned} \theta' &= [([\theta',\eta_0] \vee \theta),\eta_1] \vee \theta \\ &\leq [(\theta' \cap \eta_0) \vee \theta,\eta_1] \vee \theta \\ &= [(\theta' \cap \eta_0),\eta_1] \vee [\theta,\eta_1] \vee \theta \\ &\leq (\theta' \cap \eta_0 \cap \eta_1) \vee \theta. \end{aligned} \tag{12.3}$$

Since the reverse inclusion is also true, we actually have $\theta' = (\theta' \cap \eta_0 \cap \eta_1) \vee \theta$. By substituting this result into the equation $\theta' = [\theta',\eta_2] \vee \theta$, we similarly find that $\theta' = (\theta' \cap \eta_0 \cap \eta_1 \cap \eta_2) \vee \theta$. Proceeding inductively, repeatedly applying this argument, we eventually obtain that $\theta' = (\theta' \wedge \bigwedge_{0\leq i<n} \eta_i) \vee \theta$. This means that $\theta' = \theta$, which is the desired contradiction. The assumption that $\eta_i \not\leq \alpha$ for all i must be false, as we claimed.

Now $\mathbf{B}/\zeta \in \mathrm{HS}(\mathbf{A})$ implies $|\mathbf{B}/\zeta| \leq |\mathbf{A}|$. We shall conclude this proof that (4) $\Rightarrow$ (2) by demonstrating that each ζ-class is no larger than 2^M where M is the cardinality of the free algebra in HSP($\mathbf{A}$) on $|A|+2$ generators. This will prove that $|\mathbf{B}| \leq |\mathbf{A}| \cdot 2^M$.

Next, we observe that $[\zeta,\zeta] = 0_B$, equivalently, $[\alpha,\alpha] \leq \theta$. This is true because $\mathbf{D} \models$ (C1), so that $\theta' \wedge [\alpha,\alpha] \leq [\theta',\alpha] \leq \theta$. But if $[\alpha,\alpha] \not\leq \theta$, then $\theta' \leq [\alpha,\alpha] \vee \theta$, giving that $\theta' = (\theta' \wedge [\alpha,\alpha]) \vee \theta = \theta$ by modularity, a contradiction. Thus $[\zeta,\zeta] = 0_B$.

Denote the Gumm difference term of HSP($\mathbf{A}$) by q. Since $[\zeta,\zeta] = 0$, we know that the restriction of q to any ζ-class is a Maltsev operation, and gives that set the structure of a ternary Abelian group. Select $\langle 0, b\rangle \in \beta - 0_B$, and let $+$ and $-$ denote the Abelian group

operations on $0/\zeta$ with neutral element 0 induced by q. Letting u be any element of $\mathbf{B}$, we now proceed to prove that $|u/\zeta| \leq 2^M$. Suppose that $x, y \in u/\zeta$ and $x \neq y$. Since β is the monolith of $\mathbf{B}$, $\langle 0, b\rangle \in \mathrm{Cg}_{\mathbf{B}}(x, y)$. This means that there are elements $v_0 = 0, v_2, \ldots, v_k = b$ and unary polynomials $p_0, \ldots, p_{k-1}$ so that $\{p_i(x), p_i(y)\} = \{v_i, v_{i+1}\}$ for all i. We can apply the difference term q and manipulate its local Maltsev characteristics to shorten the chain of v_i's until $k = 2$. This means that there is a unary polynomial f so that $\{f(x), f(y)\} = \{0, b\}$. Moreover, if $f(x) = b$ and $f(y) = 0$, then we can replace f by the polynomial $b - f(z)$ so that we can assume $f(x) = 0$ and $f(y) = b$. We will bound the number of constants necessary to construct f. Let $c_0, \ldots, c_{l-1}$ be representatives of the ζ-classes with $c_0 = 0$. Note that $l \leq |\mathbf{A}|$. There are constants $r_0, \ldots, r_{m-1}$ and an $(m+1)$-ary term t so that $f(z) = t(z, \mathbf{r})$ for all $z \in \mathbf{B}$. For each $j = 1, \ldots, m-1$, select i_j so that $c_{i_j} \zeta r_j$. For notational convenience, let $s_j = c_{i_j}$. If $z \in x/\zeta$, then the matrix

$$\begin{pmatrix} t(x, \mathbf{r}) - t(x, \mathbf{r}) + 0 & t(x, \mathbf{s}) - t(x, \mathbf{s}) + 0 \\ t(z, \mathbf{r}) - t(x, \mathbf{r}) + 0 & t(z, \mathbf{s}) - t(x, \mathbf{s}) + 0 \end{pmatrix}$$

is in $M(\zeta, \zeta)$. Since the top row of the matrix is an equality, so is the bottom since $[\zeta, \zeta] = 0_B$. It follows then that $f(z) = t(z, \mathbf{r}) = t(z, \mathbf{s}) - t(x, \mathbf{s}) + 0$. Let $c_l = t(x, \mathbf{s})$. Then for all $z \in x/\zeta$ we have $f(z) = q(t(z, \mathbf{s}), c_l, c_0)$ (recall that $0 = c_0$). Since each $s_j = c_{i_j}$, we have an $(l+2)$-ary term t' so that $f(z) = t'(z, c_0, \ldots, c_l)$ for all $z \in u/\zeta$. We have proven that for all $x \neq y \in u/\zeta$ there exists an $(l+2)$-ary term t' so that $t'(x, c_0, \ldots, c_l) = 0$ if and only if $t'(y, c_0, \ldots, c_l) \neq 0$. Define Σ to be the set of all functions $t(-, c_0, \ldots, c_l)$ with t an $(l+2)$-ary term of $\mathbf{B}$. Notice that $|\Sigma| \leq |\mathbf{F}_{\mathrm{HSP}(\mathbf{A})}(l+2)| = M$ where $\mathbf{F}_{\mathrm{HSP}(\mathbf{A})}(l+2)$ is the free algebra in $\mathrm{HSP}(\mathbf{A})$ on $l+2$ generators. Define an equivalence relation $\sim$ on u/ζ by $x \sim y$ if for all $f \in \Sigma$, $f(x) = 0$ if and only if $f(y) = 0$. Then $|(u/\zeta)/\sim| \leq 2^{|\Sigma|} \leq 2^M$. However, since we can separate points of u/ζ with functions in Σ, it follows that $\sim$ is the identity relation on u/ζ, so we have that $|u/\zeta| \leq 2^M$ as desired. As stated above, it follows that

$$|B| \leq |A| \cdot 2^M \leq |A| \cdot 2^{|A|^{|A|^{|A|+2}}},$$

and this gives a finite upper bound to $\mathrm{resb}(\mathbf{A})$. We have proven that (4) $\Rightarrow$ (2).

To complete the proof of this theorem, it only remains to establish that (1) $\Rightarrow$ (3). Assume that (3) does not hold. We will prove that $\mathrm{HSP}(\mathbf{A})$ has arbitrarily large subdirectly irreducibles. There is an algebra $\mathbf{E}$ in $\mathrm{HSP}(\mathbf{A})$ with congruences β' and γ so that $\beta' \cap [\gamma, \gamma] \not\leq [\beta', \gamma]$. Let $\beta = \beta' \cap [\gamma, \gamma]$. Then $\beta \leq [\gamma, \gamma]$ and $[\beta, \gamma] < \beta$ (because otherwise, $\beta' \cap [\gamma, \gamma] = \beta = [\beta, \gamma] \leq [\beta', \gamma]$). Choose a strictly meet irreducible congruence θ which exceeds $[\beta, \gamma]$ but not β. Let θ' be the unique cover of θ. It follows that $\theta' \leq [\gamma \vee \theta, \gamma \vee \theta] \vee \theta$ and

$$[\theta', \gamma \vee \theta] \leq [\beta \vee \theta, \gamma \vee \theta] \leq [\beta, \gamma] \vee \theta = \theta,$$

because $\theta' \leq \beta \vee \theta \leq [\gamma, \gamma] \vee \theta$. Therefore, we can change notation (replacing $\mathbf{E}$ by $\mathbf{E}/\theta$) and assume that $\mathbf{E}$ is subdirectly irreducible with monolith β. We have that $\beta \leq [\gamma, \gamma]$ and $[\beta, \gamma] = 0_E$.

Let $\Delta = \Delta_{\gamma,\beta}$ be the congruence on $\mathbf{E}(\gamma)$ as in Lemma 8.6. Let $\pi_i : \mathbf{E}(\gamma) \to \mathbf{E}$ for $i = 0, 1$ be the canonical projections with $\eta_i = \ker \pi_i$. For $i = 0, 1$, let $\beta_i = \pi^{-1}(\beta)$. From Lemma 8.6 we have that $\Delta \cap \eta_i = 0_{\mathbf{E}(\gamma)}$ and $\Delta \vee \eta_i = \beta_i$.

Let $\aleph$ be an arbitrary cardinal. We will follow the tradition that $\aleph = \{\sigma : \sigma < \aleph\}$ and extend the notation for elements of $\mathbf{E}^{\aleph}$ which we have been using for finite direct powers up

to now. That is, we will represent elements of $\mathbf{E}^{\aleph}$ as bold faced vectors $\mathbf{a}$, and for $\delta \in \aleph$, we will denote the δ coordinate of $\mathbf{a}$ as a_δ. Let $\mathbf{B} = \{\mathbf{a} \in \mathbf{E}^{\aleph} : a_\delta \gamma a_\epsilon \text{ for all } \delta, \epsilon \in \aleph\}$. For any $\epsilon \in \aleph$, let $\gamma_\epsilon \in \operatorname{Con}\mathbf{B}$ be defined by $\mathbf{a}\gamma_\epsilon\mathbf{b}$ if and only if $a_\epsilon \gamma b_\epsilon$, and define β_ϵ analogously. From our definition of $\mathbf{B}$, it follows that $\gamma_\delta = \gamma_\epsilon$ for all $\epsilon, \delta \in \aleph$. We will denote this congruence as γ. For each $\epsilon \in \aleph$, let η_ϵ be the kernel of the projection of $\mathbf{B}$ to the ϵ coordinate and let $\eta'_\epsilon = \bigcap_{\delta \neq \epsilon} \eta_\delta$. Then $\eta_\epsilon \vee \eta_\delta = \gamma$ for all $\{\delta, \epsilon\} \subseteq \aleph$, $\delta \neq \epsilon$. For each $\sigma \in \aleph$, let θ_σ be defined as

$$\theta_\sigma = \{\langle \mathbf{a}, \mathbf{b}\rangle : a_\sigma \beta b_\sigma \text{ and for all } \epsilon \neq \sigma,\ a_\epsilon = b_\epsilon\}.$$

For each $\sigma \in \aleph \setminus \{0\}$, let Δ_σ be defined as

$$\Delta_\sigma = \{\langle \mathbf{a}, \mathbf{b}\rangle : \langle a_0, a_\sigma\rangle \Delta \langle b_0, b_\sigma\rangle \text{ and for all } \epsilon \notin \{0, \sigma\},\ a_\epsilon = b_\epsilon\}.$$

Finally, let $\theta = \bigvee_\sigma \theta_\sigma$ and $\kappa = \bigvee_{\sigma>0} \Delta_\sigma$.

We claim that $\theta_0 \leq \theta_\delta \vee \Delta_\delta$ and that $\theta_\delta \leq \theta_0 \vee \Delta_\delta$ for $\delta \neq 0$. Suppose that $\mathbf{a}\theta_0\mathbf{b}$. This means that $a_0 \beta b_0$ and otherwise $\mathbf{a}$ and $\mathbf{b}$ are equal. Now, since $\langle a_0, a_0\rangle \Delta \langle b_0, b_0\rangle$ we have

$$\begin{aligned} \mathbf{a} &= \langle a_0, \ldots, a_\delta, \ldots\rangle \\ &\ \eta'_\delta \langle a_0, \ldots, a_0, \ldots\rangle \\ &\ \Delta_\delta \langle b_0, \ldots, b_0, \ldots\rangle \\ &\ \eta'_\delta \langle b_0, \ldots, b_\delta, \ldots\rangle \\ &= \mathbf{b} \end{aligned} \tag{12.4}$$

so $\theta_0 \leq \eta'_\delta \vee \Delta_\delta$. If we meet with $\theta_0 \vee \theta_\delta$, then using modularity several times, observing that $\Delta_\delta \leq \theta_0 \vee \theta_\delta$ and $\eta'_\delta \wedge (\theta_0 \vee \theta_\delta) = (\eta'_\delta \wedge \theta_0) \vee \theta_\delta = \theta_\delta$ (since $\eta'_\delta \geq \theta_\delta$), we obtain that

$$\theta_0 \leq (\theta_0 \vee \theta_\delta) \wedge (\eta'_\delta \vee \Delta_\delta) = \{(\theta_0 \vee \theta_\delta) \wedge \eta'_\delta\} \vee \Delta_\delta = \theta_\delta \vee \Delta_\delta.$$

That $\theta_\delta \leq \theta_0 \vee \Delta_\delta$ can be proven similarly.

For any $\delta \neq 0 \neq \epsilon$ this gives

$$\begin{aligned} \kappa \vee \theta_\delta &= \kappa \vee \Delta_\delta \vee \Delta_\epsilon \vee \theta_\delta \\ &= \kappa \vee \Delta_\epsilon \vee \theta_0 \vee \theta_\delta \\ &= \kappa \vee \theta_\epsilon \vee \theta_0 \vee \theta_\delta \\ &\geq \theta_\epsilon \vee \theta_0. \end{aligned} \tag{12.5}$$

It follows from our definitions that $\kappa \vee \theta_\delta = \theta$ for all δ including $\delta = 0$.

We also claim that $\theta_\delta \not\leq \kappa$. We first show that $\theta_0 \not\leq \kappa$. The case for $\delta \neq 0$ then follows since $\kappa \vee \theta_0 = \kappa \vee \theta_\delta$. Since $0_E \prec \beta$ in $\operatorname{Con}\mathbf{E}$, we know that $0_B \prec \theta_\delta$ in $\operatorname{Con}\mathbf{B}$. Hence, θ_δ is compact. If $\theta_0 \leq \kappa$, then by compactness, θ_0 would be exceeded by a join of finitely many of the Δ_ϵ. We will prove by induction on $n \geq 1$ that θ_0 is not exceeded by a join of n of the Δ_ϵ. It must be that $\theta_0 \cap \Delta_\epsilon = 0_B$ for all ϵ. To see this, suppose that $\langle \mathbf{a}, \mathbf{b}\rangle \in \theta_0 \cap \Delta_\epsilon$. This means that $a_\delta = b_\delta$ for all $\delta \neq 0$ and that $\langle a_0, a_\epsilon\rangle \Delta \langle b_0, b_\epsilon\rangle = \langle b_0, a_\epsilon\rangle$. By Lemma 8.6 this means that $\langle a_0, b_0\rangle \in [\gamma, \beta] = 0_E$. Hence, we also have that $a_0 = b_0$ so $\mathbf{a} = \mathbf{b}$. Since $\theta_0 \cap \Delta_\epsilon = 0_B$ for all ϵ, it cannot be that $\theta_0 \leq \Delta_\epsilon$ for any ϵ. Now suppose that $n > 1$ and that θ_0 is not exceeded by any join of fewer than n of the Δ_ϵ. Suppose that $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are n distinct members of $\aleph$. Then we know that $\Delta_{\epsilon_1} \leq \theta_0 \vee \theta_{\epsilon_1}$, that $\bigvee_{i\geq 2} \Delta_{\epsilon_i} \leq \theta_0 \vee (\bigvee_{i\geq 2} \theta_{\epsilon_i})$,

and that $\theta_0 \cap (\bigvee_{i\geq 2} \Delta_{\epsilon_i}) = 0_B$ (since $0_B \prec \theta_0$). Also, it is not difficult to prove that the θ_δ's are independent. Putting all of this together gives

$$\begin{aligned}
\theta_0 \cap \left(\bigvee_{1\leq i\leq n} \Delta_{\epsilon_i} \right) &= \theta_0 \cap (\theta_0 \vee \theta_{\epsilon_1}) \cap \left(\bigvee_{1\leq i\leq n} \Delta_{\epsilon_i} \right) \\
&= \theta_0 \cap \left(\Delta_{\epsilon_1} \vee \left[(\theta_0 \vee \theta_{\epsilon_1}) \cap \left(\theta_0 \vee \left(\bigvee_{i\geq 2} \theta_{\epsilon_i} \right) \right) \cap \left(\bigvee_{i\geq 2} \Delta_{\epsilon_i} \right) \right] \right) \\
&= \theta_0 \cap \left(\Delta_{\epsilon_1} \vee \left[\theta_0 \cap \left(\bigvee_{i\geq 2} \Delta_{\epsilon_2} \right) \right] \right) \\
&= \theta_0 \cap \Delta_{\epsilon_1} \\
&= 0_B.
\end{aligned} \tag{12.6}$$

This means, of course, that $\theta_0 \not\leq \bigvee_{1\leq i\leq n} \Delta_{\epsilon_i}$. By induction, it cannot be that θ_0 is less than any finite join of Δ_ϵ's. It follows then that $\theta_0 \not\leq \kappa$ and that $\theta_\delta \not\leq \kappa$ for any δ.

Since $\kappa \vee \theta_\delta = \theta$ but $\theta_\delta \not\leq \kappa$ for all δ, it follows that $\kappa < \theta$. Therefore, there is a completely meet irreducible $\lambda \in \operatorname{Con}\mathbf{B}$ which contains κ and not θ. In $\operatorname{Con}\mathbf{B}$, $\eta_\delta \cap \eta'_\delta = 0_B$ and $\eta_\delta \vee \eta'_\delta = \gamma$. So the interval from η_δ to γ is isomorphic to the interval from 0 to η'_δ. Since $\mathbf{E}$ is subdirectly irreducible, the interval from η_δ to γ has a unique atom. Hence, the interval from 0 to η'_δ has a unique atom—which is θ_δ in our notation. This implies that $\lambda \cap \eta'_\delta = 0_B$. If this were not the case, then $\lambda \geq \theta_\delta$ and so $\lambda \geq \theta_\delta \vee \kappa = \theta$ which is a contradiction. We have then that $\eta_\delta \vee \eta'_\delta = \gamma$ and that $\lambda \cap \eta'_\delta = 0_B$. We claim that for all δ, $\lambda \vee \eta_\delta \not\geq \gamma$. To see this, suppose that $\lambda \vee \eta_\delta \geq \gamma$. Then $[\gamma,\gamma] \leq [\eta_\delta \vee \eta'_\delta, \eta_\delta \vee \lambda] \leq \eta_\delta \vee (\lambda \cap \eta'_\delta) = \eta_\delta$. Since $\mathbf{E} \cong \mathbf{B}/\eta_\delta$, this would imply that in $\operatorname{Con}\mathbf{E}$, $[\gamma,\gamma] = 0_E$—which is not true. Now, since $\eta_\delta \vee \eta_\epsilon = \gamma$ for all $\delta \neq \epsilon \in \aleph$, the congruences $\lambda \vee \eta_\delta$, $\delta \in \aleph$, are pairwise distinct. Therefore $\mathbf{B}/\lambda$—which is subdirectly irreducible—has at least $\aleph$ congruences. Since $\aleph$ is an arbitrary infinite cardinal, it follows that $\operatorname{HSP}(\mathbf{A})$ has no residual bound. We have proven the contrapositive of (1) $\Rightarrow$ (3). □

13 Directly representable varieties

13.1 Definition A variety $\mathcal{V}$ is *directly representable* if and only if there is a finite set $\mathcal{D}$ of finite algebras such that $\mathcal{V} = \operatorname{HSP}(\mathcal{D})$ and every finite algebra in $\mathcal{V}$ belongs to $\operatorname{IP}(\mathcal{D})$, equivalently, $\mathcal{V}$ is locally finite and has, up to isomorphism, only a finite set of finite directly indecomposable algebras. The *finite spectrum* of a class $\mathcal{K}$ of algebras is the set of positive integers n such that $\mathcal{K}$ has an n-element algebra. A class $\mathcal{K}$ of algebras is said to be *narrow* if and only if there is a finite set $\{p_0, \ldots, p_{k-1}\}$ of prime integers such that every member of the finite spectrum of $\mathcal{K}$ takes the form $\prod_{i<k} p_i^{a_i}$ for some integers a_i.

In this section, we prove that every narrow locally finite variety has permuting congruences, characterize finite algebras that generate narrow varieties, and using the commutator, characterize finite algebras that generate directly representable varieties. The results proved here are drawn from R. McKenzie [32].

13.2 Theorem *Let $\mathcal{V}$ be any locally finite variety and consider these possible properties of $\mathcal{V}$.*

(1) *$\mathcal{V}$ is directly representable.*

(2) *$\mathcal{V}$ is narrow.*

(3) *All congruences on finite algebras in $\mathcal{V}$ are uniform.*

(4) *$\mathcal{V}$ has permuting congruences.*

We have (1) $\Rightarrow$ (2) $\Rightarrow$ (3) $\Rightarrow$ (4).

Proof Clearly (1) implies (2).

To prove that (2) implies (3), suppose that $\mathcal{V}$ is a narrow variety, that $\mathbf{A}$ is a finite algebra in $\mathcal{V}$, and that $\theta \in \operatorname{Con}\mathbf{A}$. For n a positive integer, define $\mathbf{A}_n(\theta)$ to be the algebra (subalgebra of $\mathbf{A}^n$) consisting of all sequences $\mathbf{u} \in A^n$ such that $u_i \theta u_j$ for all $\{i,j\} \subseteq \{0,\dots,n-1\}$. We assume that $|\mathbf{A}/\theta| = k$ and that the k distinct θ-equivalence classes have cardinalities $a_0,\dots,a_{k-1}$. We are going to show that $a_0 = a_1 = \cdots = a_{k-1}$. Observe that

$$|\mathbf{A}_n(\theta)| = a_0^n + \cdots + a_{k-1}^n = s_n(\mathbf{a}) \tag{13.1}$$

where $\mathbf{a} = \langle a_0,\dots,a_{k-1}\rangle$. Let $\{p_0,\dots,p_{\ell-1}\}$ be a finite set of prime integers that include all the prime divisors of the members of the finite spectrum of $\mathcal{V}$, and hence all the prime divisors of the integers $s_n(\mathbf{a})$, $n \geq 1$. Let $d = \gcd\{a_i : i < k\}$ and let $\mathbf{b} = \langle b_0,\dots,b_{k-1}\rangle$ where $db_i = a_i$. Choose M to be any positive integer such that $2^M \geq k$. For $j < \ell$, put $q_j = (p_j - 1)p_j^M$. An easy calculation, based on the Euler-Fermat theorem, shows that

$$b_i^{q_j} \equiv 0, 1 \pmod{p_j^{M+1}} \quad \text{for } i < k,\ j < \ell. \tag{13.2}$$

Thus for any positive integer N, $b_i^{q_j N} \equiv 0, 1 \pmod{p_j^{M+1}}$, and we have

$$s_{q_j N}(\mathbf{b}) \equiv u_j \pmod{p_j^{M+1}} \tag{13.3}$$

where u_j is the number of $i < k$ such that b_i is prime to p_j. Note that $1 \leq u_j \leq k$, since the b_i have no positive common divisors other than 1, and $p_j^{M+1} > k$. Thus p_j^{M+1} cannot divide $s_{q_j N}(\mathbf{b})$.

Now taking $q = \prod_{j<\ell} q_j$ and N any positive integer, it follows from the above analysis that $s_{qN}(\mathbf{b})$ is not divisible by p_j^{M+1} for any $j < \ell$. Since $s_{qN}(\mathbf{b})$ is a product of powers of the p_j, it follows that

$$s_{qN}(\mathbf{b}) \leq P = \prod_{j<\ell} p_j^M. \tag{13.4}$$

If we had $b_i > 1$ for some i, then the sequence $\langle s_{qN}(\mathbf{b}) : N \geq 1\rangle$ would be a strictly increasing sequence of integers. Because the sequence is bounded, we must have $b_0 = b_1 = \cdots = b_{k-1} = 1$. Thus we conclude our proof that $\mathbf{a}$ is a constant sequence, or in other words, θ is a uniform congruence.

To prove that (3) implies (4), suppose that $\mathcal{V}$ is locally finite and its finite algebras have uniform congruences. We first observe that as an easy corollary of the proof of Theorem 6.1,

a variety has permuting congruences if and only if its free algebra on three generators has permuting congruences. Since $\mathcal{V}$ is locally finite, it therefore suffices to show that the finite algebras in $\mathcal{V}$ have permuting congruences. Thus suppose that $\mathbf{A} \in \mathcal{V}$ is finite and $\{\alpha, \beta\} \subseteq \operatorname{Con}\mathbf{A}$. Let $\mathbf{B}$ be the algebra $\mathbf{A}_2(\alpha) \leq \mathbf{A} \times \mathbf{A}$ consisting of all pairs $\langle x, y\rangle$ with $x \alpha y$. Let θ (on $\mathbf{B}$) be the congruence $\beta \times \beta|_B$, so that $\langle x, y\rangle \theta \langle u, v\rangle$ if and only if $x \beta u$ and $y \beta v$. By assumption, the congruences $\beta, \alpha \cap \beta, \theta$ are uniform. Let b, c, e be the respective block-sizes for these congruences. Choose any $a \in A$. Then $\langle a, a\rangle/\theta$ consists of all $\langle x, y\rangle \in A \times A$ such that $x \in a/\beta$ and $y \in x/(\alpha \cap \beta)$; thus $e = bc$.

Now, to conclude this proof, we assume that there are elements u, v, w in A with $u \alpha v \beta w$ and $\langle u, w\rangle \notin \beta \circ \alpha$, and we derive a contradiction. This assumption implies that there is no $x \in A$ with $\langle x, w\rangle \in \langle u, v\rangle/\theta$. Note that for any $z \in A$, if $\langle x, z\rangle \in \langle u, v\rangle/\theta$ for some $x = x_0$ then $\langle x, z\rangle \in \langle u, v\rangle/\theta$ precisely for the elements $x \in x_0/(\alpha \cap \beta)$. Thus $|\langle u, v\rangle/\theta| = b'c$ where b' is the number of $z \in v/\beta$ such that such $\langle u, z\rangle \in \beta \circ \alpha$. We have that $b' < b$ since $w \in v/\beta$ and $\langle u, w\rangle \notin \beta \circ \alpha$. Thus $|\langle u, v\rangle/\theta| = b'c < bc = |\langle u, v\rangle/\theta|$, which is the promised contradiction. □

13.3 Lemma *In a directly representable variety, every finite subdirectly irreducible algebra is Abelian or simple.*

Proof Assume that $\mathcal{V}$ is a directly representable variety. By Theorem 13.2, $\mathcal{V}$ is congruence-modular; in fact, it is a Maltsev variety. Since subdirectly irreducible algebras are directly indecomposable, $\mathcal{V}$ has a finite bound on the size of its finite subdirectly irreducible algebras. As we noted in our proof of Theorem 12.2, this implies that the locally finite variety $\mathcal{V}$ has $\operatorname{resb}(\mathcal{V}) < \omega$, and $\mathcal{V} \models$ (C1). Now let $\mathbf{A}$ be a finite subdirectly irreducible algebra in $\mathcal{V}$. To get a contradiction, we suppose that $\mathbf{A}$ is neither Abelian nor simple. Where β is the monolith of $\mathbf{A}$, this supposition means that $0_A \prec \beta < 1_A$ and $\beta \leq [1_A, 1_A]$. Applying (C1), we find that $\beta = [\beta, 1_A]$.

Choose any positive integer n. Once again, we consider the algebra $\mathbf{A}_n(\beta)$ of β-constant n-tuples. Our goal this time is to prove that $\mathbf{A}_n(\beta)$ is directly indecomposable. Since $|\mathbf{A}_n(\beta)| > 2^n$, this will contradict the ground assumption that $\mathcal{V}$ is directly representable. (This is the fourth time that we have used this fruitful construction after Section 11.)

Suppose, for sake of contradiction, that $\mathbf{A}_n(\beta)$ is not directly indecomposable. Then it possesses a pair of congruences $\langle \delta_0, \delta_1\rangle$ such that $0 < \delta_\varepsilon < 1$, $\delta_0 \vee \delta_1 = 1$, $\delta_0 \wedge \delta_1 = 0$. We write η_i for the kernel of the projection homomorphism of $\mathbf{A}_n(\beta)$ to $\mathbf{A}$ at the i coordinate (as usual) and put $\eta_i' = \bigwedge_{j \neq i} \eta_j$. We write β_i for the kernel of the homomorphism of $\mathbf{A}_n(\beta)$ to $\mathbf{A}/\beta$ through the i coordinate, so that $\beta_i = \beta_j$ for alll $\{i, j\} \subseteq \{0, \ldots, n-1\}$, and we write simply β for this congruence. The fact that $\mathbf{A} \models \beta = [\beta, 1]$ gives $\mathbf{A}_n(\beta) \models \beta = \eta_i \vee [\beta, 1]$ for each $i < n$. (Here we have used Theorem 8.3(6), again.) For $i < n$ we have $\eta_i \prec \beta$, $\beta = \eta_i \vee \eta_i'$, and $\eta_i \wedge \eta_i' = 0$. Then by modularity, $0 \prec \eta_i'$.

We can now show that for each $i < n$ and $\varepsilon \in \{0, 1\}$, if $\delta_\varepsilon \not\leq \eta_i$ then $\eta_i' \leq \delta_\varepsilon$. Indeed, if $\delta_\varepsilon \not\leq \eta_i$, then $\beta \leq \eta_i \vee \delta_\varepsilon$, giving $\beta = \eta_i \vee (\beta \wedge \delta_\varepsilon)$. Then

$$[\eta_i', \delta_{1-\varepsilon}] \leq [\beta, \delta_{1-\varepsilon}] = [\eta_i \vee (\beta \wedge \delta_\varepsilon), \delta_{1-\varepsilon}] \leq \eta_i \vee [\delta_\varepsilon, \delta_{1-\varepsilon}] = \eta_i. \tag{13.5}$$

Thus $[\eta_i', \delta_{1-\varepsilon}] \leq \eta_i' \wedge \eta_i = 0$. Then since $[\beta, 1] = [\eta_i \vee \eta_i', 1] \not\leq \eta_i$ we have

$$0 \neq [\eta_i', 1] = [\eta_i', \delta_0 \vee \delta_1] = [\eta_i', \delta_0] \vee [\eta_i', \delta_1] = [\eta_i', \delta_\varepsilon]. \tag{13.6}$$

Since η_i' is an atom, then

$$\eta_i' = [\eta_i', \delta_\varepsilon] \leq \delta_\varepsilon. \tag{13.7}$$

So indeed, $\delta_\varepsilon \not\leq \eta_i$ implies $\eta_i' \leq \delta_\varepsilon$.

Since $\bigvee_i \eta_i = \beta < 1 = \delta_0 \vee \delta_1$, then there is $\varepsilon \in \{0,1\}$ such that $\delta_\varepsilon \leq \eta_i$ holds for no i. Then $\delta_\varepsilon \geq \bigvee_i \eta_i' = \beta$. This implies that $\delta_{1-\varepsilon} \geq \eta_i'$ holds for no i since $\delta_0 \wedge \delta_1 = 0$; consequently, $\delta_{1-\varepsilon} \leq \eta_i$ for all $i < n$. But that forces $\delta_{1-\varepsilon} = 0$. This contradiction proves the lemma. □

For the final theorem of this section, we will need this lemma.

13.4 Lemma (I. Fleischer [10]) *Let $\mathbf{A} \leq \mathbf{A}_0 \times \mathbf{A}_1$ be a subdirect product, where $\mathbf{A}$ has permuting congruences. Then $\mathbf{A}$ is the equalizer of a pair of surjective homomorphisms $\pi_i : \mathbf{A}_i \to \mathbf{K}$ for some algebra $\mathbf{K}$. Consequently, if $\mathbf{A}$ has permuting congruences and $\mathbf{A}$ is a subdirect product of a finite system of simple algebras, $\mathbf{A} \leq \prod_{t\in T} \mathbf{S}_t$, then for some $W \subseteq T$, the projection of $\mathbf{A}$ into $\prod_{t\in W} \mathbf{S}_t$ is an isomorphism $\pi_W : \mathbf{A} \cong \prod_{t\in W} \mathbf{S}_t$.*

Proof This is left as an exercise for the reader. □

13.5 Theorem

(1) *In a directly representable variety, every finite algebra is isomorphic, for some $m \geq 0$, with a direct product $\mathbf{B}_0 \times \mathbf{B}_1 \times \cdots \times \mathbf{B}_m$ where $\mathbf{B}_0$ is Abelian and $\mathbf{B}_i$ is a simple non-Abelian algebra for $i \geq 1$.*

(2) *Let $\mathbf{A}$ be a finite algebra.* HSP($\mathbf{A}$) *is directly representable if and only if $\mathbf{A}$ has a Maltsev term, every subalgebra of $\mathbf{A}$ is isomorphic with a direct product of an Abelian algebra and a product of simple non-Abelian algebras, and the variety $\mathcal{A}$ generated by the collection of all Abelian direct factors of subalgebras of $\mathbf{A}$ is directly representable.*

(3) *Let $\mathbf{A}$ be a finite algebra.* HSP($\mathbf{A}$) *is narrow if and only if $\mathbf{A}$ has a Maltsev term and every subalgebra of $\mathbf{A}$ has uniform congruences.*

Proof To prove (1), suppose that $\mathcal{V}$ is directly representable and $\mathbf{A}$ is a finite algebra in $\mathcal{V}$.

By Lemma 13.3, and G. Birkhoff's subdirect representation theorem, $\mathbf{A}$ is isomorphic, for some integer $m \geq 0$, to an algebra $\mathbf{A}' \leq \prod_{i\leq m} \mathbf{B}_i$ where $\mathbf{B}_0$ is Abelian and $\mathbf{B}_i$ ($i \geq 1$) is non-Abelian and simple. (Here we have used that any subdirect product of Abelian algebras is Abelian.) We assume that m is as small as it can be.

Let η_i denote the kernel of the i coordinate projection of $\mathbf{A}'$ onto $\mathbf{B}_i$. Put $\delta_0 = \eta_0$ and $\delta_1 = \bigwedge_{1\leq i\leq m} \eta_i$. Since $\mathcal{V}$ has permuting congruences, the minimality of m and Lemma 13.4 tells us that the projection of $\mathbf{A}'$ into $\mathbf{B}_1 \times \cdots \times \mathbf{B}_m$ (an algebra isomorphic to $\mathbf{A}'/\delta_1$) is the full direct product of $\mathbf{B}_1, \ldots, \mathbf{B}_m$.

Next we claim that $\mathbf{A}'/\delta_1 \models [1,1] = 1$. This actually follows from the fact that in a modular variety, the class of algebras satisfying $[x,y] = x \wedge y$ for congruences—called "neutral algebras"—is closed under binary subdirect products. (Simple, non-Abelian algebras are neutral.) To see this, suppose that $\mathbf{E} \leq \mathbf{E}_0 \times \mathbf{E}_1$ is a subdirect product and $\mathbf{E}_i \models_{\mathrm{CON}} [x,y] = x \wedge y$. Let ρ_0, ρ_1 denote the two projection congruences on $\mathbf{E}$, and let $\{\alpha, \beta\} \subseteq \operatorname{Con}\mathbf{E}$. Then $\mathbf{E}/\rho_0$ neutral implies that

$$\alpha \wedge \beta \leq (\alpha \vee \rho_0) \wedge (\beta \vee \rho_0) = [\alpha, \beta] \vee \rho_0\,. \tag{13.8}$$

Since $[\alpha,\beta] \leq \alpha\wedge\beta$, by modularity it follows that if $[\alpha,\beta] < \alpha\wedge\beta$ then $[\alpha,\beta]\wedge\rho_0 < \alpha\wedge\beta\wedge\rho_0$. Suppose that this strict inclusion holds. Now $(\alpha\wedge\beta\wedge\rho_0)\wedge\rho_1 = 0$, hence by modularity we must have $([\alpha,\beta]\wedge\rho_0)\vee\rho_1 < (\alpha\wedge\beta\wedge\rho_0)\vee\rho_1$. But since $\mathbf{E}/\rho_1$ is neutral, we can calculate that

$$\begin{aligned}([\alpha,\beta]\wedge\rho_0)\vee\rho_1 &\geq [[\alpha,\beta],\rho_0]\vee\rho_1\\ &= [[\alpha\vee\rho_1,\beta\vee\rho_1],\rho_0\vee\rho_1]\vee\rho_1\\ &= (\alpha\vee\rho_1)\wedge(\beta\vee\rho_1)\wedge(\rho_0\vee\rho_1)\\ &\geq (\alpha\wedge\beta\wedge\rho_0)\vee\rho_1.\end{aligned} \tag{13.9}$$

This final contradiction shows that $[\alpha,\beta] = \alpha\wedge\beta$, as claimed.

Thus we have proved that $\mathbf{A}'/\delta_1 \models [1,1] = 1$. This means that $\mathbf{A}' \models 1 = \delta_1\vee[1,1]$. Since $\mathbf{B}_0$ is Abelian, then $[1,1] \leq \delta_0$. Thus $\delta_1\vee\delta_0 = 1$. Since $\delta_0\wedge\delta_1 = 0$ by definition, and since δ_0 and δ_1 must commute, then it follows that $\mathbf{A}'$ is the direct product of $\mathbf{B}_0$ and the projection of $\mathbf{A}'$ into $\prod_{i\geq 1}\mathbf{B}_i$—i.e., $\mathbf{A}' = \prod_{i\leq m}\mathbf{B}_i$.

Now let $\mathbf{A}$ be any finite algebra. To prove (2), observe that we have already proved that HSP($\mathbf{A}$) directly representable implies the truth of the other three conditions. Conversely, suppose that these conditions are valid. Let $\mathcal{A}$ be the variety generated by the Abelian direct factors of subalgebras of $\mathbf{A}$. Every finite algebra $\mathbf{B}$ in SP($\mathbf{A}$) is isomorphic to a subdirect product of subalgebras of $\mathbf{A}$, and thus is isomorphic to a subdirect product $\mathbf{B}' \leq \prod_{i\leq m}\mathbf{B}_i$ with $\mathbf{B}_0 \in \mathcal{A}$ and $\mathbf{B}_i$ simple and non-Abelian for $1 \leq i \leq m$. Choosing m minimal for $\mathbf{B}$, the above argument yields that $\mathbf{B}' = \prod_{i\leq m}\mathbf{B}_i$. Now let θ be any congruence of $\mathbf{B}'$. Write, as before, δ_0 and δ_1 for the kernels of the projections of $\mathbf{B}'$ onto $\mathbf{B}_0$ and $\mathbf{B}_1\times\cdots\times\mathbf{B}_m$, respectively. The neutrality of $\mathbf{B}'/\delta_1$ yields $\theta\vee\delta_1 = [\theta,1]\vee\delta_1 = [\theta,\delta_0]\vee\delta_1$ (by replacing 1 by $\delta_0\vee\delta_1$). Modularity gives $\theta = [\theta,\delta_0]\vee(\theta\wedge\delta_1)$ and $\theta\vee\delta_0 = \delta_0\vee(\theta\wedge\delta_1)$. Then

$$\begin{aligned}(\delta_0\vee\theta)\wedge(\delta_1\vee\theta) &= (\delta_0\vee(\theta\wedge\delta_1))\wedge(\delta_1\vee[\theta,\delta_0])\\ &= (\delta_0\vee(\theta\wedge\delta_1))\wedge\delta_1)\vee[\theta,\delta_0])\\ &= (\delta_0\wedge\delta_1)\vee(\theta\wedge\delta_1)\vee[\theta,\delta_0])\\ &\leq \theta.\end{aligned} \tag{13.10}$$

Thus

$$(\delta_0\vee\theta)\wedge(\delta_1\vee\theta) = \theta. \tag{13.11}$$

Since also,

$$(\delta_0\vee\theta)\vee(\delta_1\vee\theta) = 1, \tag{13.12}$$

it follows that $\mathbf{B}'/\theta \cong (\mathbf{B}'/(\delta_0\vee\theta))\times(\mathbf{B}'/(\delta_1\vee\theta))$, the product of an algebra in $\mathcal{A}$ and a quotient of $\mathbf{B}'/\delta_1$. Since $\mathbf{B}'/\delta_1$ is neutral, it has distributive congruence lattice, and every quotient of this algebra is isomorphic to a direct product of a subsystem of the simple algebras $\mathbf{B}_i$, $1 \leq i \leq m$.

We have now shown that every finite algebra in HSP($\mathbf{A}$) is isomorphic to a product of an algebra in $\mathcal{A}$ and a product of a system of simple non-Abelian direct factors of subalgebras of $\mathbf{A}$. Since $\mathcal{A}$ is directly representable, then $\mathcal{V}$ has, within isomorphism, only a finitely number of directly indecomposable finite algebras.

To prove (3), suppose that the finite algebra $\mathbf{A}$ has a Maltsev term and all subalgebras of $\mathbf{A}$ have uniform congruences. We first show that SP($\mathbf{A}$) is narrow, in fact, that every prime

divisor of the finite spectrum of $\mathrm{SP}(\mathbf{A})$ divides the cardinality of some subalgebra of $\mathbf{A}$. To do this, we use Lemma 13.4. If $\mathbf{B} \leq \mathbf{B}_0 \times \mathbf{B}_1$ is a subdirect product in a Maltsev variety, then $\mathbf{B}$ is the equalizer of surjective homomorphisms $\pi_i : \mathbf{B}_i \to \mathbf{K}$. If the kernel of π_1 is a uniform congruence on $\mathbf{B}_1$ with congruence classes of size c, then it is trivial to see that $|\mathbf{B}| = |\mathbf{B}_1|c$. Now suppose that that $\mathbf{F} \leq \mathbf{F}_1 \times \cdots \times \mathbf{F}_n$ is a subdirect product where $\mathbf{F}_i$ are subalgebras of $\mathbf{A}$. By induction on n, using the prece ding observation and the fact that all congruences on every $\mathbf{F}_i$ are uniform, we find that $|\mathbf{F}|$ is the product of $|\mathbf{F}_1|$ and a sequence of integers $c_2, \ldots, c_n$ where c_i is the block size of a uniform congruence on $\mathbf{F}_i$. Thus $|\mathbf{F}| = c_1 c_2 \ldots c_n$ with c_i a divisor of $|\mathbf{F}_i|$.

Now since $\mathrm{SP}(\mathbf{A})$ is narrow, our proof of Theorem 13.2, (2) $\Rightarrow$ (3), shows that the finite algebras in $\mathrm{SP}(\mathbf{A})$ have uniform congruences. If an algebra $\mathbf{B}$ has uniform congruences, then $|\mathbf{B}/\theta|$ divides $|\mathbf{B}|$ for every congruence θ. Thus $\mathrm{HSP}(\mathbf{A})$ is narrow, as claimed. □

In [44] it is proven that on any finite set A there altogether only finitely many clones F so that the algebra $\langle A, F\rangle$ satisfies the conditions of (2) in Theorem 13.5. Each of these clones is generated by its operations of $|A| + 2$ variables. To finish our presentation of this topic, we characterize, in a fashion, the directly representable affine varieties.

13.6 Theorem *For an Abelian, congruence modular variety $\mathcal{A}$, the following are equivalent:*

(1) *$\mathcal{A}$ is directly representable.*

(2) *$\mathcal{A}$ is locally finite and the polynomially equivalent variety ${}_{\mathbf{R}}\mathcal{M}$ of modules is directly representable.*

(3) *$\mathcal{A}$ is locally finite and the ring of $\mathcal{A}$, $\mathbf{R}$, is a finite ring of finite representation type.*

Proof If $\mathbf{A} \in \mathcal{A}$ and $\mathbf{M} \in {}_{\mathbf{R}}\mathcal{M}$ and $\mathcal{A}$ and $\mathbf{M}$ are polynomially equivalent, then these algebras have the same universe and the same congruence lattice $\mathbf{L}$. Hence $\mathbf{A}$ is directly indecomposable if and only if $\mathbf{M}$ is directly indecomposable, since direct decompositions in algebras with permuting congruences correspond to complement pairs of congruences—(θ, ψ) with $\theta \cap \psi = 0$, $\theta \vee \psi = 1$. If $\mathcal{A}$ is locally finite, then $\mathbf{R}$ is finite, and each of these varieties has, for each integer n, only finitely many non-isomorphic n-element algebras. (They are all homomorphic images of the free algebra on n generators in the variety.) Then $\mathcal{A}$ is directly representable if and only if ${}_{\mathbf{R}}\mathcal{M}$ is directly representable if and only if there is a finite bound to the size of finite directly indecomposable algebras in $\mathcal{A}$. A ring $\mathbf{R}$ with the property that up to isomorphism there are only finitely many finitely generated, directly indecomposable, $\mathbf{R}$-modules is said to be of finite representation type. The equivalence of (1), (2) and (3) should now be clear. □

14 Varieties with very few models

14.1 Definition For a class $\mathcal{C}$ of algebras and a cardinal k, let $\mathrm{G}_{\mathcal{C}}(k)$ denote the number of pairwise non-isomorphic members of $\mathcal{C}$ that are generated by at most k elements. We call this function, restricted to positive integral k, the *G-spectrum* (or *generative complexity*) of $\mathcal{C}$. We say that $\mathcal{C}$ *has very few models* if and only if there is a positive integer N such that for all positive integral $k > 1$, $\mathrm{G}_{\mathcal{C}}(k) \leq k^N$.

Below is the chief result of P. M. Idziak, R. McKenzie [23], which will be proved in this section.

14.2 Theorem *A locally finite, congruence modular, variety has very few models if and only if it is Abelian and polynomially equivalent to the variety of unitary modules over some finite ring of finite representation type.*

In the next two lemmas, $\mathcal{V}$ denotes a fixed, locally finite, congruence modular variety with very few models. We first show that all finite algebras in $\mathcal{V}$ are nilpotent. Then by Corollary 10.6, the finite algebras in $\mathcal{V}$ have permuting congruences. Since the free algebra on three generators in $\mathcal{V}$ is finite, it follows that $\mathcal{V}$ has a Maltsev term. Then to prove that $\mathcal{V}$ is Abelian, it suffices to show that all finite algebras in $\mathcal{V}$ are Abelian. Assuming that this fails, we show by direct construction that $\mathrm{G}_{\mathcal{V}}(k)$ is not bounded by any polynomial function of k, thus getting a contradiction. Our proofs will be modifications of those appearing in P. M. Idziak, R. McKenzie [23]. As we mentioned in Section 11, all locally finite varieties with very few models have recently been completely characterized in P. M. Idziak, R. McKenzie, M. Valeriote [24]. Each such variety consists entirely of Abelian algebras.

Notation The following notation will be used in the next two lemmas. For any sets $B \subseteq X$, and elements a, b in an algebra $\mathbf{A}$, we define a member of $\mathbf{A}^X$: $[a,b]_B$ denotes the function $f \in A^X$ such that $f(x) = b$ for $x \in B$ and $f(x) = a$ for $x \in X \setminus B$. Then for $x \in X$, we use $[a,b]_x$ to denote $[a,b]_B$ with $B = \{x\}$.

14.3 Lemma *Every finite algebra in $\mathcal{V}$ is nilpotent.*

Proof Assume that this fails. By taking a quotient of a finite non-nilpotent algebra, we can find a finite algebra $\mathbf{A} \in \mathcal{V}$ with a minimal congruence μ such that $[1,\mu] = \mu$. For $n > 0$, let X be a set of cardinality 2^n and let $\{X_{i,j} : 0 \leq i < n, 0 \leq j < 2\}$ be a system of $2n$ subsets of X so that for all $x \in X$ there is a function $p : \{0,1,\ldots,n-1\} \to \{0,1\}$ such that

$$\{x\} = \bigcap_{i<n} X_{i,p(i)}.$$

For example, we can take $X_{i,0}$ and $X_{i,1}$ to be B_i and its complement, where $B_0, \ldots, B_{n-1}$ is a set of generators of the Boolean algebra of all subsets of X.

Let $\mathbf{K}_n$ be the subalgebra of $\mathbf{A}^X$ generated by the set of all functions $[a,b]_{X_{i,0}}$ where a and b are any two elements of A and $0 \leq i \leq n-1$. (See the note on notation above.) Thus $\mathbf{K}_n$ is generated by a set of $a(a-1)n+a$ elements, where $a = |\mathbf{A}|$. We shall show that $\mathbf{K}_n$ has a set of $2^n + 1$ pairwise non-isomorphic homomorphic images. Since these are all generated by at most $a(a-1)n + a$ elements, then we can conclude that $\mathrm{G}_{\mathcal{V}}(a(a-1)n+1) \geq 2^n$. But this conclusion is obviously incompatible with the assumption that $\mathcal{V}$ has very few models.

Suppose that $X = \{x_0, \ldots, x_{2^n-1}\}$. For $0 \leq i \leq 2^n$ let θ_i be the congruence of $\mathbf{K}_n$ consisting of all pairs $\langle f, g\rangle \in \mathbf{K}_n^2$ such that $f(x_j) = g(x_j)$ for all $0 \leq j < i$. Then for $i < j$ we have $\theta_j \leq \theta_i$. We shall show that $\theta_j < \theta_i$. This will imply that $|\mathbf{K}_n/\theta_i| < |\mathbf{K}_n/\theta_j|$ so that the two quotient algebras are non-isomorphic, as desired.

Actually, given $x \in X$, we shall show that $\mathbf{K}_n$ contains two distinct functions f, g such that $f(y) = g(y)$ for all $y \in X \setminus \{x\}$. Taking $x = x_i$, this certainly implies that $\theta_i > \theta_j$ whenever $i < j \leq 2^n$.

So let $x \in X$ and write

$$\{x\} = \bigcap_{i<n} X_{i,p(i)},$$

where p is a certain function mapping $\{0, 1, \ldots, n-1\} \to \{0, 1\}$. We shall now produce, by induction on i, pairs of functions $\langle f_i, g_i \rangle \in K_n^2$ for $0 \leq i \leq n-1$, so that where $[\![f_i \neq g_i]\!]$ is the set of all $z \in X$ with $f_i(z) \neq g_i(z)$, we have

$$[\![f_i \neq g_i]\!] = \bigcap_{0 \leq j \leq i} X_{i,p(i)}.$$

Then f_{n-1}, g_{n-1} will be the desired pair of functions f, g with $[\![f \neq g]\!] = \{x\}$. The inductive construction requires that $f_i(z) \mu g_i(z)$ for all $z \in X$—i.e., $\langle f_i, g_i \rangle \in \mu_X$—and that each of f_i, g_i is constant on the set $[\![f_i \neq g_i]\!]$.

Choosing $\langle a, b \rangle \in \mu$, $a \neq b$, we put $f_0 = [a, a]_{X_{0,p(0)}}$, $g_0 = [a, b]_{X_{0,p(0)}}$ so that $\{f_0, g_0\} \subseteq K_n$, $\langle f_0, g_0 \rangle \in \mu_X$, $[\![f_0 \neq g_0]\!] = X_{0,p(0)}$ and each of f_0, g_0 is constant on $X_{0,p(0)}$. Now suppose that $i < n-1$ and we have succeeded in constructing f_i, g_i with the required properties. Let a_i, b_i be the constant value of f_i, respectively g_i, on the set $[\![f_i \neq g_i]\!]$. Since $0 \prec \mu = [1, \mu]$, then $\langle a_i, b_i \rangle$ does not belong to the center of $\mathbf{A}$. Hence there must exist a term $t(u, \bar{w})$ and tuples of elements $\bar{c}, \bar{d}$ in $\mathbf{A}$ so that

$$t(a_i, \bar{c}) = t(a_i, \bar{d}) \leftrightarrow t(b_i, \bar{c}) \neq t(b_i, \bar{d}).$$

Without losing generality, assume that $t(b_i, \bar{c}) = t(b_i, \bar{d})$ and $t(a_i, \bar{c}) \neq t(a_i, \bar{d})$. Taking γ in the Shifting Lemma (Lemma 6.6) to be the equality relation, we find that there is a Day term $m(x, y, z, u)$ such that

$$m(t(a_i, \bar{c}), t(a_i, \bar{c}), t(a_i, \bar{d}), t(a_i, \bar{d})) \neq m(t(a_i, \bar{c}), t(b_i, \bar{c}), t(b_i, \bar{d}), t(a_i, \bar{d})).$$

Let a_{i+1} be the left hand side of this inequality and b_{i+1} be the right. We can choose tuples of (two-valued) functions $\bar{h}, \bar{k}$ in K_n so that for all $z \in X_{i+1,p(i+1)}$, $\bar{h}(z) = \bar{c}$ and $\bar{k}(z) = \bar{d}$ while for all $z \in X_{i+1,1-p(i+1)}$, $\bar{h}(z) = \bar{k}(z)$. Then consider the functions

$$\begin{aligned} f_{i+1} &= m(t(f_i, \bar{h}), t(f_i, \bar{h}), t(f_i, \bar{k}), t(f_i, \bar{k})) \\ g_{i+1} &= m(t(f_i, \bar{h}), t(g_i, \bar{h}), t(g_i, \bar{k}), t(f_i, \bar{k})). \end{aligned}$$

Since $\mathcal{V} \models m(u, v, v, u) \approx u$, then $f_{i+1}(z) = g_{i+1}(z)$ for all $z \in X$ for which either $f_i(z) = g_i(z)$ or $\bar{h}(z) = \bar{k}(z)$. In particular, $f_{i+1}(z) = g_{i+1}(z)$ when $z \notin \bigcap_{j \leq i+1} X_{j,p(j)}$. On the other hand, if $z \in \bigcap_{j \leq i+1} X_{j,p(j)}$ then

$$f_{i+1}(z) = a_{i+1} \neq b_{i+1} = g_{i+1}(z).$$

The two functions f_{i+1}, g_{i+1} obviously satisfy all our requirements. This concludes our proof of the lemma. □

14.4 Lemma *Every algebra in $\mathcal{V}$ is Abelian.*

Proof We assume that the lemma is false, in order to get a contradiction. let $\mathbf{A}$ be a non-Abelian algebra in $\mathcal{V}$ of least cardinality. Then it follows that $\mathbf{A}$ is finite, is subdirectly

irreducible, and where μ is the monolith of $\mathbf{A}$, we have that $\mathbf{A}/\mu$ is Abelian. By Lemma 14.3, we have $[1_A, \mu] = 0_A$, and our assumption of least cardinality implies that $\mu = [1_A, 1_A]$.

Our proof will consist in constructing, for every structure (X, E) consisting of an equivalence relation E over a finite set X, an algebra $\mathbf{R}(X, E) = \mathbf{A}^X/\Theta_E$ (for a certain congruence Θ_E on $\mathbf{A}^X$), and proving that $\mathbf{R}(X, E_1) \cong \mathbf{R}(X, E_2)$ if and only if $(X, E_1) \cong (X, E_2)$. If $|X| = k$ then the number of non-isomorphic equivalence-relation structures (X, E) is $\pi(k)$, the number of partitions of the integer k. Thus $\mathbf{A}^k$ has at least $\pi(k)$ non-isomorphic quotient algebras.

We shall show that $\mathbf{A}^k$ is generated by a set of at most $|A|k$ many elements. Thus it will follow that $\mathrm{G}_{\mathcal{V}}(|A|^2 k) \geq \pi(k)$. But $\pi(k)$ is known to be asymptotic to

$$\frac{1}{4k\sqrt{3}} e^{\left(\pi\sqrt{2k/3}\right)}$$

(confer G. E. Andrews [1, p. 70]). Thus we have a clear contradiction to our assumption that $\mathcal{V}$ has very few models. The contradiction will establish that all algebras in $\mathcal{V}$ are Abelian.

Let $d(x, y, z)$ be a Maltsev term for $\mathcal{V}$. Suppose that $X = \{1, \ldots, k\}$. Choose $a_0 \in A$. We show that $\mathbf{A}^X$ is generated by the set

$$G = \{[a_0, b]_x : b \in A \text{ and } x \in X\},$$

thus establishing that $\mathbf{A}^X$ is $|A|k$-generated. Indeed, if $(a_1, a_2, \ldots, a_k) \in A^X$, then each of the functions $f_i = [a_0, a_i]_i$ $(0 \leq i \leq k)$ belongs to G, and therefore

$$(a_1, \ldots, a_k) = d(d(\ldots(d(d(f_1, f_0, f_2), f_0, f_3)\ldots, f_0, f_{k-1}), f_0, f_k)$$

belongs to the subalgebra generated by G.

Now let (X, E) be a finite equivalence-relation structure. Before defining Θ_E, we choose and fix a non-trivial μ-equivalence class N, and an element of N which we will denote by 0. Let $\mathbf{A}|_N$ denote the set N supplied with all the functions $f : N^m \to N$ (for any positive integral m) such that $f = g|_N$ for some polynomial operation g of the algebra $\mathbf{A}$. Since μ is an Abelian congruence, and $d|_N$ is a Maltsev operation on N, then $\mathbf{A}|_N$ is an Abelian algebra in a certain congruence modular variety. By Theorem 9.10, $\mathbf{A}_N$ is polynomially equivalent to a module $\mathbf{M}$ over a ring $\mathbf{R}$ with unit. Without any loss of generality, we can assume that 0 is the zero-element of this module.

Where $\bar{0}$ denotes the constant function in N^X with value 0, we define Θ_E to be the congruence relation of $\mathbf{A}^X$ generated by the set

$$G_E = \left\{(f, \bar{0}) : f \in N^X \text{ and } \sum_{x \in Z} f(x) = 0 \text{ for all } Z \in X/E\right\}.$$

The sums in this definition are finite sums in the module.

Since all generating pairs of Θ_E are $\bar{\mu}$-related, we get

$$\Theta_E \subseteq \bar{\mu}, \tag{14.1}$$

where $\bar{\mu}$ is the kernel of the homomorphism of $\mathbf{A}^X$ onto $(\mathbf{A}/\mu)^X$. Moreover,

$$\text{if } f^0, f^1 \in N^X \text{ then } \langle f^0, f^1\rangle \in \Theta_E \iff \sum_{x \in Z} f^0(x) = \sum_{x \in Z} f^1(x) \text{ for all } Z \in X/E. \tag{14.2}$$

To prove the "if" in (14.2), note that the assumption that $\sum_{x\in Z} f^0(x) = \sum_{x\in Z} f^1(x)$ for all $Z \in X/E$ gives $f^0 - f^1 \equiv \bar{0} \pmod{\Theta_E}$ and therefore $\langle f^0, f^1\rangle \in \Theta_E$.

Conversely, assume that $f^0, f^1 \in N^X$ and $\langle f^0, f^1\rangle \in \Theta_E$. Then the congruence permutability of the variety generated by $\mathbf{A}$ implies that there is a polynomial, say n-ary, $H(x_1, \dots, x_n)$ of $\mathbf{A}^X$ and $f_1, \dots, f_n \in N^X$ such that $\sum_{x\in Z} f_i(x) = 0$ for all $Z \in X/E$ and

$$f^0 = H(f_1, \dots, f_n), \qquad f^1 = H(\bar{0}, \dots, \bar{0}).$$

This means that there is an $n+m$-ary polynomial

$$h(x_1, \dots, x_n, y_1, \dots, y_m)$$

of $\mathbf{A}$ and $g_1, \dots, g_m \in A^X$ with

$$f^0 = h(f_1, \dots, f_n, g_1, \dots, g_m), \qquad f^1 = h(\bar{0}, \dots, \bar{0}, g_1, \dots, g_m).$$

It follows that $h(a_1, \dots, a_n, g_1(x), \dots, g_m(x)) \in N$ whenever $\{a_1, \dots, a_n\} \subseteq N$ and $x \in X$ (since N is an equivalence class of the congruence μ).

Choose $x_0 \in X$ and put $Z = x_0/E$. We apply the fact that $[\mu, 1_A] = 0_A$ to the equation

$$h(\underline{0}, \dots, \underline{0}, \bar{g}(x)) - h(0, \dots, 0, \bar{g}(x)) = h(\underline{0}, \dots, \underline{0}, \bar{g}(x_0)) - h(0, \dots, 0, \bar{g}(x_0))$$

replacing the underlined elements to obtain

$$h(u_1, \dots, u_n, \bar{g}(x)) - h(0, \dots, 0, \bar{g}(x)) = h(u_1, \dots, u_n, \bar{g}(x_0)) - h(0, \dots, 0, \bar{g}(x_0))$$

for all $x \in Z$ and $u_1, \dots, u_n \in N$. Thus

$$h(u_1, \dots, u_n, \bar{g}(x)) = h(u_1, \dots, u_n, \bar{g}(x_0)) - h(0, \dots, 0, \bar{g}(x_0)) + h(0, \dots, 0, \bar{g}(x)).$$

The map $(u_1, \dots, u_n) \mapsto h(u_1, \dots, u_n, \bar{g}(x_0)) - h(0, \dots, 0, \bar{g}(x_0))$ is a polynomial of the module $\mathbf{M}$ that maps $(0, \dots, 0)$ to 0; thus it must be of the form $(u_1, \dots, u_n) \mapsto \sum_{1\le i\le n} \lambda_i u_i$ for some $\lambda_i \in R$. Thus

$$h(u_1, \dots, u_n, \bar{g}(x)) = \sum_{i=1}^{i=n} \lambda_i u_i + h(0, \dots, 0, \bar{g}(x))$$

for all $x \in Z$ and $u_1, \dots, u_n \in N$. This implies that

$$\begin{aligned} f^0(x) = h(f_1(x), \dots, f_n(x), \bar{g}(x)) &= \sum_{i=1}^{i=n} \lambda_i f_i(x) + h(0, \dots, 0, \bar{g}(x)) \\ &= \sum_{i=1}^{i=n} \lambda_i f_i(x) + f^1(x). \end{aligned}$$

Together with $\sum_{x\in Z} f_i(x) = 0$, this gives

$$\sum_{x\in Z} f^0(x) = \sum_{x\in Z} f^1(x)$$

as required in (14.2).

Now, for a subset $B \subseteq X$ we define the following congruences of $\mathbf{A}^X$:

$$\eta_B = \left\{\langle f, g\rangle \in A^X \times A^X : f_t = g_t \text{ for all } t \in B\right\}, \qquad \eta'_B = \eta_{X\setminus B}.$$

For a congruence ϕ of $\mathbf{A}$ we put

$$\phi_B = \left\{\langle f, g\rangle \in A^X \times A^X : \langle f_t, g_t\rangle \in \phi \text{ for all } t \in B\right\}, \qquad \phi'_B = \phi_B \cap \eta'_B\,.$$

Also, we write $\eta_t, \eta'_t, \phi_t, \phi'_t$ instead of $\eta_{\{t\}}, \eta'_{\{t\}}, \phi_{\{t\}}, \phi'_{\{t\}}$, respectively. For any congruence γ of $\mathbf{A}^X$, the congruence $(\gamma \vee \Theta_E)/\Theta_E$ of $\mathbf{A}^X/\Theta_E$ will be denoted by $\tilde{\gamma}$.

Next, we observe that

$$\mu'_t \text{ is the unique atom of } \mathrm{Con}(\mathbf{A}^X) \text{ that is below } \eta'_t. \tag{14.3}$$

To see this, note that $\eta_t \vee \eta'_t = 1$ and $\eta_t \wedge \eta'_t = 0$, hence by modularity in the congruence lattice, the intervals $I[0, \eta'_t]$ and $I[\eta_t, 1]$ are transposes, and hence isomorphic. Thus the interval $I[0, \eta'_t]$, isomorphic to $\mathrm{Con}\,\mathbf{A}$, has the unique atom $\mu_t \wedge \eta'_t = \mu'_t$.

Choose and fix any $a \in N \setminus \{0\}$, and for $x \in X$ put $a^x = [0, a]_x$. Note that we have $\mu'_x = \mathrm{Cg}(\bar{0}, a^x)$ (the congruence generated by the pair $\langle\bar{0}, a^x\rangle$) since μ'_x is an atom. Then since $\langle\bar{0}, a^x\rangle \notin \Theta_E$ (by (14.2)), the covering pair $0 \prec \mu'_x$ projects up to $\Theta_E \prec \Theta_E \vee \mu'_x$ and therefore

$$\tilde{\mu}'_x \text{ is an atom in } \mathrm{Con}(\mathbf{A}^X/\Theta_E). \tag{14.4}$$

Moreover,

$$\tilde{\mu}'_t = \tilde{\mu}'_s \iff \langle t, s\rangle \in E. \tag{14.5}$$

In fact, if $\langle t, s\rangle \in E$ then (14.2) gives $\langle a^t, a^s\rangle \in \Theta_E$, which easily yields the "if" direction in (14.5).

Conversely, suppose that $\langle t, s\rangle \notin E$. If $\tilde{\mu}'_t = \tilde{\mu}'_s$, then $\langle a^t, \bar{0}\rangle \in \Theta_E \vee \mu'_s$. Then by congruence permutability, we have an element $f \in A^X$ with

$$a^t \Theta_E f \mu'_s \bar{0}.$$

Since $\langle a^t, f\rangle \in \bar{\mu}$ then $f \in N^X$. Then applying (14.2) to $a^t\,\Theta_E\,f$ and $Z = s/E$ we get that $\sum_{z\in Z} f(z) = 0$. On the other hand, from $f\mu'_s\bar{0}$ we have that $f(z) = 0$ for all $z \in X \setminus \{s\}$. Consequently, $f = \bar{0}$, i.e., $\langle a^t, \bar{0}\rangle \in \Theta_E$, which contradicts (14.2).

Next we show:

$$\text{For } \gamma \in \mathrm{Con}(\mathbf{A}^X) \text{ and } t \in X \text{ either } [\gamma, 1] \subseteq \eta_t \text{ or } \mu'_t \subseteq [\gamma, 1]. \tag{14.6}$$

Indeed, suppose that $[\gamma, 1] \not\subseteq \eta_t$. Then there is a term $\tau(x, \bar{y})$, a pair $\langle f, g\rangle \in \gamma$ and tuples $\bar{c}, \bar{d}$ in $\mathbf{A}^X$ such that $\langle\tau(f, \bar{c}), \tau(f, \bar{d})\rangle \in \eta_t$ and $\langle\tau(g, \bar{c}), \tau(g, \bar{d})\rangle \notin \eta_t$. Let $\bar{d}' = [\bar{c}, \bar{d}]_t$. Then $\tau(f, \bar{c}) = \tau(f, \bar{d}')$, $\tau(g, \bar{c}) \neq \tau(g, \bar{d}')$, so that $0 \neq [\gamma, \eta'_t] \leq \eta'_t$. Then (14.3) gives $\mu'_t \subseteq [\gamma, 1]$, as required.

Let us call a congruence of $\mathbf{A}^X/\Theta_E$ *regular* if it is the only atom below a congruence of the form $[\gamma, 1]$. We prove:

$$\text{A congruence of } \mathbf{A}^X/\Theta_E \text{ is regular} \iff \text{it is of the form } \tilde{\mu}'_t \text{ for some } t \in X. \tag{14.7}$$

First, suppose that $\alpha, \gamma \geq \Theta_E$ are such that $\tilde{\alpha}$ is the unique atom below $[\gamma/\Theta_E, 1/\Theta_E]$. Then $[\gamma, 1] \neq 0$ and we can choose t such that $[\gamma, 1] \not\leq \eta_t$. By (14.6), we have $\mu'_t \leq [\gamma, 1]$. Thus $\tilde{\mu}'_t \leq [\gamma/\Theta_E, 1/\Theta_E]$. By uniqueness of $\tilde{\alpha}$, we have that $\tilde{\alpha} = \tilde{\mu}'_t$ as required.

To prove that $\tilde{\mu}'_t$ is regular, it suffices to show that $\mu'_t \vee \Theta_E$ is the only congruence α with $\Theta_E \prec \alpha \subseteq [\eta'_t, 1] \vee \Theta_E$. Obviously, $[\eta'_t, 1] \subseteq [1, 1] \subseteq \bar{\mu}$ and so $[\eta'_t, 1] \subseteq \eta'_t \cap \bar{\mu} = \mu'_t$. Since $\eta'_t \vee \eta_t = 1$ and $\mathbf{A} \models [1, 1] \neq 0$, then $[\eta'_t, 1] \neq 0$. We now have that $[\eta'_t, 1] = \mu'_t$ since μ'_t is an atom. We know that $\Theta_E \prec \mu'_t \vee \Theta_E$ by (14.3). Thus it follows that $\tilde{\mu}'_t$ is regular.

From (14.5) and (14.7) we have that the number of E-classes in X can be recovered from $\mathbf{A}^X/\Theta_E$—it is the number of regular atoms in the congruence lattice of this algebra.

We will prove that non-isomorphic structures (X, E), (X, E') give rise to non-isomorphic algebras $\mathbf{A}^X/\Theta_E$, $\mathbf{A}^X/\Theta_{E'}$ by showing that the sizes of the equivalence classes of E are also recoverable from $\mathbf{A}^X/\Theta_E$.

Note that for the center ζ of $\mathbf{A}$ we have $\mu \leq \zeta < 1$. Obviously,

$$\mathbf{A}^X/\zeta_B \text{ is isomorphic to } (\mathbf{A}/\zeta)^B, \text{ for any } B \subseteq X. \tag{14.8}$$

We have $\Theta_E \subseteq \bar{\mu} \subseteq \zeta_X \subseteq \zeta_B$ for $B \subseteq X$, whence $(\mathbf{A}^X/\Theta_E)/(\zeta_B/\Theta_E) \cong (\mathbf{A}/\zeta)^B$ and $|B|$ is the logarithm of the cardinality of this algebra to the base $|\mathbf{A}/\zeta|$. Thus, we can finish the proof of this lemma by showing that the set of congruences of the form $\zeta_{X\setminus Z}/\Theta_E$ with Z ranging over X/E is definable in $\mathbf{A}^X/\Theta_E$.

Let us call a congruence δ of $\mathbf{A}^X/\Theta_E$ *co-regular* if δ is a maximal congruence with the property that the commutator $[\delta, 1]$ contains exactly one atom. We conclude our proof by showing:

$$\text{A congruence } \delta \text{ is co-regular } \iff \delta = \zeta_{X\setminus Z}/\Theta_E \text{ for some } Z \in X/E. \tag{14.9}$$

To begin, let $t \in Z \in X/E$. Now $\zeta_{X\setminus Z} \vee \eta_t = 1$, so $[\zeta_{X\setminus Z}, 1] \vee \eta_t = [1, 1] \vee \eta_t = \mu_t$, implying $[\zeta_{X\setminus Z}, 1] \not\leq \eta_t$. Then by (14.6), $[\zeta_{X\setminus Z}, 1] \geq \mu'_t$. On the other hand, clearly we have

$$[\zeta_{X\setminus Z}, 1] \leq \mu'_Z = \bigvee_{s\in Z} \mu'_s$$

and this combined with the above conclusion yields

$$[\zeta_{X\setminus Z}, 1] = \mu'_Z = \bigvee_{s\in Z} \mu'_s.$$

Then

$$[\zeta_{X\setminus Z}/\Theta_E, 1/\Theta_E] = [\zeta_{X\setminus Z}, 1] \vee \Theta_E = \bigvee_{s\in Z} \tilde{\mu}'_s = \tilde{\mu}'_t.$$

Next, suppose that $\Theta_E \leq \gamma$ and γ/Θ_E is co-regular. Then $[\gamma/\Theta_E, 1/\Theta_E]$ contains exactly one atom, and by (14.7), this atom must be of the form $\tilde{\mu}'_t$. Say $\tilde{\mu}'_t \leq [\gamma/\Theta_E, 1/\Theta_E]$ and $t/E = Z$. Then for $s \in X \setminus Z$, $[\gamma, 1]$ cannot contain μ'_s and so by (14.6), $[\gamma, 1] \leq \eta_s$. Since this holds for all $s \in X \setminus Z$, we have $[\gamma, 1] \leq \eta'_Z$. This is easily seen to imply that $\gamma \leq \zeta_{X\setminus Z}$. We have seen in the last paragraph that $[\zeta_{X\setminus Z}/\Theta_E, 1/\Theta_E] = \tilde{\mu}'_t$. The maximality of γ now gives that $\gamma = \zeta_{X\setminus Z}$. The results of this paragraph and the last one, combined, yield (14.9). □

The next lemma completes our proof of Theorem 14.2.

14.5 Lemma *A locally finite affine variety has very few models if and only if it is directly representable.*

Proof Assume that $\mathcal{A}$ is a locally finite, congruence modular, Abelian variety, and let $\mathbf{R}$ denote the finite ring such that $\mathcal{A}$ is polynomially equivalent with ${}_{\mathbf{R}}\mathcal{M}$.

For a positive integer n, let $\mathbf{F} = \mathbf{F}_{\mathcal{A}}(n)$ be the free algebra in $\mathcal{A}$ freely generated by $x_1, \ldots, x_n$. Let $\mathbf{M}$ be the $\mathbf{R}$-module polynomially equivalent to $\mathbf{F}$, with x_n chosen as the zero element. Every element $w \in F$ can be written as $t^{\mathbf{F}}(x_1, \ldots, x_n)$ for a term t, and the operation $t^{\mathbf{F}}$ in F can be expressed as

$$t^{\mathbf{F}}(b_1, \ldots, b_n) = \sum_{1 \leq i \leq n} r_i b_i + c$$

for some $r_i \in R$ and $c \in F$. Thus $w = \sum_{1 \leq i \leq n-1} r_i x_i + c$ is determined by the sequence $\langle r_i : 1 \leq i \leq n-1 \rangle$ and the element $t^{\mathbf{F}}(x_n, \ldots, x_n) = c$. This means that $f_n = |\mathbf{F}_{\mathcal{A}}(n)| \leq r^{n-1} f_1$ where $r = |R|$ and $f_1 = |\mathbf{F}_{\mathcal{A}}(1)|$.

Now suppose that $\mathcal{A}$ is directly representable. Let $\mathbf{D}_0, \ldots, \mathbf{D}_{k-1}$ be a list of all the directly indecomposable finite algebras in $\mathcal{A}$, up to isomorphism. If $\mathbf{B}$ is an n-generated member of $\mathcal{A}$, then $|B| \leq f_n \leq r^{n-1} f_1$. We can write

$$\mathbf{B} \cong \mathbf{D}_0^{\ell_0} \times \cdots \times \mathbf{D}_{k-1}^{\ell_{k-1}}$$

for some non-negative integers ℓ_i, and here

$$\sum_i \ell_i \leq \log_2(f_n) \leq M(n-1)$$

for some positive integer M, independently of n. The number of solutions (m_i) of the inequality $\sum_i m_i \leq M(n-1)$ is

$$\binom{M(n-1)+k}{k} \leq (M(n-1)+k)^k.$$

Thus $\mathrm{G}_{\mathcal{A}}(n) \leq (M(n-1)+k)^k$ which estabishes that $\mathcal{A}$ has very few models.

Conversely, suppose that $\mathcal{A}$ is not directly representable. Let $\mathcal{A}_p$ denote the class of all algebras in $\mathcal{A}$ that have a one-element subalgebra. Since $\mathcal{A}$ has at most $2^{f_n^2}$ non-isomorphic n-element members, then there is no finite bound on the size of the finite directly indecomposable members of $\mathcal{A}$. For $\mathbf{A} \in \mathcal{A}$, the algebra $\mathbf{A}_\nabla \in \mathcal{A}_p$ has the same corresponding module, up to isomorphism, and is directly indecomposable if and only if $\mathbf{A}$ is. (Cf. the discussion at the end of Section 9.) Where $d(k)$ denotes the number of non-isomorphic, directly indecomposable, k-generated algebras in $\mathcal{A}_p$, we thus have that $d(k)$ is unbounded.

For a fixed k, let $\mathbf{D}_0, \ldots, \mathbf{D}_{d-1}$ be pairwise non-isomorphic, directly indecomposable, k-generated members of $\mathcal{A}_p$, where $d(k) = d$. By a theorem of G. Birkhoff (see, for example, R. McKenzie, G. McNulty, W. Taylor [37], Theorem 5.3), finite algebras with permuting congruences and a one-element subalgebra have the unique factorization property. This means that if $\langle m_i \rangle_{i<d}$, $\langle m'_i \rangle_{i<d}$ are sequences of non-negative integers and $\prod_{i<d} \mathbf{D}_i^{m_i} \cong \prod_{i<d} \mathbf{D}_i^{m'_i}$ then $m_i = m'_i$ for all $i < d$. Now if $\sum_{i<d} m_i \leq n$ then $\prod_{i<d} \mathbf{D}_i^{m_i}$ is nk generated. (This follows by the same argument used in the proof of Lemma 14.4 to show that $\mathbf{A}^k$ is $|A|k$-generated.)

Thus $\mathrm{G}_{\mathcal{A}_p}(nk)$ is at least as great as the number of systems $\langle m_i\rangle_{i<d}$ of non-negative integers satisfying $\sum_{i<d} m_i \leq n$, i.e., we have

$$\mathrm{G}_{\mathcal{A}_p}(nk) \geq \binom{n+d(k)}{n}.$$

In particular, $\mathrm{G}_{\mathcal{A}}(k^2) \geq \binom{k+d(k)}{k}$. For any fixed positive integer M, we have $d(k) \geq M$ for large k, and for such k, it follows that $\mathrm{G}_{\mathcal{A}}(k^2) \geq \binom{k+M}{k}$. Since this is a polynomial of degree M in k, then $\mathrm{G}_{\mathcal{A}}(k)$ cannot be bounded for all k by a polynomial of degree $< M/2$. This ends our proof that if $\mathcal{A}$ is not directly representable then $\mathrm{G}_{\mathcal{A}}$ is not bounded by any polynomial function. □

15 Problems

15.1 Problem Is there an algorithm to determine, given a finite ring $\mathbf{R}$ with unit, whether the variety $\mathbf{R}\mathcal{M}$ of left unitary $\mathbf{R}$-modules is decidable? F. Point, M. Prest [40] and M. Prest [41] provide a starting point for the exploration of what is known about this problem.

15.2 Problem Is there an algorithm to determine, given a finite algebra $\mathbf{F}$ of finite type such that $\mathrm{HSP}(\mathbf{F})$ has modular congruence lattices, whether $\mathrm{HSP}(\mathbf{F})$ (or $\mathrm{SP}(\mathbf{F})$) is finitely axiomatizable? By a famous theorem of K. Baker [2], every finitely generated congruence distributive variety of finite type is finitely axiomatizable. R. McKenzie [36] proved that there is no algorithm to determine if $\mathrm{HSP}(\mathbf{F})$ is finitely axiomatizable where $\mathbf{F}$ ranges over all finite algebras with one binary operation.

15.3 Problem Prove or disprove that for every finite set A and every operation $m = m(x, y, z)$ over A that satisfies Maltsev's equations $m(x, x, y) \approx y \approx m(y, x, x)$, there are only countably many clones of operations on A containing m. A. A. Bulatov, P. M. Idziak [3] has results on this problem.

15.4 Problem Is it true that every congruence modular variety of finite type that is residually finite has a finite residual bound? K. Kearnes, R. Willard [29] proved that this implication holds for congruence distributive varieties of finite type, and more generally for congruence meet semi-distributive varieties of finite type.

15.5 Problem Characterize the locally finite congruence modular varieties $\mathcal{V}$ that possess first-order definable principal congruences—i.e., there is a first-order formula $\theta(x, y, u, v)$ so that for all $\mathbf{A} \in \mathcal{V}$ and $\{a, b, c, d\} \subseteq A$, $\mathbf{A} \models \theta(a, b, c, d)$ if and only if $\langle a, b\rangle$ lies in the congruence of $\mathbf{A}$ generated by $\langle c, d\rangle$.

References

[1] G. E. Andrews, *The Theory of Partitions*, Encyclopedia Math. Appl. 2, Addison-Wesley, Reading, MA, 1976.

[2] K. Baker, Finite equational bases for finite algebras in congruence-distributive equational class, *Advances in Math.* **24** (1977), 207–243.

[3] A. A. Bulatov and P. M. Idziak, Counting Maltsev clones on small sets, *Discrete Math.* **268** (1–3) (2003), 59–80.

[4] S. Burris and R. McKenzie, *Decidability and Boolean Representations*, Memoirs Amer. Math. Soc. **32** (246), 1981.

[5] G. Czédli and E. Horváth, All congruence identities implying modularity have Maltsev conditions, preprint.

[6] G. Czédli, E. Horváth, and P. Lipparini, Optimal Maltsev conditions for congruence modular varieties, preprint.

[7] A. Day, A characterization of modularity for congruence lattices of algebras, *Canadian Math. Bull.* **12** (1969), 167–173.

[8] A. Day and R. Freese, A characterization of identities implying congruence modularity, *Can. J. Math.* **32** (1980), 1140–1167.

[9] R. Dedekind, Uber die von drei modulun erzeugte dualgruppe, *Math. Ann.* **53** (1900), 371–403.

[10] I. Fleischer, A note on subdirect products, *Acta Math. Acad. Sci. Hungar.* **6** (1955), 463–465.

[11] R. Freese and R. McKenzie, Residually small varieties with modular congruence lattices, *Trans. Amer. Math. Soc.* **264** (1981), 419–430.

[12] R. Freese and R. McKenzie, *Commutator Theory for Congruence Modular Varieties*, London Mathematical Society Lecture Note Series **125**, Cambridge Univ. Press, 1987.

[13] H.-P. Gumm, Uber die losungermenge von gleichungssystemen uber allgemeinen algebren, *Math Z.* **162** (1978), 51–62.

[14] H.-P. Gumm, Algebras in permutable varieties: geometrical properties of affine algebras, *Algebra Universalis* **9** (1) (1979), 8–34.

[15] H.-P. Gumm, An easy way to the commutator in modular varieties, *Proc. Amer. Math. Soc.* **80** (1980), 393–397.

[16] H.-P. Gumm, Congruence modularity is permutability composed with distributivity, *Archiv der Math. (Basel)* **36** (1981), 569–576.

[17] H.-P. Gumm, *Geometrical Methods in Congruence Modular Varieties*, Memoirs Amer. Math. Soc. **36** (286), 1983.

[18] H.-P. Gumm and C. Herrmann, Algebras in modular varieties: Baer refinements, cancellation and isotopy, *Houston J. Math.* **5** (1979), 503–523.

[19] J. Hagemann and C. Herrmann, A concrete ideal multiplication for algebraic systems and its relation to congruence distributivity, *Arch. Math. Basel* **32** (1979), 234–245.

[20] C. Herrmann, Affine algebras in congruence modular varieties, *Acta Sci. Math. (Szeged)* **41** (1979), 119–125.

[21] D. Hobby and R. McKenzie, *The Structure of Finite Algebras*, Amer. Math. Soc. Contemporary Mathematics Series **76**, 1988 [revised edition 1996].

[22] P. M. Idziak, A characterization of finitely decidable congruence modular varieties, *Trans. Amer. Math. Soc.* **349** (1997), 903–934.

[23] P. M. Idziak and R. McKenzie, Varieties with polynomially many models, *Fundamenta Mathematicae* **170** (2001), 53–68.

[24] P. M. Idziak and R. McKenzie, and M. Valeriote, The structure of locally finite varieties with polynomially many models, preprint.

[25] B. Jónsson, Algebras whose congruence lattices are distributive, *Math. Scand.* **21** (1967), 110–121.

[26] K. A. Kearnes, On the relationship between AP, RS and CEP, *Proc. Amer. Math. Soc.* **105** (4) (1989), 827–839.

[27] K. A. Kearnes and R. McKenzie, Commutator theory for relatively modular quasivarieties, *Trans. Amer. Math. Soc.* **331** (1992), 465–502.

[28] K. A. Kearnes and A. Szendrei, The relationship between two commutators, *Inter. J. Algebra. Comput.* **8** (1998), 497–531.

[29] K. A. Kearnes and R. Willard, Residually finite, congruence meet semi-distributive varieties of finite type have a finite residual bound, *Proc. Amer. Math. Soc.* **127** (10) (1999), 2841–2850.

[30] H. Lakser and W. Taylor and S. T. Tschantz, A new proof of Gumm's theorem, *Algebra Universalis* **20** (1985), 115–122.

[31] A. I. Maltsev, On the general theory of algebraic systems, *Mat. Sbornik* **77** (1954), 3–20.

[32] R. McKenzie, Narrowness implies uniformity, *Algebra Universalis* **15** (1982), 67–85.

[33] R. McKenzie, Finite equational bases for congruence modular varieties, *Algebra Universalis* **24** (1987), 224–250.

[34] R. McKenzie and M. Valeriote, *The Structure of Decidable Locally Finite Varieties*, Birkhauser, Progress in Mathematics **79**, 1989.

[35] R. McKenzie, The residual bounds of finite algebras, *Inter. J. Algebra. Comput.* **6** (1996), 1–28.

[36] R. McKenzie, Tarski's finite basis problem is undecidable, *Inter. J. Algebra. Comput.* **6** (1996), 49–104.

[37] R. McKenzie, G. McNulty, and W. Taylor, *Algebras, Lattices, Varieties, Vol. I*, Wadsworth and Brooks/Cole, Monterey, CA, 1987.

[38] A. F. Pixley, Distributivity and permutability of congruence relations in equational classes of algebras, *Proc. Amer. Math. Soc.* **14** (1963), 105–109.

[39] A. F. Pixley, Local Malcev conditions, *Canad. Math. Bull.* **15** (1972), 559–568.

[40] F. Point and M. Prest, Decidabity for theories of modules, *J. London Math. Soc.* **38** (2) (1988), 193–206.

[41] M. Prest, *Model Theory and Modules*, London Math. Soc. Lecture Note Ser. **130**, Cambridge Univ. Press, 1988.

[42] R. W. Quackenbush, Equational classes generated by finite algebras, *Algebra Universalis* **1** (1971), 265–266.

[43] J. D. H. Smith, *Mal'cev Varieties*, Springer-Verlag, Lecture Notes in Math. **554**, New York, 1976.

[44] J.W. Snow, Generating primitive positive clones, *Algebra Universalis* **44** (2000), 169–185.

[45] W. Taylor, Some applications of the term condition, *Algebra Universalis* **14** (1982), 11–24.

[46] R. Wille, *Kongruenzklassengeometrien*, Springer-Verlag, Lecture Notes in Math. **113**, New York, 1970.

Epigroups

Lev N. SHEVRIN*

Department of Mathematics and Mechanics
Ural State University
Lenina 51, 620083 Ekaterinburg
Russia

Dedicated to Walter Douglas Munn on the occasion of his 75th birthday

Abstract

An epigroup is a semigroup in which some power of any element lies in a subgroup of the given semigroup. The class of epigroups includes a number of important classes of semigroups. Not only some particular types of epigroups but epigroups as such can serve as the subject matter of a substantial theory. In this survey the following topics concerning epigroups are considered: epigroups as unary semigroups, certain "nice" decompositions, finiteness conditions.

A semigroup S is called an *epigroup* if for any element x of S some power of x lies in some subgroup of S. The class of epigroups is very large: it contains all periodic semigroups and, in particular, all finite semigroups, all completely regular semigroups, all completely 0-simple semigroups. It also contains some important concrete semigroups; for example, the semigroup of all matrices over a division ring is an epigroup. The study of completely 0-simple, completely regular, and periodic semigroups was begun at the very outset of the development of the theory of semigroups; we mention in this regard the pioneer works [12, 77, 84]. In the course of semigroup-theoretic investigations during several decades, some properties of epigroups in general were revealed via the study of particular types of epigroups. The first work where epigroups were considered *per se* was the paper [51] in which these semigroups were called *pseudo-invertible*. This paper was, in its turn, influenced by the paper [15] in which the notion of the pseudo-inverse of an element had been introduced. Epigroups were considered later (as one of the objects or the main object of attention) in several dozens papers by different authors and under different names: *quasi-completely regular, group-bound, quasiperiodic* and some others. The term "epigroup" was suggested by the present writer in the late eighties and apparently has become fairly common by now. Being shorter and, what is especially essential, being expressed by a noun, such a term is more flexible and definitely more appropriate for a concept which pretends to be the object of a certain theory.

In this survey we present quite a number of basic facts about epigroups. Its contents are revealed by the titles of the sections and subsections. Notice that a very brief presentation

*The work was partially supported by the President Program of a support of the Leading Scientific Schools of the Russian Federation (grant 2227.2003.01) and by the Ministry of Education of the Russian Federation (grants E02–1.0–143 and 04.01.059).

V. B. Kudryavtsev and I. G. Rosenberg (eds.),
Structural Theory of Automata, Semigroups, and Universal Algebra, 331–380.

of these topics was given by the author earlier in [49, Sections A6 and A7], and, somewhat more extensively, in [102, §6].

I thank Boris Vernikov for technical assistance in preparing the text of this article.

1 Preliminaries

All standard semigroup-theoretic information can be found, for example, in Chapter A of the handbook [49]. We will recall some definitions and notation as needed. Let S be a semigroup. An element $a \in S$ is called a *group element* if a belongs to some subgroup of S. If S has a zero, then an element $a \in S$ is called a *nil-element* if $a^n = 0$ for some n. We will denote by E_S or simply by E the set of all idempotents of S, by $\operatorname{Gr} S$ the set of all group elements of S (the *group part* of S), and by $\operatorname{Reg} S$ the set of all regular elements of S. Obviously $E_S \subseteq \operatorname{Gr} S \subseteq \operatorname{Reg} S$. We will often view E as a partially ordered set with respect to the natural order: $e \leq f \iff ef = fe = e$. We say that an element a of a semigroup S *divides* an element $b \in S$, and we write $a \,|\, b$, if b belongs to the principal ideal $J(a)$. If S has a zero, we will denote by $\operatorname{Nil} S$ the set of all nil-elements of S; S is a *nilsemigroup* if $\operatorname{Nil} S = S$. A semigroup S with zero 0 is called *nilpotent* if there exists n such that $x_1 x_2 \cdots x_n = 0$ for any $x_1, x_2, \ldots, x_n \in S$; the smallest n with this property is called the *degree of nilpotency* of S. We will denote by G_e the maximal subgroup of S having the idempotent e as its identity element. Thus $\operatorname{Gr} S = \bigcup_{e \in E_S} G_e$. The equality $\operatorname{Gr} S = S$ just defines the completely regular semigroups. The set of all natural numbers is denoted by $\mathbb{N}$.

For an arbitrary subset M of a semigroup S we put

$$\sqrt{M} = \{x \in S \,|\, x^n \in M \text{ for some } n \in \mathbb{N}\}.$$

Thus the fact that S is an epigroup can be expressed by the equality $S = \sqrt{\operatorname{Gr} S}$. That S is a periodic semigroup can be expressed by the equality $S = \sqrt{E_S}$; as is known, this is equivalent to the property that, for any $a \in S$, the cyclic subsemigroup $\langle a \rangle$ generated by a is finite. If S has a zero, then $\operatorname{Nil} S = \sqrt{\{0\}}$.

For $e \in E$ we put

$$K_e = \sqrt{G_e}.$$

By definition, for any element a of an epigroup S there exists $e \in E_S$ such that $a \in K_e$. That the idempotent e is defined here uniquely is a consequence of the following property already mentioned in [51]: if $a^m \in G_e$, then $a^n \in G_e$ for any $n > m$. So one may imagine that elements of an epigroup S are pulled by idempotents to the group part of S, and each element is "attached" to a single idempotent. This shows that the set of all idempotents plays the role of a certain framework of an epigroup. The property noted means that any epigroup S is partitioned into the subsets K_e called *unipotency classes*. The partition into unipotency classes gives, so to speak, a rough structure of an arbitrary epigroup. A unipotency class of an epigroup does not have to be a subsemigroup; to find out when this takes place (and even when all the unipotency classes are subsemigroups) is the goal of Subsection 3.2.

The following well-known statement (see [51] and also [13, Theorem 2.55]) is one of the first general results concerning epigroups.

1.1 Proposition *A semigroup is a* [0-]*simple epigroup if and only if it is completely* [0-] *simple.*

A semigroup is called *stable* if, for any of its $\mathcal{J}$-classes, the sets of $\mathcal{L}$-classes and $\mathcal{R}$-classes contained in it are antichains with respect to the natural order ($L_a \leq L_b$ means $L(a) \subseteq L(b)$; $R_a \leq R_b$ means $R(a) \subseteq R(b)$); there are some properties equivalent to the one indicated in this definition. It is easy to show that any epigroup is a stable semigroup; this was done explicitly in [26]. The following statement points out some useful properties of stable semigroups, and hence of epigroups.

1.2 Proposition *In a stable semigroup:* (a) *Green's relations $\mathcal{J}$ and $\mathcal{D}$ coincide;* (b) *no subsemigroup is bicyclic;* (c) *the idempotents belonging to the same $\mathcal{D}$-class form an antichain.*

Assertion (a) here is presented, for instance, in [13, Theorem 6.45]; Assertion (b) appears explicitly, for instance, in [2]; Assertion (c) follows easily from (b) (without additional assumptions about the semigroup in question), as can be seen, for example, from the argument in the proof of Theorem 2.54 in [13].

The idempotent of the unipotency class to which an element a belongs will be denoted by e_a. The following properties were mentioned in [51].

1.3 Lemma (a) $ae_a = e_a a$; (b) $ae_a \in G_{e_a}$.

Lemma 1.3 easily implies the following assertions.

1.4 Corollary (a) *The group G_e is the kernel* (*i.e., the smallest ideal*) *of the subsemigroup $\langle K_e \rangle$ generated by K_e.* (b) *An idempotent e is the smallest element in the partially ordered set $E \cap \langle K_e \rangle$.*

The assertions in Lemma 1.3 and its Corollary pertain to the situation involving an arbitrary semigroup with idempotents. In the case of epigroups, one can state somewhat strengthened assertions by replacing the subsemigroup $\langle K_e \rangle$ by the subepigroup $\langle\langle K_e \rangle\rangle$ generated by K_e (see Subsection 2.1).

Recall that a *retract* of a semigroup S is a subsemigroup T for which there exists a homomorphism $\varphi : S \to T$ which is the identity mapping on T; such a homomorphism φ is called a *retraction*. A *retract ideal* is an ideal that is a retract. The following fact is well known.

1.5 Lemma *If T is a retract ideal of a semigroup S, then S is a subdirect product of T and the quotient semigroup S/T.*

The concepts of the index and period of an element of finite order in a semigroup can be extended to the case of elements of infinite order. The *period* of an element of infinite order will be assumed to be infinite and will be denoted by the symbol ∞. The *index* of an arbitrary element a of a semigroup S is defined as follows: if $\langle a \rangle \cap \operatorname{Gr} S \neq \emptyset$, it is the smallest natural number n such that $a^n \in \operatorname{Gr} S$; if $\langle a \rangle \cap \operatorname{Gr} S = \emptyset$, it is 0. The index of an element a will be denoted by $\operatorname{ind}(a)$. The elements of index 1 are precisely the group elements. As in the case of elements of finite order, the pair (*index, period*) will be called the *type* of a given element. Thus an element of infinite order has type (n, ∞), where n is a non-negative integer. In epigroups, and only in epigroups, all elements have indices that are natural numbers. If the indices of all elements of an epigroup S are bounded by a natural number, we say that S is an *epigroup of finite index*, and its *index* is defined to be $\max\{\operatorname{ind}(a) \mid a \in S\}$; the index of S is denoted by $\operatorname{ind} S$. The epigroups of index 1 are precisely the completely regular semigroups.

2 Epigroups as unary semigroups

2.1 Background

A *unary semigroup* is a semigroup with an additional unary operation. Note that important representatives of such algebras are completely regular and inverse semigroups; see the corresponding treatments, for instance, in [49, Sections A10 and A11]. Epigroups can be treated as unary semigroups as well. To introduce the corresponding unary operation, recall that, according to Assertion (b) of Lemma 1.3, for any element a of an epigroup we have $ae_a \in G_{e_a}$. Hence one may consider the element

$$\overline{a} = (ae_a)^{-1}, \tag{2.1}$$

where the right-hand side is the inverse in the group G_{e_a}. This element is called the *pseudo-inverse* of a, and the operation $a \mapsto \overline{a}$ is just the stated one. The following equalities hold:

$$a\overline{a} = \overline{a}a, \qquad a\overline{a}^2 = \overline{a}, \qquad a^{n+1}\overline{a} = a^n \text{ for some } n. \tag{2.2}$$

The equalities (2.2) have been used in [15] for the original definition of pseudo-inverses. The conditions (2.1) and (2.2) are equivalent in the sense that if a and $\overline{a}$ are some elements of a semigroup such that the equalities (2.2) hold, then a^n is a group element, and if $a^n \in G_e$, then $e = e_a$, and it is easy to ascertain that $\overline{a} \in G_e$ and condition (2.1) holds.

The definition of the pseudo-inverse of an element a of non-zero index by (2.1) is more convenient than by (2.2). Indeed, then we do not need to establish the uniqueness of $\overline{a}$ for a given a, the equalities (2.2) (where $n = \mathrm{ind}(a)$) follow directly from (2.1), and we can write them in the following form, which more clearly illuminates their "structural" meaning:

$$a\overline{a} = \overline{a}a = e_a, \qquad e_a\overline{a} = \overline{a}, \qquad a^n e_a = a^n. \tag{2.2$'$}$$

But the original form of the equalities (2.2) is still valuable, particularly in connection with defining epigroups by identities (see Subsection 2.2).

The idea of viewing epigroups as unary semigroups has been proposed in [104], and this approach was shown in three ways there: in formulating statements solving some problems, in posing some problems, in applying techniques. A number of results of this work are presented below in this section and, especially, in Section 3.

In a particular case when an epigroup S is in fact completely regular, the operation $a \mapsto \overline{a}$ turns into the usual unary operation $a \mapsto a^{-1}$ of such semigroups, i.e., $\overline{a}$ is none other than the inverse of a in the maximal subgroup containing a. Along with the basic unary operation $a \mapsto \overline{a}$ it is useful to consider the derivative operation $a \mapsto e_a$. The latter is linked with the former by the formula $e_a = a\overline{a}$. It is customary to denote e_a by a^ω in the theory of finite semigroups, and by a^0 in the theory of completely regular semigroups. For the general case of epigroups, we choose the notation a^ω here.

Repeated application of pseudo-inversion will be denoted by an additional bar. It follows from (2.1) that $\overline{\overline{a}} = a$ if and only if a is a group element. The same equality explains the identities $\overline{\overline{\overline{x}}} = \overline{x}$ and $\overline{\overline{x}} = x^2\overline{x}$ which were already mentioned in [15].

A *subepigroup* of an epigroup is a subset that is closed under both the operations of multiplication and pseudo-inversion. It is easy to see that the subepigroups of an epigroup are precisely the subsemigroups that are themselves epigroups. The simplest example of an

epigroup having subsemigroups which are not subepigroups is given by the infinite cyclic group (the subepigroups of a group are obviously none other than its subgroups).

Note the following two elementary properties of the principal character; the latter is a special case of a rudimentary general algebraic fact.

2.1 Observation *A homomorphic image of an epigroup is an epigroup, and any semigroup homomorphism φ of an epigroup S onto some epigroup is automatically an epigroup homomorphism, i.e., $\varphi(\overline{a}) = \overline{\varphi(a)}$ for any $a \in S$.*

2.2 Observation *Under a homomorphism of an epigroup, the inverse image of a subepigroup (in particular, the inverse image of an idempotent) is a subepigroup.*

A homomorphic image of a subsemigroup of a given semigroup is called a *divisor* (or *factor*) of the semigroup. A divisor is called a *Rees divisor* if it is the Rees quotient semigroup of some subsemigroup modulo an ideal. A divisor obtained from a subepigroup of an epigroup will be called, for brevity, an *epidivisor*. Observation 2.1 shows that any epidivisor of an epigroup is itself an epigroup. Since an ideal of an epigroup is obviously a subepigroup, among the epidivisors of an epigroup are all its principal factors.

The subepigroup generated by a subset A of a given epigroup will be denoted by $\langle\langle A\rangle\rangle$, while the single angular brackets are used to denote the subsemigroup generated by a set. (We could also suggest the alternative notation $\mathrm{ep}\langle A\rangle$, but the double angular brackets seem to contrast better with the subsemigroup notation.) If S is an epigroup and $S = \langle\langle A\rangle\rangle$ [resp., $S = \langle A\rangle$], then A will be called an *epigroup* [*semigroup*] *generating set* of S. It is worth noting that if A is an epigroup generating set, and $\overline{A} = \{\overline{a} \mid a \in A\}$, then the set $A \cup \overline{A}$ does not have to be a semigroup generating set of the given epigroup. Moreover, a finitely generated epigroup (even if it is completely regular) may have no finite semigroup generating sets: an example is the free completely regular semigroup of rank 2 (see [116][1]). We can say, however, that if S is an epigroup in which the mapping $a \mapsto \overline{a}$ is an endomorphism or an anti-endomorphism, then $S = \langle\langle A\rangle\rangle$ obviously implies $S = \langle A \cup \overline{A}\rangle$. Epigroups with these properties of the pseudo-inversion operation will be considered in Subsection 3.4.

It is natural to call an epigroup having a one-element epigroup generating set *monogenic* or, in accordance with traditional group-theoretic and semigroup-theoretic terminology, *cyclic*. We choose the latter. Consider an arbitrary cyclic epigroup $C = \langle\langle a\rangle\rangle$. Let $\mathrm{ind}(a) = n$. The equalities (2.2′) show that $\mathrm{Gr}\, C$ is the cyclic group generated by $\overline{a}$, $\mathrm{Gr}\, C = G_{e_a}$, C is a cyclic group if $n = 1$ (and only in this case), and $C \setminus \mathrm{Gr}\, C = \{a, \ldots, a^{n-1}\}$ if $n > 1$. In any event, $C = K_{e_a}$, C is commutative, and $\mathrm{Gr}\, C$ is its kernel. We see that the finite cyclic epigroups are precisely the finite cyclic semigroups whose structure and classification are textbook facts. Our aim now is to classify infinite cyclic epigroups. To do this, we need the construction described below.

Suppose T is an arbitrary semigroup, H a subsemigroup, and I an ideal of H. We take a set F of cardinality equal to that of $H \setminus I$ and disjoint from T; the image of any element

[1]Note, in order to avoid a misunderstanding, that in the work cited, as well as in a number of other works of different authors, completely regular semigroups were called *Clifford* (another variant is *Cliffordian*) *semigroups*. This good term had been proposed not later than the mid sixties (refer to the book [29]); being shorter, it seemed to be rather appropriate for the notion which became the subject of a rich theory. However, later many authors began to use the term "Clifford semigroup" only for the particular case of inverse completely regular semigroups, i.e., semilattices of groups. Somewhat more detailed terminological remarks on this matter are given in [108, p. 23].

$x \in F$ under a fixed bijection $\psi : F \to H \setminus I$ will be denoted by x'. We define an operation $\circ$ on the union $S = T \cup F$ as follows:

$$x \circ y = \begin{cases} xy & \text{if } x, y \in T, \\ xy' & \text{if } x \in T,\ y \in F, \\ x'y & \text{if } x \in F,\ y \in T, \\ x'y' & \text{if } x, y \in F \text{ and } x'y' \in I, \\ \psi^{-1}(x'y') & \text{if } x, y \in F \text{ and } x'y' \in H \setminus I. \end{cases}$$

It is easy to verify that $\circ$ is associative; hence S becomes a semigroup that obviously contains T as an ideal, and the quotient semigroup S/T is isomorphic to the Rees factor H/I. Thus S is an ideal extension of T. Figuratively speaking, S is obtained from T by "attaching" a copy of subsemigroup H so that the copy of I is identified with I and no other identifications are made. We will call S an extension of T by the Rees factor H/I, or a *duplicate extension* with *attachment* H and *base* I. A duplicate extension of the semigroup T is uniquely defined up to isomorphism by the attachment H and base I; we will denote it by $D(T, H, I)$. It is easy to see that a duplicate extension is a retract extension, i.e., T is a retract of $D(T, H, I)$. In the degenerate case when $I = H$, we have simply $D(T, H, H) = T$.

The following classification was given in [104, §1].

2.3 Proposition *The infinite cyclic epigroups are exhausted by all possible duplicate extensions of the infinite cyclic group with attachment consisting of the positive powers of its generator. Two such epigroups are isomorphic if and only if they have the same index.*

The infinite cyclic epigroup of index n will be denoted by $C_{n,\infty}$. This notation is consistent with the notation $C_{n,m}$ for the finite cyclic semigroup of type (n, m). It is also natural to call (n, ∞) the *type* of the epigroup $C_{n,\infty}$. Thus the family of all cyclic epigroups is $\{C_{n,\pi}\}$, where n runs over the set $\mathbb{N}$, and π runs over the set $\mathbb{N} \cup \{\infty\}$. The epigroup $C_{1,\pi}$ is nothing but the cyclic group of order π. A homomorphic image of the epigroup $C_{n,\infty}$ is, of course, a cyclic epigroup of index at most n; conversely, it follows easily from the description given by Proposition 2.3 that any epigroup $C_{m,\pi}$ with $m \le n$ is a homomorphic image of $C_{n,\infty}$.

The infinite cyclic epigroup $C_{n,\infty} = \langle\langle a \rangle\rangle$ is generated as a semigroup by two elements: a and $b = \overline{a}$. It can be defined in the class of all semigroups by the presentation

$$C_{n,\infty} = \langle a, b \,|\, ab = ba,\ ab^2 = b,\ a^{n+1}b = a^n \rangle.$$

The semigroup $C_{n,\infty}$ is an ideal extension of the infinite cyclic group by a cyclic nilpotent semigroup of order n. The converse is obviously false, but the extensions of the indicated type admit an exhaustive classification: if we ignore the degenerate case $n = 1$, they form a two-parameter family $\{D_{n,k}\}$, $n > 1$, $k \ge 0$, where the semigroup $D_{n,k}$ is defined for $k > 0$ by the presentation

$$D_{n,k} = \langle a, b \,|\, ab = ba,\ ab^{k+1} = b,\ a^{n+1}b^k = a^n \rangle,$$

and for $k = 0$ by the presentation

$$D_{n,0} = \langle a, b, c \,|\, ab = ba = b,\ ac = ca = c,\ bc = cb = a^n \rangle.$$

The semigroup $D_{n,k}$ is embeddable in a cyclic epigroup (and then, for example, in $C_{nk,\infty}$) if and only if $k > 0$, and this is the case if and only if the idempotent in $D_{n,k}$ is a unique element of finite order. When $k > 0$, $D_{n,k}$ is a duplicate extension of the infinite cyclic group $\langle\langle b\rangle\rangle$ with attachment $\langle b^{-k}\rangle$; when $k = 1$, we have simply $D_{n,1} = C_{n,\infty}$.

As to epigroups having a two-element epigroup generating set, it makes no sense to find their classification, because, as in the case of groups and semigroups, the following is true.

2.4 Theorem *Any countable epigroup is embeddable in a 2-generated epigroup.*

Similarly to known results concerning arbitrary countable semigroups [44, 60] and finite semigroups [32], the following is true as well.

2.5 Theorem *Any countable epigroup is embeddable in an epigroup generated by three idempotents.*

Theorems 2.4 and 2.5 were announced in [70]; their proofs are still unpublished (I must note with sorrow that the first author of [70] died several years ago). Analogous results (with a number of refinements concerning periods or indices of elements) have been obtained earlier in [69] for periodic semigroups. Note that the technique developed in [69] made it possible in passing to reprove all the mentioned results of [32, 44, 60] as well as the results of the pioneer works [19, 25, 53] relating to embeddability of countable or finite semigroups in 2-generated ones.

In some considerations concerning epigroups, it is useful to apply the following fact.

2.6 Lemma *The subsemigroup $\langle E_S\rangle$ of an epigroup S is in fact a subepigroup.*

This statement was apparently first proved (among some other facts about epigroups) in the unpublished work [115]. In published form it appeared in particular in [16, Corollary 1], and in [63, Theorem 2.2]; its very short proof uses a trick from [21] showing that the pseudo-inverse of a product of k idempotents can be expressed as a product of $k+1$ idempotents (see also [27, Theorem 1.4.18]). The main line of examination in [63] deals with the set E_S of an epigroup $S \in \mathcal{E}_n$, and E_S is treated there as a special type of a universal algebra endowed with operations t_i, $i \geq 2$, where t_i is the i-ary operation defined by the rule $t_i(e_1, e_2, \ldots, e_i) = (e_1 e_2 \cdots e_i)^\omega$. This algebra is called the *idempotent algebra* of an epigroup S, and some properties of such algebras are examined in [63]. Note in this connection that there is another type of (partial) algebraic systems which can be defined on the set of all idempotents of a semigroup, so-called *biordered sets*. Such biordered sets were characterized abstractly in [52]. In [16] the corresponding characterizations were given for epigroups properly, for periodic semigroups, for finite semigroups, and, first of all, for eventually regular semigroups. Recall that a semigroup S is said to be an *eventually regular* [17] if $S = \sqrt{\operatorname{Reg} S}$. Any epigroup is clearly eventually regular. Some conditions under which an eventually regular semigroup turns out to be an epigroup or even a periodic semigroup are given in [18]. In particular, an eventually regular semigroup S with finitely many regular $\mathcal{D}$-classes is an epigroup if (and obviously only if) S has no bicyclic subsemigroup.

Let us call a semigroup *bi-unary* if it is further endowed with two unary operations. We may view inverse epigroups as bi-unary semigroups, having in mind the operations of pseudo-inversion $x \mapsto \overline{x}$ and taking the inverse element $x \mapsto x^{-1}$. There arises a number of natural questions pertaining to the study of such algebraic systems. Leaving them out of this

article, we mention only that Corollary 3.54 below presents a "degenerate" case of bi-unary inverse epigroups—where the two signature unary operations coincide. In the general case, the connection between these operations is provided by the following fact observed in [104, §7].

2.7 Proposition *In any inverse epigroup, the operations of pseudo-inversion and taking the inverse element are permutable, i.e., the identity* $(\overline{x})^{-1} = \overline{x^{-1}}$ *holds.*

2.2 Identities of epigroups

In speaking of identities of epigroups, we have in mind the signature of unary semigroups, i.e., for an identity $u = v$, in the terms u and v both multiplication and pseudo-inversion may appear (as well as the operation $a \mapsto a^\omega$ derivative from them: $a^\omega = a\overline{a}$). One can observe the simpliest identities holding in any epigroup, which follow directly from the corresponding equalities in (2.2):

$$x\overline{x} = \overline{x}x, \qquad x(\overline{x})^2 = \overline{x}.$$

We mention further simple consequences from the definitions:

$$x^2\overline{x} = \overline{\overline{x}}, \qquad \overline{\overline{\overline{x}}} = \overline{x}, \qquad x\overline{\overline{x}} = (\overline{\overline{x}})^2.$$

Here are less trivial examples of the identities that hold in any epigroup:

$$\overline{x^\omega y^\omega x^\omega} = \overline{x^\omega y^\omega} \cdot x^\omega = x^\omega \cdot \overline{y^\omega x^\omega}, \qquad \overline{x^\omega y^\omega} \cdot \overline{y^\omega x^\omega} = (\overline{x^\omega y^\omega x^\omega})^3.$$

Can one describe all the identities which hold in any epigroup? We will denote the class of all epigroups by $\mathcal{E}$. For a class $\mathcal{X}$ of algebraic systems of one type, the *equational theory* $\operatorname{eq}\mathcal{X}$ of this class is the set of all the identities which hold in any system from $\mathcal{X}$. A *basis* of $\operatorname{eq}\mathcal{X}$, or a *basis of identities* of $\mathcal{X}$, is a subset of $\operatorname{eq}\mathcal{X}$ such that every identity from $\operatorname{eq}\mathcal{X}$ can be derived from identities of this subset. The question above may be reformulated, in particular, as the problem of finding some basis of $\operatorname{eq}\mathcal{E}$. Another way to answer it is to establish whether the theory $\operatorname{eq}\mathcal{E}$ is decidable.

The same problem is interesting also for the class $\mathcal{E}_{\text{fin}}$ of all finite semigroups treated as unary semigroups with pseudo-inversion. Observe that $\mathcal{E}_{\text{fin}}$ is a pseudovariety, i.e., it is closed under taking subsystems, homomorphic images and finite direct products. Note in this connection that the majority of important pseudovarieties of finite semigroups can be specified in the class of all finite semigroups just by epigroup identities; one may see many confirmations of this thesis in, for example, [1].

An answer to the questions discussed in the previous paragraph is given by the following theorem.

2.8 Theorem *The equational theories of the classes* $\mathcal{E}$ *and* $\mathcal{E}_{\text{fin}}$ *coincide and have a basis consisting of the identities* $x(yz) = (xy)z$, $x(\overline{yx}) = (\overline{xy})x$, $x(\overline{x})^2 = \overline{x}$, $x^2\overline{x} = \overline{\overline{x}}$, $\overline{x\overline{x}} = x\overline{x}$, $\overline{x^p} = (\overline{x})^p$, *where* p *runs over the set of all primes. The theory* $\operatorname{eq}\mathcal{E} = \operatorname{eq}\mathcal{E}_{\text{fin}}$ *is decidable.*

Along with the classes $\mathcal{E}$ and $\mathcal{E}_{\text{fin}}$, among important classes of epigroups are classes consisting of epigroups S such that $\operatorname{Gr} S = E_S$, i.e., all the maximal subgroups of S are one-element. Following widely-used terminology, we call the epigroups with this property *combinatorial*. Remark that, as applied to finite semigroups, it is common to call such semigroups *aperiodic*.

I shall not discuss to what extent this term is apt, but in the general case it is definitely unacceptable; indeed, combinatorial epigroups are obviously periodic, so, if they were called aperiodic, we would have a collision between terminology and logic. However, taking into account the traditional notation used for the class of finite semigroups with the property under consideration, we will denote the class of all combinatorial epigroups by $\mathcal{A}$ and the class of all finite combinatorial semigroups by $\mathcal{A}_{\rm fin}$.

2.9 Theorem *The equational theories of the classes $\mathcal{A}$ and $\mathcal{A}_{\rm fin}$ coincide and have a basis consisting of the identities $x(yz) = (xy)z$, $x(\overline{yx}) = (\overline{xy})x$, $x\overline{x} = \overline{x}$, $\overline{\overline{x}} = \overline{x}$, $\overline{x^p} = \overline{x}$, where p runs over the set of all primes. The theory $\operatorname{eq}\mathcal{A} = \operatorname{eq}\mathcal{A}_{\rm fin}$ is decidable.*

Theorems 2.8 and 2.9 were given in [129] together with a description of the main ideas of their proofs. (Unfortunately, a full presentation of this material by the author is impossible because of his tragic death in Black Sea in 2000). Remark that the list of identities which appeared in the original formulation of Theorem 2.9 in [129] contains an extra identity $\overline{x}\cdot\overline{x} = \overline{x}$ which can be derived from the others; indeed, from the third identity it follows that $\overline{x}\cdot\overline{\overline{x}} = \overline{\overline{x}}$, and then, in view of the fourth identity, $\overline{x}\cdot\overline{x} = \overline{x}\cdot\overline{\overline{x}} = \overline{\overline{x}} = \overline{x}$.

Note in the conclusion of this subsection that many important types of epigroups can be characterized in terms of identities; see various examples in Section 3.

2.3 On varieties of epigroups

The concept of an epigroup identity leads naturally to the concept of a variety of epigroups treated as unary semigroups. For a class of algebraic systems $\mathcal{X}$ of one type, we denote by $L(\mathcal{X})$ the lattice of all varieties contained in $\mathcal{X}$. The class $\mathcal{E}$ of all epigroups is not a variety (in other words, $L(\mathcal{E})$ has no greatest element) because it is not closed under direct products (of infinite families). Indeed, if we take the family of cyclic epigroups $\langle\langle a_i\rangle\rangle$ whose indices are unbounded, then the direct product $\prod_i\langle\langle a_i\rangle\rangle$ is not an epigroup since, for example, no power of the element $(\dots, a_i, \dots)$ belongs to a subgroup.

We denote by $\mathcal{E}_n$ the class of all epigroups of index at most n. The equalities (2.2), along with the remark that the smallest n with the property indicated in the third of the equalities is $\operatorname{ind}(a)$, establish

2.10 Proposition *The class $\mathcal{E}_n$ is a variety of unary semigroups. It is defined by the identities*

$$(xy)z = x(yz), \qquad x\overline{x} = \overline{x}x, \qquad x\overline{x}^2 = \overline{x}, \qquad x^{n+1}\overline{x} = x^n.$$

In the lattice $L(\mathcal{E})$, there is a naturally distinguished chain

$$\mathcal{E}_1 \subset \mathcal{E}_2 \subset \cdots \subset \mathcal{E}_n \subset \cdots$$

which, figuratively speaking, can be regarded as the "spine" of this lattice. The reason is the following property.

2.11 Observation *For any variety $\mathcal{V}$ of epigroups, there exists n such that $\mathcal{V} \subseteq \mathcal{E}_n$.*

Indeed, a class of epigroups that is not contained in any of the varieties $\mathcal{E}_n$ obviously contains cyclic epigroups whose indices are unbounded, so one may refer to the argument given at the end of the first paragraph of this subsection.

In view of Observation 2.11, to any variety $\mathcal{V}$ of epigroups there corresponds a smallest n such that $\mathcal{V} \subseteq \mathcal{E}_n$. This number, $\operatorname{ind}\mathcal{V}$, which is equal to the maximal index of the epigroups in $\mathcal{V}$ is naturally called the *index* of the variety $\mathcal{V}$. (We should mention here that the term "variety of finite index" has already been used in the theory of semigroups in a different sense, namely for varieties in which the nilpotency degrees of the nilpotent semigroups are uniformly bounded. The corresponding concept can be appropriately extended to varieties of epigroups and, to avoid a collision of terminology, we will call a variety with the latter property a *variety of finite degree*. A certain characterization of such varieties is given below in Corollary 3.20 of Proposition 3.18.)

The variety $\mathcal{E}_1$ is nothing but the variety of all completely regular semigroups (treated as unary semigroups). We will call any variety of such semigroups a *completely regular* (*c.r.*) *variety*; so $\mathcal{E}_1$ is the largest c.r. variety. Note that c.r. varieties are examined in several chapters of [68]. Along with c.r. varieties, among the examined varieties of epigroups are all *periodic varieties*, i.e., varieties consisting of periodic semigroups (a lot of such varieties appeared as a matter of fact in diverse investigations devoted to semigroup varieties). Indeed, in any periodic variety, an identity of the form $x^n = x^{n+m}$ holds, and the pseudo-inversion operation in this case can be expressed by a formula in the semigroup signature: for instance, $\overline{x} = x^{nm-1}$. Conversely, any class of epigroups that is a semigroup variety obviously consists of periodic semigroups. Since any non-periodic epigroup contains among its subepigroups the infinite cyclic group, and the latter generates the variety of all abelian groups, this variety is the smallest non-periodic variety of epigroups.

Any variety of epigroups that contains $\mathcal{E}_1$ will be called an *over c.r. variety*. It is clear that in studying varieties of epigroups we encounter three possible types of situations: (1) we generalize facts previously established for c.r. or periodic varieties; (2) we find properties of varieties having significant new consequences for c.r. and/or periodic varieties; (3) we consider specific properties of over c.r. varieties that are not manifested (or are trivially manifested) in c.r. and periodic varieties. Among the epigroup varieties that we consider in Section 3 are both over c.r. varieties and proper subvarieties of $\mathcal{E}_1$, and we encounter the first two of the situations just mentioned.

As for possible investigation of epigroup varieties in general, one can, of course, proceed in several directions, and it would be natural to pay attention to all of the main aspects: equational, structural, algorithmic, and so on; see, for example, the introductions in the survey articles [109,110]. Not attempting to list in detail the concrete problems and questions that arise, we limit ourselves to a few remarks.

Turning to the lattice $L(\mathcal{E})$, we can easily see that it has the same atoms as the lattice of semigroup varieties, and for any n the upper cone $\mathcal{E}_n^{\Delta} = \{\mathcal{X} \in L(\mathcal{E}) \mid \mathcal{X} \supseteq \mathcal{E}_n\}$ has a single atom, namely $\mathcal{E}_n \vee \operatorname{var} C_{n+1,1}$, which is contained in any variety in $\mathcal{E}_n^{\Delta} \setminus \{\mathcal{E}_n\}$. One of the standard properties examined for lattices of varieties (or lattices of related classes such as, for instance, pseudovarieties) is the so-called the *cover property*. By definition, a lattice L has this property if any element $x \in L$ which is not the greatest element of L has a *cover* in L, i.e., an element y such that $x < y$ and there is no element of L lying strictly between x and y. A brief survey of the main results and problems connected with the cover property in the lattice of semigroup varieties is given in [123] whose central result exhibits an example of a semigroup pseudovariety which has no cover in the lattice of all semigroup pseudovarieties. A part of the paper cited is devoted to epigroup varieties. Namely, the following fact has been established there.

2.12 Proposition *The lattice $L(\mathcal{E})$ has the cover property.*

A more significant result of [123] in this direction is concerned with a problem about the cover property for the lattices $L(\mathcal{E}_n)$; this problem was noted in [104]. To formulate the result mentioned, denote by $\mathcal{EA}_n$ the class of all epigroups from $\mathcal{E}_n$ whose idempotent generated subepigroups (see Lemma 2.6) are combinatorial. It is observed in [123] that $\mathcal{EA}_n$ is a variety of epigroups.

2.13 Theorem *For any $n > 1$ the variety $\mathcal{EA}_n$ has no cover in the lattice $L(\mathcal{E}_n)$.*

As for the case $n = 1$, the corresponding problem was first posed by the present writer more that 25 years ago (see [91, Problem 2.62b]) and still remains open; we repeat it here.

2.14 Problem *Does the lattice $L(\mathcal{E}_1)$ of all completely regular varieties have the cover property?*

Another open problem pertaining to the lattice of subvarieties of a variety of epigroups will be formulated in Subsection 3.3.

Among known results related to the lattice $L(\mathcal{E})$, one may mention also the main result of [3] which indicates certain intervals in $L(\mathcal{E})$ such that no variety within any interval of the type considered has a finite basis for its identities. The precise formulation of this statement would require a number of additional definitions and special notations, so we omit it here.

In Section 3 we give, in particular, some characterizations of varieties consisting of epigroups which have certain "nice" decompositions.

There are many questions about epigroup varieties which are waiting to be examined. Quite a number of such questions are briefly discussed at the end of §1 in [104], so we will mention only part of them here. In all likelihood, none of the lattices $L(\mathcal{E}_n)$ has coatoms, i.e. elements covered by the greatest element; it would be interesting to verify this conjecture. The intervals $[\mathcal{E}_n, \mathcal{E}_{n+1}]$ of the lattice $L(\mathcal{E})$ deserve special attention. What are their lattice properties (for example, the order types of maximal chains and the cardinalities of antichains)? What are the interactions between these intervals (for example, when $m < n$, is $[\mathcal{E}_m, \mathcal{E}_{m+1}]$ embeddable into $[\mathcal{E}_n, \mathcal{E}_{n+1}]$, and is it true that $[\mathcal{E}_m, \mathcal{E}_{m+1}]$ and $[\mathcal{E}_n, \mathcal{E}_{n+1}]$ are not isomorphic)?

We denote by $Ni\ell$ the class of all nilsemigroups, and we put $Ni\ell_n = Ni\ell \cap \mathcal{E}_n$. With each variety $\mathcal{V}$ of epigroups we can associate its "nil part" $\mathcal{V} \cap Ni\ell$, which is a subvariety of the variety $Ni\ell_{\mathrm{ind}\,\mathcal{V}}$, and its "completely regular part" $\mathcal{V} \cap \mathcal{E}_1$. It would be interesting to examine different aspects of the interrelation between a variety and its nil and completely regular parts in certain situations; some specifications of this idea are proposed in [104, §1], and the interested reader is referred to the work mentioned[2].

[2]It should be noted that the translator of the original Russian version of this work into English was not so accurate and made many mistakes in choosing appropriate words in the translation. By the way, it began with the title of the work: the original "To the theory of epigroups" has turned into "On the theory of epigroups". There are some blunders; for example, "given" was translated as "long" (Russian equivalents of these words, written in Roman letters, are "dannyj" and "dlinnyj", respectively, so only a very thoughtless translator could mess up the words and in doing so write an expression making no sense). There are several terminological slips; for example, "complete preimage" is written instead of the correct "inverse image" as well as the prefix "super" instead of the correct "over" in the situation considered. There is plainly carelessness as well; for example, "presents" is written instead of the correct "presence", and so on, and the like.

I will allow myself to give two more similar examples taken from English versions of the surveys [109] and

Let us mention, finally, some questions that can be posed about free epigroups in various varieties, including $\mathcal{E}_n$. Among such questions is that of the basis rank of $\mathcal{E}_n$ for $n > 1$ (is it equal to 2?) and the related question on the embeddability of the $\mathcal{E}_n$-free epigroup of countable rank in an $\mathcal{E}_n$-free epigroup of finite rank. (For $n = 1$ the answers are known: it is easy to show that the free completely regular semigroup of countable rank is not embeddable in any finitely generated completely regular semigroup; also, as is shown in [64], the basis rank of $\mathcal{E}_1$ is infinite.) It is natural to ask about the decidability of the word problem in $\mathcal{E}_n$-free epigroups (for $n = 1$ the decidability was established in [24,36,117]), about a description, for $n > 1$, of Green's relations and, in particular, the maximal subgroups, and about a description of the partially ordered set of idempotents. By the way, one may regard the $\mathcal{E}_n$-free epigroups as looking to some extent like the free Burnside semigroups, i.e., the free semigroups of the variety defined by the identity $x^n = x^{n+m}$ for some m. Both certain structural properties of such semigroups and the corresponding algorithmic aspects were considered in [43], and perhaps some ideas of that work could be applied to the examination of the $\mathcal{E}_n$-free epigroups.

3 Certain decompositions

3.0 Introductory remarks

A common trend in a structural theory of the algebraic systems under examination is to find conditions under which these systems can be constructed in some way from more specific ones, in particular, those having a more simple, or more rigid, or more clear structure. In our case, the role of such "building blocks" will be played by completely simple semigroups (including certain particular types of them, up to groups) as well as nilsemigroups (in particular, nilpotent semigroups). As to the ways of "assemblage", we will deal mostly with ideal extensions and decompositions into bands (including certain particular types of such constructions, for instance, the semilattice decompositions).

Recall that a semigroup is said to be a *band of subsemigroups* S_α, $\alpha \in A$, if S_α form a partition of S, and for any $\alpha, \beta \in A$, there exists $\gamma \in A$ such that $S_\alpha S_\beta \subseteq S_\gamma$. This means, in other words, that these subsemigroups (the *components* of the band under consideration) are the classes of some congruence, ρ, say, on S, and the quotient semigroup S/ρ is a semigroup of idempotents. If S/ρ is commutative, and hence is a semilattice, then S is said to be a *semilattice of semigroups* S_α; if the semilattice S/ρ is a chain, we say that S is a *chain of the corresponding semigroups*. If S/ρ is rectangular, i.e., it satisfies the identity $xyx = x$, then S is a *rectangular band of the corresponding semigroups*. A particular case of a rectangular band is given by a *left* [*right*] *bands of semigroups*; in this case, for any two components S_α, S_β of a given band, the inclusion $S_\alpha S_\beta \subseteq S_\alpha$ [$S_\alpha S_\beta \subseteq S_\beta$] holds, that is all the components are right [left] ideals. A decomposition of a semigroup into a semilattice of some subsemigroups is called a *semilattice decomposition*. Similarly, we obtain the notion of a *rectangular*, *left* and *right decomposition* as well. A rectangular decomposition is also called a *matrix decomposition*.

[110]. In the former, one of the egregious errors uses "arbitrary" instead of "derivative" (Russian equivalents are "proizvol'nyj" and "proizvodnyj", respectively, so there is no need to comment what were tranlator's attention and understanding). In the latter, "manifold" is used instead of "variety" (Russian equivalents of these words are the same, namely, "mnogoobrazie", so the translator of [110] was not aware of the subject of this survey). The reader of these translations should be vigilant in order to be able to detect probable incorrectness not due to the authors but introduced by the translator.

Both of these terms are justified by the fact that the components of a matrix decomposition can be equipped with double indices running over some sets I and Λ such that the rule

$$S_{i\lambda}S_{j\mu} \subseteq S_{i\mu}, \quad i, j \in I,\ \lambda, \mu \in \Lambda,$$

is fulfilled. If all the components of a given band belong to a certain class of semigroups $\mathcal{X}$, then we say that the corresponding semigroup is a *band* (or *decomposable into a band*) *of semigroups from* $\mathcal{X}$. For instance, we may consider bands of groups, bands of commutative semigroups, and the like. Recall in this connection that completely simple semigroups are precisely rectangular bands of groups. A semigroup is a *left* [*right*] *group* if it is a left [right] band of groups; as is known, there are several equivalent properties characterizing these semigroups, see, for instance, [49, Section A5].

Different decompositions into bands play an important role in the structural theory of semigroups. Here it is particularly worthwhile to distinguish semilattice and matrix decompositions, not only because of their clearness but also due to the following fact: any band of semigroups forming a family $\{S_\alpha\}$ is a semilattice of rectangular bands of semigroups from $\{S_\alpha\}$; that is, the components S_α can be grouped in subfamilies such that the union of the components belonging to each subfamily is a rectangular band of these components, while the initial semigroup is a semilattice of the unions mentioned. Notice that various questions relating to semilattice and matrix decompositions of semigroups are extensively examined in [65] and [66]; semilattice decompositions were studied in detail in [74], and the note [73] is devoted to a certain description of the greatest semilattice decomposition of an epigroup. Notice also that the topic "Bands of semigroups" is surveyed with due fullness in [102, §2.3] and, somewhat more briefly, in [49, Section A2]. A comprehensive survey of investigations relating to various kinds of decompositions is given in [11].

One of the distinctive features of the approach carried out in Subsections 3.2–3.5 is the primary interest in so-called *indicator characterizations*, i.e., characterizations in terms of "forbidden" objects; the latter ones in our case are mainly epidivisors. In these divisor characterizations, there appear mostly finite semigroups having an extremely clear structure. This makes such characterizations very transparent and enables us to obtain, in most cases as direct corollaries, many other criteria for the decomposability of epigroups (or, earlier, of periodic semigroups). Moreover, divisor criteria immediately yield indicator characterizations of the varieties consisting of the corresponding epigroups. These criteria were subsequently applied to the theory of varieties as well; see, for example, [79–81, 114, 127].

3.1 Archimedean epigroups

A semigroup S is called *archimedean* [*left archimedean, right archimedean*] if for any $a, b \in S$ there exists an n such that $b \mid a^n$ [resp., $a^n \in Sb$, $a^n \in bS$]. A semigroup is *bi-archimedean* if it is left and right archimedean. A semigroup is *unipotent* if it has a unique idempotent. An ideal extension of an arbitrary semigroup by a nilsemigroup [nilpotent semigroup] is called a *nil-extension* [*nilpotent extension*]. Extending to epigroups the corresponding semigroup terminology (see, for example, [110]), we will call an epigroup identity $u = v$ *heterotypical* if one of the words u, v contains a letter which does not occur in the other.

3.1 Proposition *The following conditions on a semigroup S are equivalent:*

(1) *S is an archimedean epigroup;*

(2) *S is an epigroup in which E_S is an antichain;*

(3) *S is an nil-extension of a completely simple semigroup;*

(4) *S is an epigroup satisfying some heterotypical identity;*

(5) *S is an epigroup with the identity $(x^\omega y^\omega x^\omega)^\omega = x^\omega$;*

(6) *S is an epigroup with the identity $(x^\omega y x^\omega)^\omega = x^\omega$.*

3.2 Corollary *The following conditions for a variety $\mathcal{V}$ of epigroups are equivalent:*

(1) *$\mathcal{V}$ consists of archimedean semigroups;*

(2) *one (hence each) of the identities appearing in Proposition* 3.1 *is satisfied in $\mathcal{V}$;*

(3a) *$\mathcal{V}$ contains no two-element semilattice;*

(3b) *each regular epigroup in $\mathcal{V}$ is completely simple.*

The equivalence of conditions (4)–(6) in Proposition 3.1 is an epigroup modification of a long-known property of periodic semigroups (see [10]). The equivalence of conditions (1) and (3) was noted in [74]. Condition (3) leads to the following consequence.

3.3 Observation *In an archimedean epigroup S, the set* $\operatorname{Gr} S$ *is the kernel of S, the Green's relation $\mathcal{D}$ coincides with the Rees congruence corresponding to the ideal* $\operatorname{Gr} S$*, and the relations $\mathcal{L}$ and $\mathcal{R}$ (hence also $\mathcal{H}$) are congruences.*

The one-sided version of Proposition 3.1 is given by the following.

3.4 Proposition *The following conditions on a semigroup S are equivalent:*

(1) *S is a right [left] archimedean epigroup;*

(2) *S is an epigroup in which E_S is a right [left] zero semigroup;*

(3) *S is an nil-extension of a right [left] group;*

(4) *S is an epigroup satisfying the identity $x^\omega y^\omega = y^\omega$ $[x^\omega y^\omega = x^\omega]$.*

Lemma 1.5, together with some additional elementary arguments, implies the following statement distinguishing a specific subclass of the class of archimedean epigroups.

3.5 Proposition *An epigroup S is a subdirect product of a completely regular semigroup and a nilsemigroup if and only if* $\operatorname{Gr} S$ *is a retract ideal.*

The theme of the retractability of the group part of an epigroup will again be heard at the end of Subsection 3.3 and in Subsection 3.4. In particular, Theorem 3.48 gives a characterization of epigroups appearing in Proposition 3.5 in terms of "forbidden epidivisors".

Two extreme types of unipotent epigroups are groups and nilsemigroups. It turns out that an arbitrary unipotent epigroup can be built of these polar types. This is seen in Proposition 3.6 obtained plainly by combining both the left and the right versions of Proposition 3.4, taking into account Proposition 3.5.

3.6 Proposition *The following conditions on a semigroup S are equivalent:*

(1) *S is a unipotent epigroup;*

(2) *S is a bi-archimedean epigroup;*

(3) *S is a nil-extension of a group;*

(4) *S is a subdirect product of a group and a nilsemigroup;*

(5) *S is an epigroup with the identity $x^\omega = y^\omega$.*

3.2 Conditions under which unipotency classes are subsemigroups

One of the main "characters" in this section is the five-element Brandt semigroup B_2. It is well known as a matrix semigroup

$$\left\{ \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \right\}.$$

Being a completely 0-simple inverse semigroup, it has a clear representation as the Rees matrix semigroup $\mathcal{M}^0\,[\{e\}; 2, 2; \left(\begin{smallmatrix} e & 0 \\ 0 & e \end{smallmatrix}\right)]$ over the 0-group $\{e, 0\}$. When examining structural properties of semigroups, it is convenient to define them by a presentation, also well known:

$$B_2 = \langle c, d \,|\, c^2 = d^2 = 0,\ cdc = c,\ dcd = d \rangle.$$

(To reduce the number of defining relations, we at once employ a zero element in them.) We will not make a distinction between the terms "the semigroup B_2" and "a semigroup isomorphic to B_2"; an analogous convention will be applied to other concrete semigroups that appear below. An ideal extension of an arbitrary semigroup by B_2 will be called, for brevity, a *B_2-extension*.

The semigroup B_2 is frequently encountered in various semigroup-theoretic investigations. In particular, it plays a special role in our considerations. First of all, it provides an example of a semigroup with the smallest number of elements that has a unipotency class (namely, K_0) that is not a subsemigroup. But it turns out (and it is rather surprising) that its specific presence is a characteristic feature for any epigroup with this property. Namely, the following statement (proved in [104, §2]) is true.

3.7 Proposition *A unipotency class K_e in an epigroup is not a subsemigroup if and only if the subepigroup $\langle\langle K_e \rangle\rangle$ contains a subsemigroup that is a B_2-extension of a unipotent epigroup.*

Any epigroup is a union of its unipotent (for instance, cyclic) subepigroups. If it has a partition into unipotent subepigroups, this means that such a partition is unique and its components are precisely the unipotency classes; in this case an epigroup will be called *unipotently partitionable*. In the following theorem which provides a criterion for an epigroup to be unipotently partitionable, the "if" part is delivered by Proposition 3.7, and the "only if" part follows directly from the fact that a B_2-extension of a unipotent epigroup with idempotent e contains two elements of K_e whose product is an idempotent different from e.

3.8 Theorem *An epigroup is unipotently partitionable if and only if it contains no subsemigroup that is a B_2-extension of a unipotent epigroup.*

The triviality of verifying the properties of B_2 connected with the behavior of its elements enables us to obtain as direct corollaries of this criterion (actually, of Proposition 3.7) all conditions found before (pertaining mainly to periodic semigroups) for a unipotency class to be a subsemigroup, and, in particular, conditions for an epigroup to be unipotently partitionable. We limit ourselves here to only three corollaries. The assertion of the first one was given in [71] (and in a weaker form in [128]). The assertion of the second corollary in the case of periodic semigroups was also mentioned in [71]. The third corollary, in view of condition (2) of Proposition 3.1, follows automatically from the second; it was proved (as one of the main results) in [22], and the corresponding result for periodic semigroups was given in [84].

3.9 Corollary *A torsion class K of a periodic semigroup is a subsemigroup if and only if for any $a, b \in K$ there exist natural numbers p, q such that $b^p a^q \in \langle ab \rangle$.*

3.10 Corollary *If an idempotent e of an epigroup S is such that the set of all upper bounds for e in E_S is a chain (in particular, if e is a maximal idempotent in E_S), then K_e is a subsemigroup.*

3.11 Corollary *Any archimedean epigroup is unipotently partitionable.*

The latter corollary significantly increases the supply of unipotently partitionable epigroups: among such epigroups are all the types of epigroups considered in Subsections 3.3 and 3.4. As for the varieties consisting of unipotently partitionable epigroups, see Proposition 3.18 below. It follows from Corollary 3.10 that in any epigroup with a finite number of idempotents there are unipotency classes that are subsemigroups. On the other hand, there exist epigroups that can be called antipodes of unipotently partitionable epigroups; in such an epigroup none of the unipotency classes is a subsemigroup. An example of a periodic semigroup with this property is given in [72].

If a unipotency class K_e is not a subsemigroup, then there is nothing we can say about the structure of the semigroup $\langle K_e \rangle$ (besides the properties indicated in the Corollary of Lemma 1.3). The point is that any epigroup can be embedded in an epigroup generated as a semigroup by a single unipotency class. More precisely, the following fact is true (see [104, §7]).

3.12 Proposition *Any semigroup S can be embedded in a semigroup T such that:*

(a) *T has a zero and it is generated by nil-elements of order 2;*

(b) *if S is an epigroup [periodic semigroup], then T is also an epigroup [periodic semigroup];*

(c) *if, in addition, S has finite index n, then T also has finite index, and $\operatorname{ind} T$ is equal to n or $n+1$.*

Let us return to Theorem 3.8. We now call attention to the fact that the existence among the subsemigroups of a given semigroup S of a B_2-extension of a unipotent epigroup implies the presence of B_2 among the (Rees) epidivisors of S. Indeed, it is obvious that an ideal extension of an epigroup by an epigroup is an epigroup, so a B_2-extension of a unipotent epigroup is an epigroup. The answer to the question of whether, in general, the absence of B_2 among the epidivisors is a characteristic feature of unipotently partitionable

epigroups is negative. Moreover, unipotently partitionable epigroups cannot be characterized in any way in terms of forbidden epidivisors or Rees epidivisors. The fact is that the class of such epigroups (even the finite ones) is not closed under Rees quotient semigroups. This is demonstrated by the following example.

3.13 Example Let $S = \langle c, d \,|\, cdc = c,\ dcd = d,\ c^2d = cd^2 = c^2,\ d^2c = dc^2 = d^2 \rangle$. It is easy to see that S has six elements and four torsion classes, all of which are subsemigroups: $\{c, c^2\}$, $\{d, d^2\}$, $\{cd\}$, and $\{dc\}$. The subset $I = \{c^2, d^2\}$ is an ideal of S, and $S/I \cong B_2$.

The largest homomorphically closed subclass of the class of unipotently partitionable epigroups, and only this subclass, has the divisor characterization just mentioned; it can be also transparently characterized from slightly different natural points of view. This is done in the following theorem proved in [104, §2]; the equivalence of conditions (1), (2a)–(2c), (3a)–(3c) was already stated in [96].

3.14 Theorem *The following conditions on an epigroup S are equivalent:*

(1) *all homomorphic images of S are unipotently partitionable;*

(2a) *in any homomorphic image S' with zero, the set* $\operatorname{Nil} S'$ *is a subsemigroup;*

(2b) *for any ideal I of S, the set $\sqrt{I}$ is a subsemigroup;*

(2c) *in any Rees epidivisor H' of S, the set* $\operatorname{Nil} H'$ *is a subsemigroup;*

(3a) *there is no semigroup B_2 among the epidivisors of S;*

(3b) *there is no semigroup B_2 among the Rees epidivisors of S;*

(3c) *there is no subsemigroup B_2 in any homomorphic image of S;*

(3d) *for any ideal I of S, there are no subsemigroup B_2 in S/I;*

(4) *for any subepigroup H of S,* $\operatorname{Reg} H \subseteq \langle \operatorname{Gr} H \rangle$*;*

(5) *for any $e, f, i \in E_S$, $e \,|\, i$ and $f \,|\, i$ imply $ef \,|\, i$ or $fe \,|\, i$.*

The fact that just epidivisors but not arbitrary divisors appear in conditions (3a), (3b) is essential. For (3a) this is obvious: groups trivially satisfy the conditions of Theorem 3.14, but any semigroup is a divisor of a free group of suitable cardinality. The following example shows that also in condition (3b) epidivisors cannot be replaced by arbitrary divisors.

3.15 Example Let S be the Rees matrix semigroup $\mathcal{M}\left[G; 2, 2; \begin{pmatrix} e & g^{-1} \\ g^{-1} & e \end{pmatrix}\right]$, where $G = \langle\langle g \rangle\rangle$ is the infinite cyclic group with identity element e. Put $c = (1, g, 1)$, $d = (2, g, 2)$, $Q = \{(i, g^n, \lambda) \,|\, i, \lambda \in \{1, 2\},\ n > 1\}$. It can be shown by a direct calculation that Q is an ideal of the subsemigroup $\langle c, d \rangle$, $cdc = c$, $dcd = d$, $c^2, d^2 \in Q$, $\langle c, d \rangle \setminus Q = \{c, d, cd, dc\}$, and therefore $\langle c, d \rangle / Q \cong B_2$. Thus B_2 occurs among the Rees divisors of S. But S satisfies the conditions of Theorem 3.14, which can be seen at once, for instance, from condition (2b), since S, being a completely simple semigroup, contains no proper ideals.

3.3 Epigroups decomposable into a band of archimedean semigroups

The title of this subsection shows what will be the subject of our considerations here. In view of Observation 2.2, any congruence class on an arbitrary epigroup that is a subsemigroup is a subepigroup. Therefore, the components of any decomposition of an epigroup into a band are subepigroups, so from now on we will speak about bands of epigroups with various properties. Epigroups decomposable into a band of archimedean epigroups are characterized from different points of view by Theorem 3.16 which may be considered as the culmination of this section and serves as a base for the rest of its material. To facilitate comprehension, as in Theorem 3.14, we present a rather long list of equivalent conditions grouped together by type. The reader will detect certain parallels between conditions on this list and the corresponding conditions in Theorem 3.14 which, of course, was used to prove Theorem 3.16. The proof of this theorem is given in [104, §3]. The equivalence of conditions (1a) and (4a) was stated in [95] and later mentioned also in [23]; the equivalence of (1a), (2a)–(2c), and (3a)–(3c) was stated in [97]; the equivalence of (1b), (6a) was first established in [74]; the characterization given by the first of the identities in (8) was established in [82, §II.6]. We mention also that one can find in [23] several other conditions on a semigroup that are equivalent to the condition that it be decomposable into a semilattice of archimedean epigroups; these conditions were reproduced in [5, Chapter X], where the corresponding semigroups were called GV-semigroups. Various results on semigroups decomposable into a semilattice of archimedean semigroups are contained in the survey [8].

Notice that in the case of finite semigroups, the semigroups under discussion are rather popular in certain investigations, where they are commonly presented by condition (4c); see, for instance, [1, Chapter 8] which is devoted to the study of so-called implicit operations on such semigroups and, as is noted at the beginning of this chapter, contains some of the main results of the whole book.

Along with the semigroup B_2, among the "cast of characters" in Theorem 3.16 is the related 5-element semigroup A_2. It is also well known both as the matrix semigroup

$$\left\{\begin{pmatrix}0&0\\0&0\end{pmatrix},\begin{pmatrix}1&0\\0&0\end{pmatrix},\begin{pmatrix}0&1\\0&0\end{pmatrix},\begin{pmatrix}1&0\\1&0\end{pmatrix},\begin{pmatrix}0&1\\0&1\end{pmatrix}\right\}$$

and a completely 0-simple semigroup: it is isomorphic to the Rees matrix semigroup

$$\mathcal{M}^0\left[\{e\};2,2;\begin{pmatrix}e&e\\0&e\end{pmatrix}\right]$$

over the 0-group $\{e,0\}$. As B_2, it can clearly be given by a presentation:

$$A_2=\langle c,d\,|\,c^2=0,\ d^2=d,\ cdc=c,\ dcd=d\rangle.$$

If ρ is a binary relation on a semigroup S, we denote by $\sqrt{\rho}$ the binary relation on S defined by the following condition:

$$x\,\sqrt{\rho}\,y \iff \text{there exist } m,n \text{ such that } x^m\,\rho\,y^n.$$

3.16 Theorem *The following conditions on an epigroup S are equivalent:*

(1a) *S is a band of archimedean epigroups;*

(1b) *S is a semilattice of archimedean epigroups;*

(2a) *in any homomorphic image S' with zero, the set* $\operatorname{Nil} S'$ *is an ideal;*

(2b) *for any ideal I of S, the set $\sqrt{I}$ is an ideal;*

(2c) *in any Rees epidivisor H' of S, the set* $\operatorname{Nil} H'$ *is an ideal*;

(3a) *there are no semigroups A_2, B_2 among the epidivisors of S;*

(3b) *there are no semigroups A_2, B_2 among the Rees epidivisors of S;*

(3c) *there are no semigroups A_2, B_2 among the subsemigroups of any homomorphic image of S;*

(3d) *for any ideal I of S, there are no semigroups A_2, B_2 among the subsemigroups of S/I;*

(4a) $\operatorname{Reg} S = \operatorname{Gr} S$*;*

(4b) *for any subepigroup H of S,* $\operatorname{Reg} H = H \cap \operatorname{Reg} S$*;*

(4c) *each regular $\mathcal{D}$-class in S is a (automatically completely simple) subsemigroup;*

(4d) *each completely 0-simple principal factor of S is a completely regular semigroup, i.e., it contains no non-zero nil-elements;*

(5a) *for any $e, f, i \in E_S$, $i \mid e$ and $f \mid e$ imply $if \mid e$;*

(5b) *for any $a, b \in S$ and $e \in E_S$, $a \mid e$ and $b \mid e$ imply $ab \mid e$;*

(6a) *for any $a \in S$ and $e \in E_S$, $a \mid e$ implies $a^2 \mid e$;*

(6b) *for any $a \in S$, $J(a) \subseteq \sqrt{J(a^2)}$;*

(7a) *$\sqrt{\mathcal{D}}$ is a transitive relation (and is therefore an equivalence relation);*

(7b) *$\sqrt{\mathcal{D}}$ is the smallest semilattice congruence;*

(8) *S satisfies either (hence both) of the identities*
$$((xy)^\omega(yx)^\omega(xy)^\omega)^\omega = (xy)^\omega, \quad ((xy)^\omega yx(xy)^\omega)^\omega = (xy)^\omega.$$

We mention some direct consequences of Theorem 3.16 pertaining to varieties consisting of semilattices of archimedean epigroups. Condition (8) of Theorem 3.16 provides an equational characterization of such varieties. Denote by $\mathcal{AS}_n$ the class of all epigroups in $\mathcal{E}_n$ that are decomposable into a semilattice of archimedean epigroups.

3.17 Lemma *For any n the class $\mathcal{AS}_n$ is a subvariety of $\mathcal{E}_n$.*

Note that $\mathcal{AS}_1 = \mathcal{E}_1$, but for $n > 1$ we have the strict inclusions $\mathcal{AS}_n \subset \mathcal{E}_n$.

Structural characterizations of varieties of semilattices of archimedean epigroups are provided by Proposition 3.18 which is based on Theorem 3.16 in conjunction with Theorem 3.14. As for the indicator characterization (3) here, the elimination of the semigroup A_2 (cf. condition (3a) in Theorem 3.16) is explained by the well-known and easily verified fact that B_2 is a Rees divisor of the direct product $A_2 \times A_2$.

3.18 Proposition *The following conditions for a variety $\mathcal{V}$ of epigroups are equivalent:*

(1) *$\mathcal{V}$ consists of semilattices of archimedean epigroups (i.e., in view of Observation* 2.11 *and Lemma* 3.17*, $\mathcal{V}$ is contained in some variety $\mathcal{AS}_n$);*

(1′) *$\mathcal{V}$ consists of unipotently partitionable epigroups;*

(2) *in any epigroup S with zero belonging to $\mathcal{V}$, the set* $\operatorname{Nil} S$ *is an ideal;*

(2′) *in any epigroup S with zero belonging to $\mathcal{V}$, the set* $\operatorname{Nil} S$ *is a subsemigroup;*

(3) *$\mathcal{V}$ does not contain B_2;*

(4) *each regular epigroup in $\mathcal{V}$ is completely regular;*

(4′) *each completely 0-simple epigroup in $\mathcal{V}$ is completely regular.*

In the case of periodic varieties, the equivalence of (1), (1′), and (3) was established earlier in [81] which also contains an equational characterization; the corresponding information was reproduced in [109, Theorem 7.1]. As in [81], criterion (3) of Proposition 3.18 can be employed to further study the varieties of epigroups. From this criterion and the fact (mentioned in [81]) that the variety $\operatorname{var} B_2$ has the non-modular lattice of subvarieties we obtain:

3.19 Corollary *A variety of epigroups whose lattice of subvarieties is modular consists of semilattices of archimedean epigroups.*

Another application of the same criterion leads to the following consequence.

3.20 Corollary *The following conditions for a variety $\mathcal{V}$ of epigroups are equivalent:*

(1) *$\mathcal{V}$ is a variety of finite degree;*

(1′) *each nilsemigroup in $\mathcal{V}$ is nilpotent;*

(2) *$\mathcal{V}$ consists of semilattices of nilpotent extensions of completely simple semigroups.*

Here the equivalence of (1) and (1′) is guaranteed by the theorem in [88] on the nilpotency of a nilsemigroup in which the nilpotency degrees of its nilpotent subsemigroups are uniformly bounded. The characterization given by condition (2) extends to varieties of epigroups the same condition found for periodic varieties in [81]. It would be interesting to extend the other characterizations of the periodic varieties of finite degree obtained in [81] (see also [109, Theorem 8.1]), as well as in [113] and [114], to the case of varieties of epigroups.

Let us return to Corollary 3.19. It can be considered as a starting point for a solution of the following problem already mentioned in fact in [104].

3.21 Problem *Describe the varieties of epigroups $\mathcal{V}$ such that the lattice $L(\mathcal{V})$ is modular.*

Among such varieties are all the completely regular varieties, since it is known that $L(\mathcal{E}_1)$ is modular (see [61, 62, 67]). Further, there is a description of semigroup varieties whose lattice of subvarieties is modular, and, since these varieties are periodic (it follows, for instance, from the result of [83]), they also belong to the family figuring in Problem 3.21. Moreover, in all likelihood, just the description mentioned can be extended to the case of varieties of

epigroups. Remark that this description has solved a long-known problem posed in [20]. The corresponding result had been stated in [121] and published with a proof "modulo the nil case" in [120, 126]. The original version of the proof for the nil case was given only in the dissertation [122]. A revised and improved version of the proof for this case was given in passing in the course of the proofs of certain stronger results obtained in the recent works [118, 119, 125].

In the remainder of this subsection we characterize epigroups of several well-known subclasses of the class of semilattices of archimedean epigroups. Specifically, we will mainly be interested in epigroups that are decomposable into a semilattice of nil-extensions of rectangular groups (Theorem 3.22), into a band of right archimedean epigroups (Theorem 3.28), into a semilattice of right archimedean epigroups (Theorem 3.30), into a band of unipotent epigroups (Theorem 3.31), into a semilattice of unipotent epigroups (Theorem 3.35), or into a rectangular band of unipotent epigroups (Theorem 3.39). The proofs of all these theorems are given in [104, §§4–6]. Notice that in a number of papers some of the types of epigroups considered below have been distinguished within the class of semilattices of archimedean epigroups. We mention in this regard the papers [6, 7, 22, 23, 46, 74, 75]; see also [5, Chapters IX and X] as well as some sections in [9], and the bibliography therein.

It is easy to see that if a semigroup S is decomposable into a semilattice of archimedean semigroups, then such a decomposition is unique, and it is precisely the greatest semilattice decomposition of S. Its components will be called the *archimedean components*. The next object we consider is the case when the kernels of the archimedean components of an epigroup with the property under consideration are rectangular groups. Recall that a *rectangular group* is a completely simple semigroup in which the set of all idempotents is a subsemigroup, or, equivalently, a direct product of a group, a left zero semigroup, and a right zero semigroup. Among the "cast of characters" in Theorem 3.22, along with the semigroups A_2, B_2, are the semigroups M_n, $n \in \mathbb{N}$, defined by the presentations

$$M_n = \langle e, f \mid e^2 = (efe)^n = e,\ f^2 = (fef)^n = f \rangle.$$

The semigroup M_n is isomorphic to the Rees matrix semigroup $\mathcal{M}\,[C_{1,n}; 2, 2; \left(\begin{smallmatrix} \varepsilon & \varepsilon \\ \varepsilon & \gamma \end{smallmatrix}\right)]$, where ε is the identity element and γ is a generator of the group $C_{1,n}$.

3.22 Theorem *The following conditions on an epigroup S are equivalent:*

(1) *S is a semilattice of nil-extensions of rectangular groups;*

(2a) *S satisfies the identity $(xy)^\omega(yx)^\omega(xy)^\omega = (xy)^\omega$;*

(2b) *S satisfies the identities $((xy)^\omega(yx)^\omega(xy)^\omega)^\omega = (xy)^\omega$, $x^\omega y^\omega(x^\omega y^\omega)^\omega = (x^\omega y^\omega)^\omega$;*

(3) *there are no semigroups A_2, B_2, M_p for any prime p among the epidivisors of S.*

Note that condition (2a) of Theorem 3.22 is related to the equivalent (but not formulated in terms of identities) condition $\mathfrak{Q}_5$ of [74], and condition (2b) appeared in [82, §II.6]. Note also that for an element a of an epigroup, the equality $aa^\omega = a^\omega$ holds if and only if a has finite order and period 1, i.e., $a^n = a^{n+1}$ for some natural number n. In view of this remark and condition (8) of Theorem 3.16, we see that condition (2b) of Theorem 3.22 can be transformed into the following criterion (see [5, §2 of Chapter X]):

3.23 Corollary *A semigroup S is a semilattice of nil-extensions of rectangular groups if and only if S is a semilattice of archimedean epigroups, and, for any $e, f \in E_S$, the element ef has finite order and period* 1.

Corollary 3.23 directly implies:

3.24 Corollary *If S is an epigroup that is decomposable into a semilattice of archimedean epigroups, and $\operatorname{Gr} S$ is a subsemigroup, then E_S is a subsemigroup if (and obviously only if) in each archimedean component, the set of idempotents is a subsemigroup.*

We will state two corollaries pertaining to the special case of completely regular semigroups. Recall that a completely regular semigroup S in which E_S is a subsemigroup is called an *orthogroup*. Corollary 3.24, applied to completely regular semigroups, yields the following long-known fact (see [65, Corollary IV.3.2 and Proposition IV.3.7]):

3.25 Corollary *A semigroup is an orthogroup if and only if it is a semilattice of rectangular groups.*

Condition (3) of Theorem 3.22 and Corollary 3.25 give the following criterion:

3.26 Corollary *A completely regular semigroup is an orthogroup if and only if there is no semigroup M_p for any prime p among its epidivisors.*

It is impossible to remove from Corollary 3.24 the requirement that $\operatorname{Gr} S$ be a subsemigroup, i.e., the condition that the set of idempotents is a subsemigroup distinguishes a proper subclass of the class of epigroups described by Theorem 3.22. This subclass is characterized by Proposition 3.27 (which in turn overlaps Corollary 3.26). In this proposition, there appears yet another "character"—the semigroup V defined by the presentation

$$V = \langle e, f \mid e^2 = e,\ f^2 = f,\ fe = 0 \rangle.$$

It is obvious that V consists of four elements: $V = \{e, f, ef, 0\}$.

3.27 Proposition *The following conditions on an epigroup S are equivalent:*

(1) *S is a semilattice of archimedean epigroups, and E_S is a subsemigroup;*

(2) *S satisfies the identities $(xy)^\omega(yx)^\omega(xy)^\omega = (xy)^\omega$, $(x^\omega y^\omega)^\omega = x^\omega y^\omega$;*

(3) *there are no semigroups B_2, V, M_p for any prime p among the epidivisors of S.*

Remark that in the original formulation of this proposition in [104] there was an extra forbidden epidivisor A_2 which may in reality be omitted; indeed, it is easy to see that the semigroup V occurs among the subsemigroups of A_2, so a "ban" of V implies the same for A_2. A similar remark should be made regarding the formulation of Theorem 3.53 below.

In the problem of characterizing semigroups that are decomposable into a band of right [left] archimedean epigroups, we will consider, for definiteness, the first of the dual situations.

The solution of this problem for this case is given by Theorem 3.28. We must add to the "cast of characters" the following semigroups:

$$L_2 = \langle e, f \mid ef = e,\ fe = f \rangle,$$
$$L_{3,1} = \langle a, f \mid a^2 f = a^2,\ fa = f^2 = f \rangle,$$
$$LZ(n) = \langle g, f \mid g^{n+1} = g,\ g^n f = fg = f^2 = f \rangle.$$

Obviously L_2 is merely a two-element left zero semigroup. The semigroup $L_{3,1}$ consists of four elements and is an ideal extension of a three-element left zero semigroup by the semigroup $C_{2,1}$. The semigroup $LZ(n)$ consists of $2n$ elements and is a chain of two semigroups: an n-element left zero semigroup and the cyclic group $C_{1,n}$.

3.28 Theorem *The following conditions on an epigroup S are equivalent:*

(1) *S is a band of right archimedean epigroups;*

(2) *S satisfies the identity $(x^\omega y^\omega)^\omega (xy)^\omega = (xy)^\omega$;*

(3) *there are no semigroups A_2, B_2, $L_{3,1}$, $LZ(n)$ for all $n > 1$ (it suffices to consider n being prime) among the epidivisors of S.*

The equivalence of conditions (1) and (3) of Theorem 3.28 was announced in [98]. For the particular case of completely regular semigroups, we have the following consequence.

3.29 Corollary *A completely regular semigroup is a band of right groups if and only if there is no semigroup $LZ(p)$ for any prime p among its epidivisors.*

We will say that an idempotent e of a semigroup S *switches divisors from right to left* [*from left to right*] if, for any $x \in S$, $e \in Sx$ implies $e \in xS$ [$e \in xS$ implies $e \in Sx$].

3.30 Theorem *The following conditions on an epigroup S are equivalent:*

(1) *S is a semilattice of right archimedean epigroups;*

(2a) *S satisfies the identity $(xy)^\omega (yx)^\omega = (yx)^\omega$;*

(2b) *S satisfies the identity $x^\omega (yx)^\omega = (yx)^\omega$;*

(3) *there are no semigroups A_2, B_2, L_2 among the epidivisors of S;*

(4) *each regular $\mathcal{D}$-class of S is a right group;*

(5) *each idempotent of S switches divisors from right to left;*

(6) *S satisfies any of the conditions of Theorem* 3.16 *and does not contain L_2 as a subsemigroup.*

Note that condition (5) is related to the equivalent condition $\mathfrak{Q}_7$ in [74], and the identity in (2b), as well as the identity in condition (2) of Theorem 3.35 below, appeared in [82, §II.6].

We denote by R_2, $R_{3,1}$, $RZ(n)$ the semigroups dual to L_2, $L_{3,1}$, $LZ(n)$, respectively.

3.31 Theorem *The following conditions on an epigroup S are equivalent:*

(1) *S is a band of unipotent epigroups;*

(2) *S satisfies the identity $(xy)^\omega = (x^\omega y^\omega)^\omega$;*

(3) *there are no semigroups A_2, B_2, $L_{3,1}$, $R_{3,1}$, $LZ(n)$, $RZ(n)$ for any $n > 1$ (it suffices to consider n being prime) among the epidivisors of S.*

Remark that the implication (1) $\Rightarrow$ (2) here is obvious; the rest follows from Theorem 3.28 and its dual, taking into account that, for an epigroup, the property of being unipotent is equivalent to the property of being bi-archimedean (see Proposition 3.6). For completely regular semigroups we obtain the following divisor criterion complementing other long-known criteria for decomposability into a band of groups (see, for example, [13, Vol. I, p. 129]).

3.32 Corollary *A completely regular semigroup is a band of groups if and only if there are no semigroups $LZ(n)$ and $RZ(n)$ for any $n > 1$ among its epidivisors.*

In turn, Corollary 3.32 automatically implies the following indicator characterization found in [76].

3.33 Corollary *A variety of completely regular semigroups consists of bands of groups if and only if it does not contain the semigroups $LZ(n)$ and $RZ(n)$ for any $n > 1$.*

Here is one more direct consequence of Theorem 3.31; for the case of periodic semigroups this result appeared in [71].

3.34 Corollary *Any epigroup with two idempotents is a band of unipotent epigroups.*

The next theorem follows automatically from Theorem 3.30 and its dual.

3.35 Theorem *The following conditions on an epigroup S are equivalent:*

(1) *S is a semilattice of unipotent epigroups;*

(2) *S satisfies the identity $(xy)^\omega = (yx)^\omega$;*

(3) *there are no semigroups A_2, B_2, L_2, R_2 among the epidivisors of S;*

(4) *each regular $\mathcal{D}$-class of S is a group;*

(5) *each idempotent of S switches divisors from left to right and from right to left;*

(6) *S satisfies any of the conditions of Theorem* 3.16 *and does not contain L_2 or R_2 as subsemigroups.*

We limit ourselves to three examples of well-known facts that follow directly from Theorem 3.35. Recall that a semigroup S is called *weakly commutative* if $\langle ab\rangle \cap bSa \neq \emptyset$ for all $a, b \in S$.

3.36 Corollary *The list of equivalent conditions in Theorem* 3.35 *can be augmented by the following ones:*

(7) *in S, the equality $\mathcal{K} = \sqrt{\mathcal{D}}$ holds;*

(8) *S is weakly commutative.*

Note that the equality $\mathcal{K} = \sqrt{\mathcal{D}}$ is equivalent to the inclusion $\mathcal{D} \subseteq \mathcal{K}$. In this form condition (7) occured earlier. The equivalence of (1), (7), and (8) applied to periodic semigroups was established in [50], taking into account [85]; the corresponding results were extended to epigroups in [46].

3.37 Corollary *Any epigroup with central idempotents (in particular, any commutative epigroup) is decomposable into a semilattice of unipotent epigroups.*

Periodic semigroups with central idempotents were already considered in [40,41]. Almost all of the results of these two papers can be extended to epigroups. In particular, for any semilattice A and any family of unipotent epigroups U_α ($\alpha \in A$), there exists an epigroup that is decomposable into a semilattice of these epigroups; it can be realized by a strong semilattice structure, which extends Clifford's construction describing semilattices of groups (for the latter, see, for instance, [49, Section A10]).

3.38 Corollary *A semigroup S is a chain of unipotent epigroups if and only if S is an epigroup in which E_S is a chain.*

Remark that in [5, p. 142], this result was given in a somewhat weaker form, where the property of being an epigroup was replaced by the property of being decomposable into a semilattice of archimedean epigroups (in the terminology of [5], the property of being a GV-semigroup).

3.39 Theorem *The following conditions on an epigroup S are equivalent:*

(1) *S is a rectangular band of unipotent epigroups;*

(2a) *S satisfies the identity $(xyx)^\omega = x^\omega$;*

(2b) *S satisfies the identities $(x^\omega y x^\omega)^\omega = x^\omega$, $(xy)^\omega = (x^\omega y^\omega)^\omega$;*

(3) *there are no two-element semilattice or semigroups $L_{3,1}$, $R_{3,1}$ among the epidivisors of S;*

(4a) *S is archimedean, and $\operatorname{Gr} S$ is a retract of S;*

(4b) *S is a subdirect product of a completely simple semigroup and a nilsemigroup.*

The equivalence of conditions (1) and (4b) was announced in [95], and one of the results in [22] established the equivalence of (1) and (4a) (it also appeared in [23] and in [5, §2 of Chapter IX]; condition (4a) was given in the cited papers in the following equivalent form: S is a retract extension of a completely simple semigroup by a nilsemigroup).

3.4 Epigroups whose group part is a subsemigroup

It is natural to consider epigroups in which the operation $x \mapsto \overline{x}$ is an endomorphism or an anti-endomorphism. Such epigroups are distinguished by the identities $\overline{xy} = \overline{x} \cdot \overline{y}$ and $\overline{xy} = \overline{y} \cdot \overline{x}$, respectively. In either case, the epigroup in question satisfies the identity

$$\overline{\overline{xy}} = \overline{\overline{x}} \cdot \overline{\overline{y}}. \tag{3.1}$$

Since an element a of an epigroup S belongs to $\operatorname{Gr} S$ if and only if $a = \overline{\overline{a}}$, the identity (3.1) implies, in turn, that the set $\operatorname{Gr} S$ is a subsemigroup of S. In this subsection we pay attention to all four of the mentioned restrictions on an epigroup.

We start with the last restriction defining the largest of the corresponding classes of epigroups. Note that several particular types of epigroups satisfying this condition occurred in Subsections 3.1 and 3.3. We give now an indicator characterization of the epigroups under consideration in terms of forbidden epidivisors. This result has been obtained in [124]. To formulate it, we need to augment our "cast of characters" by a new family of individual semigroups.

The following construction was presented in [78]. Let G be a group. Define a multiplication on the disjoint union $R(G)$ of G^0 with the cartesian square $G \times G$ by preserving the multiplication on G^0 and letting, for all $g \in G$, $[h, j], [k, \ell] \in G \times G$,

$$[h,j] \cdot g = [h, jg], \qquad g \cdot [h,j] = [hg^{-1}, j], \qquad [h,j] \cdot 0 = 0 \cdot [h,j] = 0,$$
$$[h,j] \cdot [k,\ell] = \begin{cases} [h,\ell] & \text{if } j = k, \\ 0 & \text{otherwise.} \end{cases}$$

Then $R(G)$ becomes a semigroup being obviously an epigroup. Recall that, as in Subsection 2.1, we denote by $C_{1,\infty}$ and $C_{1,n}$ the infinite cyclic group and the cyclic group of order n, respectively. The semigroup V was introduced in Subsection 3.3.

3.40 Theorem *For an epigroup S, the set $\operatorname{Gr} S$ is a subsemigroup if and only if there are no semigroups V, $R(C_{1,\infty})$, $R(C_{1,p})$ for any prime p among the epidivisors of S.*

The role of the semigroup V in general is clarified by the following fact established also in the paper [124].

3.41 Proposition *The semigroup V occurs among the epidivisors of an epigroup S if (and obviously only if) there exist $e, f \in E_S$ such that $ef \notin \operatorname{Gr} S$.*

Let us turn to epigroups satisfying the identity (3.1). First of all, they are characterized by the following simple property, observed in [104, §1].

3.42 Lemma *In an epigroup S, the mapping $x \mapsto \overline{\overline{x}}$ is an endomorphism if and only if the set $\operatorname{Gr} S$ is a retract. In this case, the indicated mapping is a unique retraction of S onto $\operatorname{Gr} S$.*

Part A of the next proposition clarifies a structure of such epigroups and shows that they form a subclass of the class of epigroups considered in Theorem 3.16, part B gives a sufficient condition for the property under discussion.

3.43 Proposition

A. *An epigroup S in which the set* $\operatorname{Gr} S$ *is a retract is decomposable into a semilattice of rectangular bands of unipotent epigroups.*

B. *If an epigroup S is decomposable into a band of unipotent epigroups and* $\operatorname{Gr} S$ *is a subsemigroup, then* $\operatorname{Gr} S$ *is a retract.*

The premise in A is not a necessary condition, and the conclusion in B is not a sufficient condition for an epigroup to have the group part as a retract. The first assertion is confirmed by the semigroups $LZ(n)$ for $n > 1$, which, being completely regular semigroups, trivially satisfy this condition but undecomposable into a band of unipotent epigroups (see Theorem 3.31). The second assertion is confirmed by the semigroup V which, as is easily seen, does not satisfy the identity $\overline{\overline{xy}} = \overline{\overline{x}} \cdot \overline{\overline{y}}$ but is decomposable even into a semilattice of (three) unipotent semigroups. Thus the class of epigroups under discussion lies strictly between the classes figuring in the premise in A and the conclusion in B of Proposition 3.43. In [104] the problem of finding a divisor characterization of epigroups of this class was noted. Such a characterization has been found in [124].

3.44 Theorem *For an epigroup S, the set* $\operatorname{Gr} S$ *is a retract if and only if there are no semigroups B_2, $L_{3,1}$, $R_{3,1}$, V among the epidivisors of S.*

The work [124] also contains (without proofs) divisor characterizations of epigroups whose group part is a quasiideal, or a left ideal, or an ideal, or a retract ideal. Recall that the set $Q \subseteq S$ is called a *quasiideal* if $QS \cap SQ \subseteq Q$. The characterizations mentioned involve the following semigroups supplementing our "cast of characters":

$$C = \langle a, e \,|\, e^2 = e,\ ae = ea = a,\ a^2 = 0 \rangle;$$
$$Y = \langle a, e, f \,|\, e^2 = e,\ f^2 = f,\ ef = fe = 0,\ ea = af = a \rangle;$$
$$P = \langle a, e \,|\, e^2 = e,\ ea = a,\ ae = 0 \rangle,$$

and the semigroup $\overleftarrow{P}$ which is dual to P. These semigroups have a very simple structure; C, P, and $\overleftarrow{P}$ are three-element, Y is four-element.

3.45 Theorem *For an epigroup S, the set* $\operatorname{Gr} S$ *is a quasiideal of S if and only if there are no semigroups C, Y, V among the epidivisors of S.*

3.46 Theorem *For an epigroup S, the set* $\operatorname{Gr} S$ *is a left ideal of S if and only if there are no semigroups C, P among the epidivisors of S.*

Theorem 3.46 and its dual immediately imply the following consequence.

3.47 Corollary *For an epigroup S, the set* $\operatorname{Gr} S$ *is an ideal of S if and only if there are no semigroups C, P, $\overleftarrow{P}$ among the epidivisors of S.*

By combining this corollary and Theorem 3.44, we obtain a characterization of epigroups whose group part is a retract ideal. Since the semigroup P is, as easily seen, isomorphic to a subsemigroup in both B_2 and V, this characterization is reduced to the following statement.

3.48 Theorem *For an epigroup S, the set* $\operatorname{Gr} S$ *is a retract ideal of S if and only if there are no semigroups C, P, $\overleftarrow{P}$, $L_{3,1}$, $R_{3,1}$ among the epidivisors of S.*

Let us now turn to epigroups satisfying the identity $\overline{xy} = \overline{x}\cdot\overline{y}$. To the property indicated by Lemma 3.43, the following properties have obviously to be added: all the maximal subgroups are abelian and the subsemigroup generated by the idempotents satisfies the identity $\overline{x} = x$, which is equivalent to the identity $x^3 = x$. Thus we have the following fact.

3.49 Proposition *An epigroup S satisfying the identity $\overline{xy} = \overline{x}\cdot\overline{y}$ is decomposable into a semilattice of rectangular bands of nil-extensions of abelian groups, and the subsemigroup $\langle E_S\rangle$ satisfies the identity $x^3 = x$.*

That the converse is false is demonstrated by the following example.

3.50 Example Let S be the semigroup given by the presentation

$$S = \langle a, f \mid a^3 = a^2,\ fa = f^2 = f\rangle.$$

(In [104] this semigroup was denoted by $L_{3,2}$; observe that the semigroup $L_{3,1}$ which occured several times above is a homomorphic image of the semigroup $L_{3,2}$.) Obviously S consists of five elements a, a^2, f, af, a^2f, and it is a chain of the three-element left zero semigroup $\{f, af, a^2f\}$ and the semigroup $\langle a\rangle = C_{2,1}$. Hence S trivially satisfies all the conditions of Proposition 3.49 (in particular $\langle E_S\rangle = E_S$ satisfies even the identity $x^2 = x$). However the premise is not fulfilled, since $\overline{af} = af \neq a^2f = \overline{a}\cdot\overline{f}$.

Therefore one may pose the following problem (mentioned in [104]).

3.51 Problem *Find a structural characterization (in particular, in terms of forbidden divisors) of the epigroups obeying the identity $\overline{xy} = \overline{x}\cdot\overline{y}$.*

If we limit ourselves to archimedean epigroups, then, as was shown in [104], the converse of Proposition 3.49 is true, i.e., we have

3.52 Proposition *An archimedean epigroup S satisfies the identity $\overline{xy} = \overline{x}\cdot\overline{y}$ if and only if S is a rectangular band of nil-extensions of abelian groups, and the subsemigroup $\langle E_S\rangle$ satisfies the identity $x^3 = x$.*

At last, let us turn to epigroups satisfying the identity $\overline{xy} = \overline{y}\cdot\overline{x}$. They are characterized from several points of view by the following statement proved in [104, §6].

3.53 Theorem *The following conditions on an epigroup S are equivalent:*

(1) *S satisfies the identity $\overline{xy} = \overline{y}\cdot\overline{x}$;*

(2) *S satisfies the identities $(xy)^\omega = (yx)^\omega$, $(x^\omega y^\omega)^\omega = x^\omega y^\omega$;*

(3) *there are no semigroups B_2, L_2, R_2, V among the epidivisors of S;*

(4a) *S is a semilattice of unipotent epigroups, and E_S is a subsemigroup (which is then a semilattice);*

(4b) *S is a semilattice of unipotent epigroups, and* $\operatorname{Gr} S$ *is a subsemigroup.*

Recall that a unary semigroup is called *involutory* if its unary operation is an anti-automorphism of order 2. Theorem 3.53 has the following immediate

3.54 Corollary *An epigroup S is an involutory semigroup (with respect to the operation $x \mapsto \overline{x}$) if and only if S is a semilattice of groups.*

3.5 Unipotency classes and Green's relations

We will denote by $\mathcal{K}$ the equivalence relation on an epigroup corresponding to the partition of the given epigroup into its unipotency classes. In this subsection we are going to talk about the interaction between the relation $\mathcal{K}$ and the relations $\mathcal{D}$, $\mathcal{L}$, $\mathcal{R}$, $\mathcal{H}$; among the dual situations for $\mathcal{L}$, $\mathcal{R}$, we will, for definiteness, consider $\mathcal{L}$. The results presented in this subsection are proved in [104, §7].

Note first that in any epigroup $\mathcal{K} \subseteq \sqrt{\mathcal{H}}$ and, in any semigroup, $\sqrt{\mathcal{H}} \subseteq \sqrt{\mathcal{L}} \subseteq \sqrt{\mathcal{D}}$. An example of non-trivial connections is provided by Corollary 3.36 of Theorem 3.35. By analogy with the condition $\mathcal{K} = \sqrt{\mathcal{D}}$ there, we will be interested in the equalities $\mathcal{K} = \sqrt{\mathcal{L}}$ and $\mathcal{K} = \sqrt{\mathcal{H}}$, which define wider classes of epigroups. As in the case of $\mathcal{D}$, these equalities are equivalent, respectively, to the inclusions $\mathcal{L} \subseteq \mathcal{K}$ and $\mathcal{H} \subseteq \mathcal{K}$. The equalities $\mathcal{K} = \mathcal{L}$ and $\mathcal{K} = \mathcal{H}$ are stronger conditions; they, along with the equality $\mathcal{K} = \mathcal{D}$, were considered in [46] (and, for periodic semigroups, in [85]). Since, as is easy to prove, $\mathcal{K} \subseteq \mathcal{D}$ holds in an epigroup S if and only if S is a completely regular semigroup, a characterization of the corresponding epigroups follows directly from Clifford's theorem on the decomposability of any completely regular semigroup into a semilattice of completely simple semigroups: in an epigroup S, we have $\mathcal{K} = \mathcal{D}$ [$\mathcal{K} = \mathcal{L}$, $\mathcal{K} = \mathcal{H}$] if and only if S is a semilattice of groups [a semilattice of right groups, a completely regular semigroup].

Let us turn to the condition $\mathcal{L} \subseteq \mathcal{K}$. The following is true.

3.55 Lemma *An epigroup in which $\mathcal{L} \subseteq \mathcal{K}$ is a semilattice of right archimedean epigroups.*

There arises the question: Does the property mentioned in the conclusion of Lemma 3.55 characterize the epigroups under consideration? A negative answer is provided by the following example.

3.56 Example Let S be the semigroup defined by the presentation

$$S = \langle a, g \mid a^3 = a^2,\ g^5 = g,\ g^2a = ag^4 = a,\ ga^2 = a^2,\ a^2g^2 = aga \rangle.$$

It is convenient to put $e = a^2$, $i = g^4$. It follows directly from the defining relations that e, i are idempotents, and e is a right zero. Then eg, eg^2, eg^3 are also right zeros. It is now easy to see that S consists of 16 elements and has five torsion classes: $K_i = \langle g \rangle$, $K_e = \{a, gag^2, e\}$, $K_{eg} = \{ag^3, gag, eg\}$, $K_{eg^2} = \{ag^2, ga, eg^2\}$, $K_{eg^3} = \{ag, gag^3, eg^3\}$.

The semigroup S is a chain of two right archimedean semigroups: $K_e \cup K_{eg} \cup K_{eg^2} \cup K_{eg^3}$

and K_i (the latter is simply a cyclic group of order 4). The $\mathcal{L}$-classes and $\mathcal{R}$-classes are

$$\begin{array}{llll} L_i = R_i = \langle g\rangle, & & & \\ L_e = \{e\}, & L_{eg} = \{eg\}, & L_{eg^2} = \{eg^2\}, & L_{eg^3} = \{eg^3\}, \\ L_a = \{a, ga\}, & L_{ag} = \{ag, gag\}, & L_{ag^2} = \{ag^2, gag^2\}, & L_{ag^3} = \{ag^3, gag^3\}, \\ R_e = e\langle g\rangle, & R_a = a\langle g\rangle, & R_{ga} = ga\langle g\rangle. & \end{array}$$

In particular, S has $\mathcal{L}$-equivalent elements lying in different torsion classes.

It is possible to exhibit a semigroup with the desired properties having fewer elements (e.g., the semigroup in Example 3.61), but the one in Example 3.56 is needed for one of the conclusions below. For this reason, we indicate its $\mathcal{H}$-classes: $H_i = \langle g\rangle$ and the remaining $\mathcal{H}$-classes are singletons; thus $\mathcal{H} \subseteq \mathcal{K}$.

The following lemma, together with the preceding one, "brackets" the class of epigroups under discussion.

3.57 Lemma *If S is a right archimedean epigroup or a semilattice of right archimedean epigroups in which the archimedean components are decomposable into a band of unipotent epigroups, then $\mathcal{L} \subseteq \mathcal{K}$ in S.*

The disjunction in the premise of Lemma 3.57 is also not a characteristic property of epigroups in which $\mathcal{L} \subseteq \mathcal{K}$, as is demonstrated by the example of the semigroup obtained from $R_{3,1}$ by adjoining an identity element: the realtion $\mathcal{L}$ in this semigroup coincides with the equality relation, but it is not archimedean, and its archimedean component $R_{3,1}$ is indecomposable into a band of unipotent epigroups. We see that the class of epigroups in question lies strictly between the classes appearing in the premise of Lemma 3.57 and the conclusion of Lemma 3.55. How can one characterize it in the spirit of the earlier characterizations in this section? There is a divisor characterization of this class. Indeed, it is closed under subepigroups (which is obvious) and homomorphic images (which is proved in [104, Lemma 28]). Thus we may pose the following problem (already noted in [104]).

3.58 Problem *Find a divisor characterization of epigroups in which $\mathcal{L} \subseteq \mathcal{K}$.*

It is easy to see that within each of the varieties $\mathcal{E}_n$, a direct product of epigroups such that $\mathcal{L} \subseteq \mathcal{K}$ also has this property. Therefore, in view of what was said before the formulation of Problem 3.58, the class of such epigroups is a subvariety of $\mathcal{E}_n$. An equational characterization of this class is provided by the following statement.

3.59 Proposition *The largest subvariety of $\mathcal{E}_n$ consisting of epigroups in which $\mathcal{L} \subseteq \mathcal{K}$ is defined within $\mathcal{E}_n$ by the identity $((xy)^n z)^\omega = (y(xy)^n z)^\omega$.*

Let us now turn to epigroups in which $\mathcal{H} \subseteq \mathcal{K}$. The class of such epigroups is considerably wider than the preceding one and consists not only of semilattices of archimedean epigroups or even of unipotently partitionable epigroups, as is shown by the example of the semigroup A_2, where $\mathcal{H}$ coincides with the equality relation. The connection between this class and those considered earlier in this section is demonstrated by the following statement which is an analogue of Lemma 3.57 in a more general situation.

3.60 Proposition *If S is an archimedean semigroup or a semilattice of rectangular bands of unipotent epigroups, then $\mathcal{H} \subseteq \mathcal{K}$ in S.*

Since under the condition $\mathcal{H} = \mathcal{K}$ we obtain the class of all completely regular semigroups, one might put the question: Does the inclusion $\mathcal{H} \subseteq \mathcal{K}$ hold in any epigroup? A negative answer is provided by the following example.

3.61 Example Let S be the semigroup defined by the presentation

$$S = \langle a, g \mid a^3 = a^2,\ g^5 = g,\ ga = ag^2,\ ag^4 = a,\ ga^2 = a^2 \rangle.$$

It follows from the defining relations that $g^2a = a$ and $a^2g^2 = aga$, so all the defining relations of the semigroup in Example 3.56 are satisfied. Thus S is a homomorphic image of that semigroup; S is obtained from it by a pairwise "fusing" of the following elements: ga and ag^2, a and gag^2, ag and gag^3, ag^3 and gag. This implies the "fusion" of the $\mathcal{R}$-classes R_a and R_{ga} as well as the $\mathcal{L}$-classes L_a and L_{ag^2}, L_{ag} and L_{ag^3}. Thus S has the following 2-element $\mathcal{H}$-classes: $H_a = \{a, ga\}$, $H_{ag} = \{ag, gag\}$; the elements of each of them lie in different torsion classes, e.g., $a \in K_{a^2}$, $ga \in K_{a^2g^2}$. Thus the inclusion $\mathcal{H} \subseteq \mathcal{K}$ is not fulfilled in S.

In contrast to the situation for epigroups in which $\mathcal{L} \subseteq \mathcal{K}$, the problem of finding a divisor characterization for epigroups such that $\mathcal{H} \subseteq \mathcal{K}$ cannot be posed. This is caused by the fact that the class of these epigroups is not homomorphically closed; it follows from the arguments in Examples 3.56 and 3.61: as was noted above, $\mathcal{H} \subseteq \mathcal{K}$ in the epigroup of Example 3.56, and the epigroup of Example 3.61 is a homomorphic image of the former.

4 Finiteness conditions

4.0 Introductory remarks

Given a class of algebraic systems, by a *finiteness condition* is meant any property which is possessed by all finite systems of this class. Imposing finiteness conditions is a classical approach in investigations of algebraic systems of different kinds. For semigroups, a broad group of such conditions deals with conditions formulated in terms of ideals or congruences of certain types, in particular, in terms of the partially ordered sets of principal (one-sided or two-sided) ideals. Some facts in this direction are presented in [13, §6.6]; see also works [26, 30, 31, 42] as well as [45, §3.6]. For example, in [26] the interdependence of several minimal conditions (including, by the way, the property of being an epigroup) has been clarified. In this section we shall focus attention on another group of conditions which are expressed basically in terms of subsemigroups, especially in terms of the lattice of subsemigroups. They distinguish some subclasses within the class of epigroups. A trivial finiteness condition is the property of being plainly a finite semigroup. To describe semigroups with some non-trivial finiteness conditions, we should clarify, so to speak, a character and a degree of "deviations" from the property of being a finite semigroup. For semigroups considered in Subsections 4.1 and 4.2, such deviations will always happen in the maximal subgroups only. Thus we have a complete reduction to groups here. Information on groups with the conditions under consideration is given in Subsection 4.3.

Note that the conditions in question are hereditary for subsemigroups (or subepigroups), so, as in the situations examined in Section 3, the idea of "forbidden" objects can be used for a characterization of the corresponding semigroups. These objects are now subsemigroups, and the role of forbidden subsemigroups is played by semigroups having a unique infinite basis. The absence of such subsemigroups is a crucial factor in a general scheme (presented in Subsection 4.1) embracing many concrete finiteness conditions.

For more detailed information on the matters discussed in Subsections 4.1–4.3 (including the proofs of many results), see [108, Chapter IV]. In Subsection 4.4 we will briefly touch on several finiteness conditions for nilsemigroups, predominantly calling the reader's attention to some open problems.

4.1 Finitely assembled semigroups

A semigroup S is called *finitely assembled* (*f.a.*) if the sets E_S and $S \setminus \operatorname{Gr} S$ are finite. Any f.a. semigroup is evidently an epigroup. The term "finitely assembled" is justified by the observation that such a semigroup looks as if it is assembled from a finite family of (disjoint) groups and a finite set of (non-group) elements. If in such a semigroup S all its maximal subgroups belong to a fixed class $\mathcal{X}$, we say that S is *f.a. from groups of* $\mathcal{X}$. The structure of f.a. semigroups is fairly clarified by the following description.

4.1 Proposition *A semigroup S is finitely assembled if and only if S has a finite series of ideals in which each factor is either finite or is a rectangular band of finitely many infinite groups* (*this takes place for the kernel of S only*) *or is obtained from such a band by the adjunction of a zero.*

Note that this proposition in its non-trivial part is based on some statement relating to a more general situation, namely, when a given epigroup has finitely many idempotents (this is Proposition 11.1 in [108]).

A key role in the subsequent considerations of this subsection is played by certain properties formulated in terms of bases. By a *basis* of an algebraic system is meant an irreducible (i.e., minimal) generating set of the given system. In order to distinguish bases of an epigroup regarded as a (usual) semigroup or a unary semigroup, an epigroup basis of the given epigroup will be called an *epibasis*. Recall that a semigroup of idempotents is often called a *band*.

4.2 Proposition *Any epigroup with finitely many idempotents and infinitely many non-group elements has a* (*unipotent*) *subsemigroup* [*subepigroup*] *with a unique infinite basis* [*epibasis*].

4.3 Proposition *Any epigroup with infinitely many idempotents has a subsemigroup with a unique infinite basis. Such a subsemigroup may be taken to be either a nilpotent semigroup or a band; so in both cases it is a subepigroup with a unique infinite epibasis.*

Remark that the inevitable question of whether it is possible to retain only a band in the conclusion of Proposition 4.3 has a negative answer, which the following example shows.

4.4 Example Consider a semigroup $T = \langle t_i \mid t_i^2 = t_i,\ i \in \mathbb{N} \rangle$, and denote by T_3 the set of all elements in T of the form $t_{i_1} t_{i_2} \cdots t_{i_n}$, where $n \geq 3$ and adjacent elements t_{i_k} are distinct.

Evidently T_3 is an ideal of T. It is seen that the Rees quotient semigroup $S = T/T_3$ has infinitely many idempotents; however all the maximal subbands in S are two-element: each of them consists of the zero T_3 and $\{t_i\}$ for some i.

Propositions 4.2 and 4.3 immediately imply the following

4.5 Theorem *If an epigroup has no subsemigroups with a unique infinite basis, then it is finitely assembled.*

The property in the premise of this theorem is not characteristic for finitely assembled semigroups. A trivial counter-example is given by any group having a subsemigroup with a unique infinite basis, for example, by a non-cyclic free group (which, as is well known, contains a free subsemigroup of countable rank). However, by considering the subepigroup versions of Propositions 4.2 and 4.3, we obtain another corollary of them which gives a necessary and sufficient criterion.

4.6 Theorem *Finitely assembled semigroups are exactly the epigroups without subepigroups with a unique infinite epibasis.*

For an abstract semigroup-theoretic property θ, a semigroup possessing this property will be called a *θ-semigroup*. For quite a number of finiteness conditions, it is possible to cover by a general scheme the description of semigroups satisfying these conditions. This scheme is based on Theorem 4.5, or Theorem 4.6. To expound it, we need the following requirements which may be satisfied for a property θ:

(A) θ is hereditary for subsemigroups;

(B) any θ-semigroup has no unique infinite basis;

(C) any semigroup being the union of a finite family of θ-subsemigroups is itself a θ-semigroup.

4.7 Theorem *Let θ be a finiteness condition satisfying the requirements* (A)–(C). *Then a semigroup is a θ-epigroup if and only if it is finitely assembled from θ-groups.*

When considering epigroups, one may also deal with similar requirements concerning subepigroups:

(A′) θ is hereditary for subepigroups;

(B′) any θ-epigroup has no unique infinite epibasis;

(C′) any epigroup being the union of a finite family of θ-subepigroups is itself a θ-epigroup.

4.7′ Theorem Let θ be a finiteness condition satisfying the requirements (A′)–(C′) and, in addition, such that an ideal extension of a θ-epigroup by a finite semigroup is a θ-epigroup. Then a semigroup is a θ-epigroup if and only if it is finitely assembled from θ-groups.

These theorems give, under highly general requirements on a property θ, a description of θ-epigroups effecting a complete reduction to the case of groups. So further possible advances in clarification of the structure of epigroups under consideration becomes entirely a

task of group theory. It concerns, in particular, the interrelation between different conditions. Indeed, let $\mathcal{E}(\theta)$ and $\mathcal{G}(\theta)$ denote the classes of all θ-epigroups and all θ-groups, respectively; then we have the following immediate consequence of Theorems 4.7 and 4.7′.

4.8 Proposition *Let θ_1 and θ_2 be two finiteness conditions satisfying the requirements in Theorem* 4.7 *or in Theorem* 4.7′. *Then the inclusion $\mathcal{E}(\theta_1) \subseteq \mathcal{E}(\theta_2)$ holds if (and obviously only if) $\mathcal{G}(\theta_1) \subseteq \mathcal{G}(\theta_2)$. This, in turn, implies similar conclusions for strict inclusion and for the equality relation. Furthermore, if $\mathcal{E}(\theta_1) \subset \mathcal{E}(\theta_2)$, then any θ_2-epigroup which is not a θ_1-epigroup contains a subgroup which, being a θ_2-group, is not a θ_1-group.*

The general scheme presented above has arisen as a result of several subsequent approximations. It was first proposed, for the case of periodic semigroups and in a slightly weaker form, in [93]. A strengthened variant was exhibited in [94]. Later the author found an improved variant, but, as before, for periodic semigroups only; it was announced in [105] (notice that f.a. semigroups were called *almost finite* there). Finally, the latter was extended to epigroups in [106]. Theorems 4.5 and 4.7 were announced in [99], Theorems 4.6 and 4.7′ were stated in [100].

In Subsection 4.2 we consider a number of concrete finiteness conditions covered by this general scheme, namely, by Theorem 4.7. It should be noted that in most cases the semigroups under consideration turn out to be periodic. Since a subsemigroup of a periodic semigroup is automatically an epigroup, there is no need to apply Theorem 4.7′ in such cases, and in the formulation of the corresponding specifications of Theorem 4.7 the assumption that a semigroup be a priori an epigroup may be omitted. Furthermore, since a subsemigroup of a periodic group is in fact a subgroup, the corresponding properties of groups may then be formulated in terms of subgroups. Information about groups with the conditions under discussion will be given in Subsection 4.3.

4.2 Certain concrete finiteness conditions

In the applications of Theorem 4.7 to concrete cases, one can easily verify requirements (A)–(C) (moreover, the validity of conditions (A) and (B) will always be practically obvious). Almost all of the conditions considered below are formulated in terms of the lattice Sub of all subsemigroups of a semigroup S. (In order to have the right to speak of $\operatorname{Sub} S$ as a lattice, one has to treat the empty set as a subsemigroup).

If $\operatorname{Sub} S$ satisfies the minimal [maximal] condition, we will call such a semigroup S a Min-*semigroup* [Max-*semigroup*]. It is obvious that any Min-semigroup is periodic. For this condition Theorem 4.7 turns into the following statement.

4.9 Theorem *A semigroup S is a* Min-*semigroup if and only if S is finitely assembled from groups with the minimal condition for subgroups.*

It is clear that the property of a semigroup of having only finite proper subsemigroups is stronger than the property of being a Min-semigroup. Therefore Theorem 4.9 immediately implies the following consequence in which a reduction to groups turns out to be simply "dismembered".

4.10 Corollary *In a semigroup S, all the proper subsemigroups are finite if and only if S is either a finite semigroup or an infinite group whose proper subgroups are finite.*

The assertion of this corollary for the case of commutative semigroups was proved in [35], and it was observed there that an infinite abelian group whose proper subgroups are finite is a quasicyclic group. Recall that a group G is called *quasicyclic* if it is given by the presentation

$$G = \langle a_1, a_2, \ldots, a_n \,|\, a_1^p = 1,\ a_{n+1}^p = a_n \text{ for all } n \in \mathbb{N} \rangle,$$

where p is a prime number. The characteristic property of quasicyclic groups is that the lattice of subgroups is an infinite chain (being in reality isomorphic to the chain $\mathbb{N}$ with respect to the ordinary order).

In connection with Corollary 4.10, note that an interesting development of the theme "semigroups whose proper subsemigroups have a lesser cardinality than that of the whole semigroup" was accomplished in [48]. A semigroup with the property just mentioned was called a *Jónsson semigroup* there, and the following statement was proved.

4.11 Theorem *If the generalized continuum hypothesis is assumed, then any Jónsson semigroup is a group.*

Let us turn to Max-semigroups. Such a semigroup does not have to be periodic or even an epigroup; an example is provided by the infinite cyclic semigroup. Therefore in the corresponding concrete version of Theorem 4.7 one cannot eliminate the assumption that a semigroup under consideration be an epigroup. We obtain the following criterion.

4.12 Theorem *An epigroup S is a* Max*-semigroup if and only if S is finitely assembled from* Max*-groups.*

As for Max-semigroups which are not epigroups, the situation is far from clear. One may first notice the following property which is weaker than finite assembleness.

4.13 Proposition *Any* Max*-semigroup has finitely many idempotents.*

Further information about Max-semigroups can be given for commutative semigroups. Here we have a complete description. Recall that any commutative semigroup is decomposable into a semilattice of archimedean semigroups, so one may speak about the archimedean components of such a semigroup.

4.14 Theorem *For a commutative semigroup S, the following conditions are equivalent:*

(1) *S is a* Max*-semigroup;*

(2) *S has finitely many archimedean components each of which is an ideal extension of a semigroup embeddable in an* (*abelian*) Max*-group by a finite nilpotent semigroup;*

(3) *S is embeddable in a* (*commutative*) *semigroup which is finitely assembled from* Max*-groups.*

Condition (2) here is a modified variant of the criterion formulated in [37], the criterion given by condition (3) was found in [107]. Note that certain information about commutative Max-semigroups was previously obtained in [47]. In particular, the following statement has been proved there.

4.15 Proposition *For an archimedean commutative semigroup S without idempotents, the following conditions are equivalent:*

(1) *S is a* Max*-semigroup;*

(2) *S is finitely generated;*

(3) *S is embeddable in the direct product of the infinite cyclic semigroup and a finite unipotent semigroup.*

A question concerning possible clarification of the structure of arbitrary Max-semigroups was posed by the author in the late sixties and was formulated in [105, Question 50]. Condition (3) of Theorem 4.14 suggests a plausible conjecture whose affirmation would mean obtaining a good reducing description of Max-semigroups. This conjecture is that the answer to question (a) in the following problem is affirmative. It seems advisable to consider also questions (b) and (c) in this problem which are consequently weaker from the point of view of the positive answers.

4.16 Problem

(a) *Is any* Max*-semigroup embeddable in a semigroup which is finitely assembled from* Max*-groups?*

(b) *Is any* Max*-semigroup embeddable in a finitely assembled semigroup?*

(c) *Is any* Max*-semigroup embeddable in an epigroup?*

These questions were first formulated in [107], together with the following problem.

4.17 Problem *Is any* Max*-semigroup that can be embedded in a group embeddable in a* Max*-group*?

In what follows we mention several more conditions covered by Theorem 4.7. There is no need to reproduce the special formulations of the theorem for these conditions; remark only that each of the conditions under consideration implies periodicity of a semigroup satisfying this condition. We present mostly the definitions of these conditions and note some interrelations between them. As already remarked in Subsection 4.0, all relevant details can be found in [108, Chapter IV].

A semigroup S is said to have *finite rank* [*finite breadth*] r if any finitely generated subsemigroup of S is generated by at most r elements [for any $r+1$ elements of S, at least one belongs to the subsemigroup generated by the others], and r is the least number with this property. A semigroup which is not a semigroup of finite rank [finite breadth] is called a semigroup of *infinite rank* [*infinite breadth*]; similarly we shall use the word "infinite" for to other conditions which have a numerical meaning. The formulated definition of a semigroup of finite breadth is equivalent to the definition given originally in terms of the subsemigroup lattice; for a semigroup S, the property of having finite rank cannot be expressed in terms of $\operatorname{Sub} S$. Any semigroup of finite breadth is of finite rank which is not greater than the breadth. The direct product of a countable set of cyclic groups which have distinct prime orders provides an example of a semigroup of rank 1 and of infinite breadth.

A semigroup S is said to be *narrow* [*of finite length* ℓ] if all antichains in the lattice $\operatorname{Sub} S$ are finite [all chains in $\operatorname{Sub} S$ are finite, their lengths are bounded by ℓ, and ℓ is the least number with this property]. A simple lattice-theoretic observation shows that a semigroup of finite length has finite breadth which is not greater than the length. A quasicyclic group provides an example of a semigroup of breadth 1 and infinite length. Any narrow semigroup is a Min-semigroup (it is easy to show) and has finite breadth (it follows from Theorems 4.26 and 4.22, taking into account Proposition 4.3).

A semigroup is said to be of *finite lattice dimension* d if $\operatorname{Sub} S$ has dimension d, which means that $\operatorname{Sub} S$ is embeddable in the direct product of d chains, and d is the least number with this property. Any semigroup of finite lattice dimension has finite breadth which is not greater than this dimension (this is a manifestation of a non-trivial lattice-theoretic fact). The direct product of two quasicyclic p-groups for the same p provides an example of a semigroup of breadth 2 which is not narrow and has infinite lattice dimension.

There are some other conditions of the same character. We mention only one of them here; not being included in our general scheme (namely, requirement (C) is not satisfied), it is closely connected with the conditions considered above. We will say that a semigroup S is of *finite width* w if $\operatorname{Sub} S$ has finite width w, which means that the cardinalities of all antichains in $\operatorname{Sub} S$ are bounded by w, and w is the least number with this property. Any semigroup of finite width is obviously narrow. The direct product of two quasicyclic groups for distinct primes provides an example of a narrow semigroup of infinite width.

4.18 Theorem (a) *Any semigroup of finite width is finitely assembled from groups of finite width.* (b) *A finitely assembled semigroup S in which at most one of the maximal subgroups is an infinite group of finite width, and the others are finite, is a semigroup of finite width.*

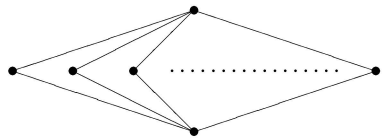
The necessary [sufficient] condition given by statement (a) [resp., (b)] is not sufficient [necessary]; the counter-examples are given in [108]. The gap between these two conditions seems to be rather narrow. This augments the interest in the following problem which was actually noted in [90] and formulated explicitly in [106, Part 1].

4.19 Problem *Find a necessary and sufficient condition describing the structure of semigroups of finite width.*

4.3 The case of groups

First of all, note that, for the subgroup lattice of a group, the minimal condition implies periodicity of the given group; the same takes place for finiteness of breadth, finiteness of dimension, the property of being narrow (and thereby for finiteness of length or width). Hence, for groups satisfying these conditions, there is no difference between subsemigroups and subgroups.

We start our considerations here by mentioning that there was the long-standing problem of whether any group of finite length is in fact finite. This problem was posed by I. Kaplansky in the forties, see [4, Problem 43]. Groups providing a negative solution have been constructed in [55–57], see also [58, Chapter 9]. In these groups every proper subgroup is of prime order, so the subgroup lattice of such a group has the following diagram.



Here the prime orders of proper subgroups can be either distinct or the same. In the latter case one obtains the answer to a question which was known as the Tarski's problem: Does there exist an infinite group whose proper subgroups have the same prime order? The groups constructed in the works mentioned above gave the solution of other problems open for a long time (see, for instance, comments at the beginning of §27 in [58], or comments on Question Γ4 in [33, §1]). The proofs of such results are based on fruitful geometric methods to study groups presented by defining relations, see [58].

Let us turn to the other conditions considered in Subsection 4.2. We will use the notation Min [resp., Max] for the minimal [maximal] condition for subsemigroups, Rn [resp., Br, Ln] for finiteness of rank [breadth, length], and Per for periodicity. We have the following implication, observed above: Ln $\Longrightarrow$ Min & Max & Br, Br $\Longrightarrow$ Rn $\Longrightarrow$ Per. There were open questions concerning more precise interrelations between these conditions. Several such questions were formulated in [89] (see also [105, Questions 51–55], and [33], where all the questions from [89] are reproduced). In [33] the examination of the interdependence of the listed conditions was completed: the questions have been solved negatively, even in their strongest form. The main result of [33] is the following theorem.

4.20 Theorem (a) *The conjunction* Min & Max & Br *does not imply* Ln. (b) *The conjunction* Min & Max & Rn *does not imply* Br. (c) *The conjunction* Min & Max *does not imply* Rn. (d) *The conjunction* Min & Br *does not imply* Max. (e) *The conjunction* Max & Br *does not imply* Min.

Previously, an example of a group constructed in [14] showed that Min does not imply Br. In [33] there are some related results as well. For instance, the problem discussed in [105, Question 56] was settled. That question was about the interrelation between Min and the property of having an at most countable subgroup lattice for periodic groups. It turned out that both implications are not true for the conditions in question, see Theorems 4 and 5 in [33] (the latter is taken from [54]). In contrast to the situation for arbitrary periodic groups, these two conditions as well as Min and Br are equivalent for locally finite groups, see Theorem 4.22 below. Recall that a group G is called *locally finite* (*l.f.*) if every finitely generated subgroup of G is finite.

A special type of l.f. groups is represented by so-called Chernikov groups. A group is said to be a *Chernikov group* if it is an extension of a direct product of finitely many quasicyclic groups by a finite group. It is well known that Chernikov groups satisfy the minimal condition for subgroups. One of the crucial results in this area is given by the following theorem proved independently in [38] and [111] (for a proof, see also [39]).

4.21 Theorem *A locally finite group satisfies the minimal condition for abelian subgroups if and only if it is a Chernikov group.*

Taking into account this theorem as well as some properties of groups of finite breadth established in [90], we obtain the equivalence of conditions (1), (2), and (4) of the next theorem. As to condition (3), it can be added in view of Theorem 13.9 of [108].

4.22 Theorem *For a locally finite group G, the following conditions are equivalent:*

(1) *G is a* Min*-group;*

(2) *G is a* Br*-group;*

(3) *the subgroup lattice of G is at most countable;*

(4) *G is a Chernikov group.*

Any Max-group is obviously *Noetherian*, i.e., satisfies the maximal condition for subgroups, but the former condition is stronger than the latter, even in the case of abelian groups, as the following statement shows.

4.23 Proposition *An infinite abelian group is a* Max*-group if and only if it is the direct product of the infinite cyclic group and a finite group.*

This statement can be generalized as follows, where the situation is completely transparent. Recall that, for some property θ, a group is said to be a *finite extension of a θ-group*, or *θ-by-finite*, if it has a normal subgroup of finite index which is a θ-group.

4.24 Theorem *Any soluble* Max*-group is necessarily cyclic-by-finite. Conversely, any cyclic-by-finite and, more generally, any* Max*-by-finite group is a* Max*-group.*

The assertions of this theorem for the case of soluble groups were proved in [107]. As to the converse assertion in the general case, it was noted in [107] as an open question. This question was reproduced also in [106, Part 1], and in [34], and it was soon settled in [59], that has been reflected in [108, Lemma 13.5.5].

The existence of the groups mentioned in Theorem 4.20 gives no hope of obtaining a satisfactory description of periodic Max-groups. As to non-periodic Max-groups, apparently there is also no hope of obtaining a transparent description: the types of such groups are too diverse, as the following statements show. These statements were proved in [34], see Theorems 1 and 2 there.

4.25 Theorem *There are continuously many non-isomorphic 2-generated torsion-free groups in each of which every maximal subsemigroup is a cyclic group, and distinct maximal subgroups intersect trivially. For any sufficiently big prime p (e.g., $p > 10^{80}$) there are continuously many non-isomorphic 2-generated non-periodic groups G such that the identity $x^p(x^py^p)^p = (x^py^p)^px^p$ holds in G; every maximal subsemigroup of G is a cyclic group either of infinite order or of order p, and distinct maximal subgroups of G intersect trivially.*

The structure of infinite narrow groups and infinite groups of finite width is described by the following statements proved in [92] and [90], respectively. One can see that such groups are automatically locally finite.

4.26 Theorem *An infinite group is narrow [of finite width] if and only if it is the direct product of a finite group and finitely many quasicyclic groups with distinct primes [a quasicyclic p-group for some prime p] not dividing the order of the finite factor.*

The next theorem was stated in [101].

4.27 Theorem *A locally finite group of finite lattice dimension is a finite extension of the direct product of finitely many quasicyclic groups for distinct primes.*

It seems quite plausible that the converse to this statement is true as well. But there is no proof by now, therefore one has to formulate the following

4.28 Problem *Does any group which is a finite extension of the direct product of finitely many quasicyclic groups for distinct primes have finite lattice dimension?*

4.4 On nilsemigroups

Being of independent interest, along with groups, as one of the most important types of unipotent epigroups, nilsemigroups occur in various semigroup-theoretic investigations. In particular, they appear as certain components of the structure of some epigroups examined; see relevant statements in Section 3. Remark that nilsemigroups could serve as the subject matter of a separate survey, and the author hopes to prepare such a survey in the future. Here we restrict our consideration only to information relating to some long-standing problems.

Note first that all the conditions considered in Subsection 4.2, as applied to nilsemigroups, are reduced to the property of simply being a finite semigroup. This may be regarded as an immediate consequence of the general result given by Theorem 4.7. But what actually happens is that the proof of this result (in essence, of Propositions 4.1 and 4.2) uses in one of the key steps the following fact about nilsemigroups established in [87].

4.29 Theorem *Any nilsemigroup whose nilpotent subsemigroups are finite is itself finite.*

Applications of this fact are based on the observation that, for a nilpotent semigroup T (which is not a singleton), the set $T \setminus T^2$ is its unique basis, and T is finite if and only if this basis is finite. Hence for many properties which are hereditary for subsemigroups, obvious arguments reduce questions discussed for nilsemigroups to their nilpotent subsemigroups.

It is easy to ascertain that a finite nilsemigroup is in fact nilpotent. Therefore the property of being a nilpotent semigroup may be regarded as the strongest non-trivial finiteness condition for the class Nil. There are diverse weaker such conditions; we will consider only a few of them.

A semigroup is called *locally nilpotent* (*l.n.*) if each of its finitely generated subsemigroups is nilpotent. Remark that if we speak about a nilpotent subsemigroup H of some semigroup S, then the zero of H does not have to be the zero of S, but it is obvious that any l.n. semigroup is unipotent and so is a nilsemigroup. Since a finitely generated nilpotent semigroup is finite, l.n. semigroups are precisely locally finite nilsemigroups. There exists an infinite 2-generated [3-generated] semigroup satisfying the identity $x^3 = 0$ [$x^2 = 0$]. Such nilsemigroups which are not l.n. can be easily constructed by using the famous so-called Thue-Morse words (see, for instance, [49, Section A3], or [45, §2.6]).

A proper subclass of the class of l.n. semigroups is formed by semigroups with the idealizer condition. Recall that the *idealizer* of a subsemigroup T in a semigroup S is the greatest subsemigroup in S containing T as an ideal. A semigroup S is said to satisfy the *idealizer condition*, or to be an *I-semigroup*, if any proper subsemigroup of S is distinct from its idealizer. A well ordered series $\{0\} = A_0 \subset A_1 \subset \cdots \subset A_\alpha \subset A_{\alpha+1} \subset \cdots \subset A_\beta = S$ of ideals of a semigroup S with zero (where, as usual, $A_\alpha = \bigcup_{\gamma<\alpha} A_\gamma$ for every limit ordinal α) is

called an *ascending annihilator series* (*a.a.s.*) of S if $SA_{\alpha+1} \cup A_{\alpha+1}S \subseteq A_\alpha$ for every $\alpha < \beta$. It is easy to see that every semigroup with an a.a.s. is an I-semigroup. In [86] it was shown that every I-semigroup is l.n. However, it is unknown whether the converse is true, so the following problem is still open.

4.30 Problem *Does the idealizer condition imply the existence of an ascending annihilator series in a given semigroup?*

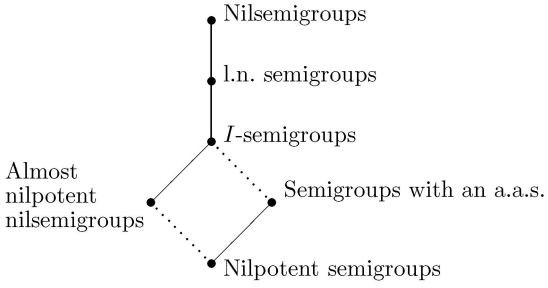
This problem arose in the early sixties and was explicitly formulated by the author later in [91, Problem 1.50a] in the following equivalent form: Does every I-semigroup have the non-trivial annihilator? Recall that the *annihilator* of a semigroup S with zero is the set $\{a \in S \,|\, ax = xa = 0 \text{ for each } x \in S\}$. The equivalence of these two formulations can be easily derived from the fact that a homomorphic image of an I-semigroup is again an I-semigroup.

Another open problem concerns semigroups all of whose proper subsemigroups are nilpotent. We call such a semigroup *almost nilpotent*. If H is a proper nilpotent subsemigroup of a semigroup S, and the zero of H is the zero of S, then H is distinct from its idealizer. Therefore every almost nilpotent nilsemigroup is an I-semigroup. As is shown in [88], almost nilpotent semigroups not being nilsemigroups are exhausted by 2-element bands and cyclic groups of prime order. For the nil case, we have the following problem.

4.31 Problem *Is any almost nilpotent nilsemigroup in fact nilpotent?*

This problem was first noted in [86] and, after that, in [88]. It was formulated later in [91, Problem 1.52], and from that time this intriguing problem attracted the attention of a number of researchers. There are results which give an affirmative answer under different additional restrictions on a semigroup; they concern Problem 4.30 as well. We mention particularly the paper [28], where, among other results, certain connections of Problems 4.30 and 4.31 with yet another problem relating to I-semigroups (namely, Problem 1.51a from [91]) was established. A survey of the results concerning the problems under discussion was given in [103]. As was noted in the work mentioned, these problems are connected: namely, a counter-example to Problem 4.31 will be at the same time a counter-example to Problem 4.30. In [103] some properties are pointed out which have to be possessed by an almost nilpotent but non-nilpotent nilsemigroup if such a semigroup exists. These properties show that a corresponding example should be rather freakish. An approach to produce such an example was also suggested in [103]; moreover, even a candidate for a counter-example was proposed there. However, in the plan of a hypothetical proof, a concrete place was indicated where the construction proposed might lead to a collapse. Such a collapse was, indeed, detected later in [112]. So, Problem 4.31, being open by now for more than 40 years, is waiting for new attempts to attack it.

The following diagram of inclusions shows the interrelations between the above-mentioned classes of semigroups.



Continuous lines correspond to strict inclusions, dotted ones mean that we have two unsolved problems here.

References

[1] J. Almeida, *Semigrupos Finitos e Álgebra Universal*, São Paulo, Universidade de São Paulo, 1992. (Portuguese; English translation: *Finite Semigroups and Universal Algebra*, Singapore, World Scientific, 1994.)

[2] L. Anderson, R. Hunter, and R. Koch, Some results on stability in semigroups, *Trans. Amer. Math. Soc.* **117** (1965), 521–529.

[3] K. Auinger and M. B. Szendrei, On identity bases of epigroup varieties, *J. Algebra* **220** (1999), 437–448.

[4] G. Birkhoff, *Lattice Theory*, Amer. Math. Soc., 2nd ed., Providence, New York, 1948.

[5] S. M. Bogdanović, *Semigroups with a System of Subsemigroups*, Novi Sad, University of Novi Sad, 1985.

[6] S. M. Bogdanović, Semigroups of Galbiati-Veronesi. I, II, in: *Proc. Conf. "Algebra and Logic", Zagreb*, 1984 (Z. Stojanović, ed.), Inst. Math. Fac. Sci. University of Novi Sad, Novi Sad, 1985, 9–20; *Facta Univ. Ser. Math. Inform.* **4** (1987), 61–66.

[7] S. M. Bogdanović and M. D. Ćirić, Semigroups of Galbiati-Veronesi. III, *Facta Univ. Ser. Math. Inform.* **4** (1989), 1–14.

[8] S. M. Bogdanović and M. D. Ćirić, Semilattices of archimedean semigroups and (completely) π-regular semigroups. I, *Filomat* (Niš) **7** (1993), 1–40.

[9] S. M. Bogdanović and M. D. Ćirić, *Polugrupe*, Niš, Prosveta, 1993. (Serbian; English preface and contents.)

[10] J. A. Chrislock, A certain class of identities on semigroups, *Proc. Amer. Math. Soc.* **21** (1969), 189–190.

[11] M. D. Ćirić and S. M. Bogdanović, Theory of greatest decompositions of semigroups (A survey), *Filomat* (Niš) **9** (1995), 385–426.

[12] A. H. Clifford, Semigroups admitting relative inverses, *Ann. Math.* **42** (2) (1941), 1037–1049.

[13] A. H. Clifford and G. B. Preston, *The Algebraic Theory of Semigroups*, Amer. Math. Soc., Providence, RI, Vol. 1, 1961 (4th ed., 1990); Vol. 2, 1967 (3rd ed., 1988).

[14] G. S. Deryabina, On infinite p-groups with cyclic subgroups, *Mat. Sb.* **124** (4) (1984), 425–504. (Russian; English translation: *Math. USSR. Sb.* **52** (1985), 481–490.)

[15] M. P. Drazin, Pseudo-inverses in associative rings and semigroups, *Amer. Math. Monthly* **65** (1958), 506–514.

[16] D. Easdown, Biordered sets of eventually regular semigroups, *Proc. London Math. Soc.* **49** (1984), 483–503.

[17] P. M. Edwards, Eventually regular semigroups, *Bull. Austral. Math. Soc.* **28** (1983), 23–38.

[18] P. M. Edwards, Eventually regular semigroups that are group-bound, *Bull. Austral. Math. Soc.* **34** (1986), 127–132.

[19] T. Evans, Embedding theorems for multiplicative systems and projective geometries, *Proc. Amer. Math. Soc.* **3** (1952), 614–620.

[20] T. Evans, The lattice of semigroup varieties, *Semigroup Forum* **2** (1971), 1–43.

[21] D. G. Fitz-Gerald, On inverses of products of idempotents in regular semigroups, *J. Austral. Math. Soc.* **13** (1972), 335–337.

[22] J. L. Galbiati and M. L. Veronesi, Sui semigruppi che sono un band di t-semigruppi, *Inst. Lombardo Accad. Sci. Lett. Rend.* **114** (1980), 217–234.

[23] J. L. Galbiati and M. L. Veronesi, On quasi completely regular semigroups, *Semigroup Forum* **29** (1984), 271–286.

[24] J. A. Gerhard, Free completely regular semigroups. Word problem, *J. Algebra* **82** (1983), 143–156.

[25] M. Hall, The word problem for semigroups with two generators, *J. Symb. Logic* **14** (1949), 115–119.

[26] T. E. Hall and W. D. Munn, Semigroups satisfying minimal conditions. II, *Glasgow Math. J.* **20** (1979), 133–140.

[27] P. M. Higgins, *Techniques of Semigroup Theory*, Oxford, Oxford University Press, 1992.

[28] I. L. Hmelnitsky, On semigroups with the idealizer condition, *Semigroup Forum* **32** (1985), 135–144.

[29] K. H. Hofmann and P. Mostert, *Elements of Compact Semigroups*, Columbus, OH, 1966.

[30] E. Hotzel, On semigroups with maximal conditions, *Semigroup Forum* **11** (1975–76), 337–362.

[31] E. Hotzel, On finiteness conditions in semigroups, *J. Algebra* **60** (1979), 352–370.

[32] J. M. Howie, Idempotent-generated semigroups: minimality results, *Prepr. and Lect. Notes Math.* (1979), 1–7.

[33] S. V. Ivanov, On some finiteness conditions in semigroup and group theory, *Semigroup Forum* **48** (1994), 28–36.

[34] S. V. Ivanov, On subsemigroup lattices of aperiodic groups, *Semigroup Forum* **48** (1994), 131–137.

[35] B. A. Jensen and O. W. Miller, Commutative semigroups which are almost finite, *Pacif. J. Math.* **27** (1968), 533–538.

[36] J. Kad'ourek and L. Polák, On the word problem for free completely regular semigroups, *Semigroup Forum* **34** (1986), 127–138.

[37] S. I. Katsman, Commutative semigroups with the maximal condition for subsemigroups, *XVII All-Union Algebra Conf. Abstracts of Reports. Part 2, Minsk, 1983*, 97. (Russian)

[38] O. H. Kegel and B. A. F. Wehrfritz, Strong finiteness conditions in locally finite groups, *Math. Z.* **117** (1970), 309–324.

[39] O. H. Kegel and B. A. F. Wehrfritz, *Locally Finite Groups*, North-Holland, Amsterdam, 1973.

[40] B. Kolibiarova, O kommutativnych perodickych pologrupach, *Mat.-Fyz. Časopis Slovensk. Akad. Vied.* **8** (1958), 127–135. (Slovak; German summary.)

[41] B. Kolibiarova, O čiastočne kommutativnych perodickych pologrupach, *Mat.-Fyz. Časopis Slovensk. Akad. Vied.* **9** (1959), 160–172. (Slovak; English summary.)

[42] I. B. Kozhukhov, On semigroups with minimal or maximal conditions on left congruences, *Semigroup Forum* **21** (1980), 337–350.

[43] A. P. do Lago and I. Simon, Free Burnside semigroups, *Theor. Informatics and Appl.* **35** (2001), 575–595.

[44] F. Levin, One variable equations over semigroups, *Bull. Austral. Math. Soc.* **2** (1970), 247–252.

[45] A. de Luca and S. Varricchio, *Finiteness and Regularity in Semigroups and Formal Languages*, Springer-Verlag, Berlin, Heidelberg, New York, 1999.

[46] B. L. Madison, T. K. Mukherjee, and M. K. Sen, Periodic properties of groupbound semigroups. I, II, *Semigroup Forum* **22** (1981), 225–234; **26** (1983), 229–236.

[47] D. B. McAlister and L. O'Carroll, Finitely generated commutative semigroups, *Glasgow Math.* **11** (1970), 134–151.

[48] R. N. McKenzie, Semigroups whose proper subsemigroups have less power, *Algebra Universalis* **1** (1971), 21–25.

[49] A. V. Mikhalev and G. F. Pilz, eds.,*The Concise Handbook of Algebra* Kluwer, Dordrecht, 2002.

[50] D. W. Miller, Some aspects of Green's relations on periodic semigroups, *Czechosl. Math. J.* **33** (1983), 537–544.

[51] W. D. Munn, Pseudo-inverses in semigroups, *Proc. Cambridge Phil. Soc.* **57** (1961), 247–250.

[52] K. S. S. Nambooripad, *Structure of Regular Semigroups. I*, Memoirs Amer. Math. Soc. **22**, 1979.

[53] B. H. Neumann, Embedding theorems for semigroups, *J. London Math. Soc.* **35** (1960), 184–192.

[54] S. V. Obraztsov, A theorem on injections of groups and its corollaries, *Mat. Sb.* **180** (1989), 529–541. (Russian; English translation: Russ. Acad. Sci. Sb. Math. **66** (1990), 541–553.)

[55] A. Yu. Ol'shanskii, Infinite groups with cyclic subgroups, *Dokl. Acad. Sci. USSR* **245** (1979), 785–787. (Russian; English translation: *Soviet Math. Dokl.* **20** (1979), 343–346.)

[56] A. Yu. Ol'shanskii, Infinite group with subgroups of prime orders, *Izv. Acad. Sci. USSR. Ser. Mat.* **44** (1980), 309–321. (Russian; English translation: *Math USSR. Izv.* **16** (1981), 272–289.)

[57] A. Yu. Ol'shanskii, Groups of bounded period with subgroups of prime orders, *Algebra i Logika* **21** (1982), 553–618. (Russian; English translation: Algebra and Logic **21** (1983) 369–418.)

[58] A. Yu. Ol'shanskii, *Geometry of Defining Relations in Groups*, Moscow, Nauka, 1989. (Russian; English translation: Kluwer, Dordrecht, 1991.)

[59] O. B. Paison, On groups with the maximal condition for subsemigroups, Manuscript, Ural State University, 1994. (Russian.)

[60] F. Pastijn, Embedding semigroups in semibands, *Semigroup Forum* **14** (1977), 247–263.

[61] F. Pastijn, The lattice of completely regular semigroup varieties, *J. Austral. Math. Soc. Ser. A* **49** (1990), 24–42.

[62] F. Pastijn, Commuting fully invariant congruences on free completely regular semigroups, *Trans. Amer. Math. Soc.* **323** (1991), 79–92.

[63] F. Pastijn, The idempotents in a periodic semigroup, *Int. J. Algebra Comput.* **6** (1996), 511–540.

[64] F. J. Pastijn and P. G. Trotter, Lattices of completely regular varieties, *Pacif. J. Math.* **119** (1985), 191–224.

[65] M. Petrich, *Introduction to Semigroups*, Merrill, Columbus, OH, 1973.

[66] M. Petrich, *Lectures in Semigroups*, Academic Press, Berlin, 1977.

[67] M. Petrich and N. R. Reilly, The modularity of the lattice of varieties of completely regular semigroups and related representations, *Glasgow Math. J.* **32** (1990), 137–152.

[68] M. Petrich and N. R. Reilly, *Completely Regular Semigroups*, John Wiley & Sons, New York, 1999.

[69] A. N. Petrov, Embedding theorems for countable periodic semigroups, *Ural. State Univ. Mat. Zap.* **14** (1) (1985), 128–140. (Russian.)

[70] A. N. Petrov and E. G. Tretyakova, Embedding theorems for countable quasiperiodic semigroups, *III All-Union Symp. on Theory of Semigroups. Abstracts of Reports*, Sverdlovsk, 1988, 72. (Russian.)

[71] A. S. Prosvirov, On periodic semigroups, *Ural. State Univ. Mat. Zap.* **8** (1) (1971), 77–94. (Russian.)

[72] A. S. Prosvirov, On periodic semigroups in which no torsion class is a subsemigroup, *II All-Union Symp. on Theory of Semigroups. Abstracts of Reports*, Sverdlovsk, 1978, 73. (Russian.)

[73] M. Putcha, Semigroups in which a power of each element lies in a subgroup, *Semigroup Forum* **5** (1973), 345–361.

[74] M. Putcha, Semilattice decompositions of semigroups, *Semigroup Forum* **6** (1973), 12–34.

[75] M. Putcha, Bands of t-archimedean semigroups, *Semigroup Forum* **6** (1973), 232–239.

[76] V. V. Rasin, On varieties of Cliffordian semigroups, *Semigroup Forum* **23** (1981), 201–220.

[77] D. Rees, On semi-groups, *Proc. Cambridge Phil. Soc.* **36** (1940), 387–400.

[78] N. R. Reilly, Minimal non-cryptic varieties of inverse semigroups, *Quart. J. Math. Oxford* **36** (2) (1985), 467–487.

[79] M. V. Sapir, Problems of Burnside type and the finite basis property in semigroup varieties, *Izv. Akad. Nauk SSSR. Ser. Matem.* **51** (1987), 319–340. (Russian; English translation: *Math. USSR. Izv.* **30** (1988), 295–314.)

[80] M. V. Sapir, Inherently not finitely based finite semigroups, *Math. Sb.* **133** (1987), 154–166. (Russian; English translation: *Math. USSR. Sb.* **61** (1988), 155–166.)

[81] M. V. Sapir and E. V. Sukhanov, On varieties of periodic semigroups, *Izv. VUZ. Matem.* **4** (1981), 48–55. (Russian; English translation: *Soviet Math. Izv. VUZ* **25** (4) (1981), 53–63.)

[82] M. P. Schützenberger, Sur le produit de concaténation non ambigu, *Semigroup Forum* **13** (1976), 47–76.

[83] R. Schwabauer, A note on commutative semigroups, *Proc. Amer. Math. Soc.* **20** (1969), 503–504.

[84] Š. Schwarz, Contribution to the theory of torsion semigroups, *Czechosl. Math. J.* **3** (1953), 7–21. (Russian.)

[85] J. T. Sedlock, Green's relations on periodic semigroups, *Czechosl. Math. J.* **19** (1969), 318–323.

[86] L. N. Shevrin, To the general theory of semigroups, *Mat. Sb.* **53** (3) (1961), 367–385. (Russian.)

[87] L. N. Shevrin, Nilsemigroups with certain finiteness conditions, *Mat. Sb.* **55** (4) (1961), 473–480. (Russian.)

[88] L. N. Shevrin, On semigroups in which all subsemigroups are nilpotent, *Sib. Mat. Zh.* **2** (1961), 936–942. (Russian.)

[89] L. N. Shevrin, Certain finiteness conditions in the theory of semigroups, *Izv. Acad. Sci. USSR. Ser. Mat.* **29** (1965), 553–566. (Russian.)

[90] L. N. Shevrin, Semigroups of finite breadth, *Theory of Semigroups and Its Applications* (Saratov) **1** (1965), 325–351. (Russian.)

[91] L. N. Shevrin, ed., *The Sverdlovsk Notebook: Unsolved Problems in the Theory of Semigroups*, Sverdlovsk, Ural State University, 1969; 2nd ed., 1979; 3rd ed., 1989. (Russian; English translation of the 1st ed.: *Semigroup Forum* **4** (1972), 274–280.)

[92] L. N. Shevrin, A supplement to the paper "Semigroups of finite breadth", *Theory of Semigroups and Its Applications* (Saratov) **2** (1970), 94–96. (Russian.)

[93] L. N. Shevrin, A general theorem on semigroups with some finiteness conditions, *Mat. Zametki* **15** (1974), 925–935. (Russian.)

[94] L. N. Shevrin, To the theory of periodic semigroups, *Izv. VUZ. Matem.* **5** (1974), 205–216. (Russian.)

[95] L. N. Shevrin, On the decomposition of a quasiperiodic semigroup into a band of archimedean semigroups, *XIV All-Union Algebra Conf. Abstracts of Reports. Part 1, Novosibirsk, 1977*, 104–105. (Russian.)

[96] L. N. Shevrin, Quasiperiodic semigroups possessing a partition into unipotent semigroups, *XVI All-Union Algebra Conf. Abstracts of Reports. Part 1, Leningrad, 1981*, 177–178. (Russian.)

[97] L. N. Shevrin, Quasiperiodic semigroups that are decomposable into a band of archimedean semigroups, *XVI All-Union Algebra Conf. Abstracts of Reports. Part 1, Leningrad, 1981*, 188. (Russian.)

[98] L. N. Shevrin, On the decomposition of quasiperiodic semigroups into a band, *XVII All-Union Algebra Conf. Abstracts of Reports. Part 2, Minsk, 1983*, 267–268. (Russian.)

[99] L. N. Shevrin, Finitely assembled semigroups, *III All-Union Symp. on Theory of Semigroups. Abstracts of Reports, Sverdlovsk, 1988*, 102. (Russian.)

[100] L. N. Shevrin, Epigroups with certain finiteness conditions, *Int. Conf. on Algebra in Memory of A. I. Mal'tsev. Abstracts of Reports on Theory of Models and Algebraic Systems, Novosibirsk, 1989*, 153. (Russian.)

[101] L. N. Shevrin, Groups whose subgroup lattice has finite dimension, *XI All-Union Symp. on Theory of Groups. Abstracts of Reports, Sverdlovsk, 1989*, 132. (Russian.)

[102] L. N. Shevrin, Semigroups, Chapter IV in: *General Algebra.* Vol. II (L. A. Skornyakov, ed.), Nauka, Moscow, 1991, 11–191. (Russian.)

[103] L. N. Shevrin, On two longstanding problems concerning nilsemigroups, in: *Semigroups with Applications* (J. M. Howie, W. D. Munn, and H. J. Weinert, eds.), World Scientific, Singapore, 1992, 222–235.

[104] L. N. Shevrin, To the theory of epigroups. I, II, *Mat. Sb.* **185** (8) (1994), 129–160; **185** (9) (1994), 153–176. (Russian; English translation: *Russ. Acad. Sci. Sb. Math.* **82** (1995), 485–512; **83** (1995), 133–154.)

[105] L. N. Shevrin and A. Ja. Ovsyannikov, Semigroups and their subsemigroup lattices, *Semigroup Forum* **27** (1983), 1–154.

[106] L. N. Shevrin and A. Ja. Ovsyannikov, *Semigroups and Their Subsemigroup Lattices*, Ural State University Press, Sverdlovsk, Part 1, 1990; Part 2, 1991. (Russian.)

[107] L. N. Shevrin and A. Ja. Ovsyannikov, Finitely assembled semigroups and ascending chain condition for subsemigroups, in: *Monash Conf. on Semigroup Theory in Honour of G. B. Preston* (T. E. Hall, P. R. Jones, and J. C. Meakin, eds.), World Scientific, Singapore, 1991, 269–284.

[108] L. N. Shevrin and A. Ja. Ovsyannikov, *Semigroups and Their Subsemigroup Lattices*, Kluwer, Dordrecht, 1996. (Revised and enlarged English translation of [106].)

[109] L. N. Shevrin and E. V. Sukhanov, Structural aspects of the theory of semigroup varieties, *Izv. VUZ. Matem.* **6** (1989), 3–39. (Russian; English translation: *Soviet Math. Izv. VUZ* **33** (6) (1989), 1–34.)

[110] L. N. Shevrin and M. V. Volkov, Identities of semigroups, *Izv. VUZ. Matem.* **11** (1985), 3–47. (Russian; English translation: *Soviet Math. Izv. VUZ* **29** (11) (1985), 1–64.)

[111] V. P. Shunkov, On locally finite groups with the maximal condition for abelian subgroups, *Algebra i Logika* **9** (1970), 579–615. (Russian.)

[112] A. N. Silkin, *On almost nilpotent nilsemigroups*, manuscript, Ural State University, 1993. (Russian.)

[113] A. V. Tishchenko, A remark on semigroup varieties of finite index, *Izv. VUZ. Matem.* **7** (1990), 79–83. (Russian; English translation: *Soviet Math. Izv. VUZ* **34** (7) (1990), 92–96.)

[114] A. V. Tishchenko and M. V. Volkov, A characterization of semigroup varieties of finite index in the language of "forbidden divisors", *Izv. VUZ. Matem.* **1** (1995), 91–99. (Russian; English translation: *Russ. Math. Izv. VUZ* **39** (1) (1995), 84–92.)

[115] E. G. Tretyakova, On quasiperiodic semigroups, manuscript, Ural State University, 1982. (Russian.)

[116] E. G. Tretyakova, On the free two-generated Clifford semigroup, *Int. Conf. on Algebra in Memory of A. I. Shirshov. Barnaul*, 1991. *Abstracts of Reports on Logic, Universal Algebras, and Applied Algebra, Novosibirsk, 1991*, 145. (Russian.)

[117] P. G. Trotter, Free completely regular semigroups, *Glasgow Math. J.* **25** (1984), 241–254.

[118] B. M. Vernikov, Semimodular and arguesian varieties of semigroups: forbidden subvarieties, *Proc. Ural State Univ.* **22** (*Matem., Mechan.*) (4) (2002), 16–42. (Russian.)

[119] B. M. Vernikov and M. V. Volkov, Semimodular and arguesian varieties of semigroups: a completion of description, *Proc. Ural State Univ.* **30** (*Matem., Mechan.*) (6) (2004), 5–36. (Russian.)

[120] M. V. Volkov, Semigroup varieties with modular subvariety lattices. I, II, III, *Izv. VUZ. Matem.* (6) (1989), 51–60; (7) (1992), 3–8; (8) (1992), 21–29. (Russian; English translation: *Soviet Math. Izv. VUZ* **33** (6) (1989), 48–58; *Russ. Math. Izv. VUZ* **36** (7) (1992), 1–6; **36** (8) (1992), 18–25.)

[121] M. V. Volkov, Semigroup varieties with modular subvariety lattices, *Dokl. Russ. Akad. Sci.* **326** (1992), 409–413. (Russian; English translation: *Russ. Acad. Sci. Dokl. Math.* **46** (1993), 274–278.)

[122] M. V. Volkov, *Identities in the Lattice of Semigroup Varieties*, Dr. Sci. dissertation, Ural State University, Ekaterinburg, 1994. (Russian.)

[123] M. V. Volkov, Covers in the lattices of semigroup varieties and pseudovarieties, in: *Semigroups, Automata and Languages* (J. Almeida, G. M. S. Gomes, and P. V. Silva, eds.), World Scientific, Singapore, 1996, 263–280.

[124] M. V. Volkov, "Forbidden divisor" characterizations of epigroups with certain properties of group elements, *RIMS Kokyuroku* **1166** (*Algebraic Systems, Formal Languages and Computations*) (2000), 226–234.

[125] M. V. Volkov, Semimodular and arguesian varieties of semigroups: identities, *Proc. Ural State Univ.* **22** (*Matem., Mechan.*) (4) (2002), 43–61. (Russian.)

[126] M. V. Volkov and T. A. Ershova, The lattice of varieties of semigroups with completely regular square, in: *Monash Conf. on Semigroup Theory in Honour of G. B. Preston* (T. E. Hall, P. R. Jones, and J. C. Meakin, eds.), World Scientific, Singapore, 1991, 306–322.

[127] M. V. Volkov and M. V. Sapir, HFB property and structure of semigroups, *Contrib. General Algebra* **6** (1988), 303–310.

[128] M. Yamada, A remark on periodic semigroups, *Bull. Shimane Univ.* **9** (1959), 1–5.

[129] I. Yu. Zhil'tsov, On epigroup identities, *Dokl. Russ. Acad. Sci.* **375** (2000), 10–12. (Russian; English translation: *Russ. Acad. Sci. Dokl. Math.* **62** (2000), 322–324.)

Algebraic classifications of regular tree languages

Magnus STEINBY

Department of Mathematics
and
Turku Centre for Computer Science
University of Turku
FIN-20014 Turku
Finland

Abstract

We review several algebraic formalisms used for characterizing families of regular tree languages. A theory based on syntactic algebras and varieties of congruences is presented in detail, including the Variety Theorem that establishes the correspondences between varieties of tree languages, varieties of finite algebras and varieties of congruences on the finitely generated term algebras. Several examples of varieties of tree languages are given. Also generalizations of this theory as well as some alternative approaches are considered.

1 Introduction

In 1976 Samuel Eilenberg [18] presented his fundamental Variety Theorem that establishes a bijective correspondence between varieties of finite monoids and certain families of regular languages that he called $*$-varieties, or alternatively, between varieties of finite semigroups and so-called $+$-varieties consisting of regular languages that do not include the empty word. The variety theorem was preceded, and inspired, by some remarkable results about certain individual families of regular languages. First of all, M. P. Schützenberger [77] had shown in 1965 that a regular language is star-free exactly in case its syntactic monoid is aperiodic, that is to say, all of its subgroups are trivial. This theorem gave a decision method for an important family of languages that arises in many seemingly independent ways. In the early 1970's several other families of regular languages had been similarly characterized. By describing exactly the families of regular languages that can be characterized this way by syntactic monoids or syntactic semigroups, the variety theorem has provided a general framework for the classification of regular languages that has retained its place till this day. Of course, there have also been some important amendments to the variety theory itself, such as Thérien's varieties of congruences [86, 87] and Pin's theory of positive varieties [64]. For the theory of string language varieties the references [2, 18, 37, 45, 63, 64] are recommended.

Algebraic classifications of regular string languages are almost exclusively based on semigroup theory via syntactic monoids or semigroups (although a classification corresponding to varieties of finite unary algebras is also possible, cf. [83, 57, 58, 59]). It appears clear that syntactic monoids or semigroups are not equally well suited for describing properties of tree languages. In fact, we shall see that many natural families of regular tree languages cannot

V. B. Kudryavtsev and I. G. Rosenberg (eds.),
Structural Theory of Automata, Semigroups, and Universal Algebra, 381–432.

be characterized by these syntactic invariants, and that there are more promising approaches. In [78, 80] and [1] the syntactic algebra of a subset of any algebra is defined in such a way that the syntactic algebra of a string language L over an alphabet X becomes its syntactic monoid when L is viewed as a subset of the free monoid X^*. Similarly, if L does not contain the empty word, then its syntactic algebra as a subset of the free semigroup X^+ is its usual syntactic semigroup. For a tree language T, defined as a set of terms over a finite set Σ of operation symbols, i.e., a ranked alphabet, and a finite set X of variables, the syntactic algebra is a Σ-algebra that is finite exactly in case T is a regular tree language. Using syntactic algebras it is possible to prove a general variety theorem that establishes a correspondence between certain families of recognizable subsets (as defined in [52]) of finitely generated free algebras over a given variety $\mathbf{V}$ of Σ-algebras and varieties of finite Σ-algebras. For the varieties of monoids and semigroups, the two cases of Eilenberg's variety theorem are obtained. On the other hand, if $\mathbf{V}$ is taken to be the class of all Σ-algebras, the free algebras are term algebras and we obtain a variety theorem that establishes a bijection between varieties of tree languages and varieties of finite Σ-algebras. It turns out that a considerable part of the known families of special regular tree languages are such varieties of tree languages.

As shown in [1] and [80], the variety theorem can be completed with varieties of congruences also for tree languages. The theory can also be generalized in various ways. Firstly, in [84] a variety theory that does not assume a fixed ranked alphabet is presented. Thus we obtain generalized varieties of tree languages that include regular tree languages over any pair of alphabets Σ and X. On the other hand, in [73] it is shown how the theory can be formulated for many-sorted tree languages. Then there is naturally the above mentioned general formulations of [1] and [78] in which tree languages are replaced by recognizable subsets of free algebras over a given variety. Moreover, there are several alternative approaches, a fact that once again shows that generalizations from strings to trees are not always obvious or unique. The natural idea of defining syntactic monoids and syntactic semigroups for tree languages was realized by Thomas [88, 89] who also characterized non-counting tree languages by their syntactic monoids. However, we shall see that syntactic monoids or semigroups are of limited use in the case of tree languages. In [22] Ésik proposes another alternative framework in which tree languages are classified by their syntactic algebraic theories—here the 'algebraic theories' are the special categories introduced by Lawvere (1963; cf. [20] or [22] for the reference). In Wilke's [92] approach the syntactic invariant used is the syntactic tree algebra of a tree language. This is a 3-sorted algebraic structure that in some sense embodies the syntactic algebra as well as the syntactic monoid, or the syntactic semigroup, of the tree language. The theory differs from those mention before in that trees are not defined directly as terms but they are represented by terms over an 'external' 3-sorted alphabet. The type of the syntactic tree algebra does not depend on the label alphabet. Instead, the label alphabet serves as a generating set of the free tree algebra in the same way as the alphabets of string languages generate the free monoids. However, the formalism is restricted to binary trees.

The theory of finite tree automata and regular tree languages can be formulated as an algebraic generalization of the theory of finite automata and regular languages in such a way that Universal Algebra becomes the natural mathematical framework for it. In Section 2 we recall some basic algebraic notions and establish the corresponding notation to be used. Section 3 is then a brief introduction to finite automata, regular languages, and Eilenberg's variety theory. It should facilitate the understanding of the corresponding tree theory and also put the subject matter of this paper into a proper perspective. In Section 4 it is shown

how *trees* are defined as terms. The next two sections introduce the basic notions of the theory of finite tree automata and regular tree languages. In Section 5 *tree recognizers* are defined as finite algebras equipped with initial state assignments and sets of final states, and then we present some algebraic characterizations of the tree languages recognized by them, the *regular tree languages*. Section 6 introduces several operations on tree languages and notes the closure properties of the family of regular tree languages under them.

Sections 7 to 12 form the core part of the paper. The presentation is based to a great extent on reference [80], but it has been updated and extended in many ways. On the other hand, some technical proofs are omitted. Section 7 introduces *varieties of tree languages*, the families of regular tree languages to which the variety theory applies. In Section 8 *syntactic congruences* and *syntactic algebras* are considered. This can be done for subsets of arbitrary algebras, not just tree languages, and then the results become immediately available in many other similar theories. In Section 9 we consider *varieties of finite algebras*, also called *pseudovarieties*. These have been studied quite extensively in Universal Algebra independently of tree languages, and some references to such work are given. *Varieties of finite congruences* are introduced in Section 10. These are certain systems of filters of the congruence lattices of finitely generated term algebras that are very convenient for defining varieties of tree languages. In Section 11 all these ingredients are brought together to yield the *Variety Theorem* that establishes bijective correspondences between varieties of tree languages, varieties of finite algebras and varieties of congruences. In Section 12 we consider several examples of varieties of tree languages. Most of these are defined through the corresponding varieties of congruences, and in some cases also the corresponding variety of finite algebras can be given. Many of these families have also been studied in connection with the problem of inferring tree languages from samples (cf. [43] and [46], for example).

The next two sections are devoted to various generalizations and alternative approaches. Hence, in Section 13 we consider *varieties of recognizable subsets* of [78] and [1], the theory of *many-sorted varieties* developed in [73], and the theory of *generalized varieties of tree languages* (*GVTLs*) of [84]. We also mention the theory of *positive varieties of tree languages* recently put forward in [60]. Section 14 begins with a discussion of *syntactic monoids* of tree languages and a couple of examples of their use, but we will also note the limitations of the approach. In particular, we note a recent important result by Salehi [71] that characterizes the GVTLs that can be defined by syntactic monoids or syntactic semigroups. Then we note the *syntactic theory* approach of [22], mentioned already above, the related work on *syntactic pre-clones* by Ésik and Weil [23], as well as the *tree algebras* of [92].

In the last section, Section 15, we consider tree languages accepted by *deterministic top-down tree recognizers*, called *DT-recognizable tree languages* for short. These form a proper subfamily of the family of all regular tree languages [47]. It is not a variety of tree languages, but its Boolean closure is (as noted in [38, 39]). However, most of the section is devoted to the *syntactic path monoids* and *syntactic path semigroups* introduced by Gécseg and Steinby [30]. These are obtained from the string languages that describe the labelled paths leading from the root to a leaf with a given label in a tree belonging to the given tree language—there is such a language for each leaf symbol. It seems possible that syntactic path monoids or semigroups are suitable for characterizing families of DT-recognizable tree languages because DT-recognizable tree languages are completely determined by their path languages (cf. [12, 90, 28, 39, 29]).

The references are naturally an essential part of a paper like this, and hopefully all the most important works are mentioned.

2 Algebraic Preliminaries

Let us begin with some general notation. First of all, the symbol := is sometimes used in definitions. Thus we write $A := B$ to indicate that A is defined to be equal to B. The basic set theoretical symbols $\cap$, $\cup$, $\subseteq$, $\subset$, ... have their usual meanings. The power set of a set U is written as $\wp U$. A family $\mathcal{F} \subseteq \wp U$ of subsets of U is a *field of sets* on U if (1) $\emptyset, U \in \mathcal{F}$, and (2) $S \cap T, S \cup T, S - T \in \mathcal{F}$ whenever $S, T \in \mathcal{F}$. The *Boolean closure* of a set $\mathcal{S} \subseteq \wp U$ of subsets of U is the least field of sets $\mathcal{F}$ on U such that $\mathcal{S} \subseteq \mathcal{F}$. A *Boolean subalgebra* of a field of sets $\mathcal{F} \subseteq \wp U$ is any subfamily $\mathcal{B} \subseteq \mathcal{F}$ of $\mathcal{F}$ that also forms a field of sets on U.

Let $\theta \subseteq A \times A$ be a relation on a set A. The fact that $(a, b) \in \theta$, for some $a, b \in A$, is often expressed by writing $a \,\theta\, b$. The *converse* of θ is the relation $\theta^{-1} = \{(b, a) \mid (a, b) \in \theta\}$ on A. The *product* of two relations θ and ρ on A is the relation $\theta \circ \rho = \{(a, c) \mid (\exists b \in A)\, a \,\theta\, b,\ b \,\rho\, c\}$ on A. The *diagonal relation* $\{(a, a) \mid a \in A\}$ and the *universal relation* $A \times A$ are denoted by Δ_A and ∇_A, respectively. A relation $\theta \subseteq A \times A$ is an *equivalence relation*, or simply an *equivalence*, on A if $\Delta_A \subseteq \theta$, $\theta^{-1} \subseteq \theta$ and $\theta \circ \theta \subseteq \theta$. The set of all equivalences on A is denoted by $\mathrm{Eq}(A)$. The *restriction* of an equivalence $\theta \in \mathrm{Eq}(A)$ to a subset $B \subseteq A$ is the equivalence $\theta_B := \theta \cap (B \times B)$ on B. For $\theta \in \mathrm{Eq}(A)$, the *quotient set* A/θ is the set $\{a/\theta \mid a \in A\}$, where $a/\theta = \{b \in A \mid a \,\theta\, b\}$ is the θ-class of $a \in A$. If A/θ is finite, θ is said to be of *finite index*, or simply to be *finite*. An equivalence $\theta \in \mathrm{Eq}(A)$ is said to *saturate* a subset H of A if H is the union of some θ-classes.

A mapping $\varphi : A \to B$ may also be viewed as a special relation between A and B. Often we write $a\varphi$ for the image $\varphi(a)$ of an element $a \in A$. This notation is used especially when φ is a homomorphism. Furthermore, for any $H \subseteq A$ and any $K \subseteq B$, we write $H\varphi$ for $\varphi(H) := \{a\varphi \mid a \in H\}$ and $K\varphi^{-1}$ for $\varphi^{-1}(K) := \{a \in A \mid a\varphi \in K\}$. If $\theta \in \mathrm{Eq}(B)$ is an equivalence on B, then $\varphi \circ \theta \circ \varphi^{-1}$ is the equivalence $\{(a_1, a_2) \in A \times A \mid a_1\varphi \,\theta\, a_2\varphi\}$ on A.

All lattice theory needed in the theories to be considered, can be found in any standard text on universal algebra, such as [9] or [32], for example. Let us just recall some notions to be used here. Partial orders will be called simply *orders*. Hence an *ordered set* is pair $(A, \leq)$ consisting of a non-empty set A and a relation $\leq$ on A that is reflexive, antisymmetric and transitive. A *lattice* is an ordered set $(A, \leq)$ such that any two elements $a, b \in A$ have a least upper bound, the *join* $a \vee b$, and a greatest lower bound, the *meet* $a \wedge b$. A *complete lattice* is an ordered set $(A, \leq)$ such that the least upper bound $\sup H$ and the greatest lower bound $\inf H$ exist for every $H \subseteq A$. It is well known that an ordered set is a complete lattice if all greatest lower bounds exist.

Some of the ordered sets to be considered here are *algebraic lattices* of the following special kind. Let $\mathcal{S} \subseteq \wp U$ be a set of subsets of a universe U. Then $(\mathcal{S}, \subseteq)$ is an algebraic lattice if $\bigcap C \in \mathcal{S}$ for all $\mathcal{C} \subseteq \mathcal{S}$, and $\bigcup \mathcal{D} \in \mathcal{S}$ for every directed $\mathcal{D} \subseteq \mathcal{S}$; $\mathcal{D}$ is *directed* if for any two sets $X, Y \in \mathcal{D}$, there is a set $Z \in \mathcal{D}$ such that $X, Y \subseteq Z$. Of course, any algebraic lattice is a complete lattice.

Finally, let us recall that a *filter* of a lattice $(A, \leq)$ is a non-empty subset F of A such that (1) $a \leq b$ and $a \in F$ imply $b \in F$, and (2) $a \wedge b \in F$ whenever $a, b \in F$. The *filter generated* by a non-empty subset $H \subseteq A$, i.e., the least filter $[H)$ containing H, can easily be shown to be the set $\{a \in A \mid (\exists n > 0)\, (\exists b_1, \ldots, b_n \in H)\, b_1 \wedge \ldots \wedge b_n \leq a\}$.

The tree automata to be considered here are essentially finite algebras of finite type. Therefore we recall some basic notions of universal algebra. For systematic introductions to the subject we refer the reader to any of the books [2, 9, 10, 32, 91], for example.

Let Σ be a set of *operation symbols*. Hence each symbol in Σ has a unique non-negative integer *arity*. For each $m \geq 0$, Σ_m denotes the set of m-ary symbols in Σ. A Σ-*algebra* $\mathcal{A}$ consists of a non-empty set A and a Σ-indexed family of operations $f^{\mathcal{A}}$ on A such that if $f \in \Sigma_m$ $(m > 0)$, then $f^{\mathcal{A}} : A^m \to A$ is an m-ary operation, and any $c \in \Sigma_0$ is realized as a constant $c^{\mathcal{A}} \in A$. We write $\mathcal{A} = (A, \Sigma)$, and call $\mathcal{A}$ *finite* if A is a finite set. A one-element Σ-algebra is said to be *trivial*.

A Σ-algebra $\mathcal{B} = (B, \Sigma)$ is a *subalgebra* of a Σ-algebra $\mathcal{A} = (A, \Sigma)$ if $B \subseteq A$ and every operation $f^{\mathcal{B}}$ of $\mathcal{B}$ is the restriction to B of the corresponding operation $f^{\mathcal{A}}$. In particular, $c^{\mathcal{B}} = c^{\mathcal{A}}$ for every $c \in \Sigma_0$. This also means that B is a *closed subset* of $\mathcal{A}$, i.e., $f^{\mathcal{A}}(b_1, \ldots, b_m) \in B$ whenever $m > 0$, $f \in \Sigma_m$ and $b_1, \ldots, b_m \in B$, and $c^{\mathcal{A}} \in B$ for every $c \in \Sigma_0$. Since there is a natural bijection between the subalgebras and the non-empty closed subsets of $\mathcal{A}$, non-empty closed subsets are also called subalgebras. Let $\mathrm{Sub}(\mathcal{A})$ be the set of all closed subsets of $\mathcal{A}$. The intersection of any set of closed subsets is again closed. Hence, any subset H of A is contained in a unique minimal closed subset $\langle H \rangle$ that, when non-empty, is called the *subalgebra generated* by H. Obviously, $\langle H \rangle \neq \emptyset$ if $H \neq \emptyset$, and $\langle \emptyset \rangle = \emptyset$ exactly in case $\Sigma_0 = \emptyset$.

A mapping $\varphi : A \to B$ is a *homomorphism* from $\mathcal{A} = (A, \Sigma)$ to $\mathcal{B} = (B, \Sigma)$, expressed by writing $\varphi : \mathcal{A} \to \mathcal{B}$, if $f^{\mathcal{A}}(a_1, \ldots, a_m)\varphi = f^{\mathcal{B}}(a_1\varphi, \ldots, a_m\varphi)$ for all $m > 0$, $f \in \Sigma_m$ and $a_1, \ldots, a_m \in A$, and $c^{\mathcal{A}}\varphi = c^{\mathcal{B}}$ for every $c \in \Sigma_0$. A homomorphism is called an *epimorphism* if it is surjective, a *monomorphism* if it is injective, and an *isomorphism* if it is bijective. The algebras $\mathcal{A}$ and $\mathcal{B}$ are *isomorphic*, $\mathcal{A} \cong \mathcal{B}$ in symbols, if there is an isomorphism $\varphi : \mathcal{A} \to \mathcal{B}$. If there is an epimorphism $\varphi : \mathcal{A} \to \mathcal{B}$, then $\mathcal{B}$ is said to be an *(epimorphic) image* of $\mathcal{A}$.

Let $\mathcal{A}$ and $\mathcal{B}$ be Σ-algebras. If $\mathcal{A}$ is an image of some subalgebra of $\mathcal{B}$, we write $\mathcal{A} \preceq \mathcal{B}$ and say that $\mathcal{A}$ is *covered* by $\mathcal{B}$, or that $\mathcal{A}$ *divides* $\mathcal{B}$. It is clear that for any Σ-algebras $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$, (1) $\mathcal{A} \preceq \mathcal{A}$ and (2) $\mathcal{A} \preceq \mathcal{B}$ and $\mathcal{B} \preceq \mathcal{C}$ imply $\mathcal{A} \preceq \mathcal{C}$. Moreover, if $\mathcal{A}$ and $\mathcal{B}$ are finite, then (3) $\mathcal{A} \preceq \mathcal{B}$ and $\mathcal{B} \preceq \mathcal{A}$ if and only if $\mathcal{A} \cong \mathcal{B}$. Of course, $\mathcal{A} \preceq \mathcal{B}$ whenever $\mathcal{A}$ is isomorphic to a subalgebra of $\mathcal{B}$ or an image of $\mathcal{B}$. Hence the $\preceq$-relation may be used as a common generalization of the subalgebra and image relations.

A *congruence* on $\mathcal{A} = (A, \Sigma)$ is an equivalence relation $\theta \subseteq A \times A$ such that for any $m > 0$, $f \in \Sigma_m$ and $a_1, \ldots, a_m, b_1, \ldots, b_m \in A$,

$$a_1 \,\theta\, b_1, \ldots, a_m \,\theta\, b_m \quad \Rightarrow \quad f^{\mathcal{A}}(a_1, \ldots, a_m) \,\theta\, f^{\mathcal{A}}(b_1, \ldots, b_m).$$

Let $\mathrm{Con}(\mathcal{A})$ denote the set of all congruences on $\mathcal{A} = (A, \Sigma)$. It includes at least the diagonal relation Δ_A and the universal relation ∇_A. For any $\theta \in \mathrm{Con}(\mathcal{A})$, the *quotient algebra* $\mathcal{A}/\theta = (A/\theta, \Sigma)$ is defined by setting

$$f^{\mathcal{A}/\theta}(a_1/\theta, \ldots, a_m/\theta) = f^{\mathcal{A}}(a_1, \ldots, a_m)/\theta$$

for all $m > 0, f \in \Sigma_m$ and $a_1, \ldots, a_m \in A$, and $c^{\mathcal{A}/\theta} = c^{\mathcal{A}}/\theta$ for each $c \in \Sigma_0$. The operations of $\mathcal{A}/\theta$ are well-defined precisely because $\theta \in \mathrm{Con}(\mathcal{A})$, and the natural mapping

$$\theta^{\natural} : A \to A/\theta,\ a \mapsto a/\theta,$$

is an epimorphism from $\mathcal{A}$ onto $\mathcal{A}/\theta$; it is called the *natural homomorphism*. Futhermore, the *kernel* $\ker\varphi = \{(a, a') \in A \times A \mid a\varphi = a'\varphi\}$ of any homomorphism $\varphi : \mathcal{A} \to \mathcal{B}$ is a congruence on $\mathcal{A}$. Moreover, if φ is an epimorphism, then $\mathcal{A}/\ker\varphi \cong \mathcal{B}$.

Let us recall that a *right congruence* on a semigroup S is an equivalence θ on S such that, for any $a, b, c \in S$, $a\,\theta\,b$ implies $ac\,\theta\,bc$. Left congruences are defined similarly, and it is clear that an equivalence on a semigroup is a congruence if and only if it is both a left and a right congruence.

A mapping $p : A \to A$ is an *elementary translation* of an algebra $\mathcal{A} = (A, \Sigma)$ if

$$p(\xi) = f^{\mathcal{A}}(a_1, \ldots, a_{i-1}, \xi, a_{i+1}, \ldots, a_m),$$

for some $m > 0$, $f \in \Sigma_m$, $1 \leq i \leq m$ and $a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_m \in A$ (and where ξ is a variable ranging over A). Let $\mathrm{ETr}(\mathcal{A})$ denote the set of all elementary translations of $\mathcal{A}$. The set $\mathrm{Tr}(\mathcal{A})$ of all *translations* of $\mathcal{A}$ is the smallest set of unary operations on A that contains the identity map $1_A : A \to A$, $a \mapsto a$, and all elementary translations, and is closed under composition. The following useful fact is easy to verify.

2.1 Lemma *Any congruence θ on a Σ-algebra $\mathcal{A} = (A, \Sigma)$ is invariant with respect to every translation $p \in \mathrm{Tr}(\mathcal{A})$ of $\mathcal{A}$, i.e., $a\,\theta\,b$ implies $p(a)\,\theta\,p(b)$. In addition, any equivalence on A invariant with respect to all elementary translations of $\mathcal{A}$, is a congruence on $\mathcal{A}$.*

The *direct product* $\prod_{i\in I} \mathcal{A}_i = (\prod_{i\in I} A_i, \Sigma)$ of an indexed family $(\mathcal{A}_i \mid i \in I)$ of Σ-algebras $\mathcal{A}_i = (A_i, \Sigma)$ is defined by setting for every $j \in I$,

$$f^{\prod_{i\in I} \mathcal{A}_i}(a_1, \ldots, a_m)(j) = f^{\mathcal{A}_j}(a_1(j), \ldots a_m(j)),$$

for any $m > 0$, $f \in \Sigma_m$, $a_1, \ldots, a_m \in \prod_{i\in I} A_i$, and $c^{\prod_{i\in I} \mathcal{A}_i}(j) = c^{\mathcal{A}_j}$ for every $c \in \Sigma_0$; recall that an element of $\prod_{i\in I} A_i$ is a mapping $a : I \to \bigcup_{i\in I} A_i$ such that $a(i) \in A_i$ for every $i \in I$. For each $j \in I$, the j^{th} *projection* $\pi_j : \prod_{i\in I} A_i \to A_j$, $a \mapsto a(j)$, is easily seen to be an epimorphism from $\prod_{i\in I} \mathcal{A}_i$ onto $\mathcal{A}_j$. A subalgebra $\mathcal{B} = (B, \Sigma)$ of $\prod_{i\in I} \mathcal{A}_i$ is a *subdirect product* of the algebras $\mathcal{A}_i = (A_i, \Sigma)$ $(i \in I)$ if $B\pi_j = A_j$ for every $j \in I$, and such a subdirect product is *proper* if no π_j is injective, that is to say the restriction of π_j to B does not define an isomorphism between $\mathcal{B}$ and $\mathcal{A}_j$. A monomorphism $\varphi : \mathcal{A} \to \prod_{i\in I} \mathcal{A}_i$ is a (*proper*) *subdirect decomposition* of $\mathcal{A} = (A, \Sigma)$ if $A\varphi$ is a (proper) subdirect product. An algebra is *subdirectly irreducible* if it has no proper subdirect decomposition. It is well-known that a non-trivial algebra $\mathcal{A} = (A, \Sigma)$ is subdirectly irreducible if and only if it has a unique minimal non-diagonal congruence, that is to say, if and only if $\bigcap\{\theta \in \mathrm{Con}(\mathcal{A}) \mid \theta \supset \Delta_A\} \supset \Delta_A$.

As usual, we write $\mathcal{A}_1 \times \cdots \times \mathcal{A}_n$ for $\prod_{i\in I} \mathcal{A}_i$ when $I = \{1, \ldots, n\}$ is finite, and for $n = 0$ this product is taken to be a trivial Σ-algebra (that is unique up to isomorphism).

For any class $\mathbf{K}$ of Σ-algebras,

(1) $\mathrm{S}(\mathbf{K})$ is the class of all algebras isomorphic to a subalgebra of some algebra in $\mathbf{K}$,

(2) $\mathrm{H}(\mathbf{K})$ is the class of all images of members of $\mathbf{K}$, and

(3) $\mathrm{P}(\mathbf{K})$ is the class of all algebras isomorphic to the direct product of members of $\mathbf{K}$.

A class $\mathbf{K}$ of Σ-algebras is a *variety* if $\mathrm{S}(\mathbf{K}), \mathrm{H}(\mathbf{K}), \mathrm{P}(\mathbf{K}) \subseteq \mathbf{K}$. The intersection of all varieties of Σ-algebras containing a given class $\mathbf{K}$ of Σ-algebras, the *variety generated* by $\mathbf{K}$, is denoted by $\mathrm{V}(\mathbf{K})$. It is well known that $\mathrm{V}(\mathbf{K}) = \mathrm{HSP}(\mathbf{K})$ $(= \mathrm{H}(\mathrm{S}(\mathrm{P}(\mathbf{K}))))$. Moreover, any variety is generated by its subdirectly irreducible members.

If Σ is a set of operation symbols and X a set of symbols disjoint from Σ, then the set $T_\Sigma(X)$ of Σ-*terms with variables in* X is the smallest set T such that

(1) $X \cup \Sigma_0 \subseteq T$, and

(2) $f(t_1, \ldots, t_m) \in T$ whenever $m > 0$, $f \in \Sigma_m$ and $t_1, \ldots, t_m \in T$.

Clearly, $T_\Sigma(X) \neq \emptyset$ if and only if $\Sigma_0 \cup X \neq \emptyset$, and in that case the ΣX-*term algebra* $\mathcal{T}_\Sigma(X) = (T_\Sigma(X), \Sigma)$ is defined by setting $c^{\mathcal{T}_\Sigma(X)} = c$ for $c \in \Sigma_0$, and $f^{\mathcal{T}_\Sigma(X)}(t_1, \ldots, t_m) = f(t_1, \ldots, t_m)$ for any $m \geq 1$, $f \in \Sigma_m$ and $t_1, \ldots, t_m \in T_\Sigma(X)$.

The term algebra $\mathcal{T}_\Sigma(X)$ is *freely generated* by X over the class of all Σ-algebras, i.e.,

(1) $\mathcal{T}_\Sigma(X)$ is generated by X, $\langle X \rangle = T_\Sigma(X)$, and

(2) for any Σ-algebra $\mathcal{A} = (A, \Sigma)$, every mapping $\alpha : X \to A$ has a unique extension to a homomorphism $\alpha_\mathcal{A} : \mathcal{T}_\Sigma(X) \to \mathcal{A}$.

The requirement that $\alpha_\mathcal{A} : \mathcal{T}_\Sigma(X) \to \mathcal{A}$ be a homomorphism such that $x\alpha_\mathcal{A} = \alpha(x)$ for every $x \in X$, gives inductively a unique value $t\alpha_\mathcal{A}$ to every $t \in T_\Sigma(X)$:

- $x\alpha_\mathcal{A} = \alpha(x)$ for any $x \in X$;
- $c\alpha_\mathcal{A} = c^{\mathcal{T}_\Sigma(X)}\alpha_\mathcal{A} = c^\mathcal{A}$ for any $c \in \Sigma_0$;
- $t\alpha_\mathcal{A} = f^{\mathcal{T}_\Sigma(X)}(t_1, \ldots, t_m)\alpha_\mathcal{A} = f^\mathcal{A}(t_1\alpha_\mathcal{A}, \ldots, t_m\alpha_\mathcal{A})$, for $t = f(t_1, \ldots, t_m)$.

The *term function* $t^\mathcal{A} : A^X \to A$ defined by a ΣX-term $t \in T_\Sigma(X)$ is now obtained by setting $t^\mathcal{A}(\alpha) = t\alpha_\mathcal{A}$ for any $\alpha : X \to A$ (i.e., $\alpha \in A^X$). For example, if $t = f(g(x), c)$, then $t^\mathcal{A}(\alpha) = t\alpha_\mathcal{A} = f^\mathcal{A}(g^\mathcal{A}(\alpha(x)), c^\mathcal{A})$.

3 Finite automata and regular languages

In this section we review some relevant parts of the theory of finite automata and regular languages. Proofs and further results can be found in [6, 17, 37, 44, 94], for example.

In what follows, an *alphabet* is a finite non-empty set of symbols called *letters*. If X is an alphabet, then X^* denotes the set of all (finite) *words* over X, e is the *empty word*, and X^+ is the set of non-empty words over X. The free monoid generated by X under the concatenation operation $(u, v) \mapsto uv$ with e as the identity is also denoted X^*. Similarly, X^+ also stands for the free semigroup generated by X. Subsets of X^* are called *languages*, and subsets of X^+ are *e-free* languages.

An X-*automaton* is a triple (A, X, δ) consisting of a finite non-empty set A of *states*, the *input alphabet* X, and a *transition function* $\delta : A \times X \to A$; for any $a \in A$ and $x \in X$, $\delta(a, x)$ is the next state of the automaton if a is the present state and x is the input letter currently read by the automaton. The transition function δ is extended to a function $\delta^* : A \times X^* \to A$ by setting $\delta^*(a, e) = a$ and $\delta^*(a, ux) = \delta(\delta^*(a, u), x)$ for all $a \in A$, $u \in X^*$ and $x \in X$. Then, for any $a \in A$ and $w \in X^*$, $\delta^*(a, w)$ is the state reached from state a after reading the input word w. An X-*recognizer* is a system $\mathbf{A} = (A, X, \delta, a_0, F)$, where (A, X, δ) is an X-automaton, $a_0 \in A$ is the *initial state*, and $F \subseteq A$ is the set of *final states*. The *language recognized* by $\mathbf{A}$ is the set $L(\mathbf{A}) = \{w \in X^* \mid \delta^*(a_0, w) \in F\}$. A language $L \subseteq X^*$ is *recognizable*, or *regular*, if $L = L(\mathbf{A})$ for some X-recognizer $\mathbf{A}$. Let $\mathrm{Rec}(X)$ denote the set of

all regular languages over X and let $\mathrm{Rec} = \{\mathrm{Rec}(X)\}_X$ be the family of all regular languages, where X ranges over all finite alphabets[1].

Let us note some closure properties of the family Rec. At the same time we define several important language operations.

3.1 Proposition *Let X and Y be any (finite non-empty) alphabets.*

(1) (Boolean operations) $\mathrm{Rec}(X)$ *forms a field of sets on* X^*.

(2) (Product) *If* $K, L \in \mathrm{Rec}(X)$, *then* $KL := \{uv \mid u \in K, v \in L\} \in \mathrm{Rec}(X)$.

(3) (Iteration) *If* $L \in \mathrm{Rec}(X)$, *then* $L^* := \{u_1 u_2 \dots u_n \mid n \geq 0, u_1, \dots, u_n \in L\} \in \mathrm{Rec}(X)$.

(4) (Quotients) *For any* $L \in \mathrm{Rec}(X)$ *and* $w \in X^*$,

  (a) $w^{-1}L := \{u \in X^* \mid wu \in L\} \in \mathrm{Rec}(X)$;
  (b) $Lw^{-1} := \{u \in X^* \mid uw \in L\} \in \mathrm{Rec}(X)$.

(5) (Homomorphisms and Inverse Homomorphisms) *For any homomorphism* $\varphi : X^* \to Y^*$,

  (a) $L \in \mathrm{Rec}(X)$ *implies* $L\varphi \in \mathrm{Rec}(Y)$;
  (b) $L \in \mathrm{Rec}(Y)$ *implies* $L\varphi^{-1} \in \mathrm{Rec}(X)$.

Let us now recall some algebraic characterizations of regular languages that have natural counterparts in the theory of regular tree languages.

3.2 Theorem (Kleene 1956) *A language over an alphabet X is regular if and only if it can be obtained from the empty language $\emptyset$ and the singleton sets $\{x\}$ $(x \in X)$ by means of the* regular operations *union $K \cup L$, product KL, and iteration L^*.*

Hence, a language is regular if and only if it is denoted by a *regular expression* that shows how it can be obtained from the elementary languages $\emptyset$ and $\{x\}$ by regular operations.

3.3 Theorem (Nerode 1958) *For any language $L \subseteq X^*$ the following are equivalent:*

(1) *L is regular;*

(2) *L is saturated by a right congruence on X^* of finite index;*

(3) *the following* Nerode congruence ϱ_L *on X^* of L is of finite index:*

$$u \,\varrho_L\, v \iff (\forall w \in X^*)(uw \in L \leftrightarrow vw \in L) \quad (u, v \in X^*).$$

The Nerode congruence of a language $L \subseteq X^*$ is easily seen to be the greatest right congruence on X^* that saturates L, and if L is regular, it yields a *minimal recognizer* of L in which the states are the ϱ_L-classes.

3.4 Theorem (Myhill 1957) *For any language $L \subseteq X^*$, the following are equivalent:*

[1] Since no set-theoretic problems can arise here, we will simply let X range over "all" alphabets. Of course, we could assume that every X is a finite non-empty subset of a given infinite set of symbols.

(1) *L is regular;*

(2) *L is saturated by a congruence on X^* of finite index;*

(3) *the following* Myhill congruence μ_L *on X^* of L is of finite index:*

$$u\,\mu_L\,v \iff (\forall s,t \in X^*)(sut \in L \leftrightarrow svt \in L) \quad (u,v \in X^*).$$

It is easy to see that μ_L is the greatest congruence on X^* that saturates L. The congruence μ_L is also called the *syntactic congruence* of L and the quotient monoid $M(L) = X^*/\mu_L$ is the *syntactic monoid* of L. If L is regular, $M(L)$ can be computed because it is isomorphic to the *transition monoid* of the minimal recognizer $\mathbf{A} = (A, X, \delta, a_0, F)$ of L, that is, the monoid formed by the maps $\delta^*(-,w) : A \to A$, $a \mapsto \delta^*(a,w)$ $(w \in X^*)$, under composition. Myhill's Theorem can be restated as follows. Condition (2) of the following corollary is often used as the definition of regularity in algebraic presentations of the theory of regular languages.

3.5 Corollary *For any language $L \subseteq X^*$, the following are equivalent:*

(1) *L is regular;*

(2) *there exist a finite monoid M, a homomorphism $\varphi : X^* \to M$ and a subset $H \subseteq M$ such that $L = H\varphi^{-1}$;*

(3) *the syntactic monoid $M(L)$ of L is finite.*

Several interesting types of regular languages with some special properties have been studied in the literature. Let us define generally a *family of regular languages* as a mapping $\mathcal{L}$ that assigns to each alphabet X a set $\mathcal{L}(X) \subseteq \mathrm{Rec}(X)$ of regular languages over X. We write $\mathcal{L} = \{\mathcal{L}(X)\}_X$ with the understanding that X ranges over all alphabets. For each such family $\mathcal{L}$ one is faced with the problem of finding an algorithm for deciding for any given X-recognizer $\mathbf{A}$ whether $L(\mathbf{A}) \in \mathcal{L}(X)$.

The *definite languages* already introduced by Kleene [42] form perhaps the first non-trivial proper sub-family of the regular languages that got an effective characterization when Perles, Rabin and Shamir [56] described the corresponding recognizers. However, a considerably harder problem was solved when Schützenberger [77] proved that a language is *star-free* if and only if its syntactic monoid is *aperiodic*, that is to say, has only trivial subgroups. The result was remarkable as this family of languages arises in many natural ways (cf. [50]), but no other decision method for it was known. Subsequently several other families were similarly characterized by properties of their syntactic monoids or syntactic semigroups. Finally, in his Variety Theorem Eilenberg [18] identified the families of languages for which such a characterization is possible. As Eilenberg's theory is also the starting point for the corresponding work on tree languages, we review its main notions and results. For systematic expositions the reader is referred to [2, 18, 63, 64]. Briefer accounts can be found in [37, 45].

A family $\mathcal{L} = \{\mathcal{L}(X)\}_X$ of regular languages is called a *$*$-variety*, or a *variety of regular languages* (VRL) if for all alphabets X and Y,

(1) $\mathcal{L}(X) \subseteq \mathrm{Rec}(X)$,

(2) $L \in \mathcal{L}(X)$ implies $X^* \setminus L \in \mathcal{L}$,

(3) $K, L \in \mathcal{L}(X)$ implies $K \cap L \in \mathcal{L}(X)$,

(4) $L \in \mathcal{L}(X)$ implies $w^{-1}L, Lw^{-1} \in \mathcal{L}(X)$ for every $w \in X^*$, and

(5) $L \in \mathcal{L}(Y)$ implies $L\varphi^{-1} \in \mathcal{L}(X)$ for every homomorphism $\varphi : X^* \to Y^*$.

In [18] the corresponding families of e-free languages are called *+-varieties*. The greatest VRL is the family Rec of all regular languages. If we exclude the VRL $\mathcal{L}$ with $\mathcal{L}(X) = \emptyset$ for every X, then the least VRL is Triv, where $\mathrm{Triv}(X) = \{\emptyset, X^*\}$ for every X. The more interesting examples include the families of *star-free*, *definite*, *reverse definite*, *generalized definite*, *locally testable* and *piecewise testable languages*.

A non-empty class $\mathbf{M}$ of finite monoids is a *variety of finite monoids* (VFM), or a *pseudovariety*, if it is closed under the forming of submonoids, images and finite direct products. For any given class $\mathbf{K}$ of finite monoids, there is a unique minimal VFM $\mathrm{V_f}(\mathbf{K})$ containing $\mathbf{K}$, the *VFM generated* by $\mathbf{K}$.

For any family $\mathcal{L} = \{\mathcal{L}(X)\}_X$ of regular languages, let

$$\mathcal{L}^\mu = \mathrm{V_f}(\{M(L) \mid L \in \mathcal{L}(X) \text{ for some } X\}),$$

and for any class $\mathbf{K}$ of finite monoids, define $\mathbf{K}^\lambda = \{\mathbf{K}^\lambda(X)\}_X$ by setting

$$\mathbf{K}^\lambda(X) = \{L \subseteq X^* \mid M(L) \in \mathbf{K}\} \quad \text{for each } X.$$

If we let **VRL** and **VFM** denote the classes of all VRLs and all VFMs, respectively, then *Eilenberg's Variety Theorem* can be stated as follows.

3.6 Theorem (Eilenberg 1976) *The mappings $\mathcal{L} \mapsto \mathcal{L}^\mu$ and $\mathbf{M} \mapsto \mathbf{M}^\lambda$ define mutually inverse isomorphisms between the lattice $(\mathbf{VRL}, \subseteq)$ of all varieties of regular languages and the lattice $(\mathbf{VFM}, \subseteq)$ of all varieties of finite monoids. In particular,*

(1) *if $\mathcal{L} \in \mathbf{VRL}$, then $\mathcal{L}^\mu \in \mathbf{VFM}$ and $\mathcal{L}^{\mu\lambda} = \mathcal{L}$, and*

(2) *if $\mathbf{M} \in \mathbf{VFM}$, then $\mathbf{M}^\lambda \in \mathbf{VRL}$ and $\mathbf{M}^{\lambda\mu} = \mathbf{M}$.*

If $\mathcal{L} = \{\mathcal{L}(X)\}_X$ is a VRL, then for any X and $L \subseteq X^*$, $L \in \mathcal{L}(X)$ if and only if $M(L) \in \mathcal{L}^\mu$. Hence, an effective characterization of $\mathcal{L}$ may be obtained by determining the VFM $\mathcal{L}^\mu$. A similar correspondence holds between the varieties of finite semigroups and the +-varieties.

Let us also recall an important addition to the Variety Theorem due to Thérien [86, 87]. For any alphabet X, let $\mathrm{FCon}(X^*)$ denote the set of congruences on X^* of finite index. Clearly $\mathrm{FCon}(X^*)$ is a filter of the congruence lattice $\mathrm{Con}(X^*)$. A *$*$-variety of congruences*, or a *variety of finite congruences* (VFC), is a family $\Gamma = \{\Gamma(X)\}_X$ of sets of congruences such that for all alphabets X and Y,

(1) $\Gamma(X) \subseteq \mathrm{FCon}(X^*)$ is a filter of $\mathrm{Con}(X^*)$, and

(2) if $\varphi : X^* \to Y^*$ is a homomorphism and $\theta \in \Gamma(Y)$, then $\varphi \circ \theta \circ \varphi^{-1} \in \Gamma(X)$.

That each VRL corresponds to a unique VFC is a useful fact as many varieties of regular languages are most naturally defined in terms of congruences of the monoids X^*.

We conclude this section by noting how finite automata can be defined as unary algebras. This approach, already propounded by J. R. Büchi and J. B. Wright in the 1950s, provides a natural passage to tree automata (cf. [8, 13, 16, 28, 79, 85], for example).

Each input letter $x \in X$ defines in an X-automaton (A, X, δ) a unary operation

$$x^{\mathcal{A}} : A \to A,\ a \mapsto \delta(a, x),$$

and these operations determine δ completely. Hence (A, X, δ) can be redefined as a unary algebra $\mathcal{A} = (A, X)$ when we view X as a set of unary operation symbols. If ε is a variable, we may identify each word $w \in X^*$ with an X-term t_w over $\{\varepsilon\}$ by setting $t_e = \varepsilon$, and $t_w = x(t_u)$ for $w = ux$ $(u \in X^*, x \in X)$. For example, $t_{xxy} = y(x(x(\varepsilon)))$. By letting w represent the term t_w, the $X\{\epsilon\}$-term algebra may be taken to be $\mathcal{T}_X(\varepsilon) = (X^*, X)$, where $x^{\mathcal{T}_X(\varepsilon)}(w) = wx$ for any $w \in X^*$ and $x \in X$. The mapping $\delta^*(-, w) : A \to A,\ a \mapsto \delta^*(a, w)$, induced by an input word $w \in X^*$ in (A, X, δ) now becomes the term function $w^{\mathcal{A}}$ defined by the term w $(= t_w)$ in the algebra $\mathcal{A} = (A, X)$. Furthermore, an X-recognizer can be defined as a system $\mathbf{A} = (\mathcal{A}, a_0, F)$, where $\mathcal{A} = (A, X)$ is a finite X-algebra, $a_0 \in A$ is the initial state, and $F \subseteq A$ is the set of final states. The language recognized by $\mathbf{A}$ is then the term set $L(\mathbf{A}) = \{w \mid w^{\mathcal{A}}(a_0) \in F\}$, and $L \subseteq X^*$ is regular if and only if $L = F\varphi^{-1}$ for some finite X-algebra $\mathcal{A} = (A, X)$, a homomorphism $\varphi : \mathcal{T}_X(\varepsilon) \to \mathcal{A}$ and a subset $F \subseteq A$.

Tree recognizers and regular tree languages are now obtained by allowing function symbols of any finite arities. Let us note that the unary interpretation described above also suggests an alternative theory of varieties of regular languages [83].

4 Trees and terms

In mathematics and computer science trees are defined in several different ways, often depending on the applications in mind. The trees to be considered here are finite, their nodes are labelled with symbols, and the branches leaving any given node have a specified order. For example, derivations in context-free grammars can be represented by such trees. For the algebraic approach to be adopted here it will be convenient to define our trees formally as terms of the kind used in algebra and logic, for example.

A *ranked alphabet* is a finite set of operation symbols, but now these symbols will also be used for labelling nodes of trees. In what follows, Σ is always a ranked alphabet. For each $m \geq 0$, the set of m-ary symbols in Σ is again denoted Σ_m. If Ω is also a ranked alphabet, $\Sigma \subseteq \Omega$ means that $\Sigma_m \subseteq \Omega_m$ for every $m \geq 0$. The union $\Sigma \cup \Omega$ may be formed if $\Sigma_m \cap \Omega_n = \emptyset$ whenever $m \neq n$, and then $(\Sigma \cup \Omega)_m = \Sigma_m \cup \Omega_m$ for every $m \geq 0$. In examples we may give a ranked alphabet in the form $\Sigma = \{f_1/m_1, \ldots, f_k/m_k\}$ indicating that Σ consists of the symbols $f_1, \ldots, f_m$ with the respective ranks $m_1, \ldots, m_k$.

In addition to ranked alphabets, ordinary finite alphabets, called *leaf alphabets*, are used for labelling leaves of trees. These will usually be denoted by X or Y. When a leaf alphabet is considered together with a ranked alphabet, the two sets are assumed to be disjoint.

Terms will be regarded as syntactic representations of trees, and Σ-terms with variables in X are called also *ΣX-trees*. Any $t \in X \cup \Sigma_0$ represents a one-node tree in which the only node is labelled with the symbol t. A composite term $f(t_1, \ldots, t_m)$ is interpreted as

a tree formed by adjoining the m trees represented by $t_1, \ldots, t_m$ to a new f-labelled root. Subsets of $T_\Sigma(X)$ are called ΣX-*tree languages*. We may also speak about Σ-*trees* and Σ-*tree languages* without specifying the leaf alphabet, or generally about *trees* and *tree languages* without mentioning any alphabet.

If $\Sigma_0 \cup X = \emptyset$, the set $T_\Sigma(X)$ is also empty, and if $\Sigma = \Sigma_0$, the only ΣX-trees are the finitely many one-node trees labelled with symbols from $\Sigma_0 \cup X$. To exclude these uninteresting trivial cases, we will tacitly assume that $\Sigma_0 \cup X \neq \emptyset$ and $\Sigma \neq \Sigma_0$.

The inductive definition of $T_\Sigma(X)$ yields a *Principle of Tree Induction* for proving assertions about ΣX-trees: a statement S holds for every ΣX-tree if

(1) S holds for every $x \in X$ and for every $c \in \Sigma_0$, and

(2) S holds for $f(t_1, \ldots, t_m)$ assuming that S holds for $t_1, \ldots, t_m$ ($m > 0$, $f \in \Sigma_m$).

Similarly, notions related to ΣX-trees may be defined recursively following the inductive definition of $T_\Sigma(X)$. As useful examples we define the set of *subtrees* $\mathrm{sub}(t)$, the *height* $\mathrm{hg}(t)$ and the *root (symbol)* $\mathrm{root}(t)$ of a ΣX-tree t:

(1) $\mathrm{sub}(t) = \{t\}$, $\mathrm{hg}(t) = 0$ and $\mathrm{root}(t) = t$ for any $t \in X \cup \Sigma_0$;

(2) $\mathrm{sub}(t) = \{t\} \cup \mathrm{sub}(t_1) \cup \cdots \cup \mathrm{sub}(t_m)$, $\mathrm{hg}(t) = \max\{\mathrm{hg}(t_1), \ldots, \mathrm{hg}(t_m)\} + 1$ and $\mathrm{root}(t) = f$ for $t = f(t_1, \ldots, t_m)$.

For example, for the ΣX-tree $t = g(f(g(y), x))$, where $f \in \Sigma_2$, $g \in \Sigma_1$ and $x, y \in X$, we get $\mathrm{hg}(t) = 3$, $\mathrm{root}(t) = g$ and $\mathrm{sub}(t) = \{t, f(g(y), x), g(y), x, y\}$.

For any $n \geq 0$, let

$$T_\Sigma(X)^{\geq n} = \{t \in T_\Sigma(X) \mid \mathrm{hg}(t) \geq n\}.$$

Similarly, let $T_\Sigma(X)^{<n}$ be the set of ΣX-trees of height $< n$.

Let ξ be a special symbol that appears neither in any ranked alphabet nor in any leaf alphabet considered. A $\Sigma(X \cup \{\xi\})$-tree in which ξ appears exactly once, is called a ΣX-*context*. The set of all ΣX-contexts is denoted by $C_\Sigma(X)$. If $p, q \in C_\Sigma(X)$, then $p \cdot q = q(p)$ is the ΣX-context obtained by replacing the ξ in q with p. Similarly, if $t \in T_\Sigma(X)$ and $p \in C_\Sigma(X)$, then $t \cdot p = p(t)$ is the ΣX-tree obtained when the ξ in p is replaced with t. The *height* $\mathrm{hg}(p)$ and the *root* $\mathrm{root}(p)$ of a ΣX-context p are defined the same way as for ΣX-trees treating ξ as a leaf symbol, i.e., $\mathrm{hg}(\xi) = 0$ and $\mathrm{root}(\xi) = \xi$. Moreover, the *depth* $\mathrm{dp}(p)$ of a ΣX-context p is the distance of the ξ-labelled leaf from the root, that is to say

(1) $\mathrm{dp}(\xi) = 0$;

(2) $\mathrm{dp}(p) = \mathrm{dp}(q) + 1$ for $p = f(t_1, \ldots, q, \ldots, t_m)$, $t_1, \ldots, t_m \in T_\Sigma(X)$ and $q \in C_\Sigma(X)$.

Finally, let us note that in many presentations no separate leaf alphabets are used, but a special set of nullary symbols is singled out when the need arises. Although this could be done without any essential loss of generality, leaf alphabets are convenient in many cases and we shall use them. If $X = \emptyset$ in the above definitions, $T_\Sigma(X)$ becomes the set T_Σ of *ground* Σ-*terms*, or *ground* Σ-*trees*, and $C_\Sigma(X)$ becomes the set C_Σ of Σ-*contexts*.

5 Finite tree recognizers and regular tree languages

A finite Σ-algebra $\mathcal{A} = (A, \Sigma)$ may be regarded as a *tree automaton* that reads ΣX-trees in a *bottom-up*, or *frontier-to-root*, fashion starting from the leaves and finishing at the root. The elements of A are then called *states*. At a leaf labelled with a nullary symbol $c \in \Sigma_0$ it starts in state $c^{\mathcal{A}}$. The starting states at leaves labelled with symbols from X are specified by a mapping $\alpha : X \to A$. If $\mathcal{A}$ has reached the m immediate descendant nodes of a node u labelled with an m-ary symbol $f \in \Sigma_m$, where $m > 0$, in states $a_1, \ldots, a_m$, respectively, then it enters u in state $f^{\mathcal{A}}(a_1, \ldots, a_m)$. Obviously, the root of a ΣX-tree t is reached in state $t\alpha_{\mathcal{A}}$, where $\alpha_{\mathcal{A}} : T_\Sigma(X) \to \mathcal{A}$ is the homomorphic extension of α. Specifying now a set of final states, we obtain the following tree recognizers that are also called (*deterministic*) *frontier-to-root tree recognizers*.

5.1 Definition A (*deterministic bottom-up*) ΣX*-recognizer* $\mathbf{A} = (\mathcal{A}, \alpha, F)$ consists of a finite Σ-algebra $\mathcal{A} = (A, \Sigma)$, an *initial assignment* $\alpha : X \to A$, and a set $F \subseteq A$ of *final states*; A is the *state set*. The ΣX-tree language *recognized* by $\mathbf{A}$ is the set

$$T(\mathbf{A}) = \{t \in T_\Sigma(X) \mid t\alpha_{\mathcal{A}} \in F\}.$$

A ΣX-tree language is called *recognizable*, or *regular*, if it is recognized by some ΣX-recognizer. Let $\mathrm{Rec}_\Sigma(X)$ be the set of all recognizable ΣX-tree languages.

We may also speak generally about *tree recognizers* without specifying the alphabets. The family of all regular tree languages is denoted by Rec. The following example illustrates some basic capabilities and a limitation of these recognizers.

5.2 Example Let $\Sigma = \{f/2, g/1\}$ and let $X = \{x, y\}$. For any $t \in T_\Sigma(X)$, we set $g^0(t) = t$ and $g^{k+1}(t) = g(g^k(t))$ for all $k \geq 0$. The ΣX-tree language

$$T_1 = \{p(f(g^m(x), g^n(y))) \mid p \in C_\Sigma(X),\ m \equiv 1 \pmod 2,\ n \geq 2\},$$

formed by the ΣX-trees that have a subtree $f(g^m(x), g^n(y))$ with m odd and $n \geq 2$, is recognized by the ΣX-recognizer $\mathbf{A} = (\mathcal{A}, \alpha, F)$ defined as follows. The state set is $A = \{a_0, a_1, b_0, b_1, b_2, a_+, a_-\}$ and the operations of $\mathcal{A}$ are defined thus:

- $g^{\mathcal{A}}(a_0) = a_1$, $g^{\mathcal{A}}(a_1) = a_0$, $g^{\mathcal{A}}(b_0) = b_1$, $g^{\mathcal{A}}(b_1) = g^{\mathcal{A}}(b_2) = b_2$,
- $g^{\mathcal{A}}(a_+) = a_+$, $g^{\mathcal{A}}(a_-) = a_-$,
- $f^{\mathcal{A}}(a_1, b_2) = a_+$, $f^{\mathcal{A}}(a_+, a) = f^{\mathcal{A}}(a, a_+) = a_+$ for all $a \in A$, and
- $f^{\mathcal{A}}(a, b) = a_-$ in all remaining cases.

Furthermore, $\alpha(x) = a_0$, $\alpha(y) = b_0$ and $F = \{a_+\}$.

On the other hand, the ΣX-tree language $T_2 = \{p(f(g^k(x), g^k(x))) \mid n \geq 0\}$ is not regular. Indeed, if $T = T(\mathbf{A})$ for a ΣX-recognizer $\mathbf{A} = (\mathcal{A}, \alpha, F)$, then $g^j(x)\alpha_{\mathcal{A}} = g^k(x)\alpha_{\mathcal{A}}$ for some $0 \leq j < k$. Now $f(g^j(x), g^k(x)) \in T(\mathbf{A})$ would follow from

$$f(g^j(x), g^k(x))\alpha_{\mathcal{A}} = f^{\mathcal{A}}(g^j(x)\alpha_{\mathcal{A}}, g^k(x)\alpha_{\mathcal{A}}) = f^{\mathcal{A}}(g^k(x)\alpha_{\mathcal{A}}, g^k(x)\alpha_{\mathcal{A}}) \in F.$$

Similarly as the non-regularity of T_2 in the above example, the following general *Pumping Lemma* also follows from the finiteness of the state sets of tree recognizers.

5.3 Lemma *For any recognizable ΣX-tree language T there is a number $n \geq 1$ such that if $t \in T$ and $\mathrm{hg}(t) \geq n$, then for some $s \in T_\Sigma(X)$ and $p, q \in C_\Sigma(X)$,*

(1) $t = s \cdot p \cdot q$,

(2) $\mathrm{dp}(p) \geq 1$, $1 \leq \mathrm{hg}(s \cdot p) \leq n$, *and*

(3) $s \cdot p^k \cdot q \in T$ *for every* $k \geq 0$.

We may choose the number of states of any ΣX-recognizer $\mathbf{A}$ such that $T = T(\mathbf{A})$ as the limit n in the Pumping Lemma. Moreover, the Pumping Lemma clearly implies that T is infinite if and only if there is a $t \in T$ such that $n \leq \mathrm{hg}(t) < 2n$. Although the algorithm suggested by this fact is not very efficient, it proves the decidability of the *Finiteness Problem* "Is $T(\mathbf{A})$ finite?". The *Emptiness Problem* "$T(\mathbf{A}) = \emptyset$?" is decidable, too. Indeed, if $\mathbf{A} = (\mathcal{A}, \alpha, F)$ is any ΣX-recognizer, then $T(\mathbf{A}) \neq \emptyset$ if and only if some final state $a \in F$ is *reachable*, i.e., $a = t\alpha_\mathcal{A}$ for some $t \in T$. Moreover, the reachable states form the subalgebra of $\mathcal{A}$ generated by $\alpha(X) = \{\alpha(x) \mid x \in X\}$, and this can always be computed.

Each regular ΣX-tree language T has a *minimal* ΣX-recognizer, unique up to isomorphism, that can be constructed from any given ΣX-recognizer $\mathbf{A} = (\mathcal{A}, \alpha, F)$ of T by first deleting all non-reachable states (the result is a *connected* recognizer) and then merging all pairs of equivalent states. For defining the equivalence of states, we need the following notion that will be used later too.

5.4 Definition Let $\mathbf{A} = (\mathcal{A}, \alpha, F)$ be any ΣX-recognizer. The *translation* $p^\mathbf{A} : A \to A$ of $\mathbf{A}$ defined by a ΣX-context $p \in C_\Sigma(X)$ is defined by setting $\xi^\mathbf{A} = 1_A$ and

$$p^\mathbf{A}(a) = f^\mathcal{A}(t_1\alpha_\mathcal{A}, \ldots, q^\mathbf{A}(a), \ldots, t_m\alpha_\mathcal{A})$$

for $p(\xi) = f(t_1, \ldots, q(\xi), \ldots, t_m)$, where $m > 0$, $f \in \Sigma_m$, $t_1, \ldots, t_m \in T_\Sigma(X)$, $q \in C_\Sigma(X)$, and $a \in A$. The set $\{p^\mathbf{A} \mid p \in C_\Sigma(X)\}$ of all translations of $\mathbf{A}$ is denoted $\mathrm{Tr}(\mathbf{A})$.

It is easy to verify that all translations of a ΣX-recognizer $\mathbf{A} = (\mathcal{A}, \alpha, F)$ are also translations of the Σ-algebra $\mathcal{A}$, and that if $\mathbf{A}$ is connected, then $\mathrm{Tr}(\mathbf{A}) = \mathrm{Tr}(\mathcal{A})$. Two states a and b of a ΣX-recognizer $\mathbf{A} = (\mathcal{A}, \alpha, F)$ are *equivalent* if for every $p \in C_\Sigma(X)$, $p^\mathbf{A}(a) \in F$ if and only if $p^\mathbf{A}(b) \in F$. The minimization theory of tree recognizers can be found in [28].

The regular tree languages are obtained in many other ways too. A standard subset construction shows that they are exactly the tree languages recognized by *nondeterministic bottom-up tree recognizers*. Also the *nondeterministic top-down tree recognizers* that process a tree starting at the root recognize exactly the regular tree languages. Furthermore, they are defined by *regular tree grammars*, certain systems of *fixed-point equations*, *monadic second order logic* etc. A regular tree grammar $G = (N, \Sigma, X, P, a_0)$ generating a regular ΣX-tree language consists of a finite set N of non-terminal symbols, the alphabets Σ and X, an initial symbol $a_0 \in N$, and a finite set P of productions, each of them of the form $a \to x$, $a \to c$ or $a \to f(a_1, \ldots, a_m)$, where $a, a_1, \ldots, a_m \in N$, $x \in X$, $c \in \Sigma_0$ and $f \in \Sigma_m$ ($m > 0$). Derivations and the tree language $T(G) = \{t \in T_\Sigma(X) \mid a_0 \Rightarrow_G^* t\}$ generated by G are defined in the usual manner. We refer the reader to [28, 29], for example, for details and further references.

Finally, let us note some algebraic characterizations of regularity that are of direct interest here. The following fact is an immediate consequence of Definition 5.1.

5.5 Proposition *A ΣX-tree language T is regular if and only if there exist a finite Σ-algebra $\mathcal{A} = (A, \Sigma)$, a homomorphism $\varphi : \mathcal{T}_\Sigma(X) \to \mathcal{A}$ and subset $F \subseteq A$ such that $T = F\varphi^{-1}$.*

We may say that an algebra $\mathcal{A} = (A, \Sigma)$ *recognizes* a ΣX-tree language T if $T = F\varphi^{-1}$ for some homomorphism $\varphi : \mathcal{T}_\Sigma(X) \to \mathcal{A}$ and a subset $F \subseteq A$. For every ΣX-tree language T there is a unique (up to isomorphism) "smallest" algebra recognizing T, and this algebra is finite exactly in case T is regular. Moreover, for a regular T the smallest algebra is the underlying algebra of the minimal ΣX-recognizer of T. These ideas can be also expressed in terms of congruences by generalizing Nerode's Theorem to tree languages.

It is easy to see that for any ΣX-tree language T, there is a greatest congruence θ_T on $\mathcal{T}_\Sigma(X)$ that saturates T. It could be called the *Nerode congruence* of T, and the following result is *Nerode's Theorem* for tree languages.

5.6 Proposition *For any ΣX-tree language T, the following three conditions are equivalent:*

(1) *$T \in \mathrm{Rec}_\Sigma(X)$;*

(2) *T is saturated by a congruence on $\mathcal{T}_\Sigma(X)$ of finite index;*

(3) *the Nerode congruence θ_T of T is of finite index.*

Proof The equivalence of (1) and (2) follows by Proposition 5.5 as follows.

(1) If $T = F\varphi^{-1}$ for some homomorphism $\varphi : \mathcal{T}_\Sigma(X) \to \mathcal{A}$ and a subset $F \subseteq A$, where $\mathcal{A} = (A, \Sigma)$ is a finite Σ-algebra, then $\ker \varphi$ is a congruence of finite index on $\mathcal{T}_\Sigma(X)$ that saturates T.

(2) If $\theta \in \mathrm{Con}(\mathcal{T}_\Sigma(X))$ is a congruence of finite index that saturates T, then the finite Σ-algebra $\mathcal{T}_\Sigma(X)/\theta$ recognizes T. Indeed, it is easy to see that $T = (T\theta^\natural)(\theta^\natural)^{-1}$ for the natural homomorphism $\theta^\natural : \mathcal{T}_\Sigma(X) \to \mathcal{T}_\Sigma(X)/\theta,\ t \mapsto t/\theta$.

Of course, (2) and (3) imply each other immediately. □

A characterization that could be regarded as a counterpart to Myhill's Theorem is obtained by considering fully invariant congruences on $\mathcal{T}_\Sigma(X)$ (cf. [79, 81]).

6 Tree language operations and closure properties

We introduce now several tree language operations and note that the family of regular tree languages is closed under most of them. As shown by the following obvious lemma, we may usually assume that all tree languages involved are over the same alphabets Σ and X.

6.1 Lemma *If Σ and Ω are ranked alphabets such that $\Sigma \subseteq \Omega$, and X and Y are leaf alphabets such that $X \subseteq Y$, then for any $T \subseteq T_\Sigma(X)$, $T \in \mathrm{Rec}_\Sigma(X)$ if and only if $T \in \mathrm{Rec}_\Omega(Y)$.*

In particular, if $S \in \mathrm{Rec}_\Sigma(X)$, $T \in \mathrm{Rec}_\Omega(Y)$, and $\Sigma_m \cap \Omega_n = \emptyset$ whenever $m \neq n$, then $S, T \in \mathrm{Rec}_{\Sigma\cup\Omega}(X \cup Y)$.

As tree languages are sets, we may apply any *Boolean operations*, i.e., the usual basic set-theoretic operations, to them.

6.2 Proposition *For any ranked alphabet Σ and any leaf alphabet X, $\mathrm{Rec}_\Sigma(X)$ is a field of sets on $T_\Sigma(X)$.*

Proof It is clear that $\emptyset, T_\Sigma(X) \in \mathrm{Rec}_\Sigma(X)$. If $S, T \in \mathrm{Rec}_\Sigma(X)$, then $S = T(\mathbf{A})$ and $T = T(\mathbf{B})$ for some ΣX-recognizers $\mathbf{A} = (\mathcal{A}, \alpha, F)$ and $\mathbf{B} = (\mathcal{B}, \beta, G)$. Let $\mathbf{C} = (\mathcal{C}, \gamma, H)$ be a new ΣX-recognizer, where $\mathcal{C} = (A \times B, \Sigma)$ is the direct product $\mathcal{A} \times \mathcal{B}$, the initial assignment is $\gamma : X \to A \times B, x \mapsto (\alpha(x), \beta(x))$, and the set of final states $H \subseteq A \times B$ is specified as appropriate. Since $t\gamma_\mathcal{C} = (t\alpha_\mathcal{A}, t\beta_\mathcal{B})$ for every $t \in T_\Sigma(X)$, it is clear that any Boolean combination of S and T can be recognized by $\mathbf{C}$ by selecting a suitable set H of final states. For example, $T(\mathbf{C}) = S - T$ if $H = F \times (B - G)$. □

There are a few different natural products of tree languages. The *tree language product* $T(x \leftarrow T_x \mid x \in X)$ of a ΣX-tree language T and an X-indexed family $(T_x \mid x \in X)$ of ΣX-tree languages is the set of all ΣX-trees that can be obtained from a tree $t \in T$ by simultaneously replacing in it every $x \in X$ with a tree from the corresponding set T_x. The different occurrences of each x may be replaced with different trees from T_x.

6.3 Proposition *If the ΣX-tree languages T and T_x $(x \in X)$ are regular, then so is their tree language product $T(x \leftarrow T_x \mid x \in X)$.*

Regular tree grammars often provide the simplest way to prove a closure result like this. It is quite easy to construct a regular tree grammar generating $T(x \leftarrow T_x \mid x \in X)$ if we are given grammars generating T and the tree languages T_x $(x \in X)$. Note that we may assume that all grammars have pairwise disjoint sets of non-terminals.

It is easy to see that the operation defined in the following corollary is a special case of the tree language product.

6.4 Corollary *For any $m > 0$ and $f \in \Sigma_m$, the* f-product

$$f(T_1, \ldots, T_m) := \{f(t_1, \ldots, t_m) \mid t_1 \in T_1, \ldots, t_m \in T_m\}$$

of any $T_1, \ldots, T_m \in \mathrm{Rec}_\Sigma(X)$, is also a regular ΣX-tree language.

For any given $z \in X$, the *z-product* $S \cdot_z T$ of two ΣX-tree languages S and T is defined as the special tree language product $T(x \leftarrow T_x \mid x \in X)$, where $T_z = S$ and $T_x = \{x\}$ for all $x \neq z$, $x \in X$. In other words, any element of $S \cdot_z T$ is obtained from some tree $t \in T$ by replacing each z-labelled leaf with some tree from S, and again different z-labelled leaves can be replaced with different members of S.

6.5 Corollary *If $S, T \in \mathrm{Rec}_\Sigma(X)$, then $S \cdot_z T \in \mathrm{Rec}_\Sigma(X)$ for every $z \in X$.*

For any $z \in X$, the *z-iteration* of a ΣX-tree language T is defined as the union

$$T^{*z} := \bigcup\{T^{k,z} \mid k \geq 0\} = \{z\} \cup T \cup (\{z\} \cup T) \cdot_z T \cup \ldots,$$

where $T^{0,z} = \{z\}$, and $T^{k,z} = T^{k-1,z} \cdot_z T \cup T^{k-1,z}$ for every $k \geq 1$.

6.6 Proposition *If $T \in \mathrm{Rec}_\Sigma(X)$, then $T^{*z} \in \mathrm{Rec}_\Sigma(X)$ for every $z \in X$.*

Any ΣX-context $p \in C_\Sigma(X)$ defines a unary operation $t \mapsto p(t)$ on $T_\Sigma(X)$. It is easy to see that these operations are exactly the translations of the ΣX-term algebra $\mathcal{T}_\Sigma(X)$. We extend them and their inverses in the natural way to tree language operations: for any $p \in C_\Sigma(X)$ and $T \subseteq T_\Sigma(X)$, let $p(T) := \{p(t) \mid t \in T\}$ and $p^{-1}(T) := \{t \in T_\Sigma(X) \mid p(t) \in T\}$.

6.7 Proposition *If $T \in \mathrm{Rec}_\Sigma(X)$, then $p(T), p^{-1}(T) \in \mathrm{Rec}_\Sigma(X)$ for every $p \in C_\Sigma(X)$. Moreover, the set $\{p^{-1}(T) \mid p \in C_\Sigma(X)\}$ is finite for every regular ΣX-tree language T.*

Proof Let $T \in \mathrm{Rec}_\Sigma(X)$ and $p \in C_\Sigma(X)$. Since $p(T) = T \cdot_\xi \{p\}$ and $p(T) \subseteq T_\Sigma(X)$, the claim $p(T) \in \mathrm{Rec}_\Sigma(X)$ holds by Corollary 6.5 and Lemma 6.1.

If $\mathbf{A} = (\mathcal{A}, \alpha, F)$ is a ΣX-recognizer such that $T(\mathbf{A}) = T$, it is clear that $p^{-1}(T) = T(\mathbf{A}_p)$ for the ΣX-recognizer $\mathbf{A}_p = (\mathcal{A}, \alpha, F_p)$, where $F_p = \{a \in A \mid p^{\mathbf{A}}(a) \in F\}$. Since the number of possible sets F_p is finite, this also proves the last assertion of the proposition. □

Since $\mathcal{T}_\Sigma(X)$ is generated by X, any homomorphism $\varphi : \mathcal{T}_\Sigma(X) \to \mathcal{T}_\Sigma(Y)$ of Σ-algebras is completely determined by the images $x\varphi \in T_\Sigma(Y)$ of the leaf symbols $x \in X$. In fact, the image $t\varphi$ of any $t \in T_\Sigma(X)$ is obtained from t by replacing every occurrence of each $x \in X$ by the ΣY-tree $x\varphi$. Again, we associate with any such φ two tree language operations:

6.8 Proposition *Let $\varphi : \mathcal{T}_\Sigma(X) \to \mathcal{T}_\Sigma(Y)$ be a homomorphism of Σ-algebras.*

(1) *If $T \in \mathrm{Rec}_\Sigma(X)$, then $T\varphi := \{t\varphi \mid t \in T\} \in \mathrm{Rec}_\Sigma(Y)$.*

(2) *If $T \in \mathrm{Rec}_\Sigma(Y)$, then $T\varphi^{-1} := \{s \in T_\Sigma(X) \mid s\varphi \in T\} \in \mathrm{Rec}_\Sigma(X)$.*

Proof Since $T\varphi = T(x \leftarrow \{x\varphi\} \mid x \in X)$, Claim (1) follows by Proposition 6.3. Assume now that $T = T(\mathbf{A})$ for some ΣY-recognizer $\mathbf{A} = (\mathcal{A}, \alpha, F)$. If $t \in T_\Sigma(X)$, then $t \in T\varphi^{-1}$ if and only if one obtains a ΣY-tree belonging to T when, for every $x \in X$, all x-labelled leaves of t are replaced with $x\varphi$. Hence, $T\varphi^{-1} = T(\mathbf{B})$ for the ΣX-recognizer $\mathbf{B} = (\mathcal{A}, \beta, F)$, where $\beta : X \to A$ is defined by the condition $\beta(x) = (x\varphi)\alpha_\mathcal{A}$ $(x \in X)$. □

We shall now consider mappings of trees that may alter the structure of a tree more radically than the homomorphisms of Σ-term algebras considered above. In what follows, Σ and Ω are ranked alphabets, and X and Y are leaf alphabets. Moreover, for any $m \geq 0$, let $\Xi_m = \{\xi_1, \ldots, \xi_m\}$ be a set of m variables that do not appear in the other alphabets.

6.9 Definition Given a mapping $\varphi_X : X \to T_\Sigma(Y)$ and, for each $m \geq 0$ such that $\Sigma_m \neq \emptyset$, a mapping $\varphi_m : \Sigma_m \to T_\Omega(Y \cup \Xi_m)$, the *tree homomorphism* $\varphi : T_\Sigma(X) \to T_\Omega(Y)$ determined by these mappings is defined as follows:

(1) $x\varphi = \varphi_X(x)$ for any $x \in X$,

(2) $c\varphi = \varphi_0(c)$ for any $c \in \Sigma_0$, and

(3) $t\varphi = \varphi_m(f)(\xi_1 \leftarrow t_1\varphi, \ldots, \xi_m \leftarrow t_m\varphi)$ for $t = f(t_1, \ldots, t_m)$ $(m > 0)$.

The tree homomorphism φ is *linear* if no variable ξ_i appears more than once in $\varphi_m(f)$ for any $m > 0$ and $f \in \Sigma_m$. It is *non-deleting* if for all $m > 0$ and $f \in \Sigma_m$, every variable $\xi_1, \ldots, \xi_m$ appears at least once in $\varphi_m(f)$.

6.10 Example If $\Sigma = \{\vee/2, \neg/1\}$, $\Omega = \{\wedge/2, \neg/1\}$, and X is any finite set of propositional variables, then ΣX-trees are propositions formed using the connectives $\vee$ (disjunction) and $\neg$ (negation), while ΩX-terms are propositions that may involve only the connectives $\wedge$ (conjunction) and $\neg$. A tree homomorphism $\varphi : T_\Sigma(X) \to T_\Omega(X)$ that converts any proposition of the first kind into an equivalent proposition of the second kind is determined by the mappings φ_X, φ_1 and φ_2 such that $\varphi_X(x) = x$ for every $x \in X$, $\varphi_1(\neg) = \neg\xi_1$, and $\varphi_2(\vee) = \neg(\neg\xi_1 \wedge \neg\xi_2)$, where we have used the conventional way of writing propositions. For example, $(x \vee \neg y)\varphi = \neg(\neg x \wedge \neg\neg y)$. The tree homomorphism is both linear and non-deleting.

A tree homomorphism $\varphi : T_\Sigma(X) \to T_\Omega(Y)$ defines two tree language operations:

(1) $T\varphi := \{t\varphi \mid t \in T\}$ $(\subseteq T_\Omega(Y))$, for any $T \subseteq T_\Sigma(X)$, and

(2) $T\varphi^{-1} := \{t \in T_\Sigma(X) \mid t\varphi \in T\}$, for any $T \subseteq T_\Omega(Y)$.

Since non-linear tree homomorphisms may form identical copies of the image of an arbitrarily large subtree, such tree homomorphisms do not necessarily preserve regularity.

6.11 Example For $\Sigma = \{f/1\}$, $\Omega = \{g/2\}$ and $X = \{x\}$, let $\varphi : T_\Sigma(X) \to T_\Omega(X)$ be the tree homomorphism determined by $\varphi_X(x) = x$ and $\varphi_1(f) = g(\xi_1, \xi_1)$. Clearly, $T_\Sigma(X)\varphi = \{x, g(x,x), g(g(x,x), g(x,x)), \dots\}$ is the non-regular set of all fully balanced ΩY-trees.

That non-linearity is crucial here is shown by the following result. On the other hand, all inverse tree homomorphisms preserve regularity.

6.12 Proposition *Let $\varphi : T_\Sigma(X) \to T_\Omega(Y)$ be a tree homomorphism*

(1) *If $T \in \mathrm{Rec}_\Omega(Y)$, then $T\varphi^{-1} \in \mathrm{Rec}_\Sigma(X)$.*

(2) *If φ is linear, then $T \in \mathrm{Rec}_\Sigma(X)$ implies $T\varphi \in \mathrm{Rec}_\Omega(Y)$.*

7 Varieties of tree languages

Let Σ be again a ranked alphabet. A *family of regular Σ-tree languages* is a mapping $\mathcal{V}$ that assigns to every leaf alphabet X a set $\mathcal{V}(X)$ of regular ΣX-tree languages. We write such a family as $\mathcal{V} = \{\mathcal{V}(X)\}_X$. For any two families of regular Σ-tree languages $\mathcal{U}$ and $\mathcal{V}$, let us set $\mathcal{U} \subseteq \mathcal{V}$ if and only if $\mathcal{U}(X) \subseteq \mathcal{V}(X)$ for every X. Unions and intersections of families of regular Σ-tree languages are defined by similar componentwise conditions.

7.1 Definition A *variety of Σ-tree languages* (Σ-VTL) is a family of regular Σ-tree languages $\mathcal{V} = \{\mathcal{V}(X)\}_X$ such that for all leaf alphabets X and Y,

(1) $\mathcal{V}(X)$ is a Boolean subalgebra of $\mathrm{Rec}_\Sigma(X)$,

(2) $T \in \mathcal{V}(X)$ implies $p^{-1}(T) \in \mathcal{V}(X)$ for any $p \in C_\Sigma(X)$, and

(3) $T \in \mathcal{V}(Y)$ implies $T\varphi^{-1} \in \mathcal{V}(X)$ for any homomorphism $\varphi : \mathcal{T}_\Sigma(X) \to \mathcal{T}_\Sigma(Y)$.

Let $\mathbf{VTL}(\Sigma)$ denote the class of all Σ-VTLs.

The defining closure properties of varieties of Σ-tree languages correspond in a natural way to those defining a $*$-variety with inverse translations replacing quotient operations. Note, however, that we require every component $\mathcal{V}(X)$ of a Σ-VTL $\mathcal{V}$ to be nonempty.

It is clear that the intersection of any family of Σ-VTLs is also a Σ-VTL. Hence, for any family of regular Σ-tree languages $\mathcal{U}$, there is a unique least Σ-VTL $\mathcal{U}^v$ such that $\mathcal{U} \subseteq \mathcal{U}^v$, the Σ-VTL *generated* by $\mathcal{V}$. The Σ-VTL generated by a given family of regular Σ-tree languages can be described as follows.

7.2 Proposition *Let $\mathcal{U}$ be a family of regular Σ-tree languages. For any leaf alphabet X, $\mathcal{U}^v(X)$ is the Boolean closure of the set all ΣX-tree languages $p^{-1}(T)\varphi^{-1}$, where $T \in \mathcal{U}(Y)$, $p \in C_\Sigma(Y)$, and $\varphi : \mathcal{T}_\Sigma(X) \to \mathcal{T}_\Sigma(Y)$ is a homomorphism, for some leaf alphabet Y.*

Proof Let $\mathcal{V} = \{\mathcal{V}(X)\}_X$ be the family of regular Σ-tree languages where, for each X, $\mathcal{V}(X)$ is the Boolean closure defined in the proposition. It is then clear that $\mathcal{U} \subseteq \mathcal{V} \subseteq \mathcal{U}^v$. Indeed, if $T \in \mathcal{U}(X)$, then $T = \xi^{-1}(T)1_{T_\Sigma(X)}^{-1} \in \mathcal{V}(X)$, and the inclusion $\mathcal{V} \subseteq \mathcal{U}^v$ is a consequence of the defining closure properties of Σ-VTLs. Hence, it suffices to verify that $\mathcal{V}$ is a Σ-VTL.

Of course, $\mathcal{V}$ satisfies Condition (1) of Definition 7.1 as each $\mathcal{V}(X)$ is defined as a Boolean closure. Because both inverse translations and inverse homomorphisms commute with all Boolean operations, i.e., $p^{-1}(S \cup T) = p^{-1}(S) \cup p^{-1}(T)$, $p^{-1}(S - T) = p^{-1}(S) - p^{-1}(T)$ etc., it suffices to verify Conditions (2) and (3) for the generating sets of the Boolean closures defining $\mathcal{V}$. Hence let us consider a set $S = p^{-1}(T)\varphi^{-1} \in \mathcal{V}(X)$, where $T \in \mathcal{U}(Y)$, $p \in C_\Sigma(Y)$ and $\varphi : \mathcal{T}_\Sigma(X) \to \mathcal{T}_\Sigma(Y)$ is a homomorphism, for some Y.

To verify Condition (2), consider any $q \in C_\Sigma(X)$. By Lemma 8.5 (to be presented in the following section) there is a ΣY-context q_φ such that $q(t)\varphi = q_\varphi(t\varphi)$ for every $t \in T_\Sigma(X)$. If we write $r = q_\varphi \cdot p$, it is easy to see that $q^{-1}(S) = r^{-1}(T)\varphi^{-1} \in \mathcal{V}(X)$.

For any homomorphism $\psi : \mathcal{T}_\Sigma(Z) \to \mathcal{T}_\Sigma(X)$, where Z is some leaf alphabet, $S\psi^{-1} = p^{-1}(T)(\psi\varphi)^{-1} \in \mathcal{V}(Z)$, and therefore $\mathcal{V}$ also satisfies Condition (3). □

It is clear that $\mathbf{VTL}(\Sigma)$ is closed under unrestricted intersections, and that the union of any directed family of Σ-VTLs is also a Σ-VTL. Hence the following proposition.

7.3 Proposition *For any ranked alphabet Σ, $(\mathbf{VTL}(\Sigma), \subseteq)$ is an algebraic lattice such that $\inf(\mathcal{F}) = \bigcap \mathcal{F}$ and $\sup(\mathcal{F}) = (\bigcup \mathcal{F})^v$, for every $\mathcal{F} \subseteq \mathbf{VTL}(\Sigma)$.*

8 Syntactic congruences and syntactic algebras

We shall see that varieties of Σ-tree languages can be characterized in terms of *syntactic algebras*. Following [78] these are defined for subsets of general algebras in such a way that for free monoids and semigroups generated by finite alphabets, we get the syntactic monoids and semigroups used in Eilenberg's variety theory. The starting point is the following general notion of recognizability [52] motivated by facts like Corollary 3.5 and Proposition 5.5.

8.1 Definition A subset $L \subseteq A$ of an algebra $\mathcal{A} = (A, \Sigma)$ is said to be *recognizable* if it is recognized by a finite Σ-algebra, i.e. there exist a finite algebra $\mathcal{B} = (B, \Sigma)$, a homomorphism $\varphi : \mathcal{A} \to \mathcal{B}$ and a subset $F \subseteq B$ such that $L = F\varphi^{-1}$. Let $\mathrm{Rec}(\mathcal{A})$ denote the set of all recognizable subsets of $\mathcal{A}$.

This notion could also be naturally defined in terms of congruences: $L \subseteq A$ is recognizable in $\mathcal{A} = (A, \Sigma)$ if and only if L is saturated by a finite congruence on $\mathcal{A}$. Of course, $\mathrm{Rec}(X^*) = \mathrm{Rec}(X)$ and $\mathrm{Rec}(\mathcal{T}_\Sigma(X)) = \mathrm{Rec}_\Sigma(X)$ for any X and any Σ. It is easy to see that $\mathrm{Rec}(\mathcal{A})$ is a field of sets on A, and some of the other closure properties of regular string or tree languages noted above also extend to general algebras (cf. [13, 82] for details and further references).

8.2 Definition The *syntactic congruence* of a subset $L \subseteq A$ of a Σ-algebra $\mathcal{A} = (A, \Sigma)$ is the relation θ_L on A defined by the condition that for any $a, b \in A$,

$$a \,\theta_L\, b \quad \text{if and only if} \quad (\forall p \in \mathrm{Tr}(\mathcal{A}))[p(a) \in L \iff p(b) \in L].$$

The *syntactic algebra* of L is the quotient algebra $\mathcal{A}/L := \mathcal{A}/\theta_L = (A/L, \Sigma)$, where $A/L := A/\theta_L = \{a/L \mid a \in A\}$. The natural homomorphism $\theta_L^\natural : \mathcal{A} \to \mathcal{A}/L$, $a \mapsto a/L$, is called the *syntactic homomorphism* of L and it is denoted by φ_L.

The fundamental properties of syntactic congruences and algebras are as follows.

8.3 Proposition *Let $L \subseteq A$ be any subset of a Σ-algebra $\mathcal{A} = (A, \Sigma)$.*

(1) *The syntactic congruence θ_L is the greatest congruence on $\mathcal{A}$ saturating L.*

(2) *A Σ-algebra $\mathcal{B} = (B, \Sigma)$ recognizes L if and only if $\mathcal{A}/L \preceq \mathcal{B}$.*

(3) *$L \in \mathrm{Rec}(\mathcal{A})$ if and only if $\mathcal{A}/L$ is finite.*

Let us also note the following fact about syntactic congruences.

8.4 Lemma *Every congruence ϱ on any algebra $\mathcal{A} = (A, \Sigma)$ is the intersection of syntactic congruences. In particular, $\varrho = \bigcap\{\theta_{a/\varrho} \mid a \in A\}$.*

The following lemma is frequently needed.

8.5 Lemma *Let $\mathcal{A} = (A, \Sigma)$ and $\mathcal{B} = (B, \Sigma)$ be Σ-algebras, and let $\varphi : \mathcal{A} \to \mathcal{B}$ be a homomorphism. For any translation $p \in \mathrm{Tr}(\mathcal{A})$ there is a translation $p_\varphi \in \mathrm{Tr}(\mathcal{B})$ such that $p(a)\varphi = p_\varphi(a\varphi)$ for every $a \in A$. If φ is surjective, then there is for every $q \in \mathrm{Tr}(\mathcal{B})$ a translation $p \in \mathrm{Tr}(\mathcal{A})$ such that $p_\varphi = q$.*

Proof For $p = 1_A$, let $p_\varphi = 1_B$. For an elementary translation $p(\xi) = f^{\mathcal{A}}(a_1, \ldots, \xi, \ldots, a_m)$, we may set $p_\varphi(\xi) = f^{\mathcal{B}}(a_1\varphi, \ldots, \xi, \ldots, a_m\varphi)$. If φ is surjective, every elementary translation of $\mathcal{B}$ can be written in this form. The proposition follows now from the fact that the remaining translations are compositions of elementary translations. □

8.6 Proposition *Let $\mathcal{A} = (A, \Sigma)$ and $\mathcal{B} = (B, \Sigma)$ be Σ-algebras.*

(1) *$\theta_{A-L} = \theta_L$, for any $L \subseteq A$.*

(2) *$\theta_K \cap \theta_L \subseteq \theta_{K \cap L}$, for any $K, L \subseteq A$.*

(3) *$\theta_L \subseteq \theta_{p^{-1}(L)}$, for any $L \subseteq A$ and any $p \in \mathrm{Tr}(\mathcal{A})$.*

(4) $\varphi \circ \theta_L \circ \varphi^{-1} \subseteq \theta_{L\varphi^{-1}}$, *for any homomorphism* $\varphi : \mathcal{A} \to \mathcal{B}$ *and any* $L \subseteq B$, *and equality holds if* φ *is an epimorphism.*

Proof The first three claims follow directly from the definition of syntactic congruences. Let us note how Lemma 8.5 yields (4): for any $a, b \in A$,

$$\begin{aligned} a\varphi \circ \theta_L \circ \varphi^{-1} b &\iff a\varphi\, \theta_L\, b\varphi \\ &\implies (\forall p \in \mathrm{Tr}(\mathcal{A}))[p_\varphi(a\varphi) \in L \leftrightarrow p_\varphi(b\varphi) \in L] \\ &\iff (\forall p \in \mathrm{Tr}(\mathcal{A}))[p(a) \in L\varphi^{-1} \leftrightarrow p(b) \in L\varphi^{-1}] \\ &\iff a\theta_{L\varphi^{-1}} b. \end{aligned}$$

Moreover if φ is an epimorphism, then the only implication also becomes an equivalence. □

The following proposition [78] generalizes some basic facts presented in [18] for semigroups replacing quotients operations by inverse translations.

8.7 Proposition *Let* $\mathcal{A} = (A, \Sigma)$ *and* $\mathcal{B} = (B, \Sigma)$ *be* Σ*-algebras.*

(1) $\mathcal{A}/(A - L) = \mathcal{A}/L$, *for any* $L \subseteq A$.

(2) $\mathcal{A}/K \cap L \preceq \mathcal{A}/K \times \mathcal{A}/L$, *for any* $K, L \subseteq A$.

(3) $\mathcal{A}/p^{-1}(L) \preceq \mathcal{A}/L$, *for any* $L \subseteq A$ *and any* $p \in \mathrm{Tr}(\mathcal{A})$.

(4) $\mathcal{A}/L\varphi^{-1} \preceq \mathcal{B}/L$, *for any homomorphism* $\varphi : \mathcal{A} \to \mathcal{B}$ *and any* $L \subseteq B$. *If* φ *is an epimorphism, then* $\mathcal{A}/L\varphi^{-1} \cong \mathcal{B}/L$.

Proof The first assertion follows directly from Proposition 8.6 (1). For (2) we note first that $B := \{(a/K, a/L) \mid a \in A\}$ defines a subalgebra $\mathcal{B} = (B, \Sigma)$ of $\mathcal{A}/K \times \mathcal{A}/L$. Since $\theta_K \cap \theta_L \subseteq \theta_{K \cap L}$, the mapping $\varphi : B \to A/K \cap L$, $(a/K, a/L) \mapsto a/K \cap L$, is well-defined, and it is easy to see that it is an epimorphism from $\mathcal{B}$ onto $\mathcal{A}/K \cap L$. Similarly (3) is obtained by noting that $\mathcal{A}/p^{-1}(L)$ is an epimorphic image of $\mathcal{A}/L$ since $\theta_L \subseteq \theta_{p^{-1}(L)}$.

Finally to prove (4) we assume first that φ is an epimorphism. Then one can verify that

$$\psi : A/L\varphi^{-1} \to B/L,\ a/L\varphi^{-1} \mapsto a\varphi/L,$$

is an isomorphism from $\mathcal{A}/L\varphi^{-1}$ onto $\mathcal{B}/L$. Firstly, that ψ is well-defined and injective follows by Lemma 8.5 thus: for any $a, b \in A$,

$$\begin{aligned} a/L\varphi^{-1} = b/L\varphi^{-1} &\iff (\forall p \in \mathrm{Tr}(\mathcal{A})[p(a) \in L\varphi^{-1} \leftrightarrow p(b) \in L\varphi^{-1}] \\ &\iff (\forall p \in \mathrm{Tr}(\mathcal{A})[p_\varphi(a\varphi) \in L \leftrightarrow p_\varphi(b\varphi) \in L] \\ &\iff (\forall q \in \mathrm{Tr}(\mathcal{B})[q(a\varphi) \in L \leftrightarrow q(b\varphi) \in L] \\ &\iff a\varphi/L = b\varphi/L \end{aligned}$$

Since φ is surjective, ψ is also surjective. Moreover, ψ is a homomorphism:

$$c^{\mathcal{A}/L\varphi^{-1}}\psi = (c^{\mathcal{A}}/L\varphi^{-1})\psi = c^{\mathcal{A}}\varphi/L = c^{\mathcal{B}}/L = c^{\mathcal{B}/L},$$

for every $c \in \Sigma_0$, and similarly

$$f^{\mathcal{A}/L\varphi^{-1}}(a_1/L\varphi^{-1}, \ldots, a_m/L\varphi^{-1})\psi = \cdots = f^{\mathcal{B}/L}((a_1/L\varphi^{-1})\psi, \ldots, (a_m/L\varphi^{-1})\psi),$$

for all $m > 0$, $f \in \Sigma_m$ and $a_1, \ldots, a_m \in A$. Hence, $\mathcal{A}/L\varphi^{-1} \cong \mathcal{B}/L$.

Consider now the general case of an arbitrary homomorphism $\varphi : \mathcal{A} \to \mathcal{B}$. Let $\mathcal{C} = \mathcal{A}\varphi\varphi_L$ be the image subalgebra in $\mathcal{B}/L$ of $\mathcal{A}$ under the homomorphism $\varphi\varphi_L : \mathcal{A} \to \mathcal{B}/L$. Then $\eta : \mathcal{A} \to \mathcal{C}$, $a \mapsto a\varphi\varphi_L$, is an epimorphism, and hence $\mathcal{A}/L\varphi^{-1}\eta\eta^{-1} \cong \mathcal{C}/L\varphi^{-1}\eta$ by the previous part of the proof. Since $L\varphi^{-1}\eta\eta^{-1} = L\varphi^{-1}$ and $\mathcal{C}$ is a subalgebra of $\mathcal{B}/L$, this implies that $\mathcal{A}/L\varphi^{-1} \preceq \mathcal{B}/L$ as required. □

The essence of the proposition is that the syntactic algebra of a subset obtained by one of the operations defining varieties of tree languages is always in every variety of algebras containing the syntactic algebras of the original subsets. The following technical lemma will also be needed.

8.8 Lemma *If $\varphi : \mathcal{A} \to \mathcal{B}$ is a homomorphism and $L \subseteq B$ is any subset of $\mathcal{B}$, then*

$$\varphi \circ \theta_L \circ \varphi^{-1} = \bigcap\{\theta_{p^{-1}(L)\varphi^{-1}} \mid p \in \mathrm{Tr}(\mathcal{B})\}.$$

Proof Let ρ denote the right hand side of the claimed equality. For every $p \in \mathrm{Tr}(\mathcal{B})$,

$$\varphi \circ \theta_L \circ \varphi^{-1} \subseteq \varphi \circ \theta_{p^{-1}(L)}\varphi^{-1} \subseteq \theta_{p^{-1}(L)\varphi^{-1}}$$

by Proposition 8.6. Hence $\varphi \circ \theta_L \circ \varphi^{-1} \subseteq \rho$. The converse holds too: for any $a, b \in A$,

$$\begin{aligned}
a\,\rho\,b &\implies (\forall p \in \mathrm{Tr}(\mathcal{B}))\, a\,\theta_{p^{-1}(L)\varphi^{-1}}\, b \\
&\implies (\forall p \in \mathrm{Tr}(\mathcal{B}))(\forall q \in \mathrm{Tr}(\mathcal{A}))\,[q(a) \in p^{-1}(L)\varphi^{-1} \;\leftrightarrow\; q(b) \in p^{-1}(L)\varphi^{-1}] \\
&\implies (\forall p \in \mathrm{Tr}(\mathcal{B}))(\forall q \in \mathrm{Tr}(\mathcal{A}))\,[q(a)\varphi \in p^{-1}(L) \;\leftrightarrow\; q(b)\varphi \in p^{-1}(L)] \\
&\implies (\forall p \in \mathrm{Tr}(\mathcal{B}))(\forall q \in \mathrm{Tr}(\mathcal{A}))\,[p(q_\varphi(a\varphi)) \in L \;\leftrightarrow\; p(q_\varphi(b\varphi)) \in L] \\
&\implies (\forall p \in \mathrm{Tr}(\mathcal{B}))[p(a\varphi) \in L \;\leftrightarrow\; p(b\varphi) \in L] \\
&\implies a\,\varphi \circ \theta_L \circ \varphi^{-1}\, b.
\end{aligned}$$

□

An algebra is called *syntactic* if it is isomorphic to the syntactic algebra of a subset of some algebra. A subset L of an algebra $\mathcal{A} = (A, \Sigma)$ is *disjunctive* if $\theta_L = \Delta_A$. The following fact was first observed for semigroups in [76], and the general form can be found in [78, 80].

8.9 Proposition *An algebra is syntactic if and only if it has a disjunctive subset.*

We should also mention the paper [41] where Kelarev and Sokratova describe the graphs such that the corresponding *graph algebras* are syntactic algebras.

If $\mathcal{A} = (A, \Sigma)$ is subdirectly irreducible, then $\{a\}$ is disjunctive for at least one $a \in A$ because $\bigcap\{\theta_{\{a\}} \mid a \in A\} = \Delta_A$ by Lemma 8.4. Hence the following result.

8.10 Proposition *Every subdirectly irreducible algebra is syntactic.*

A simple counter-example to the converse of Proposition 8.10 is provided by the 3-element mono-unary algebra $\mathcal{A} = (\{1, 2, 3\}, f)$ defined by $f^{\mathcal{A}}(1) = f^{\mathcal{A}}(2) = 2$ and $f^{\mathcal{A}}(3) = 3$; it has a proper subdirect decomposition, but $\{1, 3\}$ is a disjunctive subset.

Let us now consider syntactic congruences and syntactic algebras of tree languages. Since the Nerode congruence of a ΣX-tree language T is the greatest congruence on $\mathcal{T}_\Sigma(X)$ saturating T, it is exactly the syntactic congruence of T as a subset of the ΣX-term algebra $\mathcal{T}_\Sigma(X)$. We already denoted it earlier by θ_T, but from now on θ_T will be called the *syntactic congruence* of T. The fact that the translations of $\mathcal{T}_\Sigma(X)$ are precisely the mappings $t \mapsto p(t)$ defined by ΣX-contexts yields the following description of θ_T.

8.11 Proposition *For any ΣX-tree language T and all $s, t \in T_\Sigma(X)$,*

$$s\,\theta_T\,t \quad \text{if and only if} \quad (\forall p \in C_\Sigma(X))[p(s) \in T \leftrightarrow p(t) \in T].$$

The proposition gives an intuitive meaning to the name "syntactic congruence": two ΣX-trees s and t are *syntactically equivalent* with respect to a given ΣX-tree language T if they appear in T in exactly the same contexts. The *syntactic algebra* $\mathcal{T}_\Sigma(X)/\theta_T$ of T is also denoted $\mathrm{SA}(T)$, and the *syntactic homomorphism* of T is the mapping $\varphi_T : \mathcal{T}_\Sigma(X) \to \mathrm{SA}(T), t \mapsto t/T$.

The following facts are immediate consequences of Propositions 5.5, 5.6 and 8.3.

8.12 Corollary *A ΣX-tree language T is regular if and only if its syntactic algebra $\mathrm{SA}(T)$ is finite. Moreover, a Σ-algebra $\mathcal{A}$ recognizes T if and only if $\mathrm{SA}(T) \preceq \mathcal{A}$. Hence, $\mathrm{SA}(T)$ yields the minimal ΣX-recognizer for any $T \in \mathrm{Rec}_\Sigma(X)$.*

By combining Proposition 6.7 with Lemma 8.8, we obtain the following result.

8.13 Corollary *Let $\varphi : \mathcal{T}_\Sigma(X) \to \mathcal{T}_\Sigma(Y)$ be a homomorphism of term algebras, where X and Y are leaf alphabets. If $T \in \mathrm{Rec}_\Sigma(Y)$, then*

$$\varphi \circ \theta_T \circ \varphi^{-1} = \theta_{p_1^{-1}(T)\varphi^{-1}} \cap \cdots \cap \theta_{p_k^{-1}(T)\varphi^{-1}},$$

for some finite set $\{p_1, \ldots, p_k\}$ of ΣY-contexts.

We will use finite syntactic algebras only, and they can be described as follows.

8.14 Proposition *A finite algebra is syntactic if and only if it is isomorphic to the syntactic algebra of a regular tree language. More precisely: if $\mathcal{A} = (A, \Sigma)$ is a finite algebra and X is any leaf alphabet such that $\mathcal{A}$ is generated by a subset of size $\leq |X|$, then $\mathcal{A}$ is a syntactic algebra if and only if $\mathcal{A} \cong \mathrm{SA}(T)$ for some $T \in \mathrm{Rec}_\Sigma(X)$.*

Proof Since the condition is clearly sufficient, let us assume that $\mathcal{A}$ is syntactic. Then it has a disjunctive subset $D \subseteq A$. Let $\alpha : X \to A$ be any mapping such that $X\alpha$ contains a generating set of $\mathcal{A}$; such a mapping exists by our assumptions. Then the homomorphic extension $\alpha_\mathcal{A} : \mathcal{T}_\Sigma(X) \to \mathcal{A}$ of α is an epimorphism. By Proposition 8.7 this means that for $T = D\alpha_\mathcal{A}^{-1}$, we get $\mathrm{SA}(T) \cong \mathcal{A}/D \cong \mathcal{A}$. □

9 Varieties of finite algebras

The *pseudovarieties* of finite monoids or semigroups introduced in [19] form a central ingredient of Eilenberg's variety theory [18]. The same notion applied to general algebras of any finite type, plays the corresponding part in the theory of tree language varieties.

We begin by introducing a new algebra class operator in addition to the operators S, H and P defined in Section 2. For any class $\mathbf{K}$ of Σ-algebras, the class of all algebras isomorphic to a direct product $\mathcal{A}_1 \times \cdots \times \mathcal{A}_n$ of a finite sequence $\mathcal{A}_1, \ldots, \mathcal{A}_n$ of members of $\mathbf{K}$ is denoted by $\mathrm{P_f}(\mathbf{K})$. Note that $\mathrm{P_f}(\mathbf{K})$ always contains all trivial Σ-algebras obtained by setting $n = 0$.

9.1 Definition A class of finite Σ-algebras $\mathbf{K}$ is a *variety of finite Σ-algebras*, a Σ-VFA for short, if it is closed under the formation of subalgebras, homomorphic images and finite direct products, i.e., if $\mathrm{S}(\mathbf{K}), \mathrm{H}(\mathbf{K}), \mathrm{P_f}(\mathbf{K}) \subseteq \mathbf{K}$. Let $\mathbf{VFA}(\Sigma)$ denote the class of all Σ-VFAs.

Note that by our definition, every Σ-VFA is non-empty since it contains at least the trivial Σ-algebras The intersection of any class of varieties of finite Σ-algebras is obviously also a Σ-VFA. The Σ-VFA *generated* by a given class $\mathbf{K}$ of Σ-algebras is the least Σ-VFA containing $\mathbf{K}$ as a subclass and it is denoted $\mathrm{V_f}(\mathbf{K})$. It is also clear that the union of any directed family of Σ-VFAs is a Σ-VFA. These observations lead to the following proposition.

9.2 Proposition *For any ranked alphabet Σ, $(\mathbf{VFA}(\Sigma), \subseteq)$ is an algebraic lattice in which $\inf \mathcal{K} = \bigcap \mathcal{K}$ and $\sup \mathcal{K} = \mathrm{V_f}(\bigcup \mathcal{K})$ for every $\mathcal{K} \subseteq \mathbf{VFA}(\Sigma)$.*

Clearly the class $\mathrm{F}(\mathbf{V})$ of all finite members of any variety $\mathbf{V}$ of Σ-algebras is a Σ-VFA. However, not all varieties of finite Σ-algebras are obtained this way. Ash [4] has shown that a class $\mathbf{K}$ of finite Σ-algebras is a Σ-VFA exactly in case $\mathbf{K} = \mathrm{F}(\mathbf{V})$ for some generalized variety $\mathbf{V}$ of Σ-algebras; a *generalized variety* is a class of algebras closed under the formation of subalgebras, homomorphic images, finite direct products and (unrestricted) direct powers.

A simple modification of the well-known decomposition $\mathrm{V} = \mathrm{HSP}$ of the generated variety operator yields $\mathrm{V_f}(\mathbf{K}) = \mathrm{HS}\,\mathrm{P_f}(\mathbf{K})$. In terms of the covering relation this fact may be expressed as follows.

9.3 Proposition *If $\mathbf{K}$ is any class of finite Σ-algebras and $\mathcal{A}$ is a finite Σ-algebra, then $\mathcal{A} \in \mathrm{V_f}(\mathbf{K})$ if and only if $\mathcal{A} \preceq \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$, for some $n \geq 0$ and $\mathcal{A}_1, \ldots, \mathcal{A}_n \in \mathbf{K}$.*

Proposition 8.10 yields the following useful fact.

9.4 Lemma *Every Σ-VFA is generated by the syntactic algebras it contains.*

A Σ-VFA $\mathbf{K}$ is called *equational* if there is a set I of identities such that $\mathbf{K}$ is the class of finite Σ-algebras satisfying I, that is to say, $\mathbf{K} = \mathrm{F}(\mathrm{Mod}(I))$. In other words, a Σ-VFA $\mathbf{K}$ is equational if and only if $\mathbf{K} = \mathrm{F}(\mathbf{V})$ for some variety $\mathbf{V}$ of Σ-algebras. However, every variety of finite Σ-algebras can be defined by identities in the following sense.

9.5 Definition Let E be a countable sequence $u_0 \simeq v_0, u_1 \simeq v_1, u_2 \simeq v_2, \ldots$ of Σ-identities (over some set of variables). A Σ-algebra $\mathcal{A}$ is said to *ultimately satisfy* E, and this we express by writing $\mathcal{A} \models_u E$, if there is a number $n_0 \geq 0$ such that $\mathcal{A} \models u_n \simeq v_n$ for every $n \geq n_0$. The class of all finite Σ-algebras that ultimately satisfy E, is denoted by $\mathrm{Mod}_u(E)$. The class $\mathrm{Mod}_u(E)$ is said to be *ultimately defined* by the sequence E.

Since the subalgebras, homomorphic images and direct products of any given algebras satisfy all identities satisfied by these algebras, it is clear that $\mathrm{Mod}_u(E)$ is a Σ-VFA for every sequence E of identities, but the converse also holds. This is expressed by the following theorem by Eilenberg and Schützenberger [19] (cf. also [18]). The original proof was presented for monoids, but it can easily be extended to any varieties of finite algebras.

9.6 Theorem *For any ranked alphabet Σ, a class of finite Σ-algebras is a Σ-VFA if and only if it is ultimately defined by some sequence of identities.*

Let E be a sequence $u_0 \simeq v_0, u_1 \simeq v_1, u_2 \simeq v_2, \dots$ of Σ-identities. For each $n \geq 0$, let $E_n = \{u_n \simeq v_n, u_{n+1} \simeq v_{n+1}, \dots\}$ and let $\mathbf{V}_n = \mathrm{Mod}(E_n)$ be the variety of Σ-algebras defined by E_n. Clearly, $\mathbf{V}_0 \subseteq \mathbf{V}_1 \subseteq \mathbf{V}_2 \subseteq \dots$ and $\mathrm{Mod}_u(E) = \bigcup_{n\geq 0} \mathrm{F}(\mathbf{V}_n)$. On the other hand, it is clear that $\bigcup_{n\geq 0} \mathrm{F}(\mathbf{V}_n)$ is a Σ-VFA for any ascending chain $\mathbf{V}_0 \subseteq \mathbf{V}_1 \subseteq \mathbf{V}_2 \subseteq \dots$ of varieties of Σ-algebras. Hence we obtain the following characterization of varieties of finite algebras proved in a different way by Baldwin and Berman [5].

9.7 Proposition *For any ranked alphabet Σ, a class $\mathbf{K}$ of finite Σ-algebras is a Σ-VFA if and only if $\mathbf{K} = \bigcup_{n\geq 0} \mathrm{F}(\mathbf{V}_n)$ for some ascending chain $\mathbf{V}_0 \subseteq \mathbf{V}_1 \subseteq \mathbf{V}_2 \subseteq \dots$ of varieties of Σ-algebras.*

Proposition 9.7 can also be expressed as follows.

9.8 Corollary *For any ranked alphabet Σ, a class $\mathbf{K}$ of finite Σ-algebras is a Σ-VFA if and only if $\mathbf{K} = \bigcup_{n\geq 0} \mathbf{K}_n$ for some ascending chain $\mathbf{K}_0 \subseteq \mathbf{K}_1 \subseteq \mathbf{K}_2 \subseteq \dots$ of equational varieties of finite Σ-algebras.*

For further results on varieties of finite algebras the reader is referred to [1, 4, 5, 70], for example. A systematic study of the topic and related matters, as well many further relevant references, can be found in the monograph [2] of J. Almeida.

10 Varieties of finite congruences

A variety of tree languages may be characterized by a corresponding variety of finite algebras, but it is often most conveniently defined by giving a variety of congruences of the kind to be introduced in this section.

For any ranked alphabet Σ and any leaf alphabet X, let $\mathrm{FCon}_\Sigma(X)$ denote the set of all congruences on $\mathcal{T}_\Sigma(X)$ of finite index. Clearly, $\mathrm{FCon}_\Sigma(X)$ is a filter, and hence also a sublattice, of the congruence lattice $\mathrm{Con}(\mathcal{T}_\Sigma(X))$. By a *family of finite Σ-congruences* we mean a mapping Γ that assigns to each leaf alphabet X a subset $\Gamma(X)$ of $\mathrm{FCon}_\Sigma(X)$, and we write it as $\Gamma = \{\Gamma(X)\}_X$. Inclusions, unions and intersections of families of finite Σ-congruences are defined in the natural componentwise manner.

10.1 Definition A *variety of finite Σ-congruences*, a Σ-VFC for short, is a family of finite Σ-congruences $\Gamma = \{\Gamma(X)\}_X$ such that for all leaf alphabets X and Y,

(1) $\Gamma(X)$ is a filter of $\mathrm{FCon}_\Sigma(X)$, and

(2) for any homomorphism $\varphi : \mathcal{T}_\Sigma(X) \to \mathcal{T}_\Sigma(Y)$, $\theta \in \Gamma(Y)$ implies $\varphi \circ \theta \circ \varphi^{-1} \in \Gamma(X)$.

Let $\mathbf{VFC}(\Sigma)$ denote the class of all Σ-VFCs.

The intersection of any family of Σ-VFCs is obviously again a Σ-VFC, and the Σ-VFC *generated* by a family of finite Σ-congruences Θ is the Σ-VFC

$$\mathrm{VFC}(\Theta) = \bigcap \{\Gamma \in \mathrm{FCon}_\Sigma(X) \mid \Theta \subseteq \Gamma\}.$$

Since the union of any directed family of Σ-VFCs is also a Σ-VFC, the following holds.

10.2 Proposition *For any ranked alphabet Σ, $(\mathbf{VFC}(\Sigma), \subseteq)$ is an algebraic lattice such that $\inf \mathcal{C} = \bigcap \mathcal{C}$ and $\sup \mathcal{C} = \mathrm{VFC}(\bigcup \mathcal{C})$, for every $\mathcal{C} \subseteq \mathbf{VFC}(\Sigma)$.*

By Lemma 8.4, $\theta = \bigcap\{\theta_{t/\theta} \mid t \in T_\Sigma(X)\}$ for any $\theta \in \mathrm{Con}(\mathcal{T}_\Sigma(X))$. If $\theta \in \mathrm{FCon}_\Sigma(X)$, the intersection involves just a finite number of different θ-classes. This observation yields the following counterpart to Proposition 8.10. Recall that an element a of a lattice is *meet-irreducible*, if $a = b \wedge c$ implies $a = b$ or $a = c$.

10.3 Proposition *Every meet irreducible congruence in $\mathrm{FCon}_\Sigma(X)$ is the syntactic congruence of a regular ΣX-tree language.*

Moreover, since the filter $[H)$ generated by a non-empty subset H of any lattice L is the set $\{a \in L \mid (\exists n > 0)(\exists b_1, \ldots, b_n \in H)\, b_1 \wedge \cdots \wedge b_n \leq a\}$ (cf. [49, p. 48], for example), the above representation of a congruence $\theta \in \mathrm{FCon}_\Sigma(X)$ gives also the following important fact.

10.4 Proposition *Any Σ-VFC is generated by the syntactic congruences it contains.*

We call a Σ-VFC Γ *principal*, if for each leaf alphabet X, $\Gamma(X) = [\gamma_X)$ for some congruence $\gamma_X \in \mathrm{FCon}_\Sigma(X)$. The following characterization of these Σ-VFCs is easily verified.

10.5 Proposition *Assume that we are given a congruence $\gamma_X \in \mathrm{FCon}_\Sigma(X)$ for each leaf alphabet X. Then $\Gamma = \{[\gamma_X)\}_X$ is a principal Σ-VFC if and only if $\gamma_X \subseteq \varphi \circ \gamma_Y \circ \varphi^{-1}$ for all X and Y and every homomorphism $\varphi : \mathcal{T}_\Sigma(X) \to \mathcal{T}_\Sigma(Y)$.*

By applying the above condition to all endomorphisms $\varphi : \mathcal{T}_\Sigma(X) \to \mathcal{T}_\Sigma(X)$, the following result is obtained.

10.6 Corollary *If $\Gamma = \{[\gamma_X)\}_X$ is a principal Σ-VFC, then for each X, γ_X is a fully invariant congruence on $\mathcal{T}_\Sigma(X)$.*

Finally, let us mention the following fact noted by V. Piirainen (personal communication).

10.7 Lemma *The principal Σ-VFCs form a sublattice of $(\mathbf{VFC}(\Sigma), \subseteq)$ in which $\Gamma \vee \Delta = \{[\gamma_X \wedge \theta_X)\}_X$ and $\Gamma \wedge \Theta = \{[\gamma_X \vee \theta_X)\}_X$ for any Σ-VFCs $\Gamma = \{[\gamma_X)\}_X$ and $\Theta = \{[\theta_X)\}_X$.*

11 The Variety Theorem

In this section we present a variety theorem for tree languages. It establishes, for any given ranked alphabet Σ, bijective correspondences between varieties of Σ-tree languages, varieties of finite Σ-algebras and varieties of finite Σ-congruences via six isomorphisms between the lattices $(\mathbf{VTL}(\Sigma), \subseteq)$, $(\mathbf{VFA}(\Sigma), \subseteq)$ and $(\mathbf{VFC}(\Sigma), \subseteq)$. These isomorphisms form three mutually inverse pairs, and any possible composition of two of them is a third isomorphism belonging to the set. We give the definitions in detail and present a few proofs to show the concordance of the various concepts involved and the relevance of some of the results presented above. However, for more complete proofs we have to refer the reader to [80, 84].

11.1 Definition For any class $\mathbf{K}$ of finite Σ-algebras, let $\mathbf{K}^t$ be the family of regular Σ-tree languages and $\mathbf{K}^c$ be the family of finite Σ-congruences such that for any leaf alphabet X,

(1) $\mathbf{K}^t(X) = \{T \subseteq T_\Sigma(X) \mid \mathrm{SA}(T) \in \mathbf{K}\}$, and

(2) $\mathbf{K}^c(X) = \{\theta \in \mathrm{FCon}_\Sigma(X) \mid \mathcal{T}_\Sigma(X)/\theta \in \mathbf{K}\}$.

11.2 Lemma *If* $\mathbf{K} \in \mathbf{VFA}(\Sigma)$*, then* $\mathbf{K}^t \in \mathbf{VTL}(\Sigma)$ *and* $\mathbf{K}^c \in \mathbf{VFC}(\Sigma)$.

Proof The claim $\mathbf{K}^t \in \mathbf{VTL}(\Sigma)$ follows by Proposition 8.7 when we use the fact that $\mathbf{K}$ is a Σ-VFA. For example, if $T \in \mathbf{K}^t(X)$, then $p^{-1}(T) \in \mathbf{K}^t(X)$ for any $p \in C_\Sigma(X)$ because $\mathrm{SA}(p^{-1}(T)) \preceq \mathrm{SA}(T)$ and $\mathrm{SA}(T) \in \mathbf{K}$ imply $\mathrm{SA}(p^{-1}(T)) \in \mathbf{K}$.

To prove $\mathbf{K}^c \in \mathbf{VFC}(\Sigma)$, consider any $\theta, \rho \in \mathrm{FCon}_\Sigma(X)$. If $\theta \subseteq \rho$ and $\theta \in \mathbf{K}^c(X)$, then $\rho \in \mathbf{K}^c(X)$ because $\mathcal{T}_\Sigma(X)/\rho \preceq \mathcal{T}_\Sigma(X)/\theta \in \mathbf{K}$ implies $\mathcal{T}_\Sigma(X)/\rho \in \mathbf{K}$. Moreover if $\theta, \rho \in \mathbf{K}^c(X)$, then $\theta \cap \rho \in \mathbf{K}^c(X)$ follows from $\mathcal{T}_\Sigma(X)/\theta \cap \rho \preceq \mathcal{T}_\Sigma(X)/\theta \times \mathcal{T}_\Sigma(X)/\rho$ and $\mathcal{T}_\Sigma(X)/\theta, \mathcal{T}_\Sigma(X)/\rho \in \mathbf{K}$. Because $\mathbf{K}^c(X)$ contains at least the universal relation $\nabla_{T_\Sigma(X)}$, this means that it is a filter in $\mathrm{FCon}_\Sigma(X)$.

To verify that $\mathbf{K}^c$ also satisfies the second condition of Definition 10.1, consider any homomorphism $\varphi : \mathcal{T}_\Sigma(X) \to \mathcal{T}_\Sigma(Y)$ and assume that $\theta \in \mathbf{K}^c(Y)$. Let $\rho = \varphi \circ \theta \circ \varphi^{-1}$. It is easy to verify that $t/\rho \mapsto t\varphi/\theta$ defines a monomorphism $\psi : \mathcal{T}_\Sigma(X)/\rho \to \mathcal{T}_\Sigma(Y)/\theta$. Because $\mathcal{T}_\Sigma(Y)/\theta \in \mathbf{K}$, this implies that $\mathcal{T}_\Sigma(X)/\rho \in \mathbf{K}$, and therefore $\rho \in \mathbf{K}^c(X)$. □

11.3 Definition For any family $\mathcal{V}$ of regular Σ-tree languages, let

$$\mathcal{V}^a = \mathrm{V_f}(\{\mathrm{SA}(T) \mid T \in \mathcal{V}(X) \text{ for some } X\})$$

be the Σ-VFA generated by the syntactic algebras of the tree languages belonging to $\mathcal{V}$, and let $\mathcal{V}^c$ be the family of finite Σ-congruences such that for any leaf alphabet X,

$$\mathcal{V}^c(X) = [\{\theta_T \mid T \in \mathcal{V}(X)\})$$

is the filter of $\mathrm{FCon}_\Sigma(X)$ generated by the syntactic congruences of the members of $\mathcal{V}(X)$.

11.4 Lemma *If* $\mathcal{V} \in \mathbf{VTL}(\Sigma)$*, then* $\mathcal{V}^a \in \mathbf{VFA}(\Sigma)$ *and* $\mathcal{V}^c \in \mathbf{VFC}(\Sigma)$.

Proof As everything else is obvious, it suffices to prove that $\mathcal{V}^c$ satisfies Condition (2) of Definition 10.1. Let $\varphi : \mathcal{T}_\Sigma(X) \to \mathcal{T}_\Sigma(Y)$ be a homomorphism and assume that $\theta \in \mathcal{V}^c(Y)$. Then $\theta_{T_1} \cap \cdots \cap \theta_{T_n} \subseteq \theta$ for some $T_1, \ldots, T_n \in \mathcal{V}(Y)$. For each $i = 1, \ldots, n$,

$$\varphi \circ \theta_{T_i} \circ \varphi^{-1} = \bigcap \{\theta_{p^{-1}(T_i)\varphi^{-1}} \mid p \in \mathrm{Tr}(\mathcal{T}_\Sigma(Y))\}$$

by Lemma 8.8. Every set $p^{-1}(T_i)\varphi^{-1}$ is in $\mathcal{V}(X)$ since $\mathcal{V}$ is a Σ-VTL, and by Proposition 6.7 the number of such sets is finite. Hence each $\varphi \circ \theta_{T_i} \circ \varphi^{-1}$ is in $\mathcal{V}^c(X)$, and therefore $\varphi \circ \theta_T \circ \varphi^{-1} \in \mathcal{V}^c(X)$ as required. □

Let us now introduce the remaining two transformations.

11.5 Definition For any family Γ of finite Σ-congruences, let

$$\Gamma^a = \mathrm{V_f}(\{\mathcal{T}_\Sigma(X)/\theta \mid \theta \in \Gamma(X) \text{ for some } X\}),$$

and let Γ^t be the family of regular Σ-tree languages such that for any leaf alphabet X,

$$\Gamma^t(X) = \{T \subseteq T_\Sigma(X) \mid \theta_T \in \Gamma(X)\}.$$

The first assertion of the following lemma holds by definition, and the second assertion is easily proved by using Proposition 8.6.

11.6 Lemma *If* $\Gamma \in \mathbf{VFC}(\Sigma)$, *then* $\Gamma^a \in \mathbf{VFA}(\Sigma)$ *and* $\Gamma^t \in \mathbf{VTL}(\Sigma)$.

The following facts are completely obvious.

11.7 Lemma *The six operations introduced above are isotone, that is to say*

(1) *for any* $\mathbf{K}, \mathbf{L} \in \mathbf{VFA}(\Sigma)$, $\mathbf{K} \subseteq \mathbf{L}$ *implies* $\mathbf{K}^t \subseteq \mathbf{L}^t$ *and* $\mathbf{K}^c \subseteq \mathbf{L}^c$,

(2) *for any* $\mathcal{U}, \mathcal{V} \in \mathbf{VTL}(\Sigma)$, $\mathcal{U} \subseteq \mathcal{V}$ *implies* $\mathcal{U}^a \subseteq \mathcal{V}^a$ *and* $\mathcal{U}^c \subseteq \mathcal{V}^c$, *and*

(3) *for any* $\Gamma, \Phi \in \mathbf{VFC}(\Sigma)$, $\Gamma \subseteq \Phi$ *implies* $\Gamma^a \subseteq \Phi^a$ *and* $\Gamma^t \subseteq \Phi^t$.

To show that the six mappings considered above define isomorphisms between the lattices $(\mathbf{VTL}(\Sigma), \subseteq)$, $(\mathbf{VFA}(\Sigma), \subseteq)$ and $(\mathbf{VFC}(\Sigma), \subseteq)$, it suffices now to prove that they form mutually inverse pairs. This is stated in the following lemma which we give without proof. We compose the transformations in the natural way from left to right. For example, for any Σ-VFA $\mathbf{K}$, $\mathbf{K}^{ta}$ means $(\mathbf{K}^t)^a$.

11.8 Lemma *For any* $\mathbf{K} \in \mathbf{VFA}(\Sigma)$, $\mathcal{V} \in \mathbf{VTL}(\Sigma)$ *and* $\Gamma \in \mathbf{VFC}(\Sigma)$,

(1) $\mathbf{K}^{ta} = \mathbf{K}$ *and* $\mathbf{K}^{ca} = \mathbf{K}$,

(2) $\mathcal{V}^{at} = \mathcal{V}$ *and* $\mathcal{V}^{ct} = \mathcal{V}$,

(3) $\Gamma^{ac} = \Gamma$ *and* $\Gamma^{tc} = \Gamma$, *and moreover*

(4) $\mathbf{K}^{ct} = \mathbf{K}^t$, $\mathbf{K}^{tc} = \mathbf{K}^c$, $\mathcal{V}^{ac} = \mathcal{V}^c$, $\mathcal{V}^{ca} = \mathcal{V}^a$, $\Gamma^{ta} = \Gamma^a$ *and* $\Gamma^{at} = \Gamma^t$.

The results of the section can be summarized as the following *Variety Theorem*.

11.9 Theorem *For any ranked alphabet* Σ, *the mappings*

$$\mathbf{VFA}(\Sigma) \to \mathbf{VTL}(\Sigma),\ \mathbf{K} \mapsto \mathbf{K}^t; \qquad \mathbf{VTL}(\Sigma) \to \mathbf{VFA}(\Sigma),\ \mathcal{V} \mapsto \mathcal{V}^a,$$

the mappings

$$\mathbf{VFA}(\Sigma) \to \mathbf{VFC}(\Sigma),\ \mathbf{K} \mapsto \mathbf{K}^c; \qquad \mathbf{VFC}(\Sigma) \to \mathbf{VFA}(\Sigma),\ \Gamma \mapsto \Gamma^a,$$

and the mappings

$$\mathbf{VTL}(\Sigma) \to \mathbf{VFC}(\Sigma),\ \mathcal{V} \mapsto \mathcal{V}^c; \qquad \mathbf{VTL}(\Sigma) \to \mathbf{VFC}(\Sigma),\ \Gamma \mapsto \Gamma^t,$$

form three pairs of mutually inverse isomorphisms between the algebraic lattices $(\mathbf{VFA}(\Sigma), \subseteq)$, $(\mathbf{VTL}(\Sigma), \subseteq)$ *and* $(\mathbf{VFC}(\Sigma), \subseteq)$. *Moreover,* $\mathbf{K}^{ct} = \mathbf{K}^t$, $\mathbf{K}^{tc} = \mathbf{K}^c$, $\mathcal{V}^{ac} = \mathcal{V}^c$, $\mathcal{V}^{ca} = \mathcal{V}^a$, $\Gamma^{ta} = \Gamma^a$ *and* $\Gamma^{at} = \Gamma^t$ *for all* $\mathbf{K} \in \mathbf{VFA}(\Sigma)$, $\mathcal{V} \in \mathbf{VTL}(\Sigma)$ *and* $\Gamma \in \mathbf{VFC}(\Sigma)$.

Many well-known varieties of tree languages, as well as the associated varieties of finite algebras and finite congruences, are actually unions of chains of subvarieties or, more generally, unions of directed families of subvarieties. Moreover, it is often natural to first define these subvarieties. Therefore, we note the following facts.

11.10 Proposition *Let $\mathcal{V} = \bigcup_{i\in I} \mathcal{V}_i$ be the union of a directed family $\{\mathcal{V}_i \mid i \in I\}$ of varieties of Σ-tree languages. Then*

(1) *$\{\mathcal{V}_i{}^a \mid i \in I\}$ is a directed family of Σ-VFAs such that $\bigcup_{i\in I} \mathcal{V}_i{}^a = \mathcal{V}^a$, and*

(2) *$\{\mathcal{V}_i{}^c \mid i \in I\}$ is a directed family of Σ-VFCs such that $\bigcup_{i\in I} \mathcal{V}_i{}^c = \mathcal{V}^c$.*

Proof The families $\{\mathcal{V}_i{}^a \mid i \in I\}$ and $\{\mathcal{V}_i{}^c \mid i \in I\}$ are directed because the maps $\mathcal{U} \mapsto \mathcal{U}^a$ and $\mathcal{U} \mapsto \mathcal{U}^c$ are isotone. Therefore $\mathbf{K} = \bigcup_{i\in I} \mathcal{V}_i{}^a$ is a Σ-VFA and $\Gamma = \bigcup_{i\in I} \mathcal{V}_i{}^c$ is a Σ-VFC, and it remains to be shown that $\mathbf{K} = \mathcal{V}^a$ and $\Gamma = \mathcal{V}^c$.

Of course, $\mathbf{K} \subseteq \mathcal{V}^a$ since $\mathcal{V}_i{}^a \subseteq \mathcal{V}^a$ for every $i \in I$. For the converse inclusion it suffices to recall that $\mathcal{V}^a$ is generated by syntactic algebras $\mathrm{SA}(T)$ with $T \in \mathcal{V}(X)$ for some X. Indeed, by the definition of $\mathcal{V}$, any such T is in $\mathcal{V}_i(X)$ for some $i \in I$, and therefore $\mathrm{SA}(T) \in \mathcal{V}_i{}^a \subseteq \mathbf{K}$.

The inclusion $\Gamma \subseteq \mathcal{V}^c$ is again obvious. If $\theta \in \mathcal{V}^c(X)$ for some X, then $\theta_{T_1} \wedge \cdots \wedge \theta_{T_n} \subseteq \theta$, where $n \geq 1$ and $T_1 \in \mathcal{V}_{i_1}, \ldots, T_n \in \mathcal{V}_{i_n}$ for some $i_1, \ldots, i_n \in I$. Now $\mathcal{V}_{i_1}, \ldots, \mathcal{V}_{i_n} \subseteq \mathcal{V}_i$ for some $i \in I$, and therefore $\theta \in \mathcal{V}_i{}^c(X) \subseteq \Gamma(X)$. Hence also $\mathcal{V}^c \subseteq \Gamma$ holds. □

Let us just note without proofs the corresponding propositions for directed families of Σ-VFAs and directed families of Σ-VFCs.

11.11 Proposition *Let $\mathbf{K} = \bigcup_{i\in I} \mathbf{K}_i$ be the union of a directed family $\{\mathbf{K}_i \mid i \in I\}$ of varieties of finite Σ-algebras. Then*

(1) *$\{\mathbf{K}_i{}^t \mid i \in I\}$ is a directed family of Σ-VTLs such that $\bigcup_{i\in I} \mathbf{K}_i{}^t = \mathbf{K}^t$, and*

(2) *$\{\mathbf{K}_i{}^c \mid i \in I\}$ is a directed family of Σ-VFCs such that $\bigcup_{i\in I} \mathbf{K}_i{}^c = \mathbf{K}^c$.*

11.12 Proposition *Let $\Gamma = \bigcup_{i\in I} \Gamma_i$ be the union of a directed family $\{\Gamma_i \mid i \in I\}$ of varieties of finite Σ-congruences. Then*

(1) *$\{\Gamma_i{}^a \mid i \in I\}$ is a directed family of Σ-VFAs such that $\bigcup_{i\in I} \Gamma_i{}^a = \Gamma^a$, and*

(2) *$\{\Gamma_i{}^t \mid i \in I\}$ is a directed family of Σ-VTLs such that $\bigcup_{i\in I} \Gamma_i{}^t = \Gamma^t$.*

Often each subvariety $\mathcal{V}_i$ ($i \in I$) forming a Σ-VTL $\mathcal{V}$ as in Proposition 11.10 is obtained as follows. For every X, a finite congruence $\gamma_i(X) \in \mathrm{FCon}_\Sigma(X)$ is introduced and $\mathcal{V}_i(X)$ is defined as the set of ΣX-tree languages saturated by $\gamma_i(X)$, and then the congruences $\gamma_i(X)$ define a principal Σ-VFC $\Gamma_i = \{[\gamma_i(X))\}_X$ such that $\mathcal{V}_i = \Gamma_i{}^t$. Moreover, in many cases the congruences $\gamma_i(X)$ form, for any given X, a descending chain, and then $\{\Gamma_i \mid i \in I\}$ and $\{\mathcal{V}_i \mid i \in I\}$ form, correspondingly, ascending chains.

Let us now note some special properties of the Σ-VTLs and Σ-VFAs that correspond to principal Σ-VFCs.

11.13 Proposition *Let $\Gamma = \{[\gamma_X)\}_X$ be a principal Σ-VFC. For any leaf alphabet X, let $\mathcal{F}_X := \mathcal{T}_\Sigma(X)/\gamma_X$ and $\bar{X} := \{\bar{x} \mid x \in X\}$, where $\bar{x} = x/\gamma_X$.*

(a) *For any X, $\Gamma^t(X)$ is the set of all ΣX-tree languages saturated by γ_X.*

(b) *For any X and $T \subseteq T_\Sigma(X)$, $T \in \Gamma^t(X)$ if and only if $\mathrm{SA}(T)$ is an epimorphic image of $\mathcal{F}_X$.*

(c) *For any X, $\mathcal{F}_X$ is freely generated by $\bar{X}$ over Γ^a.*

(d) *Γ^a is the Σ-VFA generated by the algebras $\mathcal{F}_X$.*

Proof For (a) it suffices to observe that for any $T \in \mathrm{Rec}_\Sigma(X)$, $T \in \Gamma^t(X)$ if and only if $\gamma_X \subseteq \theta_T$, while $\gamma_X \subseteq \theta_T$ if and only if γ_X saturates T.

Let us now consider (b). If $T \in \Gamma^t(X)$, then $\gamma_X \subseteq \theta_T$ implies immediately that $\mathrm{SA}(T)$ is an image of $\mathcal{F}_X$. Assume now that there is an epimorphism $\varphi\colon \mathcal{F}_X \to \mathrm{SA}(T)$. For each $x \in X$ we may choose a term $t_x \in T_\Sigma(X)$ such that $(t_x/\gamma_X)\varphi = x/T$, and then the assignment $x \mapsto t_x$ can be extended to an endomorphism $\psi\colon \mathcal{T}_\Sigma(X) \to \mathcal{T}_\Sigma(X)$. Of course, $\psi\gamma_X^\natural\varphi = \varphi_T$. Moreover, if $s\,\gamma_X\,t$ for some $s, t \in T_\Sigma(X)$, then $s\psi\,\gamma_X\,t\psi$ since γ_X is fully invariant by Corollary 10.6. Hence, $s\,\gamma_X\,t$ implies $s/T = s\varphi_T = s\psi\gamma_X^\natural\varphi = t\psi\gamma_X^\natural\varphi = t\varphi_T = t/T$, that is to say, $\gamma_X \subseteq \theta_T$.

Since γ_X is a fully invariant congruence on $\mathcal{T}_\Sigma(X)$, $\mathcal{F}_X$ is a free algebra. As the class of Σ-algebras $\mathcal{A} = (A, \Sigma)$ for which any mapping $\varphi_0 : \bar{X} \to A$ can be extended to a homomorphism $\varphi : \mathcal{F}_X \to \mathcal{A}$ is a variety (cf. [32, p. 165], for example), it suffices by (b) to verify that the algebras $\mathcal{F}_Y$ belong to this class. Let Y be a leaf alphabet and consider any mapping $\varphi_0 : \bar{X} \to T_\Sigma(Y)/\gamma_Y$. Let us define a mapping $\psi_0 : X \to T_\Sigma(Y)$ by setting, for each $x \in X$, $x\psi_0 = s_x$, where s_x is any ΣY-tree such that $\bar{x}\varphi_0 = s_x/\gamma_Y$. If $\psi : \mathcal{T}_\Sigma(X) \to \mathcal{T}_\Sigma(Y)$ is the homomorphic extension of ψ_0, then

$$\varphi : T_\Sigma(X)/\gamma_X \to T_\Sigma(Y)/\gamma_Y,\ t/\gamma_X \mapsto t\psi/\gamma_Y$$

is the required extension of φ_0 to a homomorphism $\varphi : \mathcal{F}_X \to \mathcal{F}_Y$. Indeed,

(1) φ is well-defined because $\gamma_X \subseteq \psi \circ \gamma_Y \circ \psi^{-1}$,

(2) $\varphi|\bar{X} = \varphi_0$ as $\bar{x}\varphi = x\psi/\gamma_Y = x\psi_0/\gamma_Y = \bar{x}\varphi_0$ for every $\bar{x} \in \bar{X}$, and

(3) a direct computation shows that φ is a homomorphism from $\mathcal{F}_X$ to $\mathcal{F}_Y$.

It remains to prove (d). By definition, Γ^a is the Σ-VFA generated by the algebras $\mathcal{T}_\Sigma(X)/\theta$, where X is any leaf alphabet and $\theta \in [\gamma_X)$. However, $\gamma_X \subseteq \theta$ means that $\mathcal{T}_\Sigma(X)/\theta \preceq \mathcal{F}_X$, and hence Γ^a is already generated by the algebras $\mathcal{F}_X$. □

Note that part (b) of Proposition 11.13 also means that every $T \in \Gamma^t(X)$ is recognized by a ΣX-recognizer based on the algebra $\mathcal{F}_X$. Moreover, the proposition implies that the Membership Problem "$T(\mathbf{A}) \in \Gamma^t(X)$?", where $\mathbf{A}$ is any given ΣX-recognizer, is decidable for any Σ-VTL defined by a principal Σ-VFC $\Gamma = \{[\gamma_X)\}_X$ assuming that the congruences γ_X are effectively given.

12 Examples of varieties

We shall now consider several natural families of regular tree languages that form varieties and use them for illustrating the general notions and facts presented in the previous section. Some further examples will be noted in the sections to follow.

12.1 Example The greatest Σ-VTL is the family $\mathrm{Rec}_\Sigma = \{\mathrm{Rec}_\Sigma(X)\}_X$ of all regular Σ-tree languages—the required closure properties were noted in Section 6. That the corresponding Σ-VFA is the class $\mathbf{Fin}_\Sigma$ of all finite Σ-algebras follows from Lemma 9.4 and Proposition 8.14:

a Σ-VFA is generated by the syntactic algebras it contains and every finite syntactic Σ-algebra is the syntactic algebra of some regular Σ-tree language. It is also clear that Rec_Σ^c is the family $\mathrm{FCon}_\Sigma = \{\mathrm{FCon}_\Sigma(X)\}_X$ of all finite Σ-congruences.

Since we excluded the variety of regular Σ-tree languages with empty components, the least Σ-VTL is $\mathrm{Triv}_\Sigma = \{\mathrm{Triv}_\Sigma(X)\}_X$, where $\mathrm{Triv}_\Sigma(X) = \{\emptyset, T_\Sigma(X)\}$ for each X. Clearly, Triv_Σ^a is the class $\mathbf{Triv}_\Sigma$ of all trivial Σ-algebras, and $\mathrm{Triv}_\Sigma^c = \{\{\nabla_{T_\Sigma(X)}\}\}_X$.

The following notion of nilpotency of tree automata was introduced in [24]

12.2 Example Let us call a Σ-algebra $\mathcal{A} = (A, \Sigma)$ *nilpotent* if there is an element $a_0 \in A$ and a number $n \geq 0$ such that for any X, any $\alpha : X \to A$ and any $t \in T_\Sigma(X)$, if $\mathrm{hg}(t) \geq n$ then $t\alpha_\mathcal{A} = a_0$. Let $\mathbf{Nil}_\Sigma$ be the class of all finite nilpotent Σ-algebras. If $\mathcal{A} \in \mathbf{Nil}_\Sigma$, the element a_0 (obviously uniquely determined) and the smallest number n for which the above condition is satisfied, are called, respectively, the *absorbing state* and the *degree of nilpotency* of $\mathcal{A}$. It is easy to see that $\mathbf{Nil}_\Sigma$ is a Σ-VFA. In fact, for each $n \geq 0$, the finite Σ-algebras with degree of nilpotency $\leq n$ form a Σ-VFA $\mathbf{Nil}_{\Sigma,n}$, and $\mathbf{Nil}_\Sigma$ is the union of the ascending chain $\mathbf{Triv}_\Sigma = \mathbf{Nil}_{\Sigma,0} \subseteq \mathbf{Nil}_{\Sigma,1} \subseteq \mathbf{Nil}_{\Sigma,2} \subseteq \ldots$ of such sub-Σ-VFAs. For example, let $\mathcal{A} = (A, \Sigma)$ be in $\mathbf{Nil}_{\Sigma,n}$ with absorbing state a_0 and let $\varphi : \mathcal{A} \to \mathcal{B}$ be an epimorphism onto a Σ-algebra $\mathcal{B} = (B, \Sigma)$. Since φ is surjective, we may define for any $\beta : X \to B$ an $\alpha : X \to A$ such that $x\beta = x\alpha\varphi$ for every $x \in X$. Then $t\beta_\mathcal{B} = t\alpha_\mathcal{A}\varphi$ for every $t \in T_\Sigma(X)$. In particular, if $\mathrm{hg}(t) \geq n$, then $t\beta_\mathcal{B} = a_0\varphi$, and hence $\mathcal{B} \in \mathbf{Nil}_{\Sigma,n}$ with $a_0\varphi$ as the absorbing state.

As noted in [80], $\mathbf{Nil}_\Sigma^t$ is the Σ-VTL $\mathrm{Nil}_\Sigma = \{\mathrm{Nil}_\Sigma(X)\}_X$, where for each X, $\mathrm{Nil}_\Sigma(X)$ is the set of all finite ΣX-tree languages and their complements in $T_\Sigma(X)$ (the co-finite ΣX-tree languages). For showing directly that Nil_Σ is a Σ-VTL, it is useful to note that $\mathrm{hg}(t\varphi) \geq \mathrm{hg}(t)$ for any tree $t \in T_\Sigma(X)$ and any homomorphism $\varphi : \mathcal{T}_\Sigma(X) \to \mathcal{T}_\Sigma(Y)$, as this implies that $T\varphi^{-1}$ is finite for every finite $T \subseteq T_\Sigma(Y)$. A similar observation can be made about the translations $p : T_\Sigma(X) \to T_\Sigma(X)$.

Let us consider the corresponding finite congruences. For each $n \geq 0$ and any leaf alphabet X, let $\nu_{\Sigma,n}(X)$ be the equivalence on $T_\Sigma(X)$ such that

$$T_\Sigma(X)/\nu_{\Sigma,n}(X) = \{\{t\} \mid t \in T_\Sigma(X)^{<n}\} \cup \{T_\Sigma(X)^{\geq n}\}.$$

Obviously, $\nu_{\Sigma,n}(X) \in \mathrm{FCon}_\Sigma(X)$ and $\nabla_{T_\Sigma(X)} = \nu_{\Sigma,0}(X) \supseteq \nu_{\Sigma,1}(X) \supseteq \nu_{\Sigma,2}(X) \supseteq \ldots$, and the inclusions are proper whenever $T_\Sigma(X)$ is infinite. It is also clear that $\Gamma\,\mathrm{Nil}_{\Sigma,n} = \{[\nu_{\Sigma,n}(X))\}_X$ is a principal Σ-VFC. Let $\Gamma\,\mathrm{Nil}_\Sigma$ be the union of the ascending chain $\Gamma\,\mathrm{Nil}_{\Sigma,0} \subseteq \Gamma\,\mathrm{Nil}_{\Sigma,1} \subseteq \Gamma\,\mathrm{Nil}_{\Sigma,2} \subseteq \ldots$ of Σ-VFCs. Of course, $\Gamma\,\mathrm{Nil}_\Sigma \in \mathbf{VFC}(\Sigma)$.

Clearly, a ΣX-tree language T is in $\mathrm{Nil}_\Sigma(X)$ if and only if T is saturated by some $\nu_{\Sigma,n}(X)$. In fact, for any $T \in \mathrm{Nil}_\Sigma(X)$ there is a least $n \geq 0$ such that T is saturated by $\nu_{\Sigma,k}(X)$ if and only if $k \geq n$. For every $n \geq 0$, let $\mathrm{Nil}_{\Sigma,n}(X)$ be the set of all ΣX-tree languages saturated by $\nu_{\Sigma,n}(X)$. By Proposition 11.13, $\mathrm{Nil}_{\Sigma,n} = \{\mathrm{Nil}_{\Sigma,n}(X)\}_X$ is the Σ-VTL corresponding to $\Gamma\,\mathrm{Nil}_{\Sigma,n}$. Moreover, Nil_Σ is the union of the chain $\mathrm{Nil}_{\Sigma,0} \subseteq \mathrm{Nil}_{\Sigma,1} \subseteq \mathrm{Nil}_{\Sigma,2} \subseteq \ldots$ and $\mathrm{Nil}_\Sigma^c = \Gamma\,\mathrm{Nil}_\Sigma$.

To complete the picture, we note that the chain $\mathbf{Nil}_{\Sigma,0} \subseteq \mathbf{Nil}_{\Sigma,1} \subseteq \ldots$ matches the chain $\Gamma\,\mathrm{Nil}_{\Sigma,0} \subseteq \Gamma\,\mathrm{Nil}_{\Sigma,1} \subseteq \ldots$. The quotient algebra $\mathcal{N}_{\Sigma,n}(X) := \mathcal{T}_\Sigma(X)/\nu_{\Sigma,n}(X)$ is in $\mathbf{Nil}_{\Sigma,n}$ with the class $T_\Sigma(X)^{\geq n}$ as the absorbing state. It corresponds to the algebra $\mathcal{F}_X$ of Proposition 11.13. Hence, a ΣX-tree language T is in $\mathrm{Nil}_{\Sigma,n}(X)$ if and only if $\mathrm{SA}(T)$ is an image of $\mathcal{N}_{\Sigma,n}(X)$, and $\mathcal{N}_{\Sigma,n}(X)$ is freely generated by $\{x/\nu_{\Sigma,n}(X) \mid x \in X\}$ over $\mathbf{Nil}_{\Sigma,n}$.

For any given ΣX-recognizer $\mathbf{A} = (\mathcal{A}, \alpha, F)$, we may decide whether $T(\mathbf{A}) \in \mathrm{Nil}_\Sigma(X)$ by using the correspondence between Nil_Σ and $\mathbf{Nil}_\Sigma$. We may assume that $\mathbf{A}$ is minimal. Then $\mathcal{A} \cong \mathrm{SA}(T(\mathbf{A}))$, and hence $T(\mathbf{A}) \in \mathrm{Nil}_\Sigma(X)$ if and only if $\mathcal{A} \in \mathbf{Nil}_\Sigma$. The latter condition can be tested directly. It is easy to see that if an n-element algebra is nilpotent, its degree of nilpotency is $\leq n-1$. In [24] it is shown that an algebra $\mathcal{A} = (A, \Sigma)$ is nilpotent if and only if the unary algebra $(A, \mathrm{Tr}(\mathcal{A}))$ is nilpotent, where the operations are the translations of $\mathcal{A}$.

Recall that a string language $L \subseteq X^*$ is *definite* [42, 56] if there is a $k \geq 0$ such that for any word $w \in X^*$ of length $\geq k$, $w \in L$ if and only if the suffix of w of length k is in L. Heuter [33, 35] was the first to study definite tree languages, and we adopt her definition with a minor modification. That the definite Σ-tree languages form a variety of tree languages is noted in [80]. Further relevant results can be found in [21].

12.3 Example As our tree recognizers read a tree in the direction from the leaves to the root, we replace suffixes with root segments. The *k-root* $\mathrm{root}_k(t)$ ($k \geq 0$) of a ΣX-tree t is defined as follows:

(1) $\mathrm{root}_0(t) = \epsilon$ (the *empty root segment*) for every $t \in T_\Sigma(X)$.

(2) $\mathrm{root}_1(t) = \mathrm{root}(t)$ for every $t \in T_\Sigma(X)$.

(3) Let $k \geq 2$. If $\mathrm{hg}(t) \leq k$, then $\mathrm{root}_k(t) = t$. If $\mathrm{hg}(t) > k$ and $t = f(t_1, \ldots, t_m)$, then $\mathrm{root}_k(t) = f(\mathrm{root}_{k-1}(t_1), \ldots, \mathrm{root}_{k-1}(t_m))$.

For example, if $t = f(g(x), f(g(c), x))$, then $\mathrm{root}_0(t) = \epsilon$, $\mathrm{root}_1(t) = f$, $\mathrm{root}_2(t) = f(g, f)$, $\mathrm{root}_3(t) = f(g(x), f(g, x))$ and $\mathrm{root}_k(t) = t$ for every $k \geq 4$.

For any $k \geq 0$ and any X, let $\delta_{\Sigma,k}(X)$ be the equivalence on $T_\Sigma(X)$ defined by

$$s\, \delta_{\Sigma,k}(X)\, t \quad \text{if and only if} \quad \mathrm{root}_k(s) = \mathrm{root}_k(t) \qquad (s, t \in T_\Sigma(X)).$$

It is obvious that $\delta_{\Sigma,k}(X)$ is in $\mathrm{FCon}_\Sigma(X)$ with a congruence class corresponding to every possible root segment $\mathrm{root}_k(t)$. For any homomorphism $\varphi : \mathcal{T}_\Sigma(X) \to \mathcal{T}_\Sigma(Y)$, any $k \geq 0$ and any $s, t \in T_\Sigma(X)$, if $\mathrm{root}_k(s) = \mathrm{root}_k(t)$ then clearly $\mathrm{root}_k(s\varphi) = \mathrm{root}_k(t\varphi)$ also. This implies by Proposition 10.5 that $\Gamma\,\mathrm{Def}_{\Sigma,k} = \{[\delta_{\Sigma,k}(X))\}_X$ is a principal Σ-VFC for each $k = 0, 1, \ldots$, and hence $\Gamma\,\mathrm{Def}_\Sigma = \bigcup_{k\geq 0} \Gamma\,\mathrm{Def}_{\Sigma,k}$ is a Σ-VFC.

For any $k \geq 0$, a ΣX-tree language T is said to be *k-definite* if it is saturated by $\delta_{\Sigma,k}(X)$, that is to say, for any $s, t \in T_\Sigma(X)$, if $\mathrm{root}_k(s) = \mathrm{root}_k(t)$, then $s \in T$ if and only if $t \in T$. Let $\mathrm{Def}_{\Sigma,k}(X)$ denote the set of k-definite ΣX-tree languages. It follows immediately from the definition that $\mathrm{Def}_{\Sigma,k} = \{\mathrm{Def}_{\Sigma,k}(X)\}_X$ is the Σ-VTL corresponding to the Σ-VFC $\Gamma\,\mathrm{Def}_{\Sigma,k}$. A ΣX-tree language is *definite* if it is k-definite for some $k \geq 0$. Let $\mathrm{Def}_\Sigma = \bigcup_{k\geq 0} \mathrm{Def}_{\Sigma,k}$ be the Σ-VTL of all definite Σ-tree languages. Of course, $\mathrm{Def}_\Sigma = \Gamma\,\mathrm{Def}_\Sigma^t$.

For each $k \geq 0$, an algebra $\mathcal{A} = (A, \Sigma)$ is *k-definite* [21] if $\mathrm{root}_k(s) = \mathrm{root}_k(t)$ implies $s\alpha_\mathcal{A} = t\alpha_\mathcal{A}$ for every leaf alphabet X, every $\alpha : X \to A$ and all $s, t \in T_\Sigma(X)$. Let $\mathbf{Def}_{\Sigma,k}$ be the class of all finite k-definite Σ-algebras, and let $\mathbf{Def}_\Sigma = \bigcup_{k\geq 0} \mathbf{Def}_{\Sigma,k}$ be the class of all finite *definite Σ-algebras*. It is easy to verify directly that every $\mathbf{Def}_{\Sigma,k}$ is a Σ-VFA, but one can also show that $\mathbf{Def}_{\Sigma,k} = \Gamma\,\mathrm{Def}_{\Sigma,k}^a$ by considering the quotient algebras $\mathcal{D}_{\Sigma,k}(X) := \mathcal{T}_\Sigma(X)/\delta_{\Sigma,k}(X)$. Indeed, it is clear that $\mathcal{D}_{\Sigma,k}(X)$ is k-definite and that any k-definite Σ-algebra with a generating set of size $\leq |X|$ is an epimorphic image of $\mathcal{D}_{\Sigma,k}(X)$. In fact,

$\mathcal{D}_{\Sigma,k}(X)$ is generated freely over $\mathbf{Def}_{\Sigma,k}$ by $\{x/\delta_{\Sigma,k}(X) \mid x \in X\}$. It follows from the results of [33] that $\mathbf{Def}_{\Sigma,k}$ is the Σ-VFA corresponding to $\mathrm{Def}_{\Sigma,k}$.

For any given $k \geq 0$ and X, the question "$T(\mathbf{A}) \in \mathrm{Def}_{\Sigma,k}(X)$?" is again decidable for any ΣX-recognizer $\mathbf{A} = (\mathcal{A}, \alpha, F)$; assuming that $\mathbf{A}$ is minimal, one has just to check whether the algebra $\mathcal{A}$ is k-definite. On the other hand, it is not immediately clear that the question "$T(\mathbf{A}) \in \mathrm{Def}_{\Sigma}(X)$?" is decidable. However, Heuter [33, 34] has shown that if $\mathbf{A}$ has n states and $T(\mathbf{A})$ is definite, then $T(\mathbf{A})$ is $(n-1)$-definite. Furthermore, Ésik [21] has shown that $\mathcal{A}$ is k-definite if and only if the unary algebra $(A, \mathrm{Tr}(\mathcal{A}))$ is k-definite. Hence, the definiteness question of regular tree languages can be reduced to the well-known question of whether a given unary algebra is definite. Let us also note that the paper [21] contains many further results about definite tree automata. In particular, it is shown that the class of definite tree automata is closed under cascade compositions, and various equational characterizations are considered.

A language is *reverse definite* [7] if there is a $k \geq 0$ such that for any word $w \in X^*$ of length $\geq k$, $w \in L$ if and only if the prefix of w of length k is in L. In the definition of a *generalized definite* language [31] a prefix condition and a suffix condition are combined. These notions can be extended to trees as follows.

12.4 Example The counterparts of prefixes of words are subtrees. However, two differences should be noted. Firstly, a tree may have several subtrees of a given height, and secondly, to take into account the whole frontier part of a tree up to a certain depth, we have to also consider the subtrees of lesser height.

For any $k \geq 0$, let $\mathrm{sub}_k(t) = \{s \in \mathrm{sub}(t) \mid \mathrm{hg}(t) < k\}$. For example, if $f(g(c), f(x, g(c)))$ then $\mathrm{sub}_0(t) = \emptyset$, $\mathrm{sub}_1(t) = \{c, x\}$, $\mathrm{sub}_2(t) = \{c, x, g(c)\}$, $\mathrm{sub}_3(t) = \{c, x, g(c), f(x, g(c))\}$ and $\mathrm{sub}_k(t) = \mathrm{sub}(t)$ for all $k \geq 4$.

For any $k \geq 0$ and any X, let $\rho_{\Sigma,k}(X)$ be the equivalence on $T_\Sigma(X)$ defined by

$$s\,\rho_{\Sigma,k}(X)\,t \quad \text{if and only if} \quad \mathrm{sub}_k(s) = \mathrm{sub}_k(t) \qquad (s, t \in T_\Sigma(X)).$$

It is obvious that $\rho_{\Sigma,k}(X) \in \mathrm{FCon}_\Sigma(X)$ with a congruence class for every possible collection $\mathrm{sub}_k(t)$ of subtrees of height $< k$. For any homomorphism $\varphi : \mathcal{T}_\Sigma(X) \to \mathcal{T}_\Sigma(Y)$, any $k \geq 0$ and any $s, t \in T_\Sigma(X)$, if $\mathrm{sub}_k(s) = \mathrm{sub}_k(t)$ then clearly also $\mathrm{sub}_k(s\varphi) = \mathrm{sub}_k(t\varphi)$. This implies by Proposition 10.5 that $\Gamma\,\mathrm{RDef}_{\Sigma,k} = \{[\rho_{\Sigma,k}(X))\}_X$ is a principal Σ-VFC for each $k = 0, 1, \ldots$, and hence $\Gamma\,\mathrm{RDef}_\Sigma = \bigcup_{k\geq 0} \Gamma\,\mathrm{RDef}_{\Sigma,k}$ is a Σ-VFC.

The ΣX-tree languages saturated by $\rho_{\Sigma,k}(X)$ are said to be *reverse k-definite* ($k \geq 0$). Let $\mathrm{RDef}_{\Sigma,k}(X)$ denote the set of all reverse k-definite ΣX-tree languages. Then $\mathrm{RDef}_{\Sigma,k} = \{\mathrm{RDef}_{\Sigma,k}(X)\}_X$ is the Σ-VTL corresponding to the Σ-VFC $\Gamma\,\mathrm{RDef}_{\Sigma,k}$. A ΣX-tree language is *reverse definite* if it is reverse k-definite for some $k \geq 0$. Let $\mathrm{RDef}_\Sigma = \bigcup_{k\geq 0} \mathrm{RDef}_{\Sigma,k}$ be the Σ-VTL of all reverse definite Σ-tree languages. Of course, $\mathrm{RDef}_\Sigma = \Gamma\,\mathrm{RDef}_\Sigma^t$.

For any $h, k \geq 0$ and X, let $\gamma_{\Sigma,h,k}(X) = \rho_{\Sigma,h}(X) \cap \delta_{\Sigma,k}(X)$. It is again easy to see that $\Gamma\,\mathrm{GDef}_{\Sigma,h,k} = \{[\gamma_{\Sigma,h,k}(X))\}_X$ is a principal Σ-VFC for any pair $h, k \geq 0$, and that the family of these principal Σ-VFCs is directed. Therefore their union $\Gamma\,\mathrm{GDef}_\Sigma = \bigcup_{h\geq 0, k\geq 0} \Gamma\,\mathrm{GDef}_{\Sigma,h,k}$ is also a Σ-VFC.

For any $h, k \geq 0$ and X, a ΣX-tree language is called *generalized (h,k)-definite* if it is saturated by $\gamma_{\Sigma,h,k}(X)$, and it is *generalized definite* if it is generalized (h,k)-definite for some $h, k \geq 0$. Hence, $T \subseteq T_\Sigma(X)$ is generalized (h,k)-definite, if $s \in T$ if and only if $t \in T$, for

any $s, t \in T_\Sigma(X)$ such that $\mathrm{sub}_h(s) = \mathrm{sub}_h(t)$ and $\mathrm{root}_k(s) = \mathrm{root}_k(t)$. Let $\mathrm{GDef}_{\Sigma,h,k} = \{\mathrm{GDef}_{\Sigma,h,k}(X)\}$ be the Σ-VTL of all generalized (h,k)-definite Σ-tree languages, and let $\mathrm{GDef}_\Sigma = \bigcup_{h\geq 0,k\geq 0} \mathrm{GDef}_{\Sigma,h,k}$ be the Σ-VTL of all generalized definite Σ-tree languages. Obviously, $\mathrm{GDef}_{\Sigma,h,k} = \Gamma\,\mathrm{GDef}^t_{\Sigma,h,k}$, and hence also $\mathrm{GDef}_\Sigma = \Gamma\,\mathrm{GDef}^t_\Sigma$. Moreover, it is clear that

(1) $\mathrm{Triv}_\Sigma \subseteq \mathrm{GDef}_{\Sigma,h,k} \subseteq \mathrm{Rec}_\Sigma$ for all $h, k \geq 0$,

(2) $\mathrm{GDef}_{\Sigma,0,k} = \mathrm{Def}_{\Sigma,k}$,

(3) $\mathrm{GDef}_{\Sigma,h,0} = \mathrm{RDef}_{\Sigma,h}$, and

(4) $\mathrm{GDef}_{\Sigma,i,j} \subseteq \mathrm{GDef}_{\Sigma,h,k}$ whenever $i \leq h$ and $j \leq k$.

Moreover, the inclusions in (4) are proper whenever $i < j$ or $j < k$ (for any non-trivial Σ).

The problems "$T(\mathbf{A}) \in \mathrm{RDef}_\Sigma$?" and "$T(\mathbf{A}) \in \mathrm{GDef}_\Sigma$?" were shown to be decidable by Heuter [34, 35]. Wilke [92] characterizes the syntactic tree algebras (cf. Section 14) of reverse definite binary tree languages.

A language $L \subseteq X^*$ is *local (in the strict sense)* (cf. [17], for example), if there are sets $A, B \subseteq X$ and $C \subseteq X^2$ of initial letters, final letters and allowed pairs of consecutive letters, respectively, such that $L = AX^* \cap BX^* - X^*(X^2 - C)X^*$. A corresponding notion also appears in the theory of tree languages. Firstly, any regular tree language is the image of a local tree language under an alphabetic tree homomorphism. Secondly, the production trees of any context-free grammar form a local tree language, and hence every context-free language is the yield of a local tree language (cf. [28, 29]).

12.5 Example The root symbol $\mathrm{root}(t)$ of a tree t corresponds to the last letter of a word. The counterparts of pairs of consecutive letters, as well as of initial letters, are given by the following notion. The set $\mathrm{fork}(t)$ of *forks* of a ΣX-tree t is defined as follows:

(1) $\mathrm{fork}(x) = \mathrm{fork}(c) = \emptyset$ for $x \in X$ and $c \in \Sigma_0$;

(2) $\mathrm{fork}(t) = \mathrm{fork}(t_1) \cup \cdots \cup \mathrm{fork}(t_m) \cup \{f(\mathrm{root}(t_1), \ldots, \mathrm{root}(t_m))\}$ for $t = f(t_1, \ldots, t_m)$.

For example, if $t = g(f(f(c,x), g(c)))$, then $\mathrm{fork}(t) = \{g(f), f(f,g), f(c,x), g(c)\}$. The (finite) set of all possible forks appearing in any ΣX-tree is denoted by $\mathrm{fork}(\Sigma, X)$. For any $F \subseteq \mathrm{fork}(\Sigma, X)$ and $R \subseteq \Sigma \cup X$, let

$$SL(F,R) = \{t \in T_\Sigma(X) \mid \mathrm{fork}(t) \subseteq F,\ \mathrm{root}(t) \in R\}.$$

Tree languages of this form are said to be *local in the strict sense*. Let $\mathrm{SLoc}_\Sigma(X)$ be the set of all ΣX-tree languages local in the strict sense. Clearly, $\mathrm{SLoc}_\Sigma(X) \subseteq \mathrm{Rec}_\Sigma(X)$ for any Σ and X, but the family $\mathrm{SLoc} = \{\mathrm{SLoc}_\Sigma(X)\}_X$ is not a Σ-VTL because it is not closed under unions or complements. On the other hand, it is easy to see that it is closed under inverse translations and inverse homomorphisms. For example, let $\varphi : \mathcal{T}_\Sigma(X) \to \mathcal{T}_\Sigma(Y)$ be a homomorphism and let $T = SL(F,R) \in \mathrm{SLoc}_\Sigma(Y)$. Then $T\varphi^{-1} = SL(F',R')$, where $R' = (R \cap \Sigma) \cup \{x \in X \mid x\varphi \in T\}$, and F' consists of all forks $f(d_1, \ldots, d_m) \in \mathrm{fork}(\Sigma, X)$ such that $f(e_1, \ldots, e_m) \in F$ when for each $i = 1, \ldots, m$, $e_i = d_i$ if $d_i \in \Sigma$, and $e_i = \mathrm{root}(d_i\varphi)$ if $d_i \in X$. This readily implies that the Σ-VTL generated by SLoc_Σ, the family $\mathrm{Loc}_\Sigma = \{\mathrm{Loc}_\Sigma(X)\}_X$

of *local* Σ-tree languages, is simply the Boolean closure of SLoc_Σ, that is to say, $\mathrm{Loc}_\Sigma(X)$ is the Boolean closure of $\mathrm{SLoc}_\Sigma(X)$ for each X.

The family Loc_Σ can also be obtained directly as the Σ-VTL corresponding to a Σ-VFC as follows. For each X let us define the relation $\lambda_\Sigma(X)$ on $T_\Sigma(X)$ by the condition

$$s\,\lambda_\Sigma(X)\,t \quad \text{if and only if} \quad \mathrm{fork}(s) = \mathrm{fork}(t) \text{ and } \mathrm{root}(s) = \mathrm{root}(t) \qquad (s,t \in T_\Sigma(X)).$$

It is easy to see that $\lambda_\Sigma(X) \in \mathrm{FCon}_\Sigma(X)$ for each X, and that $\Gamma\,\mathrm{Loc}_\Sigma = \{[\lambda_\Sigma(X))\}_X$ is the Σ-VFC such that $\Gamma\,\mathrm{Loc}_\Sigma^t = \mathrm{Loc}_\Sigma$.

This example can be generalized by defining for each $k \geq 0$ a Σ-VTL of *k-testable tree languages*. For this we have to define in a natural way the set $\mathrm{fork}_k(t)$ of "k-forks" of a tree t; the forks defined above are then 1-forks (cf. [43]).

13 Some generalizations

In this section we consider a few generalizations of the variety theory discussed above. Some alternative approaches are then reviewed in the section to follow. Let us begin with the theory of varieties of recognizable subsets of free algebras presented in [78]. The theory generalizes also Eilenberg's theories of $*$- and $+$-varieties. The point of view adopted is that the syntactic algebra of a language L over an alphabet X is a monoid, the syntactic monoid of L, because L is regarded as a subset of the free monoid X^*, while the syntactic algebra of a ΣX-tree language T is a Σ-algebra because T is viewed as a subset of the term algebra $\mathcal{T}_\Sigma(X)$. The theory of syntactic congruences and syntactic algebras of subsets of algebras presented in Section 8 is directly applicable to the generalization to be considered here.

Let $\mathbf{V}$ be any given variety of Σ-algebras. We shall consider recognizable subsets of the free algebras over $\mathbf{V}$ generated by finite non-empty sets. Without loss of generality, we may assume that these generating sets are the finite non-empty leaf alphabets $X, Y, Z, \ldots$ used above. The free $\mathbf{V}$-algebra generated by a set U is denoted by $\mathcal{F}_\mathbf{V}(U) = (F_\mathbf{V}(U), \Sigma)$.

A *family of recognizable $\mathbf{V}$-sets* is a mapping $\mathcal{R}$ that assigns to every X a set $\mathcal{R}(X) \subseteq \mathrm{Rec}(\mathcal{F}_\mathbf{V}(X))$ of recognizable subsets of $\mathcal{F}_\mathbf{V}(X)$. We write $\mathcal{R} = \{\mathcal{R}(X)\}_X$ and define the $\subseteq$-relation as well as the unions and intersections of families of recognizable $\mathbf{V}$-sets by the natural componentwise conditions.

13.1 Definition A family of recognizable $\mathbf{V}$-sets $\mathcal{R} = \{\mathcal{R}(X)\}_X$ is a *variety of recognizable $\mathbf{V}$-sets*, a $\mathbf{V}$-VRS for short, if for all X and Y,

(1) $\mathcal{R}(X)$ is a Boolean subalgebra of $\mathrm{Rec}(\mathcal{F}_\mathbf{V}(X))$,

(2) $T \in \mathcal{R}(X)$ implies $p^{-1}(T) \in \mathcal{R}(X)$ for any $p \in \mathrm{Tr}(\mathcal{F}_\mathbf{V}(X))$, and

(3) $T \in \mathcal{R}(Y)$ implies $T\varphi^{-1} \in \mathcal{R}(X)$ for any homomorphism $\varphi : \mathcal{F}_\mathbf{V}(X) \to \mathcal{F}_\mathbf{V}(Y)$.

Let $\mathbf{VRS}(\mathbf{V})$ denote the class of all $\mathbf{V}$-VRSs.

Clearly, $(\mathbf{VRS}(\mathbf{V}), \subseteq)$ is a complete lattice. If $\mathbf{V}$ is the class of all Σ-algebras, the term algebra $\mathcal{T}_\Sigma(X)$ can be used as $\mathcal{F}_\mathbf{V}(X)$ and then the $\mathbf{V}$-varieties of recognizable sets are exactly the Σ-VTLs. Similarly, for the variety $\mathbf{Mon}$ of all monoids and the variety $\mathbf{Sg}$ of all semigroups, $\mathbf{VRS}(\mathbf{Mon})$ and $\mathbf{VRS}(\mathbf{Sg})$ are the classes of all $*$- and $+$-varieties, respectively.

Let $\mathcal{R} = \{\mathcal{R}(X)\}_X$ be a family of recognizable $\mathbf{V}$-sets. The *syntactic algebra* $\mathrm{SA}(T)$ of a set $T \subseteq F_{\mathbf{V}}(X)$ is defined to be its syntactic algebra $\mathcal{F}_{\mathbf{V}}(X)/\theta_T$ as a subset of $\mathcal{F}_{\mathbf{V}}(X)$. Of course, $\mathrm{SA}(T)$ is finite if and only if T is recognizable in $\mathcal{F}_{\mathbf{V}}(X)$. Now, let

$$\mathcal{R}^a = \mathrm{V_f}(\{\mathrm{SA}(T) \mid T \in \mathcal{R}(X) \text{ for some } X\})$$

be the Σ-VFA generated by the syntactic algebras of the subsets belonging to $\mathcal{R}$. On the other hand, for any class $\mathbf{K}$ of finite Σ-algebras, let $\mathbf{K}^r = \{\mathbf{K}^r(X)\}_X$ be the family of recognizable $\mathbf{V}$-sets such that $\mathbf{K}^r(X) = \{T \subseteq F_{\mathbf{V}}(X) \mid \mathrm{SA}(T) \in \mathbf{K}\}$ for each X. Furthermore, let $\mathbf{VFA}(\mathbf{V})$ denote the class of all Σ-VFAs contained in $\mathbf{V}$. We may now formulate the following Variety Theorem [78].

13.2 Theorem *For any variety* $\mathbf{V}$ *of* Σ*-algebras, the mappings*

$$\mathbf{VRS}(\mathbf{V}) \to \mathbf{VFA}(\mathbf{V}),\ \mathcal{R} \mapsto \mathcal{R}^a, \qquad \text{and} \qquad \mathbf{VFA}(\mathbf{V}) \to \mathbf{VRS}(\mathbf{V}),\ \mathbf{K} \mapsto \mathbf{K}^r,$$

form a pair of mutually inverse isomorphisms between the complete lattices $(\mathbf{VRS}(\mathbf{V}), \subseteq)$ *and* $(\mathbf{VFA}(\mathbf{V}), \subseteq)$.

The theorem can be proved exactly as the parts of Theorem 11.9 that concern the connection between $\mathbf{VTL}(\Sigma)$ and $\mathbf{VFA}(\Sigma)$. As noted above, the theory of syntactic algebras presented in Section 8 is directly applicable here, too, and replacing the term algebras $\mathcal{T}_\Sigma(X)$ with the free algebras $\mathcal{F}_{\mathbf{V}}(X)$ does not require any essential changes either since just the freeness property is needed.

We could complete Theorem 13.2 with varieties of finite congruences on the free algebras $\mathcal{F}_{\mathbf{V}}(X)$. As a matter of fact, varieties of congruences appear both in the related theory to be discussed next and the generalization to many-sorted mentioned thereafter.

In [1] Almeida studies various ways to describe varieties of finite algebras. Since varieties of recognizable sets and varieties of finite congruences (in our terminology) are among these, his work also includes a variety theory similar to that considered above. A detailed presentation of this theory, as well as many related matters, can be found in [2]. Here we review just the main notions and results adapting the terminology and notation to ours.

The starting point is any given Σ-VFA $\mathbf{K}$. Syntactic congruences and syntactic algebras of subsets of algebras are defined as above. A subset $L \subseteq A$ of an algebra $\mathcal{A} = (A, \Sigma)$ is said to be $\mathbf{K}$-*recognizable* if it is recognized by a member of $\mathbf{K}$, i.e., if $L = F\varphi^{-1}$ for some $\mathcal{B} = (B, \Sigma)$ in $\mathbf{K}$, a subset $F \subseteq B$ and a homomorphism $\varphi : \mathcal{A} \to \mathcal{B}$. Moreover, $L \subseteq A$ is *locally* $\mathbf{K}$-*recognizable* if $L \cap \langle H \rangle$ is $\mathbf{K}$-recognizable for every finite $H \subseteq A$.

Let $\mathbf{V} = \mathrm{V}(\mathbf{K})$ be the variety of Σ-algebras generated by $\mathbf{K}$, and for each $n \geq 1$, let $\mathcal{F}_n = (F_n, \Sigma)$ be the free $\mathbf{V}$-algebra generated by $X_n := \{x_1, \ldots, x_n\}$. Similarly let $\mathcal{F}_\omega = (F_\omega, \Sigma)$ be the free $\mathbf{V}$-algebra generated by $X_\omega := \{x_1, x_2, x_3, \ldots\}$. We may assume that $F_1 \subseteq F_2 \subseteq F_3 \subseteq \ldots \subseteq F_\omega = \bigcup_{n \geq 1} F_n$. For each $n \geq 1$, let $\mathrm{Rec}_n(\mathbf{K})$ be the set of all $\mathbf{K}$-recognizable subsets of $\mathcal{F}_n$, and let $\mathrm{Con}_n(\mathbf{K}) = \{\theta \in \mathrm{Con}(\mathcal{F}_n) \mid \mathcal{F}_n/\theta \in \mathbf{K}\}$. Moreover let $\mathrm{LRec}(\mathbf{K})$ be the set of all locally $\mathbf{K}$-recognizable subsets of $\mathcal{F}_\omega$, and let $\mathrm{LCon}(\mathbf{K})$ consist of all congruences $\theta \in \mathrm{Con}(\mathcal{F}_\omega)$ such that $\mathcal{B}/\theta_B \in \mathbf{K}$ for every finitely generated subalgebra $\mathcal{B} = (B, \Sigma)$ of $\mathcal{F}_\omega$.

It is clear that the family $(\mathrm{Rec}_n(\mathbf{K}) \mid n \geq 1)$ is essentially the $\mathbf{V}$-VRS corresponding to $\mathbf{K}$ since $\mathrm{Rec}_n(\mathbf{K}) = \mathbf{K}^r(X_n)$ for every $n \geq 1$. The locally $\mathbf{K}$-recognizable subsets of $\mathcal{F}_\omega$ represent in some sense families of sets of $\mathbf{K}$-recognizable subsets of the algebras $\mathcal{F}_n$ $(n \geq 1)$.

In particular, a subset L of F_ω is in $\mathrm{LRec}(\mathbf{K})$ if and only if $L \cap F_n \in \mathrm{Rec}_n(\mathbf{K})$ for every $n \geq 1$. Similarly $\mathrm{LCon}(\mathbf{K})$ is closely related to the family of all $\mathrm{Con}_n(\mathbf{K})$ $(n \geq 1)$.

A *variety of* $\mathbf{K}$*-languages* is a sequence $V = (V_n)_{n\geq 1}$ such that for all $n, m \geq 1$,

(1) V_n is a Boolean subalgebra of $\mathrm{Rec}_n(\mathbf{K})$,

(2) $p^{-1}(T) \in V_n$ whenever $T \in V_n$ and $p \in \mathrm{Tr}(\mathcal{F}_n)$, and

(3) $T\varphi^{-1} \in V_n$ for any $T \in V_m$ and any homomorphism $\varphi\colon \mathcal{F}_n \to \mathcal{F}_m$.

A *global variety of* $\mathbf{K}$*-languages* is a Boolean subalgebra $\mathcal{W}$ of $\mathrm{LRec}(\mathbf{K})$ closed under inverse translations and inverse endomorphisms of $\mathcal{F}_\omega$ such that if $L \in \mathrm{LRec}(\mathbf{K})$ and for every finite $H \subseteq F_\omega$ there is an $L_H \in \mathcal{W}$ for which $L \cap \langle H\rangle = L_H \cap \langle H\rangle$, then $L \in \mathcal{W}$. The last condition may be interpreted to mean that $\mathcal{W}$ contains any subset of F_ω that belongs to it locally.

A *variety of filters of* $\mathbf{K}$*-congruences* is a sequence $(\Phi_n)_{n\geq 1}$ such that for all $m, n \geq 1$,

(1) Φ_n is a filter of $\mathrm{Con}_n(\mathbf{K})$, and

(2) $\theta \in \Phi_m$ implies $\varphi \circ \theta \circ \varphi^{-1} \in \Phi_n$ for any homomorphism $\varphi\colon \mathcal{F}_n \to \mathcal{F}_m$.

Similarly, a *global variety of filters of* $\mathbf{K}$*-congruences* is a filter Φ of $\mathrm{LCon}(\mathbf{K})$ closed under inverse endomorphisms of $\mathcal{F}_\omega$ such that if $\theta \in \mathrm{LCon}(\mathbf{K})$ and for every finitely generated subalgebra $\mathcal{A} = (A, \Sigma)$ of $\mathcal{F}_\omega$ there is a $\varrho \in \Phi$ such that $\theta_A = \varrho_A$, then $\theta \in \Phi$.

Now Almeida's Variety Theorem establishes isomorphisms between the complete lattices formed by (1) all varieties of $\mathbf{K}$-languages, (2) all global varieties of $\mathbf{K}$-languages, (3) all varieties of filters of $\mathbf{K}$-congruences, (4) all global varieties of filters of $\mathbf{K}$-congruences, and (5) all Σ-VFAs contained in $\mathbf{V}$.

Salehi and Steinby [73] extend the theory of varieties of recognizable sets to *many-sorted algebras* (cf. [51] for a survey and many further references). Many-sorted algebras are much used in computer science, especially in the theory of data types. Many-sorted tree languages were introduced by Maibaum [48], and Courcelle [13, 14] has studied recognizable subsets of general many-sorted algebras. Moreover, it can be noted that Wilke's [92] tree algebras—to be considered in the next section—involve three sorts. For a given set S of sorts, there are actually two natural kinds of recognizable subset of an S-sorted algebra $\mathcal{A} = (A, \Omega)$, where $A = \langle A_s\rangle_{s\in S}$ is the S-sorted set of elements and Ω is an S-sorted ranked alphabet. Firstly, we may consider sorted subsets $L = \langle L_s\rangle_{s\in S}$ that include for each sort $s \in S$ a subset $L_s \subseteq A_s$ of that sort. On the other hand, there are the 'pure' subsets in which all elements are of the same sort. In [48] and [13, 14] sets of the second kind are considered, but in [73] the variety theory is presented primarily for sorted subsets. However when S is finite, the recognizable subsets the two types are definable in terms of each other, and in [73] a variety theorem is also derived for varieties of pure recognizable sets.

In the theories mentioned so far, all varieties considered are over a given fixed ranked alphabet. In [84] the variety theory of tree languages is generalized so that we may speak generally about, say, the variety of definite tree languages or the variety of locally testable tree languages without fixing the ranked alphabet. Corresponding to such *generalized varieties of tree languages* (GVTLs) we have *generalized varieties of finite algebras* (GVFAs) and *generalized varieties of finite congruences* (GVFCs). To define these one has to generalize the basic notions of universal algebra in such a way that algebras of different types can be considered together. For example, a generalized homomorphism from a Σ-algebra $\mathcal{A} = (A, \Sigma)$

to an Ω-algebra $\mathcal{B} = (B, \Omega)$ consists of two mappings $\iota\colon \Sigma \to \Omega$ and $\varphi\colon A \to B$ such that $\iota(\Sigma_m) \subseteq \Omega_m$ for every $m \geq 0$, and $f^{\mathcal{A}}(a_1, \ldots, a_m)\varphi = \iota(f)^{\mathcal{B}}(a_1\varphi, \ldots, a_m\varphi)$ for all $m \geq 0$, $f \in \Sigma_m$ and $a_1, \ldots, a_m \in A$. Also the definitions of syntactic congruences and syntactic algebras have to be suitably modified. All the examples of Σ-VTLs mentioned in Section 7 yield also examples of GVTLs. Thus, there are the GVTLs of definite, reverse definite, generalized definite, locally testable and aperiodic tree languages, for example. A recent example is provided by the *piecewise testable* tree languages studied by Piirainen [62]. Here piecewise testability is defined using the well-known *homeomorphic embedding* order of trees (cf. [3], for example). In fact for each $k \geq 0$, the *k-piecewise testable* tree languages form a GVTL, and the GVTL of all piecewise testable tree languages is the union of the ascending chain of these sub-GVTLs. Let us also note that if $\mathcal{V} = \{\mathcal{V}_\Sigma(X)\}_{\Sigma,X}$ is a GVTL (where Σ ranges over all ranked alphabets and X over all leaf alphabets), then for any fixed Σ we get a Σ-VTL $\mathcal{V}_\Sigma = \{\mathcal{V}_\Sigma(X)\}_X$.

A *positive $*$-variety* is defined as a generalized $*$-variety that is not required to be closed under complementation. Pin [64] shows that these families of regular languages can be characterized by *ordered syntactic monoids*, and he proves a variety theorem that connects positive $*$-varieties with varieties of finite ordered monoids. Of course there is a corresponding notion of positive $+$-varieties. Recently Petković and Salehi [60] have presented some tree language counterparts to this theory. In particular they extend the above mentioned theory of generalized varieties by establishing a correspondence between *generalized positive varieties of tree languages* and *generalized varieties of finite ordered algebras*. Any GVTL is naturally a generalized positive variety of tree languages but also other examples are presented in [60].

14 Alternative approaches

Transition semigroups of tree automata were already considered in the 1970's (for references, cf. [28, p. 123]), but syntactic monoids as a means to classify regular tree languages were introduced several years later by Thomas [88, 89]. A similar notion defined for binary trees is used in [53, 54, 55, 65].

Let us first note that the set $C_\Sigma(X)$ of all ΣX-contexts forms a monoid with respect to the product $p \cdot q = q(p)$ with ξ as the unit element. Recall also that for any ΣX-tree t and any ΣX-context p, $t \cdot p$ is the ΣX-tree $p(t)$. Following [88, 89], the *syntactic monoid congruence* of a ΣX-tree language T is now defined as the relation μ_T on $C_\Sigma(X)$ such that for any $p, q \in C_\Sigma(X)$,

$$p \;\mu_T\; q \quad \text{if and only if} \quad (\forall t \in T_\Sigma(X))(\forall r \in C_\Sigma(X))[t \cdot p \cdot r \in T \;\leftrightarrow\; t \cdot q \cdot r \in T].$$

The *syntactic monoid* $\mathrm{SM}(T)$ of T is the quotient monoid $C_\Sigma(X)/\mu_T$. It is easy to show that a ΣX-tree language is regular if and only if its syntactic monoid is finite. Of course, the *syntactic semigroup congruence* σ_T and the *syntactic semigroup* $\mathrm{SS}(T) = C^+_\Sigma(X)/\sigma_T$ are defined similarly, but restricting p and q to $C^+_\Sigma(X) := C_\Sigma(X) - \{\xi\}$.

The following facts about syntactic monoids of tree languages are due to K. Salomaa [75]:

(1) Every finite monoid is isomorphic to the syntactic monoid of a regular tree language over some (unary) ranked alphabet and a suitable leaf alphabet.

(2) The syntactic monoid of a regular tree language T is isomorphic to the translation monoid $\mathrm{Tr}(\mathrm{SA}(T))$ of the syntactic algebra of T.

(3) $\mathrm{SM}(T_\Sigma(X) - T) = \mathrm{SM}(T)$ and $\mathrm{SM}(T \cap U), \mathrm{SM}(T \cup U) \preceq \mathrm{SM}(T) \times \mathrm{SM}(U)$ for any $T, U \subseteq T_\Sigma(X)$.

(4) $\mathrm{SM}(p^{-1}(T)) \preceq \mathrm{SM}(T)$ for any $T \subseteq T_\Sigma(X)$ and any $p \in C_\Sigma(X)$.

(5) $\mathrm{SM}(T\varphi^{-1}) \preceq \mathrm{SM}(T)$ for any homomorphism $\varphi\colon \mathcal{T}_\Sigma(X) \to \mathcal{T}_\Sigma(Y)$ and $T \subseteq T_\Sigma(Y)$.

Let $\mathbf{M}$ be a variety of finite monoids. For any Σ and X, let $\mathbf{M}^t_\Sigma(X)$ be the set of all ΣX-tree languages T such that $\mathrm{SM}(T) \in \mathbf{M}$. If follows from the above facts (3)–(5) that $\mathbf{M}^t_\Sigma = \{\mathbf{M}^t_\Sigma(X)\}_X$ is a Σ-VTL.

14.1 Example A ΣX-tree language T is called *aperiodic*, or *non-counting*, if there is a number $n \geq 0$ such that for all $t \in T_\Sigma(X)$ and $p, q \in C_\Sigma(X)$, $t \cdot p^{n+1} \cdot q \in T$ if and only if $t \cdot p^n \cdot q \in T$. It is easy to verify directly that the family $\mathrm{Ap}_\Sigma = \{\mathrm{Ap}_\Sigma(X)\}_X$ of such tree languages is a Σ-VTL, but Thomas [88, 89] proved that $\mathrm{Ap}_\Sigma = \mathbf{Ap}^t$, where $\mathbf{Ap}$ is the VFM of finite *aperiodic* monoids, that is to say finite monoids in which all subgroups are trivial. He also shows that the families of aperiodic, first-order definable and star-free tree languages are all pairwise distinct although the corresponding string language families are all equal (cf. [50]). These families and related matters are studied further in [35, 36, 66, 68, 67, 69].

14.2 Example A Σ-algebra $\mathcal{A} = (A, \Sigma)$ is said to be *monotone* if there is an order $\leq$ on A such that $a_1, \ldots, a_m \leq f^{\mathcal{A}}(a_1, \ldots, a_m)$ for all $m > 0, f \in \Sigma_m$ and $a_1, \ldots, a_m \in A$. A ΣX-tree language is *monotone* if it is recognized by a finite monotone Σ-algebra. An element $u \in S$ of a semigroup S is a *divisor* of an element $s \in S$ if $s = uv$ or $s = vu$ for some $v \in S$. A subsemigroup is said to be *closed under divisors* if it contains all divisors of its elements. Finally, a subsemigroup S' of a semigroup S is called a *right-unit subsemigroup* if there is an element $s \in S$ such that $S' = \{u \in S \mid su = s\}$. Gécseg and Imreh [25] prove that a regular tree language is monotone if and only if every right-unit subsemigroup of its syntactic monoid is closed under divisors. Piirainen [61] notes that the finite monotone algebras form a GFVA and the monotone tree languages the corresponding GVTL. Moreover he shows that a regular tree language is monotone if and only if its syntactic monoid is $\mathcal{R}$-trivial.

Like the generalized variety theory discussed in the previous section, the syntactic monoid approach does not require a fixed ranked alphabet. In [84] it is noted that every family of recognizable tree languages $\mathbf{M}^t$ definable by syntactic monoids is a GVTL, but that the converse does not hold: not every GVTL is of the form $\mathbf{M}^t$. Indeed, for each VFM $\mathbf{M}$, $\mathbf{M}^t$ is just the greatest GVTL $\mathcal{V}$ such that $\mathrm{SM}(T) \in \mathbf{M}$ whenever T is in $\mathcal{V}$. For example, in [92] Wilke shows that the reverse definite tree languages cannot be characterized by syntactic monoids (or syntactic semigroups). Recently, Salehi [71] has described the GVTLs that can be characterized by syntactic monoids or syntactic semigroups. His results confirm the intuitive impression that the defining power of the syntactic monoid framework is quite limited for tree languages. In particular, it turns out that the definite tree languages cannot be defined by syntactic monoids or semigroups, contrary to what is claimed in [54]. Furthermore, in [62] Piirainen shows that the piecewise testable tree languages cannot be defined by syntactic monoids. In the paper [60] mentioned already in the previous section, Petković and Salehi also consider *ordered syntactic monoids* of tree languages and, extending the result of [71], they characterize the generalized positive varieties of tree languages that can be characterized by these.

The syntactic monoid of a regular language L is isomorphic to the transition monoid of the minimal recognizer of L, and this again is the monoid of all unary term functions of the algebra underlying the recognizer. This observation may be seen as the starting point of the classification theory by Z. Ésik [22], where everything is formulated in terms of (algebraic) theories of Lawvere (cf. [20, 22]). Such a *theory* is a category of a certain kind in which the objects are the non-negative integers. A theory T is *finitary* if for each pair of objects n, p, the set $T(n,p)$ of morphisms $n \to p$ is finite. Each ranked alphabet Σ defines a theory $\Sigma\mathrm{TM}$ of trees such that for each $n \geq 0$, we may identify $\Sigma\mathrm{TM}(n,1)$ with the set $T_\Sigma(X_n)$ of ΣX_n-trees. A tree language $L \subseteq \Sigma\mathrm{TM}(n,1)$ is then said to be *recognizable* if there exist a finitary theory T, a morphism $\varphi\colon \Sigma\mathrm{TM} \to T$ of theories and a set $P \subseteq T(n,1)$ such that $T = P\varphi^{-1}$. Ésik develops a theory of varieties of such recognizable tree languages that in its form resembles the theory described in Section 11, but syntactic algebras are replaced by *syntactic theories*, varieties of finite algebras by varieties of finitary theories, etc. The paper [22] also considers several examples, suggests some open problems, and compares the presented theory with other approaches. In particular, the relationship to the theory of generalized varieties of [84] is discussed in some detail.

As noted in [22], Ésik's theory could be reformulated in terms of *clones* (cf. [10, 15, 49]). In [23] Ésik and Weil introduce *syntactic preclones* and present a variety theorem that connects each variety of tree languages with a variety of finitary preclones. Within this framework they characterize the first-order definable tree languages as well as some families obtained by extending the FO language with various restricted second-order quantifiers.

The *tree algebra* framework of Wilke [92] is designed for languages of binary trees. The nodes of the trees considered are labelled with symbols taken from a finite label alphabet A, and each of these symbols may be used both as a binary symbol labelling an inner node or as a nullary symbol labelling a leaf. Hence A may be regarded as a (generalized) ranked alphabet such that $A_0 = A_2 = A$. Then the set T_A of *A-trees* and the set C_A of *A-contexts* are defined inductively by the following two clauses:

(1) $A \subseteq T_A$ and $\xi \in C_A$;

(2) if $a \in A$, $s, t \in T_A$ and $p \in C_A$, then $a(s,t) \in T_A$ and $a(p,t), a(t,p) \in C_A$.

To represent such trees and contexts, Wilke introduces a sorted ranked alphabet Γ with the three sorts $\mathbf{l}$ (*labels*), $\mathbf{t}$ (*trees*) and $\mathbf{c}$ (*contexts*), and the six operation symbols ι, κ, λ, ρ, η and σ. The types of these operation symbols and their interpretations in the Γ-algebra $\mathcal{T}_A = (A, T_A, C_A, \Gamma)$ of A-trees and A-contexts are as follows:

$\iota : \mathbf{l} \to \mathbf{t},\ a \mapsto a$; $\quad \kappa : \mathbf{ltt} \to \mathbf{t},\ (a,s,t) \mapsto a(s,t)$; $\quad \lambda : \mathbf{lt} \to \mathbf{c},\ (a,t) \mapsto a(\xi,t)$;

$\rho : \mathbf{lt} \to \mathbf{c},\ (a,t) \mapsto a(t,\xi)$; $\quad \eta : \mathbf{ct} \to \mathbf{t},\ (p,t) \mapsto p(t)$; $\quad \sigma : \mathbf{cc} \to \mathbf{c},\ (p,q) \mapsto p(q)$.

Hence in $\mathcal{T}_A$ the ι-operator forms from a label $a \in A$ the one-node A-tree a, the κ-operator creates from a label a and two A-trees s and t a new A-tree with an a-labelled root and s and t as the two maximal proper subtrees, etc. As a matter of fact Wilke considers the subalgebra $\mathcal{T}_A^+ = (A, T_A, C_A^+, \Gamma)$ of $\mathcal{T}_A$, where $C_A^+ := C_A - \{\xi\}$, obtained by omitting the unit A-context ξ. Each Γ-term with variables in A of sort $\mathbf{t}$ or of sort $\mathbf{c}$ represents in a natural way an A-tree or an A-context respectively, when these interpretations of the operators are adopted. In fact, it is obvious that every A-tree and every A-context has at least one such

representation. For example, the A-tree $a(b(a,a),b)$, where $a,b \in A$, is represented both by $\kappa(a,\kappa(b,\iota(a),\iota(a)),\iota(b))$ and by $\eta(\lambda(a,\iota(b)),\kappa(b,\iota(a),\iota(a)))$.

It is easy to verify that the algebra of A-trees $\mathcal{T}_A$ satisfies the identities

$$\sigma(\sigma(p,q),r) \simeq \sigma(p,\sigma(q,r)), \tag{TA1}$$
$$\eta(\sigma(p,q),t) \simeq \eta(p,\eta(q,t)), \tag{TA2}$$
$$\eta(\lambda(a,s),t) \simeq \kappa(a,t,s), \tag{TA3}$$
$$\eta(\rho(a,s),t) \simeq \kappa(a,s,t), \tag{TA4}$$

where p, q and r are variables of sort **c**, s and t are variables of sort **t**, and a is a variable of sort **l**. Any Γ-algebra satisfying the identities (TA1)–(TA4) is called a *tree algebra*. Wilke shows that $\mathcal{T}_A^+$ is the free tree algebra generated by the sorted set $\langle A,\emptyset,\emptyset\rangle$. The *syntactic tree algebra congruence* τ_T of an A-tree language $T \subseteq T_A$ is the greatest congruence on $\mathcal{T}_A^+$ saturating the sorted subset $\langle\emptyset,T,\emptyset\rangle$, and the *syntactic tree algebra* $\mathrm{STA}(T)$ of T is defined as the corresponding quotient algebra $\mathcal{T}_A^+/\tau_T$. It is not hard to see that an A-tree language T is regular if and only if its syntactic tree algebra $\mathrm{STA}(T)$ is finite. In fact, the **t**-component of $\mathrm{STA}(T)$ is in essence the syntactic algebra $\mathrm{SA}(T)$ of T while the **c**-component corresponds to the syntactic semigroup $\mathrm{SS}(T)$ of T. As an application of syntactic tree algebras, Wilke presents an elegant effective equational characterization of the reverse definite A-tree languages. The general theory of tree algebras is developed further in [74, 72].

15 Deterministic top-down tree languages

In this section we consider tree languages recognized by *deterministic top-down tree recognizers*, or *DT recognizers* for short, that read their input trees in a deterministic way starting at the root and ending at the leaves. Such tree automata are also called *deterministic root-to-frontier recognizers*, or *DR recognizers*. That DT recognizers are much weaker than the bottom-up recognizers considered above was already observed in the 1960s. However, the family of DT-recognizable tree languages is still quite interesting and it includes, for example, the sets of *production trees* of context-free grammars (cf. [29], for example). For general introductions to the topic and further references we refer the reader to [28, 29, 38, 39]. Here we shall consider the variety properties of this family of tree languages. In particular, we define the *syntactic path monoid* introduced by Gécseg and Steinby [30] as a tool for classifying DT-recognizable tree languages.

Let Σ be a ranked alphabet and X a leaf alphabet. To simplify matters, we assume that $\Sigma_0 = \emptyset$. A *deterministic top-down Σ-algebra*, a *DT Σ-algebra* for short, $\mathcal{A} = (A,\Sigma)$ consists of a non-empty set A and a Σ-indexed family of *top-down operations*

$$f^{\mathcal{A}}\colon A \longrightarrow A^m \quad (m > 0,\ f \in \Sigma_m).$$

Such a DT Σ-algebra $\mathcal{A}$ is *finite* if A is a finite set. In [27] it is shown that subalgebras, homomorphisms, congruences, quotient algebras and direct products—with their usual properties—can be defined in a natural way for DT-algebras.

15.1 Definition A *deterministic top-down ΣX-recognizer*, or a *DT ΣX-recognizer*, is a system $\mathbf{A} = (\mathcal{A}, a_0, \alpha)$, where

(1) $\mathcal{A} = (A, \Sigma)$ is a finite DT Σ-algebra,

(2) $a_0 \in A$ is the *initial state*, and

(3) $\alpha : X \to \wp A$ is the *final state assignment*.

The elements of A are the *states* of $\mathbf{A}$. To define the tree language recognized by $\mathbf{A}$, we first extend α to a mapping $\alpha_{\mathcal{A}} : T_\Sigma(X) \to \wp A$ thus:

- $x\alpha_{\mathcal{A}} = x\alpha$ for each $x \in X$;
- $f(t_1, \ldots, t_m)\alpha_{\mathcal{A}} = \{a \in A \mid f^{\mathcal{A}}(a) \in t_1\alpha_{\mathcal{A}} \times \cdots \times t_m\alpha_{\mathcal{A}}\}$.

Now the tree language *recognized* by $\mathbf{A}$ is defined as

$$T(\mathbf{A}) = \{t \in T_\Sigma(X) \mid a_0 \in t\alpha_{\mathcal{A}}\}.$$

A ΣX-tree language T is said to be *DT-recognizable* if $T = T(\mathbf{A})$ for a DT ΣX-recognizer $\mathbf{A}$. Let $\mathrm{DRec}_\Sigma = \{\mathrm{DRec}_\Sigma(X)\}_X$ where, for each X, $\mathrm{DRec}_\Sigma(X)$ is the set of all DT-recognizable ΣX-tree languages.

The process by which a DT ΣX-recognizer $\mathbf{A} = (\mathcal{A}, a_0, \alpha)$ accepts or rejects a given input tree $t \in T_\Sigma(X)$ can be described as follows:

- $\mathbf{A}$ starts at the root of t in state a_0;
- if $\mathbf{A}$ has reached node u labeled with $f \in \Sigma_m$ in state a, it continues to the i^{th} immediate successor node of u in state a_i $(1 \le i \le m)$, where $(a_1, \ldots, a_m) = f^{\mathcal{A}}(a)$;
- t is accepted if and only if $\mathbf{A}$ reaches each leaf in a state $a \in x\alpha$ matching the label x of that leaf.

The family DRec_Σ is closed under inverse translations and inverse homomorphisms, but not under all Boolean operations. Hence, it is not a variety of Σ-tree languages, but its Boolean closure BDRec_Σ is. The Σ-VTL BDRec_Σ has been studied extensively by Jurvanen [38, 39]. For example, she has established the inclusion relations between BDRec_Σ and many of the other Σ-VTLs mentioned before. In particular it is shown that BDRec_Σ is properly included in Rec_Σ (for any nontrivial Σ), a fact also established by Thomas [89] in a different way. However, it seems that the corresponding Σ-VFA is still not known.

In what follows, we shall consider a new kind of syntactic monoid introduced in [30] especially for DT-recognizable tree languages.

As shown by Gécseg and Steinby in [27], each DT ΣX-recognizer can be converted into an equivalent minimal DT ΣX-recognizer of a certain kind, and that this minimal recognizer is unique up to isomorphism. To make these statements precise, we have to introduce some new concepts. For any DT ΣX-recognizer $\mathbf{A} = (\mathcal{A}, a_0, \alpha)$ and any $a \in A$, let

$$T(\mathbf{A}, a) := \{t \in T_\Sigma(X) \mid a \in t\alpha_{\mathcal{A}}\},$$

and call state a a *0-state* if $T(\mathbf{A}, a) = \emptyset$. Clearly $T(\mathbf{A}, a) = \emptyset$ means that $\mathbf{A}$ accepts no tree starting in state a. Furthermore, a state $a \in A$ is said to be *reachable* if $a_0 \Rightarrow^* a$, where $\Rightarrow^*$ is the reflexive, transitive closure of the relation $\Rightarrow$ $(\subseteq A \times A)$ defined by

$$a \Rightarrow b \quad \text{if and only if} \quad b = \pi_i(f^{\mathcal{A}}(a)) \text{ for some } m > 0,\ f \in \Sigma_m \text{ and } 1 \le i \le m \quad (a, b \in A),$$

where $\pi_i \colon A^m \to A$ is the i^{th} projection mapping.

15.2 Definition We call a DT ΣX-recognizer $\mathbf{A} = (\mathcal{A}, a_0, \alpha)$

(1) *normalized* if, for all $m \geq 1$, $f \in \Sigma_m$ and $a \in A$, either every component in $f^{\mathcal{A}}(a) = (a_1, \ldots, a_m)$ is a 0-state or no a_i is a 0-state,

(2) *reduced* if $T(\mathbf{A}, a) = T(\mathbf{A}, b)$ implies $a = b$ $(a, b \in A)$,

(3) *connected* if all of its states are reachable, and

(4) *minimal* if it is connected and reduced.

15.3 Definition A *homomorphism* $\varphi : \mathbf{A} \to \mathbf{B}$ from a DT ΣX-recognizer $\mathbf{A} = (\mathcal{A}, a_0, \alpha)$ to a DT ΣX-recognizer $\mathbf{B} = (\mathcal{B}, b_0, \beta)$ is defined by a mapping $\varphi : A \to B$ such that

(1) $f^{\mathcal{B}}(a\varphi) = (a_1\varphi, \ldots, a_m\varphi)$ if $f^{\mathcal{A}}(a) = (a_1, \ldots, a_m)$, $m > 0$, $f \in \Sigma_m$, $a \in A$,

(2) $a_0\varphi = b_0$, and

(3) $x\beta\varphi^{-1} = x\alpha$ for every $x \in X$.

A bijective homomorphism $\varphi : \mathbf{A} \to \mathbf{B}$ is an *isomorphism* of DT ΣX-recognizers and in this case we say that $\mathbf{A}$ and $\mathbf{B}$ are *isomorphic*.

The following results of [27] show that the same facts about minimal recognizers hold for deterministic top-down tree recognizers as in the bottom-up case with the exception that uniqueness requires normalization.

15.4 Proposition *Any DT ΣX-recognizer $\mathbf{A}$ can be converted to an equivalent normalized minimal DT ΣX-recognizer $\mathbf{B}$, and this is also minimal with respect to the number of states among all DT ΣX-recognizers $\mathbf{B}$ such that $T(\mathbf{B}) = T(\mathbf{A})$. Moreover, any two equivalent minimal normalized DT ΣX-recognizers are isomorphic.*

Deterministic top-down recognizers are essentially weaker than their bottom-up counterparts because they have to accept or reject at each leaf separately without being able to combine or compare in any way the information gathered from different paths leading from the root to leaves. In fact, a DT recognizer accepts any tree t such that every path of t appears in some tree accepted by it, and this property actually characterizes the DT-recognizable tree languages as shown in [12, 90]. Since paths are also used for defining the syntactic monoids to be considered here, we now define them formally.

The *path alphabet* associated with Σ is the ordinary finite alphabet

$$\widehat{\Sigma} = \bigcup \{\Sigma_m \times \{1, \ldots, m\} \mid m > 0\}.$$

We write $(f, i) \in \widehat{\Sigma}$ as f_i. If f is the label of a node, i indicates the direction taken at that node. Hence words over $\widehat{\Sigma}$ represent paths leading from the root to a leaf in a ΣX-tree.

15.5 Definition For any $x \in X$, the set $g_x(t)$ of *x-paths* in a ΣX-tree t is defined as follows:

(1) $g_x(x) = \{e\}$;

(2) $g_x(y) = \emptyset$ for $y \in X$, $y \neq x$;

(3) $g_x(t) = f_1 g_x(t_1) \cup \cdots \cup f_m g_x(t_m)$ for $t = f(t_1, \ldots, t_m)$.

The *x-path language* of a tree language $T \subseteq T_\Sigma(X)$ is the set $T_x := \bigcup\{g_x(t) \mid t \in T\}$. The x-path languages T_x of T ($x \in X$) are collectively called the *path languages* of T. Furthermore the *path closure* of T is defined as the ΣX-tree language

$$c(T) := \{t \in T_\Sigma(X) \mid (\forall x \in X)\ g_x(t) \subseteq T_x\},$$

and T is said to be *closed* if $c(T) = T$.

It is clear that $T \mapsto c(T)$ is a closure operator on $T_\Sigma(X)$. The following lemma shows that a closed tree language is completely determined by its path languages.

15.6 Lemma *If $T \subseteq T_\Sigma(X)$ is closed and $t \in T_\Sigma(X)$, then $t \in T$ if and only if $g_x(t) \subseteq T_x$ for every $x \in X$.*

The following basic fact was proved by Courcelle [12] and by Virágh [90].

15.7 Proposition *A regular tree language is DT-recognizable if and only if it is closed.*

For example, the finite set $T = \{f(x, y), f(y, x)\}$ cannot be recognized by a DT recognizer since it is not closed. Indeed it is clear that if a DT recognizer accepts the trees $f(x, y)$ and $f(y, x)$, then it must accept also $f(x, x)$ and $f(y, y)$. On the other hand, it is easy to construct a DT recognizer for the closure $c(T) = \{f(x, y), f(y, x), f(x, x), f(y, y)\}$.

We shall use also the following simple observation.

15.8 Lemma *The path languages T_x ($x \in X$) of any DT-recognizable ΣX-tree language T are regular.*

Proof By Proposition 15.4, we may assume that $T = T(\mathbf{A})$ for a normalized DT ΣX-recognizer $\mathbf{A} = (\mathcal{A}, a_0, \alpha)$. Then for each $x \in X$, the language T_x is recognized by the $\widehat{\Sigma}$-recognizer $\mathbf{A}_x = (A, \widehat{\Sigma}, \delta, a_0, x\alpha)$, where δ is defined by setting $\delta(a, f_i) = a_i$ for all $a \in A$ and $f_i \in \widehat{\Sigma}$ assuming that $f^{\mathcal{A}}(a) = (a_1, \ldots, a_m)$. The fact that $\mathbf{A}$ is normalized is needed to prove $L(\mathbf{A}_x) \subseteq T_x$. □

The following notions were introduced in [30].

15.9 Definition The *syntactic path congruence* of a ΣX-tree language T is the relation $\widehat{\mu}_T$ on $\widehat{\Sigma}^*$ defined by

$$w_1 \,\widehat{\mu}_T\, w_2 \iff (\forall x \in X)(\forall u, v \in \widehat{\Sigma}^*)(uw_1v \in T_x \leftrightarrow uw_2v \in T_x) \quad (w_1, w_2 \in \widehat{\Sigma}^*).$$

The *syntactic path monoid* of T is the quotient monoid $\mathrm{PM}(T) = \widehat{\Sigma}^*/\widehat{\mu}_T$. The *syntactic path semigroup* $\mathrm{PS}(T)$ of T is similarly defined as a quotient monoid of $\widehat{\Sigma}^+$ restricting $\widehat{\mu}_T$ to $\widehat{\Sigma}^+$.

Since $\widehat{\mu}_T$ is the intersection of the usual syntactic congruences of the path languages of T, it really is a congruence on $\widehat{\Sigma}^*$ and the following lemma holds.

15.10 Lemma *For any ΣX-tree language T, $\widehat{\mu}_T$ is the greatest congruence on $\widehat{\Sigma}^*$ that saturates every path language T_x ($x \in X$) of T.*

The following counterpart to the usual Myhill Theorem is noted in [30].

15.11 Theorem *For any closed ΣX-tree language T the following are equivalent:*

(1) $T \in \mathrm{DRec}_\Sigma(X)$*;*

(2) *there is a congruence on $\widehat{\Sigma}^*$ of finite index that saturates all the path languages of T;*

(3) $\widehat{\mu}_T$ *is of finite index.*

Of course, the theorem yields the following immediate consequence.

15.12 Corollary *A closed tree language T is DT-recognizable if and only if the monoid* $\mathrm{PM}(T)$ *is finite.*

For any $T \in \mathrm{DRec}_\Sigma(X)$, the syntactic path monoid $\mathrm{PM}(T)$ is effectively computable because it can be shown to be isomorphic to the minimal recognizer of the family $(T_x \mid x \in X)$ of path languages of T, and this can be formed by Lemma 15.8; a recognizer of $(T_x \mid x \in X)$ is a finite automaton $\mathbf{A} = (A, \widehat{\Sigma}, \delta, a_0, (F_x \mid x \in X))$ with a family of sets of final states such that for every $x \in X$, $T_x = L(\mathbf{A}_x)$ for the $\widehat{\Sigma}$-recognizer $\mathbf{A}_x = (A, \widehat{\Sigma}, \delta, a_0, F_x)$. One may use the well-known minimization theory of Moore-automata to show that such a minimal recognizer always exists. The *transition monoid* $\mathrm{TM}(\mathbf{A})$ of $\mathbf{A}$ is formed by the mappings

$$w^{\mathbf{A}} : A \longrightarrow A,\ a \mapsto \delta(a, w), \quad (w \in \widehat{\Sigma}^*)$$

with $u^{\mathbf{A}} v^{\mathbf{A}} = (uv)^{\mathbf{A}}$ as the operation.

In conclusion, we present some examples of characterizations by path monoids.

15.13 Example A DT ΣX-recognizer $\mathbf{A} = (\mathcal{A}, a_0, \alpha)$ is *monotone* if there is an order $\leq$ on A such that $a \leq a_1, \ldots, a_m$, whenever $f^{\mathcal{A}}(a) = (a_1, \ldots, a_m)$ for some $m > 0$, $f \in \Sigma_m$ and $a, a_1, \ldots, a_m \in A$. A DT-recognizable tree language is *monotone* if it is recognized by a monotone DT recognizer. Gécseg and Imreh [25] show that a closed ΣX-tree language T is monotone if and only if every right-unit subsemigroup of $\mathrm{PM}(T)$ is closed under divisors. By the results of Piirainen [61] noted in Example 14.2, this means that T is monotone if and only if $\mathrm{PM}(T)$ is $\mathcal{R}$-trivial.

The following example is from [26] but we have slightly modified the terminology.

15.14 Example For any given DT Σ-algebra $\mathcal{A} = (A, \Sigma)$, we define the *state frontier* map $\mathrm{sf}\colon A \times T_\Sigma(X) \to A^*$ as follows:

(1) $\mathrm{sf}(a, x) = a$ for $a \in A$ and $x \in X$;

(2) $\mathrm{sf}(a, t) = \mathrm{sf}(a_1, t_1) \ldots \mathrm{sf}(a_m, t_m)$ for $t = f(t_1, \ldots, t_m)$, $a \in A$ and $f^{\mathcal{A}}(a) = (a_1, \ldots, a_m)$.

The *minimum height* $\mathrm{mh}(t)$ of a ΣX-tree t is defined by the conditions:

(1) $\mathrm{mh}(x) = 0$ for any $x \in X$;

(2) $\mathrm{mh}(t) = \min(\mathrm{mh}(t_1), \ldots, \mathrm{mh}(t_m)) + 1$ for $t = f(t_1, \ldots, t_m)$.

For any $k \geq 0$, a DT Σ-algebra $\mathcal{A} = (A, \Sigma)$ is called *frontier k-definite* if $\mathrm{sf}(a,t) = \mathrm{sf}(b,t)$ for all $a, b \in A$ whenever $\mathrm{mh}(t) \geq k$, and we call $\mathcal{A}$ *frontier definite* if it is frontier k-definite for some $k \geq 0$. A DT ΣX-recognizer $\mathbf{A} = (\mathcal{A}, \alpha, F)$ and the tree language $T(\mathbf{A})$ recognized by it are called *frontier definite* if the underlying DT Σ-algebra is frontier definite. Recall also that a semigroup S is *definite* if $Su = \{u\}$ for every idempotent $u \in S$, that is to say every idempotent of S is a right zero. Now, Gécseg and Imreh [26] show that a DT-recognizable tree language T is frontier definite if and only if its syntactic path semigroup $\mathrm{PS}(T)$ is definite.

Gécseg and Imreh [26] give also a characterization of a family of similarly defined *nilpotent DT-recognizable* tree languages. These examples may appear somewhat contrived as the families of tree languages are defined through properties of their recognizers, but it is to be noted that the characterizations give decision methods that do not require any knowledge about the parameter k. It seems that the syntactic monoid (or semigroup) approach is more useful in the case of DT-recognizable tree languages than in the case of all recognizable tree languages because DT-recognizable tree languages are determined by their path languages.

Acknowledgements

I have derived much inspiration from working with Tatjana Petković, Ville Piirainen and Saeed Salehi on topics related to this paper. I also thank all of them for reading the manuscript and for their many useful comments. Thanks are also due to the organizers of the excellent seminar that led me to write this paper, as well to the patient editors of this volume. I am especially grateful to Martin Goldstein for his careful editorial work.

References

[1] J. Almeida, On pseudovarieties, varieties of languages, filters of congruences, pseudoidentities and related topics, *Algebra Universalis* **27** (1990), 333–350.

[2] J. Almeida, *Finite Semigroups and Universal Algebra*, World Scientific, Singapore, 1995.

[3] J. Avenhaus, *Reduktionssysteme*, Springer-Verlag, Berlin, 1995.

[4] C. J. Ash, Pseudovarieties, generalized varieties and similarly described classes, *J. Algebra* **92** (1985), 104–115.

[5] J. T. Baldwin and J. Berman, Varieties and finite closure conditions, *Colloq. Math.* **35** (1976), 15–20.

[6] W. Brauer, *Automatentheorie*, B. G. Teubner, Stuttgart, 1984.

[7] J. A. Brzozowski, Canonical regular expressions and minimal state graphs of definite events, in: *Proc. Symp. Math. Theory of Automata*, Microwave Research Inst. Symp. Ser. **12**, Brooklyn, New York, 1963, 529–561.

[8] J. R. Büchi, *Finite Automata, Their Algebras and Grammars. Towards a Theory of Formal Expressions*, Springer-Verlag, New York, 1989.

[9] S. Burris and H. P. Sankappanavar, *A Course in Universal Algebra*, Springer-Verlag, New York, 1981.

[10] P. M. Cohn, *Universal Algebra*, D. Reidel, Dordrecht, 2nd revised ed., 1981.

[11] H. Comon, M. Dauchet, R. Gilleron, F. Jacquemard, D. Lugiez, S. Tison, and M. Tommasi, *Tree Automata and Applications*, `http://www.grappa.univ-lille3.fr/tata/` (project under development since 1997).

[12] B. Courcelle, A representation of trees by languages I, *Theoret. Comput. Sci.* **6** (1978), 255–279.

[13] B. Courcelle, On recognizable sets and tree-automata, in: *Resolution of Equations in Algebraic Structures, Vol. 1* (M. Nivat and H. Aït-Kaci, eds.), Academic Press, New York, 1989, 93–126.

[14] B. Courcelle, Basic notions of universal algebra for language theory and graph grammars, *Theoret. Comput. Sci.* **163** (1996), 1–54.

[15] K. Denecke and S. L. Wismath, *Universal Algebra and Applications in Theoretical Computer Science*, Chapman & Hall/CRC, Boca Raton, 2002.

[16] J. Doner, Tree acceptors and some of their applications, *J. Comput. System Sci.* **4** (1970), 406–451.

[17] S. Eilenberg, *Automata, Languages, and Machines, Vol. A*, Academic Press, New York, 1974.

[18] S. Eilenberg, *Automata, Languages, and Machines, Vol. B*, Academic Press, New York, 1976.

[19] S. Eilenberg and M.-P. Schützenberger, On pseudovarieties, *Advances in Mathematics* **19** (1976), 413–418.

[20] S. Eilenberg and J. B. Wright, Automata in general algebras, *Information and Computation* **11** (1967), 452–470.

[21] Z. Ésik, Definite tree automata and their cascade composition, *Publ. Math. Debrecen* **48**, 3–4 (1996), 243–261.

[22] Z. Ésik, A variety theory for trees and theories, *Publ. Math. Debrecen* **54** (1999), 711–762.

[23] Z. Ésik and P. Weil, On certain logically defined regular tree languages, in: *Foundations of Software Technology and Theoretical Computer Science*, Lecture Notes in Comput. Sci. **2914**, Springer, Berlin 2003, 195–207.

[24] F. Gécseg and B. Imreh, On a special class of tree automata, in: *Automata, Languages and Programming Systems*, Dept. Math., K. Marx University of Economics, Budapest, 1988, 141–152.

[25] F. Gécseg and B. Imreh, On monotone automata and monotone languages, *J. Automata, Languages and Combinatorics* **7** (2002), 71–82.

[26] F. Gécseg and B. Imreh, On definite and nilpotent DR tree languages, *J. Automata, Languages and Combinatorics* **9** (2004), 55–60.

[27] F. Gécseg and M. Steinby, Minimal ascending tree automata, *Acta Cybernetica* **4** (1978), 37–44.

[28] F. Gécseg and M. Steinby, *Tree automata*, Akadémiai Kiadó, Budapest, 1984.

[29] F. Gécseg and M. Steinby, Tree languages, in: *Handbook of Formal Languages, Vol. 3* (G. Rozenberg and A. Salomaa, eds.), Springer-Verlag, Berlin, 1997, 1–69.

[30] F. Gécseg and M. Steinby, Minimal recognizers and syntactic monoids of DR tree languages, in: *Words, Semigroups and Transductions* (M. Ito, G. Paun, and S. Yu, eds.), World Scientific, Singapore, 2001, 155–167.

[31] A. Ginzburg, About some properties of definite, reverse-definite and related automata, *IEEE Trans. Electronic Computers* **EC-15** (1966), 809–810.

[32] G. Grätzer, *Universal Algebra*, Springer-Verlag, New York, 2nd revised ed., 1979.

[33] U. Heuter, Definite tree languages, *Bulletin of the EATCS* **35** (1988), 137–142.

[34] U. Heuter, Generalized definite tree languages, in: *Math. Found. of Comput. Sci.*, Lecture Notes in Comput. Sci. **379**, Springer, Berlin, 1989, 270–280.

[35] U. Heuter, *Zur Klassifizierung regulärer Baumsprachen*, Dissertation, Faculty of Natural Science, RWTH, Aachen, 1989.

[36] U. Heuter, First-order properties of trees, star-free expressions and aperiodicity, *RAIRO Theoret. Informatics and Appl.* **25** (1991) 125–145.

[37] J. M. Howie, *Automata and Languages*, Clarendon Press, Oxford, 1991.

[38] E. Jurvanen, The Boolean closure of DR-recognizable tree languages, *Acta Cybernetica* **10** (1992), 255–272.

[39] E. Jurvanen, On tree languages defined by deterministic root-to-frontier recognizers, Ph.D. thesis, University of Turku, 1995.

[40] E. Jurvanen, A. Potthoff, and W. Thomas, Tree languages recognizable by regular frontier check, in: *Developments in language theory* (G. Rozenberg and A. Salomaa, eds.), World Scientific, Singapore 1994, 3–17.

[41] A. Kelarev and O. Sokratova, Directed graphs and syntactic algebras of tree languages, *J. Automata, Languages and Combinatorics* **6** (2001), 305–311.

[42] S. C. Kleene, Representation of events in nerve nets and finite automata, in: *Automata Studies* (C. E. Shannon and J. McCarthy, eds.), Princeton University Press, Princeton, NJ, 1956, 3–42.

[43] T. Knuutila, Inference of k-testable tree languages, in: *Advances in Structural and Syntactic Pattern Recognition* (H. Bunke, ed.), World Scientific, Singapore 1992, 109–120.

[44] D. C. Kozen, *Automata and Computability*, Springer, New York 1997.

[45] G. Lallement, *Semigroups and combinatorial applications*, John Wiley & Sons, New York, 1979.

[46] D. López, Inferencia de lenguajes de árboles, Ph.D. thesis, Universida Politécnica de Valencia, 2003.

[47] M. Magidor and G. Moran, Finite automata over finite trees, Technical Report **30**, Hebrew University, Jerusalem, 1969.

[48] T. S. E. Maibaum, A generalized approach to formal languages, *J. Comput. Systems Sci.* **8** (1974), 409–439.

[49] R. N. McKenzie, G. F. McNulty, and W. F. Taylor, *Algebras, Lattices, Varieties, Vol. I*, Wadsworth & Brooks/Cole, Monterey, CA, 1987.

[50] R. McNaughton, and S. Papert, *Counter-free automata*, MIT Press, Cambridge, MA, 1971.

[51] K. Meinke and J. V. Tucker, Universal algebra, in: *Handbook of Logic in Computer Science, Vol. 1* (S. Abramsky, D. Gabbay, and T. S. Maibaum, eds.), Clarendon Press, Oxford 1992, 189–411.

[52] J. Mezei and J. B. Wright, Algebraic automata and context-free sets, *Information and Control* **11** (1967), 3–29.

[53] M. Nivat and A. Podelski, Tree monoids and recognizability of sets of finite trees, in: *Resolution of Equations in Algebraic Theories, Vol. 1* (H. Aït-Kaci and M. Nivat, eds.), Academic Press, Boston, 1989, 351–367.

[54] M. Nivat and A. Podelski, Definite tree automata (cont'd), *Bull. EATCS* **38** (1989), 186–190.

[55] P. Péladeau and A. Podelski, On reverse and general definite tree languages, in: *Automata, Languages and Programming*, Lecture Notes in Comput. Sci. **623**, Springer, Berlin, 1992, 150–161.

[56] M. Perles, M. O. Rabin, and E. Shamir, The theory of definite automata, *IEEE Trans. Electronic Computers* **EC-12** (1963), 233–243.

[57] T. Petković, Varieties of automata and semigroups (in Serbian), Ph.D. thesis, University of Niš, 1998.

[58] T. Petković, M. Ćirić, and S. Bogdanović, Characteristic semigroups of directable automata, in: *Proceedings of Developments in Language Theory 2002* (M. Ito and M. Toyama, eds.), Lecture Notes in Computer Science **2450**, Springer, Berlin, 2003, 417–428.

[59] T. Petković, M. Ćirić, and S. Bogdanović, Unary algebras, semigroups and congruences on free semigroups, *Theoret. Comput. Sci.* **324** (2004), 87–105.

[60] T. Petković and S. Salehi, Positive varieties of tree languages, TUCS Technical Report **622**, Turku Centre for Computer Science, Turku, 2004.

[61] V. Piirainen, Monotone algebras, $\mathcal{R}$-trivial monoids and a variety of tree languages, *Bulletin of the EATCS* **84** (2004), 189–194.

[62] V. Piirainen, Piecewise testable tree languages, TUCS Technical Report **634**, Turku Centre for Computer Science, Turku 2004.

[63] J.-E. Pin, *Varieties of formal languages*, North Oxford Academic Publ., London, 1986.

[64] J.-E. Pin, Syntactic semigroups, in: *Handbook of Formal Languages, Vol. 1* (G. Rozenberg, A. Salomaa, eds.), Springer, Berlin, 1997, 679–746.

[65] A. Podelski, A monoid approach to tree automata, in: *Tree Automata and Languages* (M. Nivat and A. Podelski, eds.), Elsevier Science Publishers, Amsterdam, 1992, 41–56.

[66] A. Potthoff, *Logische Klassifizierung regulärer Baumsprachen*, Bericht Nr. 9410, Institut für Informatik und Praktische Mathematik, Christian-Albrechts-Universität, Kiel, 1994.

[67] A. Potthoff, Modulo counting quantifiers over finite trees, *Theor. Comput. Sci.* **126** (1994), 97–112.

[68] A. Potthoff, First-order logic on finite trees. *Theory and Practice of Software Development* (P. D. Mosses et al., eds.), Lecture Notes in Comput. Sci. **915**, Springer, Berlin, 1995, 125–139.

[69] A. Potthoff and W. Thomas, Regular tree languages without unary symbols are star-free, *Fundamentals of Computation Theory*, Lecture Notes in Computer Science **710**, Springer, Berlin, 1993, 396–405.

[70] J. Reiterman, The Birkhoff theorem for finite algebras, *Algebra Universalis* **14** (1985), 1–10.

[71] S. Salehi, Varieties of tree languages definable by syntactic monoids, TUCS Technical Report **619**, Turku Centre for Computer Science, Turku 2004.

[72] S. Salehi, Varieties of Tree Languages, Ph.D. thesis, University of Turku, to be presented in 2005.

[73] S. Salehi and M. Steinby, Varieties of many-sorted recognizable sets, TUCS Technical Report **626**, Turku Centre for Computer Science, Turku, 2004.

[74] S. Salehi and M. Steinby, Tree algebras and regular tree languages, in preparation.

[75] K. Salomaa, *Syntactic monoids of regular forests*, M.Sc. thesis, University of Turku, 1983, in Finnish.

[76] B. M. Schein, Homomorphisms and subdirect decompositions of semigroups, *Pacific J. Math.* **17** (1966), 529–547.

[77] M.-P. Schützenberger, On finite monoids having only trivial subgroups, *Information and Control* **8** (1965), 190–194.

[78] M. Steinby, Syntactic algebras and varieties of recognizable sets, in: *Les Arbres en algèbre et en programmation*, Université de Lille, Lille, 1979, 226–240.

[79] M. Steinby, Some algebraic aspects of recognizability and rationality, *Foundations of Computing Theory*, Lecture Notes in Comput. Sci. **117**, Springer, Berlin, 1981, 360–372.

[80] M. Steinby, A theory of tree language varieties, in: *Tree Automata and Languages* (M. Nivat and A. Podelski, eds.), North-Holland, Amsterdam, 1992, 57–81.

[81] M. Steinby, On generalizations of the Nerode and Myhill theorems, *Bulletin of the EATCS* **48** (1992), 191–198.

[82] M. Steinby, Recognizable and rational subsets of algebras, *Fundamenta Informaticae* **18** (1993), 249–266.

[83] M. Steinby, Classifying regular languages by their syntactic algebras, in: *Results and Trends in Computer Science*, Lecture Notes in Comput. Sci. **812**, Springer, Berlin, 1994, 353–364.

[84] M. Steinby, General varieties of tree languages, *Theoret. Comput. Sci.* **205** (1998), 1–43.

[85] J. W. Thatcher and J. B. Wright, Generalized finite automata theory with an application to a decision problem of second order logic, *Mathematical Systems Theory* **2** (1968), 57–81.

[86] D. Thérien, *Classification of regular congruences*, Ph.D. Thesis, University of Waterloo, 1980.

[87] D. Thérien, Classification of finite monoids: the language approach, *Theoretical Computer Science* **14** (1981), 195–208.

[88] W. Thomas, On non-counting tree languages, *Grundlagen der Theoretische Informatik* (Proc. 1. Intern. Workshop, Paderborn 1982), 234–242.

[89] W. Thomas, Logical aspects in the study of tree languages, in: *9th Colloquium on Trees in Algebra and Programming* (B. Courcelle, ed.), Cambridge University Press, London, 1984, 31–49.

[90] J. Virágh, Deterministic ascending tree automata I, *Acta Cybernetica* **5** (1980), 33–42.

[91] W. Wechler, *Universal Algebra for Computer Scientists*, Springer-Verlag, Berlin, 1992.

[92] T. Wilke, An algebraic characterization of frontier testable tree languages, *Theoret. Comput. Sci.* **154** (1996), 85–106.

[93] R. T. Yeh, Some structural properties of generalized automata and algebras, *Mathematical Systems Theory* **5** (1971), 306–318.

[94] S. Yu, Regular languages, in: *Handbook of Formal Languages, Vol. 1* (G. Rozenberg and A. Salomaa, eds.), Springer-Verlag, Berlin, 1997, 41–110.

Index