Matthias Dehmer
Frank Emmert-Streib
Alexander Mehler
Editors

Towards an Information Theory of Complex Networks

Statistical Methods and Applications

Birkhäuser

Matthias Dehmer
Frank Emmert-Streib
Alexander Mehler
Editors

Towards an Information Theory of Complex Networks

Statistical Methods and Applications

Birkhäuser

Editors
Matthias Dehmer
UMIT
Institute of Bioinformatics
and Translational Research
Eduard-Wallnöfer-Zentrum I
A-6060 Hall in Tirol
Austria
mathias.dehmer@umit.at

Frank Emmert-Streib
School of Medicine, Dentistry
and Biomedical Sciences
Center for Cancer Research and Cell Biology
Queen's University Belfast
97 Lisburn Road
Belfast BT9 7BL
United Kingdom
v@bio-complexity.com

Alexander Mehler
Faculty of Computer Science
and Mathematics
Goethe-University Frankfurt
am Main Robert-Mayer-Straße 10
P.O. Box: 154
D-60325 Frankfurt am Main
Germany
mehler@em.uni-frankfurt.de

ISBN 978-0-8176-4903-6 e-ISBN 978-0-8176-4904-3
DOI 10.1007/978-0-8176-4904-3
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2011932673

Mathematics Subject Classification (2010): 68R10, 68P30, 94C15

Printed on acid-free paper

www.birkhauser-science.com

Preface

For more than a decade, complex network analysis has evolved as a methodological paradigm for a multitude of disciplines, including physics, chemistry, biology, geography, sociology, computer science, statistics, media science, and linguistics. Researchers in these fields share an interest in information processing subject to the networking of their corresponding research object, for instance, genes, molecules, individuals, semes, memes, etc. They start with the insight that any of these research objects is *extrinsically* characterized, if not constituted, by its networking with objects of the same provenance. In this way, networks, for example, gene networks, food networks, city networks, networks of words, sentences, texts, or web documents become important research objects in more and more disciplines.

This book, in line with these research developments, presents theoretical and practical results of statistical models of complex networks in the formal sciences, the natural sciences, and the humanities. One of its goals is to advocate and promote combinations of graph-theoretic, information-theoretic, and statistical methods as a way to better understand and characterize real-world networks.

On the one hand, networks appear as paradigmatic objects of approaches throughout the natural and social sciences and the humanities. On the other hand, networks are—irrespective of their disciplinary provenance—known for characteristic distributions of graph-theoretic invariants which affect their robustness and efficiency in information processing. The main goal of this book is to further develop information-theoretic notions and to elaborate statistical models of information processing in such complex networks. In this way, the book includes first steps toward establishing a statistical information theory as a unified basis for complex network analysis across a multitude of scientific disciplines.

The book presents work on the statistics of complex networks together with applications of information theory in a range of disciplines such as quantitative biology, quantitative chemistry, quantitative sociology, and quantitative linguistics. It aims to integrate models of invariants of network topologies and dynamic aspects of information processing in these networks or by means of these networks.

Thus, the book is in support of sharing and elaborating models and methods that may help researchers get insights into complex problems emerging from interdisciplinary reasoning.

The book is divided into two parts: Chaps. 1–4 deal with formal-theoretical issues of network modeling, while Chaps. 5–13 further develop and apply these methods to empirical networks from a wide range of areas. The book starts with a theoretical contribution by *Abbe Mowshowitz* on the entropy of digraphs and infinite graphs. The aim is to provide insights into more complex graph models that go beyond the majority of network models based on finite undirected graphs. The chapter by *Nicolas Bonichon, Cyril Gavoille*, and *Nicolas Hanusse* presents an information-theoretic upper bound of planar graphs by means of the newly introduced notion of well-orderly maps. Such a technique might be useful when studying properties of the very important notion of planar graphs. *Terence Chan* and *Raymond W. Yeung* study a statistical inference problem using network models. *Richard Berkovits, Lukas Jahnke*, and *Jan W. Kantelhardt* examine phase transitions within complex networks that help to examine their structural properties.

The remainder of the book combines the theoretical stance of the first section with an empirical analysis of real networks. *Elena Konstantinova* provides a survey on information-theoretic measures used in chemical graph theory. *Prabhat K. Sahu* and *Shyi-Long Lee* develop a model of chemical graphs by example of molecular networks. Exploring the spectral characteristics of these graphs, they provide a successful classification of chemical graphs.

Biological or, more specifically, ecological networks are dealt with by *Robert E. Ulanowicz* who describes a framework of quantifying patterns of the interaction of networked trophic processes from the point of view of information theory. Ecological networks are also the focus of the chapter of *Linda J. Moniz, James D. Nichols, Jonathan M. Nichols, Evan G. Cooch*, and *Louis M. Pecora*, who provide an approach to modeling the interaction dynamics of ecosystems and their change. A comprehensive view of ontologically disparate networks is given by *Cristian R. Munteanu, J. Dorado, A. Pazos Sierra, F. Prado-Prado, L.G. Pérez-Montoto, S. Vilar, F.M. Ubeira, A. Sanchez-Gonzaléz, M. Cruz-Monteagudo, S. Arrasate, N. Sotomayor, E. Lete, A. Duardo-Sánchez, A. Díaz-López, G. Patlewicz*, and *H. González-Díaz* who use the notion of entropy centrality to compare various systems such as chemical, biological, crime, and legislative networks, thereby showing the interdisciplinary expressiveness of complex network theory.

The book continues with two contributions to linguistic networks: *Alexander Mehler* develops a framework for analyzing the topology of social ontologies as they evolve within Wikipedia and contrasts them with nonsocial, formal ontologies. *Olga Abramov* and *Tatjana Lokot* present a comparative, classificatory study of morphological networks by means of several measures of graph entropy.

Edward B. Allen discusses the measurement of the complexity and error probability of software systems represented as hypergraphs. Finally, in the chapter by *Philippe Blanchard* and *Dimitri Volchenkov*, random walks are studied as a kind of Markov process on graphs that allow insights into the dynamics of networks as diverse as city and trade and exchange networks.

With such a broad field, it is clear that the present book addresses an interdisciplinary readership. It does not simply promote transdisciplinary research. Rather, it is about interdisciplinary research that may be the starting point of developing an overarching network science.

Matthias Dehmer
Frank Emmert-Streib
Alexander Mehler

Acknowledgments

Many colleagues have provided us with input, help, and support (consciously or unconsciously) before and during the preparation of this book. In particular, we would like to thank Andreas Albrecht, Gökmen Altay, Gabriel Altmann, Alain Barrat, Igor Bass, David Bialy, Philippe Blanchard, Danail Bonchev, Stefan Borgert, Mieczysław Borowiecki, Andrey A. Dobrynin, Michael Drmota, Ramon Ferrer i Cancho, Maria and Gheorghe Duca, Maria Fonoberova, Armin Graber, Martin Grabner, Peter Gritzmann, Ivan Gutman, Peter Hamilton, Wilfried Imrich, Patrick Johnston, Elena Konstantinova, D. D. Lozovanu, Dennis McCance, Abbe Mowshowitz, Arcady Mushegian, Andrei Perjan, Armindo Salvador, Maximilian Schich, Heinz Georg Schuster, Helmut Schwegler, Andre Ribeiro, Burghard Rieger, Brigitte Senn-Kircher, Fred Sobik, Doru Stefanescu, John Storey, Shailesh Tripathi, Kurt Varmuza, Bohdan Zelinka, and Shu-Dong Zhang. Additionally, Matthias Dehmer thanks Armin Graber for strong support and providing a fruitful atmosphere at UMIT. Finally, we would like to thank our editor Tom Grasso who has been always available and helpful.

The work on the chapters of Philippe Blanchard and Dimitri Volchenkov, Olga Abramov, and Alexander Mehler have been supported by the German Federal Ministry of Education and Research (BMBF) through the project *Linguistic Networks*.[1] We gratefully acknowledge this financial support.

[1] www.linguistic-networks.net.

Contents

Contributors

Olga Abramov University of Bielefeld, Universitätsstraße 25, 33615 Bielefeld, Germany, olga.abramov@uni-bielefeld.de

Edward B. Allen Department of Computer Science and Engineering, Mississippi State University, Box 9637, Mississippi State, MS 39762, USA, edward.allen@computer.org

S. Arrasate Department of Organic Chemistry II, Faculty of Science and Technology, University of the Basque Country/Euskal Herriko Unibertsitatea, Apto. 644, 48080 Bilbao, Spain, sonia.arrasate@ehu.es

Richard Berkovits Minerva Center and Department of Physics, Bar-Ilan University, Ramat-Gan 52900, Israel, berkov@mail.biu.ac.il

Philippe Blanchard Bielefeld – Bonn Stochastic Research Center (BiBoS), University of Bielefeld, Postfach 100131, 33501 Bielefeld, Germany

Nicolas Bonichon LaBRI, University of Bordeaux, 351 Cours de la libération, 33405 Bordeaux, France, bonichon@labri.fr

Terence H. Chan Institute for Telecommunications Research, University of South Australia, Adelaide, SA 5095, Australia, hlchan6@gmail.com; terence.chan@unisa.edu.au

E.G. Cooch Department of Natural Resources, Cornell University, Ithaca, NY 14853, USA, evan.cooch@cornell.edu

M. Cruz-Monteagudo CEQA, Faculty of Chemistry and Pharmacy, UCLV, Santa Clara 54830, Cuba, gmailkelcm@yahoo.es

A. Díaz-López Department of Special Public Law, Faculty of Law, University of Santiago de Compostela, 15782 Santiago de Compostela, Spain, antonio.lopez.diaz@usc.es

J. Dorado Department of Information and Communication Technologies, Computer Science Faculty, University of A Coruña, 15071 A Coruña, Spain, julian@udc.es

A. Duardo-Sánchez Department of Special Public Law, Faculty of Law, University of Santiago de Compostela, 15782 Santiago de Compostela, Spain, aliuskaduardo@yahoo.es

Cyril Gavoille LaBRI, University of Bordeaux, 351 Cours de la libération, 33405 Bordeaux, France, gavoille@labri.fr

H. González-Díaz Department of Microbiology and Parasitology, Faculty of Pharmacy, University of Santiago de Compostela, 15782 Santiago de Compostela, Spain, humberto.gonzalez@usc.es

Nicolas Hanusse LaBRI, CNRS – University of Bordeaux, 351 Cours de la libération, 33405 Bordeaux, France, hanusse@labri.fr

Lukas Jahnke Martin-Luther-Universität Halle-Wittenberg, 06099 Halle, Germany

Jan W. Kantelhardt Martin-Luther-Universität Halle-Wittenberg, 06099 Halle, Germany

Elena Konstantinova Sabolev Institute of Mathematics, Siberian Branch of Russian Academy of Sciences, 630090 Novosibirsk, Russia, e_konsta@math.nsc.ru

Shyi-Long Lee Department of Chemistry and Biochemistry, National Chung Cheng University, Chia-Yi, 621 Taiwan, chesll@ccu.edu.tw

E. Lete Department of Organic Chemistry II, Faculty of Science and Technology, University of the Basque Country/Euskal Herriko Unibertsitatea, Apto. 644, 48080 Bilbao, Spain, esther.lete@ehu.es

Tatiana Lokot Faculty of Technology, University of Bielefeld, Universitaetsstr. 25, 33615 Bielefeld, Germany, tlokot@math.uni-bielefeld.de

Alexander Mehler Faculty of Computer Science and Mathematics, Goethe University Frankfurt am Main, D-60325 Frankfurt am Main, Germany, mehler@em.uni-frankfurt.de

L.J. Moniz Johns Hopkins University, Applied Physics Laboratory, 11100 Johns Hopkins Road, Laurel, MD 20723, USA, lindano@comcast.net

Abbe Mowshowitz Department of Computer Science, The City College of New York (CUNY), 138th Street at Convent Avenue, New York, NY 10031, USA, abbe@cs.ccny.cuny.edu

C.R. Munteanu Department of Information and Communication Technologies, Computer Science Faculty, University of A Coruña, 15071 A Coruña, Spain, cmunteaunu@udc.es

J.D. Nichols U.S. Geological Survey, Patuxent Wildlife Research Center, Laurel, MD 20708, USA, jnichols@usgs.gov

J.M. Nichols Naval Research Laboratory, Optical Sciences Division, Code 5673, Washington, DC 20375, USA, jonathan.nichols@nrl.navy.mil

G. Patlewicz Institute for Health and Consumer Protection (IHPC), Joint Research Centre (JRC), European Commission, via E. Fermi 2749–21027 Ispra (Varese), Italy

DuPont Haskell Global Centers for Health and Environmental Sciences, Newark, DE 19711, USA, Grace.Y.Tier@usa.dupont.com

Alejandro Pazos-Sierra Department of Information and Communication Technologies, Computer Science Faculty, University of A Coruña,
15071 A Coruña, Spain, apazos@udc.es

L.M. Pecora Naval Research Laboratory, Code 6362, Washington, DC 20375, USA, pecora@anvil.nrl.navy.mil

L.G. Pérez-Montoto Faculty of Pharmacy, University of Santiago de Compostela, 15782 Santiago de Compostela, Spain, lgmp2002@yahoo.es

F. Prado-Prado Faculty of Pharmacy, University of Santiago de Compostela, 15782 Santiago de Compostela, Spain, fenol1@hotmail.com

Prabhat K. Sahu Institüt für Physikalische und Theoretische Chemie, Universität Würzburg, Am Hubland, 97074 Würzburg, Germany

Department of Chemistry and Biochemistry, National Chung Cheng University, Chia-Yi, 621 Taiwan, sahu@chemie.uni-wuerzburg.de

A. Sanchez-Gonzaléz Department of Inorganic Chemistry, Faculty of Pharmacy, University of Santiago de Compostela, 15782 Santiago de Compostela, Spain, angeles.sanchez@usc.es

N. Sotomayor Department of Organic Chemistry II, Faculty of Science and Technology, University of the Basque Country/Euskal Herriko Unibertsitatea, Apto. 644, 48080 Bilbao, Spain, nuria.sotomayor@ehu.es

F.M. Ubeira Department of Microbiology and Parasitology, Faculty of Pharmacy, University of Santiago de Compostela, 15782 Santiago de Compostela, Spain,
fm.ubeira@usc.es

Robert E. Ulanowicz Department of Biology, University of Florida, Gainesville, FL 32611-8525, USA

University of Maryland Center for Environmental Science, Solomons, MD 20688-0038, USA, ulan@umces.edu

S. Vilar Faculty of Pharmacy, University of Santiago de Compostela, 15782 Santiago de Compostela, Spain, qosanti@yahoo.es

Dimitri Volchenkov The Center of Excellence Cognitive Interaction Technology (CITEC), University of Bielefeld, Postfach 100131, 33501 Bielefeld, Germany, volchenk@physik.uni-bielefeld.de

Raymond W. Yeung Department of Information Engineering, The Chinese University of Hong Kong, whyeung@ie.cuhk.edu.hk

Chapter 1
Entropy of Digraphs and Infinite Networks

A. Mowshowitz

Abstract The information content of a graph G is defined in Mowshowitz (Bull Math Biophys 30:175–204, 1968) as the entropy of a finite probability scheme associated with the vertex partition determined by the automorphism group of G. This provides a quantitative measure of the symmetry structure of a graph that has been applied to problems in such diverse fields as chemistry, biology, sociology, and computer science (Mowshowitz and Mitsou, Entropy, orbits and spectra of graphs, Wiley-VCH, 2009). The measure extends naturally to directed graphs (digraphs) and can be defined for infinite graphs as well (Mowshowitz, Bull Math Biophys 30:225–240, 1968).This chapter focuses on the information content of digraphs and infinite graphs. In particular, the information content of digraph products and recursively defined infinite graphs is examined.

Keywords Digraphs • Entropy • Infinite graphs • Information content • Networks

MSC2000 Primary 68R10; Secondary 05C20, 05C25, 05C75, 94C15, 90B10.

1 Introduction

1.1 Overview

This chapter investigates the information content of directed and infinite graphs. The information content of a finite graph (directed or undirected) is a quantitative measure based on the symmetry structure of the graph. As explained in detail

A. Mowshowitz (✉)
Department of Computer Science, The City College of New York (CUNY),
138th Street at Convent Avenue, New York, NY 10031, USA
e-mail: abbe@cs.ccny.cuny.edu

M. Dehmer et al. (eds.), *Towards an Information Theory of Complex Networks: Statistical Methods and Applications*, DOI 10.1007/978-0-8176-4904-3_1,

below, the group of symmetries of a finite graph partitions the vertex set and thus induces a unique finite probability scheme. The entropy of this scheme is taken to be the information content of the graph. This "classical" notion differs from "graph entropy" introduced in [16].

Development of the concept of entropy applied to finite graphs is discussed in [17] and [20]. The application of entropy to graphs was introduced in the 1950s soon after the appearance of Shannon's famous paper on information theory. Entropy measurement has been used as a tool for characterizing molecules and chemical structures. For example, measures characterizing the structural complexity of chemical graphs have been developed and applied in [1, 3, 6]. Most of these measures are based on graph invariants that generate an equivalence relation on the vertices or edges of a graph. The resulting equivalence classes form a partition to which a finite probability scheme [14] can be associated in a natural way. The entropy of such a scheme provides a quantitative measure of structural complexity.

Various structural features of a graph have provided the basis for entropy measures. The earliest centered on the symmetries of a graph [21]. Other features, such as branching structure in molecular graphs, have been used to define entropy measures [8]. Measures associated with graphs representing atoms and molecules have been defined and applied to problems of discriminating chemical isomers and to classifying atomic and chemical structures [7, 9, 15]. Such measures have also been used for the analysis of biological networks [13]. Degree characteristics of a graph have been used as basis for an entropy-based measure of disorder in complex networks [23]. Interest in measuring the information content of graphs has also been kindled in recent years by the growing importance of computer and social networks in modern society [10, 24]. Relationships between graph entropy-based measures, expressed as inequalities, have been demonstrated in [11].

The notion of information content can be extended to infinite graphs. The approach adopted here is to consider an infinite graph as a sequence of finite graphs. Each of the finite graphs in the sequence has a well-defined information content, and if the corresponding sequence of information content values has an unambiguous limit, that limit is defined to be the information content of the given infinite graph.

In Sect. 2, we will look into the existence of directed graphs with prescribed information content and determine the information content of certain products of directed graphs. Section 3 will focus on infinite graphs, investigating the information content of some special classes of infinite graphs, and applying results from Sect. 2 to determine the information content of infinite graphs in general. Section 4 will examine some applications of the information measure to problems in network theory.

1.2 General Definitions

Definition 1. $G = (V, E), |V| < \infty, E \subseteq \binom{V}{2}$ is called a *finite undirected graph*. If $G = (V, E), |V| < \infty$, and $E \subseteq V \times V$, then G is called a *finite directed graph*.

Definition 2. A digraph $L_n = (V, E)$ is called a *(directed) path of length n* (≥ 1), if $V = \{v_0, v_1, \cdots, v_n\}$ and $E = \{(v_0, v_1), (v_1, v_2), \cdots, (v_{n-1}, v_n)\}$. The number of vertices in L_n is $n + 1$, one more than the number of edges.

Definition 3. A digraph $C_n = (V, E)$ is called a *(directed) cycle of length n* (≥ 2), if $V = \{v_1, v_2, \cdots, v_n\}$ and $E = \{(v_1, v_2), (v_2, v_3), \cdots, (v_n, v_1)\}$. C_n has the same number (n) of vertices and edges.

Definition 4. The *complete graph* K_n has n vertices and $\binom{n}{2}$ (undirected) edges.

See [12] for additional definitions of basic concepts in graph theory.

2 Entropy of Digraphs

2.1 Definition and Examples

The automorphism group of a digraph and the measure of information content based on the group are defined below.

Definition 5. Let $G = (V, E)$ be a (directed or undirected) graph with vertex set V (with $|V| = n$), and edge set E. The *automorphism group* of G, denoted by $Aut(G)$, is the set of all adjacency preserving bijections of V.

Definition 6. Let $\{V_i | 1 \leq i \leq k\}$ be the collection of orbits of $Aut(G)$ and suppose $|V_i| = n_i \; for \; 1 \leq i \leq k$. The *entropy* or *information content* of G is given by the following formula [17]:

$$I_a(G) = -\sum_{i=1}^{k} \frac{n_i}{n} \log\left(\frac{n_i}{n}\right).$$

Figure 1.1 illustrates the computation of the information content of a digraph.

2.2 Entropy of Digraph Products

Many different binary operations on graphs and digraphs appear in the literature [19]. We will examine four such operations in some detail, namely, the sum, join, Cartesian product, and the composition. Our aim is to determine the information content of a digraph operation in relation to the information contents of the respective digraphs in the operation. Such products are useful in defining classes of digraphs with properties of interest in different applications, especially those pertaining to the analysis of networks.

Definition 7. The *sum* of G_1 and G_2 is the digraph $G_1 \cup G_2$ defined by $V(G_1 \cup G_2) = V(G_1) \cup V(G_2)$ and $E(G_1 \cup G_2) = E(G_1) \cup E(G_2)$.

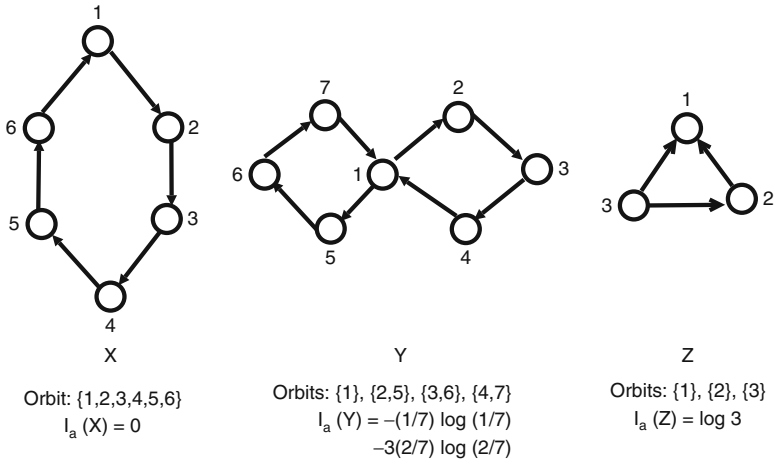

Fig. 1.1 Computation of information content

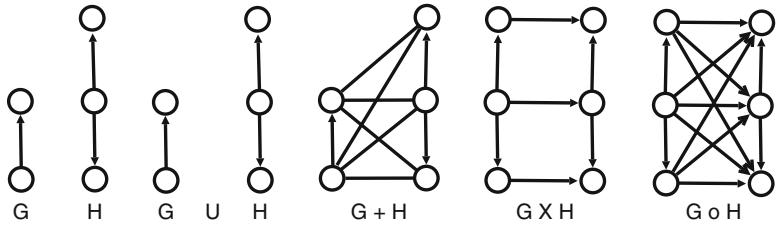

Fig. 1.2 Binary operations on digraphs

Definition 8. The *join* of G_1 and G_2 is the digraph $G_1 + G_2$ defined by $V(G_1 + G_2) = V(G_1) \cup V(G_2)$ and $E(G_1 + G_2) = E(G_1) \cup E(G_2) \cup \{[u, v] | u \in V(G_1), v \in V(G_2)\}$ where $[u, v]$ denotes the undirected edge joining u and v.

Definition 9. The *Cartesian product* of G_1 and G_2 is the digraph $G_1 \times G_2$ given by $V(G_1 \times G_2) = V(G_1) \times V(G_2)$ and $E(G_1 \times G_2) = \{(u, v) = ((u_1, u_2), (v_1, v_2)) | u_1, v_1 \in V(G_1), u_2, v_2 \in V(G_2)$, and either $u_1 = v_1$ and $(u_2, v_2) \in E(G_2)$ or $u_2 = v_2$ and $(u_1, v_1) \in E(G_1)\}$

Definition 10. Two digraphs G and H are *relatively prime with respect to the Cartesian product* if whenever G is isomorphic to $G' \times D$ and H is isomorphic to $H' \times D$, then D is the identity digraph K_1.

Definition 11. The *composition* of G_1 and G_2 is the digraph $G_1 \circ G_2$ given by $V(G_1 \circ G_2) = V(G_1) \times V(G_2)$ and $E(G_1 \circ G_2) = \{(u, v) = ((u_1, u_2), (v_1, v_2)) | u_1, v_1 \in V(G_1), u_2, v_2 \in V(G_2)$, and either $(u_1, v_1) \in E(G_1)$ or $u_1 = v_1$ and $(u_2, v_2) \in E(G_2)\}$

The foregoing operations are illustrated in Fig. 1.2.

These binary operations will be discussed with a view to characterizing the information content of digraphs resulting from their application.

2.3 Sum and Join

Theorem 1. *Let G and H be digraphs.*

(a) Suppose $Aut(G)$ has orbits V_i^G with $|V_i^G| = m_i$ for $1 \leq i \leq m$, and $Aut(H)$ has orbits V_i^H with $|V_i^H| = n_i$ for $1 \leq i \leq n$. If no component of G is isomorphic to a component of H, then

$$\begin{aligned} I_a(G \cup H) &= I_a(G + H) \\ &= \log(n+m) + \frac{1}{n+m}[nI_a(G) + mI_a(H) \\ &\qquad - n\log(n) - m\log(m)]. \end{aligned}$$

(b) If G and H are isomorphic, then $I_a(G \cup H) = I_a(G + H) = I_a(G)$. More generally, if each $G_i (1 \leq i \leq n)$ is isomorphic to G, then

$$I_a(G_1 \cup G_2 \cdots \cup G_n) = I_a(G_1 + G_2 \cdots + G_n) = I_a(G).$$

Proof. $I_a(G \cup H) = I_a(G + H)$ since the orbits of $Aut(G \cup H)$ are the same as those of $Aut(G + H)$. This is a consequence of the fact that every vertex of G is adjacent to every vertex of H in $G + H$. (a) $I_a(G \cup H) = I_a(G + H) = -\sum_{i=1}^{k} \frac{n_i}{n+m} \log(\frac{n_i}{n+m}) - \sum_{i=1}^{k} \frac{m_i}{n+m} \log(\frac{m_i}{n+m}) = \frac{1}{n+m}[\sum_{i=1}^{k} n_i \log(n+m) + \sum_{i=1}^{k} m_i \log(n+m)] + \frac{1}{n+m}[\sum_{i=1}^{k} n_i \log(n_i) - \sum_{i=1}^{k} m_i \log(m_i)] = \log(n+m) + \frac{1}{n+m}[nI_a(G) + mI_a(H) - n\log(n) - m\log(m)]$, as required. (b) See [18]. □

When the two digraphs are of equal size, the information content of their join is just one more than their average information content.

Corollary 1. *Let G and H be as in the Theorem. If $n = m$, then $I_a(G \cup H) = I_a(G + H) = \frac{1}{2}[I_a(G) + I_a(H)] + 1$.*

Proof. The result follows immediately from the Theorem by setting $m = n$ in the expression for $I_a(G \cup H) = I_a(G + H)$. □

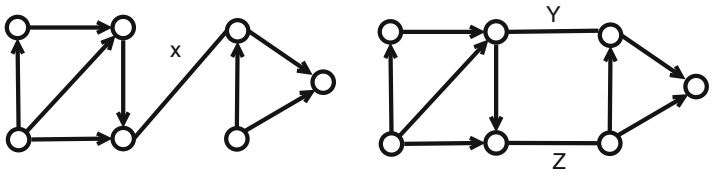

Fig. 1.3 Partial joins

Corollary 2. *Let G and H be as in the Theorem and suppose $I_a(G) = I_a(H)$. Then (i) $I_a(G \cup H) = I_a(G + H) = I_a(G) + \log(n + m) - \frac{1}{n+m}[n \log(n) + m \log(m)]$, and (ii) if in addition $n = m$, $I_a(G \cup H) = I_a(G + H) = I_a(G) = I_a(G) + 1$.*

Of particular importance to the representation of real network growth is the partial join operation.

Definition 12. A *partial join* of G_1 and G_2 for the set F is the digraph $G_1 \oplus G_2$ defined by $V(G_1 \oplus G_2) = V(G_1) \cup V(G_2)$ and $E(G_1 \oplus G_2) = E(G_1) \cup E(G_2) \cup F$, where $F \subset \{[u, v] | u \in V(G_1), v \in V(G_2)\}$.

Figure 1.3 illustrates partial join operations for different sets F.

The information content of a partial join depends on the set F. For example, if both graphs G and H are isomorphic to the directed cycle of length n and $G \oplus H$ is defined for set F consisting of a single undirected edge, $I_a(G \oplus H) = \log(n)$ since each orbit of $Aut(G \oplus H)$ consists of two of the $2n$ vertices. If there are two edges joining G and H, one of which does not join corresponding vertices of the directed n-cycles, the information content is $\log(2n)$ since $Aut(G \oplus H)$ is the trivial group in this case. Note that $I_a(G) = I_a(H) = 0$ since the automorphism group of a directed cycle with n vertices is the cyclic group of order n. Thus, it appears that $I_a(G \oplus H)$ can be expressed in terms of $I_a(G)$ and $I_a(H)$ in special cases only.

Theorem 2. *Let G_1 and G_2 be complete graphs with m and n vertices, respectively, and suppose $G = G_1 \oplus G_2$ is a partial join with $|F| = 1$.*

(a) If m is different from n, $I_a(G) = \frac{m-1}{m+n} \log\left(\frac{m-1}{m+n}\right) + \frac{n-1}{m+n} \log\left(\frac{m-1}{m+n}\right) + \frac{2}{m+n} \log(m + n)$

(b) If $m = n$, then

$$I_a(G) = \frac{1}{n}\left[(n-1)\log\left(\frac{n}{n-1}\right) + \log(n)\right].$$

Proof. Let $[x, y]$ be the edge in F where x is in G_1 and y is in G_2. If m is different from n, the partial join G has four orbits A, B, C, and D, where A consists of the $m - 1$ vertices of G_1 excluding x, B consists of the $n - 1$ vertices of G_2 excluding y, and C and D are singletons containing x and y, respectively. If $m = n$ there are two orbits with 2 and $2(n - 1)$ vertices, respectively. □

2.4 Cartesian Product and Composition

Theorem 3 ([18]). *(a)* $I_a(G \times H) \leq Ia(G) + Ia(H)$ *for any digraphs* G *and* H.
(b) Equality holds when G *and* H *are weakly connected and relatively prime with respect to the Cartesian product.*

Proof. Part (a) follows from the fact that $Aut(G \times H)$ is a subgroup $Aut(G) \times Aut(H)$. Part (b) is a consequence of the fact that $Aut(G \times H)$ is isomorphic to $Aut(G) \times Aut(H)$ if and only if digraphs G and H are relatively prime with respect to the Cartesian product. Note that being relatively prime is a sufficient but not a necessary condition for equality in the theorem. □

The information content measure is also sub-additive for the composition operation.

Theorem 4 ([18]). $I_a(G \times H) \leq Ia(G) + Ia(H)$ *for any digraphs* G *and* H.

Figure 1.4 provides examples of the information content of the Cartesian product and composition.

2.5 Existence Theorem

The join and Cartesian product can be used to construct digraphs with given information content. More precisely, for any finite probability scheme there exists a digraph with information content equal to the entropy of the scheme. This result is stated in the following theorem originally presented in [18].

Theorem 5. *Let* n *be any positive integer, and suppose* $\mathrm{P} = \{n_{ij}\}$ *is a partition of* n *where* $n_{ij} = n_i$ $(1 \leq j \leq r_i)$, $n_{i_1} \neq n_{i_2}$ $(i_1 \neq i_2)$, *and* $i = 1, 2, \cdots, k$. *Then there exists a weakly connected digraph* G *with* n *vertices such that* $Aut(G)$ *has exactly* $r = \sum_{i=1}^{k} r_i$ *orbits, and for each* n_{ij} *there is an orbit* A *with* $|A| = n_{ij}$; *and, hence,*

$$I_a(G) = H(P) = -\sum_{i=1}^{k} r_i \frac{n_i}{n} \log\left(\frac{n_i}{n}\right).$$

Proof. The proof is based on a simple construction. Let $G_i = L_{r_i - 1} \times C_{n_i}$ where L_{r_i-1} is a directed path of length $r_i - 1$ and C_{n_i} is a directed cycle of length n_i. Since the path and cycle are relatively prime with respect to the Cartesian product, the orbits of $Aut(G_i)$ are the respective products of the orbits of $Aut(L_{r_i-1})$ and $Aut(C_{n_i})$. Hence, $Aut(G_i)$ has exactly r_i orbits, each consisting of n_i elements. The digraph G formed by taking the join of the k non-isomorphic G_i has an automorphism group with orbits corresponding to the partition specified in the hypothesis of the theorem, and thus has the required information content. □

Figure 1.5 illustrates the Theorem for $n = 25$, $P = \{1^3, 2^4, 3^2, 4^2\}$.

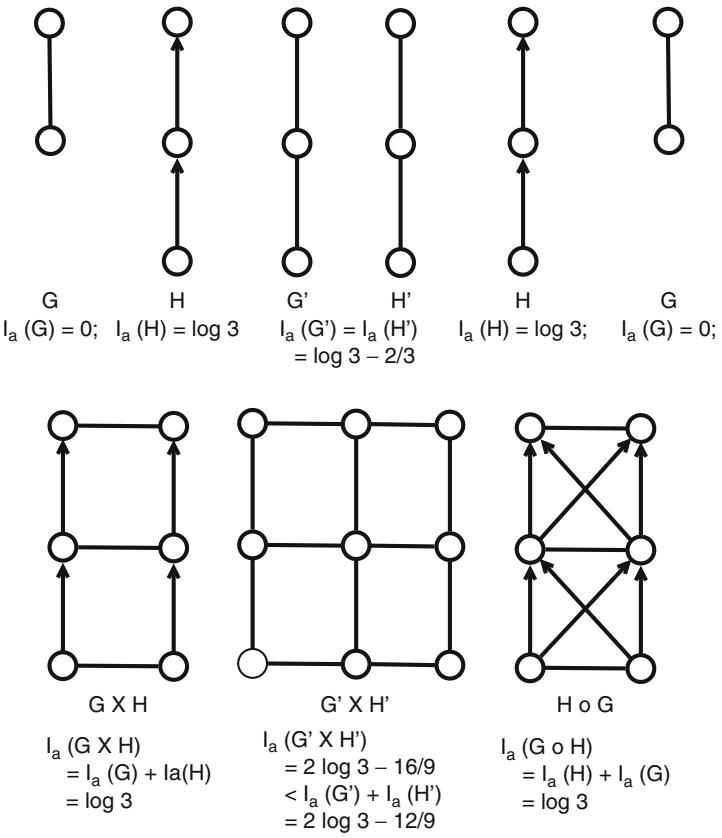

Fig. 1.4 Information content of Cartesian product and composition

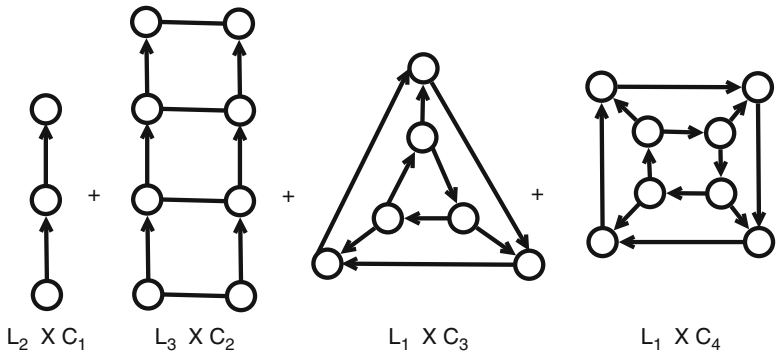

Fig. 1.5 Construction of digraph with prescribed information content

3 Entropy of Infinite Graphs

3.1 Preliminaries

Definition 13. A graph $G = (V, E)$ is *countable* if $|V \cup E|$ is countable. G is *locally finite* if the degree of every vertex of G is finite.

In what follows, we will restrict attention to countable graphs that may or may not be locally finite.

Definition 14 ([18]). Let $G = (V, E)$ be a countable graph. A sequence $\{G_n\}_{n=1}^{\infty}$ of finite graphs G_n with $V_n = V(G_n)$ and $E_n = E(G_n)$ is said to *converge* to G as a *limit* (written $\lim_{n\to\infty} G_n = G$) if $\lim_{n\to\infty} V_n = V(G)$ and $\lim_{n\to\infty} E_n = E(G)$. Note that both V and E are simply the limits of sequences of sets.

Definition 15 ([18]). A sequence $\{G_n\}_{n=1}^{\infty}$ of finite graphs G_n is a *defining sequence* for a countable graph G if $G_n \subset G_{n+1}$ for every n, and $\lim_{n\to\infty} G_n = G$. Since the limit of any monotonically increasing sequence $\{A_n\}_{n=1}^{\infty}$ of sets A_n exists and is equal to $\bigcup_{n=1}^{\infty} A_n$, every countable graph G has a defining sequence.

A defining sequence for a countable graph G with $V(G) = \{v_1, v_2, v_3, \cdots\}$ can be constructed as follows:

$V(G_1 = \{v_1\}$ and $E(G_1) = \emptyset$,
$V(G_{n+1}) = V(G_n) \cup \{v_{n+1}\}$ and $E(G_{n+1}) = E(G_n) \cup \{[v_{n+1}, u] \in E(G) | u \in V(G_n)\}$.

Definition 16 ([18]). Let $\{G_n\}_{n=1}^{\infty}$ be a defining sequence for a countable graph G. The *information content* $\hat{I}(G; G_n)$ of G *with respect to the sequence* $\{G_n\}_{n=1}^{\infty}$ is given by $\hat{I}(G; G_n)) = \lim_{n\to\infty} I_a(G_n)$ if the limit exists.

Figure 1.6 shows a countable graph with defining sequences that give rise to different information content values.

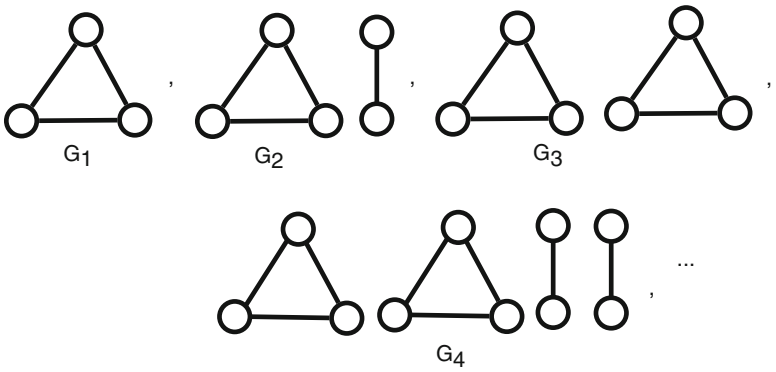

Fig. 1.6 A countable graph with more than one defining sequence

$\mathrm{i}_a(G_n) = \begin{cases} 0 & \text{if } n \text{ is odd} \\ \log(5) - \frac{3}{5}\log(3) - \frac{2}{5} & \text{if } n \text{ is even} \end{cases}$ Thus, for the subsequence S_n consisting of the odd terms, $\hat{I}(G; S_n) = 0$; and for the subsequence T_n consisting of the even terms, $\hat{I}(G; T_n) = \log(5) - \frac{3}{5}\log(3) - \frac{2}{5}$. The difference in this case is finite, but it could be infinite as shown in [18]. Using a measure that depends on the graph's defining sequence is not necessarily a disadvantage. An infinite graph can be viewed as an idealization of a growth process. Including the defining sequence in the definition allows for capturing different principles of growth in practice.

3.2 Classes of Infinite Graphs

Infinite graphs can be built up recursively with the aid of graph products. The following result makes use of the Cartesian product.

Lemma 1. *Let G be a graph with n vertices. $I_a(G \times K_2) = I_a(G)$.*

Proof. Corresponding vertices of the two copies of G are in the same orbit of $G \times K_2$, so G and $G \times K_2$ have the same number of orbits, and each orbit of $G \times K_2$ has exactly double the number of vertices as the corresponding orbit of G. Thus, if $Aut(G)$ has orbits $A_1, A_2, \cdots, A_r$ with $|A(_i)| = k_i, 1 \leq i \leq r$, then $I_a(G \times K_2) = -\sum_{i=1}^{r} \frac{2k_i}{2n} \log\left(\frac{2k_i}{2n}\right) = I_a G$. □

Suppose G is a graph with n vertices. If $Aut(G)$ is the identity group, then $I_a(G) = \log(n)$, and $I_a(G \times K_2) = \log(n)$. The sequence

$$H_1 = G,$$

$$H_{n+1} = H_n \times K_2, \text{ for } n \geq 1$$

serves as a defining sequence of an infinite graph. Since $I_a(H_n) = \log(n)$, $\lim_{n\to\infty} I_a(H_n) = \infty$.

At the other extreme is the *hypercube* H_n, which can be defined recursively as follows:

$$H_1 = K_2,$$

$$H_{n+1} = H_n \times K_2, \text{ for } n \geq 1.$$

Since the limit of the (defining) sequence $\{H_n\}_{n=1}^{\infty}$ exists, we can set $H_\infty = \lim_{n\to\infty} H_n$. Now, $I_a(H_n) = 0$ for all $n \geq 1$ which implies by the lemma that $\hat{I}(H; H_n) = 0$, i.e., the sequence of finite hypercubes yields a limit whose information content is zero. The hypercube serves as a useful model in parallel computation. A key feature in this context is the favorable maximum distance between any two vertices in the graph. This allows for placing computational units so as to minimize communication costs. The zero information content of the

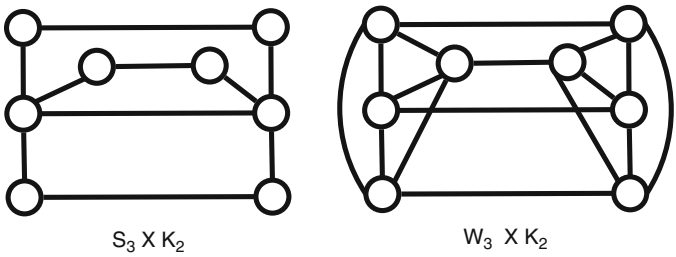

Fig. 1.7 Star and wheel products

hypercube reflects the high degree of symmetry of this graph, which allows for simultaneous placement of elements at optimal distance from each other.

Other graphs of interest, with information content between the two extremes, can be substituted for G in $G \times K_2$.

Let S^k denote the *star of order k*, a connected graph with one vertex of degree $k-1$ and $k-1$ vertices of degree 1; and let W^k denote the *wheel of order k*, a connected graph obtained from the star by joining the degree 1 vertices in a cycle of length $k-1$. Once again using the Cartesian product, we can build infinite sequences based on these simple graphs.

$$S_1^k = S^k,$$
$$S_{n+1}^k = S_n^k \times K_2, \text{ for } n \geq 1.$$

A sequence of graphs W_n^k can be defined similarly.

$I_a(S^k) = I_a(W^k) = -\frac{k-1}{k}\log\left(\frac{k-1}{k}\right) - \frac{1}{k}\left(\log\frac{1}{k}\right) = \log(k) - \frac{k-1}{k}\log(k-1)$. Denoting by S_∞ and W_∞, respectively, the infinite graphs with defining sequences $\{S_n^k\}_{n=1}^\infty$ and $\{W_n^k\}_{n=1}^\infty$, we have $\hat{I}(S_\infty; S_n^k) = \hat{I}(W_\infty; W_n^k) = \log(k) - \frac{k-1}{k}\log(k-1)$.

As k increases, almost all the vertices fall into one orbit and the information content tends to zero.

Figure 1.7 shows the Cartesian products, respectively, of the star and the wheel with K_2.

The information content of the line graph of order k is given by:

$$I_a(L^k) = \begin{cases} \log\left(\frac{k}{2}\right) & \text{if } n \text{ is even} \\ \frac{k-1}{k}\log\left(\frac{k}{2} + \frac{1}{k}\log k\right) & \text{if } n \text{ is odd.} \end{cases}$$

The information content of the line graph increases without bound, so the information content of the limit graph is infinite.

The cycle graph of order k has information content $I_a(C^k) = 0$, so the limit graph in this case has information content zero.

More complex graphs could be constructed by substituting for K_2 in the Cartesian products defining the terms in the infinite sequences considered above.

4 Applications

Preferential attachment has been studied extensively as a protocol for the growth of large-scale networks like the Internet [5]. According to this protocol, a vertex added to a network will be more likely to become attached to existing vertices of higher rather than of lower degree. The "preference" of a vertex v as a target of attachment might be expressed as the probability given by the degree of v divided by the sum of the degrees in the graph. This introduces a random element in the growth process. Perhaps the simplest way to realize a (relatively deterministic) version of growth by preferential attachment is to add a single new vertex at each iteration, connecting the new vertex to an existing one whose degree is maximal in the current graph. Call this a *type-0 preferential attachment protocol*. If the starting graph is K_1, the result is clearly a star. After the nth new vertex has been added, a star of order $n+1$ has been formed. This graph S^{n+1} has information content $\log(n+1) - \frac{n}{n+1}\log(n)$, and as noted above, this value tends to zero as n increases without bound.

A variation on this simple protocol is to add k new vertices at each iteration and attach each one of them to a different existing vertex, choosing the existing vertices in nonincreasing order of degree, beginning with one of maximal degree.

Figure 1.8 illustrates the construction process according to this protocol, and the following theorem gives the information content in the case where k equals the number of vertices in the initial graph of the sequence.

Theorem 6. *Let $\{G_n^k\}_1^\infty$ be a sequence of graphs defined as follows:*

$$G_1^k = S^k$$

G_{n+1}^k is obtained from G_n^k by adding k new vertices and joining each one to a different vertex of maximal degree in G_n^k.

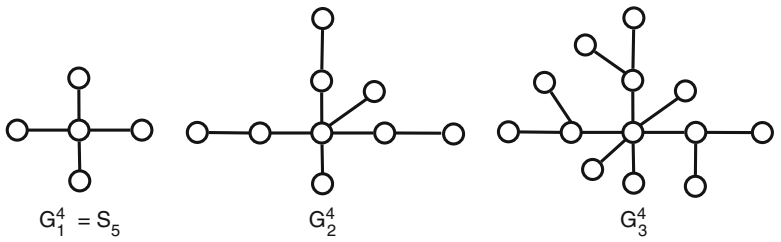

Fig. 1.8 A graph constructed with preferential attachment protocol type-0

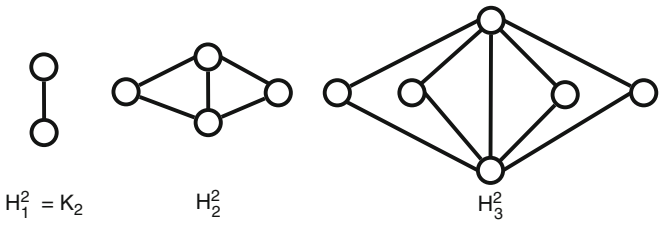

Fig. 1.9 A graph constructed with preferential attachment protocol type-1

The information content of G_n^k is given by:

$$I_a(G_n^k) = \frac{(k-1)(n-1)}{nk+1}\log\left(\frac{nk+1}{(k-1)(n-1)}\right) + \frac{k}{nk+1}\log\left(\frac{nk+1}{k}\right) + \frac{n-1}{nk+1}\log\left(\frac{nk+1}{n-1}\right) + \frac{1}{nk+1}\log(nk+1).$$

Proof. Since k vertices are added for each iteration, G_n^k, the nth graph in the sequence has $nk+1$ vertices. Let v be the vertex of highest degree in G_n^k. The orbits of $Aut(G_n^k)$ consist of the vertex v alone, the vertices of degree 1 adjacent to v, the vertices of degree > 1 adjacent to v, and the vertices of degree 1 at distance 2 from v. Thus, the orbits of $Aut(G_n^k)$ have 1, $n-1$, k, and $(k-1)(n-1)$ vertices from which the result follows. □

Corollary 3. *Let G_n^k be defined as in the theorem. Then $\hat{I}(G; G_n^k) = \log k$.*

Proof. Simplifying the expression in the theorem gives $I_a(G_n^k) = \log(nk+1) - \frac{1}{nk+1}[(k-1)(n-1)\log(k-1)(n-1) + (n-1)\log(n-1) + k\log k]$. Taking the limit as $n \to \infty$ yields $\hat{I}(G; G_n^k) = \log k$ as required. □

Greater connectivity in a network that grows by preferential attachment can be achieved by allowing the newly added vertices to be joined to more than one existing vertex [2]. Call this a *type-1 preferential attachment protocol*. This protocol is illustrated in Fig. 1.9. The following theorem gives the information content of an infinite graph that grows according to a type-1 protocol with $k = 2$.

Theorem 7. *Let $\{H_n^2\}_1^\infty$ be a sequence of graphs defined as follows:*

$$H_1^2 = K_2$$

H_{n+1}^2 is obtained from H_n^2 by adding 2 new vertices and joining each one to exactly two different vertex of maximal degree in H_n^2.