The IMA Volumes in Mathematics and its Applications

Andrew Beveridge
Jerrold R. Griggs
Leslie Hogben
Gregg Musiker
Prasad Tetali *Editors*

Recent Trends in Combinatorics

Springer

The IMA Volumes in Mathematics and its Applications

Volume 159

More information about this series at http://www.springer.com/series/811

Institute for Mathematics and its Applications (IMA)

The Institute for Mathematics and its Applications (IMA) was established in 1982 as a result of a National Science Foundation competition. The mission of the IMA is to connect scientists, engineers, and mathematicians in order to address scientific and technological challenges in a collaborative, engaging environment, developing transformative, new mathematics and exploring its applications, while training the next generation of researchers and educators. To this end the IMA organizes a wide variety of programs, ranging from short intense workshops in areas of exceptional interest and opportunity to extensive thematic programs lasting nine months. The IMA Volumes are used to disseminate results of these programs to the broader scientific community.

The full list of IMA books can be found at the Web site of the Institute for Mathematics and its Applications:

http://www.ima.umn.edu/springer/volumes.html.

Presentation materials from the IMA talks are available at

http://www.ima.umn.edu/talks/.

Video library is at

http://www.ima.umn.edu/videos/.

Fadil Santosa, Director of the IMA

Andrew Beveridge • Jerrold R. Griggs
Leslie Hogben • Gregg Musiker • Prasad Tetali
Editors

Recent Trends in Combinatorics

Springer

Editors

Andrew Beveridge
Department of Mathematics, Statistics
and Computer Science
Macalester College
St. Paul, MN, USA

Leslie Hogben
Department of Mathematics
Iowa State University
Ames, IA, USA

Prasad Tetali
School of Mathematics
and Computer Science
Georgia Institute of Technology
Atlanta, GA, USA

Jerrold R. Griggs
Department of Mathematics
University of South Carolina
Columbia, SC, USA

Gregg Musiker
School of Mathematics
University of Minnesota
Minneapolis, MN, USA

ISSN 0940-6573 ISSN 2198-3224 (electronic)
The IMA Volumes in Mathematics and its Applications
ISBN 978-3-319-24296-5 ISBN 978-3-319-24298-9 (eBook)
DOI 10.1007/978-3-319-24298-9

Library of Congress Control Number: 2016932007

Mathematics Subject Classification (2010): 05

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

Foreword

This volume is based on the research focus at the IMA during the Fall semester of 2014. The Annual Thematic Program covering this period was “Discrete Structures: Analysis and Applications”. The program was organized by Sergey Bobkov, Jerrold Griggs, Penny Haxell, Michel Ledoux, Benny Sudakov, and Prasad Tetali. Many of the topics presented in this volume were discussed in the first three workshops that took place during the year. We thank the organizers of the workshops, the speakers, workshop participants, and visitors to the IMA who contributed to the scientific life at the institute and to the successful program. In particular, we thank Andrew Beveridge, Jerrold Griggs, Leslie Hogben, Gregg Musiker, and Prasad Tetali for taking the lead to edit this volume. We also thank the National Science Foundation for its support of the IMA.

Minneapolis, MN, USA

Fadil Santosa

Preface

Combinatorics is a research field driven by collaboration, with a large number of applications to different areas of pure and applied mathematics. The Institute for Mathematics and Its Applications (IMA) is an ideal setting for such collaborations and applications to develop.

The 2014–2015 Annual Thematic Program at the IMA was *Discrete Structures: Analysis and Applications*. The program was organized by Sergey Bobkov (University of Minnesota), Jerrold Griggs (University of South Carolina), Penny Haxell (University of Waterloo), Michel Ledoux (Paul Sabatier University of Toulouse), Benny Sudakov (University of California, Los Angeles), and Prasad Tetali (Georgia Institute of Technology).

Combinatorics was the focus during Fall 2014, and this volume presents some of the research topics discussed during this intense semester. We have particularly encouraged authors to write surveys of research problems, thus making state-of-the-art results more conveniently and widely available.

This volume is organized into parts, following the themes of the three workshops held during Fall 2014:

- *Probabilistic and Extremal Combinatorics*, held September 8–12, 2014, at IMA and organized by Penny Haxell (University of Waterloo), Eyal Lubetzky (Microsoft Research), Dhruv Mubayi (University of Illinois, Chicago), and Benny Sudakov (Eidgenössische TH Zürich-Zentrum).
- *Additive and Analytic Combinatorics*, held September 29–October 3, 2014, at IMA and organized by David Conlon (University of Oxford), Ernie Croot (Georgia Institute of Technology), Van Vu (Yale University), and Tamar Ziegler (Hebrew University).
- *Geometric and Enumerative Combinatorics*, held November 10–14, 2014, at IMA and organized by Zoltan Furedi (Hungarian Academy of Sciences MTA), Jerrold Griggs (University of South Carolina), Victor Reiner (University of Minnesota, Twin Cities), and Carla Savage (North Carolina State University).

Part 1: Extremal and probabilistic combinatorics. Extremal and probabilistic combinatorics are central to modern combinatorial theory, and both have developed dramatically over the last few decades. Extremal combinatorics studies problems of finding the maximum or minimum possible cardinality of a set of finite objects satisfying certain requirements. Frequently such problems originate in other areas, such as computer science, information theory, analysis, number theory, and geometry. Probabilistic combinatorics, as the name suggests, blends combinatorics and probability. It is the foundation of the study of random graphs and other random discrete structures, and probabilistic arguments have been very powerfully applied to problems in other areas of combinatorics and in theoretical computer science. Major research topics in extremal and probabilistic combinatorics include extremal problems for graphs and set systems, Ramsey theory, random graphs, and application of probabilistic methods.

Part 2: Additive and analytic combinatorics. Additive combinatorics counts additive structures in sets; there have been exciting developments in recent years. Tools from Fourier and harmonic analysis have expanded the realm of additive combinatorics into the analytic while contributing to more effective applications. Many combinatorial ideas known to the combinatorics community can be used effectively to attack difficult problems in other areas of mathematics. For example, a famous theorem of Szemerédi on arithmetic progressions in dense sets is a key tool for the proof of the Green-Tao theorem on the existence of long arithmetic progressions in primes. The work of Breuillard, Green, and Tao, which established an analogue of the Freiman inverse theorem for noncommutative groups, is another example. This theorem was first stated and proved for integers by Freiman. Ruzsa's subsequent different proof was extended to abelian groups by Green and Ruzsa a few years ago. The extension to noncommutative groups is much more difficult. Research in additive and analytic combinatorics is also of interest to computer scientists; techniques and results have been applied to communication complexity, property testing, and the design of randomness extractors.

Part 3: Enumerative and geometric combinatorics. Geometric combinatorics studies discrete objects with geometric or topological structure, such as convex polytopes, arrangements of vectors, points, subspaces, triangulations, tilings, and partially ordered sets. Enumerative combinatorics, often called the mathematics of counting, has broad applications to probability, statistical physics, optimization, and computer science. Problems in geometric combinatorics give rise to counting problems that are sometimes difficult even to estimate and sometimes involve objects with interesting symmetry groups. Such problems often dovetail nicely with topics from enumerative combinatorics via calculations of partition functions, f-vectors, Ehrhart polynomials, and other quantities. Enumerative combinatorics also includes the study of permutation patterns, the complexity of tilings, and bijections between families of objects counted by the same numerical sequences or with related generating functions. In recent years, problems in both of these areas have stimulated the development of many new results and tools and enhanced connections with other areas of mathematics.

Discrete Structures: Analysis and Applications attracted intense interest from the mathematical sciences community, with each of the three workshops drawing more than 100 visitors and often filling Keller 3-180 to capacity. There are many other aspects to an annual thematic year at the IMA besides workshops, with the relaxed but stimulating environment of the IMA fostering new collaborations and approaches to solving problems old and new. This program drew an eclectic mix of experts and junior researchers in various aspects of combinatorics together with numerous people who apply combinatorics to other fields. This volume reflects many of the aspects of the semester, with chapters drawn from workshop talks, annual program seminars, and research interests of the many visitors.

No single volume could possibly cover all the active and important areas of combinatorics research that were presented at the IMA, and we make no claim of comprehensiveness. But we think this volume presents a reasonable selection of interesting areas, written by leading experts who have surveyed the current state of knowledge and posed conjectures and open questions to stimulate further research. We thank the authors for their generous donations of time and expertise; needless to say, without them this volume would not have been possible.

We thank the IMA for wonderfully stimulating and productive long-term visits. We believe that the IMA is a critical national resource for mathematics. The *Discrete Structures: Analysis and Applications* program will have a lasting impact on research in combinatorics and related fields, and we hope this volume will enhance that impact. We are grateful for the opportunity to be part of it.

St. Paul, MN, USA	Andrew Beveridge
Columbia, SC, USA	Jerrold R. Griggs
Ames, IA, USA	Leslie Hogben
Minneapolis, MN, USA	Gregg Musiker
Atlanta, GA, USA	Prasad Tetali

Contents

Part I Extremal and probabilistic combinatorics

Part II Additive and analytic combinatorics

Part III Enumerative and geometric combinatorics

Part I
Extremal and probabilistic combinatorics

Problems related to graph indices in trees

László Székely, Stephan Wagner, and Hua Wang

Abstract In this chapter we explore recent development on various problems related to graph indices in trees. We focus on indices based on distances between vertices, vertex degrees, or on counting vertex or edge subsets of different kinds. Some of the indices arise naturally in applications, e.g., in chemistry, statistical physics, bioinformatics, and other fields, and connections are also made to other branches of graph theory, such as spectral graph theory. We will be particularly interested in the extremal values (maxima and minima) for different families of trees and the corresponding extremal trees. Moreover, we review results for random trees, consider localized versions of different graph indices and the associated notions of centrality, and finally discuss inverse problems, where one wants to find trees for which a specific graph index has a prescribed value.

Keywords Wiener index • Randic index • Merrifield-Simmons index • Hosoya index • Random tree • Inverse problems • Dominating set • Number of subtrees

Mathematics Subject Classification (2000): Primary: 05C05, Secondary 05C12, 05C31, 05C35, 05C69, 05C70, 05C80, 92E10

1 Introduction

Enumeration of trees [27, 116] and of spanning trees [80] goes back to the XIX$^{\text{th}}$ century. The concepts of the center and centroid are also that old [77]. However, the topic of this survey really starts with the papers of Wiener [153, 154]. He noted

L. Székely (✉)
Department of Mathematics, University of South Carolina, Columbia, SC, USA
e-mail: laszlo@mailbox.sc.edu; szekely@math.sc.edu

S. Wagner
Department of Mathematical Sciences, Stellenbosch University, Stellenbosch, South Africa
e-mail: swagner@sun.ac.za

H. Wang
Department of Mathematical Sciences, Georgia Southern University, Statesboro, GA, USA
e-mail: hwang@georgiasouthern.edu

A. Beveridge et al. (eds.), *Recent Trends in Combinatorics*, The IMA Volumes
in Mathematics and its Applications 159, DOI 10.1007/978-3-319-24298-9_1

that the boiling temperatures of alkanes correlate with the sum of the lengths of the shortest paths between all pairs of vertices in the chemical graph representing the non-hydrogen atoms in the molecule. This quantity (now called the Wiener index), and its variants will be discussed in Section 2.

The Wiener index started chemical graph theory. Graphs arise as representation of atoms with vertices, bonds with edges, usually suppressing hydrogen atoms. (Such representation appeared already in [26], as pointed out in [12].) The point of chemical graph theory is to come up with graph invariants, which have predictive power for chemical properties of the molecule, if computed for the molecular graph. In chemical graph theory such invariants are called (topological) indices, as the expectation is that the shape of the molecule is the ultimate source of information. The discriminating power of an index is high, if different tree shapes tend to show distinct indices. The range of an index limits its discriminating power. A chemical graph (resp. tree) is understood to have maximum degree four, reflecting the valence of a carbon atom. Several books are fully or partly devoted to chemical graph theory [8, 132, 133]. Our survey is focused on indices of trees although many of these indices have been investigated on graphs as well. This survey does not even try to be complete, as the number of indices is exploding: one of the main organizing principles of selection is the interest of the authors.

Randić [111] introduced a very influential index, which is now named after him. Earlier it was called the branching index or connectivity index. The Randić index, which is the prototype of degree-based indices, and its variants will be reviewed in Section 4. The survey papers [58] and [94] are particularly informative about degree-based indices. There are even three books devoted to the Randić index [78, 79, 95]. Unexpectedly, the generalized Randić index turned useful in an entirely different context: to find an analogue of the Crossing Lemma to set lower bounds for the minor crossing number of graphs [14].

The Merrifield-Simmons index of a graph is the number of its independent vertex sets [103] and the Hosoya index (also called topological Z index) of a graph is the number of its independent edge sets, i.e., matchings without size restriction [66]. These quantities have been relevant in statistical physics in the hard square model [10] and the monomer-dimer model [64]. The number of subtrees of a tree came from bioinformatics [85], where the number of subtrees of a binary tree containing at least one leaf was involved in the complexity of an algorithm. The maximum of the number of subtrees among binary trees was determined in [125], and the systematic study of this quantity determined these maximum values exactly using a novel number representation [124]. The original problem of determining the maximum number of subtrees of a binary tree containing at least one leaf was also solved [123]. It turned out that for several classes of trees, those trees that minimize the Wiener index happen to maximize the number of subtrees and vice versa. A similar phenomenon is present for the Merrifield-Simmons and Hosoya indices, see Section 3. The paper [136] contains an analysis of the correlation between pairs of tree indices, and found the highest (negative) correlation between the Wiener index and the number of subtrees. An intriguing problem would be to give an "explanation" for the negative correlation between the number of subtrees and the

Wiener index—although we admit that we cannot tell criteria for a “satisfactory” explanation.

From a practical point of view, identifying the extremal structures alone is far from being sufficient for understanding the behavior of an index. A natural and important question appears to be the following.

> For a given index, what is the distribution of its value over all possible trees under given restriction (i.e., of given order, with given degree sequence, etc..)?

The answer to this question more specifically presents the behavior of a graph index and hence can be used to further examine the similarities or differences between different indices.

> How similar or different are two indices according to the distributions of their values over different categories of trees?

Random trees are a huge topic of their own right, there is a whole book [43] devoted to them. In Section 5 we focus on only a few questions about expected values and distributions of tree indices that we consider in this survey.

There are several centrality concepts for trees. Two of them, the center and the centroid, are age-old [77]. Section 6 investigates the behavior of some local versions of the indices in central positions compared to the behavior elsewhere.

Lepović and Gutman [91] made the beautiful conjecture that almost all positive integers are the Wiener indices of some trees. This conjecture was verified about a decade later, in [135] and [152]. The conference paper [55] proposed a more general problem, the inverse problem of indices in combinatorial chemistry: when we have to synthesize a new molecule with expected properties, we may want to create first molecular graphs with prescribed indices to narrow the search. Inverse problems will be discussed in Section 7.

2 Distance-Based Graph Indices

The most classic and widely used distance-based index is the *Wiener index*, named after the chemist Harry Wiener, who proposed this concept in 1947 [153, 154]. For a graph G, the Wiener index of G is

$$W(G) = \sum_{u,v \in V(G)} d(u, v)$$

where $d(u, v)$ is the distance between vertices u and v in G. The mathematical examination of this concept frequently happened independently of applications in chemistry. Because of the many acyclic molecular structures in applications of the Wiener index, the study of its behavior in trees has been of particular interest. Over the past several decades, many interesting results have been obtained regarding the Wiener index of trees under various restrictions as well as variations of the Wiener index. An early informative survey on the Wiener index is [42].

2.1 The Wiener index

Among general trees of given order, the trees with maximum and minimum Wiener index have been characterized:

Theorem 1 ([44], [99, Ex. 6.23]). *Among all trees of the same order, the star minimizes the Wiener index and the path maximizes the Wiener index.*

As we will see, the star and the path are the extremal trees for many other indices as well. In fact, [30] argues that any acceptable branching index should attain the unique minimum for a star and the unique maximum for the path (or vice versa) among trees on the same number of vertices. If the degrees of vertices in the graph correspond to valences of atoms in a molecule, then they are severely restricted. Hence it is relevant to consider the extremal values of the Wiener index when the tree has a bounded maximum degree. Fischermann, Hoffmann, Rautenbach, Székely and Volkmann [52], and independently Jelen and Triesch [76] identified the trees with minimum Wiener index among all trees (of given order) with a bounded maximum degree, and the trees with maximum Wiener index among all trees (of given order) whose vertex degrees are 1 or k.

The work in [52] was further generalized to trees with a given degree sequence. For a tree T, the *degree sequence* is simply the non-increasing sequence of vertex degrees. In [149] and [164], respectively, it is shown that the minimum Wiener index is attained by the so-called *greedy tree*.

Definition 1 (Greedy trees). Given a tree degree sequence d, the greedy tree is achieved through the following "greedy" algorithm:

i) Start with a single vertex v as the root and give v the appropriate number of children so that it has the largest degree;
ii) Label the neighbors of v as $v_1, v_2, \ldots$, and assign to them the largest available degrees such that $d(v_1) \geq d(v_2) \geq \cdots$;
iii) Label the neighbors of v_1 (except v) as $v_{11}, v_{12}, \ldots$ such that they take all the largest degrees available and that $d(v_{11}) \geq d(v_{12}) \geq \cdots$, then do the same for $v_2, v_3, \ldots$;
iv) Repeat (iii) for all the newly labeled vertices, always start with the neighbors of the labeled vertex with largest degree whose neighbors are not labeled yet.

For example, Fig. 1 shows a greedy tree with degree sequence

$$\{4,4,4,3,3,3,3,3,3,3,2,2,1,\dots,1\}.$$

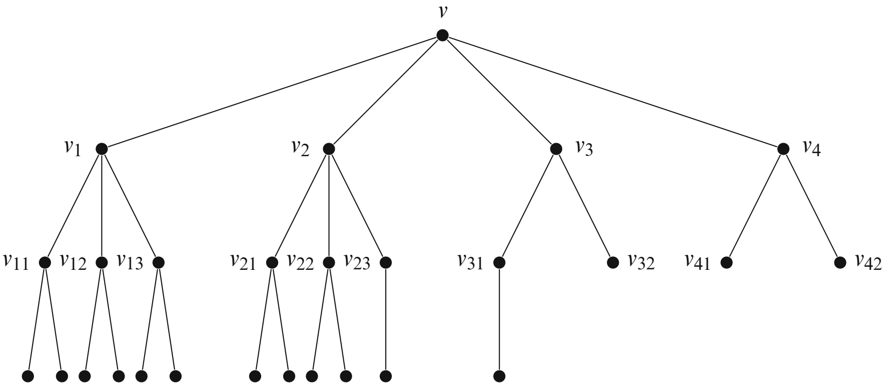

Fig. 1 A greedy tree.

Theorem 2 ([149, 164]). *Among all trees with a given degree sequence (and hence given order), the greedy tree minimizes the Wiener index.*

To maximize the Wiener index among trees with a given degree sequence turned out to be a much more difficult question. Such extremal trees were examined in detail in both [118] and [163]. While the extremal structure was already shown by Shi [117] to be a *caterpillar* (a tree for which the removal of leaves yields a path) with certain properties, the specific characteristics of such extremal trees depend on the particular degree sequence. This question was also examined as a quadratic assignment problem in [28], where an efficient algorithm was provided.

2.2 Variations of the Wiener index

In the past three decades many variations of the Wiener index have been introduced, including, but certainly not limited to the following:

- In 1993, the *hyper-Wiener index* [112] was introduced for trees and later generalized to cyclic graphs [84]:

$$WW(T) = \frac{1}{2} \sum_{u,v \in V(T)} \left(d(u,v) + d(u,v)^2 \right),$$

- the *Harary index* was defined in [71, 109]:

$$H(T) = \sum_{u,v \in V(T)} \frac{1}{d(u,v)},$$

- and the *terminal Wiener index* was proposed in [61]:

$$TW(T) = \sum_{u,v \in L(T)} d(u,v),$$

  where $L(T)$ stands for the set of leaves of T. In addition to its application in chemistry, the terminal Wiener index, being simply the sum of distances between leaves, is also found to be of importance in the study of phylogenetic trees and is known there as the *gamma index* [128].

Much work has been done on these indices. Among general trees of given order, it has been shown that the star minimizes $WW(T)$ [57, 59] and maximizes $H(T)$ [57, 158], while the path maximizes $WW(T)$ [57, 59], minimizes $TW(T)$ [61], and minimizes $H(T)$ [57, 158].

Besides finding extremal trees with the maximum (or minimum) value of these indices, a significant amount of work has also been done on characterizing trees with the first few largest or smallest values of a certain index. For a summary of such studies on the aforementioned indices we suggest Section 3 of [159].

2.3 General results and unified approaches

Given all these different graph indices defined in terms of distances with identical extremal structures, it is natural to explore unified approaches that provide more general results on such extremal questions. While studying $TW(T)$ of trees with given degree sequence, a "semi-regular" property was introduced that is satisfied by the extremal trees with a given degree sequence with respect to many indices [128]. Furthermore, satisfying this semi-regular property forces the tree to be a greedy tree.

Theorem 3 ([128]). *Among all trees of given degree sequence, $TW(T)$ is minimized by the greedy tree.*

Shortly thereafter, it was shown in [115] that if a tree is a *level-greedy tree* (defined below) with respect to any choice of root, then it satisfies the semi-regular property.

Definition 2. [Level-greedy trees] For $i = 0, 1, \ldots, H$, partition a degree sequence into multisets $\{a_{i1}, a_{i2}, \ldots, a_{i\ell_i}\}$ such that $\ell_0 = 1$, $\ell_1 = a_{01}$, and for $1 \leq i < H$

$$\ell_{i+1} = \sum_{j=1}^{\ell_i} (a_{ij} - 1).$$

Assume that the elements of each multiset are sorted, i.e., $a_{i1} \geq a_{i2} \geq \cdots \geq a_{i\ell_i}$. The level-greedy tree corresponding to this sequence of multisets is the rooted tree whose j-th vertex at level i has degree a_{ij}.

Likewise, for a degree sequence that is given as sorted multisets $\{a_{i1}, a_{i2}, \ldots, a_{i\ell_i}\}$ for $i = 0, 1, \ldots, H$ such that $\ell_0 = 2$ and for $0 \leq i < H$

$$\ell_{i+1} = \sum_{j=1}^{\ell_i} (a_{ij} - 1),$$

the level-greedy tree corresponding to this sequence is the edge-rooted tree (i.e., there are two vertices at level 0, connected by an edge) whose j-th vertex at level i has degree a_{ij}.

It is obvious that every greedy tree is also level greedy, but not necessarily vice versa. Fig. 2 shows a level-greedy tree with the level degree sequences: $\{a_{01} = 3\}$, $\{a_{11} = 5, a_{12} = 3, a_{13} = 2\}$, $\{3, 3, 3, 2, 2, 1, 1\}$, $\{2, 2, 1, 1, 1, 1, 1, 1\}$, and $\{1, 1\}$. This level-greedy tree is not greedy.

However, it was shown in [115] that a tree that is level-greedy with respect to any choice of vertex root or edge root must indeed be greedy. This led in particular to the following stability results for the extremal trees:

Theorem 4 ([115, 147]). *Let $f(x)$ be any non-negative, non-decreasing function of x. Then the graph invariant*

$$W_f(T) = \sum_{u,v \in V(T)} f(d(u,v))$$

is minimized by the greedy tree among all trees with given degree sequence.

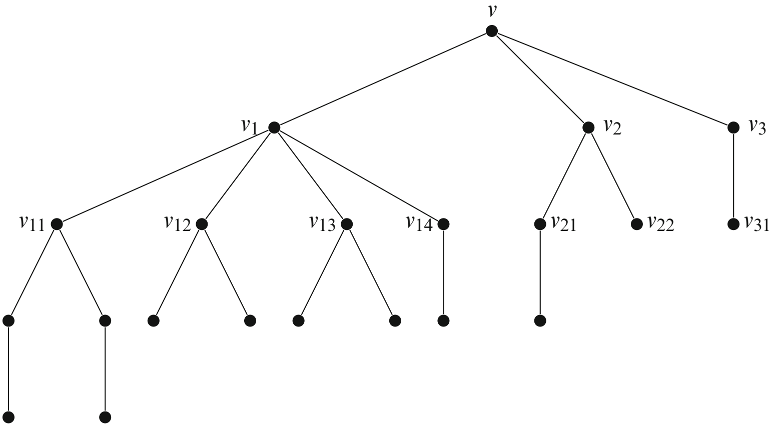

Fig. 2 A level-greedy tree.

Likewise, if $f(x)$ is any non-negative, non-increasing function of x, then the graph invariant W_f is maximized by the greedy tree among all trees with given degree sequence.

With different choices of the function f, the result above provides a general statement for various indices, including the Wiener index, the hyper-Wiener index, and the Harary index.

An important partial order of degree sequences is defined by a relation known as *majorization*:

Definition 3. For non-increasing sequences $\pi = (d_0, \ldots, d_{n-1})$ and $\pi' = (d'_0, \ldots, d'_{n-1})$, π' is said to *majorize* π if $\sum_{i=0}^{n-1} d_i = \sum_{i=0}^{n-1} d'_i$, and for $k = 0, \ldots, n-2$

$$\sum_{i=0}^{k} d_i \leq \sum_{i=0}^{k} d'_i.$$

The notion of majorization provides a means of comparing the extremal trees for different degree sequences, which yields a number of corollaries. Specifically, we have

Theorem 5 ([147]). *Let $f(x)$ be any non-negative, non-decreasing function of x, and let π and π' be two degree sequences of trees of the same length such that π' majorizes π. If $G(\pi)$ and $G(\pi')$ are the greedy trees associated with π and π', respectively, we have*

$$W_f(G(\pi)) \geq W_f(G(\pi')).$$

Likewise, if f is a non-negative, non-increasing function, then

$$W_f(G(\pi)) \leq W_f(G(\pi')).$$

Among other things, this shows that the star is always extremal, that the greedy tree associated with the sequence $(\Delta, \Delta, \ldots, \Delta, r, 1, 1, \ldots, 1)$ is extremal among trees of given order whose maximum degree is Δ (here, $r \in \{1, 2, \ldots, \Delta\}$ is chosen in such a way that the correct number of vertices is obtained), and that the greedy tree associated with the sequence $(\ell, 2, 2, \ldots, 2, 1, 1, \ldots, 1)$ is extremal among trees with exactly ℓ leaves. A similar comparison works for the Merrifield-Simmons and Hosoya indices, although the extremal trees in those cases are not greedy trees, see Section 3.

2.4 Other distance-based indices

Besides variations of the Wiener index, some other distance-based indices were also of interest. For example, the sum of distances between internal vertices and leaves is studied in [151], the sum of the *eccentricity* (the largest distance from any vertex to a given vertex) and equivalently the average eccentricity in trees are considered in [70, 121, 129]. The star, path, greedy trees, and caterpillars continue to be extremal with respect to these indices. Some lesser known distance-based indices are mentioned in [159].

3 Graph Indices Based on Counting

3.1 Independent sets and matchings

Several important graph invariants are based on counting particular sets of vertices or edges. Two prominent examples in chemical graph theory are the *Merrifield-Simmons index* and the *Hosoya index*, defined as the number of independent sets and matchings, respectively. There is a vast amount of literature devoted to the extremal values of these invariants in various families of graphs, enough to fill a survey of its own (see [143]). This includes in particular trees with various restrictions on the maximum degree, diameter, number of leaves, etc. Remarkably, one observes the general phenomenon that the graphs that maximize the Merrifield-Simmons index in some given class of graphs also minimize the Hosoya index, and vice versa. While there are many examples (and a few counterexamples) of this connection, it is still poorly understood. The paper [136] makes an attempt by studying the correlation of the two for random trees, and [53] gives a number of inequalities between the number of independent or 2-independent sets and matchings. It would be highly desirable to have a better understanding of the relation between the two indices.

The connection applies in particular to trees: unsurprisingly, if we do not impose any further restrictions, the extremal trees are the star and the path. The star S_n has the greatest number of independent sets (namely $2^{n-1} + 1$) and the smallest number of matchings (namely n), while the path P_n has the greatest number of matchings and the smallest number of independent sets. Both are Fibonacci numbers: the number of independent sets of a path with n vertices is the Fibonacci number F_{n+2} (where $F_1 = F_2 = 1$, $F_{n+1} = F_n + F_{n-1}$), while the number of matchings is F_{n+1}. This illustrates the general fact that the Hosoya index of a graph is the Merrifield-Simmons index of the line graph, since matchings of the original graph correspond exactly to independent sets of the line graph. The occurrence of Fibonacci numbers is also the reason why the number of independent sets was called the *Fibonacci number* of a graph in what is probably its earliest occurrence in the mathematical literature [110].

Let us remark that the Merrifield-Simmons index and the Hosoya index are intimately tied to two important graph polynomials. If $m(G,k)$ denotes the number of matchings of cardinality k in a graph G, the *matching polynomial* of G is defined as

$$\sum_{k\geq 0}(-1)^k m(G,k)x^{n-2k}.$$

Remarkably, the matching polynomial coincides with the characteristic polynomial (of the adjacency matrix) for trees (see, e.g., Section 5 in [47]). This is also one of the reasons for the way the matching polynomial is defined, which is perhaps somewhat less intuitive than the definition of the *matching-generating polynomial*

$$\sum_{k\geq 0} m(G,k)x^k.$$

Note that the Hosoya index, henceforth denoted by $Z(G)$, is simply the value of this polynomial at $x = 1$, i.e., $Z(G) = \sum_{k\geq 0} m(G,k)$. Since the Hosoya index is so closely related to the characteristic polynomial, it is also unsurprising that there are connections to graph invariants based on the spectrum, such as the *graph energy* (the sum of the absolute values of all eigenvalues, see the recent book [97]). Likewise, the Merrifield-Simmons index is the value of the *independence polynomial* at $x = 1$: if $i(G,k)$ is the number of independent sets of cardinality k in G, this polynomial is defined by

$$\sum_{k\geq 0} i(G,k)x^k,$$

see [92] for a survey on this polynomial. The Merrifield-Simmons index, in the following denoted by $\sigma(G)$, is of course given by $\sigma(G) = \sum_{k\geq 0} i(G,k)$.

While it is impossible to give a complete account of extremal results on the Merrifield-Simmons index and the Hosoya index of trees (the reader is again referred to [143] for a more comprehensive survey), let us state a fairly general theorem due to Andriantiana [4] that implies many other results as corollaries. Since the path and the star are extremal trees, as they are for the Wiener index, one might assume that the greedy trees (see Theorem 2 in Section 2) are also extremal again, but this is not the case.

Definition 4. Let $(d_1, d_2, \ldots, d_k, 1, 1, \ldots, 1)$ be a degree sequence of a tree, where $d_k \geq 2$, in non-increasing order. Let the associated tree $\mathscr{M}(d_1, d_2, \ldots, d_k, 1, 1, \ldots, 1)$ be defined recursively as follows: if $k \leq d_k + 1$, then $\mathscr{M}(d_1, d_2, \ldots, d_k, 1, 1, \ldots, 1)$ is the tree obtained from a star S_{d_k+1} with d_k leaves by identifying $k-1$ of its leaves with the centers of stars $S_{d_1}, S_{d_2}, \ldots, S_{d_{k-1}}$, respectively.

The internal vertices (non-leaves) are labeled $v_1, \ldots, v_k$ in such a way that their degrees are exactly $d_1, d_2, \ldots, d_k$ (in particular, v_1 is the center of the star S_{d_k+1} that started the construction).

If $k \geq d_k + 2$, we define $\mathscr{M}(d_1, d_2, \ldots, d_k, 1, 1, \ldots, 1)$ as follows: let l be the greatest integer such that v_l is a label in $\mathscr{M}(d_{d_k}, \ldots, d_{k-1}, 1, 1, \ldots, 1)$, and let s be the smallest integer such that v_s is adjacent to a leaf in $\mathscr{M}(d_{d_k}, \ldots, d_{k-1}, 1, 1, \ldots, 1)$. Now $\mathscr{M}(d_1, d_2, \ldots, d_k, 1, 1, \ldots, 1)$ is obtained from $\mathscr{M}(d_{d_k}, \ldots, d_{k-1}, 1, 1, \ldots, 1)$ by connecting a leaf that is adjacent to v_s to the centers of $d_k - 1$ disjoint stars $S_{d_1}, S_{d_2}, \ldots, S_{d_{d_k-1}}$. The centers of these stars receive the labels $v_{l+1}, \ldots, v_{l+d_k-1}$, in increasing order of degree.

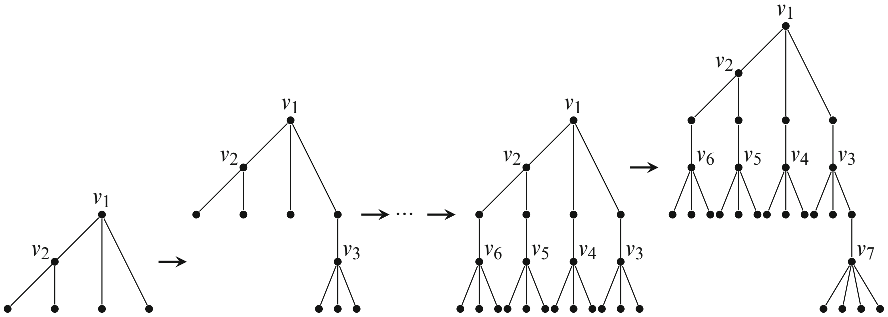

Fig. 3 Construction of the tree $\mathscr{M}(5,4,4,4,4,3,3,2,2,2,2,2,1,1,\ldots,1)$.

See Figure 3 for an example of this step-by-step construction. Note that unlike greedy trees, large and small degrees alternate in $\mathscr{M}(d_1,d_2,\ldots,d_n)$.

Theorem 6. *Among all trees with degree sequence $(d_1,d_2,\ldots,d_n)$, the tree $\mathscr{M}(d_1,\ d_2,\ldots,d_n)$ has the greatest number of independent sets and the least number of matchings (and it is unique with either of these two properties).*

In spite of the fact that the extremal trees in this scenario are no longer the greedy trees of Section 2, a majorization result (see Definition 3) akin to greedy trees holds.

Theorem 7. *If a degree sequence $(d'_1,d'_2,\ldots,d'_n)$ majorizes another degree sequence $(d_1,d_2,\ldots,d_n)$, then we have*

$$\sigma(\mathscr{M}(d_1,d_2,\ldots,d_n)) \le \sigma(\mathscr{M}(d'_1,d'_2,\ldots,d'_n))$$

and

$$Z(\mathscr{M}(d_1,d_2,\ldots,d_n)) \ge Z(\mathscr{M}(d'_1,d'_2,\ldots,d'_n)).$$

As an immediate application, one finds (in a similar way as for distance-based indices, see Theorem 5) that the tree $\mathscr{M}(\Delta,\Delta,\ldots,\Delta,r,1,1,\ldots,1)$ is extremal among trees whose maximum degree is Δ, $\mathscr{M}(\ell,2,2,\ldots,2,1,1,\ldots,1)$ is extremal among trees with exactly ℓ leaves, and $\mathscr{M}(n-D+1,2,2,\ldots,2,1,1,\ldots,1)$ is extremal among trees with diameter D (and fixed number n of vertices in all cases).

The analogous problems for the minimum of the Merrifield-Simmons index and the maximum of the Hosoya index generally appear to be harder. A complete characterization is known for trees with given maximum degree [139], and partial results are available for trees with given number of leaves [40, 160].

If one is interested in the maximum or minimum number of independent sets or matchings, it makes sense to include a restriction on the independence number or matching number. Note that in view of the König-Egerváry theorem, these two quantities are related: the sum of the matching number and the independence number in trees (and, more generally, in bipartite graphs) equals the number of vertices.

The trees of given order and independence number that maximize the number of independent sets were determined by Bruyère and Mélot [21] as part of a more general result, and a partial characterization for the minimum was given by Bruyère, Joret, and Mélot in [22]. The number of matchings in trees with given matching number, on the other hand, was studied in [67].

Instead of imposing additional restrictions on the trees, it is also natural and interesting to restrict the sets that are counted. In particular, independent sets and matchings that are *maximal* (with respect to set inclusion, i.e., not contained in any larger independent set or matching) or *maximum* (i.e., of greatest possible cardinality) have been studied. Things change considerably in that the extremal trees are no longer simply the star and the path.

While the star has the minimum number of maximal independent sets (namely 2) and the minimum number of maximum independent sets (namely 1, but it is not the only tree with this property), the path is not extremal for these two quantities. Wilf [155] showed that a tree of order n has at most $2^{(n-1)/2}$ maximal independent sets if n is odd, and $2^{(n-2)/2} + 1$ if n is even. Sagan [113] gave an alternative proof and also characterized the extremal graphs. For odd order, they are simply subdivided stars, for even order they are *batons*, obtained by attaching one or more paths of length 2 to the two ends of a path of length 1 or 3. The maximum of the number of maximum independent sets is also $2^{(n-2)/2} + 1$ if n is even, but only $2^{(n-3)/2}$ for odd n, as proven by Zito in [168]. These results have been extended further to graphs with at most r cycles [114].

For maximal and maximum matchings, the situation is even more complicated. The number of maximum matchings is trivially bounded below by 1, and this value is attained for any tree with a perfect matching. On the other hand, the trees with the greatest number of maximum matchings have one of seven different shapes (depending on the number of vertices modulo 7) composed of a repeated pattern of seven vertices, see [65]. There are also exceptions to the general pattern up to order 34 and two different extremal trees if the number of vertices is either 6 or 34. The maximum is of order $\Theta(1.391664^n)$, the exact constant being $(\frac{1}{2}(11 + \sqrt{85}))^{1/7}$. This also gives a lower bound for the maximum number of maximal matchings in a tree (since every maximum matching is automatically maximal), and an upper bound of $O(1.395337^n)$ was given for this maximum number by Górska and Skupień [56]. The exact shape of the trees with greatest number of maximal matchings and the asymptotic behavior of this greatest number remain an open problem.

Finally, we mention one further variant, namely the number of independent or maximal independent sets containing the leaves of a tree, which was studied in [156] and [157].

3.2 Subtrees and related invariants

Research on the number of *subtrees* (connected induced subgraphs of a tree) and a related quantity, the number of subtrees containing at least one leaf, started with [123–125], motivated by the study of the complexity of an algorithm in bioinformatics [85]. Remarkably, trees that maximize the number of subtrees are typically also those that minimize the Wiener index and vice versa, a similar connection as mentioned earlier for the Merrifield-Simmons index and the Hosoya index. In particular, a tree with n vertices has at least $\binom{n+1}{2}$ and at most $2^{n-1}+n-1$ subtrees (the minimum occurs for the path, the maximum for the star).

A lot of further work was done in particular on trees with given degree sequence [81, 166, 167], where the greedy trees of Theorem 2 occur again as extremal structures. In fact, as shown in [7], they have the greatest number of subtrees of any given cardinality among trees with a specific degree sequence. The situation for the minimum is more complicated, but it is known that it always occurs for a caterpillar, and partial results on the precise structure have been obtained as well [119, 165]. Other families of trees (with given number of leaves, bipartition, etc.) were considered in [93].

Some closely related concepts can be found in two much earlier papers of Jamison [72, 73], who studied the *average subtree order* (average number of vertices in a subtree). Among other things, he showed that the path has the smallest average subtree order among trees with n vertices, namely $\frac{n+2}{3}$. He posed many open questions, some of which were only resolved recently [63, 134, 144, 145] or are still open. In particular, it is not known which trees maximize the average subtree order. The star is only extremal if the number of vertices is very small, while the extremal trees for larger orders are quite “path-like” (in that they must have many vertices of degree 2). Jamison conjectured that the maximum might always be attained for a caterpillar. Since the average subtree order is also a relatively complicated parameter to calculate, not a lot of evidence for or against this conjecture is available.

As a final remark, let us mention an interesting variant obtained by counting subtrees only up to isomorphism. In this context, the path and star of order n are “equally bad,” in that they only have n non-isomorphic subtrees, which is clearly also the minimum. As it turns out, the maximum is of order $\Theta(5^{n/4})$, see [37]. A simple construction yields trees that reach this bound: take a path of length $\frac{n}{4}$ and attach a leaf and a path of length 2 to each vertex.

3.3 Dominating sets

The number of *dominating sets* is again very different from the other quantities mentioned in this section so far. An important feature is the fact that leaves can only be dominated by their neighbors or themselves, which often severely limits the possibilities. Interestingly, the number of dominating sets of any graph is always odd, see [20].

The maximum number of dominating sets among trees of order n is obtained for a star (and also for the path if the number of vertices is at most 5, see [18]), but the lower bound of $5^{n/3}$, $9 \cdot 5^{(n-4)/3}$, or $3 \cdot 5^{(n-2)/3}$ (depending on whether n is congruent to 0, 1, or 2 modulo 3) is attained for a rich class of trees: the condition is simply that each internal vertex must be adjacent to exactly two leaves, with one exception (adjacent to only one leaf) if $n \equiv 2 \bmod 3$ and one exception (adjacent to three leaves) or two exceptions (each adjacent to one leaf) if $n \equiv 1 \bmod 3$. However, there is nothing particularly special about trees in this context, as was pointed out in [141] (see also [120]): the same values are extremal for connected graphs (and even for graphs without isolated vertices), and the characterization of extremal connected graphs is also essentially the same.

Similar observations can be made for the number of *efficient dominating sets* (i.e., dominating sets with the property that no vertex is dominated by more than one of the vertices in the dominating set) and *total dominating sets*. Similar to the 3-periodicity for the total number of dominating sets, a period of 7 can be observed in the structure of the extremal trees, see [19] and [89].

Krzywkowski [86–88] studied several other types of domination and provided algorithms for listing minimal dominating sets, minimal 2-dominating sets, or minimal double dominating sets. The running time analysis of these algorithms provides upper bounds for the respective numbers.

3.4 Walks

Let $A(G)$ denote the adjacency matrix of a graph. It is well known that the entries of the k-th power $A(G)^k$ count the number of walks of length k in G starting and ending at specified vertices. In particular, the total number of closed walks of length k (each with a fixed starting vertex), which we denote by $\overline{w}_k(G)$, equals the trace of $A(G)^k$ (which in turn is equal to the k-th spectral moment), and the total number of all walks of length k, denoted by $w_k(G)$, is the sum of all entries in $A(G)^k$, which is $\mathbf{1}^T A(G)^k \mathbf{1}$.

The number of walks is closely connected to the spectrum, in particular the spectral radius, which is equal to the limits $\lim_{k\to\infty} w_k(G)^{1/k}$ and $\lim_{k\to\infty} \overline{w}_k(G)^{1/k}$; in the latter case, the limit needs to be taken only over even numbers k if G is bipartite (in particular if it is a tree). There are also explicit inequalities relating the spectral radius and the number of walks of different lengths, see [107] and the references therein. Spectral graph theory is of course a vast area of its own right, and a lot more could be said about spectral parameters of trees, but since our space is limited, let us refer to the books [29, 34, 35] and the surveys in [33].

We specifically mention two recent papers [130, 131] by Täubig et al. that also deal with inequalities for walks in graphs, especially because of an interesting conjecture on trees. It is shown in [130] that

$$w_{2a+c}(G)w_{2a+2b+c}(G) \leq w_{2a}(G)w_{2(a+b+c)}(G)$$

for all graphs G and non-negative integers a, b, c, which generalizes several inequalities proved earlier by other authors. Since $w_0(G) = |V(G)|$ and $w_1(G) = 2|E(G)|$, this implies in particular that

$$\frac{2|E(G)|}{|V(G)|} \cdot w_{k-1}(G) \leq w_k(G)$$

for all graphs G if k is even. Täubig et al. [130] conjecture that this inequality holds for *all* positive integers k if G is a tree. It is noteworthy that it is not valid for arbitrary graphs or even for the narrower class of bipartite graphs, and counterexamples are given in [130]. On the other hand, it is proven for trees in the special cases $k = 3$ and $k = 5$ in [130].

The problem of maximizing the number of closed walks in trees was motivated in particular by the aforementioned connection to spectral moments and thus also the so-called *Estrada index* [62]. Csikvári [31] proved that the path has the minimum number of closed walks of any fixed length among trees of given order, and the star has the maximum number of such walks. This was generalized to arbitrary walks by Bollobás and Tyomkyn [17], who also gave an alternative proof. Csikvári's proof is based on defining a poset structure on trees of given order, and he showed [32] that this idea could also be applied to many other tree invariants (specifically, coefficients and roots of different graph polynomials).

Trees with given degree sequence were studied in [5], and it turns out that greedy trees as defined in Section 2 are extremal once again, which has several implications on spectral invariants: they maximize all spectral moments, thus also the spectral radius (which was proven earlier in [13]) and also any invariant of the form

$$E_f(G) = \sum_i f(\lambda_i(G)),$$

where $\lambda_1(G), \lambda_2(G), \ldots$ are the eigenvalues of G and f is any entire function whose Taylor series at 0 has only non-negative coefficients. This includes the aforementioned Estrada index (corresponding to $f(x) = e^x$) as a particular instance. Again, a majorization theorem holds as well, which implies one half of Csikvári's result (extremality of the star).

It is very likely that the greedy trees also have the greatest possible total number of walks (not necessarily closed), but it seems that this is more difficult to prove (the proof for closed walks is already quite long and technical).

3.5 Rooted trees regarded as posets

A rooted tree in its standard drawing can be interpreted in a natural way as the Hasse diagram of a partially ordered set, so it makes sense to study concepts stemming from the theory of posets, such as *chains* and *antichains*.

There is a natural bijection between antichains in a rooted tree and subtrees containing the root (the leaves of such a subtree forming the corresponding antichain), so the results of [7] also show that greedy trees (with their natural root) also maximize the number of antichains, even antichains of any fixed cardinality.

Another invariant of a similar nature that was studied recently is the number of *transversals*, i.e., subsets of vertices with the property that every path from the root to a leaf contains at least one of the elements (this can be the root or leaf itself). Thus removal of a transversal "destroys" the connection between the root and the leaves, and in fact one of the motivations for the authors of [23] to study transversals was a mathematical model in counterterrorism. Among other results, the d-ary trees with minimum and maximum number of transversals are characterized in [23]: they are caterpillars and full d-ary trees, respectively.

Finally, let us mention a problem due to Klazar [83] that was only solved very recently [6]. An *infima closed set* in a poset is a set with the property that the infimum (greatest common lower bound) of any subset exists and is contained in the set as well. In the context of rooted trees, this means that the closest common ancestor of any two vertices in such a set must be contained in the set as well. Both the star and the path have "many" infima closed sets: in fact, all subsets of a path, rooted at one of its ends, are infima closed, and all sets containing the root of a star, rooted at its center, are infima closed. Klazar's question was therefore to characterize the trees with the least number of infima closed sets, and it turns out that these trees are essentially complete binary trees (all internal vertices have two children), except for the vertices that are directly adjacent to leaves: these vertices must have three leaves as children, with a bounded number of exceptions (that depend on the precise order of the tree).

4 Degree-Based Graph Indices

Randić [111] studied a new index, earlier called branching index or connectivity index,

$$R_\alpha(G) = \sum_{\{u,v\}\in E(G)} (d_G(u)d_G(v))^\alpha,$$

where the summation is over all edges of the graph, $d_G(u)$ denotes the degree of vertex u in graph G, and $\alpha = -1/2$. This is known today as Randić index, and for other values of α it is known as the generalized Randić index. (When $\alpha < 0$, the index is undefined for graphs with isolated vertices, but molecular graphs are connected.) $R_1(G)$ is also known as the second Zagreb index. (The first Zagreb index is just the sum of degree squares; more general versions of it allow any exponent instead of 2.)

There could be many graphs with the same degree sequence, and they could look strikingly different, depending on whether vertices of similar or very different degrees tend to be connected. In network science the concepts of assortativity and disassortativity describe this phenomenon, where assortativity is usually measured by the Pearson correlation coefficient of degree pairs of adjacent vertices. Assortativity is determined by the joint degree matrix [38], whose entries give the number m_{ij} of edges between degree i and degree j vertices. The Randić index can be written alternatively as

$$R_\alpha(G) = \frac{1}{2}\sum_{u\in V(G)} d_G(u)^{2\alpha+1} - \frac{1}{2}\sum_{i<j} m_{ij}(i^\alpha - j^\alpha)^2, \tag{1}$$

in which the first sum turns just n for $\alpha = -1/2$. We suspect that the power of the Randić index comes from using information from the whole joint degree matrix.

The Randić index became better known among graph theorists after the work of Bollobás and Erdős [15, 16]. They [15] proved that among graphs on n vertices with minimum degree at least 1, the star minimizes $R_{-1/2}$. [94] discusses the development of minimizing $R_{-1/2}$ among graphs on n vertices with minimum degree at least k, as this problem is not yet completely solved. On the other hand, formula (1) shows that among graphs on n vertices, the regular graphs maximize $R_{-1/2}$. Further restricting graphs to trees, the path maximizes $R_{-1/2}$ [25, 161].

Adding an edge to the graph, $R_{-1/2}$ can go up or down. This is a delicate quantity, and it is no surprise that a few published conjectures about it were refuted. Conjecture making and proving software like Graffiti [46], and AutographiX [3, 24] included $R_{-1/2}$ among the graph parameters they compared. Aouchiche et al. [3] made the conjecture that for any connected graph of order $n \geq 3$ with diameter $D(G)$,

$$R_{-1/2}(G) - D(G) \geq \sqrt{2} - \frac{n+1}{2} \quad \text{and} \quad \frac{R_{-1/2}(G)}{D(G)} \geq \frac{n-3+2\sqrt{2}}{2n-2},$$

and equality holds in any of them iff G is a path. [100] proved a stronger result that easily implies this conjecture: for any connected graph of order $n \geq 3$ with diameter $D(G)$, $R_{-1/2}(G) - \frac{1}{2}D(G) \geq \sqrt{2} - 1$, and equality holds iff G is a path.

Fajtlowicz [46] conjectured for connected graphs that $R_{-1/2}(G) \geq \frac{W(G)}{\binom{n}{2}}$, where the average distance on the RHS is written with the Wiener index in the numerator. Caporossi and Hansen [24] strengthened this conjecture to $R_{-1/2} \geq \frac{W(G)}{\binom{n}{2}} + \sqrt{n-1} + \frac{2}{n} - 2$, which, if true, is tight. They [24] also conjectured that $r(G) \leq R_{-1/2}(G)$, where $r(G)$ is the radius of the graph G, unless G is an even path, and verified this conjecture if G is a tree.

The survey [94] further discusses extremal results for $R_{-1/2}(T)$, where (a) T is a tree with given number of vertices and leaves; (b) T is a tree with given diameter and number of vertices; (c) T is a chemical tree with given number of vertices; (d) T is a chemical tree with given number of vertices and leaves. As chemical trees are the relevant trees for chemical graph theory, [60] worked out the three largest possible and the three smallest possible values of $R_{-1/2}(T)$, where T is a chemical tree with given number of vertices, and the realizing trees.

For the generalized Randić index R_α, the minimizing tree is a path for positive α and $n \geq 5$, and is a star for negative α [68]. Randić himself considered R_{-1} [111]. Building on considerable earlier work, [108] and [69] showed that the maximum of R_{-1} on trees with n vertices is $\frac{15}{56}n + o(n)$. The survey [94] further discusses extremal results for $R_{-1}(T)$, where (a) T is a tree with given number of vertices and leaves; (b) T is a tree with given number of vertices and has maximum degree 3; (c) T is a chemical tree with given number of vertices.

Gutman [58] lists a number of variants of the Randić index and notes "we have far too many descriptors, and there seems to lack a firm criterion to stop or slow down their proliferation." We include here one more, the higher order Randić index proposed in [78]:

$$^{i}R(G) = \sum_{v_0 v_1 \ldots v_i} \frac{1}{\sqrt{d_G(v_0) d_G(v_1) \cdots d_G(v_i)}},$$

where the sum is taken over paths $v_0 v_1 \ldots v_i$ of length i in the graph G. $i = 1$ gives back $R_{-1/2}(G)$, while $i = 0$ gives $\sum_v d_G(v)^{-1/2}$, the first Zagreb index with exponent $-1/2$. We note that there is some ambiguity in the definition of higher order Randić indices, as some papers allow any $v_0 v_1 \ldots v_i$ walks that do not repeat edges—for $i \geq 3$, if the graph has cycles, this may bring extra terms. Regarding the second order Randić index, [100] showed that among trees on n vertices, the star maximizes the second order Randić index, and among trees on n vertices with maximum degree 3, the path is the unique minimizer of the second order Randić index.

As for many other graph indices discussed in this survey, trees with a given degree sequence have been considered as well. The first article of this kind is [39], where the trees with given degree sequence that maximize R_1 were found. This was further generalized to arbitrary α in [148]. It is noteworthy that the extremal trees are generally not unique, and it would be interesting to find necessary and sufficient conditions on the degree sequence for the extremal tree to be unique. Finally, an even more general type of graph invariant, where $(d_G(u)d_G(v))^\alpha$ is replaced by a symmetric function $f(d_G(u), d_G(v))$ (so that the Randić index occurs as a special case) is studied in [150].

5 Random Trees

To have an idea whether the Wiener index of a tree is "small" or "large," and to facilitate statistical analysis, it is desirable to know the expectation and the variance of the Wiener index of a randomly selected tree.

Entringer et al. [45] gave a method to compute the asymptotics for the expected Wiener index of a random tree selected uniformly from a simply generated family of trees and found that the asymptotics is constant times $n^{5/2}$. Simply generated families include several important families, for example, binary trees, ordered trees, and unordered labeled trees. Asymptotic results for the average Wiener index of star-like trees were obtained in [135], and results relevant for chemical trees can be found in [41] and [137].

Janson [74] obtained asymptotics for the mean, variance, and higher moments as well as for the distribution of the Wiener index of a random tree from a simply generated family. This distribution can be expressed in terms of a Brownian excursion. He also determined the joint asymptotic distribution of the Wiener index and the internal path length, as well as asymptotics for the covariance and other mixed moments.

The study of the expected Wiener index extended to classes of random trees relevant in computer science. For them, the motivation is no longer chemical graph theory. Neininger [106] investigated the Wiener index of uniformly selected random recursive trees and random binary search trees, obtained asymptotics for the expectation, the variance, and also for the correlation and covariance with the path length. Munsonius [104] investigated the Wiener index and path length of random split trees asymptotically; Fuchs and Lee [54] obtained asymptotic expansions of moments of the Wiener index and showed a central limit law for the Wiener index of digital search trees, tries and PATRICIA tries.

Upper bounds for the tail distribution of the Wiener index in several models were obtained by [1, 51, 75], and [105].

Comparatively little work has been done on other graph invariants mentioned in this survey: [48] studies the distribution of the Randić index, proving a normal limit law for different random binary tree models. Central limit theorems for the (first) Zagreb index (i.e., the sum of the squared degrees) are proven in [49, 50].

Averages of various enumerative invariants (independent sets, matchings, subtrees, etc.) have been determined for different types of trees. A whole collection of results of this type can be found in a paper of Klazar [83]; see also [82, 101, 102, 138] for other instances. For the number of subtrees, the distribution was shown to be log-normal in a recent paper [142]. Although the results in this paper are quite general, they do not seem to apply directly to similar invariants such as the number of independent sets.

6 Localized Tree Indices and Concepts of Centrality

Recall that the function $ecc_T(.)$ tells the largest distance from a vertex in the tree T. The *center* $C(T)$ of the tree T is the set of vertices where $ecc_T(.)$ is minimized.

The *distance* $\sigma_T(.)$ *of a vertex* (unrelated to the Merrifield-Simmons index $\sigma(T)$ that was defined in Section 3) in a tree T is the sum of distances from the variable vertex to all other vertices. Clearly, $W(T) = \frac{1}{2}\sum_v \sigma_T(v)$. According to Zelinka [162], the *centroid* $CT(T)$ of a tree coincides with the set of vertices minimizing the distance function. Let $F(T)$ denote the number of subtrees of the tree T, and $F_T(v)$ denote the number of subtrees of the tree T that contain vertex v.

Let the function $F_T(.)$ assign to vertex v of the tree T the number of subtrees of T containing the vertex v. [124] defined the *subtree core* $Core(T)$ of the tree T, a new concept analogous to, but different from the concepts of center $C(T)$ and centroid $CT(T)$, where $F_T(v)$ is maximized. We see here an interaction of "local" and "global" tree indices: $\sigma_T(.)$ is the local version of the Wiener index, $F_T(.)$ is the local version of the number of subtrees; while $ecc_T(.)$ is inherently local, one can define the global version by $Ecc(T) = \sum_v ecc_T(v)$.

The center, the centroid, and the subtree core contain either a single vertex or two adjacent vertices of the tree. The reason is the following. Given the vertices along any path of a tree, the sequence of the values of $F_T(.)$ is strictly concave down ([124, proof to Theorem 9.1]), the sequence of the values of $\sigma_T(.)$ is strictly concave up ([99, Ex. 6.22]), and the sequence of the values of $ecc_T(.)$ is concave up ([99, Ex. 6.21]). Strict concavity immediately implies that $Core(T)$ and $CT(T)$ consist of one or two adjacent vertices. The fact that $C(T)$ also consists of one or two adjacent vertices was already known to Jordan [77] (see also [99, Ex. 6.21a]). It is well known that the center and centroid are generally distinct, and moreover, they can be arbitrary far from each other ([99, Ex. 6.22c]). The paper [122] investigates how far these different middle parts can be in a tree.

Behavior of ratios of local and global indices can be subtle. The investigation of extreme values of ratios started in [9], which determined extremal values of $W(T)/\sigma_T(v)$, $W(T)/\sigma_T(w)$, $\sigma_T(w)/\sigma_T(v)$, and $\sigma_T(w)/\sigma_T(u)$, where T is a tree on n vertices, v is in the centroid of the tree T, and u, w are leaves in T. The two papers [126] and [127] went further to see how far the negative correlation between distances and subtrees, discovered in [136], goes. They characterized the extremal values of $F(T)/F_T(v)$, $F(T)/F_T(w)$, $F_T(w)/F_T(v)$, and $F_T(w)/F_T(u)$, where T is a tree on n vertices, v is in the subtree core of the tree T, and u, w are leaves in T—the complete analogue of [9], changing distances to subtrees. It is instructive to look at the table in [126] to see the extremal trees corresponding to "dual" problems—there is a striking similarity.

The most recent investigation in this area [121] characterized the extremal values of $Ecc(T)/ecc_T(v)$, $Ecc(T)/ecc_T(w)$, $ecc_T(w)/ecc_T(v)$, and $ecc_T(w)/ecc_T(u)$, where T is a tree on n vertices, v is in the center of the tree T, and u, w are leaves in T—another complete analogue of [9], changing distances to eccentricities.

7 Inverse Problems

Perhaps [55] was the first paper that called for the systematic study of inverse problems in combinatorial chemistry: create molecular graphs with prescribed topological index. [91], however, asked for trees with a given Wiener index, and made the conjecture that with a finite number of exceptions all positive integers are Wiener indices of some trees.

Li et al. [96] study four indices and show that all positive integers are the number of cliques in some graphs. It discusses the first Zagreb index (sum of degree squares) and observes that it is even for any graph, and shows that all even integers $\neq 4, 8$ are the sum of squares of the degrees of some tree. An early but relevant paper [98] shows that any positive integer is the Merrifield-Simmons index of a bipartite graph, and there is a conjecture that any sufficiently large positive integer is the Merrifield-Simmons index of a tree (clearly every positive integer is the Hosoya index of some star). An interesting feature of this conjecture (which is also one reason why the problem might be very difficult) is the distribution of the Merrifield-Simmons index in residue classes: it was shown in [140] (and by a slightly different method in [2]) that for every modulus m, almost all trees (in the sense that the proportion converges to 1 as the order tends to infinity) have a number of independent sets that is divisible by m. However, every residue class modulo m occurs for some tree, as was shown by Law [90].

The Lepović-Gutman conjecture that all sufficiently large numbers are Wiener indices of some trees was proven independently in [152] and [135]. In fact, 49 exceptional numbers were found. Part of the difficulty of the problem is the large number of tree shapes—the first paper only used short caterpillars, while the second used star-like trees, which are obtained from a star by changing its leaves to some stars. Both papers approached the problem through representation of integers as values of certain quadratic forms. This approach had to use arbitrarily large degree vertices, which is not an attractive approach for chemical graph theorists. Combining their efforts the authors of these two papers [146] showed that every sufficiently large integer n is the Wiener index of some caterpillar tree with degrees at most 3, and every sufficiently large even integer is the Wiener index of some hexagon type graph. [11] conjectures that every sufficiently large integer is the Wiener index of a binary tree, a stronger form of the Lepović-Gutman conjecture.

In analogy to the Lepović-Gutman conjecture for the Wiener index, it was shown in [36] that all but 34 positive integers are the number of subtrees of some trees. In fact, the proof shows that all sufficiently large positive integers are equal to the number of subtrees of a caterpillar tree, and then bridges the gap by explicit computation. Here the quadratic form representation appears as a representation of a number as a certain kind of sum of pentagonal numbers.

The paper [36] poses a metaconjecture: if a tree index on n vertices takes sufficiently many values, and for rooted trees the index can be computed from the indices of the subtrees rooted at the children of the root with a reasonable polynomial formula (the paper gives a number of examples of such indices), and if the values of the tree index are not constrained to some residue class, then every sufficiently large positive integer is a value of this tree index.

A result of [135] is instructive to accept this conjecture: let

$$W_\lambda(T) = \sum_{u,v \in V(T)} d_T(u,v)^\lambda,$$

where λ is a positive integer, a variant of the Wiener index. If there is a star-like tree T such that $W_\lambda(T) \equiv r \bmod 2^\lambda(2^\lambda - 1)$, then all members of the residue class $r \bmod 2^\lambda(2^\lambda - 1)$— with only finitely many exceptions—are Wiener indices of trees. For $\lambda = 2, 3, 5, 6, 7, 9, 10$, this implies that all integers, with finitely many exceptions, can be written as $W_\lambda(T)$ for some star-like tree T, as all residue classes mod $2^\lambda(2^\lambda - 1)$ are covered. (For $\lambda = 4$ and all other multiples of 4, this is not the case any more.)

Acknowledgements This research was supported in part by the NSF DMS contract 1300547 and by the DARPA and AFOSR under the contract FA9550-12-1-0405 (László Székely), National Research Foundation of South Africa, grant number 96236 (Stephan Wagner), and Simons Foundation grant 245307 (Hua Wang).

References

1. T. Ali Khan, R. Neininger, Tail bounds for the Wiener index of random trees, in *Discrete Math. Theor. Comput. Sci. Proc., Proceedings of the 2007 Conference on the Analysis of Algorithms* (2007), pp. 279–289
2. N. Alon, S. Haber, M. Krivelevich, The number of F-matchings in almost every tree is a zero residue. Electron. J. Comb. **18**(1), paper no. #P30 (2011)
3. M. Aouchiche, P. Hansen, M. Zheng, Variable neighborhood search for extremal graphs. 19. Further conjectures and results about the Randić index. MATCH Commun. Math. Comput. Chem. **58**, 83–102 (2007)
4. E.O.D. Andriantiana, Energy, Hosoya index and Merrifield-Simmons index of trees with prescribed degree sequence. Discrete Appl. Math. **161**(6), 724–741 (2013)
5. E.O.D. Andriantiana, S. Wagner, Spectral moments of trees with given degree sequence. Linear Algebra Appl. **439**(12), 3980–4002 (2013)
6. E.O.D. Andriantiana, S. Wagner, Trees with minimum number of infima closed sets. In preparation (2015)
7. E.O.D. Andriantiana, S. Wagner, H. Wang, Greedy trees, subtrees and antichains. Electron. J. Comb. **20**(3), paper no. #P28 (2013)
8. A.T. Balaban, *Chemical Applications of Graph Theory* (Academic Press, New York, 1976)
9. C.A. Barefoot, R.C. Entringer, L.A. Székely, Extremal values for ratios of distances in trees. Discrete Appl. Math. **80**, 37–56 (1997)
10. R.J. Baxter, I.G. Enting, S.K. Tsang, Hard-square lattice gas. J. Stat. Phys. **22**(4), 465–489 (1980)
11. S. Bereg, H. Wang, Wiener indices of balanced binary trees. Discrete Appl. Math. **155**, 457–467 (2007)
12. N.L. Biggs, E.K. Lloyd, R.J. Wilson, *Graph Theory 1736–1936* (Clarendon Press, Oxford, 1976; Oxford University Press, Oxford, 1986)
13. T. Bıyıkoğlu, J. Leydold, Graphs with given degree sequence and maximal spectral radius. Electron. J. Comb. **15**(1), paper no. #P119 (2008)
14. D. Bokal, É. Czabarka, L.A. Székely, I. Vrt'o, General lower bounds for the minor crossing numbers of graphs. Discrete Comput. Geom. **44**, 463–483 (2010)
15. B. Bollobás, P. Erdős, Graphs of extremal weights. Ars Comb. **50**, 225–233 (1998)
16. B. Bollobás, P. Erdős, A. Sarkar, Extremal graphs for weights. Discrete Math. **200**, 5–19 (1999)
17. B. Bollobás, M. Tyomkyn, Walks and paths in trees. J. Graph Theory **70**(1), 54–66 (2012)

18. D. Bród, Z. Skupień, Trees with extremal numbers of dominating sets. Australas. J. Comb. **35**, 273–290 (2006)
19. D. Bród, Z. Skupień, Recurrence among trees with most numerous efficient dominating sets. Discrete Math. Theor. Comput. Sci. **10**(1), 43–55 (2008)
20. A.E. Brouwer, The number of dominating sets of a finite graph is odd. Preprint (2009). http://www.win.tue.nl/~aeb/preprints/domin2.pdf
21. V. Bruyère, H. Mélot, Fibonacci index and stability number of graphs: a polyhedral study. J. Comb. Optim. **18**(3), 207–228 (2009)
22. V. Bruyère, G. Joret, H. Mélot, Trees with given stability number and minimum number of stable sets. Graphs Comb. **28**(2), 167–187 (2012)
23. V. Campos, V. Chvátal, L. Devroye, P. Taslakian, Transversals in trees. J. Graph Theory **73**(1), 32–43 (2013)
24. G. Caporossi, P. Hansen, Variable neighborhood search for extremal graphs 1: the AutographiX system. Discrete Math. **212**, 29–44 (2000)
25. G. Caporossi, I. Gutman, P. Hansen, L. Pavlović, Graphs with maximum connectivity index. Comput. Biol. Chem. **27**, 85–90 (2000)
26. A. Cayley, On the mathematical theory of isomers. Philos. Mag. **47**, 444–446 (1874)
27. A. Cayley, A theorem on trees. Quart. J. Math. **23**, 376–378 (1889)
28. E. Çela, N. Schmuck, S. Wimer, G.J. Woeginger, The Wiener maximum quadratic assignment problem. Discrete Optim. **8**, 411–416 (2011)
29. F.R.K. Chung, Spectral graph theory, in *CBMS Regional Conference Series in Mathematics*, vol. 92 (American Mathematical Society, Providence, 1997)
30. L. Clark, I. Gutman, The exponent in the general Randić index. J. Math. Chem. **43** (1), 32–44 (2006)
31. P. Csikvári, On a poset of trees. Combinatorica **30**(2), 125–137 (2010)
32. P. Csikvári, On a poset of trees II. J. Graph Theory **74**(1), 81–103 (2013)
33. D.M. Cvetković, I. Gutman, (eds.), Selected topics on applications of graph spectra. Zb. Rad. (Beogr.) **14**(22), 155–174 (2011)
34. D.M. Cvetković, M. Doob, H. Sachs, *Spectra of Graphs. Theory and Applications*, 3rd edn. (Johann Ambrosius Barth, Heidelberg, 1995)
35. D.M. Cvetković, P. Rowlinson, S. Simić, *An Introduction to the Theory of Graph Spectra*. London Mathematical Society Student Texts, vol. 75 (Cambridge University Press, Cambridge, 2010)
36. É. Czabarka, L.A. Székely, S. Wagner, The inverse problem for certain tree parameters. Discrete Appl. Math. **157**(15), 3314–3319 (2009)
37. É. Czabarka, L.A. Székely, S. Wagner, On the number of nonisomorphic subtrees of a tree. (2016, submitted, arXiv:1601.00944)
38. É. Czabarka, A. Dutle, P.L. Erdős, I. Miklós, On realizations of a Joint Degree Matrix. Discrete Appl. Math. **181**, 283–288 (2015)
39. C. Delorme, O. Favaron, D. Rautenbach, Closed formulas for the numbers of small independent sets and matchings and an extremal problem for trees. Discrete Appl. Math. **130**, 503–512 (2003)
40. H. Deng, Q. Guo, On the minimal Merrifield-Simmons index of trees of order n with at least $\lfloor n/2 \rfloor + 1$ pendent vertices. MATCH Commun. Math. Comput. Chem. **60**(2), 601–608 (2001)
41. A.A. Dobrynin, I. Gutman, The average Wiener index of trees and chemical trees. J. Chem. Inf. Comput. Sci. **39**, 679–683 (1999)
42. A.A. Dobrynin, R. Entringer, I. Gutman, Wiener index of trees. Theory and applications. Acta Appl. Math. **66**, 211–249 (2001)
43. M. Drmota, *Random Trees. An Interplay Between Combinatorics and Probability* (Springer, Wien/New York, 2009)
44. R.C. Entringer, D.E. Jackson, D.A. Snyder, Distance in graphs. Czech. Math. J. **26**, 283–296 (1976)
45. R.C. Entringer, A. Meir, J.W. Moon, L.A. Székely, On the Wiener index of trees from certain families. Australas. J. Comb. **10**, 211–224 (1994)

46. S. Fajtlowicz, On conjectures of Graffiti. Discrete Math. **72**, 113–118 (1988)
47. E.J. Farrell, An introduction to matching polynomials. J. Comb. Theory Ser. B **27**(1), 75–86 (1979)
48. Q. Feng, H.M. Mahmoud, A. Panholzer, Limit laws for the Randić index of random binary tree models. Ann. Inst. Stat. Math. **60**(2), 319–343 (2008)
49. Q. Feng, Z. Hu, On the Zagreb index of random recursive trees. J. Appl. Probab. **48**(4), 1189–1196 (2011)
50. Q. Feng, Z. Hu, Phase changes in the topological indices of scale-free trees. J. Appl. Probab. **50**(2), 516–532 (2013)
51. J.A. Fill, S. Janson, Precise logarithmic asymptotics of the right tails of some limit random variables for random trees. Ann. Comb. **12**, 403–416 (2009)
52. M. Fischermann, A. Hoffmann, D. Rautenbach, L. Szekely, L. Volkmann, Wiener index versus maximum degree in trees. Discrete Appl. Math. **122**, 127–137 (2002)
53. M. Fischermann, L. Volkmann, D. Rautenbach, A note on the number of matchings and independent sets in trees. Discrete Appl. Math. **145**(3), 483–489 (2005)
54. M. Fuchs, C.-K. Lee, The Wiener index of random digital trees. SIAM J. Discrete Math. **29**(1), 586–614 (2015)
55. D. Goldman, S. Istrail, G. Lancia, A. Piccolboni, B. Walenz, Algorithmic strategies in combinatorial chemistry, in *SODA '00 Proceedings of the eleventh annual ACM-SIAM Symposium on Discrete Algorithms* (2000), pp. 275–284
56. J. Górska, Z. Skupień, Trees with maximum number of maximal matchings. Discrete Math. **307**(11–12), 1367–1377 (2007)
57. I. Gutman, A property of the Wiener number and its modifications. Indian J. Chem. **36A**, 128–132 (1997)
58. I. Gutman, Degree-based topological indices. Croat Chem. Acta **86**(4), 351–361 (2013)
59. I. Gutman, W. Linert, I. Lukovits, A.A. Dobrynin, Trees with extremal hyper-Wiener index: Mathematical basis and chemical applications. J. Chem. Inf. Comput. Sci. **37**, 349–354 (1997)
60. I. Gutman, O. Miljković, G. Caporossi, P. Hansen, Alkanes with small and large Randić connectivity indices. Chem. Phys. Lett. **306**(5), 366–372 (1999)
61. I. Gutman, B. Furtula, M. Petrović, Terminal Wiener index. J. Math. Chem. **46**, 522–531 (2009)
62. I. Gutman, H. Deng, S. Radenković, The Estrada index: an updated survey. Zb. Rad. (Beogr.) **14**(22), 155–174 (2011)
63. J. Haslegrave, Extremal results on average subtree density of series-reduced trees. J. Comb. Theory Ser. B **107**, 26–41 (2014)
64. O.J. Heilmann, E.H. Lieb, Theory of monomer-dimer systems. Commun. Math. Phys. **25**(3), 190–232 (1972)
65. C. Heuberger, S. Wagner, The number of maximum matchings in a tree. Discrete Math. **311**(21), 2512–2542 (2011)
66. H. Hosoya, Topological index. A newly proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons. Bull. Chem. Soc. Jpn. **44**(9), 2332–2339 (1971)
67. Y. Hou, On acyclic systems with minimal Hosoya index. Discrete Appl. Math. **119**(3), 251–257 (2002)
68. Y. Hu, X. Li, Y. Yuan, Trees with minimum general Randic index. MATCH Commun. Math. Comput. Chem. **52**, 119–128 (2004)
69. Y. Hu, X. Li, Y. Yuan, Solutions to two unsolved questions on the best upper bound for the Randic index R_{-1} of trees. MATCH Commun. Math. Comput. Chem. **54**, 441–454 (2005)
70. A. Ilić, On the extremal properties of the average eccentricity. Comput. Math. Appl. **64**(9), 2877–2885 (2012)
71. O. Ivanciuc, T.S. Balaban, A.T. Balaban, Reciprocal distance matrix, related local vertex invariants and topological indices. J. Math. Chem. **12**, 309–318 (1993)
72. R.E. Jamison, On the average number of nodes in a subtree of a tree. J. Comb. Theory Ser. B **35**(3), 207–223 (1983)

73. R.E. Jamison, Monotonicity of the mean order of subtrees. J. Comb. Theory Ser. B **37**(1), 70–78 (1984)
74. S. Janson, The Wiener index of simply generated random trees. Random Struct. Algoritm. **22**(4), 337–358 (2003)
75. S. Janson, P. Chassaing, The center of mass of the ISE and the Wiener index of trees. Electron. Commun. Probab. **9**, 178–187 (2004)
76. F. Jelen, E. Triesch, Superdominance order and distance of trees with bounded maximum degree. Discrete Appl. Math. **125** (2–3), 225–233 (2003)
77. C. Jordan, Sur les assemblages de lignes. J. Reine Angew. Math. **70**, 185–190 (1869)
78. L.B. Kier, L.H. Hall, *Molecular Connectivity in Chemistry and Drug Research* (Academic Press, New York, 1976)
79. L.B. Kier, L.H. Hall, *Molecular Connectivity in Structure-Activity Analysis* (Wiley, New York, 1986)
80. G. Kirchoff, Über die Auflösung der Gleichungen, auf welche man bei der Untersuchung der linearen Vertheilung galvanischer Ströme gefürt wird. Ann. Phys. Chem. **72**, 497–508 (1847)
81. R. Kirk, H. Wang, Largest number of subtrees of trees with a given maximum degree. SIAM J. Discrete Math. **22**(3), 985–995 (2008)
82. P. Kirschenhofer, H. Prodinger, R.F. Tichy, Fibonacci numbers of graphs. III. Planted plane trees, in *Fibonacci Numbers and Their Applications (Patras, 1984)* (Reidel, Dordrecht, 1986). Math. Appl. **28**, 105–120
83. M. Klazar, Twelve countings with rooted plane trees. Eur. J. Comb. **18**(2), 195–210 (1997)
84. D.J. Klein, I. Lukovits, I. Gutman, On the definition of the hyper-Wiener index for cycle-containing structures. J. Chem. Inf. Comput. Sci. **35**, 50–52 (1995)
85. B. Knudsen, Optimal Multiple Parsimony Alignment with Affine Gap Cost Using a Phylogenetic Tree. Lecture Notes in Bioinformatics, vol. 2812 (Springer, New York, 2003), pp. 433–446
86. M. Krzywkowski, Trees having many minimal dominating sets. Inform. Process. Lett. **113**(8), 276–279 (2013)
87. M. Krzywkowski, Minimal 2-dominating sets in trees. RAIRO Theor. Inform. Appl. **47**(3), 235–240 (2013)
88. M. Krzywkowski, An algorithm for listing all minimal double dominating sets of a tree. Fund. Inform. **130**(4), 415–421 (2014)
89. M. Krzywkowski, S. Wagner, Trees having few total dominating sets. In preparation (2015)
90. H.-F. Law, On the number of independent sets in a tree. Electron. J. Comb. **17**(1), paper no. #P18 (2010)
91. M. Lepović, I. Gutman, A collective property of trees and chemical trees. J. Chem. Inf. Comput. Sci. **38**, 823–826 (1998)
92. V.E. Levit, E. Mandrescu, The independence polynomial of a graph—a survey, in *Proceedings of the 1st International Conference on Algebraic Informatics* (Aristotle University of Thessaloniki, Thessaloniki, 2005), pp. 233–254
93. S. Li, S. Wang, Further analysis on the total number of subtrees of trees. Electron. J. Comb. **19**(4), paper no. #P48 (2012)
94. X. Li, Y. Shi, A survey on the Randić index. MATCH Commun. Math. Comput. Chem. **59**(1), 127–156 (2008)
95. X. Li, I. Gutman, *Mathematical Aspects of Randić-type Molecular Structure Descriptors*. Mathematical Chemistry Monographs, vol. 1 (University of Kragujevac, Kragujevac, 2006)
96. X. Li, Z. Li, L. Wang, The inverse problems for some topological indices in combinatorial chemistry. J. Comput. Biol. **10** (1), 47–55 (2003)
97. X. Li, Y. Shi, I. Gutman, *Graph Energy* (Springer, New York, 2012)
98. V. Linek, Bipartite graphs can have any number of independent sets. Discrete Math. **76** (2), 131–136 (1989)
99. L. Lovász, *Combinatorial Problems and Exercises* (North-Holland, Amsterdam, 1979)
100. L. Lu, Y. Yang, A theorem on Randić index and the diameter of a graph. Discrete Math. **311**(14), 1333–1343 (2011)

101. A. Meir, J.W. Moon, On subtrees of certain families of rooted trees. Ars Comb. **16B**, 305–318 (1983)
102. A. Meir, J.W. Moon, On maximal independent sets of nodes in trees. J. Graph Theory **12**(2), 265–283 (1988)
103. R.E. Merrifield, H.E. Simmons, *Topological Methods in Chemistry* (Wiley, New York, 1989)
104. G.O. Munsonius, On the asymptotic internal path length and the asymptotic Wiener index of random split trees. Electron. J. Probab. **16** paper no. 35, 1020–1047 (2011)
105. G.O. Munsonius, On tail bounds for random recursive trees. J. Appl. Probab. **49**, 566–581 (2012)
106. R. Neininger, The Wiener index of random trees. Comb. Probab. Comput. **11**(6), 587–597 (2002)
107. V. Nikiforov, Walks and the spectral radius of graphs. Linear Algebra Appl. **418**(1), 257–268 (2006)
108. L. Pavlovic, M. Stojanovic, X. Li, More on "Solutions to two unsolved questions on the best upper bound for the Randić Index R_{-1} of trees". MATCH Commun. Math. Comput. Chem. **58**(1), 167–182 (2007)
109. D. Plavsić, S. Nikolić, N. Trinajstić, Z. Mihalić, On the Harary index for the characterization of chemical graphs. J. Math. Chem. **12**, 235–250 (1993)
110. H. Prodinger, R.F. Tichy, Fibonacci numbers of graphs. Fibonacci Quart. **20**(1), 16–21 (1982)
111. M. Randić, Characterization of molecular branching. J. Am. Chem. Soc. **97** (23), 6609–6615 (1975)
112. M. Randić, Novel molecular descriptor for structure-property studies. Chem. Phys. Lett. **211**, 478–483 (1993)
113. B.E. Sagan, A note on independent sets in trees. SIAM J. Discrete Math. **1**(1), 105–108 (1988)
114. B.E. Sagan, V.R. Vatter, Maximal and maximum independent sets in graphs with at most r cycles. J. Graph Theory **53**(4), 283–314 (2006)
115. N.S. Schmuck, S. Wagner, H. Wang, Greedy trees, caterpillars, and Wiener-type graph invariants. MATCH Commun. Math. Comput. Chem. **68**, 273–292 (2012)
116. E. Schroeder, Vier combinatorische Probleme. Z. f. Math. Phys. **15**, 361–376 (1870)
117. R. Shi, The average distance of trees. Syst. Sci. Math. Sci. **6**(1), 18–24 (1993)
118. A.V. Sills, H. Wang, On the maximal Wiener index and related questions. Discrete Appl. Math. **160**, 1615–1623 (2012)
119. A.V. Sills, H. Wang, The minimal number of subtrees of a tree. Graphs Comb. **31**(1), 255–264 (2015)
120. Z. Skupień, Majorization and the minimum number of dominating sets. Discrete Appl. Math. **165**, 295–302 (2014)
121. H. Smith, L.A. Székely, H. Wang, Eccentricity in trees (2014, submitted, arXiv:1408.5865)
122. H. Smith, L.A. Székely, H. Wang, S. Yuan, On different "middle parts" of a tree. In preparation (2015)
123. L.A. Székely, H. Wang, Binary trees with the largest number of subtrees with at least one leaf. Congr. Numer. **177**, 147–169 (2005),
124. L.A. Székely, H. Wang, On subtrees of trees. Adv. Appl. Math. **34**, 138–155 (2005)
125. L.A. Székely, H. Wang, Binary trees with the largest number of subtrees. Discrete Appl. Math. **155**(3), 374–385 (2006)
126. L.A. Székely, H. Wang, Extremal values of ratios: distance problems vs. subtree problems in trees. Electron. J. Comb. **20**(1), #P67, 1–20 (2013)
127. L.A. Székely, H. Wang, Extremal values of ratios: distance problems vs. subtree problems in trees II. Discrete Math. **322**, 36–47 (2014)
128. L.A. Székely, H. Wang, T. Wu, The sum of distances between the leaves of a tree and the 'semi-regular' property. Discrete Math. **311**, 1197–1203 (2011)
129. Y. Tang, B. Zhou, On Average Eccentricity. MATCH Commun. Math. Comput. Chem. **67**, 405–423 (2012)
130. H. Täubig, J. Weihmann, S. Kosub, R. Hemmecke, E.W. Mayr, Inequalities for the number of walks in graphs. Algorithmica **66**(4), 804–828 (2013)

131. H. Täubig, J. Weihmann, Matrix power inequalities and the number of walks in graphs. Discrete Appl. Math. **176**, 122–129 (2014)
132. R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors* (Wiley-VCH, Weinheim, 2000)
133. N. Trinajstić, *Chemical Graph Theory* (CRC Press, Boca Raton, 1992)
134. A. Vince, H. Wang, The average order of a subtree of a tree. J. Comb. Theory Ser. B **100**(2), 161–170 (2010)
135. S. Wagner, A class of trees and its Wiener index. Acta Appl. Math. **91**(2), 119–132 (2006)
136. S. Wagner, Correlation of graph-theoretical indices. SIAM J. Discrete Math. **21**(1), 33–46 (2007)
137. S. Wagner, On the average Wiener index of degree-restricted trees. Australas. J. Comb. **37**, 187–203 (2007)
138. S. Wagner, On the number of matchings of a tree. Eur. J. Comb. **28**(4), 1322–1330 (2007)
139. S. Wagner, Extremal trees with respect to Hosoya index and Merrifield-Simmons index. MATCH Commun. Math. Comput. Chem. **57**(1), 221–233 (2007)
140. S. Wagner, Almost all trees have an even number of independent sets. Electron. J. Comb. **16**(1) paper no. #93 (2009)
141. S. Wagner, A note on the number of dominating sets of a graph. Util. Math. **92**, 25–31 (2013)
142. S. Wagner, Central limit theorems for additive tree parameters with small toll functions. Comb. Probab. Comput. **24**(1), 329–353 (2015)
143. S. Wagner, I. Gutman, Maxima and minima of the Hosoya index and the Merrifield-Simmons index: a survey of results and techniques. Acta Appl. Math. **112**(3), 323–346 (2010)
144. S. Wagner, H. Wang, Indistinguishable trees and graphs. Graphs Comb. **30**(6), 1593–1605 (2014)
145. S. Wagner, H. Wang, On the local and global means of subtree orders. J. Graph Theory **85**(2), 154–166 (2016)
146. S. Wagner, H. Wang, G. Yu, Molecular graphs and the inverse Wiener index problem, Discrete Appl. Math. **157**(7), 1544–1554 (2009)
147. S. Wagner, H. Wang, X.D. Zhang, Distance-based graph invariants of trees and the Harary index. Filomat. **27**, 39–48 (2013)
148. H. Wang, Extremal trees with given degree sequence for the Randić index. Discrete Math. **308**, 3407–3411 (2008)
149. H. Wang, The extremal values of the Wiener index of a tree with given degree sequence. Discrete Appl. Math. **156**, 2647–2654 (2009)
150. H. Wang, Functions on adjacent vertex degrees of trees with given degree sequence. Cent. Eur. J. Math. **12**, 1656–1663 (2014)
151. H. Wang, The distances between internal vertices and leaves of a tree. Eur. J. Comb. **41**, 79–99 (2014)
152. H. Wang, G. Yu, All but 49 numbers are Wiener indices of trees. Acta Appl. Math. **92**(1) 15–20 (2006)
153. H. Wiener, Structural determination of paraffin boiling points. J. Am. Chem. Soc. **69**, 17–20 (1947)
154. H. Wiener, Correlation of heats of isomerization, and differences in heats of vaporization of isomers, among the paraffin hydrocarbons. J. Am. Chem. Soc. **69**, 2636–2638 (1947)
155. H.S. Wilf, The number of maximal independent sets in a tree. SIAM J. Algebraic Discrete Methods **7**(1), 125–130 (1986)
156. I. Włoch, Trees with extremal numbers of maximal independent sets including the set of leaves. Discrete Math. **308**(20), 4768–4772 (2008)
157. I. Włoch, A. Włoch, The number of independent sets intersecting the set of leaves in trees. Ars Comb. **85**, 225–231 (2007)
158. K. Xu, Trees with the seven smallest and eight greatest Harary indices. Discrete Appl. Math. **160**, 321–331 (2012)
159. K. Xu, M. Liu, K. Das, I. Gutman, B. Furtula, A survey on graphs extremal with respect to distance-based topological indices. MATCH Commun. Math. Comput. Chem. **71**, 461–508 (2014)

160. W. Yan, L. Ye, On the maximal energy and the Hosoya index of a type of trees with many pendant vertices. MATCH Commun. Math. Comput. Chem. **53**(2), 449–459 (2005)
161. P. Yu, An upper bound on the Randić index of trees, J. Math. Study **31**, 225–230 (1998) (in Chinese)
162. B. Zelinka, Medians and peripherans of trees. Arch. Math. (Brno) **4**, 87–95 (1968)
163. X.D. Zhang, Y. Liu, M. Han, The maximum Wiener index of trees with given degree sequences. MATCH Commun. Math. Comput. Chem. **64**, 661–682 (2010)
164. X.D. Zhang, Q. Xiang, L. Xu, R. Pan, The Wiener index of trees with given degree sequences. MATCH Commun. Math. Comput. Chem. **60**, 623–644 (2008)
165. X.M. Zhang, X.D. Zhang, The minimal number of subtrees with a given degree sequence. Graphs Comb. **31**(1), 309–318 (2015)
166. X.M. Zhang, X.D. Zhang, D. Gray, H. Wang, Trees with the most subtrees—an algorithmic approach. J. Comb. **3**(2), 207–223 (2012)
167. X.M. Zhang, X.D. Zhang, D. Gray, H. Wang, The number of subtrees of trees with given degree sequence. J. Graph Theory **73**(3), 280–295 (2013)
168. J. Zito, The structure and maximum number of maximum independent sets in trees. J. Graph Theory **15**(2), 207–221 (1991)

The edit distance in graphs: Methods, results, and generalizations

Ryan R. Martin

Abstract The edit distance is a very simple and natural metric on the space of graphs. In the edit distance problem, we fix a hereditary property of graphs and compute the asymptotically largest edit distance of a graph from the property. This quantity is very difficult to compute directly but in many cases, it can be derived as the maximum of the edit distance function. Szemerédi's regularity lemma, strongly regular graphs, constructions related to the Zarankiewicz problem – all these play a role in the computing of edit distance functions. The most powerful tool is derived from symmetrization, which we use to optimize quadratic programs that define the edit distance function. In this paper, we describe some of the most common tools used for computing the edit distance function, summarize the major current results, outline generalizations to other combinatorial structures, and pose some open problems.

Keywords Edit distance • Colored regularity graphs • Quadratic programming

Mathematics Subject Classification (2010): 05C35, 05C80

1 Introduction

The edit distance in graphs was originally studied to answer two different and independent problems: one to answer questions on property testing [5] and the other, to answer a question regarding consensus trees from evolutionary biology [10]. In metabolic networks, the presence or absence of edges in a certain graph corresponds to pairs of genes which activate or deactivate one another. In evolutionary theory, avoiding forbidden induced subgraphs [23] is studied, which is equivalent to a similar edit problem of bipartite graphs or matrices. Edit distance problems with respect to more general classes of graphs are important in the algorithmic aspects of property testing [3, 5–7] and in the techniques involved in computing the speed of dense graph properties [19, 46].

R.R. Martin (✉)
Iowa State University, Ames, IA 50011-2064, USA
e-mail: rymartin@iastate.edu

A. Beveridge et al. (eds.), *Recent Trends in Combinatorics*, The IMA Volumes
in Mathematics and its Applications 159, DOI 10.1007/978-3-319-24298-9_2

The (normalized) edit metric is a metric on the set of simple, labeled n-vertex graphs. The distance between two graphs is the symmetric difference of the edge sets divided by the total number of possible edges. If $\mathrm{dist}(G, G')$ denotes the edit distance between G and G' on the same labeled vertex set, then

$$\mathrm{dist}(G, G') = |E(G) \triangle E(G')| / \tbinom{n}{2}.$$

As with any metric, we may take a *property* of graphs $\mathscr{H}$ (that is, a set of graphs), and compute the distance of a graph from that property:

$$\mathrm{dist}(G, \mathscr{H}) = \min \left\{ \mathrm{dist}(G, G') : V(G') = V(G) \right\}. \tag{1}$$

The properties that we study in this paper are *hereditary properties*. A property of graphs is hereditary if is closed under isomorphism and deletion of vertices. Alon and Stav [5] suggest that "In fact, almost all interesting graph properties are hereditary." Planarity having chromatic number at most k or not having a given H as an induced subgraph all are commonly studied hereditary properties. The property of having no graph H as an induced subgraph is called a *principal hereditary property* and we denote it by $\mathrm{Forb}(H)$. For every hereditary property $\mathscr{H}$ there exists a family of graphs $\mathscr{F}(\mathscr{H})$ ("forbidden graphs") such that $\mathscr{H} = \bigcap_{H \in \mathscr{F}(H)} \mathrm{Forb}(H)$. A hereditary property is said to be *nontrivial* if there is an infinite sequence of graphs that is in the property.

In the seminal papers by Alon and Stav [4, 5] and by Axenovich, Kézdy, and Martin [10], the fundamental question was the maximum distance of a graph G on n vertices from hereditary property $\mathscr{H}$. In fact, the maximum distance is asymptotically the same as that of the Erdős-Rényi random graph $G(n, p)$, for some value of p.

Theorem 1 (Alon-Stav [5]) *Let $\mathscr{H}$ be an arbitrary graph property. There exists $p^* = p^*_{\mathscr{H}} \in [0, 1]$ such that*

$$\max \{\mathrm{dist}(G, \mathscr{H}) : |V(G)| = n\} = \mathbb{E}[\mathrm{dist}(G(n, p^*), \mathscr{H})] + o(1). \tag{2}$$

We denote the limit of the quantity in (2) by $d^*_{\mathscr{H}}$. This is, asymptotically, the maximum distance of a graph from $\mathscr{H}$. Although $d^*_{\mathscr{H}}$ is the quantity in which we are most interested, determining its value is most often done by generalizing the result in Theorem 1. We do so by instead finding the maximum edit distance of a density-p graph from $\mathscr{H}$, for **all** values of p.

Balogh and Martin [11] introduced the *edit distance function* of a hereditary property.

Definition 2 *Let $\mathscr{H}$ be a nontrivial hereditary property of graphs. The* edit distance function of *$\mathscr{H}$ is*

$$\mathrm{ed}_{\mathscr{H}}(p) := \lim_{n \to \infty} \max \left\{ \mathrm{dist}(G, \mathscr{H}) : |V(G)| = n, |E(G)| = \left\lfloor p \tbinom{n}{2} \right\rfloor \right\}. \tag{3}$$

The existence of the limit in (3) was proven in [11].[1]

Theorem 3 (Balogh-Martin [11]) *Let $\mathscr{H}$ be an arbitrary nontrivial graph property. Then*

$$\mathrm{ed}_{\mathscr{H}}(p) = \lim_{n\to\infty} \mathbb{E}[\mathrm{dist}(G(n,p),\mathscr{H})].$$

Theorems 1 and 3 make use of Szemerédi's regularity lemma [48] but in a way that detects induced subgraphs. The idea of applying Szemerédi's regularity lemma to hereditary properties has been studied in a number of contexts, including pioneering work by Prömel and Steger [40–42], Scheinerman and Zito [46], and Bollobás and Thomason [18–20]. The essential technique is to apply the regularity lemma twice – once to the graph itself and a second time to each of the graphs induced by the non-exceptional clusters. More directly, one can use a variant of Szemerédi's regularity lemma due to Alon et al. [7] that has been used in a number of papers, including the edit distance papers [5, 11].

The edit distance function is symmetric with respect to complementation. It is easy to see that $\mathrm{ed}_{\mathrm{Forb}(H)}(p) = \mathrm{ed}_{\mathrm{Forb}(\overline{H})}(1-p)$ and, in fact, a general case is true.

Proposition 4 *Let* $\mathscr{H} = \bigcap_{H\in\mathscr{F}(\mathscr{H})} \mathrm{Forb}(H)$ *be a nontrivial hereditary property and let* $\mathscr{H}^* = \bigcap_{H\in\mathscr{F}(\mathscr{H})} \mathrm{Forb}(\overline{H})$. *Then* $\mathrm{ed}_{\mathscr{H}}(p) = \mathrm{ed}_{\mathscr{H}^*}(1-p)$.

A very similar setting to the edit distance problem was studied by Richer [43] and as further investigated by Marchant and Thomason [31, 32] regarding the two-coloring of the edges of the complete graph. Many of the most vital results for solving the edit distance problem come from this setting. In solving the problems they pose on a hereditary property $\mathscr{H}$, they obtain the function $1 - \mathrm{ed}_{\mathscr{H}}(p)$. The connection between the two settings is addressed in [32] as well as by Thomason [49] in a survey.

2 Colored regularity graphs

The key observation in computing the edit distance is that a graph can be approximated by a graph-like structure in which the clusters either behave like cliques or independent sets and the ϵ-regular pairs either behave like complete bipartite graphs, empty bipartite graphs or random graphs with density bounded away from both 0 and 1.

Alon and Stav [5] defined a *colored regularity graph (CRG) K* to be a simple complete graph, together with a partition of the vertices into white and black,

[1] It should be noted that early papers on the edit distance do not normalize the distance. That is, the distance is merely $|E(G)\Delta E(G')|$. Normalization, however, is required in order to define the edit distance function and it seems most natural to put the normalization in the metric itself, rather than doing so in order to define $\mathrm{ed}_{\mathscr{H}}$.

$V(K)=\mathrm{VW}(K)\cup\mathrm{VB}(K)$ and a partition of the edges into white, gray, and black, $E(K)=\mathrm{EW}(K)\cup\mathrm{EG}(K)\cup\mathrm{EB}(K)$.[2]

We say that a graph H embeds in K, writing $H\mapsto K$, if there is a function $\varphi: V(H)\to V(K)$ so that the following occur:

- If $h_1h_2\in E(H)$, then either $\varphi(h_1)=\varphi(h_2)\in\mathrm{VB}(K)$ or $\varphi(h_1)\varphi(h_2)\in\mathrm{EB}(K)\cup\mathrm{EG}(K)$.
- If $h_1h_2\notin E(H)$, then either $\varphi(h_1)=\varphi(h_2)\in\mathrm{VW}(K)$ or $\varphi(h_1)\varphi(h_2)\in\mathrm{EW}(K)\cup\mathrm{EG}(K)$.

A CRG K' is said to be a *sub-CRG* of K if K' can be obtained by deleting vertices of K and is a *proper sub-CRG* if $K'\neq K$.

If a graph H embeds in CRG K, then a large enough graph that is approximated by K will have an induced copy of H. This is stated and proven more formally in Section 4 of [11]. However, the main idea is that for any large graph in a hereditary property $\mathscr{H}=\bigcap_{H\in\mathscr{F}(\mathscr{H})}\mathrm{Forb}(H)$, the CRG K that approximates the graph satisfies the property that $H\not\mapsto K$ for all $H\in\mathscr{F}(\mathscr{H})$. We denote $\mathscr{K}(\mathscr{H})$ to be the subset of CRGs K such that no forbidden graph maps into K. Formally, $\mathscr{K}(\mathscr{H})=\{K : H\not\mapsto K, \forall H\in\mathscr{F}(\mathscr{H})\}$.

2.1 The *f* and *g* functions

There is a matrix associated with a CRG called $\mathbf{M}_K(p)$ that plays a role similar to the role the adjacency matrix does for graphs. We can use this matrix to help define the functions f_K and g_K, that are essential for understanding edit distance.

Definition 5 *Let K be a CRG on vertex set $\{v_1,\ldots,v_k\}$ with* VW *and* VB *denoting the white and black vertices, respectively, and* EW, EG, *and* EB *denoting the white, gray, and black edges, respectively. Let $\mathbf{M}_K(p)$ denote the matrix with entries defined as follows:*

$$m_K(p)_{ij}=\begin{cases}p, & \text{if } i\neq j \text{ and } v_iv_j\in\mathrm{EW} \text{ or } i=j \text{ and } v_i\in\mathrm{VW};\\ 0, & \text{if } i\neq j \text{ and } v_iv_j\in\mathrm{EG};\\ 1-p, & \text{if } i\neq j \text{ and } v_iv_j\in\mathrm{EB} \text{ or } i=j \text{ and } v_i\in\mathrm{VB}.\end{cases}\tag{4}$$

The functions f_K and g_K are defined as follows:

$$f_K(p)=\frac{1}{k^2}\left[p(|\mathrm{VW}|+2|\mathrm{EW}|)+(1-p)(|\mathrm{VB}|+2|\mathrm{EB}|)\right]=\frac{1}{k^2}\mathbf{1}^T\mathbf{M}_K(p)\mathbf{1}\tag{5}$$

[2] Papers by Bollobás and Thomason [18–20] and others such as [49] use the term “type” rather than CRG.

$$g_K(p) = \min\left\{\mathbf{x}^T\mathbf{M}_K(p)\mathbf{x} : \mathbf{x}^T\mathbf{1} = 1, \mathbf{x} \geq \mathbf{0}\right\}. \tag{6}$$

The vector $\mathbf{0}$ *is the all-zeroes vector,* $\mathbf{1}$ *is the all-ones vector, and vector inequalities are coordinatewise.*

Clearly, for any CRG K and any $p \in [0,1]$, we have $g_K(p) \leq f_K(p)$. Although the linearity of f_K makes proving general results about $\mathrm{ed}_{\mathcal{H}}$ possible, the g function is more useful in computing the edit distance function. In fact, if an optimal vector of (6) has a zero entry, we may obtain K' by deleting the corresponding entry and achieve $g_{K'}(p) = g_K(p)$. We say that a CRG K is *p-core* if, for any proper sub-CRG K' of K, we have $g_{K'}(p) > g_K(p)$.

The edit distance function can be defined in terms of the f and g functions:

Theorem 6 *Let* $\mathcal{H}$ *be a nontrivial hereditary property. For any* $p \in [0,1]$*,*

$$\begin{aligned} \mathrm{ed}_{\mathcal{H}}(p) &= \inf\{f_K(p) : K \in \mathcal{K}(\mathcal{H})\} = \inf\{g_K(p) : K \in \mathcal{K}(\mathcal{H})\} & (7)\\ &= \min\{g_K(p) : K \in \mathcal{K}(\mathcal{H})\}. & (8)\end{aligned}$$

Equation (7) is due to Balogh and Martin [11]. Equation (8) is from the results of Marchant and Thomason [32] and gives rise to the question as to whether only a finite set of p-core CRGs is sufficient to define the edit distance function for any nontrivial hereditary property and all $p \in [0,1]$.

There is some evidence (see Theorem 31(d) and Theorem 30(b) below) that for some hereditary properties, determining the edit distance function requires knowledge of an infinite sequence of CRGs. Nonetheless, we believe that the bulk of the edit distance function can be determined from a finite number of CRGs. That is, for any $\epsilon > 0$, we believe a finite set of CRGs can simultaneously define $\mathrm{ed}_{\mathcal{H}}$ for all $p \in [\epsilon, 1-\epsilon]$. This is Conjecture 1 in Section 8.2.

2.2 Clique spectrum

Certain colored regularity graphs play a key role in the computation of the edit distance. A *gray-edge CRG* is the CRG K with all $\binom{|V(K)|}{2}$ edges gray. The gray-edge CRG with r white vertices and s black vertices is denoted $K(r,s)$. The *clique spectrum* of $\mathcal{H}$ is the set

$$\Gamma(\mathcal{H}) \stackrel{\mathrm{def}}{=} \{(r,s) : H \not\mapsto K(r,s), \forall H \in \mathcal{F}(\mathcal{H})\}.$$

For example, if $\mathcal{H} = \mathrm{Forb}(H)$ is a hereditary property, then pairs (r,s) are in the clique spectrum of $\mathrm{Forb}(H)$ if and only if H **cannot** be partitioned into r independent sets and s cliques.

The clique spectrum has a number of useful properties. For example, it is monotone in the sense that if $(r,s) \in \Gamma(\mathcal{H})$ and $0 \leq r' \leq r$ and $0 \leq s' \leq s$, then $(r',s') \in \Gamma(\mathcal{H})$. As a result, the clique spectrum of a hereditary property can

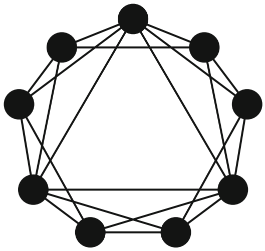

Fig. 1 A graph H_9 on 9 vertices.

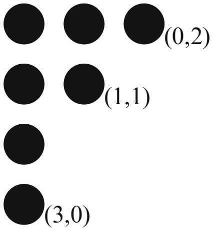

Fig. 2 The Ferrers diagram of the clique spectrum of H_9 with the extreme points labeled.

be expressed as a Ferrers diagram. An *extreme point* of the clique spectrum Γ is a pair $(r,s)\in\Gamma$ for which both $(r+1,s)\notin\Gamma$ and $(r,s+1)\notin\Gamma$. Figure 1 shows the graph H_9, and Figure 2 shows its clique spectrum, expressed as a Ferrers diagram.

Since the matrix $\mathbf{M}_{K(r,s)}(p)$ is a diagonal matrix with r entries of value p and s entries with value $1-p$, it is easy to compute that for all $p\in(0,1)$

$$g_{K(r,s)}(p)=\left(\frac{r}{p}+\frac{s}{1-p}\right)^{-1}=\frac{p(1-p)}{r(1-p)+sp}. \tag{9}$$

We have the natural convention that if $r=0$ then $g_{K(r,s)}(0)=1$ and if $s=0$ then $g_{K(r,s)}(1)=1$.

In fact, we have a more general way of computing the edit distance function if the matrix $\mathbf{M}_K(p)$ is block diagonal matrix, where the blocks correspond to a CRG notion of components.

Definition 7 *A sub-CRG, K', of a CRG K is a* component *if it is maximal with respect to the property that, for all $v,w\in V(K')$, there exists a path consisting of white and black edges entirely within K'.*

More simply, components of K are the components of the graph G with vertex set $V(K)$ and the nonedges of G the gray edges of K. This leads to the generalization of (9).

Proposition 8 (Martin [33]) *Let K be a CRG with components $K_{(1)},\ldots,K_{(\ell)}$. Then*

$$(g_K(p))^{-1}=\sum_{i=1}^{\ell}(g_{K_{(i)}}(p))^{-1}.$$

2.3 Characterization of p-core CRGs

Marchant and Thomason [32] gave a characterization of all p-core CRGs.

Theorem 9 (Marchant-Thomason [32]) *Let K be a p-core CRG.*

(a) If $p \leq 1/2$, then there are no black edges and the white edges are only incident to black vertices.
(b) If $p \geq 1/2$, then there are no white edges and the black edges are only incident to white vertices.

Consequently, if $p = 1/2$, then all edges are gray.

Theorem 9 is an essential tool and is used in most results on the edit distance function as we shall see below.

3 Estimating the edit distance function

Although it is difficult to compute the edit distance function for general hereditary properties, we can estimate the function through a variety of techniques. First, we can use the clique spectrum and (9) to construct an upper bound.

3.1 Upper bound via the clique spectrum

We begin with a trio of results with elementary proofs, followed by a result concerning the nature of the edit distance function.

Theorem 10 *Let $\mathscr{H}$ be a nontrivial hereditary property and let $\Gamma(\mathscr{H})$ denote the clique spectrum of $\mathscr{H}$. If we define*

$$\gamma_{\mathscr{H}}(p) := \min_{(r,s)\in\Gamma(\mathscr{H})} g_{K(r,s)}(p) = \min_{(r,s)\in\Gamma(\mathscr{H})} \frac{p(1-p)}{r(1-p)+sp}, \tag{10}$$

then $\mathrm{ed}_{\mathscr{H}}(p) \leq \gamma_{\mathscr{H}}(p)$.

There are three (not necessarily distinct) extreme points of a clique spectrum that are of particular interest. First, if $(r,0) \in \Gamma(\mathscr{H})$ but $(r+1,0) \notin \Gamma(\mathscr{H})$, then $r+1$ is the *chromatic number* of $\mathscr{H}$, denoted $\chi(\mathscr{H})$ or just χ, when the hereditary property is understood. Second, if $(0,s) \in \Gamma(\mathscr{H})$ but $(0,s+1) \notin \Gamma(\mathscr{H})$, then $s+1$ is the *complementary chromatic number* of $\mathscr{H}$, denoted $\overline{\chi}(\mathscr{H})$ or just $\overline{\chi}$. Note that if $\mathscr{H} = \mathrm{Forb}(H)$ for some graph H, then $\chi(\mathscr{H}) = \chi(H)$ and $\overline{\chi}(\mathscr{H}) = \chi(\overline{H})$.

We observe that if $\chi(\mathscr{H}) \geq 2$ then $(\chi-1,0) \in \Gamma(\mathscr{H})$ and if $\overline{\chi}(\mathscr{H}) \geq 2$ then $(0,\overline{\chi}-1) \in \Gamma(\mathscr{H})$. Therefore, we have the following corollary.

Corollary 11 *Let $\mathscr{H}$ be a nontrivial hereditary property with chromatic number χ and complementary chromatic number $\overline{\chi}$.*

(a) If $\chi \geq 2$, then $\mathrm{ed}_{\mathscr{H}}(p) \leq p/(\chi - 1)$.
(b) If $\overline{\chi} \geq 2$, then $\mathrm{ed}_{\mathscr{H}}(p) \leq (1-p)/(\overline{\chi} - 1)$.

The chromatic and complementary chromatic numbers of a hereditary property $\mathscr{H}$ can be defined in terms of $\mathscr{F}(\mathscr{H})$ as in Proposition 12.

Proposition 12 *Let $\mathscr{H} = \bigcap_{H \in \mathscr{F}} \mathrm{Forb}(H)$ be a nontrivial hereditary property. Then,*

(a) $\chi(\mathscr{H}) = \min\{\chi(H) : H \in \mathscr{F}\}$ and
(b) $\overline{\chi}(\mathscr{H}) = \min\{\chi(\overline{H}) : H \in \mathscr{F}\}$.

The third extreme point we address is evaluated as follows: the largest value of $r + s + 1$ such that $(r, s) \in \Gamma(\mathscr{H})$ is called the *binary chromatic number* of $\mathscr{H}$ and is denoted $\chi_B(\mathscr{H})$ or just χ_B. This quantity has appeared in the literature previously. Prömel and Steger [40–42] called $\chi_B - 1$ simply τ. Bollobás and Thomason [18, 20] called $\chi_B - 1$ the *coloring number*.

Since Theorem 9 establishes that every 1/2-core CRG is a gray-edge CRG (that is, of the form $K(r, s)$) we can compute $\mathrm{ed}_{\mathscr{H}}(1/2)$ in terms of $\chi_B(\mathscr{H})$. Combining this with other basic facts, we obtain Theorem 13.

Theorem 13 *Let $\mathscr{H}$ be a nontrivial hereditary property.*

(a) $\mathrm{ed}_{\mathscr{H}}(p)$ is continuous.
(b) $\mathrm{ed}_{\mathscr{H}}(p)$ is concave down.
(c) $\mathrm{ed}_{\mathscr{H}}(1/2) = \frac{1}{2(\chi_B(\mathscr{H})-1)}$.

Theorem 13(a) was established by Marchant and Thomason [32]. We note that a different, analysis-based proof of this is in [11]. Theorem 13(b) was proven in [11]. Theorem 13(c) was proven in [10] in the case where $\mathscr{H}$ is a principal hereditary property. More sophisticated knowledge of the edit distance function has made it a simple corollary.

Using only Corollary 11 and Theorem 13 we can already find edit distance functions for some important hereditary properties. If P_4 denotes the path on 4 vertices, then $\mathrm{ed}_{\mathrm{Forb}(P_4)}(p) = \min\{p, 1-p\}$. If C_5 denotes the cycle on 5 vertices, then $\mathrm{ed}_{\mathrm{Forb}(C_5)}(p) = \frac{1}{2}\min\{p, 1-p\}$. More about hereditary properties forbidding self-complementary graphs is below in Corollary 15.

Because $\mathrm{ed}_{\mathscr{H}}(p)$ is continuous and concave down, the function achieves its maximum on the interval $[0, 1]$. Thus, both $d^*_{\mathscr{H}}$ and $p^*_{\mathscr{H}}$ are well-defined and the coordinate $(p^*_{\mathscr{H}}, d^*_{\mathscr{H}})$ is the point at which $\mathrm{ed}_{\mathscr{H}}$ achieves its maximum value.

Note: Although $p^*_{\mathscr{H}}$ is formally defined to be a closed interval, in all but a few (very) interesting cases[3] the interval is degenerate. That is, $p^*_{\mathscr{H}}$ is usually a single value. We will often abuse notation and terminology by referring to $p^*_{\mathscr{H}} = p$, rather than $p^*_{\mathscr{H}} = [p, p]$ and instead indicate explicitly where the interval is not degenerate.

[3] See, e.g., Section 5.5.2.

3.2 Upper bound using $\chi_B(\mathscr{H})$

If $\mathscr{H} = \bigcap_{H\in\mathscr{F}} \mathrm{Forb}(H)$ is a hereditary property such that $\mathscr{F}$ contains no complete graph and no empty graph, then it is trivial that $\mathrm{ed}_{\mathscr{H}}(0) = \mathrm{ed}_{\mathscr{H}}(1) = 0$. Indeed, Proposition 12 gives that $\chi(\mathscr{H}) \geq 2$ and $\overline{\chi}(\mathscr{H}) \geq 2$. The statement then follows from the simple bounds given by Corollary 11.

Using only the $\gamma_{\mathscr{H}}$ function, we may narrow down the possible values for $p^*_{\mathscr{H}}$ and for $d^*_{\mathscr{H}}$.

Theorem 14 *Let* $\mathscr{H} = \bigcap_{H\in\mathscr{F}} \mathrm{Forb}(H)$ *with* $(r,s) \in \Gamma(\mathscr{H})$ *such that* $r+s = \chi_B(\mathscr{H}) - 1$.

(a) $\mathrm{ed}_{\mathscr{H}}(p) \leq \gamma_{\mathscr{H}}(p) \leq \frac{p(1-p)}{r(1-p)+sp}$ *for all* $p \in [0,1]$.
(b) $d^*_{\mathscr{H}} \leq \frac{1}{r+s+2\sqrt{rs}}$.
(c) $\mathrm{ed}_{\mathscr{H}}(p) \geq \min\left\{\frac{p}{\chi_B(\mathscr{H})-1}, \frac{1-p}{\chi_B(\mathscr{H})-1}\right\}$.
(d) *If* $r \leq s$, *then* $p^*_{\mathscr{H}} \in \left[\frac{r}{r+s}, \frac{1}{2}\right]$.
(e) *If* $s \leq r$, *then* $p^*_{\mathscr{H}} \in \left[\frac{1}{2}, \frac{r}{r+s}\right]$.

Theorem 14(a) comes from Theorem 10. Theorem 14(b) is simply the maximum value of $g_{K(r,s)}(p)$. Theorem 14(c) follows from Theorem 13 – continuity, concavity, and the value of $\mathrm{ed}_{\mathscr{H}}(1/2)$ – and the fact that $\mathrm{ed}_{\mathscr{H}}(0) = \mathrm{ed}_{\mathscr{H}}(1) = 0$. Theorem 14(d) and (e) follow from the fact that these are the intervals over which $g_{K(r,s)}(p) \geq 1/(2(\chi_B(\mathscr{H}) - 1))$.

Corollary 15 gives the values of $p^*_{\mathscr{H}}$ and of $d^*_{\mathscr{H}}$ if $\mathscr{H} = \mathrm{Forb}(H)$ for a self-complementary graph H.

Corollary 15 *Let* $\mathscr{H} = \bigcap_{H\in\mathscr{F}} \mathrm{Forb}(H)$ *with* $(r_1, s_1) \in \Gamma(\mathscr{H})$ *and* $(r_2, s_2) \in \Gamma(\mathscr{H})$ *(not necessarily distinct) such that* $r_1 + s_1 = r_2 + s_2 = \chi_B(\mathscr{H}) - 1$, $r_1 \leq s_1$, *and* $r_2 \geq s_2$. *Then*

$$p^*_{\mathscr{H}} = 1/2 \qquad \textit{and} \qquad d^*_{\mathscr{H}} = 1/(2(\chi_B(\mathscr{H}) - 1)).$$

In particular, if $\mathscr{H} = \mathrm{Forb}(H)$, *where* H *is a self-complementary graph, then* $p^*_{\mathscr{H}} = 1/2$ *and* $d^*_{\mathscr{H}} = 1/(2(\chi_B(H) - 1))$.

4 Symmetrization

The most powerful tool for determining the edit distance function of a hereditary property is called symmetrization. This is a term Pikhurko [38] used for a method due to Sidorenko [47]. In fact, symmetrization can be traced back to Zykov [50] and his proof of Turán's theorem. Our version of symmetrization comes directly from the matrix defined by a CRG.

Theorem 16 (Martin [33]) *Let $p \in [0,1]$ and let K be a p-core CRG with associated matrix $\mathbf{M}_K(p)$, as defined in (4). If $\mathbf{x}^*$ is an optimal solution of the quadratic program from (6), namely that $\mathbf{x}^* \geq \mathbf{0}$, $\mathbf{x}^*\mathbf{1} = 1$, and $g_K(p) = (\mathbf{x}^*)^T \mathbf{M}_K(p)\mathbf{x}^*$, then*

$$\mathbf{M}_K(p) \cdot \mathbf{x}^* = g_K(p)\mathbf{1}.$$

In addition, by virtue of K being p-core, the vector $\mathbf{x}^*$ has no zero entries and $\mathbf{x}^*$ is unique for any fixed labeling of the vertices of K.

4.1 The weighted gray degree of a vertex

In order to interpret Theorem 16, we define the *white neighborhood* of vertex v in CRG K to be $N_W(v) := \{v' \in V(K) : vv' \in \mathrm{EW}(K)\} \cup \{v : \text{if } v \in \mathrm{VW}(K)\}$. The *black neighborhood* of v is $N_B(v) := \{v' \in V(K) : vv' \in \mathrm{EB}(K)\} \cup \{v : \text{if } v \in \mathrm{VB}(K)\}$. The *gray neighborhood* of v is $N_G(v) := \{v' \in V(K) : vv' \in \mathrm{EG}(K)\}$.

If $\mathbf{x}$ is the optimum weight vector in the quadratic program (6) that defines $g_K(p)$, then the *weighted white degree* of vertex $v \in V(K)$ is $\mathrm{d}_W(v) := \sum_{v' \in N_W(v)} \mathbf{x}(v')$. The *weighted black degree* of vertex $v \in V(K)$ is $\mathrm{d}_B(v) := \sum_{v' \in N_B(v)} \mathbf{x}(v')$. The *weighted gray degree* of vertex $v \in V(K)$ is $\mathrm{d}_G(v) := \sum_{v' \in N_G(v)} \mathbf{x}(v')$.

Theorem 16 gives that, for any $v \in \mathrm{VW}(K)$,

$$p\mathrm{d}_W(v) + (1-p)\mathrm{d}_B(v) = g_K(p). \tag{11}$$

Using the characterization of p-core CRGs from Theorem 9, we can apply (11) to compute the gray degree of each vertex.

Theorem 17 (Martin [33]) *Let $p \in (0,1)$ and K be a p-core CRG with optimum weight vector $\mathbf{x}$.*

(a) If $p \leq 1/2$, then $\mathbf{x}(v) = g_K(p)/p$ for all $v \in \mathrm{VW}(K)$ and

$$\mathrm{d}_G(v) = \frac{p - g_K(p)}{p} + \frac{1-2p}{p}\mathbf{x}(v), \qquad \textit{for all } v \in \mathrm{VB}(K).$$

(b) If $p \geq 1/2$, then $\mathbf{x}(v) = g_K(p)/(1-p)$ for all $v \in \mathrm{VB}(K)$ and

$$\mathrm{d}_G(v) = \frac{1-p-g_K(p)}{1-p} + \frac{2p-1}{1-p}\mathbf{x}(v), \qquad \textit{for all } v \in \mathrm{VW}(K).$$

Most of the results below use Theorem 17 as a primary tool. Intuitively, if $g_K(p)$ is small, then $\mathrm{d}_G(v)$ is large for each vertex and so K has a large amount of gray.

However, if K has too much gray, then some $H \in \mathscr{F}(\mathscr{H})$ would map to K, which contradicts the choice of $K \in \mathscr{K}(\mathscr{H})$. This general paradigm is made more precise by knowing more about the structure of the CRGs $K \in \mathscr{K}(\mathscr{H})$.

4.2 Basic structural facts of p-core CRGs

We can use Theorem 17 to obtain some basic helpful results on certain types of CRGs:

Corollary 18 *Let $t \geq 2$ and $k \geq 2$ be integers.*

(a) Let $p \leq 1/2$ and let K be a p-core CRG on k black vertices.

(i) If K has no gray edges, then $g_K(p) = \frac{1}{k}\left[1 + (k-2)p\right]$.
(ii) If K has no gray clique of order t, then $g_K(p) > p/(t-1)$.

(b) Let $p \geq 1/2$ and let K be a p-core CRG on k white vertices.

(i) If K has no gray edges, then $g_K(p) = \frac{1}{k}\left[1 + (k-2)(1-p)\right]$.
(ii) If K has no gray clique of order t, then $g_K(p) > (1-p)/(t-1)$.

Proof By symmetry, it is sufficient to prove (a). For (a)(i) we observe that, by Theorem 9(a), all edges are white and it is easy to see that the optimum weight vector in equation (6) is constant. Thus, all vertices have the same weight and the result follows.

For (a)(ii), we use a well-worn trick, used, e.g., in [35]. Let the maximum-sized clique of K (in terms of the number of vertices) be on vertex set $\{v_1, \ldots, v_c\}$ where $c \geq 2$. For every $w \notin \{v_1, \ldots, v_c\}$ we know that wv_i is a gray edge for at most $c-1$ values of i. Using Theorem 17(a), we have

$$\sum_{i=1}^{c}\left(\frac{p - g_K(p)}{p} + \frac{1-2p}{p}\mathbf{x}(v_i)\right) \leq (c-1)\left(1 - \sum_{i=1}^{c}\mathbf{x}(v_i)\right)$$

$$(c - 3 + 1/p)\sum_{i=1}^{c}\mathbf{x}(v_i) + 1 \leq \frac{c}{p}g_K(p).$$

Since $c - 3 + 1/p \geq c - 1 > 0$, we can conclude that $g_K(p) > p/c \geq p/(t-1)$, which concludes the proof. □

Remark 19 *The bound in Corollary 18(a)(ii) can be approached by a CRG on black vertices where the gray edges induce a blow-up of K_{t-1}.*

5 Known results

5.1 Hereditary properties that forbid either a complete or an empty graph

If $\mathscr{H} \subseteq \mathrm{Forb}(K_h)$, then $\mathrm{ed}_{\mathscr{H}}(1) > 0$ and, by Proposition 4, if $\mathscr{H} \subseteq \mathrm{Forb}(\overline{K_h})$, then $\mathrm{ed}_{\mathscr{H}}(0) > 0$. We can produce bounds on the edit distance function for such properties.

Theorem 20 (Martin [33]) *Let $\mathscr{H} = \bigcap_{H\in\mathscr{F}(\mathscr{H})} \mathrm{Forb}(H)$ be a nontrivial hereditary property with $\mathscr{H} \subseteq \mathrm{Forb}(K_h)$ for some $h \geq 2$ such that*

- *m is the least positive integer such that $\mathscr{F}(\mathscr{H})$ contains a complete multipartite graph with m parts, and*
- *χ is the chromatic number of $\mathscr{H}$.*

Note $\chi \geq 2$ because $\mathscr{H}$ is nontrivial and $\chi \leq m \leq h$. Then

(a) $\mathrm{ed}_{\mathscr{H}}(p) = \frac{p}{\chi-1}$, *for all* $p \in [0, 1/2]$, *and*
(b) $\frac{1-p}{\chi-1} + \frac{2p-1}{m-1} \leq \mathrm{ed}_{\mathscr{H}}(p) \leq \min\left\{1 - p + \frac{2p-1}{m-1}, \frac{p}{\chi-1}\right\}$, *for all* $p \in [1/2, 1]$.

In particular, if $\mathscr{H} = \mathrm{Forb}(K_h)$, then $\mathrm{ed}_{\mathscr{H}}(p) = \frac{p}{h-1}$.

Remark 21 *The bound $\mathrm{ed}_{\mathscr{H}}(p) \leq p/(\chi - 1)$ in Theorem 20 was not expressed explicitly in [33] but follows directly from the concavity of the edit distance function. By Proposition 4, there are similar bounds for $\mathscr{H}$ where $\mathscr{H} \subseteq \mathrm{Forb}(\overline{K_h})$. Consequently, $\mathrm{ed}_{\mathrm{Forb}(\overline{K_h})}(p) = (1 - p)/(h - 1)$.*

5.2 C_6^* and H_9

5.2.1 Forb(C_6^*)

In [32], Marchant and Thomason address the graph C_6^*, which is the 6-cycle with a diagonal. The extreme points of the clique spectrum of $\mathrm{Forb}(C_6^*)$ are $(1, 1)$ and $(0, 2)$. Thus, if $\mathscr{H} = \mathrm{Forb}(C_6^*)$, then $\gamma_{\mathscr{H}}(p) = \min\{p(1 - p), (1 - p)/2\}$.

In fact, the edit distance function has a smaller value for $p \in (0, 1)$.

Theorem 22 (Marchant-Thomason [32]) *Let $\mathscr{H} = \mathrm{Forb}(C_6^*)$, where C_6^* is the 6-cycle with a diagonal.*

(a) $\mathrm{ed}_{\mathscr{H}}(p) = \min\left\{\frac{p}{1+2p}, \frac{1-p}{2}\right\}$, *for* $p \in [0, 1]$.
(b) $p^*_{\mathscr{H}} = 1/2$ *and* $d^*_{\mathscr{H}} = 1/4$.

The CRG that corresponds to the $p/(1 + 2p)$ part of the function has 1 white vertex, 2 black vertices, one white edge between the black vertices, and two gray edges incident to the white vertex. See Figure 3. Although the edit distance function cannot

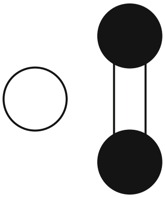

Fig. 3 The 3-vertex CRG that gives $p/(1+2p)$ in Theorem 22. The white edge is indicated, the two gray edges are not.

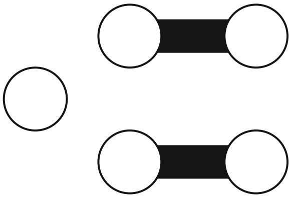

Fig. 4 The 5-vertex CRG that gives $p/(1+4p)$ in Theorem 23. The two black edges are indicated, the eight gray edges are not.

be determined by the clique spectrum, the values of $p^*_{\mathscr{H}} = 1/2$ and $d^*_{\mathscr{H}} = 1/4$ can be computed by knowing only the clique spectrum.

Using the tools in Section 4, the proof of Theorem 22 is much easier than the original proof.

Proof By Theorem 13, we may use continuity, concavity, and the knowledge of the value at $p = 1/2$ to conclude that $\mathrm{ed}_{\mathscr{H}}(p) = (1-p)/2$ for all $p \in [1/2, 1]$. Let $p \in [0, 1/2)$ and K be a p-core CRG. If $C_6^* \not\mapsto K$ and K has no white vertices, then it has no gray triangle and Corollary 18(a)(ii) gives that $g_K(p) > p/2$. If $C_6^* \not\mapsto K$ and K has one white vertex, then there are at most two black vertices, which cannot have a gray edge. Corollary 18(a)(ii) and Proposition 8 give that $g_K(p) \geq \left(p^{-1} + (1/2)^{-1}\right)^{-1} = p/(1+2p)$, as required. □

5.2.2 Forb(H_9)

Balogh and Martin [11] introduced the graph H_9, which is drawn in Figure 1. For $\mathscr{H} = \mathrm{Forb}(H_9)$, the values of $p^*_{\mathscr{H}}$ and $d^*_{\mathscr{H}}$ cannot be determined by the clique spectrum and this was established in [11]. Later, the author determined the edit distance function completely.

Theorem 23 (Martin [34]) *Let H_9 be the graph drawn in Figure 1 and let $\mathscr{H} =$* $\mathrm{Forb}(H_9)$.

(a) $\mathrm{ed}_{\mathscr{H}}(p) = \min\left\{\frac{p}{3}, \frac{p}{1+4p}, \frac{1-p}{2}\right\}$ *for* $p \in [0,1]$.
(b) $p^*_{\mathscr{H}} = \frac{1}{8}(1+\sqrt{17})$ *and* $d^*_{\mathscr{H}} = \frac{1}{8}(7-\sqrt{17})$.

The CRG that corresponds to the $p/(1+4p)$ part of the function has 5 white vertices, 2 nonadjacent white edges, and the remaining 8 edges gray. See Figure 4.

5.3 Cycles

The case of $\mathrm{Forb}(C_h)$, where C_h is a cycle on $h \geq 3$ vertices, has been widely investigated. Theorem 20 gives immediately that $\mathrm{ed}_{\mathrm{Forb}(C_3)}(p) = \mathrm{ed}_{\mathrm{Forb}(K_3)}(p) = p/2$.

In her Master's thesis, Peck almost completely settled the edit distance function for hereditary properties that forbid a cycle. Utilizing techniques inspired by the cycle arguments of Pósa [39], she determined the edit distance function for $\mathrm{Forb}(C_h)$ for odd $h \geq 5$. For even $h \geq 4$, she was able to determine enough of the function to find the maximum.

Theorem 24 (Peck [37]) *Let* $\mathscr{H} = \mathrm{Forb}(C_h)$ *where* C_h *is the cycle on* $h \geq 4$ *vertices.*

(a) If h is odd, then $\mathrm{ed}_{\mathscr{H}}(p) = \min\left\{\frac{p}{2}, \frac{p(1-p)}{1-p+(\lceil h/3\rceil-1)p}, \frac{1-p}{\lceil h/2\rceil-1}\right\}$ *for all* $p \in [0,1]$.

(b) If h is even, then $\mathrm{ed}_{\mathscr{H}}(p) = \min\left\{\frac{p(1-p)}{1-p+(\lceil h/3\rceil-1)p}, \frac{1-p}{\lceil h/2\rceil-1}\right\}$ *for all* $p \in [\lceil h/3\rceil^{-1}, 1]$.

Marchant and Thomason [32] first proved the case of $h = 4$ and, in fact, proved

Theorem 25 (Marchant-Thomason [32]) $\mathrm{ed}_{\mathrm{Forb}(C_4)}(p) = p(1-p)$ *for all* $p \in [0,1]$.

Marchant [31] proved the case for $h = 5, 7$. The cases of $h = 6, 8, 9, 10$ were first proven in [33] and, in fact, a larger range of p was proven in [33] for small even h.

Theorem 26 (Martin [33]) *Let* $\mathscr{H} = \mathrm{Forb}(C_h)$*, where* C_h *is the cycle on* $h \geq 4$ *vertices.*

(a) If $h = 6$*, then* $\mathrm{ed}_{\mathscr{H}}(p) = \min\left\{p(1-p), \frac{1-p}{2}\right\}$ *for all* $p \in [0,1]$.

(b) If $h = 8$*, then* $\mathrm{ed}_{\mathscr{H}}(p) = \min\left\{\frac{p(1-p)}{1+p}, \frac{1-p}{3}\right\}$ *for all* $p \in [0,1]$.

(c) If $h = 10$*, then* $\mathrm{ed}_{\mathscr{H}}(p) = \min\left\{\frac{p(1-p)}{1+2p}, \frac{1-p}{4}\right\}$ *for all* $p \in [1/7,1]$.

Corollary 27 *Let* $\mathscr{H} = \mathrm{Forb}(C_h)$ *where* C_h *is the cycle on* $h \geq 4$ *vertices.*

(a) If $h \not\in \{4,7,8,10,16\}$*, then*

$$p^*_{\mathscr{H}} = \frac{1}{\lceil h/2\rceil-\lceil h/3\rceil+1} \qquad \textit{and} \qquad d^*_{\mathscr{H}} = \frac{\lceil h/2\rceil-\lceil h/3\rceil}{(\lceil h/2\rceil-1)(\lceil h/2\rceil-\lceil h/3\rceil+1)}.$$

(b) If $h \in \{4, 7, 8, 10, 16\}$*, then*

$$p^*_{\mathscr{H}} = \frac{1}{1+\sqrt{\lceil h/3 \rceil - 1}} \qquad \textit{and} \qquad d^*_{\mathscr{H}} = \frac{1}{\lceil h/3 \rceil + 2\sqrt{\lceil h/3 \rceil - 1}}.$$

It is interesting that $p^*_{\mathrm{Forb}(C_h)}$ and $d^*_{\mathrm{Forb}(C_h)}$ are both rational and result from the intersection of the g functions of two p-core CRGs, except in the cases $h \in \{7, 8, 10, 16\}$.

5.4 Powers of cycles

A natural extension of hereditary properties defined by forbidding certain cycles are hereditary properties defined by forbidding certain powers of cycles. For $h \geq 2t+1$, we define C_h^t to be the graph with vertex set $\{1, \ldots, h\}$ and $ij \in E(C_h^t)$ if and only if $|i-j| \leq t \pmod{h}$. We consider the case for $t = 2$, that is, the case of the squared cycle.

For $h = 5$, C_5^2 is complete and from Theorem 20 we see that

$$\mathrm{ed}_{\mathrm{Forb}(C_5^2)}(p) = \mathrm{ed}_{\mathrm{Forb}(K_5)}(p) = p/4.$$

In the case of C_6^2, the complement is a perfect matching and we can use Proposition 4 if we know the edit distance function for $\mathrm{Forb}(M_6)$, where M_6 is the perfect matching on 6 vertices. It is easy to see that the CRGs $K \in \mathscr{F}(\mathrm{Forb}(M_6))$ have at most 2 black vertices and no pair of white vertices can have a gray edge between them, otherwise $M_6 \mapsto K$. By Theorem 9(a), we can conclude that $\mathrm{ed}_{\mathrm{Forb}(M_6)} = \frac{p(1-p)}{1+p}$ for $p \in [0, 1/2]$. Some more work verifies that $\mathrm{ed}_{\mathrm{Forb}(M_6)} = \frac{p(1-p)}{1+p}$ for $p \in [1/2, 1]$ also. Hence,

$$\mathrm{ed}_{\mathrm{Forb}(C_6^2)}(p) = \mathrm{ed}_{\mathrm{Forb}(M_6)}(1-p) = \frac{p(1-p)}{2-p}.$$

The complement of C_7^2 is simply C_7 and so Proposition 4 gives

$$\mathrm{ed}_{\mathrm{Forb}(C_7^2)}(p) = \mathrm{ed}_{\mathrm{Forb}(C_7)}(1-p) = \min\left\{\frac{p}{3}, \frac{p(1-p)}{2-p}, \frac{1-p}{2}\right\}.$$

Peck established some more values of $\mathrm{ed}_{\mathrm{Forb}(C_h^2)}(p)$ for $h \in \{8, 9, 10\}$, which we give in Theorem 28.

Theorem 28 (Peck [37]) *Let* $\mathscr{H} = \mathrm{Forb}(C_h^2)$ *where* C_h^2 *is the square of the cycle on h vertices.*

(a) If $\mathscr{H} = \mathrm{Forb}(C_8^2)$*, then* $\mathrm{ed}_{\mathscr{H}}(p) = \min\left\{\frac{p}{3}, \frac{p(1-p)}{2-p}, \frac{1-p}{2}\right\}$ *for all* $p \in [0, 1]$.

(b) If $\mathscr{H} = \mathrm{Forb}(C_9^2)$*, then* $\mathrm{ed}_{\mathscr{H}}(p) = \min\left\{\frac{p(1-p)}{2-p}, \frac{p(1-p)}{1+p}\right\}$ *for all* $p \in [0, 1]$.

(c) If $\mathscr{H} = \mathrm{Forb}(C_{10}^2)$*, then* $\mathrm{ed}_{\mathscr{H}}(p) = \min\left\{\frac{p}{3}, \frac{1-p}{3}\right\}$ *for all* $p \in [0,1]$.

(d) If $\mathscr{H} = \mathrm{Forb}(C_{11}^2)$*, then* $\mathrm{ed}_{\mathscr{H}}(p) = \min\left\{\frac{p}{3}, \frac{p(1-p)}{2}\right\}$ *for all* $p \in [0,1/2]$ *and* $\mathrm{ed}_{\mathscr{H}}(p) \leq \min\left\{\frac{p(1-p)}{2}, \frac{1-p}{3}\right\}$ *for all* $p \in [1/2,1]$.

(e) If $\mathscr{H} = \mathrm{Forb}(C_{12}^2)$*, then* $\mathrm{ed}_{\mathscr{H}}(p) = \frac{p(1-p)}{2}$ *for all* $p \in [0,1/2]$ *and* $\mathrm{ed}_{\mathscr{H}}(p) \leq \min\left\{\frac{p(1-p)}{2}, \frac{1-p}{3}\right\}$ *for all* $p \in [1/2,1]$.

Theorem 28, together with Theorem 13, is enough to determine the value of $p^*_{\mathrm{Forb}(C_h^2)}$ and of $d^*_{\mathrm{Forb}(C_h^2)}$ for $h \in \{5,\ldots,12\}$.

In work in progress, Berikkyzy, Peck, and Martin have extended the results from [37] to apply to powers of cycles, provided the number of vertices is large enough.

Theorem 29 (Berikkyzy-Martin-Peck [17]) *Let* $\mathscr{H} = \mathrm{Forb}(C_h^t)$ *where* C_h^t *is the* t^{th} *power of the cycle on* h *vertices. For* $t \geq 1$ *and* h *sufficiently large, let* $\ell_0 = \lceil h/(t+1) \rceil$, $\ell_t = \lceil h/(2t+1) \rceil$, *and* $p_0 = \ell_t^{-1}$.

(a) If $(t+1) \nmid h$*, then* $\mathrm{ed}_{\mathscr{H}}(p) = \min\left\{\frac{p}{t+1}, \frac{p(1-p)}{t(1-p)+(\ell_t-1)p}, \frac{1-p}{\ell_0-1}\right\}$ *for* $p \in [0,1]$.

(b) If $(t+1) \mid h$*, then* $\mathrm{ed}_{\mathscr{H}}(p) = \min\left\{\frac{p(1-p)}{t(1-p)+(\ell_t-1)p}, \frac{1-p}{\ell_0-1}\right\}$ *for* $p \in [p_0,1]$.

If $t = 2$, then $h \geq 13$ suffices for Theorem 29. In general, the bound that is proven to suffice is $h \geq 4t^2 + \Omega(t)$ although this is likely not best possible.

5.5 Complete bipartite graphs

5.5.1 Forb($K_{s,s}$)

The case of $\mathscr{H} = \mathrm{Forb}(K_{2,2})$ was established by Marchant and Thomason [32] where it was shown that $\mathrm{ed}_{\mathrm{Forb}(K_{2,2})}(p) = p(1-p)$ for all $p \in [0,1]$.

In the case of $\mathscr{H} = \mathrm{Forb}(K_{3,3})$, the values of $p^*_{\mathrm{Forb}(K_{3,3})} = \sqrt{2}-1$ and $d^*_{\mathrm{Forb}(K_{3,3})} = 3-2\sqrt{2}$ were established by Balogh and Martin [11].

For p not too small, the edit distance function for Forb($K_{3,3}$) coincides with $\gamma_{\mathrm{Forb}(K_{3,3})}(p) = \frac{p(1-p)}{1+p}$, but for p very small, the edit distance function is strictly smaller.

Theorem 30 (Marchant-Thomason [32]) *Let* $\mathscr{H} = \mathrm{Forb}(K_{3,3})$ *where* $K_{3,3}$ *is the complete bipartite graph with 3 vertices in each part.*

(a) $\mathrm{ed}_{\mathscr{H}}(p) = \frac{p(1-p)}{1+p}$*, for* $p \in [1/9,1]$.

(b) $\mathrm{ed}_{\mathscr{H}}(p) < \frac{p(1-p)}{1+p}$*, for* $p \in (0,1/124]$.

The CRGs used to establish Theorem 30(b) are defined by constructions due to Brown [22] to address a related Zarankiewicz problem. Specifically, for a prime power r, the constructions are $(r^2 - r)$-regular bipartite graphs on $2r^3$ vertices. Such graphs have no copy of $K_{3,3}$ and, of course, no copy of K_3. For such a graph G, we construct the CRG K for which the vertices of G are (black) vertices of K, the edges of G are gray edges of K, and the nonedges of G are white edges of K. By (5), this construction gives

$$f_K(p) = \frac{1}{2r^3}\left[1 + p\left(2r^3 - r^2 + r - 2\right)\right].$$

With $r = 19$, we obtain strict inequality for $p = 1/124$ and the continuity and concavity of the edit distance function give Theorem 30(b) for all $p \leq 1/124$.

It is also established in [32] that the value of $d^*_{\mathrm{Forb}(K_{s,s})}$ cannot be determined by the clique spectrum. The only extreme point of the clique spectrum is $(1, s-1)$ and the resulting CRG has $g_{K(1,s-1)} = \frac{p(1-p)}{1+(s-2)p} = \gamma_{\mathrm{Forb}(K_{s,s})}(p)$. The construction is a CRG $K^{(s-1)}$ on $2s - 2$ black vertices consisting of $s - 1$ disjoint white edges. It is easy to show that $K_{s,s} \not\mapsto K^{(s-1)}$ and since the g function of each component is $1/2$, Proposition 8 gives $g_{K^{(2s-2)}}(p) = 1/(2s-2)$.

So, $g_{K^{(2s-2)}}(p)$ is less than the maximum value of $\gamma_{\mathrm{Forb}(K_{s,s})}(p)$ for $s \geq 7$. We ask in Problem 2 if $1/(2s-2)$ is, indeed, the maximum value of $\mathrm{ed}_{\mathrm{Forb}(K_{s,s})}(p)$ and if that value is achieved for a positive length interval.

5.5.2 Forb($K_{2,t}$)

McKay and Martin [35] establish some surprising results for the hereditary property Forb($K_{2,t}$).

Theorem 31 (Martin-McKay [35]) *Let* $\mathscr{H} = \mathrm{Forb}(K_{2,t})$ *where* $K_{2,t}$ *is the complete bipartite graph with 2 vertices in one part and t vertices in the other part. For all* $t \geq 2$, $\gamma_{\mathscr{H}}(p) = \min\{p(1-p), (1-p)/(t-1)\}$.

(a) If $t = 3$, *then* $\mathrm{ed}_{\mathscr{H}}(p) = \min\left\{p(1-p), \frac{1-p}{2}\right\}$ *for all* $p \in [0, 1]$.

(b) If $t = 4$, *then* $\mathrm{ed}_{\mathscr{H}}(p) = \min\left\{p(1-p), \frac{7p+1}{15}, \frac{1-p}{3}\right\}$ *for all* $p \in [0, 1]$.

(c) If $t \geq 5$ *and is odd, then* $d^*_{\mathscr{H}} = \frac{1}{t+1}$ *and* $p^*_{\mathscr{H}} \supseteq \left[\frac{2t-1}{t(t+1)}, \frac{2}{t+1}\right]$.

(d) If $t \geq 9$, *there exists a* $p_0(t) < 1/2$ *such that* $\mathrm{ed}_{\mathscr{H}}(p) < p(1-p)$.

There are a number of interesting consequences arising from the study of these hereditary properties. First, we note that the CRG that gives the portion of the function in Theorem 31(b) corresponding to $(1 + 7p)/15$ results from a strongly regular graph construction.

Definition 32 *A* (n, k, λ, μ)-strongly regular graph *or* (n, k, λ, μ)-SRG *is a k-regular graph on n vertices for which each pair of adjacent vertices has exactly* λ *common neighbors and for which each pair of nonadjacent vertices has exactly* μ *common neighbors.*

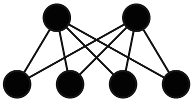

Fig. 5 The complete bipartite graph $K_{2,4}$.

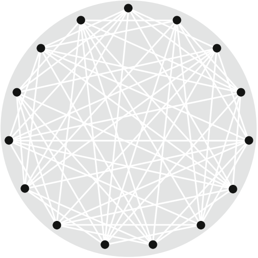

Fig. 6 The 15-vertex CRG that gives $(1+7p)/15$ in Theorem 31(b). The white edges are shown. The remaining edges are gray, and form a graph isomorphic to GQ(2, 2).

The CRG we use for Theorem 31(b) is constructed from a $(15,6,1,3)$-SRG, commonly called $\mathrm{GQ}(2,2)$. It is a member of the family of so-called generalized quadrangles. Given a $\mathrm{GQ}(2,2)$ G', the CRG K' has 15 white vertices that correspond to the vertices of the graph. An edge of K' is gray if and only if the corresponding pairs of vertices are adjacent in G'. See Figure 5 for the graph $K_{2,4}$ and Figure 6 for the 15-vertex CRG mentioned above. In [35], it is shown that $K_{2,4} \not\mapsto K'$ and $g_{K'}(p) = (1+7p)/15$.

Similar constructions from strongly regular graphs are in $\mathscr{K}(\mathrm{Forb}(K_{2,t}))$ and have a smaller g function than $\gamma_{\mathscr{H}}(p) = \min\{p(1-p), (1-p)/(t-1)\}$ for certain values of t and p.

For Theorem 31(c), the corresponding CRG $K^{(t)}$ has $t+1$ black vertices and a perfect matching of $(t+1)/2$ white edges. The remaining edges are gray. It is easy to show that $K_{2,t} \not\mapsto K^{(t+1)}$ and $g_{K^{(t+1)}}(p) = 1/(t+1)$.

For Theorem 31(d), the constructions are due to Füredi [28] to address a related Zarankiewicz problem. If q is a prime power such that $t-1$ divides $q-1$, then there exists a graph on $2(q^2-1)/(t-1)$ vertices that is q-regular with no copy of $K_{2,t}$ and no triangle. This is enough to ensure that $K_{2,t}$ does not map to the corresponding CRG. By (5), this construction gives

$$f_K(p) = \frac{t-1}{2(q^2-1)}\left[1 + p\left(\frac{2(q^2-1)}{t-1} - q - 2\right)\right].$$

With $t \geq 9$, we can find a sufficiently large prime power q so that $f_k(p) < p(1-p)$.

5.6 Split graphs

A graph H on at least two vertices is a *split graph* if there is a partition of $V(H)$ into one independent set and one clique. For a split graph with independence number $\alpha \geq 2$ and clique number $\omega \geq 2$, either $\alpha + \omega = h$ or $\alpha + \omega = h + 1$. We can compute the edit distance function of hereditary properties defined by such graphs.

Theorem 33 (Martin [34]) *Let H be a split graph which has independence number $\alpha = \alpha(H) \geq 2$ and clique number $\omega = \omega(H) \geq 2$. If $\mathscr{H} = \mathrm{Forb}(H)$, then*

$$\mathrm{ed}_{\mathscr{H}}(p) = \min\left\{\frac{p}{\omega-1}, \frac{1-p}{\alpha-1}\right\} .$$

Hence $p^*_{\mathscr{H}} = (\omega-1)/(\alpha+\omega-2)$ and $d^*_{\mathscr{H}} = 1/(\alpha+\omega-2)$.

6 Quantities related to the edit distance function

We get the following notation from Balogh et al. [14]. For a graph property[4] $\mathscr{H}$, the *labeled slice* of $\mathscr{H}$ is the set $\mathscr{H}^n$ of graphs in $\mathscr{H}$ with vertex set $\{1, \ldots, n\}$. The *labeled speed* of $\mathscr{H}$ is the function $n \mapsto |\mathscr{H}^n|$.

Theorem 34 ([12, 13, 16, 19]) *If $\mathscr{H}$ is a hereditary property of graphs, then one of the following holds:*

(i) There exist $N, k \in \mathbb{N}$ and polynomials $\{p_i(n)\}_{i=0}^k$ such that, for all $n > N$, $|\mathscr{H}^n| = \sum_{i=0}^k p_i(n) i^n$.
(ii) For some $t \in \mathbb{N}$, $t > 1$, we have $|\mathscr{H}^n| = n^{(1-1/t+o(1))n}$.
(iii) For n sufficiently large, $n^{(1+o(1))n} \leq |\mathscr{H}^n| \leq 2^{o(n^2)}$.
(iv) For some $k \in \mathbb{N}$, $k > 1$, we have $|\mathscr{H}^n| = 2^{(1-1/k+o(1))n^2/2}$.

Here $k = \chi_B(\mathscr{H}) - 1$.

This partition of hereditary properties was first discovered by Scheinerman and Zito [46]. As for the precise results, parts (i) and (ii) were established by Balogh, Bollobás, and Weinreich [16], part (iii) was also established by Balogh, Bollobás, and Weinreich [12, 13], and part (iv) was established by Bollobás and Thomason [19].

Part (iv) has been well-studied, including the $o(1)$ error term [15]. It has also been generalized. Bollobás and Thomason [20] define $c_p(\mathscr{H})$ as follows:

$$c_{\mathscr{H}}(p) = \lim_{n\to\infty} -\log_2 \Pr(G(n,p) \in \mathscr{H})/\tbinom{n}{2}. \qquad (12)$$

[4] For us, the property will be hereditary, although that is not necessary in order to define the speed.

They showed that the limit exists, based on work by Alekseev [1] (see also [2]). Thomason [49] compiled these results to show the relationship to the edit distance function:

Theorem 35 (Thomason [49]) *Let $\mathscr{H}$ be a nontrivial hereditary property and let $c_{\mathscr{H}}(p)$ be defined as in (12). Then*

$$c_{\mathscr{H}}(p) = (-\log_2 p(1-p))\,\mathrm{ed}_{\mathscr{H}}\left(\frac{\log_2(1-p)}{\log_2 p(1-p)}\right).$$

As we see, $c_{\mathscr{H}}(p)$ can be derived directly from the edit distance function.

Remark 36 *The function $c_{\mathscr{H}}(p)$ is not necessarily concave down, however $\mathrm{ed}_{\mathscr{H}}(p)$ is. Concavity is a key tool in finding the elusive lower bounds on the edit distance function which can then be used to compute lower bounds for $c_{\mathscr{H}}(p)$.*

Perhaps other functions of hereditary properties can be defined from the edit distance function. We are particularly interested in other metrics. For each positive integer n, let d be a metric on the space of graphs with vertex set $\{1,\ldots,n\}$. For any hereditary property $\mathscr{H}$, define

$$\mathrm{d}(G,\mathscr{H}) = \min\left\{\mathrm{d}(G,G') : V(G') = V(G), G' \in \mathscr{H}\right\},$$

and define the function

$$\phi_{\mathscr{H}}(p) := \limsup_{n\to\infty}\ \max\left\{\mathrm{d}(G,\mathscr{H}) : |V(G)| = n, |E(G)| = \left\lfloor p\tbinom{n}{2}\right\rfloor\right\}. \qquad (13)$$

In Question 3 from Section 8.2, we ask whether $\phi_{\mathscr{H}}(p)$ can be expressed as a function of p and $\mathrm{ed}_{\mathscr{H}}(\cdot)$ if d satisfies a natural property. Recall that dist represents the edit metric. It is clear that in order for any such result to exist, the metric d should be *continuous with respect to the edit metric*. That is, for every $\epsilon > 0$, there exists a δ such that $\mathrm{d}(G,G') < \epsilon$ whenever $\mathrm{dist}(G,G') < \delta$. This is a natural restriction because, for example, the trivial metric where $\mathrm{d}(G,G) = 0$ but $\mathrm{d}(G,G') = 1$ when $G \neq G'$ produces no useful results.

A well-studied metric is the so-called *cut metric* (or *cut norm*). Frieze and Kannan [27] introduced the cut norm and it is used extensively in the theorey of graph limits (see, e.g., Borgs et al. [21]). The cut norm is defined as follows for graphs on the same labeled vertex set $V = \{1,\ldots,n\}$:

$$\mathrm{d}_{\square}(G,G') = \max_{S,T\subset V}\ \frac{1}{n^2}\left|e_G(S,T) - e_{G'}(S,T)\right|,$$

where $e_G(S,T)$ is the number of ordered pairs (i,j) with $i \in S$ and $j \in T$ and $ij \in E(G)$. If S and T are disjoint, it counts the number of edges between S and T in G. The cut metric is useful for comparing random graphs. Although two typical graphs selected according to $G(n,p)$ have edit distance close to $2p(1-p)$, their $\mathrm{d}_{\square}$ distance is $O(1/n)$.

7 Generalizations of edit distance

Axenovich and Martin have investigated natural generalizations of the edit distance problem. The paper [8] addressed editing matrices (Section 7.1 below). The paper [9] addressed both editing the edges of multicolorings of a complete graph (Section 7.2) and editing the edges of a directed graph (Section 7.3).

7.1 Matrices

Let $\mathscr{A} = \{A_1, \dots, A_r\}$ be a partition of pairs from $[m] \times [n]$ into r nonemtpy classes. An $m \times n$ matrix $A = (a_{ij})$ is said to have a *pattern* $\mathscr{A}$ provided that $a_{ij} = a_{i'j'}$ if and only if $(i,j), (i',j') \in A_t$ for some $t \in \{1, \dots, r\}$. A pattern is *nontrivial* if $r \geq 2$. For a matrix M, if there is a submatrix M' with pattern $\mathscr{A}$, then we say that M has a *subpattern* $\mathscr{A}$.

For a pattern $\mathscr{A}$ and positive integers m, n, s, we define $\mathrm{Forb}(m, n; s, \mathscr{A})$ to be the set of all $m \times n$ matrices with at most s distinct entries and not containing subpattern $\mathscr{A}$.

For two matrices A and B of the same dimensions, we say that $\mathrm{dist}(A, B)$ is the number of positions in which A and B differ; i.e., it is the matrix Hamming distance. For a class of matrices $\mathscr{F}$ and a matrix A, all of the same dimensions, we denote $\mathrm{dist}(A, \mathscr{F}) = \min\{\mathrm{dist}(A, \mathscr{F}) : F \in \mathscr{F}\}$. Finally,

$$f(m, n; s, \mathscr{A}) := \max\left\{\mathrm{dist}(A, \mathscr{F}) : A \in \mathscr{M}(m, n; s), \mathscr{F} = \mathrm{Forb}(m, n; s, \mathscr{A})\right\}/mn. \tag{14}$$

The function f in (14) counts the maximum proportion of edits required to remove a pattern with r places from an $m \times n$ matrix with s distinct entries.[5]

Theorem 37 (Axenovich-Martin [8]) *Let s, r be positive integers, $s \geq r$. Let b_1, b_2 be positive constants such that $b_1 \leq m/n \leq b_2$. Let $\mathscr{A}$ be a nontrivial pattern with r distinct entries. Then*

$$f(m, n; s, \mathscr{A}) = (1 + o(1)) \left(\frac{s - r + 1}{s}\right).$$

Without loss of generality, the case of $s = 2$ corresponds to a $\{0, 1\}$-matrix. If the pattern has both zeros and ones, then $r = 2$ and the edit distance is $1/2$; i.e., an asymptotically most efficient editing algorithm is to make all entries zero or all entries one, whichever is most prevalent and the worst case is that there is the same

[5] In [8], f counts the number of edits, but we normalize by dividing by mn to make it consistent with the rest of this paper.

number of each. If the pattern has, say only zeroes, then the edit distance is 1 because the worst case is that the original matrix is all zeros and almost all of them must be changed to one.

The setting for matrices is identical to the case of editing the $m \times n$ complete bipartite graph in which the edges are colored with s distinct colors.

7.2 Multicolor edit distance

We will use slightly different terminology from [9] so as not to confuse it with similar notation for hypergraphs in Section 8.1. For any integer $r \geq 2$, an *r-colored graph* is pair (V,c) such that V is a finite labeled set and $c : E \to \{1,\ldots,r\}$. For r-colored graph G and $\rho \in \{1,\ldots,r\}$, we denote $E_\rho(G)$ to be the set of edges colored ρ.

If G and G' are two r-colored graphs on the same labeled vertex set, then the edit distance between them, $\mathrm{dist}(G,G')$ is the proportion of edges that receives a different color. For example, if $r = 2$, then graphs correspond to black edges and the complement corresponds to white edges. The following definitions for $r = 2$ are consistent with the graph case.

Further, we may define $\mathrm{dist}(G,\mathscr{H})$ for any hereditary property of r-colored graphs as in (1). In this setting, a property is still hereditary if it is closed under isomorphism and the deletion of vertices. For any r-colored graph H, we write $\mathrm{Forb}(H)$ to be the set of all r-colored graphs that have no copy of H. Note that "induced" is not necessary here because all edges receive a color. For an integer $r \geq 2$, a *density vector* $\mathbf{p} = (p_1,\ldots,p_r)$ is a nonnegative real vector with the property that $\sum_{\rho=1}^{r} p_i = 1$. The domain of r dimensional density vectors is the *(standard)* $(r-1)$*-simplex*.

If $\mathscr{H}$ is a hereditary property of r-colored graphs, then we may define the edit distance function parallel to (3) as follows.

$$\mathrm{ed}_{\mathscr{H}}(\mathbf{p}) := \lim_{n\to\infty} \max\left\{\mathrm{dist}(G,\mathscr{H}) : |V(G)| = n, |E_\rho(G)| = p_\rho\tbinom{n}{2}, \rho = 1,\ldots,r\right\}.$$

The limit was proven to exist in [9]. We omit floors and ceilings in defining $|E_\rho(G)|$ because they play no role in the limit.

We can also define the equivalent of CRGs in this setting. In [9], the term *type* is used, though for consistency of this paper, we will just call them r-CRGs.[6]

Definition 38 *An* r-CRG *K is a pair (U,ϕ) where U is a finite set of vertices and $\phi : U \times U \to 2^{\{1,\ldots,r\}} - \emptyset$ such that $\phi(x,y) = \phi(y,x)$ and $\phi(x,x) \neq \{1,\ldots,r\}$. The* sub-$r$-CRG induced by *$W \subseteq U$ is the r-CRG that results from deleting $U - W$.*

[6] In Section 8.1, we refer to r-CRHs when discussing the edit distance on r-uniform hypergraphs.

We say that an r-colored-graph $H = (V, c)$ embeds in r-CRG K, *and write* $H \mapsto K$, *if there is a map* $\gamma : V \to U$ *such that* $c(vv') = c_0$ *implies* $c_0 \in \phi(\gamma(v)\gamma(v'))$. *For any hereditary property* $\mathcal{H} = \bigcap_{H \in \mathcal{F}(\mathcal{H})} \mathrm{Forb}(H)$, *let* $\mathcal{K}(H)$ *be the set of r-CRGs for which none of* $\mathcal{F}(\mathcal{H})$ *embeds in that* r-CRG.

The notion of the binary chromatic number is more complicated in the r-colored graph case when $r > 2$. There are weak and strong colorings.

Definition 39 *Let* $\mathcal{H} = \bigcap_{H \in \mathcal{F}(\mathcal{H})} \mathrm{Forb}(H)$ *be a hereditary property of r-colored-graphs.*

- *An r-tuple* $(a_1, \ldots, a_r)$ *of nonnegative integers is* weakly good *if for some* $H \in \mathcal{F}(\mathcal{H})$ *the vertex set* $V(H)$ *can be partitioned into sets* $S_1, \ldots, S_r$ *such that for each* $\rho \in \{1, \ldots, r\}$ with $a_\rho \neq 0$, *the partition can be further refined* $S_\rho = V_{\rho,1} \cup \cdots \cup V_{\rho,a_\rho}$ *where each edge in* $V_{\rho,j}$ **does not** *have color* ρ.
- *An r-tuple* $(a_1, \ldots, a_r)$ *of nonnegative integers is* strongly good *if for some* $H \in \mathcal{F}(\mathcal{H})$ *the vertex set* $V(H)$ *can be partitioned into sets* $S_1, \ldots, S_r$ *such that for each* $\rho \in \{1, \ldots, r\}$ *with* $a_\rho \neq 0$, *the partitioned can be further refined* $S_\rho = V_{\rho,1} \cup \cdots \cup V_{\rho,a_\rho}$ *where each edge in* $V_{\rho,j}$ **must have** *color* ρ.

We can then define spectra and r-ary chromatic numbers based on weak and strong colorings.

Definition 40 *Let* $\mathcal{H} = \bigcap_{H \in \mathcal{F}(\mathcal{H})} \mathrm{Forb}(H)$ *be a hereditary property of r-colored-graphs.*

- *The* weak clique spectrum *of* $\mathcal{H}$ *is the set of all tuples* $(a_1, \ldots, a_r)$ *that are* **not** *weakly good. The* weak r-ary chromatic number *of* $\mathcal{H}$, *denoted* $\chi_r^{\mathrm{wk}}(\mathcal{H})$, *is the largest* $a_1 + \cdots + a_r + 1$ *such that* $(a_1, \ldots, a_r)$ *is in the weak clique spectrum of* $\mathcal{H}$.
- *The* strong clique spectrum *of* $\mathcal{H}$ *is the set of all tuples* $(a_1, \ldots, a_r)$ *that are* **not** *strongly good. The* strong r-ary chromatic number *of* $\mathcal{H}$, *denoted* $\chi_r^{\mathrm{st}}(\mathcal{H})$, *is the largest* $a_1 + \cdots + a_r + 1$ *such that* $(a_1, \ldots, a_r)$ *is in the strong clique spectrum of* $\mathcal{H}$.

The f and g functions are defined similar to the graph case.

Definition 41 *Let* $K = (\{u_1, \ldots, u_k\}, \phi)$ *be an r-CRG and for* $\mathbf{p} = (p_1, \ldots, p_r)$, *let* $\mathbf{M}_K(\mathbf{p})$ *denote the matrix with entries defined as follows:*

$$m_K(\mathbf{p})_{ij} = 1 - \sum_{\rho \in \phi(u_i, u_j)} p_\rho.$$

The functions f_K *and* g_K *are defined as follows:*

$$f_K(\mathbf{p}) = \frac{1}{k^2} \mathbf{1}^T \mathbf{M}_K(\mathbf{p}) \mathbf{1} \tag{15}$$

$$g_K(\mathbf{p}) = \min\left\{\mathbf{x}^T \mathbf{M}_K(\mathbf{p}) \mathbf{x} : \mathbf{x}^T \mathbf{1} = 1, \mathbf{x} \geq \mathbf{0}\right\}. \tag{16}$$

We say that a CRG K is $\mathbf{p}$-*core* if, for any proper sub-r-CRG K' of K, $g_{K'}(\mathbf{p}) > g_K(\mathbf{p})$.

We summarize the basic properties of this version of the edit distance function that generalize Theorems 13, 3, and 6. For $\mathbf{p} = (p_1, \ldots, p_r)$, the random r-colored graph $G(n, \mathbf{p})$ is the complete graph on n vertices in which each edge independently receives color ρ with probability p_ρ.

Theorem 42 (Axenovich-Martin [9]) *Let $\mathscr{H}$ be a hereditary property of r-colored graphs.*

(a) $\mathrm{ed}_{\mathscr{H}}(\mathbf{p})$ *is continuous over the* $(r-1)$*-simplex.*
(b) $\mathrm{ed}_{\mathscr{H}}(\mathbf{p})$ *is concave down over the* $(r-1)$*-simplex.*
(c) $\mathrm{ed}_{\mathscr{H}}(r^{-1}\mathbf{1}) \geq 1/(r(\chi_r^{\mathrm{st}}(\mathscr{H})-1))$.
(d) $\mathrm{ed}_{\mathscr{H}}(\mathbf{p}) \leq 1/(\chi_r^{\mathrm{wk}}(\mathscr{H})-1)$ *for all* $\mathbf{p}$ *in the* $(r-1)$*-simplex.*
(e) $\mathrm{ed}_{\mathscr{H}}(\mathbf{p}) = \lim_{n\to\infty} \mathbb{E}[\mathrm{dist}(G(n,\mathbf{p}), \mathscr{H})]$ *for all* $\mathbf{p}$ *in the* $(r-1)$*-simplex.*
(f) $\mathrm{ed}_{\mathscr{H}}(\mathbf{p}) = \inf\{f_K(\mathbf{p}) : K \in \mathscr{K}(\mathscr{H})\} = \inf\{g_K(\mathbf{p}) : K \in \mathscr{K}(\mathscr{H})\}$ *for all* $\mathbf{p}$ *in the* $(r-1)$*-simplex.*

Finally, we give some examples of results in the case $r = 3$.

Theorem 43 (Axenovich-Martin [9]) *Let $r = 3$ and let $\mathscr{H} = \bigcap_{H\in\mathscr{F}} \mathrm{Forb}(H)$ be a hereditary property of r-colored graphs. Let* $d^*_{\mathscr{H}} := \max\{\mathrm{ed}_{\mathscr{H}}(\mathbf{p}) : \mathbf{p}^T\mathbf{1} = 1, \mathbf{p} \geq \mathbf{0}\}$.

(a) If $\mathscr{F}$ is a family that consists of a single monochromatic triangle, then $d^*_{\mathscr{H}} = 1/2$.
(b) If $\mathscr{F}$ is a family that consists of a single triangle with two edges colored 1 and the other edge colored 2, then $d^*_{\mathscr{H}} = 1/2$.
(c) If $\mathscr{F}$ is a family that consists of two monochromatic triangles of different colors, then $d^*_{\mathscr{H}} = 1/2$.
(d) If $\mathscr{F}$ is a family that consists of all six bi-chromatic triangles, then $d^*_{\mathscr{H}} = 2/3$.
(e) If $\mathscr{F}$ is a family that consists of a single rainbow triangle, then $d^*_{\mathscr{H}} = 1/3$.

7.3 Directed edit distance

A *simple directed graph* or *digraph* G is a pair (V, c) such that V is a finite labeled set and, if $(V)_2 = V \times V - \{(v,v) : v \in V\}$ then $c : (V)_2 \to \{\bigcirc, -, \leftarrow, \rightarrow\}$ where

- $c(v,w) = c(w,v)$ if and only if $c(v,w) \in \{\bigcirc, -\}$ and
- $c(v,w) = \rightarrow$ if and only if $c(w,v) = \leftarrow$.

In the standard representation of digraphs as a pair (V, E) where $E \subseteq (V)_2$, we interpret $c(v,w) = \bigcirc$ to mean that neither (v,w) nor (w,v) is in E, $c(v,w) = -$ to mean that both (v,w) and (w,v) are in E, and $c(v,w) = \rightarrow$ to mean that $(v,w) \in E$ but $(w,v) \notin E$. We also define the following for any digraph G:

- $E_{\bigcirc}(G)$ is the set of all **unordered** pairs $\{v,w\}$ such that $c(v,w) = \bigcirc$.
- $E_{\leftarrow}(G)$ is the set of all **ordered** pairs $\{v,w\}$ such that $c(v,w) = \leftarrow$.
- $E_{\rightarrow}(G)$ is the set of all **ordered** pairs $\{v,w\}$ such that $c(v,w) = \rightarrow$.
- $E_{-}(G)$ is the set of all **unordered** pairs $\{v,w\}$ such that $c(v,w) = -$.

If $G = (V, c)$ and $G' = (V, c)$ are two digraphs on the same labeled vertex set with the same fixed palette, then the edit distance between them, $\mathrm{dist}(G, G')$ is the proportion of ordered pairs on which G and G' differ. We define $\mathrm{dist}(G, \mathscr{H})$ for any hereditary property of digraphs on any palette as in (1). A property is, of course, hereditary if it is closed under isomorphism and the deletion of vertices. For any digraph, we write $\mathrm{Forb}(H)$ to be the set of all digraphs that have no induced copy of H.

The digraph case encompasses several well-studied subclasses of digraphs. Just as the number of colors must be specified in Section 7.2, the palette must be specified for the digraph case.

Definition 44 *We say that $\mathscr{P} \subseteq \{\bigcirc, -, \leftarrow, \rightarrow\}$ is a* palette *if either none or both of "$\leftarrow$" and "$\rightarrow$" are in $\mathscr{P}$. There are 5 possible nontrivial palettes:*

(0) $\mathscr{P}_0 = \{\bigcirc, -, \leftarrow, \rightarrow\}$ *is the general case.*
(1) $\mathscr{P}_{\mathrm{compl}} = \{-, \leftarrow, \rightarrow\}$ *is the case of simple digraphs such that every pair of vertices has at least one arc between them.*
(2) $\mathscr{P}_{\mathrm{orien}} = \{\bigcirc, \leftarrow, \rightarrow\}$ *is the case of oriented graphs.*
(3) $\mathscr{P}_{\mathrm{undir}} = \{\bigcirc, -\}$ *is the usual case of simple, undirected graphs.*
(4) $\mathscr{P}_{\mathrm{tourn}} = \{\leftarrow, \rightarrow\}$ *is the case of tournaments.*

Definition 45 *A* directed density vector (p, q) *is a pair such that $p \geq 0$, $q \geq 0$, and $p + 2q \leq 1$. For different palettes, there are further restrictions.*

(0) If $\mathscr{P} = \mathscr{P}_{\mathrm{compl}}$, *then $p + 2q = 1$.*
(1) If $\mathscr{P} = \mathscr{P}_{\mathrm{orien}}$, *then $p = 0$ and $q \leq 1/2$.*
(2) If $\mathscr{P} = \mathscr{P}_{\mathrm{undir}}$, *then $q = 0$ and $p \leq 1$; i.e., the usual graph case.*
(3) If $\mathscr{P} = \mathscr{P}_{\mathrm{tourn}}$, *then $p = 0$ and $q = 1/2$.*

If $\mathscr{H}$ is a hereditary property of digraphs with palette $\mathscr{P}$, then for all directed density vectors $\mathbf{p} = (p, q)$, we define the edit distance function for hereditary property $\mathscr{H}$ as follows:

$$\mathrm{ed}_{\mathscr{H}}(\mathbf{p}) := \lim_{n\to\infty} \max\left\{\mathrm{dist}(G, \mathscr{H}) : \begin{array}{l} |V(G)| = n, |E_{-}(G)| = \lfloor p\binom{n}{2}\rfloor, \\ |E_{\leftarrow}(G)| = |E_{\rightarrow}(G)| = \lfloor q\binom{n}{2}\rfloor \end{array}\right\}.$$

The limit was proven to exist in [9].

In [9], the equivalent of CRGs (called *dir-types* in [9], but it would be natural to call them $\mathscr{P}$-dir-CRGs for palette $\mathscr{P}$) are defined as well as the notion of $H \mapsto K$ for any digraph H and any K a $\mathscr{P}$-dir-CRG. The matrix $\mathbf{M}_K(\mathbf{p})$ and functions $f_K(\mathbf{p})$ and $g_K(\mathbf{p})$ are defined analogously. In addition, the *strong directed clique spectrum*, *strong directed chromatic number* $\chi_{\mathscr{P}}^{\mathrm{st,dir}}(\mathscr{H})$, *weak directed clique spectrum*, and *weak directed chromatic number* $\chi_{\mathscr{P}}^{\mathrm{wk,dir}}(\mathscr{H})$ are defined for each palette, although for $\mathscr{P}_{\mathrm{undir}}$ and $\mathscr{P}_{\mathrm{tourn}}$ "strong" and "weak" are the same, where we use the notation $\chi_{\mathscr{P}}^{\mathrm{dir}}(\mathscr{H})$.

We will not give the detailed definitions of these quantities or of the random digraph $G(n,\mathbf{p})$. The natural notions are defined precisely in [9]. We have similar basic results for the directed case as for the multicolored case in Theorem 42.

Theorem 46 (Axenovich-Martin [9]) *Let $\mathscr{P}$ be a palette and let $\mathscr{H}$ be a hereditary property of digraphs with palette $\mathscr{P}$. Let the domain be defined as in Definition 45.*

(a) $\mathrm{ed}_{\mathscr{H}}(\mathbf{p})$ *is continuous over the domain.*
(b) $\mathrm{ed}_{\mathscr{H}}(\mathbf{p})$ *is concave down over the domain.*
(c) $\mathrm{ed}_{\mathscr{H}}(r^{-1}\mathbf{1}) \geq 1/(r(\chi_{\mathscr{P}}^{\mathrm{st,dir}}(\mathscr{H})-1))$.
(d) $\mathrm{ed}_{\mathscr{H}}(\mathbf{p}) \leq 1/(\chi_{\mathscr{P}}^{\mathrm{wk,dir}}(\mathscr{H})-1)$ *for all* $\mathbf{p}$ *in the domain.*
(e) $\mathrm{ed}_{\mathscr{H}}(\mathbf{p}) = \lim_{n\to\infty}\mathbb{E}[\mathrm{dist}(G(n,\mathbf{p}),\mathscr{H})]$ *for all* $\mathbf{p}$ *in the domain.*
(f) $\mathrm{ed}_{\mathscr{H}}(\mathbf{p}) = \inf\{f_K(\mathbf{p}) : K \in \mathscr{K}(\mathscr{H})\} = \inf\{g_K(\mathbf{p}) : K \in \mathscr{K}(\mathscr{H})\}$ *for all* $\mathbf{p}$ *in the domain.*

We give some examples involving triangles.

Theorem 47 (Axenovich-Martin [9]) *Let $\mathscr{H}$ be a hereditary property of digraphs with palette $\mathscr{P}$.*

(a) If $\mathscr{H} = \mathrm{Forb}(H_{\mathrm{dir}})$ *where* H_{dir} *is a directed triangle, then* $d^*_{\mathscr{H}} = 1/2$ *regardless of* $\mathscr{P}$.
(b) If $\mathscr{H} = \mathrm{Forb}(H_{\mathrm{tra}})$ *where* H_{tra} *is a transitive triangle, then* $\mathscr{H}$ *is a trivial hereditary property as long as* $\mathscr{P} = \mathscr{P}_{\mathrm{tourn}}$.
(c) If $\mathscr{H} = \mathrm{Forb}(H_{\mathrm{tra}})$ *where* H_{tra} *is a transitive triangle, then* $d^*_{\mathscr{H}} = 1/2$, *as long as* $\mathscr{P} \neq \mathscr{P}_{\mathrm{tourn}}$.
(d) If $\mathscr{H} = \mathrm{Forb}(H_{\mathrm{dir}}) \cap \mathrm{Forb}(H_{\mathrm{tra}})$ *where* H_{dir} *is a directed triangle and* H_{tra} *is a transitive triangle, then* $d^*_{\mathscr{H}} = 1/2$, *as long as* $\mathscr{P} \neq \mathscr{P}_{\mathrm{tourn}}$.

The case of tournaments turns out to be trivial. Theorem 47(b) is a simple consequence of Ramsey theory, a hereditary property $\mathscr{H} = \bigcap_{H\in\mathscr{F}(\mathscr{H})}\mathrm{Forb}(H)$ is nontrivial if and only if no member of $\mathscr{F}(\mathscr{H})$ is transitive. In the case of tournaments, the density vector must be $\mathbf{p} = (0,1/2)$. The edit distance function is, therefore, a constant.

Theorem 48 *Let $\mathscr{H}$ be a nontrivial hereditary property of tournaments and $\mathscr{P} = \mathscr{P}_{\mathrm{tourn}}$. Then*

$$\mathrm{ed}_{\mathscr{H}}(0,1/2) = \frac{1}{2(\chi_{\mathscr{P}}^{\mathrm{dir}}(\mathscr{H})-1)}.$$

8 Future directions

8.1 Hypergraph edit distance

Berikkyzy and the author have been investigating the extension of the edit distance problem to r-uniform hypergraphs (r-graphs). A *colored regularity hypergraph of order r (r-CRH)* is a triple (V, E, ϕ) in which V is a vertex set, E is the collection of all r-multisets on V, and $\phi : E \to \{\mathbf{W}, \mathbf{G}, \mathbf{B}\}$ with the restrictions that, (a) for any $v \in V$, $\phi(\{v, \ldots, v\}) \in \{\mathbf{W}, \mathbf{B}\}$ and (b) for any permutation $\sigma \in \Sigma_r$, $\phi(\{v_1, \ldots, v_r\}) = \phi(\{v_{\sigma(1)}, \ldots, v_{\sigma(r)}\})$. Therefore, a 2-CRH is just a CRG.

In parallel to the graph case, we can define colored homomorphisms from r-uniform hypergraphs to r-CRHs so that if r-graph H does not map to a r-CRH K, then an r-graph G which is edited according to the "recipe" defined by K will have no induced copy of H.

We can then define, for each r-CRH K, an r-linear form which we can also call $g_K(p)$. It is easy to prove, for a hereditary property $\mathscr{H}$ of r-graphs, that there is a family $\mathscr{F}(\mathscr{H})$ of r-CRHs such that

$$\mathrm{ed}_{\mathscr{H}}(p) \leq \inf_{K \in \mathscr{K}(\mathscr{H})} g_K(p). \tag{17}$$

The difficulty in extending the theory of the edit distance in graphs to hypergraphs is in proving that (17) is, in fact, an equality.

The above definition of the r-CRH would not seem to be adequate to capture the subtleties of hypergraphs. Consider the common example of a hypergraph whose 3-edges are cyclic triangles in an underlying random tournament. See, e.g., the survey of hypergraph Turán theory by Keevash [30]. This hypergraph has no copy of the tetrahedron K_4^3 but crossing triples would be gray in any 3-CRH that models it.

Strong hypergraph regularity was developed in the 3-uniform case by Frankl and Rödl [26] and then for the general r-uniform case by Gowers [29], Rödl and Skokan [44, 45], and Nagle, Rödl, and Schacht [36]. In these formulations, the notion of how overlapping hyperedges interact is captured by structures known as *complexes*. The structure of complexes inherent in strong hypergraph regularity would seem to be necessary in order to define r-CRHs and colored homomorphisms in order for the existence of a particular induced hypergraph to be determined.

The edit distance problem is, asymptotically, a general case of the Turán problem. In the context of Turán-type problems, a hypergraph property is *monotone* if it is closed under the taking of (not necessarily induced) subgraphs. Therefore, a monotone property is also hereditary. For a monotone property $\mathscr{M}$, the Turán density is $\pi(\mathscr{M}) = \limsup_{n\to\infty} \max\{|E(G)|/\binom{n}{2} : |V(G)| = n, G \in \mathscr{M}\}$. It is easy to see that

$$\pi(\mathscr{M}) = 1 - \mathrm{ed}_{\mathscr{M}}(1).$$

The Turán density for most monotone properties is not currently known, even though a great deal of work has been done on the subject.

In the graph case, it is trivial to derive $\mathrm{ed}_{\mathscr{M}}(1)$ using symmetrization. In addition, the classification of p-core CRGs established by Marchant and Thomason [32] allows for a trivial proof of the asymptotic Erdős-Stone-Simonovits result [24, 25].

In Question 4 from Section 8.2, we ask several questions that are related to a general theory of edit distance in r-uniform hypergraphs.

8.2 Open Problems

We first ask about powers of cycles and the questions left open in Section 5.4.

Question 1 *Let $\mathscr{H} = \mathrm{Forb}(C_h^t)$. What is $\mathrm{ed}_{\mathscr{H}}(p)$ for small values of p, where $t+1$ divides h? What is $\mathrm{ed}_{\mathscr{H}}(p)$ for small values of h? In particular:*

- *For $\mathscr{H} = \mathrm{Forb}(C_h)$, what is $\mathrm{ed}_{\mathscr{H}}(p)$ for even values of h and all values of p?*
- *For $\mathscr{H} = \mathrm{Forb}(C_h)$, what is $d^*_{\mathscr{H}}$ for all $t \geq 2$ and $h \geq 2t+1$?*

Next we consider complete bipartite graphs and some interesting questions from Section 5.5

Question 2 *What is $\mathrm{ed}_{\mathscr{H}}(p)$ for $\mathscr{H} = \mathrm{Forb}(K_{s,t})$? In particular:*

- *For $\mathscr{H} = \mathrm{Forb}(K_{s,s})$, is $d^*_{\mathscr{H}} = 1/(2s-2)$ if $s \geq 7$?*
- *For $\mathscr{H} = \mathrm{Forb}(K_{s,s})$, is $p^*_{\mathscr{H}}$ is an interval of positive length if $s \geq 7$?*
- *For $\mathscr{H} = \mathrm{Forb}(K_{s,t})$, which values of s and t give that $p^*_{\mathscr{H}}$ is an interval of positive length?*

Other metrics on the space of graphs are of interest, as we discussed in Section 6.

Question 3 *Let $\mathscr{H}$ be a nontrivial hereditary property of graphs.*

- *For the cut metric $\mathrm{d}_\square$, is it the case that the function $\phi_{\mathscr{H}}$, as defined in (13), can be expressed only in terms of p and of $\mathrm{ed}_{\mathscr{H}}(\cdot)$?*
- *For any metric d that is continuous with respect to the edit metric, is it the case that the function $\phi_{\mathscr{H}}$, as defined in (13), can be expressed only in terms of p and of $\mathrm{ed}_{\mathscr{H}}(\cdot)$?*

The question of the edit distance in hypergraphs is wide open, as we discussed in Section 8.1.

Question 4 *Let $\mathscr{H}$ be a nontrivial hereditary property of r-uniform hypergraphs.*

- *Is it the case that $\mathrm{ed}_{\mathscr{H}}(p) = \inf_{K\in\mathscr{K}(\mathscr{H})} g_K(p)$ for all $p \in [0,1]$?*
- *If $\mathscr{H}$ is monotone, is it the case that $\mathrm{ed}_{\mathscr{H}}(1) = \inf_{K\in\mathscr{K}(\mathscr{H})} g_K(1)$?*
- *Is there a useful form of generalizations properties do r-linear forms have?*
- *Can we provide a structural characterization for r-CRHs that are p-core, à la Theorem 9?*

The cases of $\mathscr{H} = \mathrm{Forb}(K_{3,3})$ and $\mathscr{H} = \mathrm{Forb}(K_{2,t})$ for $t \geq 9$ suggest an infinite number of CRGs are necessary to define an edit distance function, but only if one wants to compute it for p arbitrarily close to 0 (or by considering the property $\overline{\mathscr{H}}$, arbitrarily close to 1).

Conjecture 1 *Let $\mathscr{H}$ be a nontrivial hereditary property. For every $\epsilon > 0$ there exists a $\mathscr{K}' = \mathscr{K}'(\epsilon, \mathscr{H})$ such that*

$$\mathrm{ed}_{\mathscr{H}}(p) = \min\left\{g_K(p) : K \in \mathscr{K}'\right\} \qquad \textit{for all } p \in (\epsilon, 1-\epsilon).$$

We ask if the behavior we seem to observe for $\mathrm{Forb}(K_{3,3})$ and $\mathrm{Forb}(K_{2,t})$ for $t \geq 9$ – that is, that an infinite sequence of CRGs are required to compute the edit distance function for all values of p – does, in fact, occur.

Question 5 *Are there hereditary properties of graphs for which the edit distance function cannot be determined from the G functions of a finite number of CRGs?*

Finally, we conclude with an open problem for the random graph. Recall that $G(n,p)$ denotes the Erdős-Rényi random graph on n vertices with probability p.

Conjecture 2 (Martin [33]) *Fix $p_0 \in (0,1)$ and let $\mathscr{H} = \mathrm{Forb}(G(n_0, p_0))$. Then*

$$\mathrm{ed}_{\mathscr{H}}(p) = (1+o(1))\frac{2\log_2 n_0}{n_0}\min\left\{\frac{p}{-\log_2(1-p_0)}, \frac{1-p}{-\log_2 p_0}\right\}$$

with probability approaching 1 *as $n_0 \to \infty$.*

The functions that define this bound are of the form $p/(\chi - 1)$ and $(1-p)/(\overline{\chi} - 1)$. The case of $p_0 = 1/2$ was proved to be true by Alon and Stav [4].

If Conjecture 2 is true, then it implies that $p^*_{\mathscr{H}} = \frac{\log_2(1-p_0)}{\log_2 p_0(1-p_0)}$, which is only equal to p_0 itself when $p_0 \in \{0, 1/2, 1\}$. Informally, this implies the counterintuitive notion that it is harder to remove induced copies of $G(n_0, p_0)$ from $G(n, p^*_{\mathscr{H}})$ than it is to remove them from $G(n, p_0)$.

If Conjecture 2 is false, then it implies that the structure of random graphs and the behavior of editing induced graphs are quite complex and very unexpected.

Acknowledgements The author would like to extend a heartfelt appreciation many people for valuable conversations on the topic of edit distance, including Noga Alon, Ed Marchant, Richard Mycroft, Jozef Skokan, and Andrew Thomason. Thanks also to his collaborators József Balogh, Zhanar Berikkyzy, André Kézdy, Tracy McKay, and Chelsea Peck. Special appreciation to Maria Axenovich, who suggested the problem in the first place and without whom this work would not be possible.

The author would like to profoundly thank the National Security Agency (NSA) which partially supported his research via grant H98230-13-1-0226 as well as the Institute for Mathematics and its Applications (IMA). This paper was written while the author was a long-term visitor at the IMA and would not have been possible without their resources and generous support.

References

1. V.E. Alekseev, Hereditary classes and coding of graphs. Problemy Kibernet. **39**, 151–164 (1982). MR 694829 (85a:05071)
2. V.E. Alekseev, Range of values of entropy of hereditary classes of graphs. Diskret. Mat. **4**(2), 148–157 (1992). MR 1181539 (93k:05140)
3. N. Alon, Testing subgraphs in large graphs. Random Struct. Algorithm **21**(3–4), 359–370 (2002). Random structures and algorithms (Poznan, 2001). MR 1945375 (2003k:05129)
4. N. Alon, U. Stav, The maximum edit distance from hereditary graph properties. J. Comb. Theory Ser. B **98**(4), 672–697 (2008). MR 2418765 (2009d:05111)
5. N. Alon, U. Stav, What is the furthest graph from a hereditary property? Random Struct. Algorithm **33**(1), 87–104 (2008). MR 2428979 (2009c:05219)
6. N. Alon, R.A. Duke, H. Lefmann, V. Rödl, R. Yuster, The algorithmic aspects of the regularity lemma. J. Algorithm **16**(1), 80–109 (1994). MR 1251840 (94j:05112)
7. N. Alon, E. Fischer, M. Krivelevich, M. Szegedy, Efficient testing of large graphs. Combinatorica **20**(4), 451–476 (2000). MR 1804820 (2002f:05143)
8. M. Axenovich, R.R. Martin, Avoiding patterns in matrices via a small number of changes. SIAM J. Discrete Math. **20**(1), 49–54 (electronic) (2006). MR 2257244 (2007j:05015)
9. M. Axenovich, R.R. Martin, Multicolor and directed edit distance. J. Comb. **2**(4), 525–556 (2011). MR 2911190
10. M. Axenovich, A. Kézdy, R.R. Martin, On the editing distance of graphs. J. Graph Theory **58**(2), 123–138 (2008). MR 2407000 (2009c:05107)
11. J. Balogh, R.R. Martin, Edit distance and its computation. Electron. J. Comb. **15**(1), Research Paper 20, 27 (2008). MR 2383440 (2008j:05175)
12. J. Balogh, B. Bollobás, D. Weinreich, The penultimate rate of growth for graph properties. Eur. J. Comb. **22**(3), 277–289 (2001). MR 1822715 (2002b:05079)
13. J. Balogh, B. Bollobás, D. Weinreich, Measures on monotone properties of graphs. Discrete Appl. Math. **116**(1–2), 17–36 (2002). MR 1877113 (2002i:05066)
14. J. Balogh, B. Bollobás, M. Saks, V.T. Sós, The unlabelled speed of a hereditary graph property. J. Comb. Theory Ser. B **99**(1), 9–19 (2009). MR 2467814 (2009k:05105)
15. J. Balogh, B. Bollobás, M. Simonovits, The fine structure of octahedron-free graphs. J. Comb. Theory Ser. B **101**(2), 67–84 (2011). MR 2763070 (2012a:05157)
16. J. Balogh, B. Bollobás, D. Weinreich, The speed of hereditary properties of graphs. J. Comb. Theory Ser. B **79**(2), 131–156 (2000). MR 1769217 (2001d:05092)
17. Z. Berikkyzy, R.R. Martin, C. Peck, On the edit distance of powers of cycles. Submitted. arXiv: 1509.07438
18. B. Bollobás, A. Thomason, Projections of bodies and hereditary properties of hypergraphs. Bull. Lond. Math. Soc. **27**(5), 417–424 (1995). MR 1338683 (96e:52006)
19. B. Bollobás, A. Thomason, Hereditary and monotone properties of graphs, in *The mathematics of Paul Erdős, II*, Algorithms Combination, vol. 14 (Springer, Berlin, 1997), pp. 70–78. MR 1425205 (98a:05085)
20. B. Bollobás, A. Thomason, The structure of hereditary properties and colourings of random graphs. Combinatorica **20**(2), 173–202 (2000). MR 1767020 (2001e:05114)
21. C. Borgs, J.T. Chayes, L. Lovász, V.T. Sós, K. Vesztergombi, Convergent sequences of dense graphs. I. Subgraph frequencies, metric properties and testing. Adv. Math. **219**(6), 1801–1851 (2008). MR 2455626 (2009m:05161)
22. W.G. Brown, On graphs that do not contain a Thomsen graph. Canad. Math. Bull. **9**, 281–285 (1966). MR 0200182 (34 #81)
23. D. Chen, O. Eulenstein, D. Fernández-Baca, M. Sanderson, Supertrees by flipping, in *Computing and Combinatorics*. Lecture Notes in Computer Science, vol. 2387 (Springer, Berlin, 2002), pp. 391–400. MR 2064534 (2005b:92030)
24. P. Erdős, M. Simonovits, A limit theorem in graph theory. Studia Sci. Math. Hungar. **1**, 51–57 (1966). MR 0205876 (34 #5702)

25. P. Erdös, A.H. Stone, On the structure of linear graphs. Bull. Am. Math. Soc. **52**, 1087–1091 (1946). MR 0018807 (8,333b)
26. P. Frankl, V. Rödl, Extremal problems on set systems. Random Struct. Algorithm **20**(2), 131–164 (2002). MR 1884430 (2002m:05192)
27. A. Frieze, R. Kannan, Quick approximation to matrices and applications. Combinatorica **19**(2), 175–220 (1999). MR 1723039 (2001i:68066)
28. Z. Füredi, New asymptotics for bipartite Turán numbers, J. Comb. Theory Ser. A **75**(1), 141–144 (1996). MR 1395763 (97f:05096)
29. W.T. Gowers, Hypergraph regularity and the multidimensional Szemerédi theorem. Ann. Math. (2) **166**(3), 897–946 (2007). MR 2373376 (2009d:05250)
30. P. Keevash, Hypergraph Turán problems, in *Surveys in Combinatorics 2011*, The London Mathematical Society, Lecture Notes Series, vol. 392 (Cambridge University Press, Cambridge, 2011), pp. 83–139. MR 2866732
31. E.J. Marchant, *Graphs with Weighted Colours and Hypergraphs*. Ph.D. thesis, University of Cambridge, 2011
32. E. Marchant, A. Thomason, Extremal graphs and multigraphs with two weighted colours, in *Fete of Combinatorics and Computer Science*, Bolyai Society Mathematical Studies, vol. 20 (János Bolyai Mathematical Society, Budapest, 2010), pp. 239–286. MR 2797967 (2012f:05149)
33. R. Martin, The edit distance function and symmetrization Electron. J. Comb. **20**(3), Paper 26, 25 (2013). MR 3104524
34. R.R. Martin, On the computation of edit distance functions. Discrete Math. **338**(2), 291–305 (2015). MR 3279281
35. R.R. Martin, T. McKay, On the edit distance from $K_{2,t}$-free graphs. J. Graph Theory **77**(2), 117–143 (2014). MR 3246171
36. B. Nagle, V. Rödl, M. Schacht, The counting lemma for regular k-uniform hypergraphs. Random Struct. Algorithm **28**(2), 113–179 (2006). MR 2198495 (2007d:05084)
37. C. Peck, *On the Edit Distance from a Cycle- and Squared Cycle-Free Graph*. Master's thesis, Iowa State University, 2013
38. O. Pikhurko, An exact Turán result for the generalized triangle. Combinatorica **28**(2), 187–208 (2008). MR 2399018 (2009c:05161)
39. L. Pósa, Hamiltonian circuits in random graphs. Discrete Math. **14**(4), 359–364 (1976). MR 0389666 (52 #10497)
40. H.J. Prömel, A. Steger, Excluding induced subgraphs: quadrilaterals. Random Struct. Algorithm **2**(1), 55–71 (1991). MR 1099580 (92j:05157)
41. H.J. Prömel, A. Steger, Excluding induced subgraphs. III. A general asymptotic. Random Struct. Algorithm **3**(1), 19–31 (1992). MR 1139486 (93d:05065)
42. H.J. Prömel, A. Steger, Excluding induced subgraphs. II. Extremal graphs. Discrete Appl. Math. **44**(1–3), 283–294 (1993). MR 1227710 (94f:05121)
43. D.C. Richer, *Graph Theory and Combinatorial Games*. Ph.D. thesis, University of Cambridge, 2000
44. V. Rödl, J. Skokan, Regularity lemma for k-uniform hypergraphs. Random Struct. Algorithm **25**(1), 1–42 (2004). MR 2069663 (2005d:05144)
45. V. Rödl, J. Skokan, Applications of the regularity lemma for uniform hypergraphs. Random Struct. Algorithm **28**(2), 180–194 (2006). MR 2198496 (2006j:05099)
46. E.R. Scheinerman, J. Zito, On the size of hereditary classes of graphs. J. Comb. Theory Ser. B **61**(1), 16–39 (1994). MR 1275261 (95e:05061)
47. A. Sidorenko, Boundedness of optimal matrices in extremal multigraph and digraph problems. Combinatorica **13**(1), 109–120 (1993). MR 1221180 (94c:05043)
48. E. Szemerédi, Regular partitions of graphs, in *Problèmes combinatoires et théorie des graphes (Colloq. Internat. CNRS, Univ. Orsay, Orsay, 1976)*. Colloq. Internat. CNRS, vol. 260 (CNRS, Paris, 1978), pp. 399–401. MR 540024 (81i:05095)

49. A. Thomason, *Graphs, Colours, Weights and Hereditary Properties*. Surveys in combinatorics 2011, The London Mathematical Society, Lecture Notes Series, vol. 392 (Cambridge University Press, Cambridge, 2011), pp. 333–364. MR 2866736
50. A.A. Zykov, On some properties of linear complexes. Am. Math. Soc. Translat. **1952**(79), 33 (1952). MR 0051516 (14,493a)

Repetitions in graphs and sequences

Maria Axenovich

Abstract The existence of unavoidable repeated substructures is a known phenomenon implied by the pigeonhole principle and its generalizations. A fundamental problem is to determine the largest size of a repeated substructure in any combinatorial structure from a given class. The strongest notion of repetition is a pair of isomorphic substructures, such as a pair of vertex-disjoint or edge-disjoint isomorphic subgraphs or a pair of disjoint identical subsequences of a sequence. A weaker notion of repetition is a pair of substructures that have the same value on a certain set of parameters. This includes vertex-disjoint induced subgraphs of the same order and size, disjoint vertex sets with the same multiset of pairwise distances, subgraphs with the same maximum degree. This paper surveys results on unavoidable repetitions, also referred to as twins, with a focus on three asymptotically tight results obtained over the past 5 years.

Keywords Sequence • Subword • Identical subwords • Twins in sequences • Twins in graphs • Isomorphic subgraphs • Repeated sequences • Self-similarity • Divisibility of graphs • Twins in hypergraphs.

Mathematics Subject Classification (2010): 05D05, 05C35, 05C60, 05C70

1 Introduction

There are many repetitions that one can observe in nature: identical twins, two leaves on a tree that look alike, repetitive motifs in a bird's song, segments of the DNA that are identical, etc.

Discrete mathematical structures possess repetitions as well, as is implied by the pigeonhole principle and its generalizations. A fundamental problem is to determine the largest size of a repeated substructure in any combinatorial structure from a given class. Here, we shall often refer to repeated structures as twins even if these twins are not identical.

M. Axenovich (✉)
Karlsruher Institut für Technologie, Karlsruhe, Germany
e-mail: maria.aksenovich@kit.edu

A. Beveridge et al. (eds.), *Recent Trends in Combinatorics*, The IMA Volumes
in Mathematics and its Applications 159, DOI 10.1007/978-3-319-24298-9_3

The strongest notion of repetition deals with "identical" twins, i.e., with a pair of isomorphic substructures. A weaker notion of repetition is a pair of substructures with the same value on a certain set of parameters.

This survey focuses on three types of "identical" twins, which have corresponding size functions $f_v(n), f_e(n), f(n)$ and a few types of weaker twins, which have size functions such as $t(n), h(n)$.

Let $\mathcal{G}_n$ be the class of all n-vertex graphs, and $\mathcal{G}^m$ be the class of all graphs with m edges, let k be an integer. For all standard graph theoretic notions we refer the reader to the book of West [54]. All graphs here are simple graphs with no repeated edges and no loops. We define the size functions:

$$\begin{aligned}
f_v(n) &= \max\{\, k : \text{ any } G \in \mathcal{G}_n \text{ has two isomorphic vertex-disjoint induced} \\
&\qquad \text{subgraphs on } k \text{ vertices each}\}, \\
f_e(m) &= \max\{\, k : \text{ any } G \in \mathcal{G}^m \text{ has two isomorphic edge-disjoint subgraphs} \\
&\qquad \text{on } k \text{ edges each}\}, \\
f(n) &= \max\{\, k : \text{ any binary sequence with } n \text{ elements contains} \\
&\qquad \text{two disjoint identical subsequences of } k \text{ elements each}\}, \\
t(n) &= \max\{\, k : \text{ any } G \in \mathcal{G}_n \text{ has two vertex-disjoint induced subgraphs} \\
&\qquad \text{on } k \text{ vertices and with the same number of edges}\}, \\
h(n) &= \max\{\, k : \text{ any } G \in \mathcal{G}_n \text{ has two disjoint subsets of vertices} \\
&\qquad \text{whose multisets of pairwise distances are identical}\}.
\end{aligned}$$

For all of these functions, except for the last one, we now know exact asymptotic behavior. The following theorem is an easy consequence of Ramsey theorem [22, 49] and a property of random graphs, see Section 2.

Theorem 1.1.

$$f_v(n) = \Theta(\log n).$$

The next three theorems proved in 2012, 2014, and 2012, respectively, involve more sophisticated proof techniques such as random methods on graphs, regular partitions of sequences, and balanced partitions of integers.

Theorem 1.2. (Lee, Loh, and Sudakov [41]). *There are absolute positive constants c and C for which*

$$c(m \log m)^{2/3} \leq f_e(m) \leq C(m \log m)^{2/3}.$$

Theorem 1.3. (Axenovich, Person, and Puzynina [6]). *There exists an absolute constant C such that*

$$\left(1 - C\left(\frac{\log n}{\log\log n}\right)^{-1/4}\right)\frac{n}{2} \le f(n) \le \frac{n}{2} - \frac{1}{2}\log n.$$

Theorem 1.4. (Bollobás, Kittipassorn, Narayanan, and Scott [14]). *For every $\epsilon > 0$, there is a natural number $N = N(\epsilon)$ such that for all $n > N$*

$$\frac{n}{2} - \epsilon n \le t(n) \le \frac{n}{2} - \log\log n.$$

Compared to the above theorems giving tight bounds, it is still not known what is the correct asymptotic behavior of the function $h(n)$. The gap between the upper and the lower bound is large. Here, the slight $\log\log n$ improvement of the upper bound is due to a result from [5].

Theorem 1.5. (Albertson, Pach, Young [1, 5]). *There are positive constants c, c', such that for any $n > 3$,*

$$\frac{c\log n}{\log\log n} < h(n) \le \frac{n}{4} - c'\log\log n.$$

We address functions$f_v(n)$,$f_e(m)$,$f(n)$, $t(n)$, and $h(n)$ in Sections 2, 3, 4, 5, and 6, respectively. We provide some insight into proof techniques, state generalizations and open problems. At the end of the survey, we mention some other weak twin problems.

2 Vertex-disjoint isomorphic subgraphs

Among all twin problems we consider here, the problem of finding the largest order of two vertex-disjoint isomorphic subgraphs is relatively easy and Theorem 1.1 answers it asymptotically. We prove it here for completeness.

Proof of Theorem 1.1. For the lower bound, consider an n-vertex graph G. By Ramsey theorem [22, 49] there is a complete subgraph or an independent set on $c\log n$ vertices. Thus G has two vertex-disjoint induced subgraphs on $\frac{c}{2}\log n$ vertices both isomorphic to either a complete graph or an empty graph.

For the upper bound, we shall consider Erdős-Renyi random graph $G{=}G(n, 1/2)$ and follow the simple union bound approach of Lee, Loh, and Sudakov [41]. The fact that G has two vertex-disjoint isomorphic twins on t vertices each is equivalent to the existence of a $2t$-vertex subgraph H' that can be partitioned into a graph H and a vertex-disjoint isomorphic copy of H. The expected number of such graphs H' is at most $\binom{n}{2t}\binom{2t}{t}t!2^{-\binom{t}{2}}$, where the first binomial coefficient gives the number of ways to choose $2t$ vertices in G, the second counts the number of ways to split $2t$ vertices in two equal parts, $t!$ is the number of ways to permute the vertices