

# involve

a journal of mathematics

## Editorial Board

Kenneth S. Berenhaut, *Managing Editor*

Colin Adams  
Arthur T. Benjamin  
Martin Bohner  
Nigel Boston  
Amarjit S. Budhiraja  
Pietro Cerone  
Scott Chapman  
Joshua N. Cooper  
Jem N. Corcoran  
Toka Diagana  
Michael Dorff  
Sever S. Dragomir  
Joel Foisy  
Errin W. Fulp  
Joseph Gallian  
Stephan R. Garcia  
Anant Godbole  
Ron Gould  
Sat Gupta  
Jim Haglund  
Johnny Henderson  
Glenn H. Hurlbert  
Charles R. Johnson  
K. B. Kulasekera  
Gerry Ladas  
David Larson  
Suzanne Lenhart

Chi-Kwong Li  
Robert B. Lund  
Gaven J. Martin  
Mary Meyer  
Frank Morgan  
Mohammad Sal Moslehian  
Zuhair Nashed  
Ken Ono  
Yuval Peres  
Y.-F. S. Pétermann  
Jonathon Peterson  
Robert J. Plemmons  
Carl B. Pomerance  
Vadim Ponomarenko  
Bjorn Poonen  
József H. Przytycki  
Richard Rebarber  
Robert W. Robinson  
Javier Rojo  
Filip Saidak  
Hari Mohan Srivastava  
Andrew J. Sterge  
Ann Trenk  
Ravi Vakil  
Antonia Vecchio  
John C. Wierman  
Michael E. Zieve



# involve

msp.org/involve

## INVOLVE YOUR STUDENTS IN RESEARCH

*Involve* showcases and encourages high-quality mathematical research involving students from all academic levels. The editorial board consists of mathematical scientists committed to nurturing student participation in research. Bridging the gap between the extremes of purely undergraduate research journals and mainstream research journals, *Involve* provides a venue to mathematicians wishing to encourage the creative involvement of students.

## MANAGING EDITOR

Kenneth S. Berenhaut Wake Forest University, USA

## BOARD OF EDITORS

Colin Adams	Williams College, USA	Chi-Kwong Li	College of William and Mary, USA
Arthur T. Benjamin	Harvey Mudd College, USA	Robert B. Lund	Clemson University, USA
Martin Bohner	Missouri U of Science and Technology, USA	Gaven J. Martin	Massey University, New Zealand
Nigel Boston	University of Wisconsin, USA	Mary Meyer	Colorado State University, USA
Amarjit S. Budhiraja	U of N Carolina, Chapel Hill, USA	Frank Morgan	Williams College, USA
Pietro Cerone	La Trobe University, Australia	Mohammad Sal Moslehian	Ferdowsi University of Mashhad, Iran
Scott Chapman	Sam Houston State University, USA	Zuhair Nashed	University of Central Florida, USA
Joshua N. Cooper	University of South Carolina, USA	Ken Ono	Emory University, USA
Jem N. Corcoran	University of Colorado, USA	Yuval Peres	Microsoft Research, USA
Toka Diagana	Howard University, USA	Y.-F. S. Pétermann	Université de Genève, Switzerland
Michael Dorff	Brigham Young University, USA	Jonathon Peterson	Purdue University, USA
Sever S. Dragomir	Victoria University, Australia	Robert J. Plemmons	Wake Forest University, USA
Joel Foisy	SUNY Potsdam, USA	Carl B. Pomerance	Dartmouth College, USA
Errin W. Fulp	Wake Forest University, USA	Vadim Ponomarenko	San Diego State University, USA
Joseph Gallian	University of Minnesota Duluth, USA	Bjorn Poonen	UC Berkeley, USA
Stephan R. Garcia	Pomona College, USA	József H. Przytycki	George Washington University, USA
Anant Godbole	East Tennessee State University, USA	Richard Rebarber	University of Nebraska, USA
Ron Gould	Emory University, USA	Robert W. Robinson	University of Georgia, USA
Sat Gupta	U of North Carolina, Greensboro, USA	Javier Rojo	Oregon State University, USA
Jim Haglund	University of Pennsylvania, USA	Filip Saidak	U of North Carolina, Greensboro, USA
Johnny Henderson	Baylor University, USA	Hari Mohan Srivastava	University of Victoria, Canada
Glenn H. Hurlbert	Arizona State University, USA	Andrew J. Sterge	Honorary Editor
Charles R. Johnson	College of William and Mary, USA	Ann Trenk	Wellesley College, USA
K. B. Kulasekera	Clemson University, USA	Ravi Vakil	Stanford University, USA
Gerry Ladas	University of Rhode Island, USA	Antonia Vecchio	Consiglio Nazionale delle Ricerche, Italy
David Larson	Texas A&M University, USA	John C. Wierman	Johns Hopkins University, USA
Suzanne Lenhart	University of Tennessee, USA	Michael E. Zieve	University of Michigan, USA

## PRODUCTION

Silvio Levy, Scientific Editor


Cover: Alex Scorpan

See inside back cover or [msp.org/involve](http://msp.org/involve) for submission instructions. The subscription price for 2019 is US \$195/year for the electronic version, and \$260/year (+\$35, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues and changes of subscriber address should be sent to MSP.

Involve (ISSN 1444-4184 electronic, 1444-4176 printed) at Mathematical Sciences Publishers, 798 Evans Hall #3840, c/o University of California, Berkeley, CA 94720-3840, is published continuously online. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices.

Involve peer review and production are managed by EditFlow® from Mathematical Sciences Publishers.

PUBLISHED BY

 **mathematical sciences publishers**  
nonprofit scientific publishing

<http://msp.org/>

© 2019 Mathematical Sciences Publishers

# Orbigraphs: a graph-theoretic analog to Riemannian orbifolds

Kathleen Daly, Colin Gavin, Gabriel Montes de Oca,  
Diana Ochoa, Elizabeth Stanhope and Sam Stewart

(Communicated by Kenneth S. Berenhaut)

A Riemannian orbifold is a mildly singular generalization of a Riemannian manifold that is locally modeled on  $\mathbb{R}^n$  modulo the action of a finite group. Orbifolds have proven interesting in a variety of settings. Spectral geometers have examined the link between the Laplace spectrum of an orbifold and the singularities of the orbifold. One open question in this field is whether or not a singular orbifold and a manifold can be Laplace isospectral. Motivated by the connection between spectral geometry and spectral graph theory, we define a graph-theoretic analog of an orbifold called an orbigraph. We obtain results about the relationship between an orbigraph and the spectrum of its adjacency matrix. We prove that the number of singular vertices present in an orbigraph is bounded above and below by spectrally determined quantities, and show that an orbigraph with a singular point and a regular graph cannot be cospectral. We also provide a lower bound on the Cheeger constant of an orbigraph.

## 1. Introduction

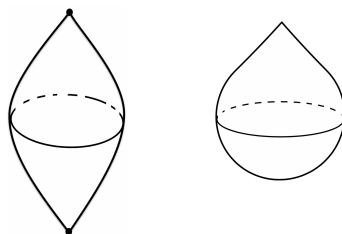
A Riemannian orbifold is a mildly singular generalization of a Riemannian manifold. A point in an  $n$ -dimensional manifold is contained in a neighborhood that is homeomorphic to  $\mathbb{R}^n$ . A point in an  $n$ -dimensional orbifold is contained in a neighborhood that is homeomorphic to a quotient of  $\mathbb{R}^n$  under the action of a finite group. Two useful examples of orbifolds to consider are the  $\mathbb{Z}_n$ -football (Figure 1, left) and the  $\mathbb{Z}_n$ -teardrop (Figure 1, right):

**Example 1.** Let  $\mathbb{Z}_n$  act on a 2-dimensional sphere by rotations generated by a  $2\pi/n$ -radian rotation about an axis passing through the center of the sphere. The quotient of the sphere under this action is the  $\mathbb{Z}_n$ -football. Points lying on the intersection of the sphere with the axis of rotation are fixed by all rotations. The images in the  $\mathbb{Z}_n$ -football of these points are the conical points at the north and

---

*MSC2010:* primary 05C50, 05C20; secondary 60J10.

*Keywords:* graph spectrum, regular graph, directed graph, orbifold.



**Figure 1.** Left: football obtained by 180-degree rotation of sphere.  
Right: teardrop orbifold.

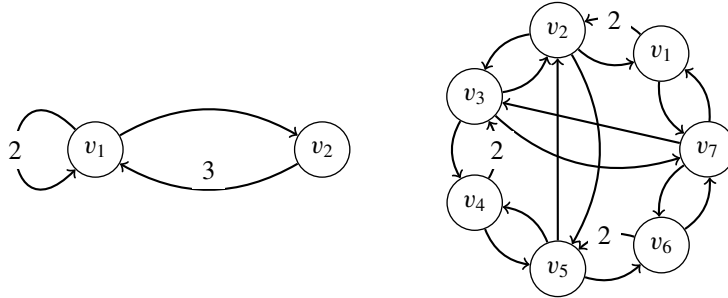
south poles of the football. If the local lift of a point in an orbifold has nontrivial isotropy, the point is called a *singular point* in the orbifold. The singular set of the  $\mathbb{Z}_n$ -football consists of the cone points at its north and south poles.

**Example 2.** The  $\mathbb{Z}_n$ -teardrop is topologically a 2-sphere except for a single point whose neighborhood is locally modeled on the cone  $\mathbb{R}^2/\mathbb{Z}_n$ , where  $\mathbb{Z}_n$  acts by rotations around a fixed point. Thus the  $\mathbb{Z}_n$ -teardrop's singular set consists of the isolated cone point. Thurston [1979] showed that unlike the  $\mathbb{Z}_n$ -football, the  $\mathbb{Z}_n$ -teardrop cannot be obtained as the quotient of a manifold under a smooth, discrete group action.

Introduced by Satake [1956] under the name *V-manifold*, and later renamed and studied as orbifolds by Thurston [1979], orbifolds have proven interesting in a variety of settings; see [Adem et al. 2007; Gordon 2012; Hodgson and Tysk 1993], for example. Of particular interest are results relating the eigenvalue spectrum of the Laplace operator on a Riemannian orbifold (an orbifold endowed with a suitably invariant Riemannian metric) to the singular set of the orbifold. For example, in the presence of a curvature hypothesis, one of us [Stanhope 2005] showed that the Laplace spectrum constrains the structure of the singular set. One fundamental orbifold spectral geometry question that remains open is whether or not the Laplace spectrum actually detects the presence of singular points.

Brooks [1991; 1999] proposes viewing  $k$ -regular graphs as combinatorial analogs of smooth manifolds. The infinite  $k$ -regular tree  $T_k$  is viewed as the graph-theoretic version of the universal cover of a finite  $k$ -regular graph. A finite  $k$ -regular graph  $\Gamma$  is studied as the quotient of  $T_k$  by the fundamental group of  $\Gamma$  in analogy to the study of quotients of the universal cover of a manifold under the action of a discrete cocompact group of isometries acting freely. In this setting Brooks obtains several results including a characterization of Ramanujan graphs, a partial converse to Sunada's theorem, and links between the spectrum of a  $k$ -regular graph and the graph's diameter and girth.

Following Brooks' analogy, observe that the action of a discrete, cocompact group of isometries which is not free yields a quotient space that is an orbifold rather than a manifold. Given the successful examination of orbifolds from the



**Figure 2.** Left: a small 3-orbigraph. Right: a 3-orbigraph with 7 vertices.

perspective of spectral geometry, we seek to extend Brooks' analogy one step further by first proposing a graph-theoretic analog of an orbifold and, second, applying the lens of spectral graph theory to orbifold graphs. References in the literature to an orbifold-like class of graphs are limited. Brooks [1999] himself describes an “orbifold graph” as a quotient of a  $k$ -regular graph under a nonfree group action. He offers orbifold graphs as a motivating idea, but chooses to “avoid entering into the technicalities of ‘orbifold graphs’.” Juan-Pineda, Lafont, Millan-Vossler and Pallekonda [Juan-Pineda et al. 2011] describe an analogy between orbifolds and objects from Bass–Serre theory [Bass 1993] called *graphs-of-groups*. Although the present work has its roots in the ideas of Brooks, the graphs that we examine here can be viewed as a generalization of the edge-index graph of a graph-of-groups.

We define an *orbigraph* to be a member of the following class of weighted, directed graphs.

**Definition 3.** An *orbigraph of degree  $k$  ( $k$ -orbigraph)* is a finite, weighted, directed graph  $\Omega$  where the adjacency matrix  $A$  of  $\Omega$  satisfies the following:

- (i)  $A_{ij} \in \mathbb{Z}_{\geq 0}$ .
- (ii)  $\sum_j A_{ij} = k$ .
- (iii)  $A_{ij} > 0$  if and only if  $A_{ji} > 0$ .

Figure 2 shows two examples of orbigraphs.

**Remark 4.** All orbigraphs discussed below will be assumed to be connected unless noted otherwise. Condition (iii) in Definition 3 implies that a connected orbigraph must be strongly connected. Nonzero diagonal entries in the adjacency matrix of an orbigraph correspond to weighted loops in the orbigraph.

In Section 2 below we demonstrate the analogy between orbigraphs and orbifolds through the following three points:

- (a) The local structure of a vertex in a  $k$ -orbigraph is that of the quotient of a  $k$ -regular graph, just as the local structure of a  $k$ -dimensional orbifold is the quotient of a  $k$ -dimensional manifold.

(b) Some vertices in an orbigraph have the same local structure as a vertex in a regular graph and some do not. This leads us to the definition of regular and singular vertices in an orbigraph — an essential piece of the analogy between orbifolds and orbigraphs.

(c) We show that some orbigraphs can be obtained as the quotient of a finite regular graph under an equitable partition and some cannot. This mirrors the fundamental fact from the geometric setting that orbifolds are divided into two classes: those that are covered by a manifold (like the football) and those that are not (like the teardrop). Indeed, the presence of singular objects that are not merely quotients of regular objects saves the study of orbifolds and orbigraphs from being simply a reduced version of a known field of study.

Section 3 connects orbigraphs to the theory of Markov chains. In Section 4 Markov chain methods are used to obtain a graph-theoretic characterization of when an orbigraph can be obtained as the quotient of a finite regular graph, and when it cannot. This characterization makes it easy to generate examples of orbigraphs with these properties, facilitating our later examination of how spectral results for orbifolds carry over to the orbigraph setting. Also using Markov chain methods we provide a lower bound on the Cheeger constant of a  $k$ -orbigraph in terms of  $k$  and the size of its vertex set. This adds a third family to the list in [Chung 2005] of families of directed graphs that satisfy similar bounds. It would be interesting to know if the bound presented here is sharp, or if an improved bound could be used to obtain a strong upper bound on the convergence of random walks on orbigraphs. Our examination of the Cheeger constant on orbigraphs is the topic of Section 5.

In Section 6 we follow the philosophy of Brooks and ask questions from the spectral geometry of orbifolds in the orbigraph setting. The orbigraph spectrum discussed here is the list of eigenvalues of the adjacency matrix of an orbigraph. Because the analogy between orbifolds and orbigraphs established in Section 2 is strong, the questions carry over naturally and we obtain several interesting results:

(a) We show that the spectrum does not detect whether or not an orbigraph can be obtained as the quotient of a finite  $k$ -regular graph. The analogous question for orbifolds is still an open problem in spectral geometry.

(b) The number of singular points in an orbigraph can be bounded both above and below by spectrally determined quantities. In the geometric setting one can seek spectral bounds on the number of components of the singular set. In dimension 2, the fifth author and Proctor [Proctor and Stanhope 2010] obtained a result of this type under a curvature hypothesis.

(c) The spectrum of an orbigraph detects the presence of singular points. As mentioned above, this question is still open in the orbifold setting.

## 2. Orbigraphs as discrete orbifolds

**2.1. Local structure of a  $k$ -orbigraph.** The local structure of an orbigraph is that of a quotient of a  $k$ -regular graph. There are multiple ways to define the quotienting process for graphs. Here quotient graphs will be formed with respect to an equitable partition. The definition given below uses the approach of Barrett, Francis and Webb [Barrett et al. 2017] to extend the definition of an equitable partition from the familiar setting of simple graphs to the more general setting of weighted directed graphs. We also follow the thorough treatment of the simple graph case in Chapter 5 of [Godsil 1993].

In what follows let  $w(u, v)$  denote the weight of directed edge  $(u, v)$ .

**Definition 5.** Let  $\Gamma$  be a graph (possibly directed, weighted, or both) and

$$\mathcal{P} = \{V_1, V_2, \dots, V_m\}$$

be a partition of its vertices:

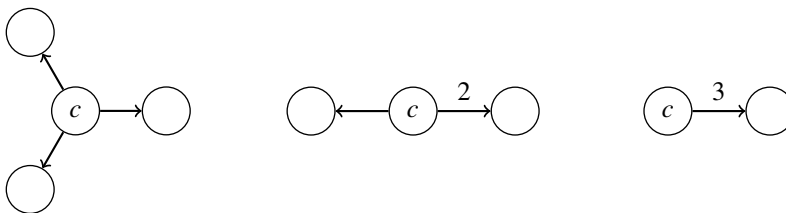
- (a) We say  $\mathcal{P}$  is an *equitable partition* if for all pairs  $i, j$  the number  $\sum_{v \in V_j} w(u, v)$  is the same for each element  $u$  in  $V_i$ .
- (b) Given an equitable partition  $\mathcal{P}$  on  $\Gamma$ , the weighted directed graph with adjacency matrix  $A_{ij} = \sum_{v \in V_j} w(u, v)$ ,  $u$  in  $V_i$ , is called the *quotient graph* of  $\Gamma$  with respect to  $\mathcal{P}$  and will be denoted by  $\Gamma/\mathcal{P}$ .

**Remark 6.** If a group  $G$  acts on a simple graph  $\Gamma$  by automorphisms, the vertex orbits of the action form an equitable partition of the vertex set of  $\Gamma$ . This type of equitable partition is called an *orbit partition*. In this case the quotient graph will be written  $\Gamma/G$ .

To discuss the local structure of an orbigraph we introduce further terms from graph theory. Note that an undirected edge  $\{v, w\}$  of weight  $n$  in a graph will be viewed as being equivalent to a pair of weight- $n$  directed edges  $(v, w)$  and  $(w, v)$ , and vice versa.

- Definition 7.**
- (a) The  *$k$ -star graph* is the complete bipartite graph  $K_{1,k}$  and will be denoted by  $S_k$ . The vertex with degree  $k$  in  $S_k$  is the *central vertex* of  $S_k$ .
  - (b) The *neighborhood* of a vertex  $v$  in an undirected graph  $\Gamma$  is the subgraph of  $\Gamma$  including the vertex  $v$ , all vertices  $w$  adjacent to  $v$ , and all edges  $\{v, w\}$ .
  - (c) The *out-neighborhood* of a vertex  $v$  in a directed graph  $\Delta$  is the directed subgraph of  $\Delta$  including vertex  $v$ , all vertices  $w$  at which edges initiating at  $v$  terminate, and all directed edges  $(v, w)$  with initial vertex  $v$ .

Because the neighborhood of each vertex in a simple  $k$ -regular graph is  $S_k$ , we view a simple  $k$ -regular graph as the graph-theoretic analog of a  $k$ -dimensional manifold.



**Figure 3.** Out-neighborhoods of the central vertex in quotients of  $S_3$ .

Let  $G$  be a group of graph automorphisms of  $S_k$  and form the quotient graph  $S_k/G$ . The central vertex  $c$  of  $S_k/G$  is the vertex in  $S_k/G$  associated to the element of the orbit partition on  $S_k$  containing the central vertex of  $S_k$ . The out-neighborhood of  $c$  in  $S_k/G$  is a weighted star graph with between 1 and  $k$  edges. The sum of the weights over all edges in the out-neighborhood of  $c$  is  $k$ .

**Example 8.** There are only three different weighted, directed graphs that arise as quotients of  $S_3$  by a group of graph automorphisms. Figure 3 illustrates the out-neighborhoods of the central vertex in each of these three quotients.

Because all row sums in the adjacency matrix of a  $k$ -orbigraph  $\Omega$  are  $k$ , the out-neighborhood of a vertex  $v$  in  $\Omega$  is identical to the outgoing neighborhood of the central vertex in some quotient of a  $k$ -star. In this way, a  $k$ -star quotient provides the local model of the neighborhood of a point in an orbigraph. Our interest in the local structure of an orbigraph at a vertex is in the number of outgoing edges and the weights of those edges. The terminal point of an outgoing edge is not important. Because of this the out-neighborhood of a vertex with a loop is taken with the loop “undone”. For example, vertex  $v_1$  in Figure 2, left, is locally modeled on the middle graph in Figure 3.

To complete our analogy between the local structure of orbifolds and the local structure of orbigraphs we observe that requirement (iii) in Definition 3 corresponds to the fact that if local neighborhoods  $U, V$  in an orbifold satisfy  $U \cap V \neq \emptyset$  then we also have  $V \cap U \neq \emptyset$ .

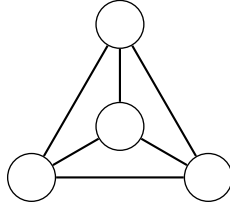
**2.2. Singular points in an orbigraph.** The key feature of the study of orbifolds that distinguishes it from manifold theory is the presence of orbifold singular points. We define a singular vertex in an orbigraph in the following way.

**Definition 9.** A vertex  $v$  of an orbigraph is *singular* if any outgoing edge from  $v$  has weight greater than 1. A vertex that is not singular is called *regular*.

We see that regular graphs contain no singular vertices, as required by our analogy between regular graphs and manifolds.

**Example 10.** Both vertices in the orbigraph in Figure 2, left, are singular. Vertices  $v_1, v_4$  and  $v_6$  in the orbigraph in Figure 2, right, are singular, and the rest are regular.





**Figure 4.** Graph diagram of  $K_4$ .

In contrast to the orbifold setting, singular points in an orbigraph are not marked with an isotropy group. However we can quantify the extent to which a vertex  $v$  is singular by noting the number of outgoing edges from  $v$  that have weight greater than 1. We can also consider the list of weights of outgoing edges from  $v$ . As mentioned in the Introduction, graphs-of-groups offer an alternative graph-theoretic interpretation of orbifolds. A graph-of-groups, in contrast to an orbigraph, has vertices that are marked with a group in a way that is analogous to an orbifold isotropy group.

**2.3. Good and bad orbigraphs.** In Example 1 we saw that the football orbifold is the quotient of a sphere under the smooth action of a finite group. In Example 2 it was asserted that the teardrop orbifold cannot be obtained as a quotient in this manner. Orbifolds that can be written as the quotient of a manifold under a smooth, discrete group action are called *good*. Otherwise they are called *bad*. Following these ideas we define *good* and *bad* orbigraphs as follows.

**Definition 11.** A  $k$ -orbigraph  $\Omega$  is said to be *good* if it can be obtained as the quotient of a finite  $k$ -regular graph  $\Gamma$  via an equitable partition on  $\Gamma$ . If an orbigraph is not good it is called *bad*.

**Example 12.** The orbigraph in Figure 2, left, is good because it is the quotient of the complete graph  $K_4$ , as presented in Figure 4, by the group  $\mathbb{Z}_3$  generated by a  $2\pi/3$ -radian rotation about the center vertex. The orbigraph in Figure 2, right, is bad. This follows from Theorem 20 below and the observation that the product of edge weights along cycle  $(v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_1)$  is 2, while the product of edge weights along the reverse cycle  $(v_1, v_7, v_6, v_5, v_4, v_3, v_2, v_1)$  is 4.

The analogy with the covering theory of topological spaces is further strengthened by the following two lemmas.

**Lemma 13.** *If  $\Omega$  is a  $k$ -orbigraph and  $\mathcal{P}$  is an equitable partition on the vertices of  $\Omega$ , then  $\Omega/\mathcal{P}$  is a  $k$ -orbigraph.*

*Proof.* Let  $A$  denote the adjacency matrix of  $\Omega/\mathcal{P}$ , where  $\mathcal{P} = \{V_1, V_2, \dots, V_m\}$ , and let  $w_\Omega(\cdot, \cdot)$  denote the weight function on directed edges in  $\Omega$ . Because  $\Omega$  is an orbigraph, we know  $w_\Omega(u, v)$  is a nonnegative integer for all vertices  $u, v$

in  $\Omega$ . Hence  $A_{ij} = \sum_{v \in V_j} w_\Omega(u, v)$ , for any  $u \in V_i$ , is a nonnegative integer. Fixing  $i \in \{1, 2, \dots, m\}$ , and taking  $u$  some element of  $V_i$ , consider the  $i$ -th row sum of  $A$ :

$$\sum_j A_{ij} = \sum_j \sum_{v \in V_j} w_\Omega(u, v) = \sum_{v \in \Omega} w_\Omega(u, v) = k.$$

Finally suppose  $A_{ij} > 0$ . Then there must a  $j \in \{1, 2, \dots, m\}$  for which any  $u \in V_i$  has  $w_\Omega(u, v) > 0$  for some  $v \in V_j$ . Because  $\Omega$  is an orbigraph, we must also have  $w_\Omega(v, u) > 0$ . Thus  $A_{ji} > 0$ .  $\square$

**Definition 14.** We say that an orbigraph  $\Omega_1$  covers an orbigraph  $\Omega_2$  if there is an equitable partition  $\mathcal{P}$  of the vertices of  $\Omega_1$  such that  $\Omega_1/\mathcal{P} = \Omega_2$ .

**Lemma 15.** *The covering relation is transitive.*

*Proof.* Suppose  $\Omega_1$  is an orbigraph with equitable partition  $\mathcal{P}_1$  such that  $\Omega_1/\mathcal{P}_1 = \Omega_2$ , and  $\Omega_2$  has an equitable partition  $\mathcal{P}_2$  such that  $\Omega_2/\mathcal{P}_2 = \Omega_3$ . We need to show there is an equitable partition  $\mathcal{P}_3$  of  $\Omega_1$  such that  $\Omega_1/\mathcal{P}_3 = \Omega_3$ . For  $i = 1, 2$  let  $A_i$  denote the adjacency matrix of orbigraph  $\Omega_i$ , and  $P_i$  denote the characteristic matrix corresponding to partition  $\mathcal{P}_i$ . By a straightforward modification of [Godsil 1993, Lemma 2.1, p. 77] to the setting of weighted, directed graphs we have that  $A_1 P_1 = P_1 A_2$  and  $A_2 P_2 = P_2 A_3$ . Thus  $A_1 P_1 P_2 = P_1 A_2 P_2 = P_1 P_2 A_3$ . We conclude  $P_1 P_2$  defines an equitable partition on  $\Omega_1$  with quotient orbigraph  $\Omega_3$ .  $\square$

As a consequence of the previous two lemmas we obtain the following.

**Corollary 16.** *The quotient of any good orbigraph must also be good.*

### 3. Orbigraphs and Markov chains

The fact that the row sum of the adjacency matrix of an orbigraph is constant provides an immediate connection between orbigraphs and Markov chains. Following [Kelly 1979], we review ideas from the theory of Markov chains and introduce notation that will be used hereafter. Matrix  $A$  will denote the adjacency matrix of a  $k$ -orbigraph  $\Omega$  with  $n$  vertices. Define  $P = (1/k)A$ . Matrix  $P$  is the transition matrix of a stationary Markov chain, as all entries of  $P$  lie in the interval  $[0, 1]$  and all rows of  $P$  sum to 1. Because the adjacency matrix of a  $k$ -orbigraph has right eigenvalue  $k$  (to see this consider the eigenvector with all entries equal to 1),  $P$  has right eigenvalue 1 and stationary distribution vector  $\pi = (\pi_1, \pi_2, \dots, \pi_n)$ , with  $\sum_{k=1}^n \pi_k = 1$ , for which  $\pi P = \pi$ . By Remark 4 we know  $\Omega$  is strongly connected so  $\pi$  is the unique stationary distribution of  $P$ .

Our first result connecting orbigraphs to Markov chains is a bound on the minimal entry of  $\pi$  in terms of the degree and number of vertices of an orbigraph.

**Lemma 17.** *Let  $\pi_m$  be a minimal entry in stationary distribution  $\pi$ . Then*

$$\pi_m \geq \frac{1}{nk^{n-1}}.$$

*Proof.* Let  $\pi_M$  denote a maximal entry in  $\pi$  and let  $c$  be the minimal nonzero value that appears as an entry in matrix  $P$ . Because  $\Omega$  is strongly connected, there is a path of length  $\ell < n$  from the  $M$ -th vertex to the  $m$ -th vertex of  $\Omega$ . This implies that  $(P^\ell)_{Mm}$  is nonzero. Using this and the fact that  $\pi P = \pi$ , we have

$$\pi_m = \sum_{k=1}^n (P^\ell)_{km} \pi_k \geq (P^\ell)_{Mm} \pi_M \geq c^\ell \pi_M \geq c^{n-1} \pi_M.$$

Because  $P$  is the transition matrix associated to an orbigraph, we have  $c \geq 1/k$ . Also, we know that  $\pi_M \geq 1/n$  because the sum of the entries of  $\pi$  is 1. Thus  $\pi_m \geq c^{n-1} \pi_M \geq 1/(nk^{n-1})$  as required.  $\square$

Here we relate the stationary distribution of a good orbigraph to that of its finite regular cover.

**Lemma 18.** *Let  $\Gamma$  be a  $k$ -regular graph with  $N$  vertices,  $\mathcal{P} = \{V_1, V_2, \dots, V_n\}$  be an equitable partition of the vertices of  $\Gamma$ , and  $P$  be the transition matrix of the orbigraph  $\Gamma/\mathcal{P}$ . Let  $|V_i|$  denote the number of vertices in partition element  $V_i$ . The stationary distribution of  $P$  is the  $n$ -tuple  $\pi$ , where  $\pi_i = (1/N)|V_i|$ .*

*Proof.* Let  $Q$  denote the transition matrix obtained by scaling the adjacency matrix of  $\Gamma$  by  $1/k$ . The result follows from the observation that the stationary distribution of  $Q$  is the  $N$ -tuple  $(1/N, 1/N, \dots, 1/N)$  and [Godsil 1993, Lemma 2.2, p. 78].  $\square$

#### 4. Characterizing good and bad orbigraphs

We use the Markov chain methods and notation from Section 3 to provide a quick way to distinguish good orbigraphs from bad orbigraphs.

**Definition 19.** An orbigraph  $\Omega$  satisfies the *balanced cycle condition* if the product of the edge weights along each directed cycle  $v_1, v_2, \dots, v_l, v_1$  in  $\Omega$  equals the product of the edge weights along the reverse directed cycle  $v_1, v_l, v_{l-1}, \dots, v_1$ .

**Theorem 20.** *An orbigraph is good if and only if it satisfies the balanced cycle condition.*

A stationary Markov chain is said to satisfy the *detailed balance equations* if

$$\pi_i P_{ij} = \pi_j P_{ji} \quad \text{for all } i, j = 1, 2, \dots, n.$$

The Markov chain analog of the balanced cycle condition from Definition 19 is called the *Kolmogorov criterion*. In particular, an orbigraph satisfies the balanced cycle condition if and only if the corresponding Markov chain satisfies the Kolmogorov criterion. We can now state a needed lemma.

**Lemma 21.** *A stationary Markov chain satisfies the detailed balance equations if and only if it satisfies the Kolmogorov criterion.*

*Proof.* This follows from combining Theorems 1.2 and 1.7 in [Kelly 1979].  $\square$

*Proof of Theorem 20.* Suppose  $\Omega$  is a good orbigraph. This implies  $\Omega = \Gamma/\mathcal{P}$ , where  $\Gamma$  is a  $k$ -regular graph and  $\mathcal{P} = \{V_1, V_2, \dots, V_n\}$  is an equitable partition on  $\Gamma$ . Scaling the adjacency matrix of  $\Gamma$  by  $1/k$  yields the symmetric transition matrix  $Q$  of a Markov chain. We relate the stationary distribution of  $Q$  to the stationary distribution of  $P$ , the transition matrix of  $\Omega$ , by Lemma 18. In particular  $\pi_i = (1/N)|V_i|$ , where  $\pi$  denotes the stationary distribution of  $P$  and  $N$  is the number of vertices in  $\Gamma$ .

The following computation confirms that  $P$  satisfies the detailed balance equations:

$$\begin{aligned} \pi_j P_{ji} &= \frac{1}{N} |V_j| P_{ji} = \frac{1}{N} |V_j| \sum_{k \in V_i} Q_{jk} = \frac{1}{N} \sum_{l \in V_j} \sum_{k \in V_i} Q_{lk} \\ &= \frac{1}{N} \sum_{k \in V_i} \sum_{l \in V_j} Q_{kl} = \frac{1}{N} |V_i| \sum_{l \in V_j} Q_{il} = \pi_i P_{ij}. \end{aligned}$$

(The argument closely follows that of [Tian and Kannan 2006, Theorem 2.16], which is given in the setting of lumpable Markov chains. It makes essential use of the fact that  $\mathcal{P}$  is an equitable partition and that  $Q$  is a symmetric matrix.) The fact that  $\Omega$  satisfies the balanced cycle condition now follows from Lemma 21.

Now suppose  $\Omega$  is an orbigraph that satisfies the balanced cycle condition. By Lemma 21,  $P$  and  $\pi$  satisfy the detailed balance equations  $\pi_i P_{ij} = \pi_j P_{ji}$ . Multiplying by  $k$  on both sides gives  $\pi_i A_{ij} = \pi_j A_{ji}$ . Because  $A$  has all nonnegative integer entries,  $\pi$  will have all nonnegative rational entries. Thus there is an integer  $m$  for which  $m\pi = (d_1, d_2, \dots, d_n)$  is a vector of nonnegative integers. This allows us to write

$$d_i A_{ij} = d_j A_{ji}, \tag{1}$$

an equality of products of nonnegative integers.

We now build a finite  $k$ -regular cover  $\Gamma$  of  $\Omega$ . Let  $X$  be the set of nonzero, nondiagonal entries of  $A$ . Let  $Y = \{A_{11} + 1, A_{22} + 1, \dots, A_{nn} + 1\}$ . Let  $c$  be the least common multiple of the integers in  $X \cup Y$ . For each  $i = 1, 2, \dots, n$  we take  $V_i$  to be a set of  $cd_i$  vertices. The disjoint union  $V_1 \sqcup V_2 \sqcup \dots \sqcup V_n$  forms the vertex set of  $\Gamma$  and gives the needed vertex partition  $\mathcal{P}$  of  $\Gamma$ .

It remains to specify adjacency in  $\Gamma$  in such a way that  $\Gamma/\mathcal{P} = \Omega$ . Suppose  $i \neq j$ . For the quotient  $\Gamma/\mathcal{P} = \Omega$  to be valid, each vertex in  $V_i$  must be adjacent to  $A_{ij}$  vertices in  $V_j$ , and each vertex in  $V_j$  must be adjacent to  $A_{ji}$  vertices in  $V_i$ . Thus the number of edges with one vertex in  $V_i$  and one vertex in  $V_j$ , which we will denote by  $e_{\{i,j\}}$ , is simultaneously  $A_{ij}|V_i|$  and  $A_{ji}|V_j|$ . The adapted detailed

balance equations from (1) show that this requirement follows from our choice for the sizes of  $V_i$  and  $V_j$  as

$$A_{ij}|V_i| = A_{ij}cd_i = A_{ji}cd_j = A_{ji}|V_j|.$$

Because  $A_{ij}$  divides  $|V_j|$  and  $A_{ji}$  divides  $|V_i|$ , we can distribute the  $e_{\{i,j\}}$  edges connecting  $V_i$  and  $V_j$  with exactly  $A_{ij}$  edges adjacent to each vertex in  $V_i$  and exactly  $A_{ji}$  edges adjacent to each vertex in  $V_j$ . Because  $A_{ii} + 1$  divides  $|V_i|$ , we can require that all elements of  $V_i$  are adjacent to exactly  $A_{ii}$  other elements of  $V_i$ . This completes the adjacency relations for  $\Gamma$ .

By construction we observe  $\Gamma/\mathcal{P} = \Omega$ . The degree of a vertex  $v$  in  $\Gamma$  is  $\sum_{j=1} A_{ij} = k$ ; thus  $\Gamma$  is  $k$ -regular. Should  $\Gamma$  fail to be connected, any connected component  $\Gamma'$  of  $\Gamma$  will satisfy  $\Gamma'/\mathcal{P} = \Omega$ .  $\square$

**Remark 22.** Corollary 16 and Theorem 20 imply that if an orbigraph  $\Omega$  satisfies the balanced cycle condition then so does any orbigraph quotient of  $\Omega$ . This stands in contrast to [Tian and Kannan 2006, Example 2.17].

## 5. Bounding the Cheeger constant of an orbigraph

Chung [2005] defined a Cheeger constant for directed graphs and obtained lower bounds on the Cheeger constant for both regular and Eulerian directed graphs. Using  $R$  to denote a  $k$ -regular directed graph on  $n$  vertices and  $E$  an Eulerian directed graph with  $m$  edges, Chung showed

$$h(R) \geq \frac{2}{kn} \quad \text{and} \quad h(E) \geq \frac{2}{m}. \quad (2)$$

Here we apply Chung's methods to obtain a lower bound on the Cheeger constant of an orbigraph. We use notation from Section 3.

Define a function  $F$  from  $\Omega$  to the nonnegative real numbers by

$$F(i, j) = \pi_i P_{ij},$$

where  $i$  and  $j$  are vertices in  $\Omega$ . This function is an example of a *circulation* on  $\Omega$ ; see [Chung 2005, Lemma 3.1]. Letting  $S$  range over all nonempty proper subsets of the vertex set of  $\Omega$ , the Cheeger constant  $h(\Omega)$  of  $\Omega$  is defined as

$$h(\Omega) = \inf_S \frac{\sum_{i \in S, j \notin S} F(i, j)}{\min\{\sum_{j \in S} F(j), \sum_{j \in \bar{S}} F(j)\}},$$

where  $F(j) = \sum_{i, i \rightarrow j} F(i, j)$  and  $\bar{S}$  is the set of vertices of  $\Omega$  that are not in  $S$ .

We have the following lower bound on the Cheeger constant of  $\Omega$ .

**Proposition 23.** *Let  $\Omega$  be a  $k$ -orbigraph with  $n$  vertices. Then*

$$h(\Omega) \geq \frac{2}{n^2 k^n}.$$

*Proof.* We begin by bounding the numerator in the expression defining the Cheeger constant (let  $\pi_m$  denote a minimal entry in  $\pi$ ):

$$\sum_{i \in S, j \notin S} F(i, j) = \sum_{i \in S, j \notin S} \pi_i P_{ij} \geq \sum_{i \in S, j \notin S} \pi_m P_{ij} \geq \frac{1}{nk^n}.$$

The last inequality follows from Lemma 17 and the observation that the smallest possible nonzero value for an entry in  $P$  is  $1/k$ .

To bound the denominator first observe that  $\sum_{j \in S} F(j)$  is no greater than the sum of the columns in  $P$  associated to the vertices in  $S$ . It is similar for  $\sum_{j \in \bar{S}} F(j)$ . Since the total sum of the entries in  $P$  is  $n$ , we have

$$\sum_{j \in S} F(j) + \sum_{j \in \bar{S}} F(j) \leq n.$$

Thus  $\min \left\{ \sum_{v \in S} F(v), \sum_{v \in \bar{S}} F(v) \right\} \leq n/2$ .

We see that for any choice of  $S$  the quotient in the definition of the Cheeger constant must be greater than or equal to  $2/(n^2 k^n)$ , completing the proof.  $\square$

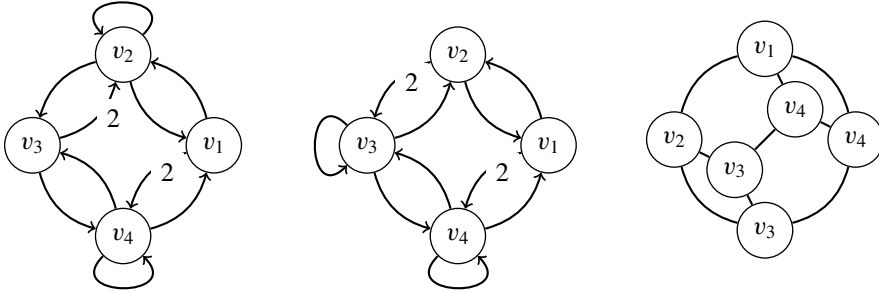
**Remark 24.** Chung uses the inequalities in (2) to obtain convergence bounds for a type of random walk on regular and Eulerian directed graphs. The presence of  $n$  in the exponent in the denominator of the orbigraph bound makes it too weak to obtain a similar orbigraph result. It would be interesting to see if a better bound on the Cheeger constant of an orbigraph, should one exist, would allow a convergence result similar to the regular and Eulerian cases.

## 6. Spectral results for orbigraphs

Because different matrices can be associated to a given graph, a variety of graph spectra are examined in spectral graph theory. Here the *spectrum* of an orbigraph  $\Omega$  is defined to be the list of eigenvalues of the adjacency matrix of  $\Omega$  with each eigenvalue repeated according to its multiplicity. We will write the spectrum of an orbigraph with  $n$  vertices as a multiset  $\{\lambda_1, \lambda_2, \dots, \lambda_n\}$ . The study of the spectral properties of directed graphs is relatively new and has yielded interesting applications, as well as directed graph analogs of familiar graph-theoretical results, including Cheeger's inequality; see [Chung 2005; Langville and Meyer 2006], for example. We focus on developing results that relate the spectrum of an orbigraph to its orbigraph structure.

**Remark 25.** Just as with  $k$ -regular graphs, the spectral radius of a  $k$ -orbigraph is  $k$ . In addition, the number of eigenvalues in the spectrum of an orbigraph (counting multiplicity) is equal to the number of vertices in the orbigraph.

**Lemma 26.** *Suppose orbigraph  $\Omega_1$  covers orbigraph  $\Omega_2$ . Then the spectrum of  $\Omega_2$  is contained in the spectrum of  $\Omega_1$  as multisets.*



**Figure 5.** The left and center orbigraphs are cospectral. The left orbigraph is bad. The center orbigraph is good as it is covered by the right-most graph using the indicated partition.

*Proof.* This follows from the argument in Lemma 2.2 of Chapter 5 in [Godsil 1993], adjusted to allow the graph carrying the equitable partition to be a weighted, directed graph.  $\square$

**Corollary 27.** *Any orbigraph with complex eigenvalues must be bad.*

*Proof.* This follows from Lemma 26 and the fact that regular graphs have real eigenvalues.  $\square$

**Theorem 28.** *The spectrum of an orbigraph does not distinguish good orbigraphs from bad orbigraphs.*

*Proof.* The orbigraph on the left in Figure 5 and the orbigraph in the center of the figure both have spectrum  $\{-2, 0, 1, 3\}$ . However the orbigraph on the left is bad and the orbigraph in the center is good. To see that the left orbigraph is bad, apply Theorem 20 and the fact that the product of the edge weights along cycle  $(v_1, v_2, v_3, v_4)$  is not equal to the product of the edge weights of this cycle reversed. The center orbigraph is good because it is covered by the 3-regular graph on the right side of Figure 5 using the indicated equitable partition.  $\square$

In the following lemma a directed edge from vertex  $v_1$  to vertex  $v_2$  of weight  $w$  is considered to contribute  $w$ -many different ways to move from  $v_1$  to  $v_2$ . The length spectrum of a graph is the finite list of nonnegative integers where the  $m$ -th number in the list counts the number of closed walks of length  $m$  present in the graph.

**Lemma 29.** *The eigenvalue spectrum of an orbigraph determines and is determined by the length spectrum of the orbigraph.*

*Proof.* Let  $\Omega$  be a  $k$ -orbigraph,  $A$  its adjacency matrix, and  $w_m$  the number of closed walks in  $\Omega$  of length  $m$ . We know that

$$w_m = \text{tr}(A^m) \quad (3)$$

because the diagonal of  $A^m$  counts the number of closed walks of length  $m$ . However

$$\mathrm{tr}(A^m) = \sum_{i=1}^n \lambda_i^m.$$

Thus the eigenvalue spectrum of  $\Omega$  uniquely determines the length spectrum of  $\Omega$ , and conversely by Newton's identities [Mead 1992] the length spectrum of  $\Omega$  uniquely determines the eigenvalue spectrum of  $\Omega$ .  $\square$

We now prove that the number of singular points in an orbigraph is bounded above and below by spectrally determined quantities.

**Theorem 30.** *Let  $\Omega$  be a  $k$ -orbigraph with  $n$  vertices. If  $s$  is the number of singular points in  $\Omega$ , then we have*

$$\frac{\sum_{i=1}^n \lambda_i^2 - nk}{k^2 - k} \leq s \leq \sum_{i=1}^n \lambda_i^2 - nk,$$

where  $\lambda_i$  are the eigenvalues of the adjacency matrix  $A$  of  $\Omega$ .

*Proof.* First note that  $\sum_{i=1}^n \lambda_i^2 = \mathrm{tr}(A^2)$  and by Lemma 29 this quantity counts the number of closed walks of length 2 in  $\Omega$ . A given vertex  $v$  in  $\Omega$  has outgoing edges with weights summing to  $k$ , each of which is matched by at least one incoming edge. This implies the number of closed walks of length 2 starting at  $v$  is at least  $k$ . Observing that there are  $n$  vertices in  $\Omega$ , we obtain  $\mathrm{tr}(A^2) \geq nk$ . Now suppose  $v_1$  is a singular vertex in  $\Omega$ . This vertex has at least one outgoing edge  $(v_1, v_2)$  of weight greater than 1. Edge  $(v_1, v_2)$  contributes at least one closed walk of length 2, beginning and ending at  $v_2$ , that has not yet been counted. We conclude that  $\mathrm{tr}(A^2) \geq nk + s$ ; thus  $s \leq \sum_{i=1}^n \lambda_i^2 - nk$ .

For the lower bound, note that each singular vertex  $v_i$  contributes  $A_{ji}(A_{ij} - 1)$  extra (i.e., beyond the initial  $k$  length-2 paths) length-2 paths based at  $v_j$ . Thus the total number of extra paths contributed by vertex  $v_i$  is  $\sum_{v_i \sim v_j} A_{ji}(A_{ij} - 1)$ . We bound this quantity in terms of  $k$ :

$$\sum_{v_i \sim v_j} A_{ji}(A_{ij} - 1) \leq \sum_{v_i \sim v_j} k(A_{ij} - 1) = k \sum_{v_i \sim v_j} A_{ij} - \sum_{v_i \sim v_j} k \leq k^2 - k.$$

Hence each singular vertex contributes at most  $k^2 - k$  extra walks of length 2, so  $s(k^2 - k) \geq \sum_{i=1}^n \lambda_i^2 - nk$ . Isolating  $s$  in this inequality completes the proof.  $\square$

**Remark 31.** The orbigraph with adjacency matrix  $kI_n$ , where  $I_n$  denotes the  $n \times n$  identity matrix, achieves the lower bound in Theorem 30 for all choices of  $k$  and  $n$ . Thus this lower bound is sharp in  $k$  and  $n$ .



**Corollary 32.** *Suppose  $\Omega$  is a  $k$ -orbigraph with  $n$  vertices. Then  $\Omega$  is isomorphic to a  $k$ -regular graph if and only if*

$$\sum_i \lambda_i^2 - nk = 0 \quad \text{and} \quad \sum_i \lambda_i = 0.$$

*Proof.* A simple  $k$ -regular graph  $\Omega$  has no self loops; thus Lemma 29 implies  $\sum_i \lambda_i = 0$ . Viewing each edge  $\{v_i, v_j\}$  in  $\Omega$  as two directed edges,  $(v_i, v_j)$  and  $(v_j, v_i)$ , we see each vertex in  $\Omega$  has exactly  $k$  closed walks of length 2. Therefore  $\sum_i \lambda_i^2 = nk$ .

Conversely, assume that  $\Omega$  is an orbigraph such that  $\sum_i \lambda_i^2 = nk$  and  $\sum_i \lambda_i = 0$ . Then by Theorem 30, we have  $s \leq 0$ . As  $s \geq 0$  we see  $s = 0$ . Thus the outgoing edges of each vertex in  $\Omega$  all have weight 1. The second condition implies  $\Omega$  has no loops. By combining pairs of directed edges  $(v_i, v_j)$  and  $(v_j, v_i)$  into a single undirected edge  $\{v_i, v_j\}$ , we obtain a simple  $k$ -regular graph.  $\square$

In the smooth setting it is not known if a manifold can have the same Laplace spectrum as a nonmanifold orbifold. We can resolve this question in the setting of orbigraphs.

**Corollary 33.** *A regular graph and an orbigraph with one or more singular points cannot be cospectral.*

*Proof.* Suppose regular graph  $\Gamma$  and orbigraph  $\Omega$  are cospectral and that  $\Omega$  contains  $s \geq 1$  singular points. By Remark 25 the largest eigenvalue in the shared spectrum of  $\Gamma$  and  $\Omega$  is the degree of regularity of each graph. Denote this largest eigenvalue by  $k$ . In addition the shared spectrum implies that each graph has the same number of vertices  $n$ . By the forward direction of Corollary 32, the fact that  $\Gamma$  is  $k$ -regular implies  $\sum_i \lambda_i^2 - nk = 0$  and  $\sum_i \lambda_i = 0$ . However the backwards direction of Corollary 32 implies  $s = 0$ , a contradiction.  $\square$

### Acknowledgements

This work was supported in part by the John S. Rogers Science Research Program and the Early Research Program at Lewis & Clark College. The authors would like to thank Omar Lopez and Yung-Pin Chen for foundational work with orbigraphs and discussions of Markov processes, respectively. We also thank the reviewer for helpful suggestions.

### References

- [Adem et al. 2007] A. Adem, J. Leida, and Y. Ruan, *Orbifolds and stringy topology*, Cambridge Tracts in Math. **171**, Cambridge Univ. Press, 2007. MR Zbl
- [Barrett et al. 2017] W. Barrett, A. Francis, and B. Webb, “Equitable decompositions of graphs with symmetries”, *Linear Algebra Appl.* **513** (2017), 409–434. MR Zbl

- [Bass 1993] H. Bass, “Covering theory for graphs of groups”, *J. Pure Appl. Algebra* **89**:1-2 (1993), 3–47. MR Zbl
- [Brooks 1991] R. Brooks, “The spectral geometry of  $k$ -regular graphs”, *J. Anal. Math.* **57** (1991), 120–151. MR Zbl
- [Brooks 1999] R. Brooks, “Non-Sunada graphs”, *Ann. Inst. Fourier (Grenoble)* **49**:2 (1999), 707–725. MR Zbl
- [Chung 2005] F. Chung, “Laplacians and the Cheeger inequality for directed graphs”, *Ann. Comb.* **9**:1 (2005), 1–19. MR Zbl
- [Godsil 1993] C. D. Godsil, *Algebraic combinatorics*, Chapman & Hall, New York, 1993. MR Zbl
- [Gordon 2012] C. Gordon, “Orbifolds and their spectra”, pp. 49–71 in *Spectral geometry* (Hanover, NH, 2010), edited by A. H. Barnett et al., Proc. Sympos. Pure Math. **84**, Amer. Math. Soc., Providence, RI, 2012. MR Zbl
- [Hodgson and Tysk 1993] C. Hodgson and J. Tysk, “Eigenvalue estimates and isoperimetric inequalities for cone-manifolds”, *Bull. Austral. Math. Soc.* **47**:1 (1993), 127–143. MR Zbl
- [Juan-Pineda et al. 2011] D. Juan-Pineda, J.-F. Lafont, S. Millan-Vossler, and S. Pallekonda, “Algebraic  $K$ -theory of virtually free groups”, *Proc. Roy. Soc. Edinburgh Sect. A* **141**:6 (2011), 1295–1316. MR Zbl
- [Kelly 1979] F. P. Kelly, *Reversibility and stochastic networks*, Wiley, Chichester, UK, 1979. MR Zbl
- [Langville and Meyer 2006] A. N. Langville and C. D. Meyer, *Google’s PageRank and beyond: the science of search engine rankings*, Princeton Univ. Press, 2006. MR Zbl
- [Mead 1992] D. G. Mead, “Newton’s identities”, *Amer. Math. Monthly* **99**:8 (1992), 749–751. MR Zbl
- [Proctor and Stanhope 2010] E. Proctor and E. Stanhope, “Spectral and geometric bounds on 2-orbifold diffeomorphism type”, *Differential Geom. Appl.* **28**:1 (2010), 12–18. MR Zbl
- [Satake 1956] I. Satake, “On a generalization of the notion of manifold”, *Proc. Nat. Acad. Sci. USA* **42** (1956), 359–363. MR Zbl
- [Stanhope 2005] E. Stanhope, “Spectral bounds on orbifold isotropy”, *Ann. Global Anal. Geom.* **27**:4 (2005), 355–375. MR Zbl
- [Thurston 1979] W. P. Thurston, “The geometry and topology of three-manifolds”, lecture notes, Princeton Univ., 1979, available at <http://www.msri.org/publications/books/gt3m>.
- [Tian and Kannan 2006] J. P. Tian and D. Kannan, “Lumpability and commutativity of Markov processes”, *Stoch. Anal. Appl.* **24**:3 (2006), 685–702. MR Zbl

Received: 2017-09-02      Revised: 2018-11-10      Accepted: 2019-01-01

daly_kathleen@bah.com	Booz Allen Hamilton, Beavercreek, OH, United States
cgavin@908devices.com	908 Devices, Campbell, CA, United States
gabem@uoregon.edu	Department of Mathematics, University of Oregon, Eugene, OR, United States
dochoa@lclark.edu	Department of Mathematical Sciences, Lewis & Clark College, Portland, OR, United States
stanhope@lclark.edu	Department of Mathematical Sciences, Lewis & Clark College, Portland, OR, United States
sams@umn.edu	School of Mathematics, University of Minnesota, Minneapolis, MN, United States

# Sparse neural codes and convexity

R. Amzi Jeffs, Mohamed Omar, Natchanon Suaysom,  
Aleina Wachtel and Nora Youngs

(Communicated by Ann N. Trenk)

Determining how the brain stores information is one of the most pressing problems in neuroscience. In many instances, the collection of stimuli for a given neuron can be modeled by a convex set in  $\mathbb{R}^d$ . Combinatorial objects known as *neural codes* can then be used to extract features of the space covered by these convex regions. We apply results from convex geometry to determine which neural codes can be realized by arrangements of open convex sets. We restrict our attention primarily to sparse codes in low dimensions. We find that intersection-completeness characterizes realizable 2-sparse codes, and show that any realizable 2-sparse code has embedding dimension at most 3. Furthermore, we prove that in  $\mathbb{R}^2$  and  $\mathbb{R}^3$ , realizations of 2-sparse codes using closed sets are equivalent to those with open sets, and this allows us to provide some preliminary results on distinguishing which 2-sparse codes have embedding dimension at most 2.

## 1. Introduction

One of the fundamental problems of convex geometry is understanding the intersection behavior of convex sets. Classical theorems in this area include Helly's theorem and its many variations, which show that the presence of lower-order intersections of convex sets in  $\mathbb{R}^d$  can force intersections of higher order; see for example [Amenta et al. 2017; Danzer et al. 1963; Eckhoff 1993; Matoušek 2002]. Recent work [Tancer 2013] on the representability of simplicial complexes provides a sharp bound on the dimension in which intersection patterns of convex sets can be realized. We consider the problem of simultaneously realizing intersection patterns along with other relationships between convex sets, such as containment. This problem is motivated by one of the challenges of mathematical neuroscience: determining how the structure of a stimulus space is represented in the brain.

Many types of neurons respond to stimuli in an environment; the set of all such stimuli is called the *stimulus space*  $X$ . Usually, we consider  $X \subset \mathbb{R}^d$ . If we are

---

MSC2010: 05C62, 52A10, 92B20.

Keywords: neural code, sparse, convexity.

considering data from  $n$  neurons  $\{1, \dots, n\}$  which respond to stimuli in  $X$ , the *receptive field* for neuron  $i$  is the subset  $U_i$  of the stimulus space  $X$  for which neuron  $i$  is highly responsive. Throughout this article, we assume the sets  $U_i$  are convex. Indeed, experimental data on many types of neurons, such as place cells [O’Keefe and Dostrovsky 1971] or orientation-tuned neurons [Hubel and Wiesel 1959], make it evident that receptive fields are often well-approximated by convex sets. Hence, for such neurons, the regions of stimulus space in which multiple neurons fire can be modeled by intersections of convex sets, and thus the mathematical theory developed by Helly, Tancer, and others can inform us about the possible arrangements of receptive fields in a given dimension.

Helly’s theorem, however, cannot inform us about all types of receptive field arrangements. For example, if  $U_i, U_j$  are receptive fields which intersect, the neural data will differentiate between  $U_i \subseteq U_j$  and  $U_i \not\subseteq U_j$ , but Helly’s theorem merely notes that  $U_i$  and  $U_j$  intersect. We thus go beyond the usual scope of convex geometry to consider the problem of finding arrangements of convex sets which fully realize the information present in the neural data, including containments. This problem was posed originally in [Curto et al. 2013b], and has been an active area of exploration in recent years. Others such as [Chen et al. 2019; Curto et al. 2017; Cruz et al. 2019; Amzi Jeffs 2018; Amzi Jeffs and Novik 2018] have approached it using methods from algebra, combinatorics, and discrete geometry, but a full solution remains out of reach. In order to address this issue, we first describe how neural data is represented mathematically.

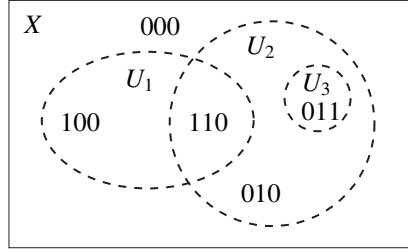
**Definition.** A *neural code* on  $n$  neurons is a set of binary firing patterns  $\mathcal{C} \subset \{0, 1\}^n$ , representing neural activity. Elements of  $\mathcal{C}$  are referred to as *codewords*.

The firing of a neuron is an all-or-nothing event, and so a codeword  $c \in \mathcal{C}$  represents a data point in which a specific set of neurons are simultaneously firing, with neuron  $i$  active if  $c_i = 1$  and inactive if  $c_i = 0$ . For example, the codeword 0011 represents a data point at which neurons 3 and 4 were active, while neurons 1 and 2 were not. In the receptive field context, the presence of this codeword in  $\mathcal{C}$  indicates that  $(U_3 \cap U_4) \setminus (U_1 \cup U_2) \neq \emptyset$ .

**Definition.** Let  $\mathcal{U} = \{U_1, \dots, U_n\}$  be a collection of sets in  $\mathbb{R}^d$ . The *associated neural code*  $\mathcal{C}(\mathcal{U}) \subseteq \{0, 1\}^n$  is the set of firing patterns representing the regions in the arrangement

$$\mathcal{C}(\mathcal{U}) \stackrel{\text{def}}{=} \left\{ c \in \{0, 1\}^n \mid \left( \bigcap_{c_i=1} U_i \right) \setminus \left( \bigcup_{c_j=0} U_j \right) \neq \emptyset \right\}.$$

Any collection of sets  $\mathcal{U}$  in  $\mathbb{R}^d$  gives rise to an associated neural code. However, as we have mentioned, the receptive fields  $U_i$  are generally presumed to be convex. One of our main motivating examples is that of place cells, whose receptive fields



**Figure 1.** An open convex realization of the code  $\mathcal{C} = \{000, 100, 010, 110, 011\}$  in  $\mathbb{R}^2$ , with each region labeled with its corresponding codeword. This shows that  $\mathcal{C}$  is an open convex realizable code with  $d(\mathcal{C}) \leq 2$ . It can be shown that, in fact,  $d(\mathcal{C}) = 1$ .

are generally seen to be convex, as explained in [Curto et al. 2017]. We additionally assume the receptive fields  $U_i$  are open, since by restricting to open sets, we force all sets in our realization to be full-dimensional; furthermore, their intersections, if nonempty, must also be full-dimensional. This allows us to avoid degenerate cases which would not be meaningful in a neural context. These assumptions are consistent with the literature [Curto et al. 2013b; 2017; Lienkaemper et al. 2017]. However, many of our proofs will require that we shift between closed and open convex sets that are associated to the same code. We therefore make the following definition:

**Definition.** If  $\mathcal{U} = \{U_1, \dots, U_n\}$  is a collection of open (respectively, closed) convex sets in  $\mathbb{R}^d$  for which  $\mathcal{C} = \mathcal{C}(\mathcal{U})$ , then we say that  $\mathcal{C}$  is *open (closed) convex realizable in  $\mathbb{R}^d$* , and that  $\mathcal{U}$  is an *open (closed) convex realization* of  $\mathcal{C}$ .

Then, for any code  $\mathcal{C}$ , we define  $d(\mathcal{C})$  to be the minimum dimension  $d$  such that  $\mathcal{C}$  has an open convex realization in  $\mathbb{R}^d$ , if such a dimension  $d$  exists. Figure 1 shows an open convex realization in  $\mathbb{R}^2$  for a code  $\mathcal{C}$  which has minimum dimension  $d(\mathcal{C}) = 1$ . If  $\mathcal{C}$  is not realizable with open convex sets in *any* dimension, we say  $d(\mathcal{C}) = \infty$ . Such codes do exist; see Figure 2.

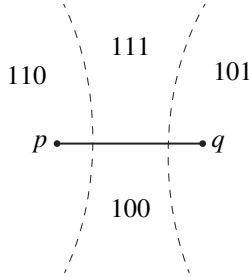
**Definition.** The *support* of a vector  $c \in \{0, 1\}^n$ , denoted by  $\text{supp}(c)$ , is the set of indices of value 1, or the set of all firing neurons:

$$\text{supp}(c) \stackrel{\text{def}}{=} \{i \mid c_i = 1\}.$$

The *support of a code*  $\mathcal{C} \subseteq \{0, 1\}^n$  is the set of the supports of its codewords:

$$\text{supp}(\mathcal{C}) \stackrel{\text{def}}{=} \{\text{supp}(c) \mid c \in \mathcal{C}\}.$$

We assume that there are instances when none of the neurons of interest are firing; hence, we will always assume that the codeword  $00 \cdots 0$  is present in any code.



**Figure 2.** The code  $\mathcal{C} = \{000, 010, 001, 110, 101\}$  is not open convex realizable in  $\mathbb{R}^d$  for any  $d < \infty$ . If it were, we could pick points  $p \in (U_1 \cap U_2) \setminus U_3$  and  $q \in (U_1 \cap U_3) \setminus U_2$ . The line segment  $\overline{pq}$  is contained in  $U_1$  by convexity; to move from  $p$  to  $q$  along  $\overline{pq}$ , we must leave  $U_2$  and enter  $U_3$ . If we leave  $U_2$  before entering  $U_3$  that would indicate the presence of codeword 100, which is not in the code; if we enter  $U_3$  before leaving  $U_2$  that would indicate the codeword 111, which is not in the code. Since all sets are open, these are the only possibilities.

**Example.** Let  $\mathcal{C} = \{000, 101, 110, 111\}$ . Then  $\text{supp}(101) = \{1, 3\}$ ,  $\text{supp}(111) = \{1, 2, 3\}$ , and  $\text{supp}(\mathcal{C}) = \{\emptyset, \{1, 3\}, \{1, 2\}, \{1, 2, 3\}\}$ .

Recent work, for example [Lin et al. 2014], shows the utility and importance of sparsity in neural codes. For practical reasons, our definition of “sparse” differs slightly from the usual low average weight definition often used in coding literature; see for example [Curto et al. 2013a]. We use instead a low maximum weight definition:

**Definition.** A code  $\mathcal{C}$  is  $k$ -sparse if  $|\text{supp}(c)| \leq k$  for all  $c \in \mathcal{C}$ .

We begin the program of studying  $k$ -sparse codes by focusing on 2-sparse codes, where there is already rich mathematics to be found. Our fundamental motivating questions are the following:

**Question 1.1.** Which 2-sparse codes are open convex realizable?

**Question 1.2.** If  $\mathcal{C}$  is an open convex realizable 2-sparse code, what is its minimum embedding dimension  $d(\mathcal{C})$ ?

Our main result is the following characterization of which 2-sparse codes have open convex realizations, including a dimensional bound.

**Theorem 1.3.** A 2-sparse code  $\mathcal{C}$  has an open convex realization if and only if  $\text{supp}(\mathcal{C})$  is intersection-complete. Furthermore, if  $\mathcal{C}$  is realizable then  $d(\mathcal{C}) \leq 3$ .

This answers our first question in its entirety, and partially answers the second. Note that in this result there is no room for generality in terms of sparsity; there are

3-sparse codes that are realizable but not intersection-complete; see for example the code  $\mathcal{C} = \{0, 1\}^3 \setminus \{001\}$  in [Curto et al. 2013b]. In Section 2, we will prove Theorem 1.3 using several lemmas. In particular we show in Lemma 2.6 that for such codes it is equivalent to find a closed convex realization, as it may be transformed to an open convex realization in  $\mathbb{R}^2$  or  $\mathbb{R}^3$ . It immediately follows from this and [Tancer 2013] that any 2-sparse code has a convex open realization in  $\mathbb{R}^3$ . In Section 3, we consider the second question in more detail, and exhibit a class of 2-sparse codes with  $d \leq 2$ , as well as a class with  $d = 3$ .

## 2. Realizability of 2-sparse codes

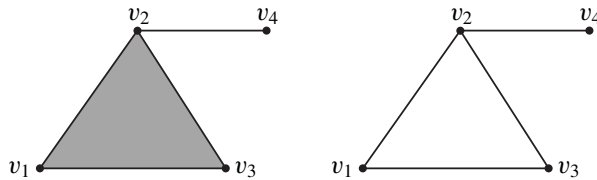
This section is dedicated to proving Theorem 1.3, which establishes that a 2-sparse code is realizable precisely when its support is intersection-complete and, for such codes  $\mathcal{C}$ ,  $d(\mathcal{C}) \leq 3$ . In order to prove this theorem, we make use of the simplicial complex of a code, which is introduced below.

**Definition.** A *simplicial complex* on a finite set  $S$  is a family  $\Delta$  of subsets of  $S$  such that if  $X \in \Delta$  and  $Y \subseteq X$ , then  $Y \in \Delta$ .

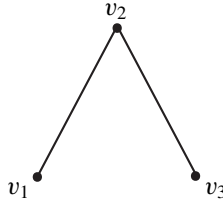
In this paper, the set  $S$  under consideration will most often be  $[n] = \{1, \dots, n\}$ . In a situation where  $S = \{v_1, \dots, v_n\}$ , we will typically refer to any set in a simplicial complex on  $S$  by its set of indices.

**Definition.** The *simplicial complex of a code  $\mathcal{C}$*  is the smallest simplicial complex containing  $\text{supp}(\mathcal{C})$ ; this is denoted by  $\Delta(\mathcal{C})$ . The  *$k$ -skeleton* of a simplicial complex  $\Delta$  is the simplicial complex  $\Delta_k$  given by the collection of sets in  $\Delta$  of size at most  $k + 1$ ; see Figure 3.

If  $\mathcal{C}$  is 2-sparse, then  $\Delta(\mathcal{C})$  consists only of 0-, 1-, and 2-element sets. We can therefore think of  $\Delta(\mathcal{C})$  as a graph, with 1-element sets corresponding to vertices and 2-element sets as edges between them. Note that since  $\Delta(\mathcal{C})$  is a simplicial complex, if  $\{i, j\} \in \Delta(\mathcal{C})$ , then both  $\{i\}$  and  $\{j\}$  must be in  $\Delta(\mathcal{C})$  as well; hence this association is well-defined. The formal relationship between 2-sparse codes and graphs is captured by the following definition.



**Figure 3.** At left, a geometric representation of simplicial complex on  $S = \{v_1, v_2, v_3, v_4\}$  with  $\Delta = \{\emptyset, \{1\}, \{2\}, \{3\}, \{4\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{2, 4\}, \{1, 2, 3\}\}$ . At right, a geometric representation of the 1-skeleton of  $\Delta$ .



**Figure 4.** The graph  $G_C$  for  $\mathcal{C} = \{000, 100, 010, 110, 011\}$ ; see Figure 1 for a realization of  $\mathcal{C}$ .

**Definition.** Let  $\mathcal{C} \subset \{0, 1\}^n$  be a neural code. The *graph of  $\mathcal{C}$* , denoted by  $G_C$ , is the graph whose vertex set is  $[n]$ , with  $i$  adjacent to  $j$  if  $\{i, j\} \subseteq \text{supp}(c)$  for some  $c \in \mathcal{C}$ ; see Figure 4.

Note that  $G_C$  is the 1-skeleton of  $\Delta(\mathcal{C})$ . In particular, for a 2-sparse code,  $\Delta(\mathcal{C})$  and  $G_C$  contain exactly the same information because  $\Delta(\mathcal{C})$  is equal to its 1-skeleton.

As we saw in Figure 2, there exist 2-sparse codes that are not convex in any dimension. The following lemma generalizes the obstruction presented in that figure.

**Lemma 2.1.** *Let  $\mathcal{C}$  be a 2-sparse code. If  $\mathcal{C}$  has a convex open realization in any dimension, then  $\text{supp}(\mathcal{C})$  is intersection-complete.*

*Proof.* Suppose  $\mathcal{C}$  is a 2-sparse code with open convex realization  $\mathcal{U} = \{U_1, \dots, U_n\}$ . Since  $\mathcal{C}$  is 2-sparse,  $|\text{supp}(c)| \in \{0, 1, 2\}$  for every  $c \in \mathcal{C}$ . If  $|\text{supp}(c)|$  is at most 1, then  $\text{supp}(c) \cap \text{supp}(c') \in \text{supp}(\mathcal{C})$  for any  $c' \in \mathcal{C}$ , because the intersection is either  $\emptyset$  or  $\text{supp}(c)$ . It then remains to show that  $\text{supp}(c) \cap \text{supp}(c') = \{i\} \in \text{supp}(\mathcal{C})$  when  $\text{supp}(c) = \{i, j\}$  and  $\text{supp}(c') = \{i, k\}$  with  $j \neq k$ . In this case,  $U_i \cap U_j$  and  $U_i \cap U_k$  are nonempty so there exist points  $p \in U_i \cap U_j$  and  $q \in U_i \cap U_k$ . Consider the line segment  $\overline{pq}$  connecting  $p$  and  $q$ . Since  $U_i$  is convex,  $\overline{pq}$  is contained in  $U_i$ . For each  $m \in [n] \setminus \{i\}$ , consider the set  $L_m = \overline{pq} \cap U_i \cap U_m$ ; note that any two such sets are disjoint, and that  $L_j$  and  $L_k$  are nonempty. If the sets  $\{L_m\}$  partition the line  $\overline{pq}$ , then this would disconnect  $\overline{pq}$  in the subspace topology, but as  $\overline{pq}$  is connected, this is impossible. Thus, there must be some point on  $\overline{pq}$  which is contained in  $U_i$  only. The existence of this point implies  $\{i\} \in \text{supp}(\mathcal{C})$  as desired.  $\square$

The conclusion of the previous lemma is that it is necessary for open convex realizable 2-sparse codes to be intersection complete. In fact, this property characterizes 2-sparse codes with an open convex realization; this is the content of Theorem 1.3. To prove Theorem 1.3, we will use a method of repeatedly making geometric augmentations to existing realizations; in order to make such augmentations without changing the underlying code, we must ensure that subset containment relations between sets are maintained. In the 2-sparse case, the following definition encapsulates the key relationships that must be maintained:

**Definition.** Let  $\mathcal{U} = \{U_1, \dots, U_n\}$  be a collection of sets in  $\mathbb{R}^d$ . For any ordered pair  $(U_i, U_j)$  we distinguish three possible relations between  $U_i$  and  $U_j$ :



**Type A** (disjointness):  $U_i \cap U_j = \emptyset$ ; i.e.,  $\{i, j\} \not\subseteq \text{supp}(c)$  for any  $c \in \mathcal{C}$ .

**Type B** (containment):  $U_j \subseteq U_i$ ; i.e., there exists a codeword  $c \in \mathcal{C}(\mathcal{U})$  so that  $\{i, j\} \subseteq \text{supp}(c)$  and any codeword whose support contains  $j$  must also have  $i$  in its support.

**Type C** (proper intersection):  $U_i \cap U_j$  is nonempty and  $U_j \setminus U_i$  is nonempty; i.e., there exist codewords  $c_1, c_2 \in \mathcal{C}(\mathcal{U})$  so that  $\{i, j\} \subseteq \text{supp}(c_1)$ ,  $j \in \text{supp}(c_2)$  and  $i \notin \text{supp}(c_2)$ .

The type-A, type-B and type-C set relationships effectively characterize the structure of a 2-sparse code; indeed, 2-sparse codes are completely determined by the pairwise relationships of the sets in any realization. We explicitly state this in the following proposition.

**Proposition 2.2.** *Let  $\mathcal{U}$  and  $\mathcal{U}'$  be collections of sets in  $\mathbb{R}^d$  so that  $\mathcal{C}(\mathcal{U})$  and  $\mathcal{C}(\mathcal{U}')$  are both 2-sparse. Then  $\mathcal{C}(\mathcal{U}) = \mathcal{C}(\mathcal{U}')$  if and only if for every ordered pair  $(i, j)$  the relation between  $U_i$  and  $U_j$  is the same as the relation between  $U'_i$  and  $U'_j$ .*

We now introduce the geometric underpinnings of the augmentations we will apply to realizations of codes. In these definitions, we make use of the idea of an  $\varepsilon$ -ball around a point  $p$  ( $B_\varepsilon(p) = \{x \in \mathbb{R}^d \mid \|x - p\| < \varepsilon\}$ ), the interior of a set  $A$  ( $\text{int}(A) = \{x \in A \mid B_\varepsilon(x) \subseteq A \text{ for some } \varepsilon > 0\}$ ), and the closure of a set ( $\bar{A} = \{x \in \mathbb{R}^d \mid x \text{ is a limit point of } A\}$ ).

**Definition.** Given  $\varepsilon > 0$  and  $A \subset \mathbb{R}^d$ , the *trim* of  $A$  by  $\varepsilon$  is the set

$$\text{trim}(A, \varepsilon) \stackrel{\text{def}}{=} \text{int}\{p \in \mathbb{R}^d \mid B_\varepsilon(p) \subseteq A\}.$$

The *inflation* of  $A$  by  $\varepsilon$  is the set

$$\text{inflate}(A, \varepsilon) \stackrel{\text{def}}{=} \{a + x \mid a \in A, x \in \mathbb{R}^d \text{ with } \|x\| < \varepsilon\}.$$

If  $\mathcal{A} = \{A_1, \dots, A_n\}$  is a collection of sets, then

$$\begin{aligned} \text{trim}(\mathcal{A}, \varepsilon) &= \{\text{trim}(A_1, \varepsilon), \dots, \text{trim}(A_n, \varepsilon)\}, \\ \text{inflate}(\mathcal{A}, \varepsilon) &= \{\text{inflate}(A_1, \varepsilon), \dots, \text{inflate}(A_n, \varepsilon)\}. \end{aligned}$$

**Proposition 2.3.** *For any convex set  $A \subset \mathbb{R}^d$  and  $\varepsilon > 0$ , the following statements hold:*

- (1)  $\text{trim}(A, \varepsilon)$  is an open convex set.
- (2)  $\overline{\text{trim}(A, \varepsilon)}$  is contained in the interior of  $A$ .
- (3)  $\text{inflate}(A, \varepsilon)$  is an open convex set.

*Proof.* For (1), we need only prove convexity, and we may assume  $\text{trim}(A, \varepsilon)$  is nonempty. Let  $p$  and  $q$  be points in  $\text{trim}(A, \varepsilon)$ ; then  $B_\varepsilon(p)$  and  $B_\varepsilon(q)$  are contained

in  $A$ , and hence so is the convex hull of their union. This convex hull contains the line segment  $\overline{pq}$ . For (2), note that  $\text{trim}(A, \varepsilon) \subseteq \text{trim}(A, \varepsilon/2) \subseteq \text{int}(A)$ . Finally, (3) follows from the fact that  $A$  is convex and  $\{x \in \mathbb{R}^d \mid \|x\| < \varepsilon\}$  is open and convex.  $\square$

We now show that open convex realizations of 2-sparse codes can be trimmed down to give another open convex realization.

**Lemma 2.4.** *Given a 2-sparse code  $\mathcal{C}$  with an open convex realization  $\mathcal{U} = \{U_1, \dots, U_n\}$ , there exists some  $\varepsilon > 0$  so that  $\text{trim}(\mathcal{U}, \varepsilon)$  is also a realization of  $\mathcal{C}$ .*

*Proof.* Our method is as follows: For each set  $U_i$ , we find an  $\varepsilon_i$  such that  $\text{trim}(U_i, \varepsilon_i) \neq \emptyset$ , and for each pair  $\{i, j\}$  we find an  $\varepsilon_{ij}$  such that  $\text{trim}(\{U_i, U_j\}, \varepsilon_{ij})$  preserves their relationship type (type A, type B or type C). We then let  $\varepsilon$  be the minimum of all  $\varepsilon_i$  and  $\varepsilon_{ij}$ , and show that  $\text{trim}(\mathcal{U}, \varepsilon)$  is a realization of the original code  $\mathcal{C}$ .

To start, for each  $i$  with  $U_i$  nonempty, there must be some point  $p$  and  $\delta_i > 0$  with  $B_{\delta_i}(p) \subseteq U_i$ . Let  $\varepsilon_i = \delta_i/2$ . Let  $\varepsilon_1 = \min_{i \in [n]} \varepsilon_i$ . Now, for each pair  $\{i, j\}$ , we choose  $\varepsilon_{ij}$  depending on the relationship type between  $U_i$  and  $U_j$ :

Type A: If  $U_i \cap U_j = \emptyset$ , set  $\varepsilon_{ij} = \min\{\varepsilon_i, \varepsilon_j\}$ .

Type B: If  $U_i = U_j$ , set  $\varepsilon_{ij} = \min\{\varepsilon_i, \varepsilon_j\}$ . If  $U_i \subsetneq U_j$ , note that  $U_j \setminus U_i$  has nonempty interior. Thus there exists some point  $p$  and some  $\delta_{ij} > 0$  with  $B_{\delta_{ij}}(p) \subseteq U_j \setminus U_i$ . Let  $\varepsilon_{ij} = \min\{\delta_{ij}/2, \varepsilon_i\}$ .

Type C: If  $U_i \cap U_j \neq \emptyset$ , but neither  $U_i \subseteq U_j$  nor  $U_j \subseteq U_i$  is true, note that  $U_i \cap U_j$  is open and therefore there exist a point  $p$  and  $\varepsilon' > 0$  with  $B_{\varepsilon'}(p) \subseteq U_i \cap U_j$ . There exist also points  $p_i, p_j$  in  $U_i \setminus U_j, U_j \setminus U_i$  respectively, with corresponding  $\hat{\varepsilon}$  and  $\tilde{\varepsilon}$  such that  $B_{\hat{\varepsilon}}(p_i) \subseteq U_i \setminus U_j$  and  $B_{\tilde{\varepsilon}}(p_j) \subseteq U_j \setminus U_i$ . Pick  $\varepsilon_{ij} = \min\{\varepsilon_i, \varepsilon_j, \hat{\varepsilon}/2, \tilde{\varepsilon}/2, \varepsilon'/2\}$ .

Let  $\varepsilon_2 = \min_{i,j} \varepsilon_{ij}$ , and finally, let  $\varepsilon = \min\{\varepsilon_1, \varepsilon_2\}$ . Since  $\text{trim}(U, \varepsilon) \subset U$ , and originally there were no triple intersections, by construction it is impossible for  $\text{trim}(\mathcal{U}, \varepsilon)$  to have triple intersections. Thus,  $\mathcal{C}(\text{trim}(\mathcal{U}, \varepsilon))$  is still 2-sparse. We now show that  $\mathcal{C}(\text{trim}(\mathcal{U}, \varepsilon)) = \mathcal{C}$ .

If the codeword with support  $\{i, j\}$  is in  $\mathcal{C}(\text{trim}(\mathcal{U}, \varepsilon))$ , then

$$\text{trim}(U_i, \varepsilon) \cap \text{trim}(U_j, \varepsilon) \neq \emptyset.$$

As  $\text{trim}(U, \varepsilon) \subset U$ , this implies  $U_i \cap U_j \neq \emptyset$ . Since  $\mathcal{C}$  is 2-sparse, the codeword with support  $\{i, j\}$  is in  $\mathcal{C}$ . On the other hand, if the codeword with support  $\{i, j\}$  is in  $\mathcal{C}$ , then  $U_i \cap U_j \neq \emptyset$ , and so we are in a case of type A, B or C above. By our choice of  $\varepsilon$ , we ensure that in each case  $\text{trim}(U_i, \varepsilon) \cap \text{trim}(U_j, \varepsilon) \neq \emptyset$ , and hence (as the code is 2-sparse) the codeword with support  $\{i, j\}$  is in  $\mathcal{C}(\text{trim}(\mathcal{U}, \varepsilon))$ .

If a codeword with support  $\{i\}$  is in  $\mathcal{C}(\text{trim}(\mathcal{U}, \varepsilon))$ , then

$$\text{trim}(U_i, \varepsilon) \setminus \bigcup_{j \in [n], j \neq i} \text{trim}(U_j, \varepsilon) \neq \emptyset.$$

We then know that  $U_i \setminus \bigcup_{j \in [n], j \neq i} U_j \neq \emptyset$ . If it were not, then we would have  $U_i \subseteq \bigcup_{j \in I} U_j$  for some index set  $I$ . However, this is impossible: If  $|I| = 1$ , then  $U_i \subseteq U_j$ , but then  $\text{trim}(U_i, \varepsilon) \subseteq \text{trim}(U_j, \varepsilon)$ . If  $|I| > 1$ , then  $U_i \subseteq \bigcup_{j \in I} U_j$ . But then the 2-sparsity of  $\mathcal{C}$  means we would see the codewords  $\{i, j\}$  and  $\{i, k\}$  in  $\mathcal{C}$  for  $j, k \in I$  but not their intersection  $\{i\}$ , contradicting Lemma 2.1. Hence, the codeword with support  $\{i\}$  is in  $\mathcal{C}$ .

Now, suppose a codeword with support  $\{i\}$  is in  $\mathcal{C}$ , and let  $J = \{j \mid U_i \cap U_j \neq \emptyset\}$ . If  $|J| \leq 1$  then we are in a case of type A, B, or C above, and by our choice of  $\varepsilon$  we know there is a codeword with support  $\{i\}$  in  $\mathcal{C}(\text{trim}(\mathcal{U}, \varepsilon))$ . If  $|J| \geq 2$ , let  $j, k \in J$ . Then by our choice of  $\varepsilon$ , we know  $\text{trim}(U_i, \varepsilon) \cap \text{trim}(U_j, \varepsilon) \neq \emptyset$  and  $\text{trim}(U_i, \varepsilon) \cap \text{trim}(U_k, \varepsilon) \neq \emptyset$ , and hence the codewords with supports  $\{i, j\}$  and  $\{i, k\}$  are in  $\text{trim}(\mathcal{U}, \varepsilon)$ . By Lemma 2.1, we know the codeword with support  $\{i\}$  is also in  $\mathcal{C}(\text{trim}(\mathcal{U}, \varepsilon))$ .  $\square$

Next, we show that a closed convex realization of a 2-sparse codes can be inflated to create an open convex realization.

**Lemma 2.5.** *Let  $\mathcal{C}$  be a 2-sparse code with a closed convex realization  $\mathcal{V} = \{V_1, \dots, V_n\}$  in which every set is bounded. Then there exists some  $\varepsilon > 0$  such that  $\text{inflate}(\mathcal{V}, \varepsilon)$  is an open convex realization of  $\mathcal{C}$ .*

*Proof.* Consider the partial ordering on  $\mathcal{V}$  given by set inclusion. We will use this ordering to inflate the sets in  $\mathcal{V}$  iteratively (possibly by different  $\varepsilon$  factors) and then argue that we can obtain a uniform  $\varepsilon$  for which  $\text{inflate}(\mathcal{V}, \varepsilon)$  is an open convex realization of  $\mathcal{C}$ . In this iterative process, if  $V_i = V_j$  for any  $i \neq j$ , we apply the process simultaneously to  $V_i$  and  $V_j$ . As such, it is sufficient for our proof to assume  $V_i \neq V_j$  for any  $i \neq j$ .

To start, begin with a fixed index  $i$  for which  $V_i$  is maximal in  $\mathcal{V}$  with respect to inclusion. All sets in  $\mathcal{V}$  are closed and bounded, so for any  $j$  with  $V_i \cap V_j = \emptyset$ ,  $V_i$  has positive distance  $d_{i,j}$  to  $V_j$ . Let  $\delta_i = \min_{V_i \cap V_j = \emptyset} d_{i,j}$ . Now if there are  $j, k \neq i$  with  $V_j \cap V_k \neq \emptyset$ , then  $V_i$  has positive distance  $d_{i,j,k}$  to  $V_j \cap V_k$ ; take  $\delta'_i$  to be the minimum of all such  $d_{i,j,k}$ . Furthermore, let  $\delta''_i > 0$  be such that for all  $j$  with  $V_j \not\subseteq V_i$ , we have  $V_j \not\subseteq \text{inflate}(V_i, \delta''_i)$ . Finally, choose  $\varepsilon_i < \min\{\frac{1}{2}\delta_i, \frac{1}{2}\delta'_i, \frac{1}{2}\delta''_i\}$ . These choices help guarantee that no new pairwise or triple intersections are created, and no new containments are created.

If we replace  $V_i$  by  $\text{inflate}(V_i, \varepsilon_i)$ , then the code is still 2-sparse, and the three subset relationship types for the ordered pairs  $(V_i, V_j)$  where  $j \neq i$  are maintained:

Type A: Disjointness is preserved since  $\varepsilon_i$  is at most half the distance from  $V_i$  to any set disjoint from it.

Type B: Containment is preserved since we are only making  $V_i$  bigger.

Type C: Proper intersection is preserved by our choice of  $\varepsilon_i$ .

By a similar argument, the subset relationship of the ordered pair  $(V_j, V_i)$  for any  $j \neq i$  is also preserved after replacing  $V_i$  by  $\overline{\text{inflate}(V_i, \varepsilon_i)}$ . Thus replacing  $V_i$  by  $\overline{\text{inflate}(V_i, \varepsilon_i)}$  yields a new realization of  $\mathcal{C}$ .

For any subsequent step in our iterative process, choose a set  $V_i \in \mathcal{V}$  for which every member of the set  $\{V_j \in \mathcal{V} \mid V_j \supset V_i\}$  has already been inflated. Choose  $\varepsilon_i$  in the same way as previously described with the additional caveat that if  $V_i \subseteq V_j$  then  $\varepsilon_i < \varepsilon_j$ . A similar argument shows that replacing  $V_i$  by  $\overline{\text{inflate}(V_i, \varepsilon_i)}$  yields a new realization of  $\mathcal{C}$ . Once we have inflated every set in the realization we can let  $\varepsilon = \min_{i \in [n]} \varepsilon_i$  and observe that  $\text{inflate}(\mathcal{U}, \frac{1}{2}\varepsilon)$  is an open convex realization of  $\mathcal{C}$ .  $\square$

This result allows us to prove the useful fact that for 2-sparse codes, open and closed convex realizations exist interchangeably, and we can build either type of realization from the other.

**Lemma 2.6.** *Let  $\mathcal{C}$  be a 2-sparse code. Then  $\mathcal{C}$  has an open convex realization in  $\mathbb{R}^d$  if and only if  $\mathcal{C}$  has a closed convex realization in  $\mathbb{R}^d$ .*

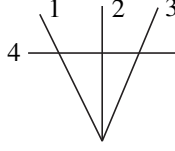
*Proof.* First, let  $\mathcal{U}$  be an open convex realization of  $\mathcal{C}$ . Applying Lemma 2.4, there is an  $\varepsilon > 0$  such that  $\mathcal{U}' = \text{trim}(\mathcal{U}, \varepsilon)$  is an open realization of  $\mathcal{C}$ . Since the closure of each  $U'_i$  is contained in  $U_i$  (by Proposition 2.3),  $\mathcal{U}'$  is an open convex realization of  $\mathcal{C}$  in which two sets intersect if and only if their closures do. Let  $\mathcal{V} = \{\overline{U}'_1, \dots, \overline{U}'_n\}$ . No triple intersections exist in  $\mathcal{V}$  since these would correspond to triple intersections in  $\mathcal{U}$ . Thus by Proposition 2.2 it suffices to show that  $\mathcal{V}$  preserves the relations between sets in  $\mathcal{U}'$ . Disjointness is preserved since sets in  $\mathcal{U}'$  intersect if and only if their closures do. Containment is preserved under taking closures. Lastly, proper intersection is preserved, since if  $U_i \setminus U_j$  is nonempty then there are limit points of  $U_i$  that are not limit points of  $U_j$ .

For the reverse direction, let  $\mathcal{V}$  be a closed convex realization of  $\mathcal{C}$ . For every nonempty intersection  $V_i \cap V_j$ , let  $p_{i,j}$  be a point in this intersection. Furthermore, if some set  $V_i$  is not contained in any other  $V_j$ , let  $p_i \in V_i \setminus \bigcup_{j \neq i} V_j$ . Then set  $V$  to be the convex hull of all these  $p_i$ 's and  $p_{i,j}$ 's. Replacing each  $V_i$  by  $V_i \cap V$  yields a realization of  $\mathcal{C}$  in which every set is closed, convex, and bounded. Applying Lemma 2.5, we obtain an open convex realization of  $\mathcal{C}$  in  $\mathbb{R}^d$ .  $\square$

Although it may not be immediately clear from the proof, the condition that  $\mathcal{C}$  is 2-sparse is necessary for Lemma 2.6 to hold. The 2-sparse condition is in fact best possible, since there exist 3-sparse codes which have closed convex realizations in  $\mathbb{R}^2$ , but for which open convex realizations exist only in  $\mathbb{R}^3$  or higher. One such example is the code

$$\mathcal{C} = \{0000, 1000, 0100, 0010, 0001, 1110, 1001, 0101, 0011\}.$$

Figure 5 shows a closed realization of this code in  $\mathbb{R}^2$ , but it has no open realization in  $\mathbb{R}^2$ ; see [Curto et al. 2017] for more details.



**Figure 5.** A closed realization of a code in  $\mathbb{R}^2$  that has no open realization in  $\mathbb{R}^2$ .

Even more strikingly, there exist codes with a closed convex realization in  $\mathbb{R}^2$  that have no open convex realization in *any* dimension; see [Lienkaemper et al. 2017] for an example of such a code on five neurons. This emphasizes how special realizations of 2-sparse codes are.

We can now use the previous lemmas to relate the convexity of a 2-sparse code  $\mathcal{C}$  to the convexity of its associated simplicial complex  $\Delta(\mathcal{C})$ . We first need a technical lemma.

**Lemma 2.7.** *Let  $\mathcal{U}$  be an open convex realization of a 2-sparse code  $\mathcal{C}$ . Then if  $U_j \not\subseteq U_k$  for any  $k \neq j$ , there is a point  $p \in \partial U_j \setminus \bigcup_{k \neq j} U_k$ .*

*Proof.* Recall that for any set  $U \subset \mathbb{R}^d$ ,  $\partial U$  is the boundary of  $U$ . Consider the sets  $\{\partial U_j \cap U_k\}_{k \neq j}$ . These sets are disjoint: if not, then there exists  $p \in (\partial U_j \cap U_k) \cap (\partial U_j \cap U_\ell)$ . As  $p \in U_k \cap U_\ell$ , there exists  $\varepsilon > 0$  with  $B_\varepsilon(p) \subseteq U_k \cap U_\ell$ . But then  $B_\varepsilon(p) \cap U_j \neq \emptyset$ , as  $p \in \partial U_j$ , so  $U_j \cap U_k \cap U_\ell \neq \emptyset$  contradicting that  $\mathcal{C}$  is 2-sparse.

Now, note that the disjoint sets  $\{\partial U_j \cap U_k\}_{k \neq j}$  are open in the subspace topology with respect to  $\partial U_j$ , and hence they cannot partition  $\partial U_j$  since  $\partial U_j$  is connected. Thus, there exists  $p \in \partial U_j \setminus \bigcup_{k \neq j} U_k$ .  $\square$

**Lemma 2.8.** *Let  $\mathcal{C}$  be a 2-sparse code and let  $d \geq 2$ . Then  $\mathcal{C}$  has an open convex realization in  $\mathbb{R}^d$  if and only if  $\text{supp}(\mathcal{C})$  is intersection-complete and  $\Delta(\mathcal{C})$  has an open convex realization in  $\mathbb{R}^d$ .*

*Proof.* For the forward direction, we know from Lemma 2.1 that if  $\mathcal{C}$  has a realization then  $\text{supp}(\mathcal{C})$  is intersection-complete. We will show that given a realization  $\mathcal{U}$  of  $\mathcal{C}$ , we can construct a realization of  $G_{\mathcal{C}}$ . Since  $\mathcal{C}$  is 2-sparse, we know  $\mathcal{C}$  and  $\Delta(\mathcal{C})$  must already contain the same 2-element sets, so we will show that we can adjust the realization of  $\mathcal{C}$  to obtain any singletons  $\{i\}$  which appear in  $\Delta(\mathcal{C})$  but not in  $\mathcal{C}$ .

Let  $\{i\} \in \Delta(\mathcal{C}) \setminus \text{supp}(\mathcal{C})$ . If there exist  $j, k$  such that  $\{i, j\}$  and  $\{i, k\}$  are both in  $\text{supp}(\mathcal{C})$ , then as  $\text{supp}(\mathcal{C})$  is intersection-complete, we know  $\{i\} \in \text{supp}(\mathcal{C})$ . Thus, there must be exactly one  $j$  such that  $\{i, j\} \in \text{supp}(\mathcal{C})$ . Note immediately that in the realization  $\mathcal{U}$  we have  $U_i \subseteq U_j$  since  $\{i, j\}$  is the only set in the support where  $i$  appears. It suffices to transform  $\mathcal{U}$  so that  $U_i$  and  $U_j$  intersect, but  $U_i$  also contains points not in any other set in the realization.

If we have  $U_j \subseteq U_i$ , then  $U_i = U_j$  so  $U_j \cap U_k = \emptyset$  for any other  $k$ , and we can replace  $U_j$  with an open ball properly contained in  $U_i$  to obtain the desired result.

Otherwise,  $U_j$  may intersect many other sets in the realization, but cannot be contained in them, since this would imply a triple intersection between the containing set  $U_k$ ,  $U_j$ , and  $U_i$ . Apply Lemma 2.4 to obtain  $\varepsilon > 0$  for which  $\mathcal{U}' = \text{trim}(\mathcal{U}, \varepsilon)$  is an open realization of  $\mathcal{C}$ . Define the sets  $V_k = \partial U'_j \cap \bar{U}'_k$ ; note that each  $V_k$  is closed. Furthermore, these sets are disjoint, since if  $p \in V_k \cap V_\ell$ , then  $p \in U_j \cap U_k \cap U_\ell$  in the original realization which is impossible for a 2-sparse code. Since  $\partial U'_j$  is connected and the  $V_k$  are disjoint closed sets,  $\bigcup_{k \neq j} V_k \subsetneq \partial U'_j$ ; let  $p \in \partial U'_j \setminus \bigcup_{k \neq j} V_k$ . Then  $p$  has positive distance to all sets  $U'_k$  with  $k \neq j$  so there is some  $\varepsilon' > 0$  with  $B_{\varepsilon'}(p) \cap U'_k = \emptyset$  for all  $k \neq j$ . Replacing  $U'_i$  with  $B_{\varepsilon'}(p)$  will create a realization of a code  $\mathcal{C}'$  with  $\text{supp}(\mathcal{C}') = \text{supp}(\mathcal{C}) \cup \{i\}$ . Repeating this step as many times as necessary, we obtain a realization of  $\Delta(\mathcal{C})$ .

For the reverse, suppose  $\mathcal{U}$  is an open convex realization of  $\Delta(\mathcal{C})$ . Note that if  $\{i, j\} \in \text{supp}(\Delta(\mathcal{C}))$ , it is also in  $\text{supp}(\mathcal{C})$  since  $\mathcal{C}$  is 2-sparse. Now, suppose  $\{i\} \in \text{supp}(\Delta(\mathcal{C})) \setminus \text{supp}(\mathcal{C})$ . Then there is at most one  $j \neq i$  such that  $\{i, j\} \in \text{supp}(\mathcal{C})$  as  $\mathcal{C}$  is intersection-complete. If there is such a  $j$ , replace  $U_i$  with  $U_i \cap U_j$  which is an open convex set; if there is no such  $j$ , then remove  $U_i$  entirely. This gives a convex realization of  $\Delta(\mathcal{C}) \setminus \{i\}$ , and we can repeat this operation as many times as necessary to obtain a realization of  $\mathcal{C}$ .  $\square$

The above lemma can be summarized as follows: realizing a 2-sparse code and realizing its simplicial complex are equivalent, as long as  $\text{supp}(\mathcal{C})$  is intersection-complete. This equivalence is our main tool in proving Theorem 1.3 and obtaining a complete classification of which 2-sparse codes are convex in  $\mathbb{R}^3$ .

*Proof of Theorem 1.3.* The fact that any open convex realizable 2-sparse code must have  $\text{supp}(\mathcal{C})$  that is intersection-complete follows directly from Lemma 2.1. For the reverse direction, since  $\text{supp}(\mathcal{C})$  is intersection-complete, we know by Lemma 2.8 that it is sufficient to find an open convex realization for  $\Delta(\mathcal{C})$ . As  $\mathcal{C}$  is 2-sparse, Lemma 2.6 tells us that it suffices to find a closed convex realization for  $\Delta(\mathcal{C})$ . Since  $\Delta(\mathcal{C})$  is a 1-dimensional simplicial complex, a construction of [Tancer 2013] (see the proof of Theorem 3.1 therein) leads to a closed convex realization of a 1-dimensional simplicial complex in  $\mathbb{R}^3$ . This proves the desired result.  $\square$

Theorem 1.3 makes it very straightforward to check whether a 2-sparse code has an open convex realization in  $\mathbb{R}^3$ . The challenge that lies ahead is determining the minimal embedding dimension for a given 2-sparse code. We begin investigating this problem in the next section.

### 3. Dimension of 2-sparse codes

We noted early on that for 2-sparse codes, the simplicial complex  $\Delta(\mathcal{C})$  and the graph  $G_{\mathcal{C}}$  of the pairwise intersections of the code capture the same information. In this section, we will make heavy use of this correspondence, and construct realizations

of various 2-sparse codes using graph-theoretic methods. Hence, throughout this discussion we will often refer to “realizations” of a graph  $G_C$ . It is important to note that while a graph is the intersection graph of its realization, finding convex sets whose intersection graph is the graph of concern is not sufficient here. In particular, if a collection of convex sets has a triple with nonempty intersection then it is not, for our purposes, a realization of any graph, since graphs only encode intersections of order 2.

Our main result, Theorem 1.3, shows that any intersection-complete 2-sparse code can be realized in dimension  $d \leq 3$ . In this section, we begin the program of classifying 2-sparse codes based on minimal embedding dimension. We focus on distinguishing codes of dimension  $d = 3$  from codes of dimension  $d \leq 2$ ; note that the general problem of distinguishing 1-dimensional codes has been solved [Rosen and Zhang 2017]. Recall from Lemma 2.8 that realizing a 2-sparse code  $\mathcal{C}$  is equivalent to realizing its simplicial complex  $\Delta(\mathcal{C})$  (and therefore, its graph  $G_C$ ), so throughout this section we refer to realizing  $G_C$  rather than  $\mathcal{C}$  itself. Our main contribution is that while the dimension of certain graphs can be bounded, we find that the traditional 2-dimensional graph-theoretic distinction (planarity) is not necessary for  $G_C$  to represent a 2-dimensional code. In particular, in Proposition 3.1, we observe  $d(\mathcal{C}) \leq 2$  if  $G_C$  is planar, and in Proposition 3.2 if  $G_C$  is not planar, one can construct a *related* graph whose code has minimal embedding dimension 3. However, planarity does not strictly govern minimal embedding dimension, as any complete or complete bipartite graphs are realizable in  $\mathbb{R}^2$ .

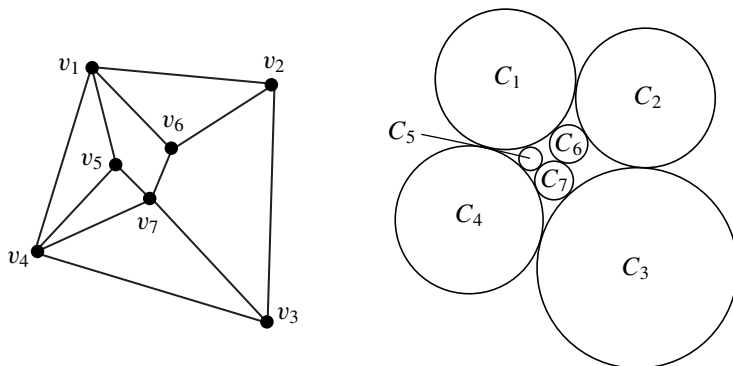
The following proposition describes some common graphs which do have 2-dimensional convex realizations, including planar graphs.

**Proposition 3.1.** *The following graphs have an open convex realization in  $\mathbb{R}^2$ :*

- (1) *planar graphs,*
- (2) *the complete  $k$ -partite graph  $K_{n_1, n_2, \dots, n_k}$  with part sizes  $n_1, n_2, \dots, n_k$ ,*
- (3) *any graph  $G$  with vertex set  $\{v_1, v_2, \dots, v_n, u_1, \dots, u_k\}$  where the induced subgraph on the vertices  $v_1, v_2, \dots, v_n$  is complete and  $\{v_1, v_2, \dots, v_n\} \supseteq N_G(u_k) \supseteq N_G(u_{k-1}) \supseteq \dots \supseteq N_G(u_1)$ .*

*Proof.* In all cases, we find a closed convex realization of the given graph  $G$ , which by Lemma 2.6 implies the existence of an open convex realization. For (1), we first recall the circle packing theorem, which says that for any planar graph  $G$  with vertex set  $\{v_1, \dots, v_n\}$ , there exist disjoint disks  $C_1, C_2, \dots, C_n$  in  $\mathbb{R}^2$  such that  $C_i$  is tangent to  $C_j$  if and only if  $v_i$  is adjacent to  $v_j$ , and  $C_i \cap C_j = \emptyset$  otherwise. See Figure 6 for an illustration of how these disks are constructed.

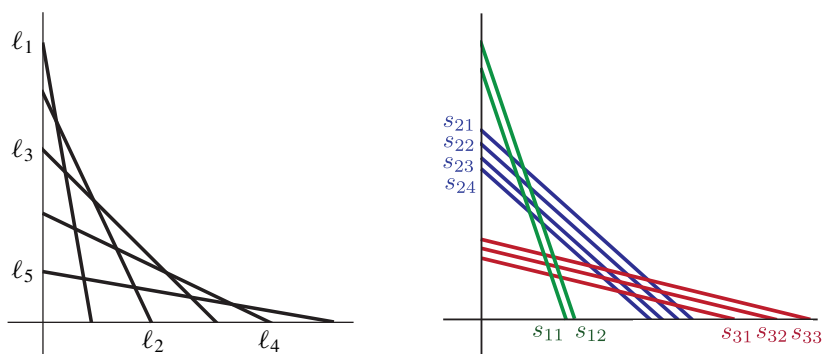
For (2), we first find a realization for the complete graph  $K_n = K_{1,1,\dots,1}$  ( $n$  copies of 1 here). Consider the line segments  $\ell_1, \ell_2, \dots, \ell_n$ , where  $\ell_i$  has endpoints  $(i, 0)$



**Figure 6.** A planar graph  $G$  and the corresponding closed realization using the circle packing theorem.

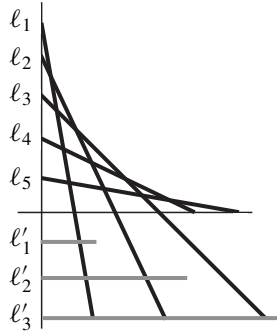
and  $(0, n + 1 - i)$ , and observe that  $\ell_i \cap \ell_j \neq \emptyset$  for any  $i \neq j$ . Moreover, no three of these lines are concurrent. This gives a closed convex realization of  $K_n$ . Now to realize  $K_{n_1, n_2, \dots, n_k}$ , start with a closed convex realization of  $K_k$  as constructed in the realization of (2). Replace each line segment  $\ell_i$  with  $n_i$  disjoint parallel translates of  $\ell_i$  that are arbitrarily close in distance to  $\ell_i$ , and call these segments  $s_{i1}, s_{i2}, \dots, s_{in_i}$ . Observe that by construction,  $s_{ij} \cap s_{ij'} = \emptyset$  for any  $j \neq j'$ . Moreover,  $s_{ij} \cap s_{i'j'} \neq \emptyset$  for  $i \neq i'$  because  $\ell_i \cap \ell_{i'} \neq \emptyset$  and  $s_{ij}$  and  $s_{i'j'}$  are arbitrarily close and parallel to  $\ell_i$  and  $\ell_{i'}$  respectively. Moreover, if any three line segments  $s_{ij}, s_{i'j'}, s_{i''j''}$  had a point in common, then  $\ell_i, \ell_{i'}, \ell_{i''}$  would, which they don't. Hence the union of the sets  $\{s_{i1}, s_{i2}, \dots, s_{in_i}\}_{i=1}^k$  gives a closed convex realization of  $K_{n_1, n_2, \dots, n_k}$ . See Figure 7 for examples of the constructions in the proof of (2).

It remains to prove (3). Without loss of generality, we assume  $N_G(u_k) = \{v_1, v_2, \dots, v_r\}$ , indexed in such a way that each set  $N_G(u_j)$  is  $\{v_1, v_2, \dots, v_s\}$



**Figure 7.** A closed convex realization of  $K_5$  (left) and a closed convex realization of  $K_{2,4,3}$  (right), as constructed in the proof of Proposition 3.1.





**Figure 8.** A closed convex realization of the graph  $G$  with vertices  $v_1, v_2, v_3, v_4, v_5, u_1, u_2, u_3$ , where the induced graph on  $v_1, \dots, v_5$  is complete, and  $N(u_3) = \{v_1, v_2, v_3\}$ ,  $N(u_2) = \{v_1, v_2\}$  and  $N(u_1) = \{v_1\}$ .

for some  $s$ . To realize  $G$ , first start with a realization of  $K_n$  as in the proof of (2), where  $v_j$  is represented by  $\ell_j$  for each  $j$ . Now, extend each line segment  $\ell_j$  for  $1 \leq j \leq r$  so that  $(0, j)$  remains as an endpoint, the slope remains the same, but the lower endpoint has  $y$ -coordinate  $-k$ . Then, for each  $s$  with  $1 \leq s \leq k$ , introduce a line segment  $\ell'_s$  that lies on the line in the  $xy$ -plane given by  $y = s$ , and only intersects the line segments in the set  $\{\ell'_j \mid j \in N_G(u_s)\}$ . The line segments  $\ell_1, \dots, \ell_n, \ell'_1, \dots, \ell'_k$  give a closed realization of  $G$ . See Figure 8 for an example of this construction.  $\square$

Thus far, we have exhibited classes of graphs that can be realized in  $\mathbb{R}^2$ , including any planar and some nonplanar graphs  $G_C$ . We now show how to adjust any nonplanar graph by edge subdivision to create a new graph that cannot be realized in  $\mathbb{R}^2$ .

**Proposition 3.2.** *Let  $G$  be a nonplanar graph. Let  $G'$  be the graph obtained from  $G$  by replacing each edge  $v_i v_j$  by a length-2 path  $v_i, v_{ij}, v_j$  (we refer to this as the edge subdivision of  $G$  throughout). Then  $G'$  does not have an open convex realization in  $\mathbb{R}^2$ , and hence its minimal embedding dimension is 3.*

*Proof.* Suppose by contradiction that  $G'$  has an open convex realization in  $\mathbb{R}^2$ . Let the graph  $G$  have vertex set  $\{v_1, v_2, \dots, v_n\}$ , so  $G'$  has as its vertices  $\{v_i \mid i = 1, \dots, n\}$  together with vertices  $\{v_{ij} \mid v_i v_j \in E(G)\}$ , where for any  $i, j$ , the vertex  $v_{ij}$  is adjacent only to  $v_i$  and  $v_j$ . Suppose the open convex realization  $\mathcal{U}$  of  $G'$  consists of the sets  $\{U_i\}$  and  $\{U_{ij}\}$ , where for any  $i$ ,  $U_i$  is the open convex set corresponding to  $v_i$ , and for any  $i \neq j$  with  $v_i v_j \in E(G)$ ,  $U_{ij}$  is the open convex set corresponding to  $v_{ij}$ .

First, for all  $i = 1, \dots, n$  select a point  $p_i$  in  $U_i$  that does not lie in any other sets in  $\mathcal{U}$ . Then, for every pair  $i, j$  such that  $v_i$  and  $v_j$  are adjacent in  $G$ , note that  $U_i \cap U_{ij}$  and  $U_j \cap U_{ji}$  are nonempty, so we can also select points  $x_{ij}$  and  $x_{ji}$  in  $U_i \cap U_{ij}$  and  $U_j \cap U_{ji}$ , respectively. Let the line segment  $x_{ij} x_{ji}$  intersect  $\partial U_i$

and  $\partial U_j$  at points  $p_{ij}$  and  $p_{ji}$ , respectively. Define the path  $P_{ij}$  from  $p_i$  to  $p_j$  by concatenating the line segments  $p_i p_{ij}$ ,  $p_{ij} p_{ji}$ , and  $p_{ji} p_j$  in that order.

Now consider another pair of indices  $k, l$ . We claim that two different paths  $P_{ij}$  and  $P_{kl}$  can only intersect at the points  $p_i, p_j, p_k$  or  $p_l$ , if anywhere. To show this, it is enough to show that among any pair of line segments, one chosen from  $\{p_i p_{ij}, p_{ij} p_{ji}, p_{ji} p_j\}$  and one from  $\{p_k p_{kl}, p_{kl} p_{lk}, p_{lk} p_l\}$ , their intersection (if it exists), must be one of the points  $p_i, p_j, p_k$  or  $p_l$ . We split this into three cases:

First, consider the intersection of  $p_i p_{ij}$  and  $p_k p_{kl}$ . If  $i = k$  then the two segments can only intersect at  $p_i$ , unless  $j = l$ , in which case the segments were the same segments to begin with. If  $i \neq k$ , then observe that  $p_i p_{ij} \in U_i$ ,  $p_k p_{kl} \in U_k$  and  $U_i \cap U_k$  is empty because  $v_i$  and  $v_k$  are not adjacent in  $G'$ . A similar argument establishes our desired result when the pair of segments in question are  $\{p_i p_{ij}, p_{kl} p_k\}$ ,  $\{p_{ij} p_i, p_{kl} p_k\}$  and  $\{p_{ij} p_i, p_k p_{kl}\}$ .

Second, consider the intersection of  $p_{ij} p_{ji}$  and  $p_{kl} p_{lk}$ . Notice that  $p_{ij} p_{ji} \subseteq U_{ij}$  and  $p_{kl} p_{lk} \subseteq U_{kl}$ . Since  $v_{ij}$  and  $v_{kl}$  are not adjacent in  $G'$ ,  $U_{ij} \cap U_{kl}$  is empty, so the two paths in question cannot intersect.

Finally, consider the intersection of  $p_i p_{ij}$  and  $p_{kl} p_{lk}$ . Suppose that  $i = k$ . When  $j = l$ , the segments in question are  $p_i p_{ij}$ ,  $p_{ij} p_{ji}$  but these are from the same path  $P_{ij}$  so we need not consider this situation. When  $j \neq l$ ,  $p_i p_{ij} \subseteq U_i \cup U_{ij}$ , and  $p_{il} p_{li} \subseteq U_{il} \setminus U_i$ . Since  $j \neq l$ ,  $U_{ij} \cap U_{il} = \emptyset$ , and hence  $(U_i \cup U_{ij}) \cap (U_{il} \setminus U_i) = \emptyset$ , so the two segments in question do not intersect. A similar argument establishes the result when  $j = l$ . It remains to establish the desired result when  $i \neq l, k$ . Suppose for a contradiction that  $p_i p_{ij}$  intersects  $p_{kl} p_{lk}$ . Since  $p_i p_{ij} \subseteq U_i \cup \partial U_i$ , and  $p_{kl} p_{lk} \subseteq U_{lk}$ , this implies  $(U_i \cup \partial U_i) \cap U_{lk}$  is nonempty. However, this is impossible because  $U_i \cap U_{lk} = \emptyset$  (because  $v_i$  and  $v_{lk}$  are not adjacent in  $G'$ ) and  $\partial U_i \cap U_{lk} = \emptyset$ .

The above argument establishes that two distinct paths  $P_{ij}$ ,  $P_{kl}$  can only intersect at their endpoints. Construct a graph  $G''$  on the same vertex set as  $G$  with two vertices adjacent precisely when they are adjacent in  $G$ , but with each edge  $v_i v_j$  drawn precisely along the path  $P_{ij}$ . The graph  $G''$  is a planar embedding of  $G$ , contradicting that  $G$  is not planar.  $\square$

#### 4. Future directions

This paper initiated the program of studying  $k$ -sparse codes, with a full characterization of the structure of 2-sparse codes. Section 2 was dedicated to a topological and analytic investigation of such codes in order to achieve a full characterization of realizability through Theorem 1.3, which additionally told us that any realizable 2-sparse code has minimal embedding dimension at most 3. Section 3 then began the study of differentiating 2-sparse codes by embedding dimension through Propositions 3.1 and 3.2. The most pressing questions are how these investigations,

which relied heavily on the graph-like structure of these codes, could generalize when  $k > 2$ .

**Question 4.1.** *For a particular  $k$ , how can we characterize which  $k$ -sparse codes are realizable? More specifically, given a positive integer  $\ell$ , for which  $k$ -sparse codes is  $d(\mathcal{C}) = \ell$ ?*

In investigating the minimum embedding dimension of a  $k$ -sparse code, certain dimension bounds can be used. For example, suppose  $\mathcal{C}$  is a  $k$ -sparse code with  $\Delta = \Delta(\mathcal{C})$ , and let  $f_d(\Delta)$  be the number of codewords in  $\Delta$  with support size  $d + 1$ . Then, by applying the fractional Helly theorem, we find  $k > f_d(\Delta) / \binom{n-1}{d}$ ; this was noted in [Curto et al. 2017]. Similar to this, many known bounds rely solely on the combinatorial information in the code and in particular the simplicial complex  $\Delta(\mathcal{C})$ . While often dimension bounds are the best known results, a more specific investigation in [Rosen and Zhang 2017] gives a full characterization of 1-dimensional codes. Our work thus focuses on distinctions between dimensions 2 and 3 for 2-sparse codes, as a beginning step towards a characterization of 2-dimensional codes.

However, in addressing the question of whether a  $k$ -sparse code is realizable at all, an investigation into the topology can provide insight beyond what is apparent from the combinatorics. This is especially evident from the developments in Section 2. The key idea there was shifting from one realization of a code to another by shrinking or expanding sets. Indeed, this method has been applied with more generality and great success in [Cruz et al. 2019]. The question then for  $k$ -sparse codes for  $k > 2$  is what analogous topological operations to realizations preserve the underlying code.

**Question 4.2.** *Given a convex realization  $\mathcal{U} = \{U_1, \dots, U_n\}$  of a code  $\mathcal{C}$  in  $\mathbb{R}^d$ , what topological maps can be applied to the sets  $U_i$  so that the resulting sets still form a convex realization of  $\mathcal{C}$ ?*

### Acknowledgments

The authors thank the Dean's Office and the Department of Mathematics at Harvey Mudd College for their summer research support, and thank Carina Curto, Chad Giusti, Elizabeth Gross, Vladimir Itskov, Bill Kronholm and Anne Shiu for fruitful conversations.

### References

- [Amenta et al. 2017] N. Amenta, J. A. De Loera, and P. Soberón, "Helly's theorem: new variations and applications", pp. 55–95 in *Algebraic and geometric methods in discrete mathematics* (San Antonio, 2015), edited by H. A. Harrington et al., Contemp. Math. **685**, Amer. Math. Soc., Providence, RI, 2017. MR Zbl
- [Amzi Jeffs 2018] R. Amzi Jeffs, "Morphisms of neural codes", preprint, 2018. arXiv

- [Amzi Jeffs and Novik 2018] R. Amzi Jeffs and I. Novik, “Convex union representability and convex codes”, preprint, 2018. [arXiv](#)
- [Chen et al. 2019] A. Chen, F. Frick, and A. Shiu, “Neural codes, decidability, and a new local obstruction to convexity”, *SIAM J. Appl. Algebra Geom.* **3**:1 (2019), 44–66. [MR](#)
- [Cruz et al. 2019] J. Cruz, C. Giusti, V. Itskov, and B. Kronholm, “On open and closed convex codes”, *Discrete Comput. Geom.* **61**:2 (2019), 247–270. [MR](#) [Zbl](#)
- [Curto et al. 2013a] C. Curto, V. Itskov, K. Morrison, Z. Roth, and J. L. Walker, “Combinatorial neural codes from a mathematical coding theory perspective”, *Neural Comput.* **25**:7 (2013), 1891–1925. [MR](#)
- [Curto et al. 2013b] C. Curto, V. Itskov, A. Veliz-Cuba, and N. Youngs, “The neural ring: an algebraic tool for analyzing the intrinsic structure of neural codes”, *Bull. Math. Biol.* **75**:9 (2013), 1571–1611. [MR](#) [Zbl](#)
- [Curto et al. 2017] C. Curto, E. Gross, J. Jeffries, K. Morrison, M. Omar, Z. Rosen, A. Shiu, and N. Youngs, “What makes a neural code convex?”, *SIAM J. Appl. Algebra Geom.* **1**:1 (2017), 222–238. [MR](#) [Zbl](#)
- [Danzer et al. 1963] L. Danzer, B. Grünbaum, and V. Klee, “Helly’s theorem and its relatives”, pp. 101–180 in *Convexity* (Seattle, 1961), edited by V. Klee, Proc. Symp. Pure Math. **7**, Amer. Math. Soc., Providence, RI, 1963. [MR](#) [Zbl](#)
- [Eckhoff 1993] J. Eckhoff, “Helly, Radon, and Carathéodory type theorems”, pp. 389–448 in *Handbook of convex geometry, A*, edited by P. M. Gruber and J. M. Wills, North-Holland, Amsterdam, 1993. [MR](#) [Zbl](#)
- [Hubel and Wiesel 1959] D. H. Hubel and T. N. Wiesel, “Receptive fields of single neurons in the cat’s striate cortex”, *J. Physiology* **148**:3 (1959), 574–591.
- [Lienkaemper et al. 2017] C. Lienkaemper, A. Shiu, and Z. Woodstock, “Obstructions to convexity in neural codes”, *Adv. Appl. Math.* **85** (2017), 31–59. [MR](#) [Zbl](#)
- [Lin et al. 2014] A. C. Lin, A. M. Bygrave, A. de Calignon, T. Lee, and G. Miesenböck, “Sparse, decorrelated odor coding in the mushroom body enhances learned odor discrimination”, *Nature Neurosci.* **17**:4 (2014), 559–568.
- [Matoušek 2002] J. Matoušek, *Lectures on discrete geometry*, Graduate Texts in Math. **212**, Springer, 2002. [MR](#) [Zbl](#)
- [O’Keefe and Dostrovsky 1971] J. O’Keefe and J. Dostrovsky, “The hippocampus as a spatial map: preliminary evidence from unit activity in the freely-moving rat”, *Brain Res.* **34**:1 (1971), 171–175.
- [Rosen and Zhang 2017] Z. Rosen and Y. X. Zhang, “Convex neural codes in dimension 1”, preprint, 2017. [arXiv](#)
- [Tancer 2013] M. Tancer, “Intersection patterns of convex sets via simplicial complexes: a survey”, pp. 521–540 in *Thirty essays on geometric graph theory*, edited by J. Pach, Springer, 2013. [MR](#) [Zbl](#)

Received: 2017-10-26

Revised: 2018-10-24

Accepted: 2018-12-05

rajeffs@uw.edu

Department of Mathematics, University of Washington,  
Seattle, WA, United States

omar@g.hmc.edu

Department of Mathematics, Harvey Mudd College,  
Claremont, CA, United States

nsuaysom@g.hmc.edu

Harvey Mudd College, Claremont, CA, United States

awachtel@g.hmc.edu

Harvey Mudd College, Claremont, CA, United States

neyoungs@colby.edu

Department of Mathematics and Statistics, Colby College,  
Waterville, ME, United States

# The number of rational points of hyperelliptic curves over subsets of finite fields

Kristina Nelson, József Solymosi, Foster Tom and Ching Wong

(Communicated by Kenneth S. Berenhaut)

We prove two related concentration inequalities concerning the number of rational points of hyperelliptic curves over subsets of a finite field. In particular, we investigate the probability of a large discrepancy between the numbers of quadratic residues and nonresidues in the image of such subsets over uniformly random hyperelliptic curves of given degrees. We find a constant probability of such a high difference and show the existence of sets with an exceptionally large discrepancy.

## 1. Introduction

Let  $q$  be a prime power and let  $\mathbb{F}_q$  be the finite field with  $q$  elements. A curve  $E: y^2 = f(x)$  (together with a point at infinity  $\mathcal{O}$ ) is called an *elliptic curve* over  $\mathbb{F}_q$  if  $f(x) \in \mathbb{F}_q[x]$  is a cubic polynomial having distinct roots in the algebraic closure  $\bar{\mathbb{F}}_q$  of  $\mathbb{F}_q$ . The set of *rational points* of  $E$  in  $\mathbb{F}_q$  is

$$E(\mathbb{F}_q) = \{(x, y) \in \mathbb{F}_q \times \mathbb{F}_q : y^2 = f(x)\} \cup \{\mathcal{O}\}.$$

Suppose that  $q$  is odd. Using the fact that there are  $(q-1)/2$  invertible quadratic residues and  $(q-1)/2$  nonresidues in  $\mathbb{F}_q$ , one can approximate the size of  $E(\mathbb{F}_q)$  as follows. For each  $x \in \mathbb{F}_q$ , the probability of  $f(x)$  being a nonzero square in  $\mathbb{F}_q$ , and hence contributing two points to  $E(\mathbb{F}_q)$ , is about  $\frac{1}{2}$ . With probability about  $\frac{1}{2}$  there is no point in  $E(\mathbb{F}_q)$  having the first coordinate  $x \in \mathbb{F}_q$ . Therefore,  $\#E(\mathbb{F}_q)$  is expected to be close to  $q+1$ . Indeed, Hasse [1936] proved that the error in this estimate is at most  $2\sqrt{q}$ :

$$|\#E(\mathbb{F}_q) - (q+1)| \leq 2\sqrt{q}.$$

Knowledge of  $\#E(\mathbb{F}_q)$  is crucial in elliptic curve cryptography (ECC), which is considered to be more efficient than the classical cryptosystems, like RSA [Rivest

*MSC2010:* 68Q87, 68R05.

*Keywords:* hyperelliptic curves, finite fields.

Solymosi was supported by NSERC and the Hungarian National Research Development and Innovation Fund K 119528.

et al. 1978]. The security of ECC depends on the difficulty of solving the elliptic curve discrete logarithm problem (ECDLP). The best known algorithm to solve ECDLP in finite fields is Pollard's rho algorithm [1975], which requires  $O(\sqrt{p})$  time complexity, where  $p$  is the prime factor of  $q$ . However, some well-studied classes of elliptic curves are not good candidates for ECC. For instance, if the number of rational points of an elliptic curve  $E$  in  $\mathbb{F}_p$  is exactly  $p$ , where  $p$  is a prime, then the running time of solving the ECDLP is  $O(\log p)$ ; see [Semaev 1998]. Using verifiably random elliptic curves in ECC improves security since randomly generated curves are unlikely to be part of a weak class. Hyperelliptic curves can also be used in cryptography; see [Cohen et al. 2006] for more details. However, the verifiability of random hyperelliptic curves is much harder; see [Hess et al. 2001; Satoh 2009].

In this paper, we investigate the behaviour of random hyperelliptic curves over subsets  $S$  of  $\mathbb{F}_q$ . We are interested in the hyperelliptic curves  $E : y^2 = f(x)$  where  $f(x)$  is a polynomial in  $\mathbb{F}_q[x]$  of degree  $4k - 1$  ( $k \geq 1$ ) having distinct roots in  $\overline{\mathbb{F}}_q$ . Denote by  $E(\mathbb{F}_q, S)$  the rational points of  $E$  in  $\mathbb{F}_q$  where the  $x$ -coordinate is in  $S$ ; i.e.,

$$E(\mathbb{F}_q, S) = \{(x, y) \in S \times \mathbb{F}_q : y^2 = f(x)\}.$$

We remark that the point at infinity  $\mathcal{O}$  is not included in  $E(\mathbb{F}_q, S)$ . The approximation we have described for  $\#E(\mathbb{F}_q)$  suggests that the expected value of  $\#E(\mathbb{F}_q, S)$  is about  $\#S$ . For random hyperelliptic curves  $E$  over  $\mathbb{F}_q$ , the probability that the error  $|\#E(\mathbb{F}_q, S) - \#S|$  is small has been extensively studied; see [Pelekis and Ramon 2017; Schmidt et al. 1995] for example.

On the other hand, it is easy to see that there exist many hyperelliptic curves of any (positive) even degree so that the error  $|\#E(\mathbb{F}_p, S) - \#S|$  is very large. Indeed, the error is about  $\#S$  when  $f(x)$  is the square of any nonconstant polynomial in  $\mathbb{F}_q[x]$  for any  $S \subset \mathbb{F}_p$ .

However, an error bound is not obvious in the case of hyperelliptic curves of odd degree, which we study in the probabilistic setting. Equivalently, we examine the difference between the numbers of quadratic residues and nonresidues in the image multiset  $f(S)$ . Using  $4k$ -wise independence, we show that all subsets  $S$  of  $\mathbb{F}_q$  behave similarly, in the sense that the interested discrepancy is proportional to  $\sqrt{\#S}$  and has a positive probability which depends only on the degree of the curve.

**Theorem 1.** *Given a positive integer  $k$  and  $\varepsilon > 0$ , there exist  $\delta > 0$  and a threshold  $N$  such that the following holds: for every odd prime power  $q > N$ , if a curve  $E : y^2 = f(x)$  is chosen uniformly at random among all hyperelliptic curves of degree  $4k - 1$  over  $\mathbb{F}_q$ , then with a probability at least  $(4\pi^{3/2}/e^3)2^{-2k} - \varepsilon$ , we have*

$$|\#E(\mathbb{F}_q, S) - \#S| > \delta\sqrt{\#S}$$

for any set  $S \subset \mathbb{F}_q$  with  $\#S \geq N$ .

**Theorem 2.** *Given a positive integer  $k$ , there exist a threshold  $N$  and  $\varepsilon > 0$  such that the following holds: for every odd prime power  $q > N$ , if a curve  $E : y^2 = f(x)$  is chosen uniformly at random among all hyperelliptic curves of degree  $4k - 1$  over  $\mathbb{F}_q$ , then with a probability at least  $\varepsilon$ , we have*

$$|\#E(\mathbb{F}_q, S) - \#S| > 0.8577\sqrt{k}\sqrt{\#S}$$

for any set  $S \subset \mathbb{F}_q$  with  $\#S \geq N$ .

These two theorems imply that one can expect large deviation of magnitude  $\sqrt{\#S}$ . In the last section, we show that for small sets  $S$  of prime fields  $\mathbb{F}_p$ , the error is often much larger.

## 2. Preliminaries

Throughout this section, let  $q$  be an odd prime power and let  $n, k$  be positive integers such that  $4k < n \leq q$ . Suppose  $S = \{s_1, \dots, s_n\} \subset \mathbb{F}_q$ , and

$$f(x) = \sum_{j=0}^{4k-1} a_j x^j \in \mathbb{F}_q[x]$$

is chosen uniformly at random.

We denote by  $\#QR$ ,  $\#NR$  and  $\#R$  the numbers of  $s_i \in S$  such that  $f(s_i)$  is an invertible quadratic residue, a quadratic nonresidue and zero in  $\mathbb{F}_q$ , respectively. Then,  $n = \#QR + \#NR + \#R$ . It follows that, provided the curve  $E : y^2 = f(x)$  forms a hyperelliptic curve of degree  $4k - 1$  over  $\mathbb{F}_q$ , the discrepancy we are interested in is

$$|\#E(\mathbb{F}_q, S) - n| = |2\#QR + \#R - n| = |\#QR - \#NR|. \quad (1)$$

This suggests we look at the random variables

$$X_i = \left( \frac{f(s_i)}{q} \right),$$

where  $\left( \frac{a}{q} \right)$  is the Legendre symbol defined as

$$\left( \frac{a}{q} \right) = \begin{cases} 0 & \text{if } a \text{ is the zero in } \mathbb{F}_q, \\ 1 & \text{if } a \text{ is a nonzero square in } \mathbb{F}_q, \\ -1 & \text{otherwise.} \end{cases}$$

We note that among all polynomials  $f(x) \in \mathbb{F}_q[x]$  of degree at most 3, only a small fraction fail to form elliptic curves. Indeed, the exceptions, where  $f(x)$  has degree strictly less than 3 or has multiple roots, contribute  $q^3 + q^2(q - 1)$  of all the  $q^4$  polynomials considered. When  $q$  is large, such exceptions are negligible. This situation generalizes to hyperelliptic curves.

**Lemma 3.** *Let  $q$  be a prime power,  $k$  be a positive integer and  $\mathbb{F}_q[x]_{4k-1}$  be the set of polynomials in  $\mathbb{F}_q[x]$  of degree at most  $4k - 1$ . Then at most a  $2/q$  fraction of the polynomials in  $\mathbb{F}_q[x]_{4k-1}$  fail to define a hyperelliptic curve of degree  $4k - 1$ .*

*Proof.* A polynomial in  $\mathbb{F}_q[x]$  defines a hyperelliptic curve precisely when it is separable, or equivalently when it is square-free because finite fields are perfect. As shown in [Carlitz 1932], the number of monic square-free polynomials in  $\mathbb{F}_q[x]$  of degree  $4k - 1$  is  $q^{4k-1} - q^{4k-2}$ . Thus, accounting for scaling, there are  $(q - 1)(q^{4k-1} - q^{4k-2})$  polynomials in  $\mathbb{F}_q[x]$  that define a hyperelliptic curve of degree  $4k - 1$ . Therefore, the fraction of those polynomials in  $\mathbb{F}_q[x]$  of degree at most  $4k - 1$  that do not is

$$\frac{q^{4k} - (q - 1)(q^{4k-1} - q^{4k-2})}{q^{4k}} = \frac{2q^{4k-1} - q^{4k-2}}{q^{4k}} < \frac{2}{q}. \quad \square$$

Hence, the probability that, among all hyperelliptic curves of degree  $4k - 1$  over  $\mathbb{F}_q$ , the discrepancy (1) is larger than some  $\delta\sqrt{n}$  is at least the probability that, among all polynomials of degree at most  $4k - 1$  over  $\mathbb{F}_q$ , the absolute value of the sum of the random variables  $X_i$  is larger than the same  $\delta\sqrt{n}$  minus  $2/q$ ; i.e.,

$$\mathbb{P}(|\#E(\mathbb{F}_q, S) - n| > \delta\sqrt{n}) \geq \mathbb{P}\left(\left|\sum_{i=1}^n X_i\right| > \delta\sqrt{n}\right) - \frac{2}{q}. \quad (2)$$

In the next two subsections, we will first estimate the higher moments

$$\mathbb{E}_j := \mathbb{E}\left(\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i\right)^j\right), \quad \text{where } 1 \leq j \leq 4k,$$

by finding their main order, and then give lower bounds on the interested probabilities involving the random variables  $X_i$ .

**2.1. Estimating  $\mathbb{E}_{2k}$  and  $\mathbb{E}_{4k}$ .** Since  $f(x) \in \mathbb{F}_q[x]$  is a random polynomial of degree at most  $4k - 1$ , the random variables  $X_i$  exhibit  $4k$ -wise independence. Indeed, by solving a system of linear equations, the number of polynomials  $f(x)$  in  $\mathbb{F}_q[x]$  of degree at most  $4k - 1$  satisfying

$$f(s_{i_1}) = r_1, \quad f(s_{i_2}) = r_2, \quad \dots, \quad f(s_{i_\ell}) = r_\ell$$

is exactly  $q^{4k-\ell}$ , given  $\ell \leq 4k$ ,  $r_1, \dots, r_\ell \in \mathbb{F}_q$  and distinct  $i_1, \dots, i_\ell \in \{1, \dots, n\}$ . Thus,

$$\begin{aligned} \mathbb{E}(X_{i_1}^{h_1} \cdots X_{i_\ell}^{h_\ell}) &= \sum_{r_1, \dots, r_\ell \in \mathbb{F}_q} \mathbb{P}(f(s_{i_1}) = r_1, \dots, f(s_{i_\ell}) = r_\ell) \left(\frac{r_1}{q}\right)^{h_1} \cdots \left(\frac{r_\ell}{q}\right)^{h_\ell} \\ &= \sum_{r_1, \dots, r_\ell \in \mathbb{F}_q} \frac{q^{4k-\ell}}{q^{4k}} \left(\frac{r_1}{q}\right)^{h_1} \cdots \left(\frac{r_\ell}{q}\right)^{h_\ell} \end{aligned}$$



$$\begin{aligned}
&= \left[ \sum_{r_1 \in \mathbb{F}_q} \frac{1}{q} \left( \frac{r_1}{q} \right)^{h_1} \right] \cdots \left[ \sum_{r_\ell \in \mathbb{F}_q} \frac{1}{q} \left( \frac{r_\ell}{q} \right)^{h_\ell} \right] \\
&= \left[ \sum_{r_1 \in \mathbb{F}_q} \mathbb{P}(f(s_{i_1})=r_1) \left( \frac{r_1}{q} \right)^{h_1} \right] \cdots \left[ \sum_{r_\ell \in \mathbb{F}_q} \mathbb{P}(f(s_{i_\ell})=r_\ell) \left( \frac{r_\ell}{q} \right)^{h_\ell} \right] \\
&= \mathbb{E}(X_{i_1}^{h_1}) \cdots \mathbb{E}(X_{i_\ell}^{h_\ell}).
\end{aligned}$$

We also note that the random variables  $X_i$  only take the values 0, 1,  $-1$ , and so  $X_i^{2h-1} = X_i$  and  $X_i^{2h} = X_i^2$  for all  $h \geq 1$ . Also, by convention,  $X_i^0 = 0$ . Therefore we have

$$\begin{aligned}
\mathbb{E}(X_i^{2h-1}) &= \mathbb{E}(X_i) = \sum_{r \in \mathbb{F}_q} \mathbb{P}(f(s_i)=r) \left( \frac{r}{q} \right) = \sum_{r \in \mathbb{F}_q} \frac{1}{q} \left( \frac{r}{q} \right) = 0, \\
\mathbb{E}(X_i^{2h}) &= \mathbb{E}(X_i^2) = \sum_{r \in \mathbb{F}_q} \mathbb{P}(f(s_i)=r) \left( \frac{r}{q} \right)^2 = \sum_{r \in \mathbb{F}_q} \frac{1}{q} \left( \frac{r}{q} \right)^2 = 1 - \frac{1}{q}.
\end{aligned}$$

To summarize the above two observations, we have the following lemma:

**Lemma 4.** *Let  $\ell \leq 4k$ , let  $h_1, \dots, h_\ell$  be positive integers, and let  $i_1, \dots, i_\ell$  be distinct numbers from  $\{1, \dots, n\}$ . Then,*

$$\mathbb{E}(X_{i_1}^{h_1} \cdots X_{i_\ell}^{h_\ell}) = \begin{cases} (1 - 1/q)^\ell & \text{if } h_1, \dots, h_\ell \text{ are all even numbers,} \\ 0 & \text{otherwise.} \end{cases}$$

Before we estimate the general  $\mathbb{E}_j$ , let us compute  $\mathbb{E}_6$  (when  $k \geq 2$ ) as a toy version:

$$\begin{aligned}
\mathbb{E}_6 &= \mathbb{E} \left( \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \right)^6 \\
&= \frac{1}{n^3} \left( \sum_{i=1}^n \mathbb{E}(X_i^6) + \frac{6!}{4!2!} \sum_{i \neq j} \mathbb{E}(X_i^4 X_j^2) + \frac{6!}{2!2!2!} \sum_{i < j < k} \mathbb{E}(X_i^2 X_j^2 X_k^2) \right) \\
&= \frac{1}{n^3} \left( n \left( 1 - \frac{1}{q} \right) + 15n(n-1) \left( 1 - \frac{1}{q} \right)^2 + 90 \binom{n}{3} \left( 1 - \frac{1}{q} \right)^3 \right) \\
&= 15 \left( 1 - \frac{1}{q} \right)^3 - \frac{15}{n} \left( 1 - \frac{1}{q} \right)^2 \left( 2 - \frac{3}{q} \right) + \frac{1}{n^2} \left( 1 - \frac{1}{q} \right) \left( 16 - \frac{45}{q} + \frac{30}{q^2} \right).
\end{aligned}$$

We derive in the lemma below how the number 15 in the leading term can be expressed in terms of  $j = 6$ .

**Lemma 5.** *For  $1 \leq j \leq 4k$ , we have*

$$\mathbb{E}_j = \begin{cases} \frac{j!}{2^{j/2}(j/2)!} + O_j \left( \frac{1}{n} \right) & \text{as } n \rightarrow \infty, \text{ if } j \text{ is an even number,} \\ 0 & \text{otherwise.} \end{cases}$$

*Proof.* If  $j$  is an odd number, then every term in the multinomial expansion has at least one odd index, and hence vanishes by Lemma 4.

Suppose now that  $j$  is an even integer. Using the multinomial theorem and Lemma 4, we have

$$\begin{aligned}\mathbb{E}_j &= \frac{1}{n^{j/2}} \mathbb{E} \left( \left( \sum_{i=1}^n X_i \right)^j \right) = \frac{1}{n^{j/2}} \mathbb{E} \left( \sum_{h_1+\dots+h_n=j} \frac{j!}{h_1! \dots h_n!} \prod_{t=1}^n X_t^{h_t} \right) \\ &= \frac{1}{n^{j/2}} \sum_{\substack{h_1+\dots+h_n=j \\ h_i \text{ even}}} \frac{j!}{h_1! \dots h_n!} \mathbb{E} \left( \prod_{t=1}^n X_t^{h_t} \right) \\ &= \frac{1}{n^{j/2}} \sum_{\substack{h_1+\dots+h_n=j \\ h_i \text{ even}}} \frac{j!}{h_1! \dots h_n!} \left( 1 - \frac{1}{q} \right)^{\#\{i: h_i > 0\}} = \frac{1}{n^{j/2}} \sum_{m=1}^{j/2} \left( 1 - \frac{1}{q} \right)^m H(j, m),\end{aligned}$$

where

$$H(j, m) = \sum_{\substack{h_1+\dots+h_n=j \\ h_i \text{ even} \\ \#\{i: h_i > 0\}=m}} \frac{j!}{h_1! \dots h_n!} = \binom{n}{m} \sum_{\substack{h'_1+\dots+h'_m=j \\ h'_i > 0 \text{ even}}} \frac{j!}{h'_1! \dots h'_m!}$$

is a polynomial (with integer coefficients) in  $n$  of degree  $m$ . Therefore, the leading term of  $\mathbb{E}_j$  comes from the summand where  $m = j/2$ . In this case,  $h'_i = 2$  for every  $1 \leq i \leq j/2$  and so

$$H(j, j/2) = \binom{n}{j/2} \frac{j!}{2^{j/2}}$$

has leading term

$$\frac{j!}{(j/2)! 2^{j/2}} n^{j/2}.$$

It follows that

$$\begin{aligned}\mathbb{E}_j &= \frac{1}{n^{j/2}} \left( \left( 1 - \frac{1}{q} \right)^{j/2} \frac{j!}{(j/2)! 2^{j/2}} n^{j/2} + \dots \right) \\ &= \left( 1 - \frac{1}{q} \right)^{j/2} \frac{j!}{(j/2)! 2^{j/2}} + O_j \left( \frac{1}{n} \right) = \frac{j!}{(j/2)! 2^{j/2}} + O_j \left( \frac{1}{n} \right)\end{aligned}$$

as  $n \rightarrow \infty$ . □

In particular, for each fixed  $k$ ,

$$\mathbb{E}_{2k} = \frac{(2k)!}{2^k k!} + O_k \left( \frac{1}{n} \right)$$

is bounded uniformly in  $n \geq 1$ . As a consequence, one can have the following estimates, which will be used later in our proof, using Stirling's approximation. For

all fixed  $k \geq 1$ , we have

$$\sqrt[2k]{\mathbb{E}_{2k}} \geq \sqrt{\frac{2k}{e}} + O_k\left(\frac{1}{n}\right) \quad (3)$$

and

$$\frac{\mathbb{E}_{2k}^2}{\mathbb{E}_{4k}} \geq \left(\frac{\sqrt{2\pi}}{e}\right)^3 2^{1/2-2k} + O_k\left(\frac{1}{n}\right) \quad (4)$$

as  $n \rightarrow \infty$ .

## 2.2. Lower bounds for the probabilities.

**Proposition 6.** *Under the setting stated in the beginning of this section, we have*

$$\mathbb{P}\left(\left|\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i\right| > \delta\right) \geq \frac{(\mathbb{E}_{2k} - \delta^{2k})^2}{\mathbb{E}_{4k} - 2\delta^{2k}\mathbb{E}_{2k} + \delta^{4k}} \quad (5)$$

for any  $0 < \delta < \frac{1}{2}$ , and

$$\mathbb{P}\left(\left|\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i\right| \geq \sqrt[2k]{\mathbb{E}_{2k}} - \varepsilon^{1/2-o(1)}\right) \geq \varepsilon > 0 \quad (6)$$

as  $\varepsilon \rightarrow 0$ .

*Proof.* Let  $c \geq 1$  be a parameter to be determined. Using the second-moment Markov inequality, one can show that for  $0 < \lambda < c^{2k}$ ,

$$\begin{aligned} \mathbb{P}\left(\left|\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i\right| > \sqrt[2k]{c^k - \sqrt{\lambda}}\right) &= \mathbb{P}\left(\left(\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i\right)^{2k} - c^k > -\sqrt{\lambda}\right)\right) \\ &\geq \mathbb{P}\left(\left|\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i\right)^{2k} - c^k\right| < \sqrt{\lambda}\right) \\ &\geq 1 - \frac{1}{\lambda} \mathbb{E}\left(\left(\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i\right)^{2k} - c^k\right)^2\right) \\ &= 1 - \frac{c^{2k} - 2c^k\mathbb{E}_{2k} + \mathbb{E}_{4k}}{\lambda}. \end{aligned} \quad (7)$$

To prove (5), we take  $\lambda = (c^k - \delta^{2k})^2$ , where  $\delta > 0$  is small. Maximizing the right-hand side of (7) over  $c$ , we see that the maximum is

$$1 - \frac{c^{2k} - 2c^k\mathbb{E}_{2k} + \mathbb{E}_{4k}}{(c^k - \delta^{2k})^2} = \frac{(\mathbb{E}_{2k} - \delta^{2k})^2}{\mathbb{E}_{4k} - 2\delta^{2k}\mathbb{E}_{2k} + \delta^{4k}},$$

when

$$c^k = \frac{\mathbb{E}_{4k} - \delta^{2k}\mathbb{E}_{2k}}{\mathbb{E}_{2k} - \delta^{2k}}.$$

Now we prove (6). To make

$$\mathbb{P}\left(\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i\right| > \sqrt[2k]{c^k - \sqrt{\lambda}}\right) \geq \varepsilon,$$

we take

$$\lambda = \frac{c^{2k} - 2c^k \mathbb{E}_{2k} + \mathbb{E}_{4k}}{1 - \varepsilon}.$$

Since we require  $c^{2k} > \lambda$ , it follows that

$$c^{2k} - 2c^k \mathbb{E}_{2k} + \mathbb{E}_{4k} < c^{2k} - c^{2k} \varepsilon,$$

and therefore

$$\eta := \varepsilon c^k < 2\mathbb{E}_{2k} - \frac{\mathbb{E}_{4k}}{c^k} < 2\mathbb{E}_{2k}.$$

To compute the leading terms of  $\sqrt[2k]{c^k - \sqrt{\lambda}}$  as  $\varepsilon \rightarrow 0$ , we first use the binomial series to expand the numerator of  $\sqrt{\lambda}$  as

$$c^k \sqrt{1 - \left(\frac{2\mathbb{E}_{2k}}{c^k} - \frac{\mathbb{E}_{4k}}{c^{2k}}\right)} = c^k \left(1 - \mathbb{E}_{2k} \frac{1}{c^k} + \frac{\mathbb{E}_{4k} - \mathbb{E}_{2k}^2}{2} \frac{1}{c^{2k}} + O\left(\frac{1}{c^{3k}}\right)\right) \quad (8)$$

as  $c \rightarrow \infty$ . Indeed, the bracket inside the square root in (8) is small in view of Lemma 5. To get  $\sqrt{\lambda}$ , we multiply (8) by

$$\frac{1}{\sqrt{1 - \varepsilon}} = 1 + \frac{1}{2}\varepsilon + \frac{3}{8}\varepsilon^2 + O(\varepsilon^3).$$

Substituting  $c^k = \eta/\varepsilon$ , we have

$$\begin{aligned} c^k - \sqrt{\lambda} &= \frac{\eta}{\varepsilon} \left[ 1 - \left(1 + \frac{1}{2}\varepsilon + \frac{3}{8}\varepsilon^2 + O(\varepsilon^3)\right) \left(1 - \frac{\mathbb{E}_{2k}}{\eta} \varepsilon + \frac{\mathbb{E}_{4k} - \mathbb{E}_{2k}^2}{2\eta^2} \varepsilon^2 + O\left(\frac{\varepsilon^3}{\eta^3}\right)\right) \right] \\ &= \mathbb{E}_{2k} - \frac{1}{2}\eta + \left(\frac{\mathbb{E}_{2k}^2 - \mathbb{E}_{4k}}{2} + \frac{\mathbb{E}_{2k}}{2}\eta - \frac{3}{8}\eta^2\right) \frac{\varepsilon}{\eta} + O\left(\frac{\varepsilon^2}{\eta^2}\right). \end{aligned}$$

We may now take  $\eta$  satisfying  $\sqrt{\varepsilon} \ll \eta \ll 1$  so that the terms in the last line are indeed arranged in decreasing order of magnitude. Therefore,

$$\sqrt[2k]{c^k} - \sqrt[2k]{\lambda} = \sqrt[2k]{\mathbb{E}_{2k} - \varepsilon^{1/2-o(1)}} = \sqrt[2k]{\mathbb{E}_{2k}} - \varepsilon^{1/2-o(1)}$$

as  $\varepsilon \rightarrow 0$ , establishing (6). □

### 3. Proofs of the theorems

*Proof of Theorem 1.* Write  $n = \#S$ , as in Section 2. Given  $\varepsilon > 0$ , we choose  $N$  large enough so that  $2/N < \varepsilon/3$ , and the error appearing in (4) has an absolute value less than  $\varepsilon/3$ .

Since  $\mathbb{E}_{4k} > \mathbb{E}_{2k} \geq \frac{1}{2}$ , there exists a small  $\delta > 0$  such that

$$\left| \frac{(1 - \delta^{2k}/\mathbb{E}_{2k})^2}{1 - 2\delta^{2k}(\mathbb{E}_{2k}/\mathbb{E}_{4k}) + \delta^{4k}(1/\mathbb{E}_{4k})} - 1 \right| < \frac{\varepsilon}{3} \frac{\mathbb{E}_{4k}}{\mathbb{E}_{2k}^2}.$$

Together with (2), (5) and (4), we have

$$\begin{aligned} \mathbb{P}(|\#E(\mathbb{F}_q, S) - n| > \delta\sqrt{n}) &\geq \mathbb{P}\left(\left|\sum_{i=1}^n X_i\right| > \delta\sqrt{n}\right) - \frac{2}{q} \\ &\geq \frac{\mathbb{E}_{2k}^2}{\mathbb{E}_{4k}} \frac{(1 - \delta^{2k}/\mathbb{E}_{2k})^2}{1 - 2\delta^{2k}(\mathbb{E}_{2k}/\mathbb{E}_{4k}) + \delta^{4k}(1/\mathbb{E}_{4k})} - \frac{\varepsilon}{3} \\ &\geq \frac{\mathbb{E}_{2k}^2}{\mathbb{E}_{4k}} - \frac{\varepsilon}{3} - \frac{\varepsilon}{3} \geq \left(\frac{\sqrt{2\pi}}{e}\right)^3 2^{1/2-2k} - \varepsilon, \end{aligned}$$

as desired.  $\square$

*Proof of Theorem 2.* Similarly we write  $n = \#S$ . Using the estimate (3), we choose  $N$  so large and  $\varepsilon$  so small that the following lower bound implied by (6) is large:

$$\sqrt[2k]{\mathbb{E}_{2k}} - \varepsilon^{1/2-o(1)} > 0.8577\sqrt{k}.$$

Here 0.8577 is a number strictly smaller than  $\sqrt{2/e}$ . Now, increasing  $N$  if necessary, we also have  $2/N < \varepsilon/2$ . Then, by (2) and (6), we have

$$\begin{aligned} \mathbb{P}(|\#E(\mathbb{F}_q, S) - n| > 0.8577\sqrt{k}\sqrt{n}) &\geq \mathbb{P}\left(\left|\sum_{i=1}^n X_i\right| > 0.8577\sqrt{k}\sqrt{n}\right) - \frac{2}{q} \\ &\geq \mathbb{P}\left(\left|\sum_{i=1}^n X_i\right| > (\sqrt[2k]{\mathbb{E}_{2k}} - \varepsilon^{1/2-o(1)})\sqrt{n}\right) - \frac{\varepsilon}{2} \\ &\geq \frac{\varepsilon}{2}. \end{aligned} \quad \square$$

#### 4. Sets with exceptionally large discrepancy

So far we have considered sets of arbitrarily large size. We will show, as one may expect, that if  $n$  is a constant, then for each prime  $p$  large enough, there is a probability  $\alpha > 0$  that the error is much larger than  $\sqrt{n}$  for  $\beta_n^{(p)}$  of the subsets  $S \subset \mathbb{F}_p$  of size  $n$ . In particular, for each  $n$ , there is a probability  $2^{-n-1}$  that a randomly chosen subset  $S \subset \mathbb{F}_p$  of size  $n$  has the following property — a randomly chosen monic separable cubic  $f$  over  $\mathbb{F}_p$  has a probability  $2^{-n-1}$  so that  $f(S)$  consists only of nonzero quadratic residues or quadratic nonresidues.

Let  $\mathcal{F}$  be the set of monic, separable cubics over  $\mathbb{F}_p$ . Note that  $\#\mathcal{F} = p^3 - p^2$ . Let  $m, n$  be constants independent of  $p$  such that  $n - 2m > \sqrt{n}$ . We construct a bipartite graph  $G$  with  $\binom{p}{n}$  “ $S$ -vertices” in one partition, each associated with a

set  $S \subset \mathbb{F}_p$  of size  $n$ , and  $p^3 - p^2$  “ $f$ -vertices” in the other, each associated with an  $f \in \mathcal{F}$ . We draw an edge between the vertex corresponding to  $f$  and the vertex corresponding to  $S$  when

$$\left| \sum_{s_i \in S} \left( \frac{f(s_i)}{p} \right) \right| \geq n - 2m.$$

Fix  $f \in \mathcal{F}$ , and let  $\mathcal{Q} \subset \mathbb{F}_p$  be the set of points mapped by  $f$  to a nonzero quadratic residue, and  $\mathcal{N} \subset \mathbb{F}_p$  be those points mapped to a nonresidue. Let  $p/2 + A_f$  be the size of the larger of these two sets. Then the degree of the vertex associated to  $f$  in  $G$  is at least

$$\binom{p/2 - A_f}{m} \binom{p/2 + A_f}{n - m}. \quad (9)$$

By Hasse’s theorem we have  $A_f \leq \sqrt{p}$ , and so (9) is bounded below by

$$\binom{p/2 - \sqrt{p}}{m} \binom{p/2 - \sqrt{p}}{n - m} = \binom{p}{n} \left[ \binom{n}{m} 2^{-n} + o(1) \right]$$

as  $p \rightarrow \infty$ . Thus the number of edges in our graph,  $E$ , is at least

$$\binom{p}{n} \left[ \binom{n}{m} 2^{-n} + o(1) \right] (p^3 - p^2).$$

Now if only  $\beta \binom{p}{n}$  of the  $S$ -vertices achieve degree at least  $\alpha(p^3 - p^2)$ , then we have

$$E \leq \beta \binom{p}{n} (p^3 - p^2) + \left( \binom{p}{n} - \beta \binom{p}{n} \right) \alpha (p^3 - p^2),$$

and so

$$\beta \geq \frac{1}{1 - \alpha} \left[ \binom{n}{m} 2^{-n} - \alpha + o(1) \right] > 0$$

as  $p \rightarrow \infty$ , provided that  $\alpha > 0$  is small enough.

## References

- [Carlitz 1932] L. Carlitz, “The arithmetic of polynomials in a Galois field”, *Amer. J. Math.* **54**:1 (1932), 39–50. MR Zbl
- [Cohen et al. 2006] H. Cohen, G. Frey, R. Avanzi, C. Doche, T. Lange, K. Nguyen, and F. Vercauteren (editors), *Handbook of elliptic and hyperelliptic curve cryptography*, Chapman & Hall, Boca Raton, FL, 2006. MR Zbl
- [Hasse 1936] H. Hasse, “Zur Theorie der abstrakten elliptischen Funktionenkörper, III: Die Struktur des Meromorphismenrings, die Riemannsche Vermutung”, *J. Reine Angew. Math.* **175** (1936), 193–208. MR Zbl
- [Hess et al. 2001] F. Hess, G. Seroussi, and N. P. Smart, “Two topics in hyperelliptic cryptography”, pp. 181–189 in *Selected areas in cryptography* (Toronto, 2001), edited by S. Vaudenay and A. M. Youssef, Lecture Notes in Comput. Sci. **2259**, Springer, 2001. MR Zbl
- [Pelekis and Ramon 2017] C. Pelekis and J. Ramon, “Hoeffding’s inequality for sums of dependent random variables”, *Mediterr. J. Math.* **14**:6 (2017), art. id. 243. MR Zbl

- [Pollard 1975] J. M. Pollard, “A Monte Carlo method for factorization”, *Nordisk Tidskr. Informationsbehandling* **15**:3 (1975), 331–334. MR Zbl
- [Rivest et al. 1978] R. L. Rivest, A. Shamir, and L. Adleman, “A method for obtaining digital signatures and public-key cryptosystems”, *Comm. ACM* **21**:2 (1978), 120–126. MR Zbl
- [Satoh 2009] T. Satoh, “Generating genus two hyperelliptic curves over large characteristic finite fields”, pp. 536–553 in *Advances in cryptology: EUROCRYPT 2009* (Cologne, 2009), edited by A. Joux, Lecture Notes in Comput. Sci. **5479**, Springer, 2009. MR Zbl
- [Schmidt et al. 1995] J. P. Schmidt, A. Siegel, and A. Srinivasan, “Chernoff–Hoeffding bounds for applications with limited independence”, *SIAM J. Discrete Math.* **8**:2 (1995), 223–250. MR Zbl
- [Semaev 1998] I. A. Semaev, “Evaluation of discrete logarithms in a group of  $p$ -torsion points of an elliptic curve in characteristic  $p$ ”, *Math. Comp.* **67**:221 (1998), 353–356. MR Zbl

Received: 2018-01-19

Revised: 2018-06-21

Accepted: 2018-07-28

krisn@math.berkeley.edu

*Department of Mathematics, University of California,  
Berekeley, CA, United States*

solymosi@math.ubc.ca

*Department of Mathematics, University of British Columbia,  
Vancouver, BC, Canada*

foster@math.ubc.ca

*Department of Mathematics, University of British Columbia,  
Vancouver, BC, Canada*

ching@math.ubc.ca

*Department of Mathematics, University of British Columbia,  
Vancouver, BC, Canada*





# Space-efficient knot mosaics for prime knots with mosaic number 6

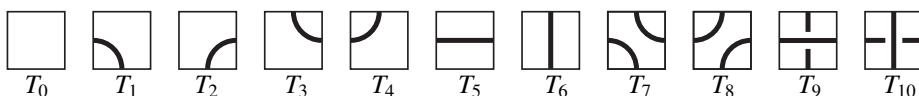
Aaron Heap and Douglas Knowles

(Communicated by Kenneth S. Berenhaut)

In 2008, Kauffman and Lomonaco introduced the concepts of a knot mosaic and the mosaic number of a knot or link  $K$ , the smallest integer  $n$  such that  $K$  can be represented on an  $n$ -mosaic. In 2018, the authors of this paper introduced and explored space-efficient knot mosaics and the tile number of  $K$ , the smallest number of nonblank tiles necessary to depict  $K$  on a knot mosaic. They determine bounds for the tile number in terms of the mosaic number. In this paper, we focus specifically on prime knots with mosaic number 6. We determine a complete list of these knots, provide a minimal, space-efficient knot mosaic for each of them, and determine the tile number (or minimal mosaic tile number) of each of them.

## 1. Introduction

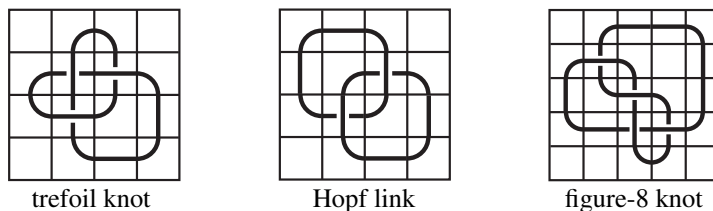
Mosaic knot theory was first introduced in [Lomonaco and Kauffman 2008] and was later proven to be equivalent to tame knot theory in [Kuriya and Shehab 2014]. The idea of mosaic knot theory is to create a knot or link diagram on an  $n \times n$  grid using *mosaic tiles* selected from the collection of 11 tiles shown below. The knot or link projection is represented by arcs, line segments, or crossings drawn on each tile. These tiles are identified, respectively, as  $T_0, T_1, T_2, \dots, T_{10}$ . Tile  $T_0$  is a blank tile, and we refer to the rest collectively as nonblank tiles.



A *connection point* of a tile is a midpoint of a tile edge that is also the endpoint of a curve drawn on the tile. A tile is *suitably connected* if each of its connection points touches a connection point of an adjacent tile. An  $n \times n$  *knot mosaic*, or *n-mosaic*, is an  $n \times n$  matrix whose entries are suitably connected mosaic tiles. As is customary in the literature of knot mosaic theory, the term “knot mosaic” is used

*MSC2010:* primary 57M25; secondary 57M27.

*Keywords:* knots, knot mosaic, mosaic number, tile number, crossing number, space-efficient.



**Figure 1.** Examples of knot mosaics.

for the mosaic, even when the resulting diagram on the mosaic depicts a link. See Figure 1 for some examples.

When listing prime knots with crossing number 10 or less, we will use the Alexander–Briggs notation, matching the table of knots in [Rolfsen 1976]. This notation names a knot according to its crossing number with a subscript to denote its order amongst all knots with that crossing number. For example, the  $7_4$  knot is the fourth knot with crossing number 7 in Rolfsen’s table of knots. For knots with crossing number 11 or higher, we use the Dowker–Thistlethwaite name of the knot. This also names a knot according to its crossing number, with an “a” or “n” to distinguish the alternating and nonalternating knots and a subscript that denotes the lexicographical ordering of the minimal Dowker–Thistlethwaite notation for the knot. For example  $11a_7$  is the seventh alternating knot with crossing number 11, and  $11n_3$  is the third nonalternating knot with crossing number 11. For more details on these and other relevant information on traditional knot theory, we refer the reader to [Adams 1994].

The *mosaic number* of a knot or link  $K$  is the smallest integer  $n$  for which  $K$  can be represented as an  $n$ -mosaic. The mosaic number has previously been determined for every prime knot with crossing number 8 or less. For details, see [Lee, Ludwig, Paat, and Peiffer 2018]. In particular, it is known that the unknot has mosaic number 2, the trefoil knot has mosaic number 4, the  $4_1$ ,  $5_1$ ,  $5_2$ ,  $6_1$ ,  $6_2$ , and  $7_4$  knots have mosaic number 5, and all other prime knots with crossing number 8 or less have mosaic number 6. In this paper, we determine the rest of the prime knots that have mosaic number 6, which includes prime knots with crossing numbers from 9 up to 13. This confirms, in the case where the mosaic number is  $m = 6$ , a result of [Howards and Kobin 2018], where they find that the crossing number is bounded above by  $(m - 2)^2 - 2$  if  $m$  is odd, and by  $(m - 2)^2 - (m - 3)$  if  $m$  is even. We also determine that not all knots with crossing number 9 (or higher) have mosaic number 6.

Another number associated to a knot mosaic is the *tile number of a mosaic*, which is the number of nonblank tiles used to create the mosaic. From this we get an invariant called the *tile number  $t(K)$  of a knot or link  $K$* , which is the least number of nonblank tiles needed to construct  $K$  on a mosaic of any size. In [Heap

and Knowles 2018], the authors explored the tile number of a knot or link and determined strict bounds for the tile number of a prime knot  $K$  in terms of the mosaic number  $m \geq 4$ . Specifically, if  $m$  is even, then  $5m - 8 \leq t(K) \leq m^2 - 4$ . If  $m$  is odd, then  $5m - 8 \leq t(K) \leq m^2 - 8$ . It follows immediately that the tile number of the trefoil knot must be 12, and the tile number of the prime knots mentioned above with mosaic number 5 must be 17. The authors also listed several prime knots with mosaic number 6 that have the smallest possible tile number  $t(K) = 22$ , which we summarize in Theorem 1. In this paper, we confirm that this list is complete. Knot mosaics in which the tile number is realized for each of these mosaics are given in [Heap and Knowles 2018] and also in the table of mosaics in the online supplement of this paper.

**Theorem 1** [Heap and Knowles 2018]. *The following knots have the given tile numbers:*

- (a) *Tile number 4: unknot.*
- (b) *Tile number 12: trefoil knot.*
- (c) *Tile number 17:  $4_1, 5_1, 5_2, 6_1, 6_2, 7_4$ .*
- (d) *Tile number 22:  $6_3, 7_1, 7_2, 7_3, 7_5, 7_6, 7_7, 8_1, 8_2, 8_3, 8_4, 8_7, 8_8, 8_9, 8_{13}, 9_5, 9_{20}$ .*

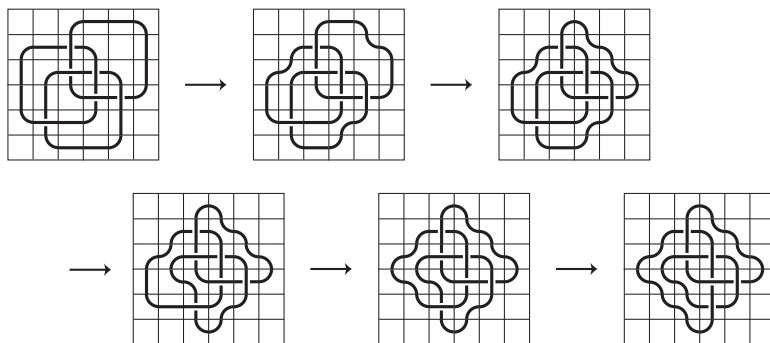
Finally, in [Heap and Knowles 2018], the authors determine all of the possible layouts for any prime knot on an  $n$ -mosaic for  $n \leq 6$ . In this paper, we complete that work by determining which prime knots can be created from those layouts.

We also point out that throughout this paper we make significant use of the software package Knotscape [Thistlethwaite and Hoste 1999] to verify that a given knot mosaic represents a specific knot. Without this program, we would not have been able to complete the work.

## 2. Space-efficient knot mosaics

Two knot mosaic diagrams are of the *same knot type* (or *equivalent*) if we can change one to the other via a sequence of *mosaic planar isotopy moves* that are analogous to the planar isotopy moves for standard knot diagrams. An example of this is shown in Figure 2. A complete list of all of these moves is given and discussed in [Lomonaco and Kauffman 2008; Kuriya and Shehab 2014]. We will make significant use of these moves throughout this paper, as we attempt to reduce the tile number of mosaics in order to construct knot mosaics that use the least number of nonblank tiles.

A knot mosaic is called *minimal* if it is a realization of the mosaic number of the knot depicted on it. That is, if a knot with mosaic number  $m$  is depicted on an  $m$ -mosaic, then it is a minimal knot mosaic. A knot mosaic is called *reduced*



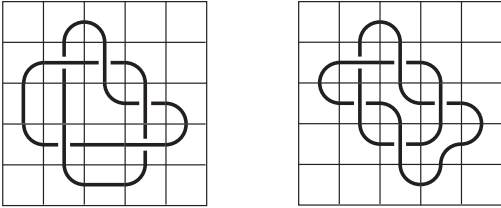
**Figure 2.** Example of mosaic planar isotopy moves.

if there are no unnecessary, reducible crossings in the knot mosaic diagram. See [Adams 1994] for more on reduced knot diagrams.

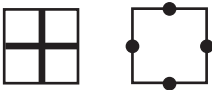
We have already defined the tile number of a mosaic and the tile number of a knot or link. A third type of tile number is the *minimal mosaic tile number*  $t_M(K)$  of a knot or link  $K$ , which is the smallest number of nonblank tiles needed to construct  $K$  on a minimal mosaic. That is, it is the smallest possible tile number of all possible minimal mosaic diagrams for  $K$ . Much like the crossing number of a knot cannot always be realized on a minimal mosaic (such as the  $6_1$  knot), the tile number of a knot cannot always be realized on a minimal mosaic. Note that the tile number of a knot or link  $K$  is certainly less than or equal to the minimal mosaic tile number of  $K$ ; that is,  $t(K) \leq t_M(K)$ . The fact that the tile number of a knot is not necessarily equal to the minimal mosaic tile number of the knot is confirmed later in Theorem 8. However, for prime knots, it is shown in [Heap and Knowles 2018] that  $t_M(K) = t(K)$  when  $t_M(K) \leq 27$ .

A knot  $n$ -mosaic is *space-efficient* if it is reduced and the tile number is as small as possible on an  $n$ -mosaic without changing the knot type of the depicted knot, meaning that the tile number cannot be decreased through a sequence of mosaic planar isotopy moves. A knot mosaic is *minimally space-efficient* if it is minimal and space-efficient. The first four knot mosaics of the Borromean rings depicted in Figure 2 are not space-efficient because we can decrease the tile number through the depicted mosaic planar isotopy moves. In Figure 3, both mosaics are knot mosaic diagrams of the  $5_1$  knot. The first knot mosaic is not space-efficient, but the second knot mosaic is minimally space-efficient.

In addition to the original 11 tiles  $T_0$ – $T_{10}$ , we will also make use of *nondeterministic tiles*, such as those in Figure 4, when there are multiple options for the tiles that can be placed in specific tile locations of a mosaic. For example, if a tile location must contain a crossing tile  $T_9$  or  $T_{10}$  but we have not yet chosen which, we will use the nondeterministic crossing tile. Similarly, if we know that a tile



**Figure 3.** Space-inefficient and minimally space-efficient knot mosaics of  $5_1$ .

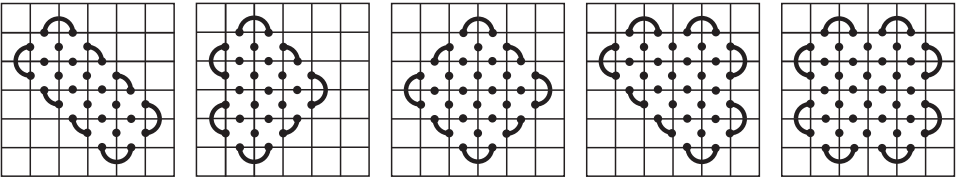


**Figure 4.** Nondeterministic crossing tile and a nondeterministic tile with four connection points.

location must have four connection points but we do not know if the tile is a double arc tile ( $T_7$  or  $T_8$ ) or a crossing tile ( $T_9$  or  $T_{10}$ ), we will indicate this with a tile that has four connection points.

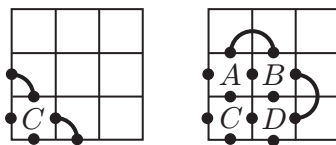
In [Heap and Knowles 2018], the authors provide the possible tile numbers (and the layouts that result in these tile numbers) for all prime knots on a space-efficient 6-mosaic.

**Theorem 2** [Heap and Knowles 2018]. *If we have a space-efficient 6-mosaic of a prime knot  $K$  for which either every column or every row is occupied, then the only possible values for the tile number of the mosaic are 22, 24, 27, and 32. Furthermore, any such mosaic of  $K$  is equivalent (up to symmetry) to one of the following mosaics:*



In order to determine all prime knots with mosaic number 6 and their minimal mosaic tile numbers, we need to determine which prime knots can be depicted on a knot mosaic with one of the layouts above. To help us with this, we make a few simple observations. All of these are easy to verify, and any rotation or reflection of these scenarios is also valid.

Consider the upper, right  $3 \times 3$  corner of any space-efficient mosaic of a prime knot with mosaic number 6 and tile number 22, 27, or 32. (That is, we are



**Figure 5.** A partially filled block and a filled block, respectively.

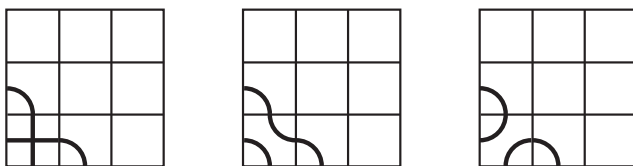
considering every option except those with tile number 24.) It must be one of the two options in Figure 5. All other  $3 \times 3$  corners are a rotation of one of these. We will refer to the first option as a *partially filled block* and the second option as a *filled block*.

**Observation 1.** In any space-efficient 6-mosaic of a prime knot, the tile in position  $C$  of a partially filled block is either a crossing tile or double arc  $T_7$ .

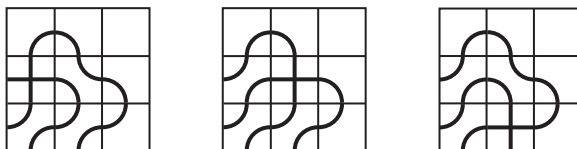
This is easy to see, as it must be a tile with four connection points, and the only space-efficient mosaics that results from using the double arc  $T_8$  are composite knots or links with more than one component. In Figure 6, the first two examples are valid possibilities, but the third one is not.

**Observation 2.** In any space-efficient 6-mosaic of a prime knot, there must be at least two crossing tiles in a filled block.

If there are no crossing tiles in positions  $A$ ,  $B$ ,  $C$ , and  $D$  of the mosaic, then the mosaic is not space-efficient or it is a link with more than one component. Each one that is not a link reduces to one of the last two partially filled block options in Figure 6. If there is only one crossing tile and it is in position  $A$ ,  $B$ , or  $D$ , then the mosaic is not space-efficient. For each option, if we fill the remaining tile positions with double arc tiles so that the block is suitably connected and we avoid the obvious inefficiencies we get the options shown in Figure 7. They are equivalent to each other via a simple mosaic planar isotopy move that rolls the crossing through each of these positions, and they all reduce to the first partially filled block in Figure 6. If there is only one crossing tile and it is in position  $C$ , then the mosaic is also not space-efficient and reduces to either of the first two options in Figure 6.

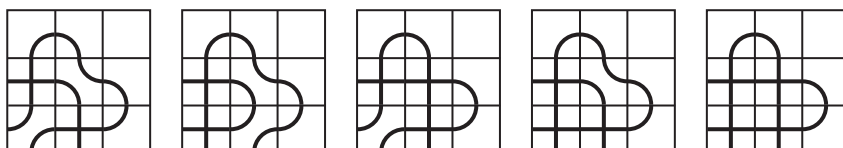


**Figure 6.** The first two examples are the only valid possibilities for a partially filled block.



**Figure 7.** Suitably connected filled blocks with one crossing in position  $A$ ,  $B$ , or  $D$ . None are space-efficient.

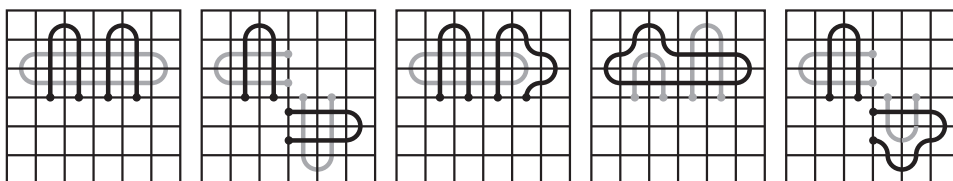
**Observation 3.** In a filled block in any space-efficient 6-mosaic of a prime knot, there are only two distinct possibilities for two crossing tiles, two distinct possibilities for three crossing tiles, and one possibility for four crossing tiles and they are shown below:



We will refer to the five filled blocks in Observation 3 together with the first two partially filled blocks in Figure 6 (and reflections and rotations of them) as *building blocks*. The observations provide a way for us to easily build all of the space-efficient 6-mosaics, as long as the tile number is 22, 27, or 32, but not 24.

**Observation 4.** In any space-efficient 6-mosaic of a prime knot, there is at most one of the filled block with four crossing tiles or the filled block with two crossings in positions  $A$  and  $C$ .

It is quite simple to verify that if there is more than one filled block with four crossings or more than one filled block with two crossings in positions  $A$  and  $C$ , the resulting mosaic must be a link with more than one component. If we use the indicated filled building block with two crossing tiles together with a filled block with four crossing tiles, the resulting mosaic will also be a link with more than one component. Several examples of these are pictured in Figure 8 with the second link component in each mosaic colored differently from the first link component.



**Figure 8.** These layouts will always be multicomponent links.

### 3. All prime knots with mosaic number 6

We are now ready to determine the tile number of every prime knot with mosaic number 6. Theorem 2 says that the only possible tile numbers are 22, 24, 27, and 32. In order to determine which knots have these tile numbers, we simply compile a list of the prime knots that can fit within each of the layouts given in Theorem 2. Because we already know the tile number of every prime knot with crossing number 7 or less, we can restrict our search to knots with crossing number 8 or more. The process is simple, and the above observations help us tremendously. If the tile number is 22, 27, or 32, we use the building blocks. In the case of the mosaics with tile number 24, we look at all possible placements, up to symmetry, of eight or more crossing tiles within the mosaics and fill the remaining tile positions with double arc tiles so as to avoid composite knots and nonreduced knots. Once the mosaics are completed, we then eliminate any links, any duplicate layouts that are equivalent to others via obvious mosaic planar isotopy moves, and any mosaics for which the tile number can easily be reduced by a simple mosaic planar isotopy move. Finally, we use Knotscape to determine what knots are depicted in the mosaic by choosing the crossings so that they are alternating, as well as all possible nonalternating combinations. We provide minimally space-efficient knot mosaics for every prime knot with mosaic number less than or equal to 6 in the table of knots in the online supplement.

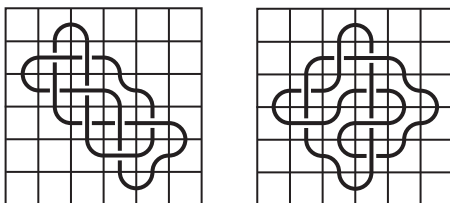
We have already listed several prime knots with tile number 22 in Theorem 1. This next theorem asserts that the list is complete.

**Theorem 3.** *The only prime knots  $K$  with tile number  $t(K) = 22$  are*

- (a)  $6_3$ ,
- (b)  $7_1, 7_2, 7_3, 7_5, 7_6, 7_7$ ,
- (c)  $8_1, 8_2, 8_3, 8_4, 8_7, 8_8, 8_9, 8_{13}$ ,
- (d)  $9_5$ , and  $9_{20}$ .

In order to obtain the minimally space-efficient knot mosaic for  $7_3$ , we had to use eight crossings. None of the possible minimally space-efficient knot mosaics with 22 nonblank tiles and exactly seven crossings produced  $7_3$ . The fewest number of nonblank tiles needed to represent  $7_3$  with only seven crossings is 24, and one such mosaic is given in Figure 9, along with a minimally space-efficient mosaic of  $7_3$  with eight crossings. In summary, on a minimally space-efficient knot mosaic, for the tile number (or minimal mosaic tile number) to be realized, it might not be possible for the crossing number to be realized. This is also the case with  $8_1, 8_3, 8_7, 8_8$ , and  $8_9$ , as nine crossing tiles are required to represent these knots on a mosaic with tile number 22.



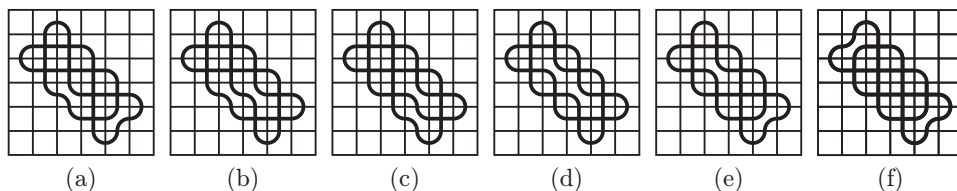


**Figure 9.** The  $7_3$  knot as a minimally space-efficient knot mosaic with eight crossing tiles and as a knot mosaic with seven crossing tiles.

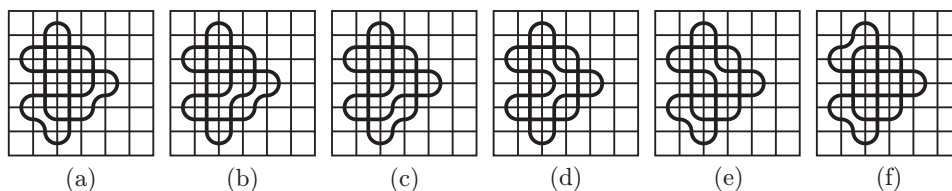
*Proof.* We simply build the first two tile configurations (both with 22 nonblank tiles) in Theorem 2 using the  $3 \times 3$  building blocks, eliminate any that do not satisfy the observations, choose specific crossing types, and see what we get. Whatever prime knots with eight or more crossings are missing are the ones we know cannot have tile number 22.

We begin with the first mosaic layout given in Theorem 2. Up to symmetry, there are only six possible configurations of this layout with eight crossings, and they are given in Figure 10. Notice that some of these are links that can be eliminated, including Figures 10(d) and (f). Furthermore, Figures 10(b) and (c) are equivalent to each other via a mosaic planar isotopy move that shifts one of the crossing tiles to a diagonally adjacent tile position. This leaves us with only three possible distinct configurations of eight crossings from this first layout, Figures 10(a), (b), and (e).

Now we do the same thing with the second mosaic layout given in Theorem 2 with 22 nonblank tiles. Up to symmetry, there are six possible configurations of this layout with eight crossings, and they are given in Figure 11. Again, Figures 11(d) and (f) are links, and Figures 11(b) and (c) are equivalent to each other. This leaves us again with only three possible configurations of eight crossings from this second layout, and they are Figures 11(a), (b), and (e). Moreover, each one of these is equivalent to the corresponding mosaics in Figure 10 via a few mosaic planar isotopy moves that shift the crossings in the lower-left building block into the lower-right building block of the mosaic.



**Figure 10.** Possible placements of eight crossing tiles in the first layout with tile number 22.

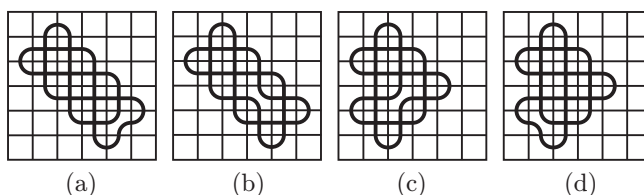


**Figure 11.** Possible placements of eight crossing tiles in the second layout with tile number 22.

This leaves us with only three distinct possible layouts for a minimally space-efficient  $6 \times 6$  mosaic with eight crossings and tile number 22. If we choose crossings for the configuration in Figure 10(a) so that they are alternating, we get the  $8_{13}$  knot. If we choose crossings for the configuration in Figure 10(b) so that they are alternating, we get the  $8_4$  knot. Finally, if we choose crossings for the configuration in Figure 10(e) so that they are alternating, we get the  $8_2$  knot. If we examine all possible nonalternating choices for each one, all of the resulting knots have crossing number 7 or less. (The minimally space-efficient knot mosaic for  $7_3$  must have eight crossing tiles and can be obtained by a choice of nonalternating crossings within any of the three distinct possible layouts in Figure 10.)

Now we go through the same process using nine crossing tiles. Up to symmetry, there are only four possible configurations of these layouts with nine crossings, and they are given in Figure 12. The mosaic in Figure 12(c) is equivalent to the mosaic in Figure 12(b) via a few mosaic planar isotopy moves that shift the crossings in the lower-left building block into the lower-right building block of the mosaic. This leaves us with only three possible configurations of nine crossing tiles.

If we choose crossings for the configuration in Figure 12(a) so that they are alternating, we get the  $9_{20}$  knot. If we examine all possible nonalternating choices for the crossings, most of the resulting knots have crossing number 7 or less, but we do get some additions to our list of prime knots with tile number 22 and crossing number 8. In particular, we get  $8_7$ ,  $8_8$ , and  $8_9$ . (We also get  $8_4$ , which was previously obtained with only eight crossings.) If we choose crossings for the configuration in Figure 12(b) so that they are alternating, we get the  $9_5$  knot. Again, if we examine



**Figure 12.** Possible placements of nine crossings with tile number 22.

the possible nonalternating choices for the crossings, we get two additional prime knots with tile number 22 and crossing number 8, and they are  $8_1$  and  $8_3$ . Finally, if we choose crossings for the configuration in Figure 12(d), we get the exact same knots as we did for Figure 12(a).

By Observation 4, we cannot place more than nine crossing tiles on any mosaic with 22 nonblank tiles. We have now found every possible prime knot with tile number 22 and eight or more crossings, and they are exactly those listed in the theorem. All other prime knots with crossing number at least 8 must have tile number larger than 22.  $\square$

We now know precisely which prime knots have tile number 22 or less. Our next goal is to determine which prime knots have tile number 24.

**Theorem 4.** *The only prime knots  $K$  with tile number  $t(K) = 24$  are*

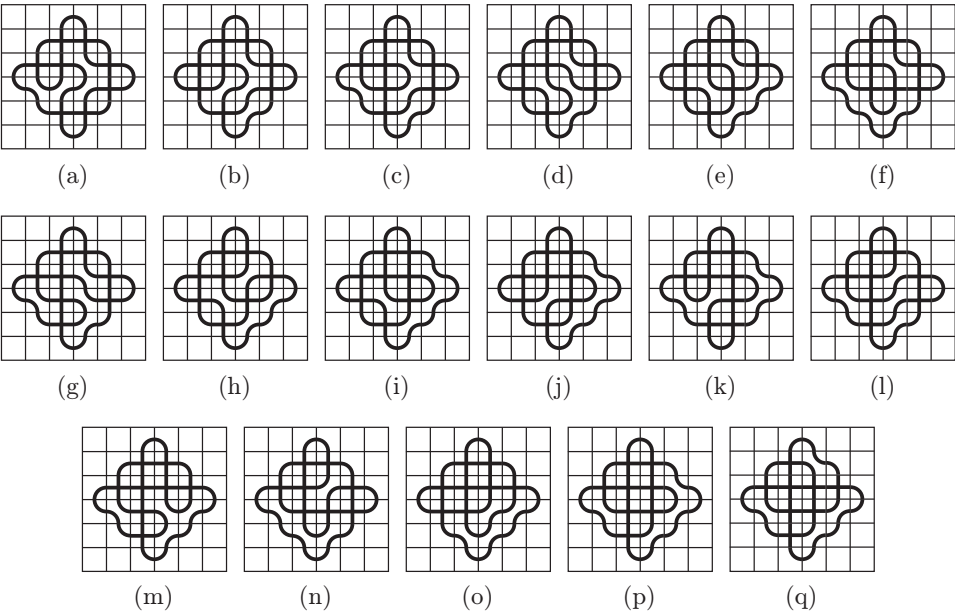
- (a)  $8_5, 8_6, 8_{10}, 8_{11}, 8_{12}, 8_{14}, 8_{16}, 8_{17}, 8_{18}, 8_{19}, 8_{20}, 8_{21},$
- (b)  $9_8, 9_{11}, 9_{12}, 9_{14}, 9_{17}, 9_{19}, 9_{21}, 9_{23}, 9_{26}, 9_{27}, 9_{31},$
- (c)  $10_{41}, 10_{44}, 10_{85}, 10_{100}, 10_{116}, 10_{124}, 10_{125}, 10_{126}, 10_{127}, 10_{141}, 10_{143}, 10_{148},$   
 $10_{155}$  and  $10_{159}.$

We will show that  $8_6$  must have nine crossing tiles to fit on a mosaic with tile number 24. None of the possible minimally space-efficient knot mosaics with exactly eight crossings produce these knots. Similarly, the minimally space-efficient mosaics for  $9_{12}, 9_{19}, 9_{21},$  and  $9_{26}$  require 10 crossings.

*Proof.* We search for all of the prime knots that have tile number 24. In this particular case, the observations at the beginning of this section do not apply, meaning we cannot use the building blocks as we did in the proof of Theorem 3. We know from Theorem 2 that any prime knot with tile number 24 has a space-efficient mosaic, like the third layout there. We simply look at all possible placements of eight or more crossings within that layout, choose the type of each crossing, and keep track of the resulting prime knots.

First, we look at all possible placements, up to symmetry, of eight crossings within the mosaic and, we fill the remaining tile positions with double arc tiles so as to avoid composite knots and unnecessary loops. After eliminating any links and any duplicate layouts that are equivalent to others via simple mosaic planar isotopy moves, we get 17 possible layouts, which are shown in Figure 13. Not all of these will result in distinct knots, and in most cases it is not difficult to see that they will result in the same knot. However, we include all of them here because they differ by more than just simple symmetries or simple mosaic planar isotopy moves.

Choosing specific crossings so that the knots are alternating, we obtain only 14 distinct knots as shown in the following table:

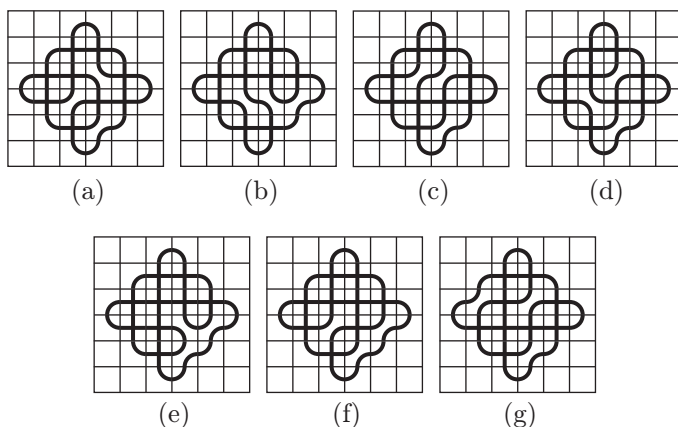


**Figure 13.** Only possible layouts, after elimination, with eight crossing tiles for a prime knot with tile number 24.

Figure 13	knot	Figure 13	knot
(a)	$8_1$	(j)	$8_{11}$
(b), (c)	$8_2$	(k)	$8_{12}$
(d)	$8_4$	(l)	$8_{13}$
(e)	$8_5$	(m), (n)	$8_{14}$
(f), (g)	$8_7$	(o)	$8_{16}$
(h)	$8_8$	(p)	$8_{17}$
(i)	$8_{10}$	(q)	$8_{18}$

Not all of these have tile number 24. We already know  $8_1$ ,  $8_2$ ,  $8_4$ ,  $8_7$ ,  $8_8$ , and  $8_{13}$  have tile number 22. Each of the others have tile number 24. The nonalternating knots  $8_{19}$ ,  $8_{20}$ , and  $8_{21}$  are obtained by choosing nonalternating crossings in a few of these. Those pictured in the table of knots come from the layout in Figure 13(p). Mosaics for all of these are given in the table of knots in the online supplement. The only knots with crossing number 8 that we have not yet found are  $8_6$  and  $8_{15}$ , and now we know that they cannot be represented with eight crossings and 24 nonblank tiles.

We now turn our attention to mosaics with nine crossings. Just as before, we look at all possible placements, up to symmetry, of nine crossings, eliminate any composite knots, unnecessary loops, links and any duplicate layouts that are equivalent to others via simple mosaic planar isotopy moves. In the end, we get

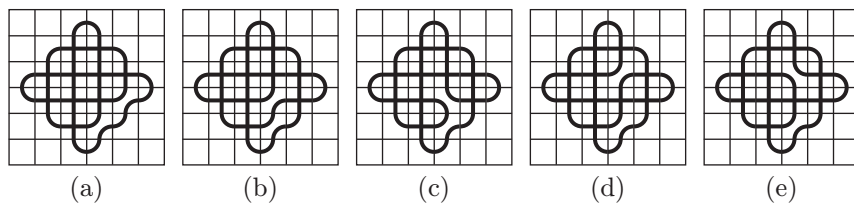


**Figure 14.** Only possible layouts, after elimination, with nine crossing tiles for a prime knot with tile number 24.

seven possible layouts, which are shown in Figure 14. Choosing specific crossings for each layout, in order, so that the knots are alternating, we obtain the seven knots  $9_8$ ,  $9_{11}$ ,  $9_{14}$ ,  $9_{17}$ ,  $9_{23}$ ,  $9_{27}$ , and  $9_{31}$ , all of which have tile number 24. If we look at all possible choices for nonalternating crossings, the only knot with tile number 24 that arises but did not show up with only eight crossing tiles is the  $8_6$  knot, whose knot mosaic in the table of knots comes from the layout in Figure 14(a). All other prime knots that arise using nonalternating crossings have been exhibited as a minimally space-efficient mosaic with fewer crossings or fewer nonblank tiles.

Now we do the same for 10 crossings. Again, we observe all possible placements of 10 crossings on the third mosaic in Theorem 2, and after eliminating any links and duplicate layouts up to reflection, rotation, or equivalencies via simple mosaic planar isotopy moves, we end up with five possible layouts, shown in Figure 15.

We begin with Figure 15(a). Choosing specific crossings so that the knot is alternating, we obtain the  $10_{116}$  knot. If we look at all possible choices for nonalternating crossings, the only prime knots that we get with tile number 24 are the nonalternating knots  $10_{124}$ ,  $10_{125}$ ,  $10_{141}$ ,  $10_{143}$ ,  $10_{155}$ , and  $10_{159}$ . We do the same with Figure 15(b) and get the alternating knot  $10_{100}$ . For the nonalternating choices, we get almost all of the same ones we just obtained, but we do not get any new additions to our list of knots. For Figure 15(c), with alternating crossings we get  $10_{41}$ , and with nonalternating crossings we get  $9_{19}$  and  $9_{21}$  as the only new additions to our list. Neither of these came from considering only nine crossings. Now we observe the mosaic in Figure 15(d). By alternating the crossings, we obtain  $10_{44}$ , and by using nonalternating crossings, the only new additions to our list are  $9_{12}$  and  $9_{26}$ . Finally, we end with Figure 15(e). Assigning alternating crossings, we get  $10_{85}$ , and assigning nonalternating crossings, we get  $10_{126}$ ,  $10_{127}$ , and  $10_{148}$ .



**Figure 15.** Only possible layouts, after elimination, with 10 crossing tiles for a prime knot with tile number 24.

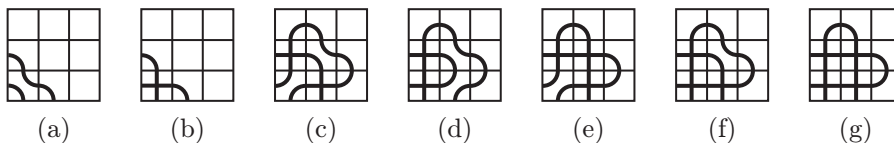
Finally, we can place 11 or 12 crossing tiles into the layout with 24 nonblank tiles, but the space-efficient results will always be a link with more than one component. Therefore, no minimally space-efficient prime knot mosaics arise from this consideration. We have considered every possible placement of crossing tiles on the third layout in Theorem 2 and have found every possible prime knot with tile number 24 and eight or more crossings, and they are exactly those listed in the theorem. Minimally space-efficient mosaics for all of these knots are given in the table of knots in the online supplement. All other prime knots with crossing number at least 8 must have tile number larger than 24.  $\square$

We now know precisely which prime knots have tile number less than or equal to 24, and we are ready to determine which prime knots with mosaic number 6 have tile number 27. We see our first occurrence of knots with crossing number larger than 10, and we use the Dowker–Thistlethwaite name of the knot.

**Theorem 5.** *The only prime knots  $K$  with mosaic number 6, tile number  $t(K) = 27$ , and minimal mosaic tile number  $t_M(K) = 27$  are*

- (a)  $8_{15}$ ,
- (b)  $9_1, 9_2, 9_3, 9_4, 9_7, 9_9, 9_{13}, 9_{24}, 9_{28}, 9_{37}, 9_{46}, 9_{48}$ ,
- (c)  $10_1, 10_2, 10_3, 10_4, 10_{12}, 10_{22}, 10_{28}, 10_{34}, 10_{63}, 10_{65}, 10_{66}, 10_{75}, 10_{78}, 10_{140}, 10_{142}, 10_{144}$ ,
- (d)  $11a_{107}, 11a_{140}$ , and  $11a_{343}$ .

Notice that this theorem is only referring to prime knots with mosaic number 6. There are certainly prime knots with tile number 27 and mosaic number 7 that are not included in this theorem. Also, the requirement that the tile number equals the minimal mosaic tile number is necessary here. As far as we know now (and will verify below), there are knots with mosaic number 6 and tile number 27 which have minimal mosaic number 32. Some of these are listed in the next theorem. Finally, notice that up to this point we have determined the tile number for every prime knot with crossing number 8 or less.



**Figure 16.** The seven building blocks resulting from the observations at the beginning of this section.

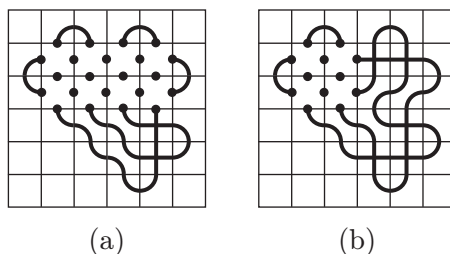
Again we claim that the minimally space-efficient mosaics for  $9_3$ ,  $9_4$ ,  $9_{13}$ ,  $9_{37}$ ,  $9_{46}$ , and  $9_{48}$  must have 10 crossing tiles. The minimally space-efficient mosaics for  $9_7$ ,  $9_9$ , and  $9_{24}$  must have 11 crossing tiles. None of the possible minimally space-efficient knot mosaics with exactly nine crossing tiles produce these knots. Similarly, the minimally space-efficient mosaics for  $10_1$ ,  $10_3$ ,  $10_{12}$ ,  $10_{22}$ ,  $10_{34}$ ,  $10_{63}$ ,  $10_{65}$ ,  $10_{78}$ ,  $10_{140}$ ,  $10_{142}$ , and  $10_{144}$  require 11 crossing tiles.

*Proof.* Similar to what we did in the proof of Theorem 3, we search for all of the prime knots that have mosaic number 6 and tile number 27, which have a space-efficient mosaic as depicted in the fourth layout of Theorem 2. We simply build this layout using the  $3 \times 3$  building blocks that result from the observations at the beginning of this section, shown again in Figure 16. We then choose specific crossing types for each crossing tile and see what knots we get.

For bookkeeping purposes, we note that the knot  $8_{15}$  has tile number 27, and this is the only knot with crossing number 8 for which we have not previously found the tile number. A minimally space-efficient mosaic for it is included in the table of knots in the online supplement. We now know the tile number for every prime knot with crossing number 8 or less, and from here we restrict our search to mosaics with nine or more crossing tiles.

Before we get started placing crossing tiles, we make a few more simple observations that apply to this particular case and help us reduce the number of possible configurations. Observe that if we place a partially filled building block with no crossing adjacent to the filled building block with two crossing tiles in Figure 16(c), the resulting mosaic will always reduce to a mosaic with tile number 22. The same result holds if the two blocks are not adjacent and one of the adjacent blocks is the filled building block with three crossings depicted in Figure 16(e). The mosaics in Figure 17 exhibit these scenarios. The same result also holds if the partially filled building block with one crossing is combined with two of the filled building blocks with two crossing tiles shown in Figure 16(c). Depending on the placement of these two filled blocks, the result will be equivalent to either Figure 17(a) or Figure 17(b) via a simple mosaic planar isotopy move that shifts the crossing in the partially filled block to another block.

First, we consider nine crossing tiles with the above observations in mind, together with the observations at the beginning of this section. Up to symmetry, there are

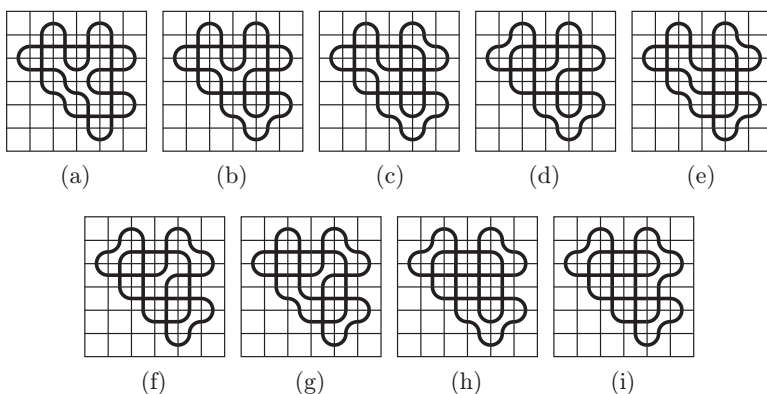


**Figure 17.** These two mosaics are not minimally space-efficient.

only nine possible configurations of the building blocks after we eliminate the links, duplicate layouts that are equivalent to others via simple mosaic planar isotopy moves, and any mosaics for which the tile number can easily be reduced by a simple mosaic planar isotopy move. They are shown in Figure 18. Not all of these will result in distinct knots, and in several cases it is not difficult to see that they will result in the same knot. However, we include all of them here because they differ by more than just symmetries or a simple mosaic planar isotopy move.

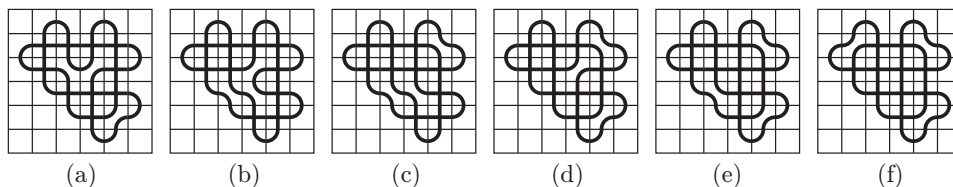
Choosing specific crossings so that the knots are alternating, we obtain only seven distinct knots. The only ones with tile number 27 are Figure 18(a), which gives the  $9_1$  knot, Figure 18(b), which gives us  $9_2$ , and Figures 18(h) and (i), which give us  $9_{28}$ . Each of the remaining layouts give knots with tile number less than 27. In particular, Figures 18(c) and (d) are  $9_8$ , Figures 18(e) and (f) are  $9_{17}$ , and Figure 18(g) is  $9_{20}$ . None of these configurations give nonalternating knots with crossing number 9.

Second, we do the same for 10 crossings. Again, we use the building blocks to build all possible configurations of the crossings, and up to symmetry, there are only



**Figure 18.** Only possible layouts, after elimination, with nine crossing tiles for a prime knot with tile number 27.





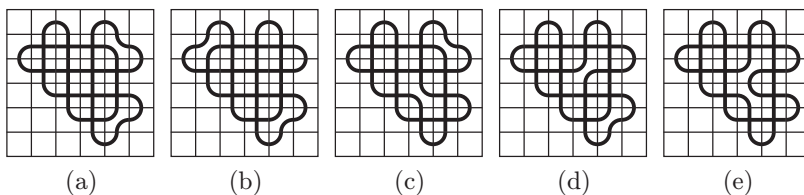
**Figure 19.** Only possible layouts, after elimination, with 10 crossing tiles for a prime knot with tile number 27.

six possibilities after eliminating any links and duplicate layouts that are equivalent via simple mosaic planar isotopy moves. These are shown in Figure 19.

Choosing specific crossings so that the knots are alternating, we obtain only five distinct knots, all of which have tile number 27. In particular, Figure 19(a) becomes the  $10_2$  knot, Figure 19(b) becomes  $10_4$ , Figures 19(c) and (d) become  $10_{28}$ , Figure 19(e) becomes  $10_{66}$ , and Figure 19(f) becomes  $10_{75}$ . Choosing nonalternating crossings, we also get some knots with crossing number 9, but we do not obtain any nonalternating knots with crossing number 10. We can get  $9_3$  from Figure 19(a),  $9_4$  from Figure 19(b),  $9_{13}$  from Figure 19(c), and  $9_{37}$ ,  $9_{46}$ , and  $9_{48}$  from Figure 19(f). All other knots that are obtained by considering nonalternating crossings can be drawn with fewer crossings or a lower tile number.

Third, we consider the case where the mosaic has 11 crossing tiles. In this instance, we end up with the five possible layouts shown in Figure 20, and again, not all of these are distinct. Choosing alternating crossing in each layout results in three distinct knots with crossing number 11. Figures 20(a) and (b) become  $11a_{107}$ , Figures 20(c) and (d) become  $11a_{140}$ , and Figure 20(e) becomes  $11a_{343}$ . (Note that, for knots with crossing number greater than 10, we are using the Dowker–Thistlethwaite name of the knot.) Choosing nonalternating crossings in each of the layouts results in several knots with crossing number 9 or 10. In particular, we can obtain the knots  $9_{24}$ ,  $10_{63}$ ,  $10_{65}$ ,  $10_{78}$ ,  $10_{140}$ ,  $10_{142}$ , and  $10_{144}$  from Figure 20(a). We can obtain  $9_7$ ,  $9_9$ ,  $10_{12}$ ,  $10_{22}$ , and  $10_{34}$  from Figure 20(c). And we can obtain  $10_1$  and  $10_3$  from Figure 20(e). All of these are shown in the table of knots in the online supplement. All other knots that are obtained by considering nonalternating crossings can be drawn with fewer crossings or a lower tile number.

Finally, by Observation 4 we do not need to consider 12 or more crossing tiles in this layout, as no minimally space-efficient prime knot mosaics arise from this consideration. We have considered every possible placement of nine or more crossing tiles on the fourth layout in Theorem 2 and have found every possible prime knot with mosaic number 6 and tile number 27. They are exactly those listed in the theorem. All other prime knots with crossing number at least 9 and mosaic number 6 must have minimal mosaic tile number 32.  $\square$



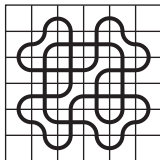
**Figure 20.** Only possible layouts, after elimination, with 11 crossing tiles for a prime knot with tile number 27.

Now we know the tile number for every prime knot with crossing number less than or equal to 8. Theorems 3, 4, and 5 tell us the tile number of some of the prime knots with crossing numbers 9, 10, and 11. Furthermore, we know that all other prime knots with mosaic number 6 must have minimal mosaic tile number 32 but not necessarily tile number 32. One problem that complicates the next step is that, as of the writing of this paper, we do not know the mosaic number of all prime knots with crossing number 9 or more. That is, we do not know all prime knots with mosaic number 6. For this reason, we need to go through the same process as we did in the preceding proofs to determine which prime knots have mosaic number 6 and minimal mosaic tile number 32. By doing this, we will also be able to determine which prime knots have mosaic number greater than 6. The good news is that this is the final step in determining which prime knots have mosaic number 6 or less and determining the tile number or minimal mosaic tile numbers of all of these.

**Theorem 6.** *The only prime knots  $K$  with mosaic number 6 and minimal mosaic tile number  $t_M(K) = 32$  are*

- (a)  $9_{10}, 9_{16}, 9_{35}$ ,
- (b)  $10_{11}, 10_{20}, 10_{21}, 10_{61}, 10_{62}, 10_{64}, 10_{74}, 10_{76}, 10_{77}, 10_{139}$ ,
- (c)  $11a_{43}, 11a_{44}, 11a_{46}, 11a_{47}, 11a_{58}, 11a_{59}, 11a_{106}, 11a_{139}, 11a_{165}, 11a_{166}, 11a_{179},$   
 $11a_{181}, 11a_{246}, 11a_{247}, 11a_{339}, 11a_{340}, 11a_{341}, 11a_{342}, 11a_{364}, 11a_{367}$ ,
- (d)  $11n_{71}, 11n_{72}, 11n_{73}, 11n_{74}, 11n_{75}, 11n_{76}, 11n_{77}, 11n_{78}$ ,
- (e)  $12a_{119}, 12a_{165}, 12a_{169}, 12a_{373}, 12a_{376}, 12a_{379}, 12a_{380}, 12a_{444}, 12a_{503}, 12a_{722},$   
 $12a_{803}, 12a_{1148}, 12a_{1149}, 12a_{1166}$ ,
- (f)  $13a_{1230}, 13a_{1236}, 13a_{1461}, 13a_{4573}$ ,
- (g)  $13n_{2399}, 13n_{2400}, 13n_{2401}, 13n_{2402}$ , and  $13n_{2403}$ .

Notice again our restriction to prime knots with mosaic number 6. Additionally, notice that this theorem only refers to the minimal mosaic tile number of the knot, not the tile number. Again, this is because we only know that these two numbers are equal when they are less than or equal to 27. Some of these knots may have (and actually do have) tile number less than 32.



**Figure 21.** Only possible layout, after elimination, with nine crossing tiles for a prime knot with minimal mosaic tile number 32.

We claim that the minimally space-efficient mosaics for  $9_{10}$ ,  $9_{16}$ ,  $10_{20}$ ,  $10_{21}$ , and  $10_{77}$  need 11 crossing tiles. The minimally space-efficient mosaics for  $9_{35}$ ,  $10_{11}$ ,  $10_{62}$ ,  $10_{64}$ ,  $10_{74}$ ,  $10_{139}$ ,  $11a_{106}$ ,  $11a_{139}$ ,  $11a_{166}$ ,  $11a_{181}$ ,  $11a_{341}$ ,  $11a_{342}$ , and  $11a_{364}$  need 12 crossing tiles. And the minimally space-efficient mosaics for  $10_{61}$ ,  $10_{76}$ ,  $11a_{44}$ ,  $11a_{47}$ ,  $11a_{58}$ ,  $11n_{76}$ ,  $11n_{77}$ ,  $11n_{78}$ ,  $11a_{165}$ ,  $11a_{246}$ ,  $11a_{339}$ ,  $11a_{340}$ ,  $12a_{119}$ ,  $12a_{165}$ ,  $12a_{169}$ ,  $12a_{376}$ ,  $12a_{379}$ ,  $12a_{444}$ ,  $12a_{803}$ ,  $12a_{1148}$ , and  $12a_{1166}$  need 13 crossing tiles.

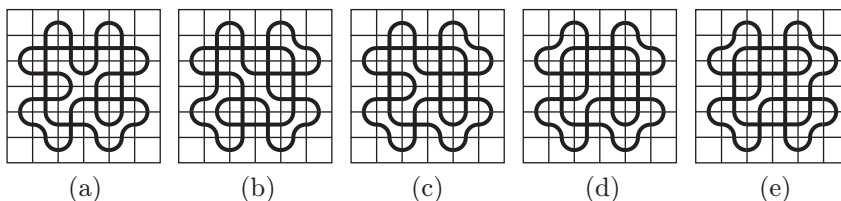
*Proof.* We simply go through the same process that we did in the previous proof. We search for all of the prime knots that have mosaic number 6 and minimal mosaic tile number 32. Whatever prime knots that do not show up in this process and that we have not previously determined the tile number for must have mosaic number greater than 6. We know from Theorem 2 that any prime knot with mosaic number 6 and minimal mosaic tile number 32 has a space-efficient mosaic with the fifth and final layout shown there.

As we have done several times previously, we use the building blocks to achieve all possible configurations, up to symmetry, of nine or more crossings within this mosaic. For this particular layout, we can only use the filled blocks, not the partially filled blocks. We can eliminate any layouts that do not meet the requirements of the observations, any multicomponent links, any duplicate layouts that are equivalent to others via simple mosaic planar isotopy moves, and any mosaics for which the tile number can easily be reduced by a simple mosaic planar isotopy move.

First, in the case of nine crossings, after we eliminate the unnecessary layouts we end up with only one possibility, and it is shown in Figure 21. However, once we choose specific crossings in an alternating fashion, it is the knot  $9_8$ , which has tile number 24. Nothing new arises from considering nonalternating crossings either.

Second, we do the same for 10 crossings, and we end up with five possible layouts, shown in Figure 22. Choosing alternating crossings in each one, we again fail to get any prime knots with minimal mosaic tile number 32. Figure 22(a) is  $10_1$ , Figure 22(b) and (c) are  $10_{34}$ , and Figures 22(d) and (e) are  $10_{78}$ . Nothing new arises from considering nonalternating crossings either.

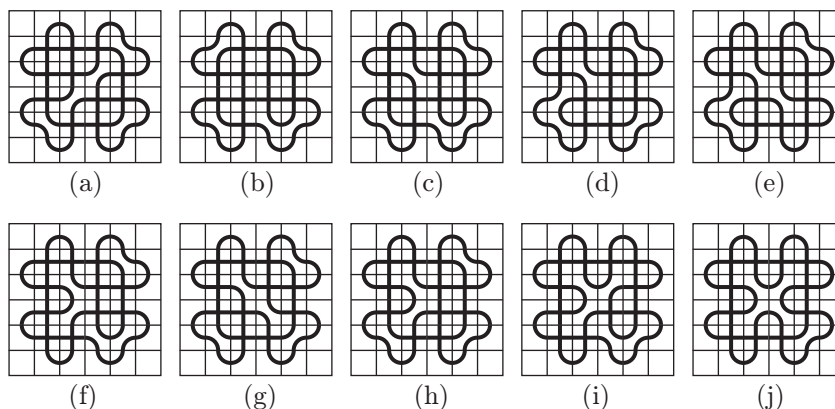
Third, we consider the case where the mosaic has 11 crossing tiles. In this instance, we end up with the 10 possible layouts shown in Figure 23. With alternating



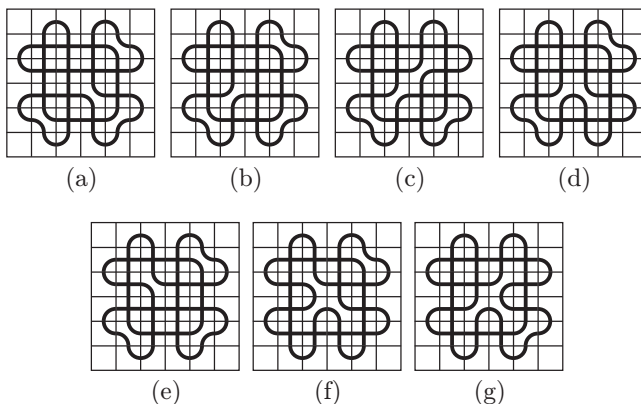
**Figure 22.** Only possible layouts, after elimination, with 10 crossing tiles for a prime knot with minimal mosaic tile number 32.

crossings, the first layout is  $11a_{140}$ , which we already know has tile number 27. The remaining layouts, given alternating crossings, lead to six distinct knots with minimal mosaic tile number 32, and with nonalternating crossings we get 10 additional knots that have minimal mosaic tile number 32. In particular, Figure 23(b) with alternating crossings is  $11a_{43}$  and with nonalternating crossings can be made into  $11n_{71}$ ,  $11n_{72}$ ,  $11n_{73}$ ,  $11n_{74}$ , and  $11n_{75}$ . Figures 23(c) and (d) are  $11a_{46}$  when using alternating crossings and can be made into  $9_{16}$  or  $10_{77}$  with nonalternating crossings. Figures 23(e) and (f) are  $11a_{59}$  when using alternating crossings and can be made into  $10_{20}$  with nonalternating crossings. Figures 23(g) and (h) are  $11a_{179}$  when using alternating crossings and can be made into  $9_{10}$  or  $10_{21}$  with nonalternating crossings. Figure 23(i) with alternating crossings is  $11a_{247}$ , and Figure 23(j) with alternating crossings is  $11a_{367}$ . Neither of these last two provide new knots to our list when considering nonalternating crossings.

Fourth, we consider the possibilities where the mosaic has 12 crossing tiles. In this case, we end up with the seven possible layouts shown in Figure 24. With alternating crossings, these layouts lead to five distinct knots with minimal mosaic



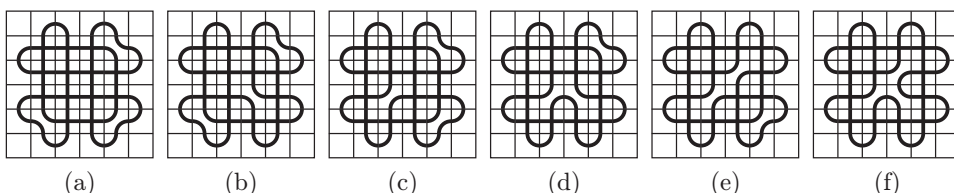
**Figure 23.** Only possible layouts, after elimination, with 11 crossing tiles for a prime knot with minimal mosaic tile number 32.



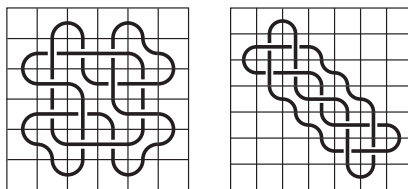
**Figure 24.** Only possible layouts, after elimination, with 12 crossing tiles for a prime knot with minimal mosaic tile number 32.

tile number 32, and with nonalternating crossings we get 13 additional knots that have minimal mosaic tile number 32. In particular, Figures 24(a) and (b) with alternating crossings are  $12a_{373}$  and with nonalternating crossings can be made into  $10_{62}$ ,  $10_{64}$ ,  $10_{139}$ ,  $11a_{106}$ , or  $11a_{139}$ . Figures 24(c) and (d) are  $12a_{380}$  when using alternating crossings and can be made into  $10_{11}$ ,  $11a_{166}$ , or  $11a_{341}$  with nonalternating crossings. Figure 24(e) is  $12a_{503}$  when using alternating crossings and can be made into  $9_{35}$ ,  $10_{74}$ , or  $11a_{181}$  with nonalternating crossings. Figure 24(f) is  $12a_{722}$  when using alternating crossings and can be made into  $11a_{364}$  with nonalternating crossings. Figure 24(g) with alternating crossings is  $12a_{1149}$  and with nonalternating crossings can be  $11a_{342}$ .

Fifth, we consider what happens when we place 13 crossing tiles on the mosaic. In this instance, we end up with the six possible layouts shown in Figure 25. With alternating crossings, the layouts lead to four distinct knots with minimal mosaic tile number 32, and with nonalternating crossings we get 26 additional knots that have minimal mosaic tile number 32. In particular, Figure 25(a) with alternating crossings is  $13a_{1230}$  and with nonalternating crossings can be made into  $11a_{44}$ ,  $11a_{47}$ ,  $11n_{76}$ ,  $11n_{77}$ ,  $11n_{78}$ ,  $12a_{119}$ ,  $13n_{2399}$ ,  $13n_{2400}$ ,  $13n_{2401}$ ,  $13n_{2402}$ , or  $13n_{2403}$ .



**Figure 25.** Only possible layouts, after elimination, with 13 crossing tiles for a prime knot with minimal mosaic tile number 32.



**Figure 26.** The  $9_{10}$  knot represented as a minimally space-efficient 6-mosaic with minimal mosaic tile number 32 and as a space-efficient 7-mosaic with tile number 27.

Figures 25(b) and (c) are  $13a_{1236}$  when using alternating crossings and can be made into  $10_{61}$ ,  $10_{76}$ ,  $11a_{58}$ ,  $11a_{165}$ ,  $11a_{340}$ ,  $12a_{165}$ ,  $12a_{376}$ , or  $12a_{444}$  with nonalternating crossings. Figures 25(d) and (e) are  $13a_{1461}$  when using alternating crossings and can be made into  $11a_{246}$ ,  $11a_{339}$ ,  $12a_{169}$ ,  $12a_{379}$ , or  $12a_{1148}$  with nonalternating crossings. Figure 25(f) is  $13a_{4573}$  when using alternating crossings and can be made into  $12a_{803}$  or  $12a_{1166}$  with nonalternating crossings.

Finally, by Observation 4, we do not need to consider 14 or more crossing tiles in this layout. We have considered every possible placement of nine or more crossing tiles on the final layout of Theorem 2 and have found every possible prime knot with mosaic number 6 and minimal mosaic tile number 32.  $\square$

Because of the work we have completed, we now know every prime knot with mosaic number 6 or less. We also know the tile number or minimal mosaic tile number of each of these prime knots. In the table of knots in online supplement, we provide minimally space-efficient knot mosaics for all of these. These preceding theorems lead us to the following interesting consequences.

**Corollary 7.** *The prime knots with crossing number at least 9 not listed in Theorems 3, 4, 5, or 6 have mosaic number 7 or higher.*

**Theorem 8.** *The tile number of a knot is not necessarily equal to the minimal mosaic tile number of a knot.*

*Proof.* According to Theorem 6, the minimal mosaic tile number for  $9_{10}$  is 32. However, on a 7-mosaic, this knot can be represented using only 27 nonblank tiles, as depicted in Figure 26. Also note that, as a 7-mosaic, this knot could be represented with only nine crossings, whereas 11 crossings were required to represent it as a 6-mosaic.  $\square$

## References

- [Adams 1994] C. C. Adams, *The knot book*, Freeman, New York, 1994. MR Zbl
- [Heap and Knowles 2018] A. Heap and D. Knowles, “Tile number and space-efficient knot mosaics”, *J. Knot Theory Ramifications* **27**:6 (2018), art. id. 1850041. MR Zbl

- [Howards and Kobin 2018] H. Howards and A. Kobin, “Crossing number bounds in knot mosaics”, *J. Knot Theory Ramifications* **27**:10 (2018), art. id. 1850056. MR Zbl
- [Kuriya and Shehab 2014] T. Kuriya and O. Shehab, “The Lomonaco–Kauffman conjecture”, *J. Knot Theory Ramifications* **23**:1 (2014), art. id. 1450003. MR Zbl
- [Lee, Ludwig, Paat, and Peiffer 2018] H. J. Lee, L. Ludwig, J. Paat, and A. Peiffer, “Knot mosaic tabulation”, *Involve* **11**:1 (2018), 13–26. MR Zbl
- [Lomonaco and Kauffman 2008] S. J. Lomonaco and L. H. Kauffman, “Quantum knots and mosaics”, *Quantum Inf. Process.* **7**:2-3 (2008), 85–115. MR Zbl
- [Rolfsen 1976] D. Rolfsen, *Knots and links*, Math. Lecture Series **7**, Publish or Perish, Berkeley, CA, 1976. MR Zbl
- [Thistlethwaite and Hoste 1999] M. Thistlethwaite and J. Hoste, “Knotscape, a program for the study of knots”, 1999, available at <http://www.math.utk.edu/~morwen/knotscape.html>.

Received: 2018-03-28      Revised: 2018-10-04      Accepted: 2018-12-27

heap@geneseo.edu	<i>Department of Mathematics, State University of New York at Geneseo, Geneseo, NY, United States</i>
dougdknowles@gmail.com	<i>Department of Mathematics, Dartmouth College, Hanover, NH, United States</i>





# Shabat polynomials and monodromy groups of trees uniquely determined by ramification type

Naiomi Cameron, Mary Kemp, Susan Maslak, Gabrielle Melamed,  
Richard A. Moy, Jonathan Pham and Austin Wei

(Communicated by Vadim Ponomarenko)

A dessin d'enfant or dessin is a bicolored graph embedded into a Riemann surface. Acyclic dessins can be described analytically by preimages of Shabat polynomials and algebraically by their monodromy groups. We determine the Shabat polynomials and monodromy groups of planar acyclic dessins that are uniquely determined by their ramification types.

## 1. Introduction

Popularized by Grothendieck in his “Esquisse d'un programme”, the theory of *dessins* reaches across and connects multiple disciplines, including graph theory, topology, geometry, algebra and complex analysis. Our motivation for this paper is rooted in one of the fundamental questions in the theory of dessins — that is, how to distinguish classes of dessins by means of topological, algebraic or combinatorial invariants. In this paper, we focus our attention on this question by studying dessins which are also trees. Since such dessins by any measure might be considered among the simplest, it is worthwhile to have a complete catalog of the Belyi maps and monodromy groups to which they correspond.

Our main objective in this paper is to determine the Shabat polynomials (up to isomorphism) and monodromy groups corresponding to every known planar connected acyclic dessin uniquely determined by its ramification type, the complete list of which was given in [Shabat and Zvonkin 1994]. We begin in Section 1 by providing the main result of the paper, followed by definitions and notation needed to describe the class of dessins with which we are concerned, as well as some necessary background about Shabat polynomials and wreath products. Readers already acquainted with these subjects may wish to read Section 1A and skip Section 1B. In Section 2 we provide a unique (up to isomorphism) Shabat polynomial

---

*MSC2010:* primary 11G32, 14H57; secondary 20E22.

*Keywords:* dessins d'enfant, Shabat polynomials, monodromy groups, Belyi maps, trees, wreath products.

for each ramification type corresponding to exactly one (planar) bicolored tree; in Section 3 we provide the monodromy groups for each such ramification type. In Section 4, we suggest future directions that may be taken from the results presented here.

**1A. Main results.** Here, we state the main result of the paper in the following theorem. The remainder of this section provides the background and preliminaries for the rest of the paper. Theorem 1.1 lists the ramification types which correspond to exactly one dessin which is a tree, along with the associated monodromy groups and Shabat polynomials. Theorem 1.1 contains every such ramification type, as asserted in [Shabat and Zvonkin 1994]. In Sections 2 and 3, we argue that Theorem 1.1 lists the correct Shabat polynomials and monodromy groups.

**Theorem 1.1.** *The following list includes all seven ramification types (degrees of black vertices followed by degrees of white vertices) that produce exactly one dessin which is a tree (see [Shabat and Zvonkin 1994]). Each ramification type given on the list is followed by (a) the Shabat polynomial (unique up to isomorphism) and (b) the monodromy group for the dessin.*

(1)  $[r; 1^r]$

(a)  $z^r$

(b)  $C_r$

(2)  $[2^r, 1; 2^r, 1]$

(a)  $\frac{1}{2}(1 + \cos((2r + 1) \arccos(z)))$

(b)  $D_{2(2r+1)}$ , where  $D_m$  denotes the dihedral group of order  $m$

(3)  $[2^r; 2^{r-1}, 1^2]$

(a)  $\frac{1}{2}(1 + \cos(2r \arccos(z)))$

(b)  $D_{2(2r)}$

(4)  $[s^{r-1}, t; r, 1^{(r-1)(s-1)+(t-1)}]$  for  $r > 1, t > 0$

(a)  $(1 - z)^t \left( \sum_{k=0}^{r-1} \binom{t}{s}_k \frac{z^k}{k!} \right)^s$

(b)  $\begin{cases} C_r \wr C_s, & s = t, \\ S_{n/d} \wr C_d, & s \neq t, r \text{ even}, \\ A_{n/d} \wr C_d, & s \neq t, r \text{ odd and } \frac{t}{d} \text{ is odd}, \\ (A_{n/d})^d \rtimes C_{2d}, & s \neq t, r \text{ odd, } \frac{t}{d} \text{ even}, \end{cases}$

where  $n = s(r - 1) + t, d = \gcd(s, t)$ .

(5)  $[r, t, 1^{r+t-2}; 2^{r+t-1}]$ ,  $r, t > 1$

(a)  $4z^r(1 - z)^t \left( \sum_{j=0}^{r-1} \binom{t-1+j}{t-1} z^j \right) \left( \sum_{j=0}^{t-1} \binom{r-1+j}{r-1} \binom{r+t-1}{r+j} (-1)^j z^j \right)$

$$(b) \begin{cases} A_{2r-1} \times C_2, & r = t, \text{ } r \text{ odd}, \\ S_{2r-1} \times C_2, & r = t, \text{ } r \text{ even}, \\ A_{r+t-1} \wr C_2, & r \neq t, \text{ both odd}, \\ R_2, & r \neq t, \text{ both even}, \\ S_{r+t-1} \wr C_2, & r \neq t, \text{ else}, \end{cases}$$

where  $R_m$  denotes the index-2 subgroup of  $S_{n/m} \wr C_m$  such that, for all  $(\tau_1, \dots, \tau_m, g) \in R_m$ , the permutation  $\tau_1 \tau_2 \cdots \tau_m$  is even.

$$(6) [r^2, 1^{4r-3}; 3^{2r-1}]$$

$$(a) -3\sqrt{3} i S_r(z)(1 - S_r(z))(S_r(z) - \frac{1}{2}(1 - i\sqrt{3}))$$

$$(b) \begin{cases} A_{2r-1} \wr C_3, & r \text{ odd}, \\ R_3, & r \text{ even} \end{cases}$$

$$(7) [3^3, 1^5; 2^7]$$

$$(a) -\frac{4}{531441}(z-1)z^3(2z^2+3z+9)^3(8z^4+28z^3+126z^2+189z+378)$$

$$(b) A_7 \wr C_2$$

**1B. Background and preliminaries.** We begin by providing a terse exploration of the object known as a *dessin*. For more detailed and comprehensive literature on the subject, see [Shabat and Zvonkin 1994; Wood 2006]. For the purposes of this paper, we begin with the observation that dessins may be realized by meromorphic functions known as *Belyi maps*. The arithmetic dynamics of these Belyi maps have been studied in some cases [Anderson et al. 2018].

**Definition 1.2.** Let  $X$  be a compact Riemann surface. A *Belyi map* is a meromorphic function  $F : X \rightarrow \mathbb{P}^1(\mathbb{C})$  that is unramified outside of  $\{0, 1, \infty\}$ . That is, all critical values of  $F$  are contained in  $\{0, 1, \infty\}$ . Here we may consider  $\mathbb{P}^1(\mathbb{C})$  as just  $\mathbb{C} \cup \{\infty\}$ .

Grothendieck's notion of a *dessin d'enfant* or *dessin* for short is a way to combinatorially characterize Belyi maps. If  $F$  is a Belyi map, then  $F^{-1}([0, 1])$ , that is, the preimage of the interval  $[0, 1]$ , has the structure of a bicolored connected graph embedded in  $X$ . The basic structure of the bicolored graph  $\Delta_F$  associated with a Belyi map  $F$  is given when we identify  $F^{-1}(0)$  as the set of black vertices,  $F^{-1}(1)$  as the set of white vertices,  $F^{-1}((0, 1))$  as the set of edges and  $F^{-1}(\mathbb{P}^1(\mathbb{C}) - [0, 1])$  as the set of faces. Note that the degrees of the black and white vertices of  $\Delta_F$  correspond to the multiplicities of the roots of  $F$  and  $F - 1$ , respectively. Furthermore, the dessin  $\Delta_F$  recovered from a Belyi map  $F$  is planar if and only if  $F$  is defined on  $\mathbb{P}^1(\mathbb{C})$ , while  $\Delta_F$  is a tree if and only if  $F$  is a polynomial. Throughout this paper, we assume  $X = \mathbb{P}^1(\mathbb{C})$ .

These structure of  $\Delta_F$  can be captured by the notion of a *dessin*, the relatively simple combinatorial characterization given by Grothendieck.

**Definition 1.3.** A *dessin d'enfant* or *dessin* is a connected bicolored graph equipped with a cyclic ordering of the edges (oriented counterclockwise) around each vertex.

Given a Belyi map  $F$ , it is not difficult to use the procedure described above to visualize the dessin  $\Delta_F$  to which  $F$  corresponds. However, recovering a Belyi map from a given dessin is a much more difficult proposition. Given a dessin  $\Delta_F$ , a corresponding Belyi map  $F$  can be determined (uniquely up to isomorphism over  $\mathbb{C} \cup \{\infty\}$ ) by considering the degrees of the vertices of  $\Delta_F$  and the resulting system of polynomial equations involving roots and poles of  $F$ . Various methods of calculating Belyi maps may be found in [Couveignes 1994; Matiyasevich 1996; Schneps 1994; Sijtsling and Voight 2014].

**Definition 1.4.** A *Shabat polynomial* is a polynomial  $F : \mathbb{C} \rightarrow \mathbb{C}$  whose critical values are contained in  $\{0, 1\}$ .

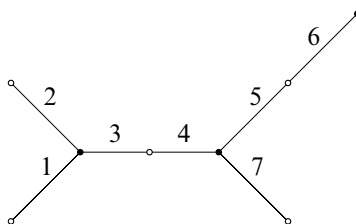
That is, a Shabat polynomial is a Belyi map which has only one pole (which is at infinity); hence, its corresponding dessin will be a tree. (Shabat polynomials can be defined more broadly as in [Shabat and Zvonkin 1994] as generalized Chebyshev polynomials which have at most two critical values. Without loss of generality, we choose in this paper to identify the two critical values 0 and 1.)

**Definition 1.5.** We say that two Shabat polynomials  $F, G$  are isomorphic if there exist  $\alpha \in \mathbb{C}^\times$  and  $\beta \in \mathbb{C}$  such that  $F(z) = G(\alpha z + \beta)$ .

Assume we have a dessin which is a tree and we label the edges with the numbers  $1, 2, \dots, n$ . We can associate the dessin with a pair of permutations  $\sigma_0, \sigma_1 \in S_n$ , where  $n$  is number of edges, such that the cycles of  $\sigma_0$  correspond to the cyclic ordering (read counterclockwise) of the edges around the black vertices and the cycles of  $\sigma_1$  correspond to the ordering (read counterclockwise) of the edges around the white vertices. For example, see Figure 1, where we have a bicolored tree, whose edges are labeled  $1, 2, \dots, 7$  inducing a pair of permutations  $\sigma_0, \sigma_1 \in S_7$  associated with the black and white vertices, respectively. In general, by  $\sigma_0$  (respectively,  $\sigma_1$ ), we mean the product of the cycle permutations associated with the edges about all of the black (respectively, white) vertices. The group that  $\sigma_0$  and  $\sigma_1$  generate is a central focus of this paper.

**Definition 1.6.** The *monodromy group* of a dessin with  $n$  edges is  $\langle \sigma_0, \sigma_1, \sigma_\infty \rangle$ , the group generated by  $\sigma_0, \sigma_1, \sigma_\infty \in S_n$ , where  $\sigma_0, \sigma_1$  are as described in the preceding paragraph and  $\sigma_\infty$  is such that  $\sigma_0 \sigma_1 \sigma_\infty = 1$ .

We remark that since  $\sigma_\infty = (\sigma_0 \sigma_1)^{-1}$ , we may remove it from the generating set for the monodromy group, but we keep it in the definition to be consistent with the wider literature on this subject, which goes well beyond the consideration of Shabat polynomials. For the remainder of the paper, when we refer to the generators of the monodromy group, we are talking about  $\sigma_0$  and  $\sigma_1$ . When a dessin is connected,



**Figure 1.** A dessin determined by the pair of permutations  $\sigma_0 = (1, 3, 2)(4, 7, 5)$  and  $\sigma_1 = (3, 4)(5, 6)$  whose monodromy group  $\langle \sigma_0, \sigma_1 \rangle$  is isomorphic to  $\text{GL}_3(\mathbb{F}_2)$ , a transitive subgroup of  $S_7$ .

its monodromy group will be a transitive subgroup of  $S_n$ , where  $n$  is the number of edges in the dessin.

To every dessin, we may associate an invariant known as its *ramification type*. The ramification type of a dessin with  $n$  edges is given by the three partitions of  $n$  corresponding to the degrees of the black vertices, the degrees of the white vertices and the degrees of the faces. In the case of a dessin having one face, the latter partition is simply  $n = n$ . Since we focus exclusively on dessins with one face in this paper, we will omit from the notation for ramification type the last partition corresponding to the degrees of the faces.

**Definition 1.7.** The *ramification type* of a dessin with  $n$  edges (and exactly one face) consists of the two partitions of  $n$

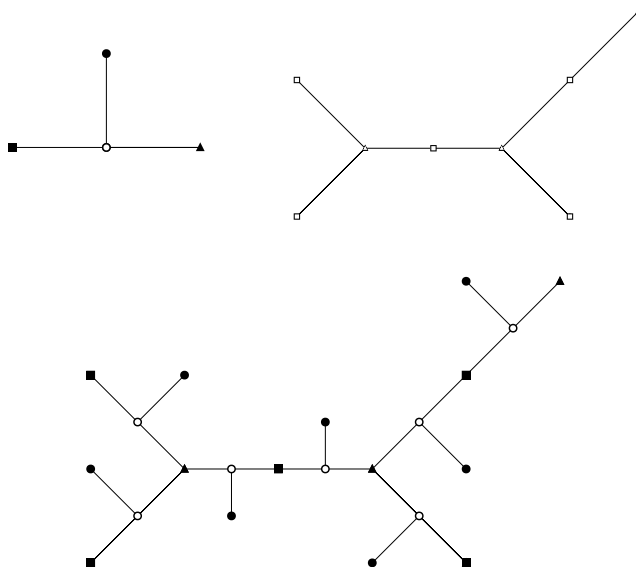
$$[b_1^{\beta_1} b_2^{\beta_2} \dots b_k^{\beta_k}; w_1^{\alpha_1} w_2^{\alpha_2} \dots w_\ell^{\alpha_\ell}]$$

written in exponential notation, where  $b_1, b_2, \dots, b_k$  are the distinct degrees of the black vertices,  $w_1, w_2, \dots, w_\ell$  are the distinct degrees of the white vertices,  $\beta_i$  is the number black vertices of degree  $b_i$  and  $\alpha_i$  is the number white vertices of degree  $w_i$ .

Note that  $b_1^{\beta_1} b_2^{\beta_2} \dots b_k^{\beta_k}$  and  $w_1^{\alpha_1} w_2^{\alpha_2} \dots w_\ell^{\alpha_\ell}$  are both partitions of  $n$ , where  $n$  is the number of edges, and these two partitions correspond to the cycle type of  $\sigma_0$  and  $\sigma_1$ , respectively.

While each dessin has a unique ramification type, one may ask how many distinct dessins (or equivalently nonisomorphic Shabat polynomials) are associated with a given ramification type. Our focus in this paper will be narrowed to ramification types which admit unique dessins.

We sometimes use the concept of tree composition to decompose a dessin into smaller dessins. Composition will also help us compute new Shabat polynomials as it corresponds with the usual polynomial composition. It is an easy exercise in calculus to show that the composition of two Shabat polynomials is again a Shabat polynomial.



**Figure 2.** Top, left:  $P$ , with two vertices marked square  $\square$  and triangle  $\triangle$ . Top, right:  $Q$ , with black vertices marked  $\square$ , white vertices marked  $\triangle$ . Bottom: The composition  $P \circ Q$  of two dessins  $P$ ,  $Q$ .

Many of the dessins that we study can be constructed by a composition process given by Adrianov and Zvonkin [1998]. Given two dessins,  $P$  and  $Q$ , we begin the composition  $P \circ Q$  by first distinguishing two vertices of  $P$ —label them with a square and a triangle. The vertices of  $Q$  will be preimages of the square and triangle, so we mark every black vertex of  $Q$  with a square and similarly every white vertex of  $Q$  with a triangle. The process of composition is as follows:

- (1) Replace each edge of  $Q$  with the union of the path from the square to the triangle in  $P$  along with every branch connected to that path.
- (2) Adjoin to each square (resp., triangle) vertex of  $Q$  the union of every branch connected to the square (triangle) in  $P$  except for the one in the path to the triangle (square). Do this as many times as the degree of the vertex.

The resulting graph should resemble  $n$  copies of  $P$  arranged in the shape of  $Q$ , where  $n$  is the number of edges of  $Q$ . We demonstrate this process in Figure 2.

**Remark 1.8.** Let  $G_P$ ,  $G_Q$  denote the respective monodromy groups of  $P$  and  $Q$ . According to a theorem of Adrianov and Zvonkin [1998], the monodromy group of  $P \circ Q$  is a subgroup of  $G_Q \wr G_P$ , where  $\wr$  denotes the wreath product.

This process also gives a way to compute Shabat polynomials. If  $p, q$  are the respective Shabat polynomials of  $P, Q$  such that  $p(0), p(1) \in \{0, 1\}$  then the Shabat polynomial of  $P \circ Q$  is  $p \circ q$  (where  $\circ$  denotes the conventional composition

of functions, i.e.,  $(f \circ g)(x) = f(g(x))$ ). Later on, when we compute Shabat polynomials of more complicated dessins, we will make extensive use of this fact.

We will often call upon the idea of the wreath product of groups to describe our monodromy groups. The composition process produces dessins whose monodromy groups are subgroups of wreath products. While there are numerous examples for which the containment is proper, often equality of the groups is achieved. As far as the present authors can tell, the exact conditions that ensure equality are not known.

**Definition 1.9.** Let  $d$  be a positive integer. Let  $G \leq S_d$  and  $H$  be groups. Let  $K$  be the direct product of  $d$  copies of  $H$ . If  $h = (h_1, \dots, h_d) \in K$ , then we define the action of  $\sigma \in G$  on  $K$  by  $\sigma \cdot h = (h_{\sigma(1)}, \dots, h_{\sigma(d)})$ . The *wreath product* of  $H$  by  $G$  is the semidirect product  $K \rtimes G$  with respect to the action above, and we denote this group by  $H \wr G$ .

In this paper,  $G$  is typically  $C_d$ , the cyclic group of order  $d$ .

## 2. Shabat polynomials for trees uniquely determined by ramification type

In this section, we summarize the list of Shabat polynomials (up to isomorphism) corresponding to dessins which are trees and are uniquely determined by ramification type. The complete list of ramification types for such dessins was given in [Shabat and Zvonkin 1994]. For the Shabat polynomials corresponding to these ramification types, we adopt the convention described in Definition 1.7.

**Proposition 2.1.** *The ramification types  $[r; 1^r]$ ,  $[2^r, 1; 2^r, 1]$ ,  $[2^r; 2^{r-1}, 1^2]$  have respective Shabat polynomials*

$$z^r, \quad \frac{1}{2}(1 + \cos((2r+1) \arccos(z))), \quad \frac{1}{2}(1 + \cos((2r) \arccos(z))),$$

*all unique up to isomorphism.*

This result is already well known in the literature and can be found on pages 3–4 of [Shabat and Zvonkin 1994]. See Figure 3.

**Proposition 2.2** [Adrianov 2007]. *Up to isomorphism, the unique Shabat polynomial for the ramification type  $[s^{r-1}, t; r, 1^{(r-1)(s-1)+(t-1)}]$  is*

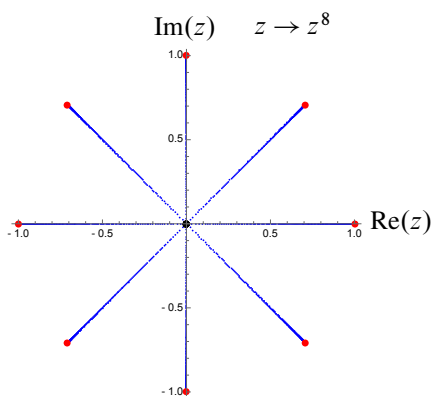
$$F(z) = (1-z)^t \left( \sum_{k=0}^{r-1} \binom{t}{s}_k \frac{z^k}{k!} \right)^s,$$

*where*

$$(a)_k = a(a+1)(a+2) \cdots (a+k-1)$$

*denotes the Pochhammer symbol.*

The proof for this proposition can be found in [Adrianov 2007].



**Figure 3.** The dessin with ramification type  $[8; 1^8]$ .

**Proposition 2.3.** *Let  $r > 1$ . Up to isomorphism, the Shabat polynomial for the tree having ramification type*

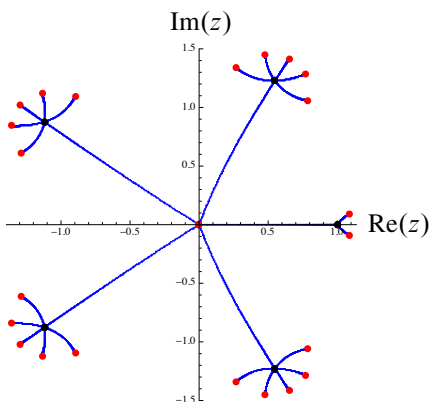
$$[r, t, 1^{r+t-2}; 2^{r+t-1}]$$

*with a black vertex of degree  $r$  located at  $z = 0$  and a black vertex of degree  $t$  located at  $z = 1$  is given by*

$$F(z) = 4z^r \binom{r+t-1}{r} {}_2F_1(t-1, r; r+1; z) \\ \times \left( 1 - (1-z)^t z^r \binom{r+t-1}{t-1} {}_2F_1(1, r+t; r+1; z) \right),$$

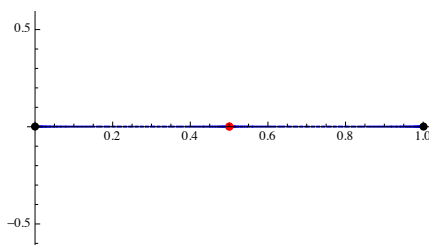
*where  ${}_2F_1$  is the hypergeometric function defined by*

$${}_2F_1(a, b; c; z) = \sum_{n=0}^{\infty} \frac{(a)_n (b)_n}{(c)_n} \frac{z^n}{n!}.$$

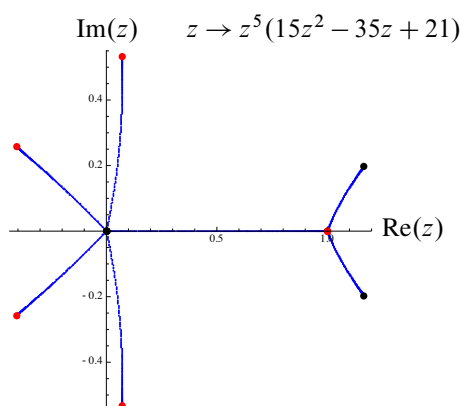


**Figure 4.** The dessin obtained by the Shabat polynomial given in Proposition 2.2 when  $s = 6$ ,  $r = 5$ ,  $t = 3$ .





**Figure 5.** The dessin (path graph) obtained by the Shabat polynomial  $\beta(z) = 4z(1 - z)$ .



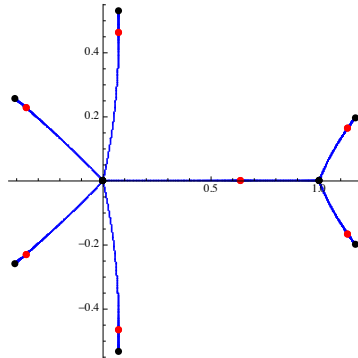
**Figure 6.** The tree obtained by the Shabat polynomial in Proposition 2.2 where  $s = 1$ ,  $r = 3$ ,  $t = 5$ .

*Proof.* Let  $S_{r,t}(z)$  be the Shabat polynomial for the ramification type  $[t, 1^{r-1}; r, 1^{t-1}]$ . By Proposition 2.2, with  $s = 1$ ,

$$S_{r,t}(z) = (1 - z)^t \sum_{j=0}^{r-1} \binom{t-1+j}{t-1} z^j.$$

Consider the map  $\beta(z) = 4z(1 - z)$  with the dessin  $\Delta_\beta$  (see Figure 5) and  $S_{r,t}(z)$  with the dessin  $\Delta_S$  (see Figure 6). The composition  $\beta(z) \circ S_{r,t}(z)$  is a Shabat polynomial that produces the dessin obtained by coloring the vertices of  $\Delta_S$  to black and adding a white vertex of degree 2 inside every edge (in other words, replacing every edge of  $\Delta_S$  with  $\Delta_\beta$ ). Note the number of edges in  $S_{r,t}(z)$  is  $r + t - 1$ . The composition produces the new dessin  $\Delta_F$  (see Figure 7) and Shabat polynomial  $F(z) = \beta(z) \circ S_{r,t}(z)$  with ramification type  $[r, t, 1^{r+t-2}; 2^{r+t-1}]$ , and therefore  $F(z)$  equals

$$4z^r(1 - z)^t \left( \sum_{j=0}^{r-1} \binom{t-1+j}{t-1} z^j \right) \left( \sum_{j=0}^{t-1} \binom{r-1+j}{r-1} \binom{r+t-1}{r+j} (-1)^j z^j \right),$$



**Figure 7.** The tree obtained by the Shabat polynomial in Proposition 2.3 with  $r = 5, t = 3$ .

which can be rewritten in terms of hypergeometric functions, as in the statement of the present proposition.  $\square$

**Proposition 2.4.** *The Shabat polynomial for the unique tree having ramification type*

$$[r^2, 1^{4r-3}; 3^{2r-1}]$$

*with two black vertices of degree  $r$  located at  $z = 0$  and  $z = 1$  is given by*

$$F(z) = (T \circ S_r)(z),$$

*where*

$$T(z) = -3i\sqrt{3}z(1-z)(z+\rho), \quad \rho = \frac{1}{2}(-1 + i\sqrt{3}),$$

*and*

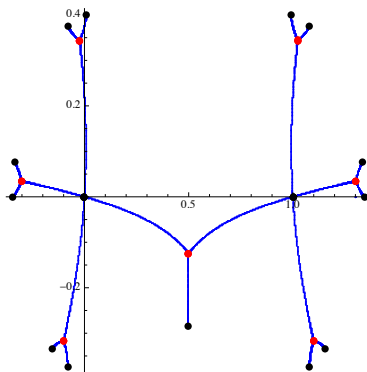
$$S_r(z) = (1-z)^r \sum_{j=0}^{r-1} \binom{r-1+j}{r-1} z^j.$$

*$F(z)$  is unique up to isomorphism.*

*Proof.* First we will show that  $T(z) := -3i\sqrt{3}z(1-z)(z+\rho)$  corresponds to a 3-star with a white center and black leaves at  $z = 0$  and  $z = 1$ . Considering  $T(z)$ , we see immediately three distinct roots of multiplicity 1 at  $z = 0, 1, \frac{1}{2}(1 - i\sqrt{3})$  representing three black leaves in  $\Delta_F$ . Next we consider the derivative of  $T(z)$ ,

$$T'(z) = -3i\sqrt{3}(\rho + 2(1-\rho)z - 3z^2),$$

which has a single root of multiplicity 2 (note that the discriminant of  $T'(z)$  is zero). Since the multiplicity of the black vertices is 1, we may assume that the multiple root in  $T'(s)$  must refer to a root of multiplicity 3 in  $F(z) - 1$ , representing the white vertex of degree 3. Therefore,  $T(z)$  must be a 3-star with black



**Figure 8.** An illustration of the tree derived from the Shabat polynomial in Proposition 2.4 where  $r = 4$ .

leaves at  $z = 1$  and  $z = 0$ . We can now use the idea of composition to replace every edge of the tree having Shabat polynomial  $S_r(z) := S_{r,r}(z)$ , where  $S_{r,t}(z)$  is the polynomial as defined in the proof of Proposition 2.3, with the 3-star by computing the composition  $(T \circ S_r)(z)$ . This will add a white vertex of degree 3 and an additional black leaf for every edge. Note that  $S_r(z)$  corresponds to a tree with  $2r - 1$  edges and  $4r - 2$  vertices. Therefore  $\Delta_F$  will have  $2r - 1$  white vertices of degree 3 and  $4r - 3$  black leaves, in addition to the two black vertices of degree  $r$ .

Note: An anonymous referee pointed out that we may go one step further here by letting  $z' := i\sqrt{3}z - \rho^2$ . A quick computation shows that  $\overline{S_r(z')} = S_r(1 - z')$ . One can also show that  $S_r(z) = 1 - S_r(1 - z)$  using the following argument. Observe that 0 is a root of order  $r$  of  $S_r(z)$  and  $1 - S_r(1 - z)$ . Further observe that 1 is a root of order  $r$  of  $S_r(z) - 1$  and  $1 - S_r(1 - z) - 1$ . Thus we deduce that  $S_r(z) = 1 - S_r(1 - z)$  using the uniqueness of the Shabat polynomial from Proposition 2.2. Hence,  $\overline{S_r(z')} = 1 - S_r(z')$ . A few simple calculations yield the equality  $\overline{T(S_r(z'))} = T(S_r(z'))$ , which implies  $T(S_r(z')) \in \mathbb{Q}[z]$ .  $\square$

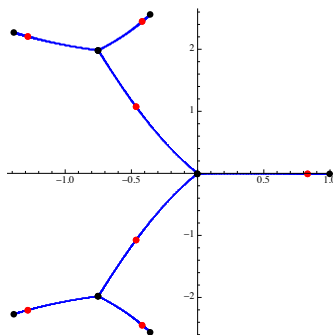
**Proposition 2.5.** *For the tree with ramification type  $[3^3, 1^5; 2^7]$ , a black vertex of degree 3 at  $z = 0$  and a black vertex of degree 1 at  $z = 1$ , the Shabat polynomial is*

$$F(z) = -\frac{4}{531441}(z-1)z^3(2z^2+3z+9)^3(8z^4+28z^3+126z^2+189z+378).$$

*Proof.* We can write  $F(z) = (\beta \circ f)(z)$ , where

$$\beta(z) = 4z(1-z) \quad \text{and} \quad f(z) = -\frac{1}{729}(z-1)(9+3z+2z^2)^3,$$

which is the Shabat polynomial for ramification type  $[3^2, 1; 3, 1^4]$  obtained by letting  $r = 3$ ,  $s = 3$ ,  $t = 1$  in Proposition 2.2.  $\square$



**Figure 9.** An illustration of the tree described in Proposition 2.5.

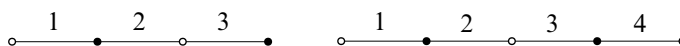
### 3. Monodromy groups for trees uniquely determined by ramification type

In this section, we provide proofs for the monodromy groups associated with each ramification type listed in Theorem 1.1. In all of our proofs, we proceed by choosing a particular labeling of the edges of the dessin. Though the monodromy group does not depend on the choice of labels, some choices better illustrate how  $\sigma_0$  and  $\sigma_1$  generate the monodromy group.

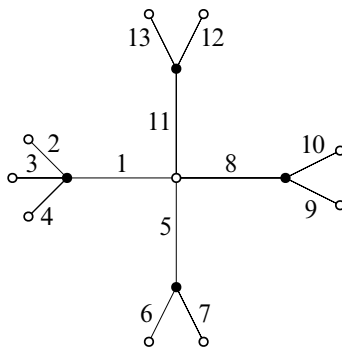
**Proposition 3.1.** *The ramification types  $[r; 1^r]$ ,  $[2^r, 1; 2^r, 1]$ , and  $[2^r; 2^{r-1}, 1^2]$  have respective monodromy groups  $C_r$ ,  $D_{2(2r+1)}$ , and  $D_{2(2r)}$ , where  $D_m$  denotes the dihedral group of order  $m$ .*

*Proof.* The first ramification type gives the  $r$ -star dessin with monodromy group generated by an  $r$ -cycle and the identity permutation. It follows that the monodromy group is the cyclic group  $C_r$ . The second and third ramification types yield the path dessins with  $2r + 1$  and  $2r$  edges respectively. We handle these two cases simultaneously, since the argument is essentially the same. The dessins in Figure 10 are examples of path dessins.

In both cases, the generators of the groups  $\sigma_0$  and  $\sigma_1$  have order 2, and the respective  $\sigma_\infty$ 's have order  $2r + 1$  and  $2r$ . Since in this case  $\sigma_\infty = (\sigma_0\sigma_1)^{-1} = \sigma_1\sigma_0$ , we may view the monodromy group as  $\langle \sigma_0, \sigma_\infty \rangle$ . We let  $n$  denote the order of  $\sigma_\infty$ ; note that  $n$  is either  $2r + 1$  or  $2r$  depending on the ramification type. The relations  $\sigma_0^2 = \sigma_\infty^r = 1$  and  $\sigma_0\sigma_\infty = (\sigma_0\sigma_1)\sigma_0 = (\sigma_1\sigma_0)^{-1}\sigma_0 = (\sigma_\infty)^{-1}\sigma_0$  hold. The conclusion is that the monodromy groups of these dessins are isomorphic to the dihedral groups of order  $2n$ .  $\square$



**Figure 10.** The path dessins of 3 and 4 edges, respectively.



**Figure 11.** An example of a dessin from Proposition 3.2, where  $r = 4$ ,  $s = 3$ ,  $t = 4$ .

**Proposition 3.2.** Assume  $r > 1$ . The ramification type  $[s^{r-1}, t; r, 1^{(r-1)(s-1)+(t-1)}]$  has  $n = (r-1)s + t$  edges and a unique tree with monodromy group  $G$ , with

$$G \cong \begin{cases} C_r \wr C_s, & s = t, \\ S_{n/d} \wr C_d, & s \neq t, r \text{ even}, \\ A_{n/d} \wr C_d, & s \neq t, r \text{ is odd and } \frac{t}{d} \text{ is odd}, \\ (A_{n/d})^d \rtimes C_{2d}, & s \neq t, r \text{ odd}, \frac{t}{d} \text{ even}, \end{cases}$$

where  $d = \gcd(s, t)$ .

*Proof.* The ramification type  $[s^{r-1}, t; r, 1^{(r-1)(s-1)+(t-1)}]$  produces a tree of diameter 4 with  $n = (r-1)s + t$  edges in the nondegenerate cases. See Figure 11.

In general,  $\sigma_0$  is the product of one  $t$ -cycle and  $(r-1)$ -many  $s$ -cycles and  $\sigma_1$  is an  $r$ -cycle. We label our edges so that we compute the permutations  $\sigma_0, \sigma_1, \sigma_\infty$  as

$$\sigma_0 = (1, \dots, t)(t+1, \dots, t+s) \cdots (t+(r-2)s+1, \dots, t+(r-1)s),$$

$$\sigma_1 = (1, t+1, t+s+1, t+2s+1, \dots, t+(r-2)s+1),$$

$$\sigma_\infty^{-1} = \sigma_0 \sigma_1 = (1, 2, \dots, n).$$

(Note that we go left to right when computing permutation products.)

Case 1:  $s = t \implies G = C_r \wr C_s$ . Assume  $s = t$ . Then our dessin is the composition of an  $s$ -star with an  $r$ -star, which means  $G$  is a subgroup of  $C_r \wr C_t$  by Remark 1.8. Define  $\tau_i := \sigma_0^{-i} \sigma_1 \sigma_0^i$ . Referring to the above where we already computed  $\sigma_0$  and  $\sigma_1$ , we see

$$\begin{aligned} \tau_0 &= (1, t+1, 2t+1, \dots, (r-1)t+1) = \sigma_1, \\ \tau_1 &= (2, t+2, 2t+2, \dots, (r-1)t+2), \\ &\vdots \\ \tau_{t-1} &= (t, 2t, 3t, \dots, rt). \end{aligned}$$

Each  $\tau_i$  is an  $r$ -cycle and generates  $C_r$ . Since the  $\tau_i$ 's partition  $\{1, 2, \dots, rt\}$ , they must commute with each other and we see that together they generate  $C_r^t$ . Also,  $\sigma_0$  is a product of  $t$ -cycles satisfying  $\sigma_0^{-1} \tau_i \sigma_0 = \tau_{i+1}$ , where the subscripts are reduced modulo  $t$ . These relations are sufficient to recognize that  $G$  contains  $\langle \sigma_0, \tau_1, \tau_2, \dots, \tau_{t-1} \rangle \cong C_r \wr C_t$ .

Case 2:  $s \neq t$ ,  $\gcd(s, t) = 1 \implies G = A_n$  for  $r, t$  odd and  $G = S_n$  otherwise. Assume that  $\gcd(s, t) = 1$ , with  $s$  or  $t > 1$ . It is known that a permutation group containing  $(1, 2, 3)$  and  $(1, 2, \dots, n)$  contains  $A_n$ ; see Lemma A.1. Our goal is to show that  $A_n \leq G \leq S_n$  and then use a parity argument to determine which containment is improper. Given that  $\sigma_0 \sigma_1 = (1, 2, \dots, n) \in G$ , we proceed to show  $(1, 2, 3) \in G$ .

Assume  $t = 1$  and  $s > 1$ . We claim  $\rho := (\sigma_0^{-1} \sigma_1^{-1} \sigma_0)(\sigma_\infty \sigma_1 \sigma_\infty^{-1}) = (1, 2, 3)$ . Since  $t = 1$ , we know  $\sigma_0$  is a product of  $(r - 1)$   $s$ -cycles, while  $\sigma_1, \sigma_1^{-1}$  remain  $r$ -cycles. We see that

$$\begin{aligned} \rho &= (\sigma_0^{-1} \sigma_1^{-1} \sigma_0)(\sigma_\infty \sigma_1 \sigma_\infty^{-1}) \\ &= (1, (r-2)s+3, \dots, 2s+3, s+3, 3)(2, 3, s+3, 2s+3, \dots, (r-2)s+3). \end{aligned}$$

One may verify that  $\rho(1) = 2$ ,  $\rho(2) = 3$ ,  $\rho(3) = 1$  and, for  $k > 3$ ,  $\rho(k) = k$ . It follows that  $A_n \leq G$ .

If  $t = 2$ , we have  $\sigma_0^s = (1, 2) \in G$ . Since  $G$  contains the transposition  $(1, 2)$  and the cycle  $(1, 2, \dots, n)$ , we have  $S_n \leq G$ .

Now suppose  $t \geq 3$ , we first set  $k$  to be the smallest positive integer such that  $k \equiv 0 \pmod{s}$  and  $k \equiv -1 \pmod{t}$ . The existence of such a number is guaranteed by the Chinese remainder theorem. We claim  $\rho := (\sigma_1^{-1} \sigma_0^k \sigma_1) \sigma_0^k (\sigma_1^{-1} \sigma_0^{-2k} \sigma_1) = (1, 2, 3)$ . Notice that

$$(\sigma_1^{-1} \sigma_0^k \sigma_1) \sigma_0^k (\sigma_1^{-1} \sigma_0^{-2k} \sigma_1) = (t+1, t, \dots, 3, 2)(1, t, \dots, 3, 2)(t+1, 2, 3, \dots, t)^2.$$

One may verify that  $\rho(1) = 2$ ,  $\rho(2) = 3$ ,  $\rho(3) = 1$  and  $\rho(k) = k$  for  $k > 3$ . Thus  $\rho = (1, 2, 3) \in G$  and therefore  $A_n \subseteq G$ .

For every triple  $s, t$  such that  $\gcd(s, t) = 1$  and  $s$  or  $t > 1$ , we have shown that  $A_n \subseteq G$ . Since we also have  $G \leq S_n$ , by index considerations  $G$  is either the symmetric or alternating group of appropriate order. Otherwise if  $r$  or  $t$  is even,  $\sigma_0$ , being the product of a  $t$ -cycle and  $(r-1)$   $s$ -cycles, is an odd permutation (note  $s$  must be odd if  $t$  is even), so  $G \cong S_n$ . Since both  $\sigma_0$  and  $\sigma_1$  are even permutations when  $r$  and  $t$  are odd, we deduce that  $G \leq A_n$  and thus the double inclusion gives us  $G \cong A_n$ .

Case 3: In this final case, we assume  $\gcd(s, t) = d > 1$ . This tree is the composition  $P \circ Q$ , where  $P$  is the  $d$ -star and  $Q$  is the dessin corresponding to the passport

$$\left[ \left( \frac{s}{d} \right)^{r-1}, \frac{t}{d}; r, 1^{(r-1)(s/d-1)+(t/d-1)} \right].$$

Hence, the monodromy group  $G$  is a subgroup of the wreath product  $G_Q \wr C_d$ , where  $G_Q$  is the monodromy group for  $Q$ .

Consider the partition of  $\{1, \dots, n\}$  into the  $d$  sets

$$\{1, d+1, \dots, n-d+1\}, \quad \{2, d+2, \dots, n-d+2\}, \quad \dots, \quad \{d, 2d, \dots, n\},$$

each of size  $\frac{n}{d}$ , and denote them by  $P_1, \dots, P_d$  respectively. Recall that  $\sigma_0$  is the disjoint product of a  $t$ -cycle and  $(r-1)$   $s$ -cycles, and moreover every element in  $\{1, 2, \dots, n\}$  is moved by exactly one of these cycles under the canonical group action. Because  $d$  divides both  $s$  and  $t$ , we know  $\tau := \sigma_0^d$  is the disjoint product of  $d$   $\frac{t}{d}$ -cycles and  $d(r-1)$   $\frac{s}{d}$ -cycles. Moreover, each disjoint cycle of  $\tau$  permutes elements in exactly one of the  $P_i$  while fixing the rest. Similarly, because  $d$  divides  $n$ , we know  $\sigma_\infty^d$  is the disjoint product of  $d$   $\frac{n}{d}$ -cycles, and each disjoint cycle of  $\sigma_\infty^d$  likewise permutes elements in exactly one of the  $P_i$ . Note that  $\sigma_1$  permutes only the elements of  $P_1$ .

Let  $k$  be the smallest positive integer such that  $k$  satisfies  $k \equiv 0 \pmod{\frac{s}{d}}$  and  $k \equiv -1 \pmod{\frac{t}{d}}$ . One may verify that  $\rho := \sigma_1^{-1} \tau^k \sigma_1 \tau^k \sigma_1^{-1} \tau^{-2k} \sigma_1 = (1, d+1, 2d+1)$ . (Note that in the case where  $t = d$ , we let  $\rho := (\tau^{-1} \sigma_1^{-1} \tau)(\sigma_\infty^d \sigma_1 \sigma_\infty^{-d})$  and proceed with the same argument.)

We can conclude that the subgroup

$$N = \langle \rho, \sigma_\infty^{-d} \rho \sigma_\infty^d, \sigma_\infty^{-2d} \rho \sigma_\infty^{2d}, \dots, \sigma_\infty^{-(n-d)} \rho \sigma_\infty^{n-d}, \sigma_1 \rangle$$

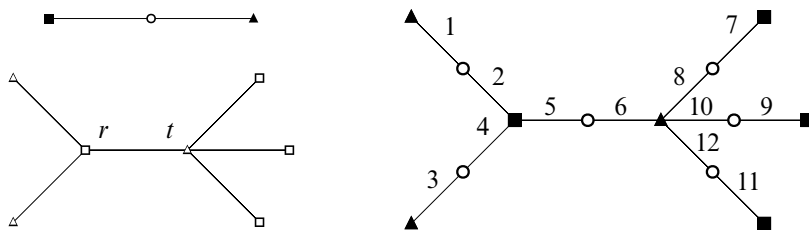
is isomorphic to  $S_{n/d}$  when  $r$  is even and isomorphic to  $A_{n/d}$  when  $r$  is odd (see Lemma A.4). Furthermore, we observe that  $N, \sigma_\infty^{-1} N \sigma_\infty, \sigma_\infty^{-2} N \sigma_\infty^2, \dots, \sigma_\infty^{-d+1} N \sigma_\infty^{d-1}$  are all isomorphic to  $S_{n/d}$  or  $A_{n/d}$  (depending on whether  $r$  is even or odd) and  $\sigma_\infty^{-i+1} N \sigma_\infty^{i+1}$  permutes elements of  $P_i$ . Hence

$$H := \langle N, \sigma_\infty^{-1} N \sigma_\infty, \sigma_\infty^{-2} N \sigma_\infty^2, \dots, \sigma_\infty^{-d+1} N \sigma_\infty^{d-1} \rangle \cong \begin{cases} (S_{n/d})^d & \text{if } r \text{ even,} \\ (A_{n/d})^d & \text{if } r \text{ odd.} \end{cases}$$

One can check that  $\sigma_0^{-1} H \sigma_0 = H$  and  $\sigma_1^{-1} H \sigma_1 = H$ . Hence,  $H \triangleleft G$ . Observe that  $\sigma_1 \in H$ . Therefore,  $H \sigma_\infty$  generates the quotient group  $G \setminus H$ . When  $r$  is even, the smallest power of  $\sigma_\infty$  in  $H$  is  $d$ , when  $r$  is odd and  $\frac{n}{d}$  is odd, the smallest power of  $\sigma_\infty$  in  $H$  is  $d$ , and when  $r$  is odd and  $\frac{n}{d}$  is even, the smallest power of  $\sigma_\infty$  in  $H$  is actually  $2d$ . (Note that when  $r$  is odd, the parities of  $\frac{t}{d}$  and  $\frac{n}{d}$  are the same.) In order to show that  $G$  is isomorphic to a semidirect product, we will use the splitting lemma. In our case, if we can find an element of  $H \sigma_\infty$  of order  $d$  or  $2d$  (depending on the case), we have shown  $G$  is a semidirect product.

First we consider the case where  $r$  is even. In this case, observe that

$$\begin{aligned} & (d, 2d, 3d, \dots, n)^{-1} (1, 2, 3, 4, \dots, n) \\ &= (1, 2, \dots, d)(d+1, d+2, \dots, 2d) \cdots (n-d+1, n-d+2, \dots, n). \end{aligned}$$



**Figure 12.**  $P$  and  $Q$  on the left;  $P \circ Q$  on the right.

Hence, there is an element of order  $d$  in  $G \setminus H$  and  $G \cong (S_{n/d})^d \rtimes C_d$  by the splitting lemma for semidirect products, and in fact  $G \cong S_{n/d} \wr C_d$ . This also shows that  $H\sigma_\infty$  contains an element of order  $d$  in the case where  $r$  and  $\frac{n}{d}$  are odd since the cycle  $(d, 2d, 3d, \dots, n)$  is an element of  $A_n$  in this case. Therefore, when  $\frac{n}{d}$  and  $r$  are odd,  $G \cong (A_{n/d})^d \rtimes C_d$  and in fact  $G \cong A_{n/d} \wr C_d$ .

Now we consider the case where  $r$  is odd and  $\frac{n}{d}$  is even. Observe that

$$\begin{aligned} & (2d, 3d, \dots, n)^{-1}(1, 2, 3, \dots, n) \\ &= (1, 2, 3, \dots, 2d)(2d+1, 2d+2, \dots, 3d) \cdots (n-d+1, n-d+2, \dots, n). \end{aligned}$$

Hence, there is an element of order  $2d$  in  $G \setminus H$  and thus  $G \cong (A_{n/d})^d \rtimes C_{2d}$ .  $\square$

**Proposition 3.3.** *Let  $r, t > 1$ . The ramification type  $[r, t, 1^{r+t-2}; 2^{r+t-1}]$  produces a unique tree with monodromy group  $G$ , where*

$$G \cong \begin{cases} A_{2r-1} \times C_2, & r = t, r \text{ odd}, \\ S_{2r-1} \times C_2, & r = t, r \text{ even}, \\ A_{r+t-1} \wr C_2, & r \neq t, \text{ both odd}, \\ R_2, & r \neq t, \text{ both even}, \\ S_{r+t-1} \wr C_2, & r \neq t, \text{ else}, \end{cases}$$

where  $R_2$  denotes the index-2 subgroup of  $S_{r+t-1} \wr C_2$  such that  $\tau_1 \tau_2$  is an even permutation for all  $(\tau_1, \tau_2, g) \in R_2$ .

*Proof.* First, we note that this dessin is the composition  $P \circ Q$ , where  $P$  is the 2-star and  $Q$  is the dessin of Proposition 3.2 with  $s = 1$ . See Figure 12.

Let  $G_Q = \langle (1, 2, \dots, r), (r, r+1, \dots, r+t-1) \rangle$  be the monodromy group of  $Q$ . By Proposition 3.2, we know that

$$G_Q \cong \begin{cases} A_{r+t-1}, & r, t \text{ both odd}, \\ S_{r+t-1}, & \text{otherwise}. \end{cases}$$

The dessin with ramification type  $[r, t, 1^{r+t-2}; 2^{2r-1}]$  is the composition of  $P$  and  $Q$ , and so its monodromy group  $G$  satisfies  $G \leq G_Q \wr C_2$  by Remark 1.8. We consider  $G$  in two cases:  $r = t$  and  $r \neq t$ .



Case 1:  $r \neq t$ . In the first case, we have  $r \neq t$ . We label our edges in such a way that

$$\begin{aligned}\sigma_0 &= (1, 2, \dots, r)(\bar{r}, \overline{r+1}, \dots, \overline{r+t-1}), \\ \sigma_1 &= (1, \bar{1})(2, \bar{2}) \cdots (r+t-1, \overline{r+t-1}).\end{aligned}$$

Note that  $\sigma_0$  is the disjoint product of an  $r$ -cycle with a  $t$ -cycle; call these cycles  $\pi_1$  and  $\pi_2$  respectively. Consider the embedding  $\phi : G \rightarrow S_{r+t-1} \wr C_2$  given by

$$\begin{aligned}\sigma_0 &\mapsto (\pi_1, \pi_2, 0), \\ \sigma_1 &\mapsto (\text{id}, \text{id}, 1).\end{aligned}$$

Note that  $\sigma_1^{-1}\sigma_0\sigma_1$  is mapped to  $(\pi_2, \pi_1, 0)$ . Apply Lemma A.5 to  $n = r+t-1$  (assume  $n \geq 5$  for now),  $\pi_1, \pi_2 \in S_{r+t-1}$ . We have  $G_Q = \langle \pi_1, \pi_2 \rangle \geq A_n$  as noted above. Lemma A.5 implies

$$A_{r+t-1} \wr C_2 \leq \phi(G) \leq S_{r+t-1} \wr C_2.$$

When  $r, t$  are odd, both  $\pi_1$  and  $\pi_2$  are even permutations, and we see that  $\phi(G) \cong A_{r+t-1} \wr C_2$ . When  $r$  and  $t$  have different parity, we know  $\langle \pi_1, \pi_2 \rangle \cong S_{r+t-1}$ , so  $\phi(G) \cong S_{r+t-1} \wr C_2$ . When  $r, t$  are both even, for any  $(\rho_1, \rho_2, g) \in \phi(G)$ ,  $\rho_1$  and  $\rho_2$  will share the same parity. Since we can take  $\rho_1 = \pi_1$ , an odd permutation, we see that  $\phi(G)$  is properly contained in between  $A_{r+t-1} \wr C_2$  and  $S_{r+t-1} \wr C_2$ . It is in fact the group  $R_2$  described earlier after Theorem 1.1. In the finite number of cases where  $r+t-1 < 5$ , one can verify the result by hand.

Case 2:  $r = t$ . In the second case, we consider  $r = t$ . We can label our dessin in such a way that

$$\begin{aligned}\sigma_0 &= (1, 2, \dots, r)(\bar{1}, \bar{2}, \dots, \bar{r}), \\ \sigma_1 &= (1, \overline{r+1})(2, \overline{r+2}) \cdots (r-1, \overline{2r-1})(r, \bar{r})(r+1, \bar{1}) \cdots (2r-1, \overline{r-1}).\end{aligned}$$

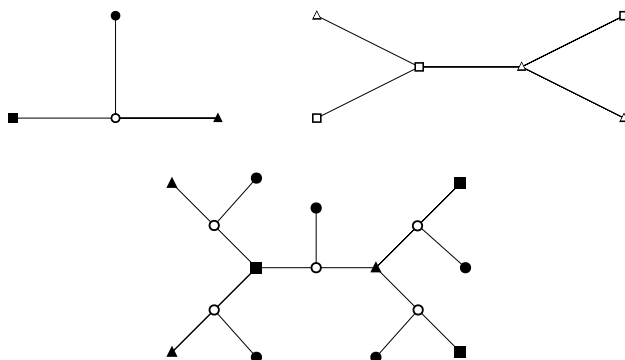
Observe that

$$\begin{aligned}\sigma_\infty^{(2r-1)} &= (1, \bar{1}) \cdots (2r-1, \overline{2r-1}), \\ \tau_1 &= \sigma_1 \sigma_0 \sigma_1^{-1} = (r, r+1, \dots, 2r-1)(\bar{r}, \overline{r+1}, \dots, \overline{2r-1}), \\ \tau_2 &= \sigma_\infty^{(2r-1)} \sigma_1 = (\bar{1}, \overline{r+1})(1, r+1) \\ &\quad \cdot (\bar{2}, \overline{r+2})(2, r+2) \cdots (\overline{r-1}, \overline{2r-1})(r-1, 2r-1)(r, \bar{r}),\end{aligned}$$

and  $G_Q = \langle \sigma_\infty^{(2r-1)}, \tau_1, \tau_2 \rangle$  is a subgroup of  $S_{2r-1} \times \mathbb{Z}_2$ . Furthermore,

$$\tau_3 = \tau_2 \sigma_1 \sigma_0 \sigma_1^{-1} \tau_2^{-1} = (1, 2, \dots, r)(\bar{1}, \bar{2}, \dots, \bar{r}).$$

By Proposition 3.2, we see that  $\langle \tau_1, \tau_3 \rangle$  is  $S_{2r-1}$  if  $r$  even and  $A_{2r-1}$  if  $r$  odd, and thus we have our result.  $\square$



**Figure 13.** Top, left:  $P$ , with vertices marked. Top, right:  $Q$ , with vertices marked. Bottom: An example of the composition for  $r = 3$ .

**Proposition 3.4.** *The ramification type  $[r^2, 1^{4r-3}; 3^{2r-1}]$  produces a unique tree with monodromy group  $G$ , where*

$$G \cong \begin{cases} A_{2r-1} \wr C_3, & r \text{ odd}, \\ R_3, & r \text{ even}, \end{cases}$$

where  $R_3$  denotes the index-2 subgroup of  $S_{2r-1} \wr C_3$  such that  $\tau_1 \tau_2 \tau_3$  is an even permutation for all  $(\tau_1, \tau_2, \tau_3, g) \in R_3$ .

*Proof.* The procedure here is similar to the proof of the previous proposition. We observe that this dessin is the composition  $P \circ Q$ , where  $P$  is the 3-star with ramification type  $[1^3; 3]$  and  $Q$  is the dessin from Proposition 3.2 where  $s = 1$ ,  $r = t$ . See Figure 13.

We can label the dessin so that

$$\begin{aligned} \sigma_0 &= (1, 2, \dots, r)(\bar{r}, \overline{r+1}, \dots, \overline{2r-1}), \\ \sigma_1 &= (1, \bar{1}, \hat{1})(2, \bar{2}, \hat{2}) \cdots (2r-1, \overline{2r-1}, \widehat{2r-1}). \end{aligned}$$

Note that  $\sigma_0$  is the product of two  $r$ -cycles (call them  $\pi_1$  and  $\pi_2$  respectively) and that  $\sigma_1$  is the product of  $(2r-1)$  3-cycles. Consider the embedding  $\phi : G \rightarrow S_{2r-1} \wr C_3$  defined by

$$\begin{aligned} \sigma_0 &\mapsto (\pi_1, \pi_2, \text{id}, 0), \\ \sigma_1 &\mapsto (\text{id}, \text{id}, \text{id}, 1). \end{aligned}$$

Under this homomorphism, successive conjugations of  $\sigma_0$  by  $\sigma_1$  are mapped to  $(\text{id}, \pi_1, \pi_2, 0)$  and  $(\pi_2, \text{id}, \pi_1, 0)$ . Applying Lemma A.5 to  $\pi_1, \pi_2$ , and  $\phi(G)$ , we have  $A_{2r-1} \wr C_3 \leq \phi(G)$ . When  $r$  is odd, both  $\pi_1$  and  $\pi_2$  are even permutations, so  $A_{2r-1} \wr C_3 \geq \phi(G)$ , giving a double inclusion. When  $r$  is even, we consider the quotient group

$$(S_{2r-1} \wr C_3) / (A_{2r-1} \wr C_3) \cong C_2 \times C_2 \times C_2.$$

Observe that when  $r$  is even,  $\phi(G) \leq R_3$  and  $(\pi_1, \pi_2, \text{id}, 0)$  is equal to  $(1, 1, 0)$  in the quotient group  $\phi(G)/A_{2r-1} \wr C_3$ . We similarly have  $(0, 1, 1)$  and  $(1, 0, 1)$  in the quotient group. Hence, we see that  $\phi(G)$  is an index-2 subgroup of  $S_{2r-1} \wr C_3$  and thus  $\phi(G) \geq R_3$ .  $\square$

**Proposition 3.5.** *The ramification type  $[3^3, 1^5; 2^7]$  produces a unique tree with monodromy group  $G \cong A_7 \wr C_2$ .*

*Proof.* This is a sporadic case that may be verified by hand.  $\square$

#### 4. Future directions

The reader will notice that there are some obvious pathways left open by this paper. In Theorem 1.1 each entry refers to a tree uniquely determined by ramification type. For each entry there exists a Shabat polynomial with rational coefficients. However, we were not able to find a closed form expression for the coefficients of the rational Shabat polynomial given for the tree with ramification type  $[r^2, 1^{4r-3}; 3^{2r-1}]$ .

As for another direction of further inquiry, we note that the present paper focuses exclusively on (planar) trees uniquely determined by ramification type. However, we know that there exists an exhaustive list of ramification types that produce exactly two distinct trees, and perhaps there are other such lists for ramification types that produce larger numbers of trees [Shabat and Zvonkin 1994]. At the very least, it would be interesting to see the complete list of monodromy groups for ramification types that produce two trees in comparison with the completion of Theorem 1.1. Finally, it would also be of interest to see similar results for classes of dessins having at least one cycle or for dessins with genus greater than 1.

#### Appendix

In this section we prove a few technical results used in the paper. We learned of the following results (Lemmas A.1, A.2, A.3, A.4) and their proofs from Keith Conrad. Recall that we multiply permutations left to right.

**Lemma A.1.** *For  $n \geq 5$ , the subgroup generated by  $(1, 2, 3)$  and  $(1, 2, \dots, n)$  contains  $A_n$ .*

We prove this lemma through a sequence of lemmas.

**Lemma A.2.** *For  $n \geq 5$ , every element of  $A_n$  is a product of 3-cycles.*

*Proof.* The set of 3-cycles is a conjugacy class that is a subset of  $A_n$ . Therefore, the subgroup generated by the set of 3-cycles is a normal subgroup of  $A_n$ . Since  $A_n$  is simple for  $n \geq 5$ , we conclude that the set of 3-cycles generates  $A_n$  and every element of  $A_n$  is a product of 3-cycles.  $\square$

**Lemma A.3.** *For  $n \geq 5$ , the group  $A_n$  is generated by elements of the form  $(1, 2, k)$ .*

*Proof.* First observe that  $A_n$  is generated by 3-cycles of the form  $(1, i, j)$ . This is easily seen by observing that for any 3-cycle  $(a, b, c)$  not containing 1, we have  $(a, b, c) = (1, b, c)(1, a, b)$ . By Lemma A.2 we see that  $A_n$  is generated by 3-cycles of the form  $(1, i, j)$ .

Now we consider the 3-cycles of the form  $(1, 2, k)$ . Since  $(1, 2, k)^{-1} = (1, k, 2)$ , any 3-cycle with 1 and 2 is generated by 3-cycles of the form  $(1, 2, k)$ . For a 3-cycle  $(1, i, j)$  not containing 2, we have  $(1, i, j) = (1, 2, j)(1, 2, i)(1, 2, j)(1, 2, j)$ . Hence, every element of  $A_n$  is generated by elements of the form  $(1, 2, k)$ .  $\square$

**Lemma A.4.** *For  $n \geq 5$ , the consecutive 3-cycles  $(i, i+1, i+2)$  with  $1 \leq i \leq n-2$  generate  $A_n$ .*

*Proof.* This can be shown to be true for  $A_5$  by computation. We proceed to prove this for  $n > 5$  by induction.

Assume this is true for  $A_n$ . Consider  $A_{n+1}$ . By induction, we know that cycles of the form  $(i, i+1, i+2)$  generate the elements  $(1, 2, k)$  for  $3 \leq k \leq n$ . Therefore, by Lemma A.3, we need only show that we can generate  $(1, 2, n+1)$  in order to show that cycles of the form  $(i, i+1, i+2)$  generate  $A_{n+1}$ . Observe that  $(1, 2, n+1) = (1, 2, n)(1, 2, n-1)(n-1, n, n+1)(1, 2, n)(1, 2, n-1)$  and thus we have proven our result.  $\square$

Now we proceed with the proof of Lemma A.1

*Proof of Lemma A.1.* Let  $\sigma = (1, 2, \dots, n)$ . Observe that

$$\sigma^{-k}(1, 2, 3)\sigma^k = (\sigma^k(1), \sigma^k(2), \sigma^k(3)) = (k+1, k+2, k+3)$$

if  $0 \leq k \leq n-3$ . Thus by Lemma A.4,  $(1, 2, 3)$  and  $(1, 2, \dots, n)$  generate a subgroup that contains  $A_n$ .  $\square$

**Lemma A.5.** *Suppose that  $\pi_0, \pi_1 \in S_n$  with  $\langle \pi_0, \pi_1 \rangle \geq A_n$  with  $n \geq 5$ .*

- (1) *If  $|\pi_0| \neq |\pi_1|$ , then  $\Gamma = \langle (\pi_0, \pi_1), (\pi_1, \pi_0) \rangle$  must contain  $A_n \times A_n$ .*
- (2)  *$\Gamma = \langle (\pi_0, \pi_1, \text{id}), (\text{id}, \pi_0, \pi_1), (\pi_1, \text{id}, \pi_0) \rangle$  must contain  $A_n \times A_n \times A_n$ .*

*Proof.* Suppose that  $\text{id} \neq \rho \in A_n$ . Observe that  $\langle \tau^{-1}\rho\tau : \tau \in A_n \rangle$  is a normal subgroup of  $A_n$ . If  $n \geq 5$ , then  $A_n$  is simple and therefore,  $A_n = \langle \tau^{-1}\rho\tau : \tau \in A_n \rangle$ .

First, we consider statement (1). Suppose that  $(\rho, \text{id}) \in \Gamma$ . We want to show that  $A_n \times \langle \text{id} \rangle$  is a subgroup of  $\Gamma$ . There is a homomorphism  $\text{proj} : S_n \times S_n \rightarrow S_n$ , which is a projection from the first component. Since  $A_n \leq \langle \pi_0, \pi_1 \rangle$ , we have  $\text{proj}(\Gamma) \geq A_n$ . Therefore, for all  $\tau \in A_n$  there exists  $\tau' \in S_n$  such that  $(\tau, \tau') \in \Gamma$ . Conjugating  $(\rho, \text{id})$  by all  $(\tau, \tau')$  shows that  $A_n \times \langle \text{id} \rangle \leq \Gamma$ . Note that the same argument can be used to show  $\langle \text{id} \rangle \times A_n \leq \Gamma$  via projection in the other component. Statement (1) then follows as long as  $\rho \neq \text{id}$  exists. Furthermore, the argument to establish statement (2) would proceed in an identical fashion, presuming  $\rho \neq \text{id}$  exists.

To establish existence of  $\rho$  in the case of statement (1), we claim that there exists an element of the form  $(\rho, \text{id}) \in \Gamma$ , where  $\rho \neq \text{id}$ . Without loss of generality, assume  $|\pi_0| > |\pi_1|$ , and then consider  $(\pi_0, \pi_1)^{|\pi_1|}, (\pi_1, \pi_0)^{|\pi_1|}$ , in which case we may let  $\rho = \pi_0^{|\pi_1|}$ .

Now we prove such an element exists in the case of statement (2) for  $n > 2$ . If  $|\pi_0| \neq |\pi_1|$ , then the proof is analogous to the argument for statement (1). Otherwise  $|\pi_0| = |\pi_1| = r$  and we want to find some element  $\pi \in A_n$  such that  $|\pi| \nmid r$ . One can show that such a  $\pi$  exists by proving that, for  $n > 2$ , there must be some prime  $q$  not dividing  $|\pi_0| = r$ . One can show  $q$  exists by using the fact that

$$n < \sum_{\substack{p \leq n \\ p \text{ prime}}} p \quad \text{for } n > 2.$$

Using all three generators of  $\Gamma$ , one can produce the element  $(\pi_0^{k_1}, \pi, \pi_1^{k_2}) \in A_n^3$ , where  $k_1, k_2 \in \mathbb{Z}$ . By raising this element to the  $r$ -th power, we produce the element  $(\text{id}, \pi^r, \text{id}) \in \Gamma$  and let  $\rho = \pi^r$ .  $\square$

**Corollary A.6.** *Let  $H$  be a simple group. Suppose  $\pi_0, \pi_1 \in S_n$  with  $\langle \pi_0, \pi_1 \rangle \geq H$ .*

- (1) *If  $|\pi_0| \neq |\pi_1|$ , then  $\Gamma = \langle (\pi_0, \pi_1), (\pi_1, \pi_0) \rangle$  must contain  $H \times H$ .*
- (2)  *$\Gamma = \langle (\pi_0, \pi_1, \text{id}), (\text{id}, \pi_0, \pi_1), (\pi_1, \text{id}, \pi_0) \rangle$  must contain  $H \times H \times H$ .*

**Remark A.7.** In [Adrianov et al. 1997], Adrianov, Kochetkov, and Suvorov classify all the possible primitive, and thus simple, monodromy groups of plane trees.

## Acknowledgements

We thank the Willamette University Mathematics Consortium REU for providing a beautiful working environment, as well as the generous support of NSF Grant #1460982. We also wish to thank Edray Goins and an anonymous referee for helpful input on the content of this paper.

## References

- [Adrianov 2007] N. M. Adrianov, “On the generalized Chebyshev polynomials corresponding to plane trees of diameter 4”, *Fundam. Prikl. Mat.* **13**:6 (2007), 19–33. In Russian; translated in *J. Math. Sci. (NY)* **158**:1 (2009), 11–21. MR Zbl
- [Adrianov and Zvonkin 1998] N. Adrianov and A. Zvonkin, “Composition of plane trees”, *Acta Appl. Math.* **52**:1-3 (1998), 239–245. MR Zbl
- [Adrianov et al. 1997] N. M. Adrianov, Y. Y. Kochetkov, and A. D. Suvorov, “Plane trees with special primitive edge rotation groups”, *Fundam. Prikl. Mat.* **3**:4 (1997), 1085–1092. In Russian. MR Zbl
- [Anderson et al. 2018] J. Anderson, I. I. Bouw, O. Ejder, N. Girgin, V. Karemaker, and M. Manes, “Dynamical Belyi maps”, pp. 57–82 in *Women in numbers Europe, II* (Leiden, Netherlands, 2016), edited by I. I. Bouw et al., Assoc. Women Math. Ser. **11**, Springer, 2018. MR Zbl

- [Couveignes 1994] J.-M. Couveignes, “Calcul et rationalité de fonctions de Belyi en genre 0”, *Ann. Inst. Fourier (Grenoble)* **44**:1 (1994), 1–38. MR Zbl
- [Matiyasevich 1996] Y. V. Matiyasevich, “Computer evaluation of generalized Chebyshev polynomials”, *Vestnik Moskov. Univ. Ser. I Mat. Mekh.* **1996**:6 (1996), 59–61. In Russian; translated in *Moscow Univ. Math. Bull.* **51**:6 (1996), 39–40. MR Zbl
- [Schneps 1994] L. Schneps (editor), *The Grothendieck theory of dessins d’enfants* (Luminy, France, 1993), London Math. Soc. Lecture Note Series **200**, Cambridge Univ. Press, 1994. MR Zbl
- [Shabat and Zvonkin 1994] G. Shabat and A. Zvonkin, “Plane trees and algebraic numbers”, pp. 233–275 in *Jerusalem combinatorics ’93*, edited by H. Barcelo and G. Kalai, Contemp. Math. **178**, Amer. Math. Soc., Providence, RI, 1994. MR Zbl
- [Sijsling and Voight 2014] J. Sijsling and J. Voight, “On computing Belyi maps”, pp. 73–131 in *Numéro consacré au trimestre “Méthodes arithmétiques et applications”, automne 2013*, Publ. Math. Besançon Algèbre Théorie Nr. **2014/1**, Presses Univ. Franche-Comté, Besançon, 2014. MR Zbl
- [Wood 2006] M. M. Wood, “Belyi-extending maps and the Galois action on dessins d’enfants”, *Publ. Res. Inst. Math. Sci.* **42**:3 (2006), 721–737. MR Zbl

Received: 2018-05-02      Revised: 2018-11-03      Accepted: 2018-11-15

ncameron@lclark.edu	Lewis & Clark College, Portland, OR, United States
maroldkemp@gmail.com	Occidental College, Los Angeles, CA, United States
susan.m.maslak@gmail.com	Ave Maria University, Ave Maria, FL, United States
gmelamed@hawaii.edu	University of Hawaii at Manoa, Honolulu, HI, United States
rmoy@leeuniversity.edu	Lee University, Cleveland, TN, United States
jonatdp1@uci.edu	University of California, Irvine, CA, United States
abw22014@myemail.pomona.edu	Pomona College, Claremont, CA, United States

# On some edge Folkman numbers, small and large

Jenny M. Kaufmann, Henry J. Wickus and Stanisław P. Radziszowski

(Communicated by Kenneth S. Berenhaut)

Edge Folkman numbers  $F_e(G_1, G_2; k)$  can be viewed as a generalization of more commonly studied Ramsey numbers.  $F_e(G_1, G_2; k)$  is defined as the smallest order of any  $K_k$ -free graph  $F$  such that any red-blue coloring of the edges of  $F$  contains either a red  $G_1$  or a blue  $G_2$ . In this note, first we discuss edge Folkman numbers involving graphs  $J_s = K_s - e$ , including the results  $F_e(J_3, K_n; n+1) = 2n-1$ ,  $F_e(J_3, J_n; n) = 2n-1$ , and  $F_e(J_3, J_n; n+1) = 2n-3$ . Our modification of computational methods used previously in the study of classical Folkman numbers is applied to obtain upper bounds on  $F_e(J_4, J_4; k)$  for all  $k > 4$ .

## 1. Overview

For a graph  $F$ , we say that  $F \rightarrow (G_1, G_2)$  if in any red-blue coloring of the edges of  $F$ , there exists a red  $G_1$  or a blue  $G_2$ . The classical Ramsey numbers can be defined using this arrowing notation as  $R(G_1, G_2) = \min\{n \mid K_n \rightarrow (G_1, G_2)\}$ . If graph  $F$  is  $K_k$ -free and  $F \rightarrow (G_1, G_2)$ , then we write  $F \rightarrow (G_1, G_2; k)$ . If graph  $G_i$  is complete, we may write  $|V(G_i)|$  in place of  $G_i$ ; for example, instead of  $F \rightarrow (K_s, K_t; k)$  we could write  $F \rightarrow (s, t; k)$ . Given graphs  $G_1, G_2$  and an integer  $k > 1$ , we define the set of edge Folkman graphs by

$$\mathcal{F}_e(G_1, G_2; k) = \{F \mid F \rightarrow (G_1, G_2) \text{ and } K_k \not\subseteq F\},$$

and we will denote by  $\mathcal{F}_e(G_1, G_2; k; m)$  the set of such Folkman graphs with  $m$  vertices. The *edge Folkman number*  $F_e(G_1, G_2; k)$  is the smallest  $m$  such that  $\mathcal{F}_e(G_1, G_2; k; m)$  is nonempty. A theorem by Folkman [1970] states that if  $k > \max\{s, t\}$ , then  $F_e(s, t; k) = F_e(K_s, K_t; k)$  exists. One may easily notice that for graphs  $G_1$  and  $G_2$ , if  $k > R(G_1, G_2)$ , then  $F_e(G_1, G_2; k) = R(G_1, G_2)$ . Henceforth, in the sequel we will focus on the cases for  $k \leq R(G_1, G_2)$ .

In general, the Ramsey numbers  $R(G, H)$  are difficult to compute, and  $F_e(G, H; k)$  for  $k < R(G, H)$  still more so. The graph  $J_3 = P_3$ , however, leads to much

MSC2010: 05C55.

Keywords: Folkman numbers, Ramsey numbers.

Research supported by the NSF Research Experiences for Undergraduates Program (#1358583) held at Rochester Institute of Technology during the summer of 2016.

easier cases. The arrowing  $F \rightarrow (J_3, H)$  is equivalent to the question, “Does the removal of every matching  $sK_2$  from  $F$  leave a subgraph containing  $H$ ?” In Section 2, we present constructions which witness upper bounds on  $F_e(J_3; K_n; n+1)$ ,  $F_e(J_3; J_n; n+1)$ , and  $F_e(J_3; J_n; n)$ , and then we show that these bounds are tight.

In Section 3, we use computational methods modified from prior work on  $F_e(3, 3; 4)$  to determine values of Folkman numbers  $F_e(J_4, J_4; k)$  for  $k > 6$ , and bounds on  $F_e(J_4, J_4; k)$  for  $k = 5, 6$ . These are obtained with the help of techniques used in satisfiability (SAT) and MAX-CUT, both of which are well-studied problems in computer science. The cases of  $F_e(J_4, J_4; k)$  lie between the much-studied  $F_e(3, 3; k)$  and little-studied  $F_e(4, 4; k)$ . We also present up-to-date history of bounds on the former, namely  $F_e(3, 3; 4)$ .

## 2. Arrowing $(J_3, K_n)$ and $(J_3, J_n)$

Let the graph  $K_{2_n}$  denote the complete graph  $K_{2n}$  with removed perfect matching; i.e.,  $K_{2_n} = K_{2n} - nK_2$ .

**Proposition 1.** *For all  $n \in \mathbb{N}$ ,  $n \geq 2$ , we have  $K_{2_{n-1}} + K_1 \rightarrow (J_3, K_n)$ .*

*Proof.* We will first show that, for each  $n \geq 2$ , in any red-blue edge coloring of  $K_{2_{n-1}}$  avoiding red  $J_3 = P_3$ , every vertex  $v \in V(K_{2_{n-1}})$  belongs to a blue  $K_{n-1}$ . We proceed by induction. The claim is obvious for  $n = 2$ . Next, consider any red-blue coloring of  $K_{2_n}$  avoiding red  $J_3$ . Fix any  $v_1 \in V(K_{2_n})$ , and let  $v_2$  be the vertex not adjacent to  $v_1$ . If  $v_1$  is redly adjacent to some vertex  $w_1$ , then let  $\{w_1, w_2\}$  be nonadjacent; otherwise, choose an independent set  $\{w_1, w_2\}$  arbitrarily, but  $v_1 \notin \{w_1, w_2\}$ . The restriction of this coloring to  $K_{2_n} - \{v_1, v_2\} = K_{2_{n-1}}$  is a red-blue coloring avoiding red  $J_3$ , so by induction  $w_2$  is part of some blue  $K_{n-1} \subset K_{2_n} - \{v_1, v_2\}$ . Since  $v_1$  is adjacent to all vertices in  $K_{2_n} - \{v_1, v_2\}$  and is blue adjacent to all its vertices, possibly except  $w_1$ , together with this blue  $K_{n-1}$  it forms a blue  $K_n$ . By induction, the statement holds for all  $n$ .

Similarly, we prove the statement of the proposition by induction. Clearly, any red-blue edge coloring of  $K_{2_1} + K_1$  has either a red  $J_3$  or a blue  $K_2$ . For  $n \geq 3$ , consider any red-blue coloring of the graph  $K_{2_{n-1}} + K_1$  without any red  $J_3$ . Let  $\{x\} = V(K_1)$ . If any vertex  $v$  is redly adjacent to  $x$ , choose an independent set  $\{v_1, v_2\}$  so that  $v_2 = v$ ; otherwise, choose an independent set  $\{v_1, v_2\}$  arbitrarily. We have shown that in the restriction of this coloring to  $K_{2_{n-1}}$ ,  $v_1$  is in a blue  $K_{n-1}$ . Vertex  $v_2$  cannot be part of this  $K_{n-1}$ . Since  $x$  is adjacent to all vertices in  $V(K_{2_{n-1}})$ , and is blue adjacent to all such vertices (except perhaps  $v_2$ ), it is in a blue  $K_n$ . Thus,  $K_{2_{n-1}} + K_1 \rightarrow (J_3, K_n)$ .  $\square$

**Theorem 2.** *For all  $k > n \geq 2$  we have  $F_e(J_3, K_n; k) = 2n - 1$ .*

*Proof.* We notice that  $R(J_3, K_n) = 2n - 1$ , as listed in [Radziszowski 2017]. For  $k = n + 1$ , this gives the lower bound  $2n - 1 \leq F_e(J_3, K_n; n + 1)$ , while Proposition 1



provides a witness for the upper bound. For larger  $k$  the claim follows directly from definitions since  $F_e(J_3, K_n; k)$  is nonincreasing in  $k$ .  $\square$

**Theorem 3.** *For all  $n \geq 3$  we have*

$$F_e(J_3, J_n; k) = \begin{cases} 4 & \text{if } k = n = 3, \\ 2n - 3 & \text{if } k > n > 2, \\ 2n - 1 & \text{if } k = n \text{ and } n > 3. \end{cases}$$

*Proof.* For the special case of  $k = n = 3$ , it can be easily checked that  $K_{1,3} \rightarrow (J_3, J_3)$ ; hence it gives the upper bound. Clearly, three vertices are not enough for a suitable Folkman graph, so  $F_e(J_3, J_3; 3) = 4$ .

For the case  $k > n > 2$ , as in Theorem 3, the lower bound  $F_e(J_3, J_n; n+1) \geq 2n-3$  for any  $k \geq n$  follows from  $R(J_3, J_n) = 2n-3$ ; see [Radziszowski 2017]. For the upper bound, we will prove that  $K_{2n-3} + K_3 \rightarrow (J_3, J_n)$ . Consider any red-blue coloring of the graph  $K_{2n-3} + K_3$  avoiding red  $J_3$ . Let  $\{x, y, z\} = V(K_3)$  and let  $e$  be the edge  $\{x, y\}$ . By Proposition 1, the restriction of this coloring to the subgraph  $K_{2n-2} + K_1 = K_{2n-3} + (K_3 - e)$  must include a blue  $K_{n-1}$ . Since  $K_{n-1} \not\subset K_{2n-3} + K_1$ , this blue  $K_{n-1}$  must include exactly one of  $x$  or  $y$ ; without loss of generality it includes  $x$  and not  $y$ . But in the original coloring,  $y$  is blue adjacent to all or all but one of the vertices in the blue  $K_{n-1}$ , so  $y$  is part of a blue  $J_n$ . Hence  $F_e(J_3, J_n; k) = 2n-3$  for all  $k > n$ .

Finally we consider the case of  $k = n$  for  $n > 3$ . Consider any  $K_n$ -free graph  $G$  with  $|V(G)| = 2n-2$ . Color the edges of  $G$  as follows: take a maximum matching  $R \subseteq E(G)$ , color all of its edges in red, and color all edges in  $G - R$  blue. This coloring contains no red  $J_3$ . We will show that either it contains no blue  $J_n$ , or that  $G \subseteq K_{n-2} + nK_1$ .

Suppose that  $G$  contains a blue  $J_n$  and let  $S \subset V(G)$  be the vertices of the  $J_n$ . Since  $G$  does not contain  $K_n$ , there exist nonadjacent vertices  $x, y \in S$ . Every edge in  $R$  must be incident to a vertex in  $\bar{S} = V(G) - S$ , implying that  $|R| \leq |\bar{S}| = n-2$ . Now consider any pair of adjacent vertices  $s, t \in S$  (one of which may be  $x$  or  $y$ ). Since  $s$  and  $t$  are adjacent, at least one must be incident to a red edge, since otherwise we could add the edge  $\{s, t\}$  to  $R$  and obtain a matching larger than  $R$ . Since  $|R| \leq |S| - 2$ , there exist two vertices in  $S$  neither of which is incident to a red edge; then these vertices must be  $x$  and  $y$ . Furthermore, any other vertex in  $S$  is adjacent to  $x$  and  $y$ , so it must be incident to some red edge. Therefore,  $|R| = n-2 = |\bar{S}|$ .

For any two vertices  $s', t' \in \bar{S}$ , there exist vertices  $s, t \in S$  distinct from  $x$  and  $y$ , such that  $\{s, s'\}$  and  $\{t, t'\}$  are red edges. We must have that  $s'$  and  $t'$  are nonadjacent, since otherwise we could obtain a matching larger than  $R$  by taking  $R$ , removing edges  $\{s, s'\}$  and  $\{t, t'\}$ , and replacing them with edges  $\{x, s\}$ ,  $\{y, t\}$ , and  $\{s', t'\}$ . Additionally, if (without loss of generality)  $x$  is adjacent to  $s' \in \bar{S}$ , then we could

obtain a matching larger than  $R$  by replacing edge  $\{s, s'\}$  with edges  $\{x, s'\}$  and  $\{y, s\}$ . Thus, the vertex set  $\bar{S} \cup \{x, y\}$  does not induce any edges, implying that  $G \subseteq K_{n-2} + nK_1$ .

We can edge color  $K_{n-2} + nK_1$  in a way that avoids red  $J_3$  and blue  $J_n$  simply by coloring only one edge in the  $K_{n-2}$  red. Thus,  $K_{n-2} + nK_1 \not\rightarrow (J_3, J_n)$ . Then there is no graph  $G$  on  $2n - 2$  vertices such that  $G \rightarrow (J_3, J_n; n)$ , which gives the lower bound  $F_e(J_3, J_n; n) \geq 2n - 1$ . For the upper bound we consider the graph  $K_{2n-1} + K_1$ . Let  $\{x\} = V(K_1)$  and let vertices  $v_1, v_2$  be nonadjacent. By Proposition 1, any red-blue coloring of  $K_{2n-1} + K_1$  with no red  $J_3$  contains a blue  $K_n$ . This blue  $K_n$  can include at most one of  $v_1, v_2$ , and therefore at most one of  $\{v_1, x\}$  and  $\{v_2, x\}$ . Hence, consider the subgraph  $K_{2n-2} + \bar{K}_3 \subset K_{2n-1} + K_1$  constructed by removing the edges  $\{v_1, x\}$  and  $\{v_2, x\}$ . Next, observe that any coloring of  $K_{2n-2} + \bar{K}_3$  with no red  $J_3$  therefore contains a blue  $J_n$ . So  $K_{2n-2} + \bar{K}_3 \rightarrow (J_3, J_n)$ , and thus,  $F_e(J_3, J_n; n) = 2n - 1$ .  $\square$

### 3. Folkman numbers $F_e(J_4, J_4; k)$

**3.1. Cases for  $k \geq 6$ .** In order to find upper bounds on  $F_e(J_4, J_4; k)$  for  $k \geq 6$  we reduced the corresponding arrowings to instances of the Boolean satisfiability (SAT) problem, which has been extensively studied. In particular, this approach had been previously used by Shetler, Wurtz, and the third author to test arrowing of  $(K_3, J_4)$ . We applied it instead to the question of whether  $G \not\rightarrow (J_4, J_4)$ , as follows: We map the edges  $E(G)$  to the variables of a Boolean formula  $\phi_G$ , so that the color of an edge  $e$  is represented by the value of its corresponding Boolean variable. Then for each  $J_4$  consisting of edges  $e_1, e_2, e_3, e_4, e_5$ , we add to  $\phi_G$  two clauses,

$$(e_1 + e_2 + e_3 + e_4 + e_5) \wedge (\bar{e}_1 + \bar{e}_2 + \bar{e}_3 + \bar{e}_4 + \bar{e}_5).$$

Then  $G \not\rightarrow (J_4, J_4)$  if and only if  $\phi_G$  is satisfiable. We solved many such instances of satisfiability problem for formulas  $\phi_G$  with the SAT-solver MiniSAT [Eén and Sörensson 2004]. The results of these computations lead to the next theorem.

**Theorem 4.** *It holds that*

$$F_e(J_4, J_4; k) = \begin{cases} 10 & \text{for } k \geq 8, \\ 11 & \text{for } k = 7, \end{cases}$$

and  $11 \leq F_e(J_4, J_4; 6) \leq 14$ .

*Proof.* It is known that  $R(J_4, J_4) = 10$ , see [Chvátal and Harary 1972], and hence  $F_e(J_4, J_4; k) \geq 10$  for all  $k \geq 4$ , and  $F_e(J_4, J_4; k) = 10$  for  $k \geq 11$ . A computation using MiniSAT determined that the graph  $G = K_4 + K_{2,2,2}$  satisfies  $G \rightarrow (J_4, J_4)$ . Since  $|V(G)| = 10$  and  $G$  is  $K_8$ -free, using previous comments we obtain that

$F_e(J_4, J_4; 8) = 10$ . Because  $F_e(J_4, J_4; k)$  is nonincreasing in  $k$ , we also obtain that  $F_e(J_4, J_4; k) = 10$  for  $k = 9$  and  $k = 10$ .

To find the lower bound for  $F_e(J_4, J_4; 7)$ , we tested all nonisomorphic graphs on 10 vertices found with nauty [McKay and Piperno 2014]. We ignored graphs containing  $K_7$  and those which are  $K_5$ -free (since it would contradict  $F_e(3, 3; 5) = 15$  [Piwakowski et al. 1999]). Testing exhaustively all 1806547 such graphs via  $\phi_G$  with MiniSAT revealed that  $\mathcal{F}_e(J_4, J_4; 7; 10) = \emptyset$ , and thus  $F_e(J_4, J_4; 7) \geq 11$ . A computation using MiniSAT determined that the graph  $F = K_2 + K_{3,2,2,2}$  satisfies  $F \rightarrow (J_4, J_4)$ . Since  $|V(F)| = 11$  and  $F$  is  $K_7$ -free, much as before we obtain  $F_e(J_4, J_4; 7) \leq 11$ . Lastly, we determined using MiniSAT that the graph  $H = C_5 + K_{3,3,3}$  satisfies  $H \rightarrow (J_4, J_4)$ . Since  $|V(H)| = 14$  and  $H$  is  $K_6$ -free, we have  $F_e(J_4, J_4; 6) \leq 14$ .  $\square$

The exact value of  $F_e(J_4, J_4; 6)$  possibly could be determined as above with a larger effort using similar computational techniques.

**3.2.  $F_e(J_4, J_4; 5)$  and MAX-CUT.** Our attempts to use MiniSAT to find a graph  $G$  witnessing an upper bound on  $F_e(J_4, J_4; 5)$  were unsuccessful, as the SAT-solver slowed down significantly when we tested larger graphs. However, we managed to obtain the bound  $F_e(J_4, J_4; 5) \leq 1297$  using a modification of an idea and computational approach of Dudek and Rödl [2008] for studying  $F_e(3, 3; 4)$ , which itself is based on an idea of Goodman [1959].

For a red-blue coloring of a graph  $G$ , we define  $T_{\text{diff}}(v)$  and  $T_{\text{same}}(v)$ , respectively, to be the number of triangles containing  $v$  in which the edges incident to  $v$  are different colors or the same color. Let  $t$  be the number of triangles in  $G$ , and let  $m$  be the number of monochromatic triangles in  $G$ . In each nonmonochromatic triangle, there are two vertices  $v_1, v_2$  for which the edges incident to it are different colors. Then  $\sum_{v \in G} T_{\text{diff}}(v) = 2(t - m)$  counts each nonmonochromatic triangle in  $G$  twice. Furthermore,  $\sum_{v \in G} T_{\text{same}}(v) = t + 2m$  gives the number of nonmonochromatic triangles plus three times the number of monochromatic triangles. Therefore,

$$6m = 2 \sum_{v \in G} T_{\text{same}}(v) - \sum_{v \in G} T_{\text{diff}}(v). \quad (1)$$

Observe that if  $3m > |E(G)|$ , then the ratio of edges in monochromatic triangles to edges is greater than 1, implying that there is some edge  $e$  which is part of two distinct monochromatic triangles. Therefore, if for every red-blue coloring of  $G$  we have

$$2|E(G)| < 2 \sum_{v \in G} T_{\text{same}}(v) - \sum_{v \in G} T_{\text{diff}}(v), \quad (2)$$

then  $G \rightarrow (J_4, J_4)$ .

We now recall a method for linking arrowing triangles to the MAX-CUT problem, first proposed by Dudek and Rödl [2008]. Let  $H_G$  be the graph created as follows: We map every edge  $e$  of  $G$  to vertex  $v_e$  of  $H$ , so that  $V(H_G) = E(G)$ . Then for any two vertices  $v_e, v_f$  in  $V(H_G)$ , we add the edge  $\{v_e, v_f\}$  if and only if their corresponding edges  $e$  and  $f$  are a part of some triangle in  $G$ . Note that any red-blue coloring of  $E(G)$  corresponds to a bipartition  $V(H_G) = B \cup R$  of vertices of  $H_G$ , inducing an edge cut  $C$ , for which any nonmonochromatic triangle in  $G$  has exactly two edges in  $C$ . For any graph  $F$ , let  $\text{MC}(F) = \text{MAX-CUT}(F)$  denote the maximum number of edges in  $F$  between the partite sets of any bipartition of  $V(F)$ . Letting  $M_C(H_G)$  be the size of the cut  $C$ , we have

$$M_C(H_G) = \sum_{v \in G} T_{\text{diff}}(v) \leq \text{MC}(H_G). \quad (3)$$

Clearly, any edge in  $H_G$  has both endpoints in the same partite set  $B$  or  $R$  if and only if it is not in  $C$ . The above considerations lead to the following theorem.

**Theorem 5.** *If  $\text{MC}(H_G) < 2t(G) - 2|E(G)|/3$ , then  $G \rightarrow (J_4, J_4)$ .*

*Proof.* For any graph  $G$  whose edges are arbitrarily colored red and blue, consider the cut  $C$  of  $H_G$  as described above. Using (1) and (3), one can easily show that

$$\sum_{v \in G} T_{\text{same}}(v) = |E(H_G)| - M_C(H_G) = 3t - M_C(H_G).$$

Now from the assumption we have  $2|E(G)| < 2(3t - M_C(H_G)) - (M_C(H_G))$ . Finally, using (2) and its implication we conclude that  $G \rightarrow (J_4, J_4)$ .  $\square$

For large graphs  $H$ , finding tight upper bounds for  $\text{MC}(H)$  is computationally expensive. For this reason, we used the following weakening of Theorem 5 for vertex-transitive graphs  $G$ . Its advantage is that it allows us to detect conditions for which Theorem 5 can be applied much faster.

**Theorem 6.** *Let  $G$  be a vertex-transitive  $d$ -regular graph, where  $G_v$  denotes the graph induced in  $G$  by the neighbors of vertex  $v$ . If we have*

$$\text{MC}(G_v) < \frac{2}{3}|E(G_v)| - \frac{1}{3}d,$$

*then  $G \rightarrow (J_4, J_4)$ .*

*Proof.* This is following the same argument as in an alternative approach to bounding Folkman numbers used by Lu [2008] and Spencer [1988]. Here, however, with an additional term  $d/3$ , we need to use the observation made above between equalities (1) and (2).  $\square$

MAX-CUT is among Karp's original 21 NP-hard problems [1972]. In order to find good bounds on  $\text{MC}(H_G)$  and  $\text{MC}(G_v)$  for graphs  $G$  of our interest, we used the eigenvalue and semidefinite programming approximations of MAX-CUT. This

approach was used by several authors, including Lu [2008], Dudek and Rödl [2008], and Lange et al. [2014] to obtain upper bounds on  $F_e(3, 3; 4)$  (see Section 3.3 for a historical summary).

We applied Theorems 5 and 6 to many graphs of different types. We found an interesting positive instance using the following construction described in [Lu 2008]. For positive integers  $n$  and  $s$ ,  $s < n$ , define  $S = \{s^i \pmod{n} \mid i = 0, 1, \dots, n-1\}$ . Then, if  $n-1 \in S$ , let  $L(n, s)$  be the graph with vertex set  $\mathbb{Z}_n$  and edge set  $\{\{x, y\} \mid x - y \in S\}$ . Clearly, the graphs  $L(n, s)$  are vertex-transitive.

**Theorem 7.**  $F_e(J_4, J_4; 5) \leq 1297$ .

*Proof.* For the graph  $L(1297, 8)$ , which is 216-regular, we found that it satisfies the assumptions of both Theorems 5 and 6, using two MAX-CUT bounding methods: the eigenvalue method and the SDP approach. We used our Java library and associated programs, including the `eigs` function in Matlab and the SDP solver SDP-LR [Helmberg and Rendl 2000]. An easy (computer) test shows that the graph  $L(1297, 8)$  is  $K_5$ -free, and hence it is a witness of the upper bound.  $\square$

We wish to note that recently (and after this work was completed) a much better bound of 51 on  $F_e(J_4, J_4; 5)$  was obtained in [Xu et al. 2018]. The latter bound did not require any computations. We also would like to recall the bound on  $F_e(J_4, J_4; 4)$  obtained in [Lu 2008], as follows.

**Proposition 8.**  $F_e(J_4, J_4; 4) \leq 30193$ .

The bound in Proposition 8 is mentioned by Lu [2008] in his paper on  $F_e(3, 3; 4)$  as a side result, without any comments on the approach. However, we communicated with the author who confirmed that the main idea of his approach was similar to one in this work.

**3.3. History of the Folkman number  $F_e(3, 3; 4)$ .** Table 1 below summarizes the history of bounds on the edge Folkman number  $F_e(3, 3; 4) = F_e(K_3, K_3; 4)$ , which is the smallest unknown classical Folkman number, sometimes also called *the most wanted*. This table builds on an earlier Table 5 by Xu and the third author [Xu and Radziszowski 2016], where further extensive comments about the progress related to  $F_e(3, 3; 4)$  can be found. The new entries in Table 1 here are lower bounds 13, 14 and 20. The bound  $F_e(3, 3; 4) \geq 14$  can be obtained as follows: removal of any independent set of three vertices from any graph in  $\mathcal{F}_e(3, 3; 4)$  must yield a 5-chromatic  $K_4$ -free graph, but Nenov [1984] proved (without using computer algorithms) that any such graph has at least 11 vertices.  $F_e(3, 3; 4) \geq 13$  is implied in the same way by an earlier result of Nenov [1983]. In contrast, the currently best-known lower bound of 20 was obtained by Bikov and Nenov [2017] using CPU-intensive computations.

year	lower/upper bounds	who/what
1967	any?	[Erdős and Hajnal 1967]
1970	exist	[Folkman 1970]
1972	11 –	implicit in [Lin 1972], implied by $F_e(3, 3; 5) \geq 10$
1975	– $10^{10}$ ?	Erdős [1975] offers \$100 for proof
1983	13 –	implied by a result of [Nenov 1983]
1984	14 –	implied by a result of [Nenov 1984]
1986	– $8 \cdot 10^{11}$	[Frankl and Rödl 1986]
1988	– $3 \cdot 10^9$	[Spencer 1988]
1999	16 –	Piwakowski, Radziszowski and Urbański, implicit in [Piwakowski et al. 1999]
2007	19 –	[Radziszowski and Xu 2007]
2008	– 9697	[Lu 2008]
2008	– 941	[Dudek and Rödl 2008]
2012	– 100?	Graham offers \$100 for proof
2014	– 786	Lange, Radziszowski and Xu [Lange et al. 2014]
2017	20 –	[Bikov and Nenov 2017]

**Table 1.** History of bounds on the Folkman number  $F_e(3, 3; 4)$ .

For any graph  $G$  with  $t$  triangles and graph  $H_G$  as defined in Section 3.2, one can easily observe that  $G \rightarrow (K_3, K_3)$  if and only if  $\text{MC}(H_G) < 2t$ ; see also [Dudek and Rödl 2008]. Thus, computational techniques to find upper bounds for MAX-CUT may lead to good upper bounds on  $F_e(3, 3; 4)$ , including the first such result by Dudek and V. Rödl [2008]. Lange, Xu, and the third author used the SDP MAX-CUT approximation to obtain an upper bound on  $\text{MC}(H_G)$  for a particular  $K_4$ -free graph  $G$  on 786 vertices, and used it to show that  $G \rightarrow (K_3, K_3)$ .

We made numerous attempts to lower this bound by trying to find a smaller  $K_4$ -free graph  $G$  for we could obtain the bound  $\text{MC}(H_G) < 2t$ . Among the graphs tested were the graphs  $G(n, r)$  considered in [Dudek and Rödl 2008], the graphs  $L(n, s)$  from [Lu 2008], and their variations. In particular, we tested a generalization of  $L(n, s)$  to Galois fields  $\text{GF}(p^k)$ , in addition to graphs constructed by adjoining various pairs of circulant graphs in a variety of ways. Our efforts have convinced us that these methods are unlikely to yield any major improvement on this bound.

The well-known  $K_4$ -free graph  $G_{127} = L(127, 5)$  was studied by several authors; see for example [Radziszowski and Xu 2007; Xu and Radziszowski 2016]. In particular, it was conjectured by Exoo that  $G_{127} \rightarrow (K_3, K_3)$ . Needless to say, we

were not successful in proving Exoo's conjecture, because otherwise it would imply that  $F_e(3, 3; 4) \leq 127$ .

**Computations.** Some of the results in this paper were found through the use of various computational methods. This involved a large library of functions, including graph manipulation, construction of various types of graphs, and tests for graph arrowing. Graphs were represented in a variety of ways, including two-dimensional Boolean arrays, lists of edges for sparse graphs, and the g6-format of [McKay and Piperno 2014]. Our code was written in Java and executed on Unix and Windows systems. For our final results, Matlab and SDP-LR [Helmberg and Rendl 2000; Rendl et al. 2010] were used to calculate eigenvalue and SDP MAX-CUT approximations, respectively. MiniSAT [Eén and Sörensson 2004] was used to solve satisfiability problems. We also made use of lists of nonisomorphic graphs with special properties found with nauty [McKay and Piperno 2014].

### Acknowledgment

We are grateful to an anonymous reviewer whose insightful comments led to a better and more complete presentation of this paper.

### References

- [Bikov and Nenov 2017] A. Bikov and N. Nenov, "The edge Folkman number  $F_e(3, 3; 4)$  is greater than 19", *Geombinatorics* **27**:1 (2017), 5–14. MR Zbl
- [Chvátal and Harary 1972] V. Chvátal and F. Harary, "Generalized Ramsey theory for graphs, II: Small diagonal numbers", *Proc. Amer. Math. Soc.* **32** (1972), 389–394. MR Zbl
- [Dudek and Rödl 2008] A. Dudek and V. Rödl, "On the Folkman number  $f(2, 3, 4)$ ", *Experiment. Math.* **17**:1 (2008), 63–67. MR Zbl
- [Eén and Sörensson 2004] N. Eén and N. Sörensson, "An extensible SAT-solver", pp. 502–518 in *Theory and applications of satisfiability testing* (Santa Margherita Ligure, Italy), edited by E. Giunchiglia and A. Tacchella, Lect. Notes in Comput. Sci. **2919**, Springer, 2004. Zbl
- [Erdős 1975] P. Erdős, "Problems and results on finite and infinite graphs", pp. 183–192 in *Recent advances in graph theory* (Prague, 1974), edited by M. Fiedler, Academia, Prague, 1975. MR Zbl
- [Erdős and Hajnal 1967] P. Erdős and A. Hajnal, "Research problem 2-5", *J. Combin. Theory* **2** (1967), 104.
- [Folkman 1970] J. Folkman, "Graphs with monochromatic complete subgraphs in every edge coloring", *SIAM J. Appl. Math.* **18** (1970), 19–24. MR Zbl
- [Frankl and Rödl 1986] P. Frankl and V. Rödl, "Large triangle-free subgraphs in graphs without  $K_4$ ", *Graphs Combin.* **2**:2 (1986), 135–144. MR Zbl
- [Goodman 1959] A. W. Goodman, "On sets of acquaintances and strangers at any party", *Amer. Math. Monthly* **66** (1959), 778–783. MR Zbl
- [Helmberg and Rendl 2000] C. Helmberg and F. Rendl, "A spectral bundle method for semidefinite programming", *SIAM J. Optim.* **10**:3 (2000), 673–696. MR Zbl

- [Karp 1972] R. M. Karp, “Reducibility among combinatorial problems”, pp. 85–103 in *Complexity of computer computations* (New York, 1972), edited by R. E. Miller and J. W. Thatcher, Plenum, New York, 1972. MR Zbl
- [Lange et al. 2014] A. R. Lange, S. P. Radziszowski, and X. Xu, “Use of max-cut for Ramsey arrowing of triangles”, *J. Combin. Math. Combin. Comput.* **88** (2014), 61–71. MR Zbl
- [Lin 1972] S. Lin, “On Ramsey numbers and  $\mathbb{K}_r$ -coloring of graphs”, *J. Combin. Theory Ser. B* **12** (1972), 82–92. MR Zbl
- [Lu 2008] L. Lu, “Explicit construction of small Folkman graphs”, *SIAM J. Discrete Math.* **21**:4 (2008), 1053–1060. MR Zbl
- [McKay and Piperno 2014] B. D. McKay and A. Piperno, “Practical graph isomorphism, II”, *J. Symbolic Comput.* **60** (2014), 94–112. MR Zbl
- [Nenov 1983] N. D. Nenov, “Zykov numbers and some of their applications in Ramsey theory”, *Serdica* **9**:2 (1983), 161–167. In Russian. MR Zbl
- [Nenov 1984] N. D. Nenov, “The chromatic number of any 10-vertex graph without 4-cliques is at most 4”, *C. R. Acad. Bulgare Sci.* **37**:3 (1984), 301–304. In Russian. MR Zbl
- [Piwakowski et al. 1999] K. Piwakowski, S. P. Radziszowski, and S. Urbański, “Computation of the Folkman number  $F_e(3, 3; 5)$ ”, *J. Graph Theory* **32**:1 (1999), 41–49. MR Zbl
- [Radziszowski 2017] S. Radziszowski, “Small Ramsey numbers”, *Electronic J. Combinator.* Dynamic Surveys (2017), art. id. DS1; revision 15.
- [Radziszowski and Xu 2007] S. P. Radziszowski and X. Xu, “On the most wanted Folkman graph”, *Geombinatorics* **16**:4 (2007), 367–381. MR
- [Rendl et al. 2010] F. Rendl, G. Rinaldi, and A. Wiegele, “Solving max-cut to optimality by intersecting semidefinite and polyhedral relaxations”, *Math. Program. Ser. A* **121**:2 (2010), 307–335. MR Zbl
- [Spencer 1988] J. Spencer, “Three hundred million points suffice”, *J. Combin. Theory Ser. A* **49**:2 (1988), 210–217. Correction in **50** (1989), 323. MR Zbl
- [Xu and Radziszowski 2016] X. Xu and S. P. Radziszowski, “On some open questions for Ramsey and Folkman numbers”, pp. 43–62 in *Graph theory: favorite conjectures and open problems, I*, edited by R. Gera et al., Springer, 2016. MR Zbl
- [Xu et al. 2018] X. Xu, M. Liang, and S. Radziszowski, “A note on upper bounds for some generalized Folkman numbers”, *Discuss. Math. Graph Theory* (online publication January 2018).

Received: 2018-06-03      Revised: 2018-10-23      Accepted: 2018-11-29

jennak@princeton.edu

*Department of Mathematics, Princeton University,  
Princeton, NJ, United States*

hwickus@gmail.com

*Department of Mathematics and Computer Science,  
DeSales University, Center Valley, PA, United States*

spr@cs.rit.edu

*Department of Computer Science, Rochester Institute  
of Technology, Rochester, NY, United States*



# Weighted persistent homology

Gregory Bell, Austin Lawson, Joshua Martin,  
James Rudzinski and Clifford Smyth

(Communicated by Józef H. Przytycki)

We introduce weighted versions of the classical Čech and Vietoris–Rips complexes. We show that a version of the Vietoris–Rips lemma holds for these weighted complexes and that they enjoy appropriate stability properties. We also give some preliminary applications of these weighted complexes.

## 1. Introduction

Topological data analysis (TDA) provides a means for the power of algebraic topology to be used to better understand the shape of a data set. In the traditional approach to TDA, isometric balls of a fixed radius  $r > 0$  are centered at each data point in some ambient Euclidean space. One then constructs the nerve of the union of these balls and computes the simplicial homology of this nerve. Computationally, this approach is infeasible for large data sets or high-dimensional data, so instead one computes the so-called Vietoris–Rips complex, which is the flag complex over the graph obtained by placing an edge between any pair of vertices that are at distance no more than  $2r$  from each other. The key idea of TDA is to allow the radius of these balls to vary and to compute simplicial homology for each value of this radius to create a topological profile of the space. This profile is encoded in either a barcode or a persistence diagram. Topological features such as holes or voids that exist for a relatively large interval of radii are said to persist and are believed to be more important than more transient features that exist for very short intervals of radii. (There are, however, important exceptions to this rule of thumb; see [Bendich et al. 2016].)

In the traditional model, the radius of each ball is the same and can be modeled by the linear function of time  $r(t) = rt$ . In this paper, we consider a model of computing persistent homology in which the radius of each ball is allowed to be a

---

*MSC2010:* primary 55N35; secondary 55U99, 68U10.

*Keywords:* persistent homology, weighted persistent homology, stability, Vietoris–Rips complex, Čech complex, interleaving, bottleneck distance, persistence diagram.

Smyth was supported by NSA MSP Grant H98230-13-1-0222 and by a grant from the Simons Foundation (Grant Number 360486, CS).

different monotonic function  $r_x(t)$  at each point  $x$ . In this way we can emphasize certain data points by assigning or *weighting* them with larger and/or more quickly growing balls and de-emphasize others by weighting them with smaller and/or more slowly growing balls. This is appropriate in the case of a noisy data set, for instance, as an alternative to throwing away data that fails to meet some threshold of significance. Various other methods of enhancing persistence with weights have been considered; see, e.g., [Buchet et al. 2016; Edelsbrunner and Morozov 2013; Petri et al. 2013; Ren et al. 2017; 2018].

The weighted model we propose fits into the framework of generalized persistence in the sense of [Bubenik et al. 2015]. We show that it enjoys many of the properties familiar from the techniques of traditional persistent homology. We prove a weighted Vietoris–Rips lemma (Theorem 3.2) that relates our weighted Čech and Rips complexes in the same way that they are related in the case of isometric balls. We also show that the persistent homology computed over weighted complexes is stable with respect to small perturbations of the rates of growth and/or the points in the data set (Theorem 4.1). Moreover, packages for computing persistent homology such as Javaplex [Adams et al. 2014] and Perseus [Mischaikow and Nanda 2013] are capable of handling our weighted persistence with the same complexity as unweighted persistence by merely adjusting inputs to the package functions.

As a proof of concept, we apply our methods to the Modified National Institute of Standards and Technology (MNIST) data set of handwritten digits translated into pixel information. Our method proves more effective than isometric persistence in finding the number 8 from among these handwritten digits. (We chose 8 for its unique 1-dimensional homology among these digits.) We found our methods to be 95.8% accurate as opposed to isometric persistence’s 92.07% accuracy. This experiment was chosen to demonstrate the performance of weighted persistence over usual persistence, but it should be noted that neither method approaches the accuracy of state-of-the art computer vision and we make no claim that we are improving on known methods.

In Section 2, we provide the background definitions that are needed for what follows and describe our weighted persistence model. In Section 3 we prove the weighted Vietoris–Rips lemma and indicate how persistent homology packages can be used to compute weighted persistence. In Section 4 we establish our stability results. Our experiments on MNIST data appear in Section 5. We end with some remarks and questions for further study.

## 2. Preliminaries

We begin by defining some terminology and setting our notation. We will assume some familiarity with simplicial homology and the basic ideas of topological data analysis. For details, we refer to [Edelsbrunner and Harer 2010; Rotman 1988].

In algebraic topology, simplicial homology is a tool that assigns to any simplicial complex  $K$  a collection of  $\mathbb{Z}$ -modules  $H_0(K), H_1(K), \dots$ , called *homology groups*, in such a way that the rank of  $H_n(K)$  describes the number of “ $n$ -dimensional holes” in  $K$ . For our purposes, we replace the standard definition in terms of  $\mathbb{Z}$ -modules with vector spaces (usually over the field with two elements, for ease of computation). We therefore refer to *homology vector spaces* instead of homology groups. We do not attempt to define  $H_n(K)$  here, but instead refer to any text in algebraic topology, such as [Rotman 1988].

Let  $\mathcal{U}$  be a collection of sets. We define the *nerve*  $\mathcal{N}(\mathcal{U})$  to be the abstract simplicial complex with vertex set  $\mathcal{U}$  with the property that the subset  $\{U_0, U_1, \dots, U_n\}$  of  $\mathcal{U}$  spans an  $n$ -simplex in  $\mathcal{N}$  whenever  $\bigcap_{i=0}^n U_i \neq \emptyset$ .

Let  $(X, d)$  be a metric space. We define  $B_r(x) = \{y \in X \mid d(x, y) < r\}$  and  $\bar{B}_r(x) = \{y \in X \mid d(x, y) \leq r\}$  to be the open and closed balls of radius  $r$  about  $x$ , respectively. (Note that we’re abusing notation since in a general metric space  $\bar{B}_r(x)$  is not necessarily the closure of the open ball, usually denoted by  $B_r(x)$ .) We most often consider examples where  $X$  is a subset of  $\mathbb{R}^d$  and  $d(x, y) = \|x - y\|$  is the Euclidean distance between  $x$  and  $y$ . For a real number  $r \geq 0$ , we define the *Čech complex of  $X$  at scale  $r$*  by  $\check{\text{Cech}}(r) = \mathcal{N}\{\bar{B}_r(x) \mid x \in X\}$ .

We generalize this construction by allowing the radius of the ball around each element  $x$  to depend on  $x$ . Let  $\mathbf{r} : X \rightarrow [0, \infty)$  be any function. We define the *weighted  $\mathbf{r}$ -Čech complex*  $\check{\text{Cech}}(\mathbf{r})$  of  $X$  by  $\check{\text{Cech}}(\mathbf{r}) = \mathcal{N}\{\bar{B}_{\mathbf{r}(x)}(x)\}$ .

In practice, it is difficult to determine whether an intersection of balls is nonempty. A much simpler construction to use is the Vietoris–Rips complex. For a given parameter  $r \geq 0$  the *Vietoris–Rips complex* is the flag complex of the 1-skeleton of the Čech complex; i.e., a collection of  $n + 1$  balls forms an  $n$ -simplex in the Vietoris–Rips complex if and only if the balls are pairwise intersecting. For the Vietoris–Rips complex we identify each ball with its center, so that the *Vietoris–Rips complex at scale  $r$*  is  $\text{VR}(r) = \{\sigma \subset X \mid \text{diam}(\sigma) \leq 2r\}$ . Similarly, if  $\mathbf{r} : X \rightarrow [0, \infty)$ , the *weighted  $\mathbf{r}$ -Vietoris–Rips complex* is

$$\text{VR}(\mathbf{r}) = \{\sigma \subset X \mid d(x, y) \leq \mathbf{r}(x) + \mathbf{r}(y) \text{ for all } x, y \in \sigma \text{ with } x \neq y\}.$$

Fix  $\mathbf{r} : X \rightarrow [0, \infty)$  and consider the simplicial complex  $\check{\text{Cech}}(\mathbf{r})$  (or  $\text{VR}(\mathbf{r})$ ). Using simplicial homology with field coefficients, one can associate homology vector spaces  $H_*(\check{\text{Cech}}(\mathbf{r}))$  to these simplicial complexes. Whenever  $t_0 \leq t_1$  there is a natural inclusion map of simplicial complexes given by  $\iota : \check{\text{Cech}}(t_0\mathbf{r}) \rightarrow \check{\text{Cech}}(t_1\mathbf{r})$  (or the corresponding inclusion of the Vietoris–Rips complexes). By functoriality, there is an induced linear map on homology  $\iota_* : H_*\check{\text{Cech}}(t_0\mathbf{r}) \rightarrow H_*\check{\text{Cech}}(t_1\mathbf{r})$ .

Let  $X \subset \mathbb{R}^d$  be finite. Although we defined the weighted complexes above for any function  $\mathbf{r} : X \rightarrow [0, \infty)$ , we want to study the persistence properties of these weighted complexes. For example, in the case of the weighted Čech complex, we

want to study the evolution of homology as the radii of the balls grow to infinity. One straightforward way to do this would be to simply scale our weighted complexes linearly in the same way that one usually scales the isometric balls in persistent homology. We prefer a more flexible approach, which we describe in terms of radius functions.

Let  $\mathcal{C}_+^1 = \mathcal{C}_+^1([0, \infty))$  denote the collection of differentiable bijective functions  $\phi : [0, \infty) \rightarrow [0, \infty)$  with positive first derivative. By a *radius function* on  $X$  we mean a function  $\mathbf{r} : X \rightarrow \mathcal{C}_+^1$ . We denote the image function  $\mathbf{r}(x)$  by  $r_x$ .

For  $t \geq 0$ , we define the Čech and Vietoris–Rips complexes at scale  $t$  by

$$\check{\text{Cech}}_{\mathbf{r}}(t) = \mathcal{N}\{\bar{B}_{r_x(t)}(x)\}$$

and

$$\text{VR}_{\mathbf{r}}(t) = \{\sigma \subset X \mid d(x, y) \leq r_x(t) + r_y(t) \text{ for all } x, y \in \sigma \text{ with } x \neq y\},$$

respectively. We define the *entry function*,

$$f_{X, \mathbf{r}}(y) = \min_{x \in X} \{r_x^{-1}(d(y, x))\}. \quad (1)$$

This function captures the scale  $t$  at which the point  $y \in \mathbb{R}^d$  is first captured by some ball  $\bar{B}_{r_x(t)}(x)$ ; we have  $f_{X, \mathbf{r}}(y) = t$  if and only if  $y \in \bar{B}_{r_x(t)}(x)$  for some  $x$  in  $X$  and  $y \notin \bigcup_{x \in X} B_{r_x(t)}(x)$ . Thus we have the following proposition.

**Proposition 2.1.** *Let  $X$  be a finite subset of some Euclidean space  $\mathbb{R}^d$ . Suppose that  $\mathbf{r}$  and  $f_{X, \mathbf{r}}$  are defined as above. Then,*

$$f_{X, \mathbf{r}}^{-1}([0, t]) = \bigcup_{x \in X} \bar{B}(x, r_x(t)).$$

It follows from the nerve lemma, see for example [Hatcher 2002, Corollary 4G.3], that  $\check{\text{Cech}}_{\mathbf{r}}(t)$  is homotopy equivalent to  $f_{X, \mathbf{r}}^{-1}([0, t])$ .

### 3. A weighted Vietoris–Rips lemma

The Vietoris–Rips complex is much easier to compute than the Čech complex in high dimensions. To determine whether  $n + 1$  balls form an  $n$ -simplex in the Čech complex, we must check whether the balls intersect, a computationally complex problem. To determine whether  $n + 1$  balls  $B_{r_i}(x_i)$  form a simplex in the Vietoris–Rips complex is computationally easy; only  $\binom{n+1}{2}$  conditions  $d(x_i, x_j) \leq r_i + r_j$  need be checked. Furthermore, if there are  $m$  points in  $X$ , it may be necessary to check all  $2^m$  subcollections of balls to determine the Čech complex, whereas determining the Rips complex will only require checking  $\binom{m}{2}$  pairs of points.

Our weighted Čech and Vietoris–Rips complexes are similar in spirit to weighted alpha complexes [Edelsbrunner and Harer 2010, III.4]. Both constructions seek to

permit “balls” with different sizes. Our constructions are simpler from a conceptual standpoint since the alpha complexes are built as subcomplexes of the Delaunay complex, which comes from the Voronoi diagram. Moreover, our complexes are computationally simple; indeed our method of finding weighted Vietoris–Rips complexes requires only marginally more computation than the unweighted Vietoris–Rips complex.

In particular, Javaplex and Perseus can compute regular (unweighted) persistent homology given input of a distance matrix  $M$  with  $M_{i,j} = d(x_i, x_j)$ . Inputting  $M_{i,j} = d(x_i, x_j)/(r_i + r_j)$  allows these packages to compute the persistent homology with  $r_{x_i}(t) = r_i t$  in the same time.

In computational problems it is common to use the Vietoris–Rips complex instead of the Čech complex to simplify the calculational overhead. The following theorem justifies this decision by saying that the Vietoris–Rips complex is “close” to the weighted Čech complex.

The classical Vietoris–Rips lemma can be stated as follows:

**Theorem 3.1** [de Silva and Ghrist 2007]. *Let  $X$  be a set of points in  $\mathbb{R}^d$  and let  $t > 0$ . Then*

$$\text{VR}(t') \subseteq \check{\text{Cech}}(t) \subseteq \text{VR}(t)$$

*whenever  $0 < t' \leq t(\sqrt{2d/(d+1)})^{-1}$ .*

The main result of this section is an extension of this result to the weighted case.

**Theorem 3.2** (weighted Vietoris–Rips lemma). *Let  $X$  be a set of points in  $\mathbb{R}^d$ . Let  $\mathbf{r} : X \rightarrow (0, \infty)$  be the corresponding weight function and let  $t > 0$ . Then*

$$\text{VR}(t'\mathbf{r}) \subseteq \check{\text{Cech}}(t\mathbf{r}) \subseteq \text{VR}(t\mathbf{r})$$

*whenever  $0 < t' \leq t(\sqrt{2d/(d+1)})^{-1}$ .*

*Proof.* The second containment  $\check{\text{Cech}}(t\mathbf{r}) \subseteq \text{VR}(t\mathbf{r})$  follows from the fact that the weighted Vietoris–Rips complex is the flag complex of the weighted Čech complex.

To show that  $\text{VR}(t'\mathbf{r}) \subset \check{\text{Cech}}(t\mathbf{r})$ , we suppose there is some finite collection  $\sigma = \{x_k\}_{k=0}^\ell \subseteq \mathbb{R}^d$  with  $\ell > 0$  that is a simplex in  $\text{VR}(t'\mathbf{r})$  and show that this is also a simplex in  $\check{\text{Cech}}(t\mathbf{r})$ . We have  $\|x_i - x_j\|_2 \leq t'(\mathbf{r}(x_i) + \mathbf{r}(x_j))$  whenever  $i \neq j$ .

Define a function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  by

$$f(y) = \max_{0 \leq j \leq \ell} \left\{ \frac{\|x_j - y\|_2}{\mathbf{r}(x_j)} \right\}.$$

Clearly,  $f$  is continuous and  $f(y) \rightarrow \infty$  as  $\|y\|_2 \rightarrow \infty$ . Thus  $f$  attains a minimum (say at  $y_0$ ) on some compact set containing  $\text{Conv}(\{x_k\}_{k=0}^\ell)$ . (Here  $\text{Conv}(S)$  is the convex hull of the set  $S \subseteq \mathbb{R}^d$ .) We must have  $\|x_i - y_0\|_2/\mathbf{r}(x_i) = f(y_0)$  for at least

one of the vertices  $x_i$ . By reordering the vertices, we may assume that

$$\begin{aligned} f(y_0) &= \frac{1}{\mathbf{r}(x_j)} \|x_j - y_0\|_2 & \text{if } 0 \leq j \leq n, \\ f(y_0) &> \frac{1}{\mathbf{r}(x_j)} \|x_j - y_0\|_2 & \text{if } n < j \leq \ell. \end{aligned}$$

Let

$$\begin{aligned} g(y) &= \max_{0 \leq j \leq n} \left\{ \frac{1}{\mathbf{r}(x_j)} \|x_j - y\|_2 \right\}, \\ h(y) &= \max_{n < j \leq \ell} \left\{ \frac{1}{\mathbf{r}(x_j)} \|x_j - y\|_2 \right\}. \end{aligned}$$

Now we wish to show that  $y_0 \in \text{Conv}(\{x_j\}_{j=0}^n)$ . To this end we apply the separation theorem [Matoušek 2002] to obtain: either  $y_0 \in \text{Conv}(\{x_j\}_{j=0}^n)$  or there is a  $v \in \mathbb{R}^d$  and a  $C < 0$  such that  $v x_j \geq 0$  for all  $0 \leq j \leq n$  and  $v y_0 < C$ . Thus if  $y_0 \notin \text{Conv}(\{x_j\}_{j=0}^n)$  there is a  $v \in \mathbb{R}^d$  such that  $v(x_j - y_0) > 0$  for  $0 \leq j \leq n$ . We suppose that there is such a  $v$  and derive a contradiction.

Since

$$\|x_j - (y_0 + \lambda v)\|_2^2 = \|x_j - y_0\|_2^2 - 2\lambda v(x_j - y_0) + \lambda^2 \|v\|_2^2$$

for each  $0 \leq j \leq n$ , it follows that  $g(y_0 + \lambda v) < f(y_0)$  for all  $0 < \lambda < \lambda_1$ , where

$$\lambda_1 = \min_{0 \leq j \leq n} \left\{ \frac{2v(x_j - y_0)}{\|v\|_2^2} \right\}.$$

Since  $h(y)$  is continuous and  $h(y_0) < f(y_0)$ , there exists a  $\lambda_2$  such that  $h(y_0 + \lambda v) < f(y_0)$  for  $0 < \lambda < \lambda_2$ . Thus, there exists a  $\lambda > 0$  such that

$$f(y_0 + \lambda v) = \max\{g(y_0 + \lambda v), h(y_0 + \lambda v)\} < f(y_0),$$

contradicting the minimality of  $y_0$ .

By Carathéodory's theorem [Matoušek 2002] and reordering of vertices if necessary,  $y_0$  is a convex combination of some subcollection of vertices  $\{x_j\}_{j=0}^m$ , where  $m \leq \min\{d, n\}$ . It is not possible that  $m = 0$ . If so, then  $y_0 = x_0$  and

$$f(y_0) = \frac{1}{\mathbf{r}(x_0)} \|x_0 - y_0\|_2 = 0$$

and  $f$  is identically zero. Since  $\sigma$  has dimension at least 1, it contains a vertex  $x_1 \neq x_0$ . It follows that

$$f(y_0) = f(x_0) > \frac{1}{\mathbf{r}(x_1)} \|x_1 - x_0\|_2 > 0,$$

which is a contradiction.

Let  $\hat{x}_j = x_j - y_0$  for all  $0 \leq j \leq m$ . Note that

$$\|\hat{x}_j\|_2^2 = \mathbf{r}(x_j)^2 f(y_0)^2. \quad (2)$$

Since  $y_0 \in \text{Conv}(\{x_j\}_{j=0}^m)$ , we know  $y_0 = \sum_{j=0}^m a_j x_j$  for some set of nonnegative real numbers  $a_0, \dots, a_m$  that sum to 1. Thus  $\sum_{j=0}^m a_j \hat{x}_j = 0$ . By relabeling, we may assume that  $a_0 \mathbf{r}(x_0) \geq a_j \mathbf{r}(x_j)$  when  $j > 0$ . Necessarily  $a_0 > 0$  (otherwise  $a_j = 0$  for all  $0 \leq j \leq m$ , a contradiction). Then,

$$\hat{x}_0 = - \sum_{j=0}^m \frac{a_j}{a_0} \hat{x}_j$$

and so

$$\mathbf{r}(x_0)^2 f(y_0)^2 = \|\hat{x}_0\|_2^2 = - \sum_{j=0}^m \frac{a_j}{a_0} \hat{x}_0 \hat{x}_j.$$

Among the indices  $1, 2, \dots, m$ , there is some  $j_0$  such that

$$\frac{1}{d} \mathbf{r}(x_0)^2 f(y_0)^2 \leq \frac{1}{m} \mathbf{r}(x_0)^2 f(y_0)^2 \leq - \frac{a_{j_0}}{a_0} \hat{x}_0 \hat{x}_{j_0}. \quad (3)$$

We must have  $a_{j_0} > 0$ . (Otherwise,  $f(y_0) = 0$ , which, as shown earlier, is a contradiction.) By reordering, we may assume  $j_0 = 1$ . Putting (1) and (2) together, we find

$$\begin{aligned} f(y_0)^2 \left( \mathbf{r}(x_0)^2 + \frac{2a_0 \mathbf{r}(x_0)^2}{a_1 d} + \mathbf{r}(x_1)^2 \right) \\ &= f(y_0)^2 \mathbf{r}(x_0)^2 + \frac{2a_0 f(y_0)^2 \mathbf{r}(x_0)^2}{a_1 d} + f(y_0)^2 \mathbf{r}(x_1)^2 \\ &\leq \|\hat{x}_0\|_2^2 - 2\hat{x}_0 \hat{x}_1 + \|\hat{x}_1\|_2^2 \\ &= \|\hat{x}_0 - \hat{x}_1\|_2^2 \\ &= \|x_0 - x_1\|_2^2 \\ &\leq (t'(\mathbf{r}(x_0) + \mathbf{r}(x_1)))^2. \end{aligned}$$

We will now show that

$$\frac{f(y_0)^2}{t'} \leq \frac{(\mathbf{r}(x_0)^2 + \mathbf{r}(x_1)^2)^2}{\mathbf{r}(x_0)^2 + 2a_0 \mathbf{r}(x_0)^2 / (a_1 d) + \mathbf{r}(x_1)^2} \leq \frac{2d}{d+1}.$$

It suffices to show, after cross-multiplying the right-hand inequality, that

$$\left( d - 1 + 4 \frac{a_0}{a_1} \right) \mathbf{r}(x_0)^2 - 2(d+1) \mathbf{r}(x_0) \mathbf{r}(x_1) + (d-1) \mathbf{r}(x_1)^2 \geq 0.$$

Since

$$\frac{a_0}{a_1} \geq \frac{\mathbf{r}(x_1)}{\mathbf{r}(x_0)}$$

we get

$$\begin{aligned} & \left(d - 1 + 4\frac{a_0}{a_1}\right) \mathbf{r}(x_0)^2 - 2(d+1)\mathbf{r}(x_0)\mathbf{r}(x_1) + (d-1)\mathbf{r}(x_1)^2 \\ & \geq \left(d - 1 + 4\frac{\mathbf{r}(x_1)}{\mathbf{r}(x_0)}\right) \mathbf{r}(x_0)^2 - 2(d+1)\mathbf{r}(x_0)\mathbf{r}(x_1) + (d-1)\mathbf{r}(x_1)^2 \\ & = (d-1)(\mathbf{r}(x_0) - \mathbf{r}(x_1))^2 \geq 0, \end{aligned}$$

as desired. Our assumption that  $t' \leq t(\sqrt{2d/(d+1)})^{-1}$  implies  $f(y_0) \leq t$  and thus

$$y_0 \in \bigcap_{i=0}^{\ell} \bar{B}_{tr(x_i)}(x_i).$$

Therefore  $\sigma \in \check{\text{Cech}}(tr)$  and we are done.  $\square$

#### 4. Stability

In this section we discuss the stability of our weighted persistence. Let  $X$  and  $Y$  be finite subsets of  $\mathbb{R}^d$  with corresponding radii functionals  $\mathbf{r} : X \rightarrow \mathcal{C}_+^1$  and  $\mathbf{s} : Y \rightarrow \mathcal{C}_+^1$ . Informally, we show that if  $(X, \mathbf{r})$  and  $(Y, \mathbf{s})$  are “close”, i.e., are small perturbations of each other, then the corresponding entry functions  $f_{X,\mathbf{r}}$  and  $f_{Y,\mathbf{s}}$ , see (1), are also “close” and hence the associated persistence diagrams must also be “close”. We’ll now make the definitions of these various types of closeness precise.

Let  $\eta \subseteq X \times Y$  be a relation such that for every  $x \in X$  there is a  $y \in Y$  with  $(x, y) \in \eta$  and for every  $y \in Y$  there is an  $x \in X$  with  $(x, y) \in \eta$ . We measure the closeness of  $X$  and  $Y$  with respect to  $\eta$  by

$$\|\eta\| := \max_{(x,y) \in \eta} d(x, y).$$

If  $L$  is any compact set and  $h : L \rightarrow \mathbb{R}$  is continuous let

$$\|h\|_L := \max_{x \in L} |h(x)|.$$

Let  $K$  be a compact subset of  $\mathbb{R}^d$  that contains  $X \cup Y$ . The closeness of  $\mathbf{r}$  and  $\mathbf{s}$  is measured by

$$D(\mathbf{r}, \mathbf{s})_{\eta, K} := \max_{(x,y) \in \eta} \|\mathbf{r}_x^{-1} - \mathbf{s}_y^{-1}\|_{[0, \text{diam}(K)]}.$$

The closeness of  $f_{X,\mathbf{r}}$  and  $f_{Y,\mathbf{s}}$  is measured by  $\|f_{X,\mathbf{r}} - f_{Y,\mathbf{s}}\|_K$ . We also define  $S(\mathbf{r})_K := \max_{x \in X} \|(\mathbf{r}_x^{-1})'\|_{[0, \text{diam}(K)]}$ .

As is common, we measure the closeness of persistence diagrams by the bottle-neck distance. We’ll give the definition of this metric in the remarks leading up to Theorem 4.5.



**Theorem 4.1.** *In the above notation we have the following bound on entry functions (see (1)):*

$$\|f_{X,r} - f_{Y,s}\|_K \leq D(\mathbf{r}, \mathbf{s})_{\eta,K} + \|\eta\| \max(S(\mathbf{r})_K, S(\mathbf{s})_K)$$

*Proof.* There is some point  $z$  in the compact set  $K$  and some points  $x \in X$  and  $y \in Y$  so that

$$\|f_{X,r} - f_{Y,s}\|_K = |f_{X,r}(z) - f_{Y,s}(z)| = |\mathbf{r}_x^{-1}(d(z, x)) - \mathbf{s}_y^{-1}(d(z, y))|.$$

We first suppose  $\mathbf{r}_x^{-1}(d(z, x)) \geq \mathbf{s}_y^{-1}(d(z, y))$ . Let  $x' \in X$  such that  $(x', y) \in \eta$ . Since  $f_{X,r}$  is a minimum,  $\mathbf{r}_{x'}^{-1}(d(z, x')) \geq \mathbf{r}_x^{-1}(d(z, x))$  and we have

$$\begin{aligned} \|f_{X,r} - f_{Y,s}\|_K &\leq |\mathbf{r}_{x'}^{-1}(d(z, x')) - \mathbf{s}_y^{-1}(d(z, y))| \\ &\leq |\mathbf{r}_{x'}^{-1}(d(z, x')) - \mathbf{s}_y^{-1}(d(z, x'))| + |\mathbf{s}_y^{-1}(d(z, x')) - \mathbf{s}_y^{-1}(d(z, y))|. \end{aligned} \quad (4)$$

Since  $d(z, x') \in [0, \text{diam}(K)]$ ,

$$|\mathbf{r}_{x'}^{-1}(d(z, x')) - \mathbf{s}_y^{-1}(d(z, x'))| \leq D(\mathbf{r}, \mathbf{s})_{\eta,K}.$$

Since  $|d(z, x') - d(z, y)| \leq d(x', y) \leq \|\eta\|$  we apply the mean value theorem to obtain the bound

$$|\mathbf{s}_y^{-1}(d(z, x')) - \mathbf{s}_y^{-1}(d(z, y))| \leq \|\eta\| \|(\mathbf{s}_y^{-1})'\|_{[0, \text{diam}(K)]} \leq \|\eta\| \max(S(\mathbf{r})_K, S(\mathbf{s})_K).$$

Together, these last two bounds give the bound of the theorem. A similar argument gives the same bound if  $\mathbf{r}_x^{-1}(d(z, x)) \leq \mathbf{s}_y^{-1}(d(z, y))$ .  $\square$

If one has free choice of the perturbed set  $(Y, \mathbf{s})$  it is clear that  $\|f_{X,r} - f_{Y,s}\|_K$  can be made arbitrarily large. This could be done, say by adding a point to  $Y$  that is arbitrarily far from any point in  $X$  or by making one  $\mathbf{s}_y$  arbitrarily larger than any  $\mathbf{r}_x$ . The upper bound of Theorem 4.1 is also a bound on how extreme such perturbations may be.

We have the following immediate corollary of Theorem 4.1.

**Corollary 4.2.** *If the radii functions are all linear, i.e., if there are positive constants  $r_x$  and  $s_y$  for all  $x \in X$  and  $y \in Y$  such that  $r_x(t) = r_x t$  and  $s_y(t) = s_y t$ , then*

$$\|f_{X,r} - f_{Y,s}\|_K \leq \text{diam}(K) \max_{(x,y) \in \eta} \left| \frac{1}{r_x} - \frac{1}{s_y} \right| + \|\eta\| \max \left( \max_{x \in X} \frac{1}{r_x}, \max_{y \in Y} \frac{1}{s_y} \right).$$

For our next two corollaries, let  $X$  and  $Y$  have the same cardinality and let  $m : X \rightarrow Y$  be a bijection. We now consider each point  $x \in X$  as being perturbed to a point  $m(x) \in Y$  and hence set  $\eta = \{(x, m(x)) : x \in X\}$ . We have the following point-stability result in which the points are perturbed but the weight functions stay the same.

**Corollary 4.3** (point-stability). *If only the locations of the points are perturbed and the radius functions stay the same, i.e.,  $s_{m(x)}(t) = r_x(t)$  for all  $x \in X$ , then*

$$\|f_{X,r} - f_{Y,s}\|_K \leq \max_{x \in X} d(x, m(x)) \|(\mathbf{r}_x^{-1})'\|_{[0, \text{diam}(K)]}.$$

*Proof.* We follow the proof of Theorem 4.1. Take  $x' \in X$  such that  $m(x') = y$ . Then  $S_y = \mathbf{r}_{x'}$  and the first term in the upper bound in inequality (4) is 0. Since the second term in that upper bound is bounded above by  $d(x', m(x')) \|(\mathbf{r}_{x'}^{-1})'\|_{[0, \text{diam}(K)]}$ , the bound of the corollary holds.  $\square$

The next corollary is a weight-function stability result concerning a case in which the points stay the same ( $Y = X$  and  $m(x) = x$ ) but the weight functions are perturbed.

**Corollary 4.4** (weight-function stability). *If only the radii functions are perturbed and the points stay the same, then*

$$\|f_{X,r} - f_{X,s}\|_K \leq \max_{x \in X} \|\mathbf{r}_x^{-1} - \mathbf{s}_x^{-1}\|_{[0, \text{diam}(K)]}.$$

*Proof.* Again following the proof of Theorem 4.1 we take  $x' = m(x') = y$ . Now the second term in the upper bound of (4) is 0 and the first term is  $|\mathbf{r}_y^{-1} - \mathbf{s}_y^{-1}|$ , where  $t = d(z, y)$ . The corollary follows.  $\square$

We now show the stability of the persistence diagrams of  $f_{X,r}$  under perturbations of  $X$  and  $\mathbf{r}$ . Let  $f : K \rightarrow [0, \infty)$  be a real-valued function on a compact set  $K \subseteq \mathbb{R}^d$ . The *persistence diagram* of  $f$ ,  $\text{dgm}(f)$ , is a multiset of points in  $[0, +\infty]^2$  recording the appearance and disappearance of homological features in  $f^{-1}([0, t])$  as  $t$  increases. Each point  $(b, d)$  in the diagram tracks a single homological feature, recording the scale  $t = b$  at which the feature first appears and the scale  $t = d$  at which it disappears [Edelsbrunner and Harer 2010]. It should also be noted that if one considers the birth-death pair as an interval, we obtain the *barcode* as seen in [Zomorodian and Carlsson 2005] (see Figures 2 and 3). Given two functions  $f, g : K \rightarrow [0, \infty]$ , let  $P = \text{dgm}(f)$  and  $Q = \text{dgm}(g)$  be the corresponding persistence diagrams (where as usual we include all points along the diagonal in  $P$  and  $Q$ ). We let  $N$  denote the set of all bijections from  $P$  to  $Q$ . We recall that the *bottleneck distance* between the diagrams [Edelsbrunner and Harer 2010] is given by

$$d_B(\text{dgm}(f), \text{dgm}(g)) = \inf_{\gamma \in N} \sup_{x \in P} \|x - \gamma(x)\|_\infty.$$

**Theorem 4.5** [Cohen-Steiner et al. 2007, Theorem 6.9]. *Suppose  $\mathcal{X}$  is a triangulable space and that  $f : \mathcal{X} \rightarrow \mathbb{R}$  and  $g : \mathcal{X} \rightarrow \mathbb{R}$  are tame, continuous functions. If  $|f - g|$  is bounded, then for each  $n$*

$$d_B(\text{dgm}_n(f), \text{dgm}_n(g)) \leq \|f - g\|_\infty,$$

where  $d_B$  denotes the bottleneck distance and  $\text{dgm}_n(f)$  denotes the  $n$ -th persistence diagram of the filtration of  $f$ .

We refer to [Edelsbrunner and Harer 2010] for the technical definitions of tame and triangulable. Note that as our spaces are nerves of balls around finite collections of points, they are finite simplicial complexes. Hence they are triangulable and only admit tame functions. Thus for our setting we get the following corollary.

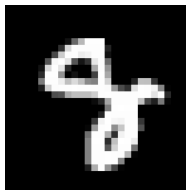
**Corollary 4.6.** *Let  $X$  and  $Y$  be finite subsets of  $\mathbb{R}^d$  and let  $\mathbf{r} : X \rightarrow \mathcal{C}_+^1$  and  $\mathbf{s} : Y \rightarrow \mathcal{C}_+^1$ . Suppose that  $\eta \subseteq X \times Y$  is a relation as above and  $K$  is a compact subset of  $\mathbb{R}^d$  containing  $X$  and  $Y$ . Then for each  $n$ ,*

$$d_B(\text{dgm}_n(f_{X,\mathbf{r}}), \text{dgm}_n(f_{Y,\mathbf{s}})) \leq D(\mathbf{r}, \mathbf{s})_{\eta,K} + \|\eta\| \max(S(\mathbf{r})_K, S(\mathbf{s})_K).$$

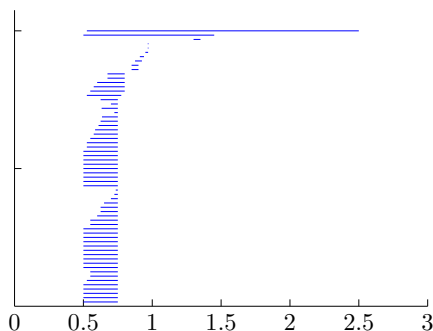
## 5. MNIST 8's recognition

In this section, we give an application of weighted persistence to a simple computer vision problem. We apply our methods to the Modified National Institute of Standards and Technology (MNIST) data set of handwritten digits. We should emphasize that this application is simply a proof of concept; our methods to detect the handwritten number 8 fall well short of state-of-the-art methods [Cireřan et al. 2012].

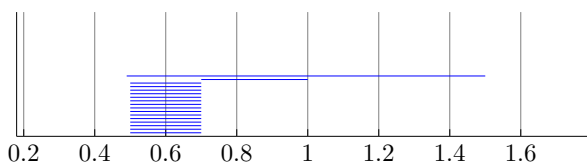
The MNIST data set consists of handwritten digits (0 through 9) translated into pixel information. Each data point contains a label and 784 other values ranging from 0 to 255 that correspond to a 28 by 28 grid of pixels. The values 0 through 255 correspond to the intensity of the pixels in gray-scale with 0 meaning completely black and 255 meaning completely white. Considering the digits from 0 through 9, unweighted persistence would easily be able to classify these numbers as having zero, one, or two holes, provided they are written precisely; however, real handwritten digits present a challenge. Consider an 8 as in Figure 1. Unweighted persistence would pick up on two holes, but one of those holes might be slightly too small and ultimately considered insignificant; see Figure 3. Our methods are able to pick up on both holes and would count them as significant; see Figure 2. We chose to work with the digit 8 due to its unique homology.



**Figure 1.** An 8 converted to a 28 by 28 grid of pixels.



**Figure 2.** Weighted persistence on the image from Figure 1 produces a barcode that clearly has two long bars in dimension 1.



**Figure 3.** Unweighted persistence on the image from Figure 1 produces a barcode that has one long bar (in 1-homology). The second-longest bar is hard to distinguish (in length) from the rest.

To begin, we convert each 28 by 28 to a set of points in the plane. We treat the location of a value in the matrix as a location in the plane. That is, the value in the  $i$ -th row,  $j$ -th column corresponds to the point  $(i, j)$ . The weight on each point is exactly its corresponding pixel intensity. Using this set of points and corresponding weights we calculate persistent homology via weighted Rips complexes. We test this method's performance against the unweighted case where all nonzero pixel values have the uniform weight of 1; again we calculate persistence in this case via Rips complexes.

We compare weighted persistence to unweighted persistence by measuring the accuracy of classifying 8's. Notice in the barcodes that the deciding factor in determining an 8 is the ability to distinguish the length of the second longest bar from the length of the third longest and smaller bars. For this reason, we consider the ratio of the third longest bar to the second longest bar. We will say (arbitrarily) that a barcode represents an 8 if this ratio is less than  $\frac{1}{2}$ . For each of the 42,000 handwritten digits in the MNIST data set, we compute both weighted and unweighted persistence and collect the predictions. We obtain the confusion matrices as in Table 1.

Notice that the weighted persistence has an accuracy rate of 95.8% whereas unweighted persistence had an accuracy of 92.07%. A full summary can be seen

	weighted persistence		unweighted persistence	
	predicted not 8	predicted 8	predicted not 8	predicted 8
not 8	36487	1450	35869	2068
is 8	633	3430	1261	2802

**Table 1.** The confusion matrices show that weighted persistence outperforms its unweighted counterpart.

	weighted persistence	unweighted persistence
accuracy	0.9504	0.9207
sensitivity	0.9618	0.9455
specificity	0.8442	0.6896
pos. predicted value	0.9829	0.9660
neg. predicted value	0.7029	0.5754
prevalence	0.9033	0.9033
balanced accuracy	0.9030	0.8176

**Table 2.** Weighted and unweighted persistence compared.

in Table 2. We view this result as promising for potential future applications of weighted persistence.

## 6. Concluding remarks and open questions

The method of weighted persistence satisfies the appropriate Vietoris–Rips lemma, is stable under small perturbations of the points, or the weights, or both, and can be successfully applied to data such as the MNIST data set to improve upon usual persistence. Furthermore, it is just as easy to calculate weighted persistence for balls growing at linear rates as it is to calculate regular persistence. We conclude the paper with some further observations and questions.

One can imagine weighted persistence as interpolating between two extreme approaches to a data set that is partitioned into data  $D$  and noise  $N$ . More precisely, we consider a noisy data set  $X$ . Various methods exist to filter  $X$  into data  $D$  and noise  $N$ . Traditional persistence can be applied to  $D \cup N$  in two ways. We can either assign the same radius to every point of  $D \cup N$  or we can throw the points of  $N$  out entirely and compute persistence on  $D$  alone. Using weighted persistence, we can assign the radius 0 to each point of  $N$  and compute weighted persistence of  $D \cup N$ . It is easy to see that this will differ from persistence of  $D$  itself only in dimension 0. By gradually increasing the  $N$ -radii from 0 to 1, our stability results can be interpreted as producing a continuum of barcodes/persistence diagrams that

interpolate between the usual persistence applied to  $D$  and the usual persistence applied to  $D \cup N$  (in dimensions above 0); see [Lawson 2016].

As mentioned in the Introduction, weighted persistence fits into the framework of generalized persistence in the sense of [Bubenik et al. 2015]. This direction was explored in detail in [Martin 2016].

Finally, it would be interesting to apply weighted persistence to the MNIST data set to determine its effectiveness in distinguishing the 1-homology of the other nine digits. One complication is that the number 4 presents an interesting challenge since it is appropriate to write it both as a simply connected space and as a space with nontrivial  $H_1$ . Distinguishing 1-homology creates three clusters of digits from which we could use other machine-learning techniques to create an ensemble and make accurate predictions.

## References

- [Adams et al. 2014] H. Adams, A. Tausz, and M. Vejdemo-Johansson, “javaPlex: a research software package for persistent (co)homology”, pp. 129–136 in *International Congress on Mathematical Software 2014* (Seoul, 2014), edited by H. Hong and C. Yap, Lecture Notes in Comput. Sci. **8592**, 2014. Software available at <http://appliedtopology.github.io/javaplex>. Zbl
- [Bendich et al. 2016] P. Bendich, J. S. Marron, E. Miller, A. Pieloch, and S. Skwerer, “Persistent homology analysis of brain artery trees”, *Ann. Appl. Stat.* **10**:1 (2016), 198–218. MR
- [Bubenik et al. 2015] P. Bubenik, V. de Silva, and J. Scott, “Metrics for generalized persistence modules”, *Found. Comput. Math.* **15**:6 (2015), 1501–1531. MR Zbl
- [Buchet et al. 2016] M. Buchet, F. Chazal, S. Y. Oudot, and D. R. Sheehy, “Efficient and robust persistent homology for measures”, *Comput. Geom.* **58** (2016), 70–96. MR Zbl
- [Cireşan et al. 2012] D. Cireşan, U. Meier, and J. Schmidhuber, “Multi-column deep neural networks for image classification”, pp. 3642–3649 in *2012 IEEE Conference on Computer Vision and Pattern Recognition* (Providence, RI, 2012), IEEE, Piscataway, NJ, 2012.
- [Cohen-Steiner et al. 2007] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer, “Stability of persistence diagrams”, *Discrete Comput. Geom.* **37**:1 (2007), 103–120. MR Zbl
- [Edelsbrunner and Harer 2010] H. Edelsbrunner and J. L. Harer, *Computational topology: an introduction*, Amer. Math. Soc., Providence, RI, 2010. MR Zbl
- [Edelsbrunner and Morozov 2013] H. Edelsbrunner and D. Morozov, “Persistent homology: theory and practice”, pp. 31–50 in *European Congress of Mathematics* (Kraków, 2012), edited by R. Latała et al., Eur. Math. Soc., Zürich, 2013. MR Zbl
- [Hatcher 2002] A. Hatcher, *Algebraic topology*, Cambridge Univ. Press, 2002. MR Zbl
- [Lawson 2016] A. Lawson, *Multi-scale persistent homology*, Ph.D. thesis, University of North Carolina at Greensboro, 2016, available at <https://search.proquest.com/docview/1806825235>.
- [Martin 2016] J. Martin, *Multiradial (multi)filtrations and persistent homology*, master’s thesis, University of North Carolina at Greensboro, 2016, available at <https://search.proquest.com/docview/1816997091>.
- [Matoušek 2002] J. Matoušek, *Lectures on discrete geometry*, Graduate Texts in Math. **212**, Springer, 2002. MR Zbl

- [Mischaikow and Nanda 2013] K. Mischaikow and V. Nanda, “Morse theory for filtrations and efficient computation of persistent homology”, *Discrete Comput. Geom.* **50**:2 (2013), 330–353. MR Zbl
- [Petri et al. 2013] G. Petri, M. Scolamiero, I. Donato, and F. Vaccarino, “Topological strata of weighted complex networks”, *PLOS One* **8**:6 (2013), art. id. e66506.
- [Ren et al. 2017] S. Ren, C. Wu, and J. Wu, “Computational tools in weighted persistent homology”, preprint, 2017. arXiv
- [Ren et al. 2018] S. Ren, C. Wu, and J. Wu, “Weighted persistent homology”, *Rocky Mountain J. Math.* **48**:8 (2018), 2661–2687. MR Zbl
- [Rotman 1988] J. J. Rotman, *An introduction to algebraic topology*, Graduate Texts in Math. **119**, Springer, 1988. MR Zbl
- [de Silva and Ghrist 2007] V. de Silva and R. Ghrist, “Coverage in sensor networks via persistent homology”, *Algebr. Geom. Topol.* **7** (2007), 339–358. MR Zbl
- [Zomorodian and Carlsson 2005] A. Zomorodian and G. Carlsson, “Computing persistent homology”, *Discrete Comput. Geom.* **33**:2 (2005), 249–274. MR Zbl

Received: 2018-06-09

Revised: 2018-09-20

Accepted: 2018-11-29

gcbell@uncg.edu

Department of Mathematics and Statistics, The University of North Carolina at Greensboro, Greensboro, NC, United States

azlawson@uncg.edu

Department of Mathematics and Statistics, The University of North Carolina at Greensboro, Mount Airy, NC, United States

jmmart27@uncg.edu

Department of Mathematics and Statistics, The University of North Carolina at Greensboro, Greensboro, NC, United States

jerudzin@uncg.edu

Department of Mathematics and Statistics, The University of North Carolina at Greensboro, Greensboro, NC, United States

cdsmlyth@uncg.edu

Department of Mathematics and Statistics, The University of North Carolina at Greensboro, Greensboro, NC, United States





# Leibniz algebras with low-dimensional maximal Lie quotients

William J. Cook, John Hall, Vicky W. Klima and Carter Murray

(Communicated by Ravi Vakil)

Every Leibniz algebra has a maximal homomorphic image that is a Lie algebra. We classify cyclic Leibniz algebras over an arbitrary field. Such algebras have the 1-dimensional abelian Lie algebra as their maximal Lie quotient. We then give examples of Leibniz algebras whose associated maximal Lie quotients exhaust all 2-dimensional possibilities.

## 1. Introduction

The theory of Leibniz algebras has blossomed since the pioneering work [Loday 1993]. Transitioning from Lie to Leibniz algebras is similar to transitioning from commutative to noncommutative rings. Both transitions drop one defining property, leading to many new and interesting structures. In a Leibniz algebra we keep a version of the Jacobi identity but no longer assume that multiplication is alternating, and hence it is not necessarily skew-symmetric either. To truly understand an algebraic structure one needs a varied collection of illuminating examples. In this paper we seek to provide a small collection of examples of non-Lie (left) Leibniz algebras.

In [Scofield and Sullivan 2014] the authors provide a classification of cyclic Leibniz algebras over the complex field. We offer a variant of their proof which avoids the use of  $n$ -th roots and thus provides a complete classification of cyclic Leibniz algebras over arbitrary fields. In addition, we construct two classes of non-cyclic Leibniz algebras with nonisomorphic 2-dimensional maximal Lie quotients, exhausting all possibilities for such quotients.

The paper is structured as follows: after providing some background in Section 2, we use Section 3 to construct and classify all cyclic Leibniz algebras over an arbitrary field. The next two sections present examples of Leibniz algebras with both nonabelian (Section 4) and abelian (Section 5) 2-dimensional maximal Lie quotients.

---

*MSC2010:* primary 17A32; secondary 17A60.

*Keywords:* Leibniz algebra, cyclic Leibniz algebra, low-dimensional examples.

## 2. Background

Let  $\mathbb{F}$  be a field. For our purposes it suffices to consider only finite-dimensional vector spaces over  $\mathbb{F}$ .

**Definition 2.1.** Let  $L$  be a vector space equipped with a bilinear map  $[\cdot, \cdot] : L \times L \rightarrow L$ , called a *bracket*, such that for all  $x, y, z \in L$  the (left) *Leibniz identity*  $[x, [y, z]] = [[x, y], z] + [y, [x, z]]$  holds. Then  $L$  is called a (left) *Leibniz algebra*.

Briefly, a (left) Leibniz algebra is an algebra whose left multiplication operators are derivations. Similarly we could assume that right multiplication operators are derivations and define the notion of a right Leibniz algebra. Just as with many other algebraic constructions our choice of left versus right is arbitrary. All of our results for left Leibniz algebras can easily be translated to results for right Leibniz algebras. For the remainder of the paper Leibniz algebra will mean left Leibniz algebra.

Notice that the Leibniz identity could replace the Jacobi identity in the definition of a Lie algebra. In fact, the left Leibniz identity, the corresponding right Leibniz identity  $[[y, z], x] = [y, [z, x]] + [[y, x], z]$ , and the Jacobi identity  $[[x, y], z] + [[y, z], x] + [[z, x], y] = 0$  are all equivalent if we assume our bracket is bilinear and alternating, that is,  $[x, x] = 0$  for all  $x$ . We refer the reader to [Demir et al. 2014] for more details concerning basic definitions related to Leibniz algebras.

**Definition 2.2.** For  $L$  a Leibniz algebra,  $\text{Leib}(L) = \text{span}_{\mathbb{F}}\{[x, x] \mid x \in L\}$ .

We have that  $L$  is a Lie algebra if and only if  $\text{Leib}(L) = \{0\}$ . Notice that  $\text{Leib}(L)$  is a (two-sided) ideal of  $L$ . Moreover,  $L/\text{Leib}(L)$  is the largest quotient of  $L$  that is a Lie algebra. Specifically, if  $I$  is any ideal of  $L$  such that  $L/I$  is a Lie algebra, then  $\text{Leib}(L) \subseteq I$ . Here we use the term *ideal* in the familiar Lie algebra sense: a subalgebra  $I$  of a Leibniz algebra  $L$  is a (two-sided) ideal of  $L$  if and only if  $[L, I]$  and  $[I, L]$  are both contained in  $I$ . We write  $I \triangleleft L$  when  $I$  is an ideal of  $L$ .

Many other definitions extend directly from Lie to Leibniz algebras. As a second example, we say  $L$  is an *abelian* Leibniz algebra if and only if  $[L, L] = \{0\}$ , that is, if  $[x, y] = 0$  for all  $x, y \in L$ . The definitions of nilpotency and solvability also carry over without modification.

**Definition 2.3.** Recall that  $L^1 = L$  and  $L^{j+1} = [L, L^j]$  for  $j \geq 1$  gives us the *lower central series*.  $L$  is *nilpotent* of class  $n$  if  $L^{n+1} = \{0\}$  but  $L^n \neq \{0\}$ . In particular,  $L$  is *nilpotent* if  $L^n = \{0\}$  for some  $n \geq 1$ . Likewise,  $L^{(0)} = L$  and  $L^{(j+1)} = [L^{(j)}, L^{(j)}]$  for  $j \geq 0$  gives us the *derived series*.  $L$  is *solvable* if  $L^{(n)} = \{0\}$  for some  $n \geq 0$ .

The proofs of many basic results given in introductory Lie algebra texts such as [Erdmann and Wildon 2006] apply just as well to Leibniz algebras. In particular, abelian implies nilpotent and nilpotent implies solvable. Recall that  $\text{rad}(L)$  is the largest solvable ideal of  $L$ . As with Lie algebras, this is just the sum of all ideals  $I$

of  $L$  such that  $I$  itself is a solvable algebra. Likewise,  $\text{nil}(L)$  is the largest nilpotent ideal.

The notion of internal direct sum for Leibniz algebras also carries over from Lie theory. As with Lie algebras, if  $L = L_1 \oplus \cdots \oplus L_n$  is an internal direct sum of Leibniz algebras, each  $L_i$  is in fact an ideal of  $L$  and  $L$  is isomorphic to the external direct sum of Leibniz algebras  $L_1, \dots, L_n$ , defined in the obvious way.

**Definition 2.4.** Let  $L$  be a Leibniz algebra with subalgebras  $L_1, \dots, L_n$ . We write  $L = L_1 \oplus \cdots \oplus L_n$ , an internal direct sum of Leibniz algebras, if  $L = L_1 \oplus \cdots \oplus L_n$  as subspaces and  $[x, y] = 0$  for any  $x \in L_i$  and  $y \in L_j$ , where  $i \neq j$ .

It is not hard to show that for  $I_j \triangleleft L_j$ , we have

$$(L_1 \oplus \cdots \oplus L_n)/(I_1 \oplus \cdots \oplus I_n) \cong (L_1/I_1) \oplus \cdots \oplus (L_n/I_n)$$

with the direct sum on the right an external direct sum. Likewise,

$$Z(L_1 \oplus \cdots \oplus L_n) = Z(L_1) \oplus \cdots \oplus Z(L_n),$$

$$\text{Leib}(L_1 \oplus \cdots \oplus L_n) = \text{Leib}(L_1) \oplus \cdots \oplus \text{Leib}(L_n),$$

$$[L_1 \oplus \cdots \oplus L_n, L_1 \oplus \cdots \oplus L_n] = [L_1, L_1] \oplus \cdots \oplus [L_n, L_n].$$

Some important definitions from Lie theory require minor modifications as we move to Leibniz algebras. For example, if we apply the Lie theory definitions of simple and semisimple algebras directly to Leibniz algebras, both simple and semisimple Leibniz algebra would necessarily be Lie and thus there would be nothing new to consider. We modify these definitions for Leibniz algebras as follows:

**Definition 2.5.** Let  $L$  be a Leibniz algebra.  $L$  is *simple* if and only if  $[L, L] \neq \text{Leib}(L)$  and  $\{0\}$ ,  $\text{Leib}(L)$ , and  $L$  are the only ideals of  $L$ .  $L$  is *semisimple* if and only if  $\text{rad}(L) = \text{Leib}(L)$ .

When  $L$  is also a Lie algebra,  $\text{Leib}(L) = \{0\}$ , so these definitions collapse back down to the usual definitions for a Lie algebra. In fact, these definitions guarantee that  $L$  is simple (resp. semisimple) as a Leibniz algebra if and only if  $L/\text{Leib}(L)$  is simple (resp. semisimple) as a Lie algebra.

When working with Lie algebras, taking powers of elements is uninteresting:  $x^1 = x$  and then  $x^2 = [x, x] = 0$  because of the alternating axiom. In Leibniz algebras much more is possible. We fix the notation  $x^1 = x$ ,  $x^2 = [x, x]$ , and in general,  $x^{n+1} = [x, x^n]$  for  $n \geq 1$ . Consider the following basic, well known result:

**Lemma 2.6.** Let  $L$  be a Leibniz algebra and  $x, y \in L$ . Then  $[[x, x], y] = 0$  and more generally  $[x^n, y] = 0$  for all  $n \geq 2$ . Moreover, the only potentially nonzero  $n$ -th power of  $x$  is

$$x^n = \underbrace{[x, [x, \dots, [x, x] \cdots ]]}_{n \text{ times}}.$$

*Proof.* The Leibniz identity states that  $[x, [x, y]] = [[x, x], y] + [x, [x, y]]$  so that  $0 = [[x, x], y]$ . Assume inductively that  $[x^n, z] = 0$  for any  $z \in L$  and some  $n \geq 2$ . The Leibniz identity states that  $[x, [x^n, y]] = [[x, x^n], y] + [x^n, [x, y]]$ . By our inductive hypothesis, we have  $[x, 0] = [x^{n+1}, y] + 0$  so that  $[x^{n+1}, y] = 0$ .

Finally, the only first and second powers of  $x$  are  $x^1 = x$  and  $x^2 = [x, x]$ . Third powers of  $x$  can be written either as  $x^3$  or  $[[x, x], x] = 0$ . Assume that all  $k$ -th powers of  $x$  other than  $x^k$  are 0 where  $1 \leq k < n$  and let  $w$  be some  $n$ -th power of  $x$ . Then  $w = [u, v]$ , where  $u$  and  $v$  are  $k$ -th and  $\ell$ -th powers of  $x$  such that  $k + \ell = n$ . By induction, if  $u \neq 0$  and  $v \neq 0$ , we must have  $u = x^k$  and  $v = x^\ell$ . So either  $k \geq 2$  and thus  $w = [u, v] = [x^k, v] = 0$  or  $k = 1$  and we have  $w = [u, v] = [x, x^\ell] = x^{\ell+1} = x^n$ .  $\square$

We can see that generally Leibniz algebras are not power associative. Notice that for a right Leibniz algebra we would have that the only potentially nonzero powers would be of the form  $[[\cdots [x, x], \dots, x], x]$ . This means that if an algebra was both a left and right Leibniz algebra, the only nonzero power could be  $x^2 = [x, x]$ . In fact,  $L = \text{span}_{\mathbb{F}}\{x, x^2\}$ , where  $[x, x] = x^2$ ,  $[x, x^2] = [x^2, x] = [x^2, x^2] = 0$ , gives an example of a simultaneously left and right Leibniz algebra which is not a Lie algebra.

### 3. Cyclic Leibniz algebras

A cyclic Leibniz algebra is a Leibniz algebra that can be generated from a single element. We do not consider cyclic Lie algebras since the only cyclic Lie algebras are either the trivial algebra  $\{0\}$  or the 1-dimensional abelian Lie algebra. Scofield and Sullivan [2014] have classified complex cyclic Leibniz algebras. In this section, we give a similar construction which allows us to classify cyclic (left) Leibniz algebras over an arbitrary field.

**Definition 3.1.** Let  $L$  be a Leibniz algebra.  $L$  is *cyclic* if and only if there exists some  $x \in L$  such that  $L = \langle x \rangle = \text{span}_{\mathbb{F}}\{x^k \mid k = 1, 2, \dots\}$ . If  $L = \langle x \rangle$ , we call  $x$  a *generator* of  $L$ .

The trivial algebra  $\{0\} = \langle 0 \rangle$  is cyclic. Likewise, any 1-dimensional algebra is cyclic as it is generated by any nonzero element.

Let  $L \neq \{0\}$  be a cyclic (left) Leibniz algebra and fix a generator  $x \neq 0$ . By definition  $L = \langle x \rangle = \{x^k \mid k = 1, 2, \dots\}$  and since  $L$  is finite-dimensional, we must have that  $\{x, x^2, \dots, x^{n+1}\}$  is linearly dependent for some  $n \geq 1$ . Let  $n$  be the smallest such power. This means that  $\{x, x^2, \dots, x^n\}$  is linearly independent and  $x^{n+1}$  can be written as a linear combination of  $\{x, \dots, x^n\}$ . Consequently all higher powers of  $x$  can be written as a linear combination of  $x, x^2, \dots, x^n$ . Thus  $\beta = \{x, x^2, \dots, x^n\}$  is a basis for  $L$  and so  $\dim(L) = n$ .

We have  $x^{n+1} \in L = \langle x \rangle = \text{span}_{\mathbb{F}}\{x, x^2, \dots, x^n\}$ . Let  $x^{n+1} = \sum_{i=1}^n c_i x^i$ , where  $c_i \in \mathbb{F}$ . When  $\dim(L) = n > 1$ , Lemma 2.6 guarantees  $0 = [x, 0] = [x, [x^n, x]]$ .

Applying the Leibniz identity and Lemma 2.6 once more yields

$$0 = [x, [x^n, x]] = [[x, x^n], x] + [x^n, x^2] = [x^{n+1}, x] + 0 = c_1 x^2 + \sum_{i=2}^n c_i [x^i, x] = c_1 x^2.$$

Since  $\dim(L) = n > 1$ , we conclude  $x^2 \neq 0$  and thus  $c_1 = 0$ . Therefore,  $x^{n+1} = \sum_{i=2}^n c_i x^i$ , a summation that does not involve  $i = 1$ .

It turns out that the necessary condition  $x^{n+1} = \sum_{i=2}^n c_i x^i$  for some  $c_2, \dots, c_n \in \mathbb{F}$  is also sufficient for any  $n$ -dimensional cyclic Leibniz algebra  $L = \langle x \rangle$ .

**Proposition 3.2.** *Fix  $n \geq 1$  and  $c_2, \dots, c_n \in \mathbb{F}$  and let  $L = \text{span}_{\mathbb{F}}\{x, x^2, \dots, x^n\}$  be an  $n$ -dimensional vector space. Define a bilinear operation on the basis  $\{x, x^2, \dots, x^n\}$  as follows:  $[x, x^j] = x^{j+1}$  for  $1 \leq j < n$ ,  $[x, x^n] = \sum_{i=2}^n c_i x^i$ , and  $[x^k, x^\ell] = 0$  for  $k \geq 2, 1 \leq \ell \leq n$ . Then  $L = \langle x \rangle$  is a cyclic Leibniz algebra.*

*Proof.* Clearly  $L$  is a cyclic algebra equipped with a bilinear operation. It just remains to verify the Leibniz identity. It is enough to do so on our basis. We note that when  $n = 1$ , we have  $x^{n+1} = x^2 = 0$  and the Leibniz identity is

$$[x, [x, x]] = [x, 0] = 0 = 0 + 0 = [0, x] + [x, 0] = [[x, x], x] + [x, [x, x]].$$

Assume  $n > 1$  and let  $1 \leq i, j, k \leq n$ .

If  $i \geq 2$ , then

$$[x^i, [x^j, x^k]] = 0 = 0 + 0 = [0, x^k] + [x^j, 0] = [[x^i, x^j], x^k] + [x^j, [x^i, x^k]].$$

If  $i = 1$  and  $j = 1$ , then

$$[x, [x, x^k]] = 0 + [x, [x, x^k]] = [x^2, x^k] + [x, [x, x^k]] = [[x, x], x^k] + [x, [x, x^k]].$$

If  $i = 1$  and  $2 \leq j < n$ , then

$$\begin{aligned} [x, [x^j, x^k]] &= [x, 0] = 0 = 0 + 0 \\ &= [x^{j+1}, x^k] + [x^j, x^{k+1}] = [[x, x^j], x^k] + [x^j, [x, x^k]]. \end{aligned}$$

If  $i = 1$  and  $j = n > 1$ , then

$$\begin{aligned} [x, [x^n, x^k]] &= [x, 0] = 0 \\ &= \sum_{m=2}^n c_m [x^m, x^k] = [x^{n+1}, x^k] + 0 = [[x, x^n], x^k] + [x^n, [x, x^k]]. \end{aligned}$$

Notice that here we used the fact that our sum begins at  $m = 2$  so  $[x^m, x^k] = 0$ .  $\square$

For  $n > 0$  fix a cyclic Leibniz algebra  $L$  with basis  $\beta = \{x, x^2, \dots, x^n\}$ . Next, we will further investigate the structure of this algebra by considering  $\text{Leib}(L)$  and the derived series of  $L$ . Note that by definition  $x^2 \in \text{Leib}(L)$ . But then since  $\text{Leib}(L)$  is an ideal of  $L$ ,  $x^j \in \text{Leib}(L)$  for all  $j \geq 2$ . Since brackets among

elements of  $L$  never result in an element involving  $x$  itself, we conclude  $\text{Leib}(L) = \text{span}\{x^2, x^3, \dots, x^n\} = [L, L]$ , an abelian Leibniz algebra of dimension  $n - 1$ . It quickly follows that the derived series for  $L$  is given by

$$L^{(0)} = L \supsetneq L^{(1)} = [L, L] = \text{span}\{x^2, x^3, \dots, x^n\} \supsetneq L^{(2)} = \{0\}.$$

The series goes to zero and thus cyclic Leibniz algebras are always solvable.

We next consider the lower central series of the cyclic Leibniz algebra  $L = \langle x \rangle$  with basis  $\beta = \{x, x^2, \dots, x^n\}$  and  $x^{n+1} = \sum_{i=2}^n c_i x^i$ . First consider the case when  $x^{n+1} = 0$ , that is, when  $c_2 = c_3 = \dots = c_n = 0$ . Then keeping in mind that only left multiplication by  $x$  can yield a nonzero result, we have

$$\begin{aligned} [L, \text{span}\{x^m, x^{m+1}, \dots, x^n\}] &= \text{span}\{[x, x^m], [x, x^{m+1}], \dots, [x, x^n]\} \\ &= \text{span}\{x^{m+1}, \dots, x^n\}. \end{aligned}$$

This means that  $L^j = \text{span}\{x^j, \dots, x^n\}$  for  $1 \leq j \leq n$  and  $L^{n+1} = \{0\}$ . In other words,  $L$  is nilpotent of class  $n$ .

Next assume that  $x^{n+1} \neq 0$ . In particular, assume  $c_j = 0$  for all  $j < k$  and  $c_k \neq 0$ . Let  $1 \leq m \leq k$  and consider  $[L, \text{span}\{x^m, \dots, x^n\}]$ . Again, only left multiplication by  $x$  yields a nonzero result so that

$$[L, \text{span}\{x^m, \dots, x^n\}] = \text{span}\{x^{m+1}, \dots, x^n, x^{n+1}\}.$$

If  $m < k$ , we have  $x^{n+1} = \sum_{\ell=k}^n c_\ell x^\ell \in \text{span}\{x^{m+1}, \dots, x^n\}$  so that

$$[L, \text{span}\{x^m, \dots, x^n\}] = \text{span}\{x^{m+1}, \dots, x^n\}.$$

If  $m = k$ , we have  $x^{n+1} = c_k x^k + \sum_{\ell=k+1}^n c_\ell x^\ell$  with  $c_k \neq 0$ . Thus

$$\text{span}\{x^{m+1}, \dots, x^{n+1}\} = \text{span}\{x^{k+1}, \dots, x^{n+1}\} = \text{span}\{x^k, \dots, x^n\}$$

and in this case  $[L, \text{span}\{x^k, \dots, x^n\}] = \text{span}\{x^k, \dots, x^n\}$ . In particular,

$$[L, \text{span}\{x^m, \dots, x^n\}] = \text{span}\{x^{\min(k, m+1)}, \dots, x^n\}.$$

This means  $L^m = \text{span}\{x^m, \dots, x^n\}$  for  $1 \leq m < k$  and  $L^k = L^{k+1} = \dots = \text{span}\{x^k, \dots, x^n\}$ . Proposition 3.3 summarizes our findings.

**Proposition 3.3.** *Let  $L$  be an  $n$ -dimensional cyclic Leibniz algebra. Then either  $L$  is nilpotent of class  $n$  or  $L \supsetneq L^2 \supsetneq \dots \supsetneq L^k = L^{k+1} = \dots \neq \{0\}$  for some  $2 \leq k \leq n$ . In this case, we say that  $L$  is cyclic of type  $k$ . Moreover, let  $x$  be any generator for  $L$ . Then  $L$  is nilpotent if and only if  $x^{n+1} = 0$ . If  $L$  is not nilpotent and is of type  $k$ , then  $x^{n+1} = \sum_{\ell=k}^n c_\ell x^\ell$  for some  $c_k, \dots, c_n \in \mathbb{F}$  and  $c_k \neq 0$ . In particular, nilpotency and type do not depend on the choice of generator.*

As we turn our attention towards a classification of cyclic Leibniz algebras, again let  $L \neq \{0\}$  be an  $n$ -dimensional cyclic Leibniz algebra generated by  $x$  with basis  $\beta = \{x, x^2, \dots, x^n\}$  and  $x^{n+1} = \sum_{j=2}^n c_j x^j$ . Using an approach introduced in [Batten Ray et al. 2014], we consider the left multiplication operator  $\mathcal{L}_x : L \rightarrow L$  defined by  $\mathcal{L}_x(z) = [x, z]$ . We have  $\mathcal{L}_x(x^j) = x^{j+1}$  for  $1 \leq j < n$  and  $\mathcal{L}_x(x^n) = \sum_{j=2}^n c_j x^j$ . Thus we get the following coordinate matrix relative to the basis  $\beta$ :

$$[\mathcal{L}_x]_\beta = \begin{bmatrix} 0 & 0 & \cdots & \cdots & 0 & 0 & 0 \\ 1 & 0 & \cdots & \cdots & 0 & 0 & c_2 \\ \vdots & \ddots & \ddots & & \vdots & \vdots & \vdots \\ \vdots & & \ddots & \ddots & 0 & 0 & \vdots \\ 0 & & & \ddots & 1 & 0 & c_{n-1} \\ 0 & 0 & \cdots & \cdots & 0 & 1 & c_n \end{bmatrix}.$$

The matrix  $[\mathcal{L}_x]_\beta$  is the companion matrix to the polynomial

$$p(t) = t^n - c_n t^{n-1} - \cdots - c_2 t$$

and thus the linear operator  $\mathcal{L}_x$  has characteristic polynomial  $p(t)$ . Note that the polynomial  $p(t)$  is in direct correspondence with our defining relation for  $x^{n+1}$ .

Suppose that  $y = \sum_{i=1}^n b_i x^i \in L$ . Then

$$\mathcal{L}_y(x^j) = \left[ \sum_{i=1}^n b_i x^i, x^j \right] = \sum_{i=1}^n b_i [x^i, x^j] = b_1 [x, x^j] = b_1 x^{j+1}$$

since  $[x^i, x^j] = 0$  for  $i \geq 2$ . This means  $[\mathcal{L}_y]_\beta = b_1 [\mathcal{L}_x]_\beta$ . With only small, obvious modifications, the standard approach to determining the characteristic polynomial for a companion matrix, see, for example, [Hoffman and Kunze 1971, Theorem 1, page 228], shows that the matrix  $[\mathcal{L}_y]_\beta$ , and thus the linear operator  $\mathcal{L}_y$ , has characteristic polynomial

$$t^n - b_1 c_n t^{n-1} - b_1^2 c_{n-1} t^{n-2} - \cdots - b_1^{n-1} c_2 t.$$

Note that if  $y$  is a generator for  $L$ , using the correspondence between the characteristic polynomial of  $\mathcal{L}_y$  and our defining relation for  $y^{n+1}$ , we see  $y^{n+1} = \sum_{i=2}^n b_1^{n-i} c_i y^i$ .

In summary for  $n \geq 2$  and any  $(c_2, \dots, c_n) \in \mathbb{F}^{n-1}$  there is an  $n$ -dimensional cyclic Leibniz algebra  $L$  with generator  $x$  such that  $\{x, x^2, \dots, x^n\}$  is a basis for  $L$  and  $x^{n+1} = \sum_{j=2}^n c_j x^j$ . If  $y$  is any other generator with  $y = \sum_{i=1}^n b_i x^i$  then  $\{y, y^2, \dots, y^n\}$  is a basis for  $L$  and  $y^{n+1} = \sum_{j=2}^n b_1^{n-j} c_j y^j$ . For  $n \geq 2$ , define an equivalence relation on  $\mathbb{F}^{n-1}$  such that  $(c_2, \dots, c_n) \sim (b^{n-1} c_2, b^{n-2} c_3, \dots, b c_n)$  for any  $b \in \mathbb{F}$ . Denote the equivalence classes as  $[(c_2, \dots, c_n)]$ . This equivalence relation allows a simple classification of cyclic Leibniz algebras.

**Theorem 3.4.** *Up to isomorphism the only cyclic Leibniz algebras of dimensions 0 and 1 are the trivial  $\{0\}$  algebra and the 1-dimensional abelian Lie algebra. For  $n \geq 2$ , up to isomorphism there is exactly one  $n$ -dimensional cyclic Leibniz algebra associated with each equivalence class  $[(c_2, \dots, c_n)]$ , where  $(c_2, \dots, c_n) \in \mathbb{F}^{n-1}$ .*

The nilpotent cyclic Leibniz algebras are associated with the class  $[(0, \dots, 0)] = \{(0, \dots, 0)\}$ . Cyclic Leibniz algebras of type  $k$  are associated with the class  $[(0, \dots, 0, c_k, \dots, c_n)]$  for some  $c_k, \dots, c_n \in \mathbb{F}$  with  $c_k \neq 0$ . In this case,  $\dim(L^k) = n - k + 1$  and  $L^k = L^{k+1} = \dots$ .

The classification of complex cyclic Leibniz algebras obtained in [Scofield and Sullivan 2014] split isomorphism classes of cyclic Leibniz algebras into cases of nilpotent or type  $k$ . For algebras of type  $k$ , they insist on a normalized generator such that  $c_k = 1$ . Note that their equivalence class  $[(c_{k+1}, \dots, c_n)]$  corresponds to our class  $[(0, \dots, 0, 1, c_{k+1}, \dots, c_n)]$ . By avoiding this normalization we no longer need the existence of roots of unity and our equivalence relation is much simpler.

As in our construction, Batten Ray et al. [2014] identify the matrix for the left multiplication operator as a companion matrix to the polynomial  $p(t)$ . They use this observation as a tool to develop several important properties of cyclic Leibniz algebras. In particular they give a construction of the unique Cartan subalgebra for each cyclic Leibniz algebra,  $L$ , and in the process describe all maximal subalgebras of  $L$  as well as the minimal ideals of  $L$  and the unique maximal ideal of  $L$ .

#### 4. A class of non-Lie, noncyclic Leibniz algebras

In this section we introduce a class of noncyclic Leibniz algebras and study their properties. Fix some  $n \geq 1$  and let  $L$  be the  $(n+1)$ -dimensional vector space with basis  $\beta = \{x, x^2, \dots, x^n, y\}$ . To determine a bilinear operation on  $L$  it is enough to specify how multiplication works on basis elements.

**Example 4.1.** Let  $L$  be the algebra with basis  $\beta = \{x, x^2, \dots, x^n, y\}$  and the bilinear bracket defined on the basis elements as follows:

- (1)  $[x, x^j] = x^{j+1}$ ,  $1 \leq j < n$ .
- (2)  $[x, x^n] = x^{n+1} = 0$ .
- (3)  $[x^k, x^j] = [x^k, y] = 0$  for all  $2 \leq k \leq n$  and  $1 \leq j \leq n$ .
- (4)  $[x, y] = x$ ,  $[y, x^j] = -jx^j$  for  $1 \leq j \leq n$ .
- (5)  $[y, y] = 0$ .

To see that  $L$  is a Leibniz algebra, we need to verify that the Leibniz identity holds. First, notice that  $\langle x \rangle = \text{span}\{x, x^2, \dots, x^n\}$  forms an  $n$ -dimensional cyclic, nilpotent Leibniz subalgebra. Likewise,  $\langle y \rangle = \text{span}\{y\}$  forms a 1-dimensional cyclic



Leibniz subalgebra which is an abelian Lie algebra. Thus we only need to check the Leibniz identity among triples of basis elements which involve both  $x$  and  $y$ .

First, we consider triples that involve two occurrences of  $y$ :

- For  $1 \leq j \leq n$ ,

$$[y, [y, x^j]] = 0 + [y, [y, x^j]] = [0, x^j] + [y, [y, x^j]] = [[y, y], x^j] + [y, [y, x^j]].$$

- For  $2 \leq j \leq n$ ,

$$[x^j, [y, y]] = [x^j, 0] = 0 = 0 + 0 = [0, y] + [y, 0] = [[x^j, y], y] + [y, [x^j, y]],$$

and for  $j = 1$ ,

$$[x, [y, y]] = [x, 0] = 0 = x - x = [x, y] + [y, x] = [[x, y], y] + [y, [x, y]].$$

- For  $2 \leq j \leq n$ ,

$$[y, [x^j, y]] = [y, 0] = 0 = 0 + 0 = [0, y] + [x^j, 0] = [[y, x^j], y] + [x^j, [y, y]],$$

and for  $j = 1$ ,

$$[y, [x, y]] = [y, x] = -x = -[x, y] + 0 = [-x, y] + [x, 0] = [[y, x], y] + [x, [y, y]].$$

Finally, we consider triples that involve one occurrence of  $y$ :

- Note that  $[y, x^j] = -jx^j$  holds even when  $j = n+1$  since  $x^{n+1} = 0$ . Let  $1 \leq k \leq n$ .

- For  $2 \leq j \leq n$ ,

$$[y, [x^j, x^k]] = [y, 0] = 0 = -j[x^j, x^k] = [-jx^j, x^k] + 0 = [[y, x^j], x^k] + [x^j, [y, x^k]],$$

and for  $j = 1$ ,

$$\begin{aligned} [y, [x, x^k]] &= [y, x^{k+1}] = -(k+1)x^{k+1} \\ &= [-x, x^k] + [x, -kx^k] = [[y, x], x^k] + [x, [y, x^k]]. \end{aligned}$$

- For  $2 \leq j \leq n$ ,

$$[x^j, [y, x^k]] = 0 = 0 + 0 = [0, x^k] + [y, 0] = [[x^j, y], x^k] + [y, [x^j, x^k]],$$

and for  $j = 1$ ,

$$\begin{aligned} [x, [y, x^k]] &= [x, -kx^k] = -kx^{k+1} = x^{k+1} - (k+1)x^{k+1} \\ &= [x, x^k] + [y, x^{k+1}] = [[x, y], x^k] + [y, [x, x^k]]. \end{aligned}$$

- For  $2 \leq j \leq n$ ,

$$[x^j, [x^k, y]] = 0 = 0 + [x^k, 0] = [[x^j, x^k], y] + [x^k, [x^j, y]],$$

and for  $j = 1$  and  $k \geq 2$ ,

$$[x, [x^k, y]] = [x, 0] = 0 = [x^{k+1}, y] + 0 = [[x, x^k], y] + [x^k, [x, y]].$$

When  $j = k = 1$ ,

$$[x, [x, y]] = 0 + [x, [x, y]] = [[x, x], y] + [x, [x, y]].$$

We use the remainder of this section to investigate the structure of the Leibniz algebra  $L$  described in Example 4.1. Let us begin by determining the lower central series of  $L$ ,  $\text{Leib}(L)$ , and the derived series for  $L$ . Since none of the brackets output a  $y$ ,  $[L, L]$  must be contained in  $\langle x \rangle = \text{span}\{x, x^2, \dots, x^n\}$ . We have seen that  $[-y, x] = x \in [L, L]$  and therefore  $\langle x \rangle \subseteq [L, L]$  and hence  $L^2 = [L, L] = \langle x \rangle$ . In fact, it follows by induction that  $L^k = \langle x \rangle$  for  $k \geq 2$ . We then have the lower central series

$$L = \text{span}\{x, x^2, \dots, x^n, y\} \supsetneq L^2 = L^3 = \dots = \text{span}\{x, x^2, \dots, x^n\} \neq \{0\},$$

and thus  $L$  is not nilpotent.

Next observe  $B = \text{span}\{x^j \mid j \geq 2\}$  is an abelian ideal of codimension 2 in  $L$  so that  $B \subseteq \text{Leib}(L)$ . Also,  $L/B$  is a Lie algebra and thus  $\text{Leib}(L) \subseteq B$ . Therefore  $\text{Leib}(L) = B = \text{span}\{x^j \mid j \geq 2\}$ . Furthermore, since

$$[x + \text{Leib}(L), y + \text{Leib}(L)] = [x, y] + \text{Leib}(L) = x + \text{Leib}(L),$$

we have that  $L/\text{Leib}(L)$  is the nonabelian 2-dimensional Lie algebra. In addition, the derived series is given by

$$L^{(0)} = L \supsetneq L^{(1)} = \langle x \rangle \supsetneq L^{(2)} = \text{Leib}(L) = \text{span}\{x^j \mid j \geq 2\} \supsetneq L^{(3)} = \{0\}$$

and thus  $L$  is solvable.

Could it be that  $L$  is simply a sum of cyclic Leibniz algebras? Recall that for a cyclic Leibniz algebra  $C$ ,  $C/\text{Leib}(C)$  is the 1-dimensional abelian Lie algebra. Thus if  $M = C_1 \oplus \dots \oplus C_\ell$  is a Leibniz algebra direct sum of cyclic Leibniz algebras  $C_1, \dots, C_\ell$ , then

$$\begin{aligned} M/\text{Leib}(M) &= (C_1 \oplus \dots \oplus C_\ell)/(\text{Leib}(C_1) \oplus \dots \oplus \text{Leib}(C_\ell)) \\ &\cong (C_1/\text{Leib}(C_1)) \oplus \dots \oplus (C_\ell/\text{Leib}(C_\ell)) \end{aligned}$$

and so  $M/\text{Leib}(M)$  is a direct sum of 1-dimensional abelian Lie algebras. In other words,  $M/\text{Leib}(M)$  is the  $\ell$ -dimensional abelian Lie algebra. Since  $L/\text{Leib}(L)$  is not abelian,  $L$  is neither cyclic nor a (Leibniz algebra) direct sum of cyclic subalgebras.

Also, since  $L$  is solvable,  $L = \text{rad}(L)$  and so  $L$  is (unsurprisingly) not semisimple. Additionally,  $\text{span}\{x^m, x^{m+1}, \dots, x^n\}$  for  $1 \leq m \leq n$  are easily seen to be ideals. In particular,  $\text{span}\{x, x^2, \dots, x^n\}$  is an ideal distinct from  $\{0\}$ ,  $\text{Leib}(L)$ , and  $L$  so that  $L$  is not simple. In summary:

**Theorem 4.2.** *The Leibniz algebra  $L = \text{span}\{x, x^2, \dots, x^n, y\}$  with bracket structure given in Example 4.1 is not nilpotent, semisimple, or simple. But  $L$  is solvable. Its maximal Lie algebra homomorphic image,  $L/\text{Leib}(L)$ , is the nonabelian 2-dimensional Lie algebra. Consequently  $L$  is not a (Leibniz algebra) direct sum of cyclic Leibniz algebras.*

## 5. Adjoining a module

In this section we offer a second class of examples. By first extending the familiar Lie algebra construction of adjoining a module to an algebra to the context of Leibniz algebras and then considering adjoining a cyclic module to a nilpotent cyclic Leibniz algebra, we obtain a class of algebras with similar properties to those of the previous section except here we will have that the maximal Lie algebra homomorphic image is abelian.

**Definition 5.1.** Let  $L$  be a Leibniz algebra and  $M$  a vector space over  $\mathbb{F}$  equipped with bilinear maps  $[\cdot, \cdot]: L \times M \rightarrow M$  and  $[\cdot, \cdot]: M \times L \rightarrow M$  (a left and a right action) such that for all  $a, b \in L$  and  $m \in M$  the following hold:

- (1)  $[a, [b, m]] = [[a, b], m] + [b, [a, m]].$
- (2)  $[a, [m, b]] = [[a, m], b] + [m, [a, b]].$
- (3)  $[m, [a, b]] = [[m, a], b] + [a, [m, b]].$

We note that if  $L$  is a Lie algebra with  $L$ -module  $M$  and action  $x \cdot m$  for  $x \in L$  and  $m \in M$ , then the left action  $[x, m] = x \cdot m$  and the right action  $[m, x] = -x \cdot m$  turn  $M$  into a module viewing  $L$  as merely a Leibniz algebra.

**Example 5.2.** Let  $L = \text{span}\{x, x^2, \dots, x^n\}$  be the  $n$ -dimensional nilpotent cyclic Leibniz algebra. Consider the vector space  $M = \text{span}(\beta)$  with basis  $\beta = \{y_1, y_2, \dots, y_n\}$ . Let  $2 \leq j \leq n$  and  $1 \leq k \leq n$  and define  $[x^j, y_k] = 0$ . When  $k < n$ , define  $[x, y_k] = y_{k+1}$  and let  $[x, y_n] = 0$ . For convenience let  $y_{n+1} = 0$  so that  $[x, y_k] = y_{k+1}$  for all  $1 \leq k \leq n$ . Finally, let  $[y_k, x^j] = 0$  for all  $1 \leq j \leq n$  and  $1 \leq k \leq n$ . In other words, the right action of  $L$  on  $M$  is trivial, whereas  $x$  acts in cyclic fashion on the left.

With these definitions,  $M$  is an  $L$ -module. To see this we must verify the relations in Definition 5.1. In relation (1), all terms are zero unless  $a = b = x$ . In this case relation (1) becomes  $[x, [x, m]] = [[x, x], m] + [x, [x, m]]$ , which is clearly true since  $[[x, x], m] = [x^2, m] = 0$ . Relations (2) and (3) hold because all terms are zero as they each involve the trivial right action of  $L$ .

We show in the following proposition that for  $L$  a Leibniz algebra and  $M$  an  $L$ -module, the vector space direct sum  $L \oplus M$  becomes a Leibniz algebra if for  $x_1, x_2 \in L$  and  $m_1, m_2 \in M$  we define  $[x_1 + m_1, x_2 + m_2] = [x_1, x_2] + [x_1, m_2] + [m_1, x_2]$ . Notice that in the definition of the bracket on  $L \oplus M$ ,  $[x_1, x_2]$  is the bracket in  $L$ ,  $[x_1, m_2]$  is the left action of  $L$  on  $M$ , and  $[m_1, x_2]$  is the right action of  $L$  on  $M$ .

**Proposition 5.3.** *Let  $L$  be a Leibniz algebra and  $M$  an  $L$ -module. The vector space direct sum  $L \oplus M$  becomes a Leibniz algebra if for  $x_1, x_2 \in L$  and  $m_1, m_2 \in M$  we define  $[x_1 + m_1, x_2 + m_2] = [x_1, x_2] + [x_1, m_2] + [m_1, x_2]$ . Moreover,  $L$  is a subalgebra and  $M$  is an abelian ideal of  $L \oplus M$ .*

*Proof.* It is obvious that the bracket on  $L \oplus M$  is bilinear. We need to verify the Leibniz identity. Let  $x_1, x_2, x_3 \in L$  and  $m_1, m_2, m_3 \in M$ . Consider the following brackets:

$$\begin{aligned}
 & \underbrace{[x_1 + m_1, [x_2 + m_2, x_3 + m_3]]}_{\text{LM}_A} \\
 &= [x_1 + m_1, [x_2, x_3] + [x_2, m_3] + [m_2, x_3]] \\
 &= \underbrace{[x_1, [x_2, x_3]]}_{\text{Leibniz}_A} + \underbrace{[x_1, [x_2, m_3]]}_{1_A} + \underbrace{[x_1, [m_2, x_3]]}_{2_A} + \underbrace{[m_1, [x_2, x_3]]}_{3_A}, \\
 & \underbrace{[[x_1 + m_1, x_2 + m_2], x_3 + m_3]}_{\text{LM}_B} \\
 &= [[x_1, x_2], x_3 + m_3] + [[x_1, m_2], x_3 + m_3] + [[m_1, x_2], x_3 + m_3] \\
 &= \underbrace{[[x_1, x_2], x_3]}_{\text{Leibniz}_B} + \underbrace{[[x_1, x_2], m_3]}_{1_B} + \underbrace{[[x_1, m_2], x_3]}_{2_B} + \underbrace{[[m_1, x_2], x_3]}_{3_B}, \\
 & \underbrace{[x_2 + m_2, [x_1 + m_1, x_3 + m_3]]}_{\text{LM}_C} \\
 &= [x_2 + m_2, [x_1, x_3]] + [x_2 + m_2, [x_1, m_3]] + [x_2 + m_2, [m_1, x_3]] \\
 &= \underbrace{[x_2, [x_1, x_3]]}_{\text{Leibniz}_C} + \underbrace{[m_2, [x_1, x_3]]}_{2_C} + \underbrace{[x_2, [x_1, m_3]]}_{1_C} + \underbrace{[x_2, [m_1, x_3]]}_{3_C}.
 \end{aligned}$$

The module axioms 1, 2, and 3 for  $M$  guarantee that  $1_A = 1_B + 1_C$ ,  $2_A = 2_B + 2_C$ , and  $3_A = 3_B + 3_C$ . The Leibniz identity for  $L$  guarantees that  $\text{Leibniz}_A = \text{Leibniz}_B + \text{Leibniz}_C$ . Putting these together we see that  $\text{LM}_A = \text{LM}_B + \text{LM}_C$  and so the Leibniz identity holds on  $L \oplus M$ .  $\square$

Taking  $L$  and  $M$  as defined in Example 5.2, let

$$K = L \oplus M = \text{span}\{x, x^2, \dots, x^n, y_1, \dots, y_n\}.$$

We have that  $K$  is a Leibniz algebra using the above construction and can now investigate the structure of this algebra.

For  $x \in L$  and  $m \in M$ , we have  $[x + m, x + m] = [x, x] + [x, m] + [m, x]$ . Therefore,

$$\text{Leib}(L \oplus M) = \text{Leib}(L) \oplus \text{span}\{[x, m] + [m, x] \mid x \in L \text{ and } m \in M\},$$

where  $\oplus$  represents a vector space direct sum. Furthermore, we know that  $\text{Leib}(L) = \text{span}\{x^2, \dots, x^n\}$  and all brackets (i.e., actions) between  $L$  and  $M$  either output 0 or something in  $\text{span}\{y_2, \dots, y_n\}$ . In fact,

$$[x, y_k] + [y_k, x] = y_{k+1} + 0 = y_{k+1} \in \text{span}\{[x, m] + [m, x] \mid x \in L \text{ and } m \in M\}$$

for  $1 \leq k \leq n$ . Therefore,  $\text{Leib}(K) = \text{span}\{x^2, \dots, x^n, y_2, \dots, y_n\}$ .

Next we explicitly calculate the lower central series for  $K$ . First, looking at the brackets for  $K$  we see that they never output any power of  $x$  smaller than  $x^2$  and never output  $y_1$ . Thus  $[K, K] \subseteq \text{span}\{x^2, \dots, x^n, y_2, \dots, y_n\}$ . But by definition,  $\text{Leib}(K) \subseteq [K, K]$ . Therefore,  $[K, K] = \text{Leib}(K) = \text{span}\{x^2, \dots, x^n, y_2, \dots, y_n\}$ . We claim that  $K^\ell = \text{span}\{x^\ell, \dots, x^n, y_\ell, \dots, y_n\}$  for  $1 \leq \ell \leq n$  and  $\{0\} = K^{n+1} = K^{n+2} = \dots$  so that  $K$  is nilpotent of class  $n$ . We proceed by induction; notice that  $[x, K^\ell] = \text{span}\{x^{\ell+1}, \dots, x^{n+1}, y_{\ell+1}, \dots, y_{n+1}\}$ , where for convenience we let  $x^m = y_m = 0$  for  $m > n$ . Also,  $[x^j, K^\ell] = [y, K^\ell] = \{0\}$  for  $j \geq 2$ . The result follows and from it we observe that  $L$  is nilpotent.

Note that we could forgo the explicit construction of the lower central series and still arrive at the nilpotency of  $K$  by applying a theorem of [Bosko et al. 2011]. Every left multiplication by an element of  $L$  on  $K$  is nilpotent and trivially left multiplication on  $K$  by elements from  $M$  is nilpotent. Therefore since  $L \cup M$  is a Lie set (i.e., it is closed under brackets and spans  $K$ ), Jacobson's refinement of Engel's theorem for Leibniz algebras [Bosko et al. 2011] shows  $K = L \oplus M$  is nilpotent.

Next we examine the structure of the cyclic subalgebras of  $K$ . Let

$$z = \sum_{i=1}^n a_i x^i + \sum_{j=1}^n b_j y_j \in K.$$

Then

$$\begin{aligned} z^2 = [z, z] &= a_1 \sum_{i=1}^{n-1} a_i x^{i+1} + a_1 \sum_{j=1}^{n-1} b_j y_{j+1} = \sum_{i=2}^n a_1 a_{i-1} x^i + \sum_{j=2}^n a_1 b_{j-1} y_j, \\ z^3 = [z, z^2] &= a_1 \sum_{i=2}^{n-1} a_1 a_{i-1} x^{i+1} + a_1 \sum_{j=2}^{n-1} a_1 b_{j-1} y_{j+1} = \sum_{i=3}^n a_1^2 a_{i-2} x^i + \sum_{j=3}^n a_1^2 b_{j-2} y_j. \end{aligned}$$

In general,

$$z^\ell = \sum_{i=\ell}^n a_1^{\ell-1} a_{i-\ell+1} x^i + \sum_{j=\ell}^n a_1^{\ell-1} b_{j-\ell+1} y_j \quad \text{for } 1 \leq \ell \leq n \text{ and } z^\ell = 0 \text{ for } \ell > n.$$

As a consequence, if  $a_1 = 0$ , then  $z^2 = 0$ . If  $a_1 \neq 0$  and  $1 \leq \ell \leq n$  then the coefficient of  $x^\ell$  in  $z^\ell$  is  $a_1^{\ell-1} a_{\ell-\ell+1} = a_1^{\ell-1} \neq 0$ . In all cases  $z^{n+1} = 0$  and thus by Proposition 3.3 all cyclic subalgebras,  $\langle z \rangle$ , are nilpotent. For  $n > 1$  they are either trivial ( $z = 0$ ), 1-dimensional ( $z \neq 0$  but  $a_1 = 0$ ), or  $n$ -dimensional ( $a_1 \neq 0$ ). For  $n = 1$ , they are

either trivial or 1-dimensional. Our understanding of the cyclic subalgebras of  $K$  plays a key role in understanding the structure of this Leibniz algebra.

**Theorem 5.4.** *The Leibniz algebra  $K = \text{span}\{x, x^2, \dots, x^n, y_1, y_2, \dots, y_n\}$  with brackets given in Example 5.2 and Proposition 5.3 is neither semisimple nor simple. But  $K$  is nilpotent of class  $n$  and solvable. Its maximal Lie algebra homomorphic image,  $K/\text{Leib}(K)$ , is the 2-dimensional abelian Lie algebra. Also, for  $n > 1$ ,  $K$  is not a (Leibniz algebra) direct sum of cyclic Leibniz algebras.*

*Proof.* We have already seen that  $K$  is nilpotent. Since  $K$  is nilpotent, it is also solvable. Referring back to definitions, it is obvious that  $K$  is neither simple nor semisimple. By definition,

$$K/\text{Leib}(K) = \text{span}\{x + \text{Leib}(K), y_1 + \text{Leib}(K)\}.$$

Notice that

$$[x + \text{Leib}(K), y_1 + \text{Leib}(K)] = [x, y_1] + \text{Leib}(K) = y_2 + \text{Leib}(K) = 0 + \text{Leib}(K),$$

since  $y_2 \in \text{Leib}(K)$ . Hence  $K/\text{Leib}(K)$  is the 2-dimensional abelian Lie algebra.

Suppose that  $K$  is a (Leibniz algebra) direct sum of cyclic Leibniz algebras. We have seen previously that if  $C = C_1 \oplus \dots \oplus C_\ell$  is a direct sum of cyclic algebras then

$$C/\text{Leib}(C) = C_1/\text{Leib}(C_1) \oplus \dots \oplus C_\ell/\text{Leib}(C_\ell)$$

and that each  $C_i/\text{Leib}(C_i)$  is the 1-dimensional abelian algebra. Thus if  $K$  is a (Leibniz algebra) direct sum of cyclic subalgebras, it must be a sum of exactly  $\dim(K/\text{Leib}(K)) = 2$  subalgebras. Considering that cyclic subalgebras of  $K$  have dimensions 0, 1, and  $n$  and that  $\dim(K) = 2n$ , we must have two cyclic subalgebras of dimension  $n$ . Suppose that  $K = \langle z_1 \rangle \oplus \langle z_2 \rangle$ , where

$$z_1 = \sum_{i=1}^n a_i x^i + \sum_{j=1}^n b_j y_j, \quad z_2 = \sum_{i=1}^n c_i x^i + \sum_{j=1}^n d_j y_j.$$

Since these are  $n$ -dimensional subalgebras we must have  $a_1 \neq 0$  and  $c_1 \neq 0$ . But then

$$[z_1, z_2] = a_1 \sum_{i=1}^{n-1} c_i x^{i+1} + a_1 \sum_{j=1}^{n-1} d_j y_{j+1}.$$

Notice that the coefficient of  $x^2$  in  $[z_1, z_2]$  is  $a_1 c_1 \neq 0$ . Since  $[z_1, z_2] \neq 0$ , this is not a Leibniz algebra direct sum (contradiction).  $\square$

Note that when  $n = 1$ ,  $K = \text{span}\{x, y_1\}$  where  $[x, x] = [x, y_1] = [y_1, x] = [y_1, y_1] = 0$  so  $K$  is the 2-dimensional abelian Lie algebra and is in this trivial situation a direct sum of cyclic subalgebras. For example, one such decomposition is  $K = \langle x \rangle \oplus \langle y_1 \rangle$ .

## Acknowledgment

We would like to thank the referee for excellent suggestions.

## References

- [Batten Ray et al. 2014] C. Batten Ray, A. Combs, N. Gin, A. Hedges, J. T. Hird, and L. Zack, “Nilpotent Lie and Leibniz algebras”, *Comm. Algebra* **42**:6 (2014), 2404–2410. MR Zbl
- [Bosko et al. 2011] L. Bosko, A. Hedges, J. T. Hird, N. Schwartz, and K. Stagg, “Jacobson’s refinement of Engel’s theorem for Leibniz algebras”, *Involve* **4**:3 (2011), 293–296. MR Zbl
- [Demir et al. 2014] I. Demir, K. C. Misra, and E. Stitzinger, “On some structures of Leibniz algebras”, pp. 41–54 in *Recent advances in representation theory, quantum groups, algebraic geometry, and related topics* (New Orleans, 2012), edited by P. N. Achar et al., Contemp. Math. **623**, Amer. Math. Soc., Providence, RI, 2014. MR Zbl
- [Erdmann and Wildon 2006] K. Erdmann and M. J. Wildon, *Introduction to Lie algebras*, Springer, 2006. MR Zbl
- [Hoffman and Kunze 1971] K. Hoffman and R. Kunze, *Linear algebra*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1971. MR Zbl
- [Loday 1993] J.-L. Loday, “Une version non commutative des algèbres de Lie: les algèbres de Leibniz”, *Enseign. Math. (2)* **39**:3–4 (1993), 269–293. MR Zbl
- [Scofield and Sullivan 2014] D. Scofield and S. M. Sullivan, “Classification of complex cyclic Leibniz algebras”, preprint, 2014. arXiv

Received: 2018-08-31      Revised: 2018-10-09      Accepted: 2019-01-01

cookwj@appstate.edu	Department of Mathematical Sciences, Appalachian State University, Boone, NC, United States
john.hall@uky.edu	Department of Mathematics, University of Kentucky, Lexington, KY, United States
klimavw@appstate.edu	Department of Mathematical Sciences, Appalachian State University, Boone, NC, United States
murraycg@appstate.edu	Department of Mathematical Sciences, Appalachian State University, Boone, NC, United States





# Spectra of Kohn Laplacians on spheres

John Ahn, Mohit Bansil, Garrett Brown,  
Emilee Cardin and Yunus E. Zeytuncu

We study the spectrum of the Kohn Laplacian on the unit spheres in  $\mathbb{C}^n$  and revisit Folland's classical eigenvalue computation. We also look at the growth rate of the eigenvalue counting function in this context. Finally, we consider the growth rate of the eigenvalues of the perturbed Kohn Laplacian on the Rossi sphere in  $\mathbb{C}^2$ .

## 1. Introduction

**Background.** The unit sphere  $\mathbb{S}^{2n-1} \subset \mathbb{C}^n$  is a CR manifold (of hypersurface type) with the CR structure induced from the ambient space. By following the standard setting we define the tangential Cauchy–Riemann complex with the operators  $\bar{\partial}_b$  and  $\bar{\partial}_b^*$  on the spaces of square integrable  $(0, q)$ -forms  $L^2_{(0,q)}(\mathbb{S}^{2n-1})$ . (For simplicity we restrict our attention to  $(0, q)$  forms instead of  $(p, q)$  forms.) The Kohn Laplacian (or  $\bar{\partial}_b$ -Laplacian)

$$\square_b = \bar{\partial}_b \bar{\partial}_b^* + \bar{\partial}_b^* \bar{\partial}_b$$

is a linear, closed, densely defined self-adjoint operator from  $L^2_{(0,q)}(\mathbb{S}^{2n-1})$  to itself. The analytic properties of this second-order differential operator are closely related to the geometry of the underlying manifold (although we work here on  $\mathbb{S}^{2n-1}$ , the same setup works on other CR manifolds). We refer the reader to [Chen and Shaw 2001, Chapter 7] for the details of this setup.

**Spherical harmonics.** We now list definitions and theorems that are needed in the rest of the paper. For a detailed study of spherical harmonics we refer the reader to [Axler et al. 1992].

We say a complex polynomial  $p(z)$  is homogeneous of degree  $k$  if  $p(\lambda z) = \lambda^k p(z)$  for all  $z \neq 0$ . Similarly,  $p(z, \bar{z})$  is called homogeneous of bidegree  $(p, q)$  if  $f(\lambda_1 z, \lambda_2 \bar{z}) = \lambda_1^p \lambda_2^q p(z, \bar{z})$  for all  $z \neq 0$ . We say a twice-differentiable function  $f$

*MSC2010:* primary 32V05; secondary 32V30.

*Keywords:* Kohn Laplacian, spherical harmonics, Gershgorin's circle theorem.

This work is supported by NSF (DMS-1659203). The work of Cardin is also partially supported by a grant from the Simons Foundation (#353525).

is harmonic if  $\Delta f = 0$ , where the Laplacian is defined by

$$\Delta f = 4 \sum_{i=1}^n \frac{\partial^2 f}{\partial z_i \partial \bar{z}_i}.$$

A spherical harmonic is the restriction to  $\mathbb{S}^{2n-1}$  of a complex polynomial that is harmonic on  $\mathbb{C}^n$ . We use  $\mathcal{H}_k(\mathbb{C}^n)$  to denote the space of all harmonic, homogeneous polynomials of degree  $k$  on  $\mathbb{C}^n$  and  $\mathcal{H}_{p,q}(\mathbb{C}^n)$  for the space of all harmonic, homogeneous polynomials of bidegree  $(p, q)$ . Similarly we use  $\mathcal{H}_k(\mathbb{S}^{2n-1})$  and  $\mathcal{H}_{p,q}(\mathbb{S}^{2n-1})$  to denote the restrictions of these spaces on  $\mathbb{S}^{2n-1}$ . The following decomposition theorem is fundamental in our study of  $\square_b$  on  $L^2(\mathbb{S}^{2n-1})$ .

**Theorem 1.1** [Klima 2004, Theorem 3.7]. *The spaces  $\mathcal{H}_{p,q}(\mathbb{S}^{2n-1})$  are pairwise orthogonal, and*

$$L^2(\mathbb{S}^{2n-1}) = \bigoplus_{p,q=0}^{\infty} \mathcal{H}_{p,q}(\mathbb{S}^{2n-1}).$$

By using a standard counting argument one obtains the following formula for the dimensions of the spaces of spherical harmonics.

**Lemma 1.2** [Klima 2004, Corollary 3.10]. *For  $p, q \geq 1$ ,*

$$\begin{aligned} \dim(\mathcal{H}_{p,q}(\mathbb{S}^{2n-1})) &= \binom{n+p-1}{p} \binom{n+q-1}{q} - \binom{n+p-2}{p-1} \binom{n+q-2}{q-1} \\ &= \frac{(n-1)(n+p+q-1)}{pq} \binom{n+p-2}{p-1} \binom{n+q-2}{q-1}. \end{aligned}$$

**Notation.** In the rest of the note we use the standard  $\Omega$  and  $O$  notation to denote asymptotic lower and upper bounds, respectively. That is, given two functions  $f$  and  $g$ , we say  $f = \Omega(g)$  if there exists a constant  $c > 0$  such that  $f(x) \geq cg(x)$  as  $x \rightarrow \infty$ . Similarly,  $f = O(g)$  if there exists  $c > 0$  such that  $f(x) \leq cg(x)$  as  $x \rightarrow \infty$ . Finally, we say  $f = \Theta(g)$  if  $f = \Omega(g)$  and  $f = O(g)$ .

**Results.** Folland [1972] computed the eigenvalues and eigenforms of  $\square_b$  on  $L^2_{(0,q)}(\mathbb{S}^{2n-1})$  by using unitary representations.

**Theorem 1.3.**  $\mathcal{H}_{p,q}(\mathbb{S}^{2n-1})$  is an eigenspace for  $\bar{\partial}_b^* \bar{\partial}_b$  with the associated eigenvalue  $2q(p+n-1)$ .

In Section 2 of this note we go over these computations on the space of square integrable functions (i.e.,  $L^2(\mathbb{S}^{2n-1})$ ) by using spherical harmonics and present eigenvalue computations in an accessible way. This more elementary approach enables us to write code<sup>1</sup> in SymPy that computes the eigenvalues of  $\square_b$  and other similar second-order differential operators defined on  $L^2(\mathbb{S}^{2n-1})$ . Furthermore, by

<sup>1</sup>The code can be downloaded at <https://goo.gl/kBsUzA>.

using the explicit forms of the eigenvalues and formulas for the dimensions of spherical harmonic subspaces of  $L^2(\mathbb{S}^{2n-1})$ , we study the growth rate for the counting function of the eigenvalues. For  $m \in \mathbb{Z}$ , let  $N(m)$  be the number of eigenvalues of  $\square_b$  on  $L^2(\mathbb{S}^{2n-1})$  that are less than or equal to  $m$ , counting multiplicity.

**Theorem 1.4.** *There exists a real  $c > 0$  so that  $\frac{1}{c}m^n \leq N(m) \leq cm^n$ ; that is,  $N(m) = \Theta(m^n)$ .*

In other words, here we prove that

$$\limsup_{m \rightarrow \infty} \frac{N(m)}{m^n} \in (0, \infty).$$

It would be interesting to compute the exact limit and check if it is related to the surface area of  $\mathbb{S}^{2n-1}$ . Indeed, in the case of the Laplace–Beltrami operator, Weyl’s law states that this ratio is the surface area of  $\mathbb{S}^{2n-1}$ .

In addition to the induced CR structure from the ambient manifold, one can define different intrinsic CR structures on a given manifold; see [Bogges 1991, Chapter 8]. The most famous example of these abstract CR manifolds is the Rossi sphere. It is known that the Rossi sphere is not globally CR embeddable into any  $\mathbb{C}^n$  [Burns 1979]. This can be seen by explicitly studying the perturbed Kohn Laplacian (defined by the abstract CR structure) and looking at its essential spectrum. In [Abbas et al. 2019], the authors studied the bottom of the spectrum of the perturbed Kohn Laplacian by using spherical harmonics. In the last section of this note we continue this study and provide the growth rate of the largest eigenvalues from each subspace of spherical harmonics.

## 2. Eigenvalues of $\square_b$ on $L^2(\mathbb{S}^{2n-1})$

**Explicit eigenvalue computation.** Since  $\bar{\partial}_b^*$  is identically zero on  $L^2(\mathbb{S}^{2n-1})$ ,  $\square_b$  simplifies on  $L^2(\mathbb{S}^{2n-1})$  as

$$\square_b = \bar{\partial}_b^* \bar{\partial}_b.$$

Before we compute the eigenvalues we present the operators  $\bar{\partial}_b$  and  $\bar{\partial}_b^*$  in coordinate forms. A smooth differential 1-form  $\omega$  on  $\mathbb{S}^{2n-1}$  can be expressed as

$$\omega = \sum_{k=1}^n (A_k dz_k + B_k d\bar{z}_k) = A_1 dz_1 + B_1 d\bar{z}_1 + \cdots + A_n dz_n + B_n d\bar{z}_n,$$

where  $A_k, B_k \in C^\infty(\mathbb{C}^n)$ . As computed in [Folland 1972], for a smooth function  $f$  on  $\mathbb{S}^{2n-1}$  we have

$$\bar{\partial}_b f = \sum_{i=1}^n \left( \frac{\partial f}{\partial \bar{z}_i} - z_i \sum_{a=1}^n \bar{z}_a \frac{\partial f}{\partial \bar{z}_a} \right) d\bar{z}_i.$$

Furthermore, following the normalization of inner products as in [Folland 1972] we have

$$\langle d\bar{z}_i, d\bar{z}_j \rangle = 2\delta_{ij} \quad \text{and} \quad \langle dz_i, d\bar{z}_j \rangle = 0.$$

Using integration by parts, we obtain the following expression for the adjoint operator.

**Lemma 2.1.** *For a smooth 1-form  $\omega = \sum_{k=1}^n (A_k dz_k + B_k d\bar{z}_k)$ ,*

$$\bar{\partial}_b^* \omega = -2 \sum_{i=1}^n \left( \frac{\partial}{\partial z_i} B_i - \sum_{a=1}^n \frac{\partial}{\partial z_a} z_a \bar{z}_i B_i \right).$$

*Proof.* Let  $g$  be a smooth function on  $\mathbb{S}^{2n-1}$ . Since we are working on a compact manifold, we don't get any boundary terms when we integrate by parts:

$$\begin{aligned} & \left\langle \bar{\partial}_b^* \left( \sum_{k=1}^n (A_k dz_k + B_k d\bar{z}_k) \right), g \right\rangle \\ &= \left\langle \sum_{k=1}^n (A_k dz_k + B_k d\bar{z}_k), \bar{\partial}_b g \right\rangle \\ &= \left\langle \sum_{k=1}^n A_k dz_k + \sum_{k=1}^n B_k d\bar{z}_k, \sum_{i=1}^n \left( \frac{\partial g}{\partial \bar{z}_i} - z_i \sum_{a=1}^n \bar{z}_a \frac{\partial g}{\partial \bar{z}_a} \right) d\bar{z}_i \right\rangle \\ &= 2 \sum_{i=1}^n \left\langle B_i, \frac{\partial g}{\partial \bar{z}_i} - z_i \sum_{a=1}^n \bar{z}_a \frac{\partial g}{\partial \bar{z}_a} \right\rangle \\ &= 2 \sum_{i=1}^n \left( \left\langle B_i, \frac{\partial g}{\partial \bar{z}_i} \right\rangle - \sum_{a=1}^n \left\langle B_i, z_i \bar{z}_a \frac{\partial g}{\partial \bar{z}_a} \right\rangle \right) \\ &= 2 \sum_{i=1}^n \left( - \left\langle \frac{\partial}{\partial z_i} B_i, g \right\rangle + \sum_{a=1}^n \left\langle \frac{\partial}{\partial z_a} z_a \bar{z}_i B_i, g \right\rangle \right) \\ &= -2 \sum_{i=1}^n \left( \left\langle \frac{\partial}{\partial z_i} B_i, g \right\rangle - \sum_{a=1}^n \left\langle \frac{\partial}{\partial z_a} z_a \bar{z}_i B_i, g \right\rangle \right) \\ &= -2 \sum_{i=1}^n \left\langle \frac{\partial}{\partial z_i} B_i - \sum_{a=1}^n \frac{\partial}{\partial z_a} z_a \bar{z}_i B_i, g \right\rangle \\ &= \left\langle -2 \sum_{i=1}^n \left( \frac{\partial}{\partial z_i} B_i - \sum_{a=1}^n \frac{\partial}{\partial z_a} z_a \bar{z}_i B_i \right), g \right\rangle. \end{aligned}$$

By comparing the beginning and ending of the identity we prove the lemma.  $\square$

Before we look at the action of  $\square_b$  on a square integrable function we look at the action of two other operations on the spherical harmonics.

**Lemma 2.2.** *If  $f \in \mathcal{H}_{p,q}(\mathbb{S}^{2n-1})$ , then*

$$\sum_{k=1}^n z_k \frac{\partial f}{\partial z_k} = pf \quad \text{and} \quad \sum_{k=1}^n \bar{z}_k \frac{\partial f}{\partial \bar{z}_k} = qf.$$

*Proof.* Consider a polynomial  $f \in \mathcal{H}_{p,q}$ . So  $f$  is harmonic homogeneous of bidegree  $p, q$ . Then for each monomial term  $g = z_1^{\alpha_1} \cdots z_n^{\alpha_n} \bar{z}_1^{\beta_1} \cdots \bar{z}_n^{\beta_n}$  of  $f$ , we have

$$\begin{aligned} \sum_{k=1}^n z_k \frac{\partial g}{\partial z_k} &= \sum_{k=1}^n (\alpha_k) g = \left( \sum_{k=1}^n \alpha_k \right) g = pg, \\ \sum_{k=1}^n \bar{z}_k \frac{\partial g}{\partial \bar{z}_k} &= \sum_{k=1}^n (\beta_k) g = \left( \sum_{k=1}^n \beta_k \right) g = qg. \end{aligned}$$

So each monomial term  $g$  is scaled by  $p$  or  $q$ . By the linearity of differential operators,  $f$  is scaled by  $p$  or  $q$  as well.  $\square$

By combining the lemmas above we obtain the eigenvalues of  $\square_b$ .

**Theorem 1.3.**  $\mathcal{H}_{p,q}(\mathbb{S}^{2n-1})$  is an eigenspace for  $\bar{\partial}_b^* \bar{\partial}_b$  with the associated eigenvalue  $2q(p+n-1)$ .

*Proof.* For  $f \in \mathcal{H}_{p,q}(\mathbb{S}^{2n-1})$ ,

$$\begin{aligned} \bar{\partial}_b^* \bar{\partial}_b f &= \bar{\partial}_b^* \left[ \sum_{i=1}^n \left( \frac{\partial f}{\partial \bar{z}_i} - z_i \sum_{a=1}^n \bar{z}_a \frac{\partial f}{\partial \bar{z}_a} \right) d\bar{z}_i \right] \\ &= \bar{\partial}_b^* \left[ \sum_{i=1}^n \left( \frac{\partial f}{\partial \bar{z}_i} - z_i qf \right) d\bar{z}_i \right] \\ &= -2 \sum_{i=1}^n \left[ \frac{\partial}{\partial z_i} \left( \frac{\partial f}{\partial \bar{z}_i} - z_i qf \right) - \sum_{a=1}^n \frac{\partial}{\partial z_a} z_a \bar{z}_i \left( \frac{\partial f}{\partial \bar{z}_i} - z_i qf \right) \right] \\ &= -2 \sum_{i=1}^n \left[ \left( \frac{\partial^2 f}{\partial z_i \partial \bar{z}_i} - \frac{\partial}{\partial z_i} z_i qf \right) - \sum_{a=1}^n \left( \frac{\partial}{\partial z_a} z_a \bar{z}_i \frac{\partial f}{\partial \bar{z}_i} - \frac{\partial}{\partial z_a} z_a \bar{z}_i z_i qf \right) \right] \\ &= -2 \sum_{i=1}^n \frac{\partial^2 f}{\partial z_i \partial \bar{z}_i} + 2 \sum_{i=1}^n \frac{\partial}{\partial z_i} z_i qf + 2 \sum_{i=1}^n \sum_{a=1}^n \frac{\partial}{\partial z_a} z_a \bar{z}_i \frac{\partial f}{\partial \bar{z}_i}. \end{aligned}$$

We start with the first term. Because  $f$  is harmonic, we know

$$0 = \Delta(f) = 4 \sum_{i=1}^n \frac{\partial^2 f}{\partial z_i \partial \bar{z}_i}.$$

Thus, we have

$$0 = \sum_{i=1}^n \frac{\partial^2 f}{\partial z_i \partial \bar{z}_i} = -2 \sum_{i=1}^n \frac{\partial^2 f}{\partial z_i \partial \bar{z}_i}.$$

For the second and third terms, we apply the product rule:

$$\begin{aligned} 2 \sum_{i=1}^n \frac{\partial}{\partial z_i} z_i q f &= 2q \sum_{i=1}^n \frac{\partial}{\partial z_i} z_i f \\ &= 2q \sum_{i=1}^n \left( z_i \frac{\partial f}{\partial z_i} + f \right) \\ &= 2q \left[ \sum_{i=1}^n z_i \frac{\partial f}{\partial z_i} + \sum_{i=1}^n f \right] = 2q(p+n)f, \\ 2 \sum_{i=1}^n \sum_{a=1}^n \frac{\partial}{\partial z_a} z_a \bar{z}_i \frac{\partial f}{\partial \bar{z}_i} &= 2 \sum_{a=1}^n \frac{\partial}{\partial z_a} z_a \sum_{i=1}^n \bar{z}_i \frac{\partial f}{\partial \bar{z}_i} \\ &= 2 \sum_{a=1}^n \frac{\partial}{\partial z_a} z_a q f \\ &= 2q \sum_{a=1}^n \left( z_a \frac{\partial f}{\partial z_a} + f \right) = 2q(p+n)f. \end{aligned}$$

Now recall that on  $\mathbb{S}^{2n-1}$  we have  $z_1 \bar{z}_1 + \cdots + z_n \bar{z}_n = 1$ . Thus,

$$\sum_{a=1}^n \sum_{i=1}^n z_i \bar{z}_i f = \sum_{a=1}^n f = nf.$$

We also go over the following explicit computation (again by using linearity we can assume  $f$  is a monomial and  $f = z_1^{\alpha_1} \cdots z_n^{\alpha_n} \bar{z}_1^{\beta_1} \cdots \bar{z}_n^{\beta_n}$ ):

$$\begin{aligned} \sum_{a=1}^n z_a \frac{\partial}{\partial z_a} \sum_{i=1}^n z_i \bar{z}_i f &= \sum_{a=1}^n z_a \frac{\partial}{\partial z_a} (z_1 \bar{z}_1 + \cdots + z_n \bar{z}_n) f \\ &= \sum_{a=1}^n z_a \left( \frac{\partial}{\partial z_a} z_1 \bar{z}_1 f + \cdots + \frac{\partial}{\partial z_a} z_a \bar{z}_a f + \cdots + \frac{\partial}{\partial z_a} z_n \bar{z}_n f \right) \\ &= \sum_{a=1}^n z_a \left( \frac{\alpha_a}{z_a} z_1 \bar{z}_1 f + \cdots + \frac{\alpha_a + 1}{z_a} z_a \bar{z}_a f + \cdots + \frac{\alpha_a}{z_a} z_n \bar{z}_n f \right) \\ &= \sum_{a=1}^n ((\alpha_a) z_1 \bar{z}_1 f + \cdots + (\alpha_a + 1) z_a \bar{z}_a f + \cdots + (\alpha_a) z_n \bar{z}_n f) \end{aligned}$$

$$\begin{aligned}
&= \sum_{i=1}^n (\alpha_1 + \cdots + (\alpha_i + 1) + \cdots + \alpha_n) z_i \bar{z}_i f \\
&= \sum_{i=1}^n (p+1) z_i \bar{z}_i f = (p+1) \sum_{i=1}^n z_i \bar{z}_i f = (p+1) f.
\end{aligned}$$

We are now ready to compute the fourth term of the  $\bar{\partial}_b^* \bar{\partial}_b f$  expansion:

$$\begin{aligned}
-2 \sum_{i=1}^n \sum_{a=1}^n \frac{\partial}{\partial z_a} z_a \bar{z}_i z_i q f &= -2q \left( \sum_{a=1}^n \frac{\partial}{\partial z_a} z_a \sum_{i=1}^n z_i \bar{z}_i f \right) \\
&= -2q \left( \sum_{a=1}^n \left( z_a \frac{\partial}{\partial z_a} + I \right) \sum_{i=1}^n z_i \bar{z}_i f \right) \\
&= -2q \left( \sum_{a=1}^n z_a \frac{\partial}{\partial z_a} \sum_{i=1}^n z_i \bar{z}_i f + \sum_{a=1}^n \sum_{i=1}^n z_i \bar{z}_i f \right) \\
&= -2q(p+1+n)f.
\end{aligned}$$

Returning to our original computation of  $\bar{\partial}_b^* \bar{\partial}_b f$ , we now have

$$\begin{aligned}
&\bar{\partial}_b^* \bar{\partial}_b f \\
&= -2 \sum_{i=1}^n \frac{\partial^2 f}{\partial z_i \partial \bar{z}_i} + 2 \sum_{i=1}^n \frac{\partial}{\partial z_i} z_i q + 2 \sum_{i=1}^n \sum_{a=1}^n \frac{\partial}{\partial z_a} z_a \bar{z}_i \frac{\partial f}{\partial \bar{z}_i} - 2 \sum_{i=1}^n \sum_{a=1}^n \frac{\partial}{\partial z_a} z_a \bar{z}_i z_i q \\
&= 0 + 2q(p+n)f + 2q(p+n)f - 2q(p+1+n)f \\
&= 2q(p+n-1)f.
\end{aligned}$$

□

**Asymptotics of counting function.** We now look at the counting function  $N(m)$ .

**Definition 2.3.** For  $m \in \mathbb{Z}$ , let  $N(m)$  be the number of eigenvalues of  $\square_b$  on  $L^2(\mathbb{S}^{2n-1})$  that are less than or equal to  $m$ , counting multiplicity.

Similar functions and relations between their asymptotics and geometry of the underlying manifold were studied in [Métivier 1981; Fu 2005; 2008]. In particular in some cases the growth rate of  $N(m)$  carries information about the type of the manifold [Fu 2005; 2008]. Furthermore, in the case of the Laplace–Beltrami operator, Weyl’s law states that the limit of the ratio  $N(m)/m^n$  gives the surface area of  $\mathbb{S}^{2n-1}$ . Before we state our result, we recall Lemma 1.2.

**Lemma 1.2.** For  $p, q \geq 1$ ,

$$\dim(\mathcal{H}_{p,q}(\mathbb{S}^{2n-1})) = \frac{(n-1)(n+p+q-1)}{pq} \binom{n+p-2}{p-1} \binom{n+q-2}{q-1}.$$

Note that ignoring multiplicity would induce a function with linear growth. Indeed for any even  $\hat{m}$  with  $m \geq \hat{m} > 2(n-1)$ , we can solve  $\hat{m} = 2q(p+n-1)$  after fixing  $q = 1$ . Additionally, by convention, we set  $N(m) = 0$  when  $m < 0$ .

We note that when  $n = 1$ , the eigenvalue of  $\bar{\partial}_b^* \bar{\partial}_b$  is equal to 0. Indeed, when  $n = 1$  and when  $p$  and  $q$  are both nonzero, Lemma 1.2 gives us that the dimension of  $\mathcal{H}_{p,q}$  is 0. This is because the only harmonic homogeneous polynomials on  $\mathbb{C}$  are of the form  $z^p$  or  $\bar{z}^q$ , which belong to  $\mathcal{H}_{p,0}$  or  $\mathcal{H}_{0,q}$ , respectively. Thus,  $\mathcal{H}_{p,q}$  is nontrivial only when either  $p$  or  $q$  is zero. However, on such spaces, the eigenvalue of  $\bar{\partial}_b^* \bar{\partial}_b$  on  $\mathcal{H}_{p,q}$  is 0.

**Lemma 2.4.** *There exists a real constant  $c > 0$  so that  $cm^n \leq N(m)$ ; that is,  $N(m) \in \Omega(m)$ .*

*Proof.* Fix even  $m$ ; then  $N(m) - N(m-2)$  is the multiplicity of the eigenvalue  $m$ , since all the eigenvalues are even by Theorem 1.3. This requires computing the sum of the dimensions of all  $\mathcal{H}_{p,q}(\mathbb{S}^{2n-1})$  such that the pair  $(p, q)$  satisfies the equation  $E(p, q) = m$ , where  $E(p, q) = 2q(p+n-1)$ . Now assuming  $m > 2(n-1)$ , there exists a positive integer solution  $p = \hat{p}$  to  $E(p, q) = m$  when  $q = 1$ . Define the solution set  $A = \{(p, q) \mid E(p, q) = m\}$ . Then we have

$$N(m) - N(m-2) = \sum_{(p,q) \in A} \dim \mathcal{H}_{p,q} \geq \dim \mathcal{H}_{\hat{p},1}.$$

Note that  $\dim \mathcal{H}_{\hat{p},1} = \Omega(m^{n-1})$ , which follows from Lemma 1.2. Namely, since asymptotically  $\hat{p} = m/2$ , we have

$$\begin{aligned} \dim \mathcal{H}_{\hat{p},1(\mathbb{S}^{2n-1})} &= \frac{(n-1)(n+\hat{p})}{\hat{p}} \binom{n+\hat{p}-2}{n-1} \binom{n-1}{n-1} \\ &\geq \binom{n+\hat{p}-2}{n-1} \geq \frac{1}{(n-1)!} \hat{p}^{n-1} \\ &= \Omega\left(\frac{m}{2}\right)^{n-1} = \Omega(m^{n-1}). \end{aligned}$$

Putting it all together, we have

$$\begin{aligned} 2N(m) &\geq N(m) + N(m-1) = \sum_{j=0}^m (N(j) - N(j-2)) \\ &\geq \sum_{j=0}^m \Omega(j^{n-1}) \geq \Omega(m^n). \end{aligned} \quad \square$$

**Lemma 2.5.** *There exists a real constant  $c > 0$  so that  $N(m) \leq cm^n$ ; that is,  $N(m) = O(m^n)$ .*

*Proof.* Again, fix an even  $m$  and inspect  $N(m) - N(m-2)$ . Note that asymptotically, we can let our eigenvalue equation be  $E(p, q) = 2qp$ . Thus, asymptotically we



have

$$N(m) - N(m-2) = \sum_{(p,q) \in A} \dim \mathcal{H}_{p,q} \lesssim \sum_{(p,q) \in A} (p+q)(pq)^{n-2} = \sigma(m) O(m^{n-2}),$$

where  $\sigma(m)$  is the sum of all divisors of  $m$ . Thus, we have

$$N(m) \lesssim \sum_{x \leq m} 2x^{n-2} \sigma(x) \lesssim 2m^{n-2} \sum_{x \leq m} \sigma(x) = O(m^n).$$

The last equality follows since  $\sum_{x \leq m} \sigma(x) = O(m^2)$ . A proof of this fact can be found in Chapter 3.6 of [Apostol 1976].  $\square$

By combining the last two lemmas we obtain the following statement.

**Theorem 1.4.** *There exists a real  $c > 0$  so that  $\frac{1}{c}m^n \leq N(m) \leq cm^n$ ; that is,  $N(m) = \Theta(m^n)$ .*

We note that the constants in Lemma 2.4, Lemma 2.5, and Theorem 1.4 do depend on the dimension  $n$ . This dependence also agrees with the explicit constants calculated by Weyl for the Laplace–Beltrami operator.

### 3. Spectra of other second-order differential operators on $L^2(\mathbb{S}^{2n-1})$

Another interesting class of second-order differential operators are sum of squares operators  $\mathcal{M}_b$ , introduced in the fourth chapter of [Klima 2004]. These operators capture *half* of the action of  $\square_b$  on  $\mathbb{S}^3$ ; in higher dimensions they lead to the study of various possible perturbations of  $\square_b$ .

We define the sum of squares operator  $\mathcal{M}_b$  on  $L^2(\mathbb{S}^{2n-1})$  as

$$\mathcal{M}_b = -(M_{12}\bar{M}_{12} + M_{13}\bar{M}_{13} + \cdots + M_{1n}\bar{M}_{1n}),$$

where  $M_{1k} = \bar{z}_1(\partial/\partial z_k) - \bar{z}_k(\partial/\partial z_1)$  and  $\bar{M}_{1k} = z_1(\partial/\partial \bar{z}_k) - z_k(\partial/\partial \bar{z}_1)$ . Note that one can easily consider  $M_{ik}$  for  $i \neq 1$ ; for simplicity we focus on the case  $i = 1$ .

For any  $f \in \mathcal{H}_{p,q}(\mathbb{S}^{2n-1})$ , the specific degrees of the  $z_k, \bar{z}_k$  may vary. For example, both  $z_1^2 z_2 \bar{z}_1^3 \bar{z}_2^2$  and  $z_1 z_2^2 \bar{z}_1^3 \bar{z}_2^3$  are in  $\mathcal{H}_{3,5}(\mathbb{S}^3)$ . In previous arguments, such specificity was unnecessary, but we find that for  $\mathcal{M}_b$ , the eigenvalues can directly depend on the exact degrees of the  $z_k, \bar{z}_k$ . To that end, for nonnegative integer tuples  $p = (p_1, \dots, p_n)$  and  $q = (q_1, \dots, q_n)$ , we use  $\mathcal{H}_{p,q}^*(\mathbb{C}^n)$  to denote the space of all harmonic, homogeneous polynomials where  $p_k$  is the degree of  $z_k$ , and  $q_k$  is the degree of  $\bar{z}_k$ . Then we use  $\mathcal{H}_{p,q}^*(\mathbb{S}^{2n-1})$  to denote the restriction of this space on  $\mathbb{S}^{2n-1}$ . For example, now  $z_1^2 z_2 \bar{z}_1^3 \bar{z}_2^2 \in \mathcal{H}_{(2,1),(3,2)}^*(\mathbb{S}^3)$  but  $z_1 z_2^2 \bar{z}_1^3 \bar{z}_2^3 \in \mathcal{H}_{(1,2),(2,3)}^*(\mathbb{S}^3)$ . Note that  $\mathcal{H}_{p,q}^*(\mathbb{S}^{2n-1})$  is a subspace of  $\mathcal{H}_{\bar{p},\bar{q}}(\mathbb{S}^{2n-1})$ , where  $\bar{p} = \sum_{i=1}^n p_i$  and  $\bar{q} = \sum_{i=1}^n q_i$ . Now for certain  $\mathcal{H}_{p,q}^*(\mathbb{S}^{2n-1})$ , we have the following result.

**Lemma 3.1.** *Consider two nonnegative integer tuples  $p = (p_1, \dots, p_n)$  and  $q = (q_1, \dots, q_n)$ . Suppose that for each  $1 \leq k \leq n$ , at least one of  $p_k$  or  $q_k$  is 0. Then the eigenvalue of  $\mathcal{M}_b$  on  $\mathcal{H}_{p,q}^*(\mathbb{S}^{2n-1})$  is*

$$p_1 \sum_{k=2}^n q_k + q_1 \sum_{k=2}^n p_k + (n-1)q_1 + \sum_{k=2}^n q_k.$$

*Proof.* Take  $f \in \mathcal{H}_{p,q}^*(\mathbb{S}^{2n-1})$ , where  $p_k = 0$  or  $q_k = 0$  for each  $k$ . By linearity, we can inspect the action of each  $-M_{1k}\bar{M}_{1k}$  piece of  $\mathcal{M}_b$  on  $f$ . We have

$$\begin{aligned} -M_{1k}\bar{M}_{1k}f &= -\left(\bar{z}_1 \frac{\partial}{\partial z_k} - \bar{z}_k \frac{\partial}{\partial z_1}\right) \left(z_1 \frac{\partial}{\partial \bar{z}_k} - z_k \frac{\partial}{\partial \bar{z}_1}\right) f \\ &= -\bar{z}_1 \frac{\partial}{\partial z_k} z_1 \frac{\partial}{\partial \bar{z}_k} f + \bar{z}_1 \frac{\partial}{\partial z_k} z_k \frac{\partial}{\partial \bar{z}_1} f + \bar{z}_k \frac{\partial}{\partial z_1} z_1 \frac{\partial}{\partial \bar{z}_k} f - \bar{z}_k \frac{\partial}{\partial z_1} z_k \frac{\partial}{\partial \bar{z}_1} f \\ &= -z_1 \bar{z}_1 \frac{\partial}{\partial z_k} \frac{\partial}{\partial \bar{z}_k} f + \bar{z}_1 \frac{\partial}{\partial \bar{z}_1} \frac{\partial}{\partial z_k} z_k f + \bar{z}_k \frac{\partial}{\partial \bar{z}_k} \frac{\partial}{\partial z_1} z_1 f - z_k \bar{z}_k \frac{\partial}{\partial z_1} \frac{\partial}{\partial \bar{z}_1} f \\ &= 0 + \bar{z}_1 \frac{\partial}{\partial \bar{z}_1} \frac{\partial}{\partial z_k} z_k f + \bar{z}_k \frac{\partial}{\partial \bar{z}_k} \frac{\partial}{\partial z_1} z_1 f - 0 \\ &= \bar{z}_1 \frac{\partial}{\partial \bar{z}_1} \left(z_k \frac{\partial}{\partial z_k} + I\right) f + \bar{z}_k \frac{\partial}{\partial \bar{z}_k} \left(z_1 \frac{\partial}{\partial z_1} + I\right) f \\ &= \bar{z}_1 \frac{\partial}{\partial \bar{z}_1} z_k \frac{\partial}{\partial z_k} f + \bar{z}_1 \frac{\partial}{\partial \bar{z}_1} f + \bar{z}_k \frac{\partial}{\partial \bar{z}_k} z_1 \frac{\partial}{\partial z_1} f + \bar{z}_k \frac{\partial}{\partial \bar{z}_k} f \\ &= q_1 p_k f + q_1 f + q_k p_1 f + q_k f. \end{aligned}$$

Thus, we have

$$\begin{aligned} \mathcal{M}_b(f) &= \sum_{k=2}^n -M_{1k}\bar{M}_{1k}f = \sum_{k=2}^n (q_1 p_k + q_1 + q_k p_1 + q_k) f \\ &= \left( \sum_{k=2}^n q_1 p_k + \sum_{k=2}^n q_1 + \sum_{k=2}^n q_k p_1 + \sum_{k=2}^n q_k \right) f \\ &= \left( q_1 \sum_{k=2}^n p_k + (n-1)q_1 + p_1 \sum_{k=2}^n q_k + \sum_{k=2}^n q_k \right) f. \quad \square \end{aligned}$$

The above lemma tells us that  $z_1^2 z_2 \bar{z}_1^3 \bar{z}_2^2 \in \mathcal{H}_{(2,1),(3,2)}^*(\mathbb{S}^3)$  has eigenvalue  $2(2) + 3(1) + (2-1)(3) + (2) = 12$ . On the other hand,  $z_1 z_2^2 \bar{z}_1^2 \bar{z}_2^3 \in \mathcal{H}_{(1,2),(3,2)}^*(\mathbb{S}^3)$  has eigenvalue  $1(2) + 3(2) + (2-1)(3) + (2) = 13$ . More generally, the lemma tells us that  $\mathcal{H}_{p,0}(\mathbb{S}^{2n-1})$  is in the null space of  $\mathcal{M}_b$  for all  $p \in \mathbb{N}$ . Furthermore, the eigenvalue of  $\mathcal{M}_b$  on  $\mathcal{H}_{0,q}^*(\mathbb{S}^{2n-1})$  is  $(n-1)q_1 + q_2 + \dots + q_n$ . On other  $\mathcal{H}_{p,q}(\mathbb{S}^{2n-1})$  spaces, computational results suggest that we have integer eigenvalues, and matrix representations follow a pattern as well. We will leave the investigation

of other eigenvalues to a future study. We invite the interested reader to see other computational results by downloading our code.<sup>2</sup>

#### 4. Eigenvalues of $\square_b^t$ on the Rossi sphere

Previously in [Abbas et al. 2019], the authors studied the spectrum of the perturbed Kohn Laplacian  $\square_b^t$  on the Rossi sphere. They obtained an upper bound for the lowest eigenvalue for  $\square_b^t$  on each  $\mathcal{H}_k(\mathbb{S}^3)$ . In our project, we look at the asymptotics of the spectrum of the (perturbed) Kohn Laplacian on the Rossi sphere, in particular the asymptotics of  $\lambda_k^{\max}$ , the maximum eigenvalue of  $\square_b^t$  on  $\mathcal{H}_k(\mathbb{S}^3)$ .

In [Abbas et al. 2019] the authors prove tridiagonal representation results for spaces of homogeneous polynomials of odd degree,  $\mathcal{H}_{2k-1}(\mathbb{S}^3)$ . However, their proof actually works for arbitrary degrees,  $\mathcal{H}_k(\mathbb{S}^3)$ . We restate the steps to construct the tridiagonal matrix representations here, and one can refer to [Abbas et al. 2019] for details. We first recall the definition of differential operators  $\mathcal{L}$ ,  $\bar{\mathcal{L}}$ , and  $\square_b^t$  on  $L^2(\mathbb{S}^3)$ .

**Definition 4.1.** We define  $\mathcal{L}$  and  $\bar{\mathcal{L}}$  as

$$\begin{aligned}\mathcal{L} &= \bar{z}_1 \frac{\partial}{\partial z_2} - \bar{z}_2 \frac{\partial}{\partial z_1}, \\ \bar{\mathcal{L}} &= z_1 \frac{\partial}{\partial \bar{z}_2} - z_2 \frac{\partial}{\partial \bar{z}_1}, \\ \square_b^t &= -\mathcal{L}_t \frac{1 + |t|^2}{(1 - |t|^2)^2} \bar{\mathcal{L}}_t.\end{aligned}$$

The motivation for these operators arises from the CR-manifold  $(\mathbb{S}^3, \mathcal{L}_t)$ , which is not CR-embeddable [Rossi 1965]. Note that  $\mathcal{L}_t = \mathcal{L} + t\bar{\mathcal{L}}$  and  $|t| < 1$ .

**Theorem 4.2** [Abbas et al. 2019]. *Let  $\{f_0, \dots, f_k\}$  be an orthogonal basis for  $\mathcal{H}_{0,k}(\mathbb{S}^3)$ . Then  $\{\bar{\mathcal{L}}^\sigma f_0, \dots, \bar{\mathcal{L}}^\sigma f_k\}$  is an orthogonal basis for  $\mathcal{H}_{\sigma,k-\sigma}(\mathbb{S}^3)$ .*

The proof of Theorem 4.2 follows from induction on inner products. The main two steps are the fact that  $-\mathcal{L}$  is the adjoint of  $\bar{\mathcal{L}}$ , and that  $\mathcal{L}\bar{\mathcal{L}}$  scales elements of  $\mathcal{H}_{p,q}(\mathbb{S}^3)$  by a constant factor based on their bidegree.

Now one can consider an orthogonal basis  $\{f_0, \dots, f_k\}$  for  $\mathcal{H}_{0,k}(\mathbb{S}^3)$  and define the following two subspaces for even  $k$ :

$$\begin{aligned}V_i &= \text{span}\{f_i, \bar{\mathcal{L}}^2 f_i, \bar{\mathcal{L}}^4 f_i, \dots, \bar{\mathcal{L}}^{k-2} f_i, \bar{\mathcal{L}}^k f_i\}, \\ W_i &= \text{span}\{\bar{\mathcal{L}} f_i, \bar{\mathcal{L}}^3 f_i, \bar{\mathcal{L}}^5 f_i, \dots, \bar{\mathcal{L}}^{k-3} f_i, \bar{\mathcal{L}}^{k-1} f_i\},\end{aligned}$$

<sup>2</sup>The code can be downloaded at <https://goo.gl/kBsUzA>.

and similarly for odd  $k$ :

$$\begin{aligned} V_i &= \text{span}\{f_i, \bar{\mathcal{L}}^2 f_i, \bar{\mathcal{L}}^4 f_i, \dots, \bar{\mathcal{L}}^{k-3} f_i, \bar{\mathcal{L}}^{k-1} f_i\}, \\ W_i &= \text{span}\{\bar{\mathcal{L}} f_i, \bar{\mathcal{L}}^3 f_i, \bar{\mathcal{L}}^5 f_i, \dots, \bar{\mathcal{L}}^{k-2} f_i, \bar{\mathcal{L}}^k f_i\}. \end{aligned}$$

The motivation to define such spaces follows by inspecting the expanded form of  $\square_b^t$ , which is equal to  $\mathcal{L}\bar{\mathcal{L}} + \bar{\mathcal{L}}\mathcal{L} + \mathcal{L}^2 + \bar{\mathcal{L}}^2$  up to constants. Previous work has shown that  $\mathcal{L}\bar{\mathcal{L}}$  and  $\bar{\mathcal{L}}\mathcal{L}$  scale elements of  $\mathcal{H}_{p,q}(\mathbb{S}^3)$  by a constant factor, and the actions of  $\mathcal{L}^2$  and  $\bar{\mathcal{L}}^2$  suggest that invariant subspaces will involve basis elements that differ by  $2j$  applications of  $\bar{\mathcal{L}}$ . Indeed, it was shown in [Abbas et al. 2019] that  $\square_b^t$  is invariant on  $V_i$  and  $W_i$ . On these finite-dimensional invariant subspaces one can obtain a matrix representation for the second-order operator  $\square_b^t$ .

**Theorem 4.3** [Abbas et al. 2019]. *The matrix representation of  $\square_b^t$ ,  $m(\square_b^t)$ , on  $V_i, W_i \subset \mathcal{H}_k(\mathbb{S}^3)$  is*

$$h \begin{pmatrix} d_1 & u_1 & & & \\ -\bar{t} & d_2 & u_2 & & \\ & -\bar{t} & d_3 & \ddots & \\ & & \ddots & \ddots & u_{k-1} \\ & & & -\bar{t} & d_k \end{pmatrix},$$

where  $h$  is a constant and on  $V_i$ ,

$$\begin{aligned} u_j &= -4t \cdot (j)(2j-1)(k-j)(2k-1-2j), \\ d_j &= (2j-1)(2k+1-2j) + 4|t|^2(j-1)(k+1-j); \end{aligned}$$

on  $W_i$ ,

$$\begin{aligned} u_j &= -4t \cdot (j)(2j+1)(k-j)(2k-1-2j), \\ d_j &= 4(j)(k-j) + |t|^2(2j-1)(2k+1-2j). \end{aligned}$$

Moreover, the matrix above is similar to

$$B = \begin{pmatrix} d_1 & c_1 & & & \\ c_1 & d_2 & c_2 & & \\ & c_2 & d_3 & \ddots & \\ & & \ddots & \ddots & c_{k-1} \\ & & & c_{k-1} & d_k \end{pmatrix},$$

where  $c_j = (-\bar{t} \cdot u_j)^{1/2} = |t|\sqrt{-u_j/t}$ .

After recalling these results, we also introduce the numerical range of a matrix.

**Definition 4.4.** Given an  $n \times n$  square matrix  $A$ , we define its numerical range  $W(A) = \{\langle Ax, x \rangle \mid x \in \mathbb{C}^n, \|x\| = 1\}$ .

Also recall that  $\lambda_k^{\max}$  denotes the maximum eigenvalue of  $m(\square_b^t)$  on  $\mathcal{H}_k(\mathbb{S}^3)$ . We first prove the following lower bound.

**Lemma 4.5.** *There exists a real constant  $c > 0$  so that  $\frac{1}{c}k^2 \leq \lambda_k^{\max}$ ; that is,  $\lambda_k^{\max} = \Omega(k^2)$ .*

*Proof.* For a square matrix  $A$ ,  $\sup W(A)$  is an upper bound for the eigenvalues of  $A$ . Furthermore, if  $A$  is Hermitian then the maximum eigenvalue equals  $\sup W(A)$ .

Let  $A = m(\square_b^t)$  on  $W_i$ . By the above discussion, since  $A$  is similar to a Hermitian matrix  $B$ , it suffices to show that  $\sup W(B) = \Omega(k^2)$ .

Fix  $x = e_{k/2}$  for  $k$  even, and  $x = e_{(k+1)/2}$  for  $k$  odd. Since  $\langle Be_i, e_j \rangle = a'_{ij}$ , by the above matrix representation we have that for  $k$  even

$$\begin{aligned} \langle Be_{k/2}, e_{k/2} \rangle &= B_{k/2, k/2} = d_{k/2} \\ &= 4\left(\frac{k}{2}\right)\left(k - \frac{k}{2}\right) + |t|^2\left(2\frac{k}{2} - 1\right)\left(2k + 1 - 2\frac{k}{2}\right) \\ &= k^2 + |t|^2(k-1)(k+1) \\ &= \Omega(k^2). \end{aligned}$$

A similar result follows for  $k$  odd. Now since  $\langle Be_{k/2}, e_{k/2} \rangle \in W(B)$ , we have  $\sup W(B) = \Omega(k^2)$ .  $\square$

For the lower bound we invoke Gershgorin's circle theorem.

**Theorem 4.6** [Gershgorin 1931]. *Suppose  $A$  is a complex square matrix, and  $R_i$  is the sum of the absolute values of the off-diagonal entries in the  $i$ -th row. Then every eigenvalue of  $A$  must lie within one of the closed discs  $D(a_{ii}, R_i) \subset \mathbb{C}$ .*

Recall that  $m(\square_b^t)$  on  $V_i, W_i$  is similar to the real symmetric matrix  $B$ . Since  $B$  is Hermitian, Theorem 4.6 will give us interval bounds on the real line. Furthermore, the tridiagonal structure of  $B$  makes these bounds tight.

**Lemma 4.7.** *There exists a real constant  $c > 0$  so that  $\lambda_k^{\max} \leq ck^2$ ; that is,  $\lambda_k^{\max} = O(k^2)$ .*

*Proof.* Applying Theorem 4.6 on  $B$ , we have

$$D(b_{ii}, R_i) = (d_i - (c_{i-1} + c_i), d_i + (c_{i-1} + c_i)),$$

since the  $i$ -th row of  $B$  has only two off-diagonal entries,  $c_{i-1}$  and  $c_i$ , both of which are nonnegative by Theorem 4.3. Note that for the extremal cases of the first and last rows, the radii of these discs will involve only one off-diagonal entry. Now it suffices to show that an upper bound for  $M_i = d_i + c_{i-1} + c_i$  is  $O(k^2)$ . By inspection,  $c_{i-1}$ , and  $c_i$  are  $O(k^2)$  because  $u_{i-1}, u_i$  are  $O(k^4)$ . Since  $d_i$  is  $O(k^2)$  as well, we have our result.  $\square$

By combining the last two lemmas we obtain the following statement.

**Theorem 4.8.** *There exists a real  $c > 0$  so that  $\frac{1}{c}k^2 \leq \lambda_k^{\max} \leq ck^2$ ; that is,  $\lambda_k^{\max} = \Theta(k^2)$ .*

In addition to the asymptotics  $\lambda_k^{\max}$ , we computed  $\lambda_k^{\max}$  explicitly by using SymPy. Similar codes also work to compute the largest eigenvalues of other operators, such as  $\mathcal{M}_b$ , on finite-dimensional invariant spaces.

Finally we note that, in this section we studied perturbed Kohn Laplacians on  $\mathbb{S}^3$ . One can define similar perturbations on higher-dimensional spheres and investigate the corresponding spectra. Although in higher dimensions the Boutet de Monvel theorem [1975] guarantees embeddability of strongly pseudoconvex abstract CR manifolds, it would be still worthwhile to compute the distribution of eigenvalues.

### Acknowledgements

We thank the anonymous referee for constructive comments. This research was conducted at the NSF REU Site (DMS-1659203) in Mathematical Analysis and Applications at the University of Michigan-Dearborn. We would like to thank the National Science Foundation, National Security Agency, and University of Michigan-Dearborn for their support.

### References

- [Abbas et al. 2019] T. Abbas, M. M. Brown, A. Ramasami, and Y. E. Zeytuncu, “Spectrum of the Kohn Laplacian on the Rossi sphere”, *Involve* **12**:1 (2019), 125–140. MR Zbl
- [Apostol 1976] T. M. Apostol, *Introduction to analytic number theory*, Springer, 1976. MR Zbl
- [Axler et al. 1992] S. Axler, P. Bourdon, and W. Ramey, *Harmonic function theory*, Graduate Texts in Math. **137**, Springer, 1992. MR Zbl
- [Boggess 1991] A. Boggess, *CR manifolds and the tangential Cauchy–Riemann complex*, CRC Press, Boca Raton, FL, 1991. MR Zbl
- [Boutet de Monvel 1975] L. Boutet de Monvel, “Intégration des équations de Cauchy–Riemann induites formelles”, exposé 9 in *Séminaire Goulaouic–Lions–Schwartz 1974–1975*, Centre Math., École Polytech., Paris, 1975. MR Zbl
- [Burns 1979] D. M. Burns, Jr., “Global behavior of some tangential Cauchy–Riemann equations”, pp. 51–56 in *Partial differential equations and geometry* (Park City, UT, 1977), edited by C. I. Byrnes, Lecture Notes in Pure and Appl. Math. **48**, Dekker, New York, 1979. MR Zbl
- [Chen and Shaw 2001] S.-C. Chen and M.-C. Shaw, *Partial differential equations in several complex variables*, AMS/IP Studies in Adv. Math. **19**, Amer. Math. Soc., Providence, RI, 2001. MR Zbl
- [Folland 1972] G. B. Folland, “The tangential Cauchy–Riemann complex on spheres”, *Trans. Amer. Math. Soc.* **171** (1972), 83–133. MR Zbl
- [Fu 2005] S. Fu, “Hearing pseudoconvexity with the Kohn Laplacian”, *Math. Ann.* **331**:2 (2005), 475–485. MR Zbl
- [Fu 2008] S. Fu, “Hearing the type of a domain in  $\mathbb{C}^2$  with the  $\bar{\partial}$ -Neumann Laplacian”, *Adv. Math.* **219**:2 (2008), 568–603. MR Zbl
- [Gershgorin 1931] S. Gershgorin, “Über die Abgrenzung der Eigenwerte einer Matrix”, *Izv. Akad. Nauk SSSR* **1931**:6 (1931), 749–754. In Russian. Zbl

- [Klima 2004] O. Klima, *Analysis of a subelliptic operator on the sphere in complex  $n$ -space*, master's thesis, University of New South Wales, 2004, available at <https://tinyurl.com/klimamast>.
- [Métivier 1981] G. Métivier, "Spectral asymptotics for the  $\bar{\partial}$ -Neumann problem", *Duke Math. J.* **48**:4 (1981), 779–806. MR Zbl
- [Rossi 1965] H. Rossi, "Attaching analytic spaces to an analytic space along a pseudoconcave boundary", pp. 242–256 in *Proc. Conf. Complex Analysis* (Minneapolis, 1964), edited by A. Aeppli et al., Springer, 1965. MR Zbl

Received: 2018-09-05      Accepted: 2018-12-26

jtahn@bowdoin.edu	<i>Bowdoin College, Brunswick, ME, United States</i>
bansilmo@msu.edu	<i>Michigan State University, East Lansing, MI, United States</i>
garrettbrown@college.harvard.edu	<i>Harvard University, Cambridge, MA, United States</i>
elcardin@email.wm.edu	<i>College of William and Mary, Williamsburg, VA, United States</i>
zeytuncu@umich.edu	<i>Department of Mathematics and Statistics, University of Michigan-Dearborn, Dearborn, MI, United States</i>





# Pairwise compatibility graphs: complete characterization for wheels

Matthew Beaudouin-Lafon, Serena Chen, Nathaniel Karst,  
Denise Sakai Troxell and Xudong Zheng

(Communicated by Ann N. Trenk)

A simple graph  $G$  is a pairwise compatibility graph (PCG) if there exists an edge-weighted tree  $T$  with positive weights and nonnegative numbers  $d_{\min}$  and  $d_{\max}$  such that the leaves of  $T$  are exactly the vertices of  $G$ , and  $uv$  is an edge in  $G$  if and only if the sum of weights of edges on the unique path between  $u$  and  $v$  in  $T$  is at least  $d_{\min}$  and at most  $d_{\max}$ . We show that a wheel on  $n$  vertices is a PCG if and only if  $n \leq 8$ , settling an open problem proposed by Calamoneri and Sinimeri (*SIAM Review* **58**:3 (2016), 445–460). Our approach is based on unavoidable binary classifications of the edges in the complement of wheels that are PCGs. (Note: during the review process of our work, we learned that the same result has been obtained independently with an alternative proof.)

## 1. Introduction

Edge-weighted rooted trees are common graph models used in phylogenetics, a branch of biology that studies the evolutionary history and relationships of sets of taxa, i.e., organisms sharing similar characteristics (e.g., species, populations). In such a phylogenetic tree, a leaf represents a taxon, an internal vertex represents a possible common ancestor of its descendant leaves, and the weight of an edge may be interpreted as the length of the evolutionary history separating the species or populations represented by its two incident vertices. One of the first illustrations of a phylogenetic tree appeared in Charles Darwin’s groundbreaking work [1859].

In computational biology, the problem of reconstructing an optimal phylogenetic tree from a given set of taxa is complex [Calamoneri and Sinimeri 2016], and so researchers have focused on constrained instances of this problem. For example, since very large and very small distances between pairs of taxa in the evolutionary history may have a negative impact on the performance of reconstruction algorithms, bounding these distances is a natural constraint [Kearney et al. 2003]. In graph-theoretical

---

MSC2010: 05C12, 05C78.

Keywords: pairwise compatibility graph, PCG, phylogenetic tree, wheel.

terms, let  $G$  be a graph where each vertex represents a taxon and  $uv$  be an edge in  $G$  if the evolutionary distance between vertices  $u$  and  $v$  is within an acceptable range. One is interested in finding an edge-weighted tree  $T$  with positive weights and nonnegative numbers  $d_{\min}$  and  $d_{\max}$  such that the set of leaves of  $T$  is exactly the set of vertices of  $G$ , and  $uv$  is an edge in  $G$  if and only if the sum of weights of edges on the unique path between  $u$  and  $v$  in  $T$  is at least  $d_{\min}$  and at most  $d_{\max}$ . If such  $T$ ,  $d_{\min}$  and  $d_{\max}$  exist, then we say that  $G$  is a *pairwise compatibility graph* (PCG) with *witness tree*  $T$  bounded by  $d_{\min}$  and  $d_{\max}$ , or simply  $G = \text{PCG}(T, d_{\min}, d_{\max})$ . For any two vertices  $u$  and  $v$  in  $G$  (not necessarily adjacent),  $d(u, v)$  will denote the sum of weights of the edges on the unique path in  $T$  between the leaves  $u$  and  $v$  (for simplicity, we omitted the subscript in  $d_T(u, v)$  which is traditionally used to denote the weighted distance between any pair of vertices  $u$  and  $v$  in  $T$ ).

The literature suggests that the PCG recognition problem is difficult, and it has been conjectured to be NP-hard [Durocher et al. 2015]. Since no complete characterization of PCGs is currently known, a large portion of the existing research has focused on determining whether particular graphs are PCGs or not. The following are some examples of the known classes of PCGs: graphs with at most seven vertices [Calamoneri et al. 2013a; Phillips 2002]; bipartite graphs with at most eight vertices [Mehnaz and Rahman 2013]; cycles, single-chord cycles, cacti, tree power graphs, Steiner and phylogenetic  $k$ -power graphs [Mehnaz and Rahman 2013; Yanhaona et al. 2009]; trees, ladders, triangle-free outerplanar 3-graphs [Salma et al. 2013]; Dilworth 2 graphs [Calamoneri and Petreschi 2014]; split matrogenic graphs and certain superclasses [Calamoneri et al. 2013b]. Some particular graphs that are not PCGs have also been identified: a nonbipartite circular arc graph on 8 vertices, a bipartite graph on 15 vertices, and a planar graph on 20 vertices [Yanhaona et al. 2009; 2010]. Recently, two results involving the complement  $G^c$  of a graph  $G$  provided additional tools in the study of PCGs [Hossain et al. 2017]: if  $G^c$  is acyclic then  $G$  is a PCG; if  $G^c$  contains two vertex-disjoint chordless cycles without an edge simultaneously incident to both cycles, then  $G$  is not a PCG. One instance relevant to our work is the class of *k-leaf power graphs* which are PCGs, where  $d_{\min} = 0$  and  $d_{\max} = k$ . It is well known that these graphs are strongly chordal, i.e., chordal and sun-free [Farber 1983]; however, the converse is not true [Bibelnieks and Dearing 1993]. In fact, no complete characterization of  $k$ -leaf power graphs is known except when  $k \leq 4$  [Brandstädt and Le 2006; Brandstädt et al. 2008; Dom et al. 2004; 2005; Rautenbach 2006].

From the references above and from our recent experience, we have learned that many of the existing results concerning the PCG recognition problem required determination and clever, nontrivial approaches to generate witness trees or to show that none exist. Nevertheless, the efforts behind these approaches may not be readily apparent since they often describe witness trees without providing a clear

discussion of what drives their particular structures. Perhaps for these reasons there are still many open problems in the area, as mentioned in the comprehensive survey [Calamoneri and Sinaimeri 2016], including the following:

**Open Problem 1.** Find other graph classes that do not belong to the PCG class.

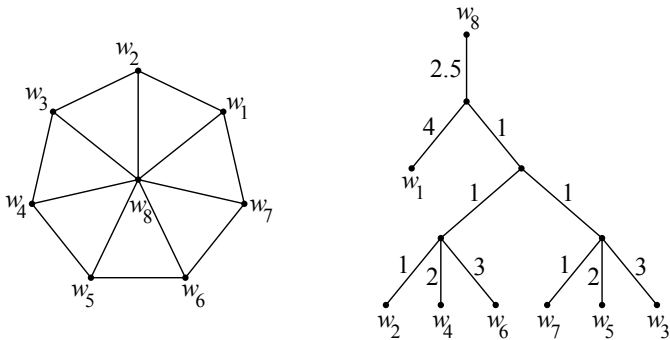
**Open Problem 2.** It is not known whether or not wheels on at least eight nodes are PCGs.

We add one more class for Open Problem 1 while settling Open Problem 2 in our main result:

**Theorem 1.1.** *Wheels on  $n$  vertices are PCGs if and only if  $n \leq 8$ .*

We will be using the following notation throughout this work. The *wheel*  $W_n$  with order  $n \geq 4$  has vertices  $w_1, w_2, \dots, w_n$ , edges  $w_i w_n$  for  $i = 1, 2, \dots, n - 1$ , edges  $w_i w_{i+1}$  for  $i = 1, 2, \dots, n - 2$ , and edge  $w_1 w_{n-1}$ . The cycle induced by the vertices  $w_1, w_2, \dots, w_{n-1}$  is called the *rim* of the wheel.

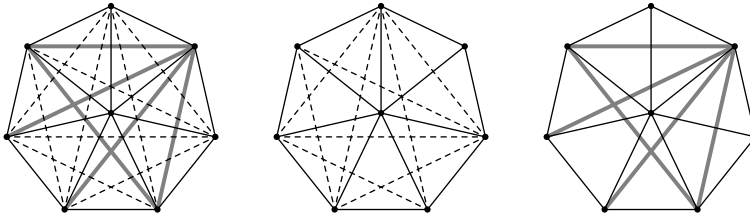
Figure 1 shows the wheel  $W_8$  and a witness tree  $T$  bounded by  $d_{\min} = 5.5$  and  $d_{\max} = 7.5$ , that is,  $W_8 = \text{PCG}(T, 5.5, 7.5)$ . This claim can be easily verified using the information in Table 1, where for each entry  $(i, j)$ , the corresponding column header is  $d(w_i, w_j)$  for  $T$  in Figure 1 (pairs in bold correspond to the edges in  $W_8$ ).



**Figure 1.** Wheel  $W_8$  on the left and a witness tree  $T$  on the right with  $W_8 = \text{PCG}(T, 5.5, 7.5)$ .

3	4	5	5.5	6	6.5	7	7.5	8	9
(2,4)	(2,6)	(2,5)	<b>(2,8)</b>	<b>(2,3)</b>	<b>(1,8)</b>	<b>(1,2)</b>	<b>(3,8)</b>	(1,4)	(1,3)
(5,7)	(2,7)	(3,5)	<b>(7,8)</b>	<b>(4,5)</b>	<b>(4,8)</b>	<b>(1,7)</b>	<b>(6,8)</b>	(1,5)	(1,6)
	(3,7)	(4,6)		<b>(6,7)</b>	<b>(5,8)</b>	<b>(3,4)</b>		(3,6)	
		(4,7)			<b>(5,6)</b>				

**Table 1.** For each entry  $(i, j)$ , the corresponding column header is  $d(w_i, w_j)$  for  $T$  in Figure 1 (pairs in bold correspond to the edges in  $W_8$ ).

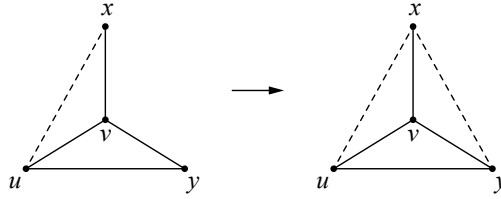


**Figure 2.** Edges in  $W_8$  are solid black, light edges in  $(W_8)^c$  are dashed and heavy edges in  $(W_8)^c$  are thick gray.

Generating this witness tree for  $W_8$  was far from trivial. A brute-force computation approach was infeasible due to the large number of trees with eight leaves and the infinite number of choices for their edge-weights and bounds. We relied on potential binary classifications of the edges in the complement  $(W_8)^c$  of  $W_8$ , more specifically, which edge  $uv$  in  $(W_8)^c$  could be *light*, i.e.,  $d(u, v) < d_{\min}$ , and which could be *heavy*, i.e.,  $d(u, v) > d_{\max}$ . Using general results that do not require the knowledge of an exact witness tree and bounds, we generated the configuration of light and heavy edges given on the left-most graph in Figure 2, where edges in  $W_8$  are solid black, light edges in  $(W_8)^c$  are dashed, and heavy edges in  $(W_8)^c$  are thick gray. The center and right-most graphs in this figure are provided for clarity and show  $W_8$  together with only light and with only heavy edges, respectively. The exact steps to obtain this configuration are omitted, as they are similar to the steps presented in the proof of Theorem 2.6 in Section 2. From this configuration, we were able to obtain the witness tree  $T$  and bounds in Figure 1 by inspection.

Recall that all graphs with at most seven vertices are PCGs. Theorem 1.1 will follow, given that we have shown here that  $W_8$  is also a PCG and will show in Section 2 that no  $W_n$  for  $n \geq 9$  is a PCG.

During the review process of our work, we learned that Theorem 1.1 has been verified independently in the arXiv manuscript [Baiocchi et al. 2017] which was later presented as the conference extended abstract [Baiocchi et al. 2018]. In [Baiocchi et al. 2018], the edges of a PCG are colored black, and edges in the complement are colored red if they are light and white if they are heavy. Several forbidden tricolored structures are identified. The general approach assumes that  $W_n$  for  $n \geq 9$  is a PCG and these forbidden structures are used in an exhaustive case discussion to reach a contradiction. Our approach is similar in the sense that it focuses on certain unavoidable binary configurations of edges and, indeed, one of the forbidden structures identified in [Baiocchi et al. 2018] (namely  $\mathbf{f-c}(2K_2)a$ , coincides with the configuration  $H_5$  described in our Lemma 2.4). Nevertheless, we believe our proof streamlines the case discussion by generating a sequence of unavoidable light edges until the forbidden configuration  $H_5$  is achieved.



**Figure 3.** Configuration  $H_1$  (left), where  $xy$  is an edge in  $G^c$  and  $d(u, v) \geq d(v, x)$ , implies  $H_2$  (right), as shown in Lemma 2.2.

## 2. Wheels with more than eight vertices are not PCGs

Key to our discussion is the following useful result that allows for distance comparisons between certain pairs of leaves in general edge-weighted trees.

**Result 2.1** [Yanhaona et al. 2010]. Let  $T$  be an edge-weighted tree and let  $u, v, x$  be three leaves in  $T$  such that  $d(u, v) = \max\{d(u, v), d(v, x), d(x, u)\}$ . If  $y$  is a leaf other than  $u, v, x$ , then  $d(x, y) \leq d(u, y)$  or  $d(x, y) \leq d(v, y)$ .  $\square$

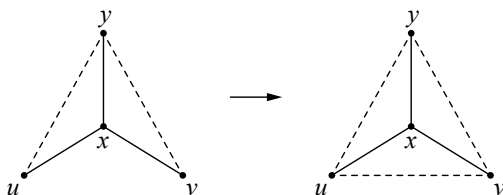
We will apply Result 2.1 to the witness trees of certain PCGs in Lemmas 2.2, 2.3 and 2.4. These lemmas will be vital tools used to show that  $W_n$  is not a PCG when  $n \geq 9$  in Theorem 2.6. We first extend the definitions of light and heavy edges mentioned in Section 1 to general PCGs; that is, given  $G = \text{PCG}(T, d_{\min}, d_{\max})$ , we say that an edge  $uv$  in  $G^c$  is *light* if  $d(u, v) < d_{\min}$  and is *heavy* if  $d(u, v) > d_{\max}$ . Any future figures will continue using the conventions given in Figure 2: edges in  $G$  are solid black, light edges in  $G^c$  are dashed, and heavy edges in  $G^c$  are thick gray.

**Lemma 2.2.** *Let  $G = \text{PCG}(T, d_{\min}, d_{\max})$ . If  $G$  and  $G^c$  contain the edges in the configuration  $H_1$  in Figure 3 (left), where  $xy$  is an edge in  $G^c$  and  $d(u, v) \geq d(v, x)$ , then  $xy$  must be light as indicated in the configuration  $H_2$  in Figure 3 (right).*

*Proof.* Since  $d(u, v) \geq d(v, x)$  and  $xu$  is light, we have  $d(u, v) = \max\{d(u, v), d(v, x), d(x, u)\}$ . By Result 2.1,  $d(x, y) \leq d(u, y)$  or  $d(x, y) \leq d(v, y)$ . But  $d(u, y) \leq d_{\max}$  and  $d(v, y) \leq d_{\max}$  because  $uy$  and  $vy$  are edges in  $G$ , therefore  $d(x, y) \leq d_{\max}$ . This latter inequality combined with the fact that  $xy$  is an edge in  $G^c$  implies  $d(x, y) < d_{\min}$  and therefore  $xy$  is light.  $\square$

**Lemma 2.3.** *Let  $G = \text{PCG}(T, d_{\min}, d_{\max})$ . If  $G$  and  $G^c$  contain the edges in the configuration  $H_3$  in Figure 4 (left), where  $uv$  is an edge in  $G^c$ , then  $uv$  must be light as indicated in the configuration  $H_4$  in Figure 4 (right).*

*Proof.* Suppose by contradiction that  $uv$  is heavy. Since  $xu$  and  $vx$  are edges in  $G$ , we must have  $d(x, u) \leq d_{\max}$  and  $d(v, x) \leq d_{\max}$ ; hence  $d(u, v) = \max\{d(u, v), d(v, x), d(x, u)\}$  and by Result 2.1,  $d(x, y) \leq d(u, y)$  or  $d(x, y) \leq d(v, y)$ . But  $uy$



**Figure 4.** Configuration  $H_3$  (left), where  $uv$  is an edge in  $G^c$ , implies  $H_4$  (right), as shown in Lemma 2.2.

and  $vy$  are light, that is,  $d(u, y) < d_{\min}$  and  $d(v, y) < d_{\min}$ , which would imply  $d(x, y) < d_{\min}$ , contradicting the fact that  $d(x, y) \geq d_{\min}$  as  $xy$  is an edge in  $G$ .  $\square$

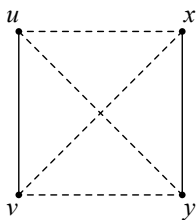
For later discussions, it is important to note the differences between the vertex labels in Figures 3 and 4 (e.g.,  $v$  and  $x$  are the only vertices of degree 3 in each respective figure). These labels were chosen so that Result 2.1 could be readily applied in the proofs of Lemmas 2.2 and 2.3, respectively.

**Lemma 2.4.** *Let  $G = \text{PCG}(T, d_{\min}, d_{\max})$ .  $G$  and  $G^c$  cannot contain the configuration  $H_5$  of Figure 5.*

*Proof.* Suppose by contradiction that  $G$  and  $G^c$  contain the configuration  $H_5$ . Since  $xu$  and  $vx$  are light, we must have  $d(x, u) < d_{\min}$  and  $d(v, x) < d_{\min}$ . But  $uv$  is an edge in  $G$ , so we have  $d(u, v) \geq d_{\min}$ . Hence  $d(u, v) = \max\{d(u, v), d(v, x), d(x, u)\}$  and by Result 2.1,  $d(x, y) \leq d(u, y)$  or  $d(x, y) \leq d(v, y)$ . But  $uy$  and  $vy$  are light, that is,  $d(u, y) < d_{\min}$  and  $d(v, y) < d_{\min}$ , which would imply  $d(x, y) < d_{\min}$ , contradicting the fact that  $d(x, y) \geq d_{\min}$  as  $xy$  is an edge in  $G$ .  $\square$

In the proof of Theorem 2.6, we will assume by contradiction that  $W_n$  is a PCG for some  $n \geq 9$  and apply Lemmas 2.2 and 2.3 repeatedly until a contradiction to Lemma 2.4 is reached. To be able to set this argument in motion, we need to verify the existence of a particular light edge. For each  $p = 2, 3, \dots, n-3$ , we define a  $p$ -light edge in  $(W_n)^c$  to be a light edge with ends connected by a path on the rim of  $W_n$  with exactly  $p$  edges (note that a  $p$ -light edge is also an  $(n-p-1)$ -light edge).

**Lemma 2.5.** *If  $n \geq 5$  and  $W_n = \text{PCG}(T, d_{\min}, d_{\max})$ , then there exists a  $p$ -light edge for each  $p = 2, 3, \dots, n-3$ .*



**Figure 5.** Forbidden configuration  $H_5$  in Lemma 2.4.

*Proof.* Let  $W_n = \text{PCG}(T, d_{\min}, d_{\max})$  with  $n \geq 5$ . Since  $p$ -light edges are  $(n-p-1)$ -light edges, it is enough to verify the lemma for  $p = 2, 3, \dots, \lfloor (n-1)/2 \rfloor$ . We will proceed by induction on  $p$ .

The rim of  $W_n$  is a chordless cycle; hence  $W_n$  is not chordal and consequently not strongly chordal. Recall from Section 1 that  $k$ -leaf power graphs are strongly chordal so  $W_n$  is not a  $k$ -leaf power graph; that is,  $d_{\min} > 0$  and there exists at least one light edge (if there are no light edges, then  $uv$  would be an edge in  $G$  if and only if  $0 \leq d(u, v) \leq d_{\max}$ , and hence  $W_n$  would be a  $k$ -leaf power graph). Choose a light edge with ends that minimize the distance on the rim of  $W_n$  (i.e., the number of edges on the shortest path between these ends using only edges on the rim) over all light edges, and let  $m$  be this smallest distance. We may assume without loss of generality that  $w_1 w_{m+1}$  is this selected light edge and  $d(w_{m+1}, w_n) \geq d(w_1, w_n)$  (if not, rotate and/or reverse the labels on the rim). Clearly,  $2 \leq m \leq \lfloor (n-1)/2 \rfloor$ . If  $m > 2$ , since  $w_1 w_m$  is an edge in  $(W_n)^c$  and  $d(w_{m+1}, w_n) \geq d(w_1, w_n)$ , then applying Lemma 2.2 with  $u = w_{m+1}$ ,  $v = w_n$ ,  $x = w_1$ ,  $y = w_m$  would imply  $xy = w_1 w_m$  is light with ends connected by a path on the rim with  $m-1$  edges, which contradicts the minimality of  $m$ . Hence  $m = 2$ ; that is,  $w_1 w_3$  is light with ends connected by the path  $w_1 w_2 w_3$  on the rim. Thus, there is a 2-light edge in  $W_n$ , and the basis of the induction has been established.

Assume for  $2 \leq p < \lfloor (n-1)/2 \rfloor$  that there exists a  $p$ -light edge and we will show that there exists a  $(p+1)$ -light edge, concluding our inductive argument. Rotate and/or reverse the labels on the rim so that  $w_1 w_{p+1}$  is this  $p$ -light edge and  $d(w_{p+1}, w_n) \geq d(w_1, w_n)$ . Note that since  $n \geq 5$  and  $p < \lfloor (n-1)/2 \rfloor$ , we have  $p+2 < \lfloor (n-1)/2 \rfloor + 2 \leq n-1$  so  $w_1 w_{p+2}$  is an edge in  $(W_n)^c$ . Applying Lemma 2.2 with  $u = w_{p+1}$ ,  $v = w_n$ ,  $x = w_1$ ,  $y = w_{p+2}$  we conclude that  $xy = w_1 w_{p+2}$  is a  $(p+1)$ -light edge, and so our induction is complete.  $\square$

We can confirm that this lemma holds in the instance of  $W_8$  presented in Figure 2; for example,  $w_1 w_3$  is a 2- and 5-light edge, and  $w_1 w_4$  is a 3- and 4-light edge.

Applications of Lemma 2.2 similar to the two discussed in the proof of Lemma 2.5 will occur multiple times in the proof of Theorem 2.6, and so we will use the abbreviated notation  $(i, j, k) \xrightarrow{2,2} (j, k)$  to indicate that  $w_j w_k$  is an edge in  $(W_n)^c$ ,  $d(w_i, w_n) \geq d(w_j, w_n)$ , and setting  $u = w_i$ ,  $v = w_n$ ,  $x = w_j$ ,  $y = w_k$  we have the configuration  $H_1$  in Figure 3 (left); therefore applying Lemma 2.2 implies  $xy = w_j w_k$  is light. With this notation, the two applications of Lemma 2.2 in the proof of Lemma 2.5 would simply read  $(m+1, 1, m) \xrightarrow{2,2} (1, m)$  and  $(p+1, 1, p+2) \xrightarrow{2,2} (1, p+2)$ , respectively. In the same spirit, we also define the abbreviated notation  $(i, j, k) \xrightarrow{2,3} (i, j)$  to indicate that  $w_i w_j$  is an edge in  $(W_n)^c$  and setting  $u = w_i$ ,  $v = w_j$ ,  $x = w_n$ ,  $y = w_k$  we have the configuration  $H_3$  in Figure 4 (left); therefore applying Lemma 2.3 implies  $uv = w_i w_j$  is light.

**Theorem 2.6.** *If  $n \geq 9$ , then  $W_n$  is not a PCG.*

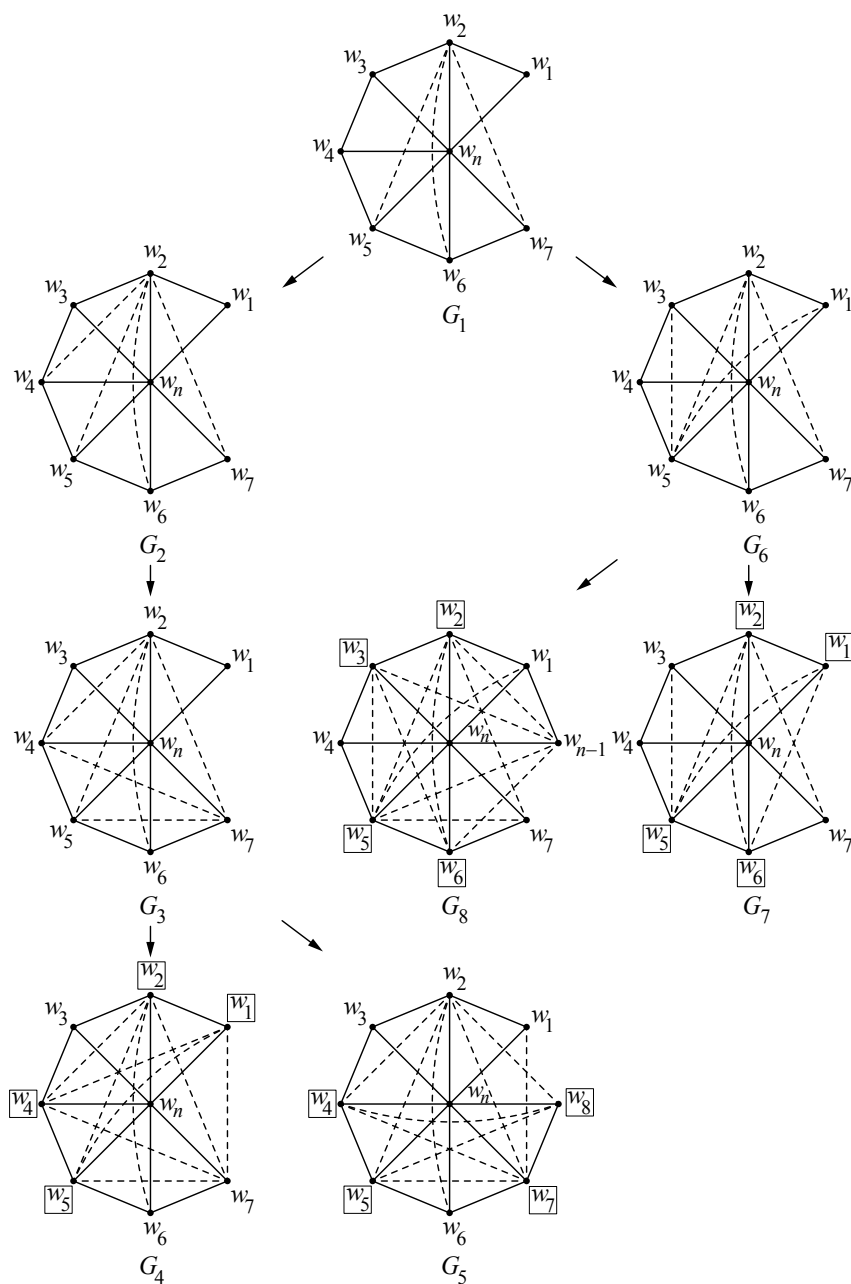
*Proof.* Let  $n \geq 9$  and suppose by contradiction that  $W_n = \text{PCG}(T, d_{\min}, d_{\max})$ . From Lemma 2.5, there exists a 4-light edge. We may assume without loss of generality that  $w_2w_6$  is a 4-light edge and that  $d(w_6, w_n) \geq d(w_2, w_n)$  (if not, rotate and/or reverse the labels on the rim). The proof proceeds by adding light edges forced by Lemmas 2.2 and 2.3 until we reach the configuration  $H_5$  featured in Figure 5, which would contradict Lemma 2.4. We begin by observing that  $(6, 2, 5) \xrightarrow{2,2} (2, 5)$  and  $(6, 2, 7) \xrightarrow{2,2} (2, 7)$ . The three current light edges are shown in the configuration  $G_1$  of Figure 6. We split the discussion into two cases:

Case 1: Suppose  $d(w_5, w_n) \geq d(w_2, w_n)$ . Hence  $(5, 2, 4) \xrightarrow{2,2} (2, 4)$ , with current light edges shown in the configuration  $G_2$  of Figure 6. In addition,  $(4, 7, 2) \xrightarrow{2,3} (4, 7)$  and  $(5, 7, 2) \xrightarrow{2,3} (5, 7)$ , with current light edges shown in the configuration  $G_3$  of Figure 6. Let us first examine the subcase where  $d(w_7, w_n) < d(w_2, w_n)$ . Since  $n \geq 9$ , we have that  $w_1w_7$  is an edge in  $(W_n)^c$  and is in fact a light edge, since  $(2, 7, 1) \xrightarrow{2,2} (7, 1)$ . We then have  $(1, 4, 7) \xrightarrow{2,3} (1, 4)$  and  $(1, 5, 7) \xrightarrow{2,3} (1, 5)$ . The current light edges are shown in the configuration  $G_4$  of Figure 6 and therefore we reached the configuration  $H_5$  with  $u = w_1$ ,  $v = w_2$ ,  $x = w_4$ ,  $y = w_5$  (boxed vertices), a contradiction. We now focus on the remaining subcase where  $d(w_7, w_n) \geq d(w_2, w_n)$  and reset our current light edges to those shown in configuration  $G_3$  of Figure 6. First observe that  $(7, 2, 8) \xrightarrow{2,2} (2, 8)$ . We then have  $(4, 8, 2) \xrightarrow{2,3} (4, 8)$  and  $(5, 8, 2) \xrightarrow{2,3} (5, 8)$ . The current light edges are shown in the configuration  $G_5$  of Figure 6 and therefore we reached the configuration  $H_5$  with  $u = w_4$ ,  $v = w_5$ ,  $x = w_7$ ,  $y = w_8$  (boxed vertices), a contradiction.

Case 2: Suppose  $d(w_5, w_n) < d(w_2, w_n)$  and reset our current light edges to those shown in configuration  $G_1$  of Figure 6. Hence  $(2, 5, 1) \xrightarrow{2,2} (5, 1)$  and  $(2, 5, 3) \xrightarrow{2,2} (5, 3)$  with current light edges shown in the configuration  $G_6$  of Figure 6. Let us first examine the subcase where  $d(w_5, w_n) \geq d(w_1, w_n)$ , thus  $(5, 1, 6) \xrightarrow{2,2} (1, 6)$ . The current light edges are shown in the configuration  $G_7$  of Figure 6 and therefore we reached the configuration  $H_5$  with  $u = w_1$ ,  $v = w_2$ ,  $x = w_5$ ,  $y = w_6$  (boxed vertices), a contradiction. We now focus on the remaining subcase where  $d(w_5, w_n) < d(w_1, w_n)$  and reset our current light edges to those shown in configuration  $G_6$  of Figure 6. First observe that  $(1, 5, n-1) \xrightarrow{2,2} (5, n-1)$ . We then have  $(2, n-1, 5) \xrightarrow{2,3} (2, n-1)$  and  $(3, n-1, 5) \xrightarrow{2,3} (3, n-1)$ . Now we have  $(6, n-1, 2) \xrightarrow{2,3} (6, n-1)$  (note that  $w_6w_{n-1}$  is an edge in  $(W_n)^c$  since  $n \geq 9$ ) and can finally conclude that  $(3, 6, n-1) \xrightarrow{2,3} (3, 6)$ . The current light edges are shown in the configuration  $G_8$  of Figure 6 and therefore we reached the configuration  $H_5$  with  $u = w_2$ ,  $v = w_3$ ,  $x = w_5$ ,  $y = w_6$  (boxed vertices), a contradiction.

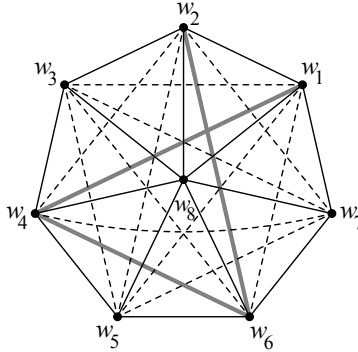
Since contradictions were reached in all possible cases, the theorem holds.  $\square$





**Figure 6.** Configurations from the proof of Theorem 2.6.

A series of steps based on Lemmas 2.2, 2.3, and 2.4, similar to the ones described in the proof of Theorem 2.6, could be applied to  $W_8$  to construct complete configurations of light and heavy edges that do not contain  $H_5$  of Figure 5. After



**Figure 7.** Invalid configuration for  $W_8$  and  $(W_8)^c$  in Lemma 2.7.

exhaustive case discussions (omitted for the sake of brevity), we found only two of these configurations, namely the configurations in Figures 2 and 7. The former allowed us to prove that  $W_8$  is a PCG as shown in Section 1. Interestingly, the latter is not a valid configuration for  $W_8$  and  $(W_8)^c$  as verified in Lemma 2.7.

**Lemma 2.7.** *If  $W_8 = \text{PCG}(T, d_{\min}, d_{\max})$ , then its corresponding light and heavy edges cannot be described by the configuration in Figure 7 (up to rotating and/or reversing the vertex labels on the rim).*

*Proof.* Suppose the lemma does not hold. We examine three cases:

Case 1:  $d(w_1, w_2) = \max\{d(w_1, w_2), d(w_2, w_8), d(w_8, w_1)\}$ . Apply Result 2.1 with  $u = w_1$ ,  $v = w_2$ ,  $x = w_8$ ,  $y = w_5$  to conclude  $d(w_8, w_5) \leq d(w_1, w_5)$  or  $d(w_5, w_8) \leq d(w_2, w_5)$ . But  $w_1w_5$  and  $w_2w_5$  are light which would imply  $d(w_8, w_5) < d_{\min}$ , contradicting the fact that  $w_8w_5$  is an edge in  $W_8$ .

Case 2:  $d(w_2, w_8) = \max\{d(w_1, w_2), d(w_2, w_8), d(w_8, w_1)\}$ . Apply Result 2.1 with  $u = w_2$ ,  $v = w_8$ ,  $x = w_1$ ,  $y = w_4$  to conclude  $d(w_1, w_4) \leq d(w_2, w_4)$  or  $d(w_1, w_4) \leq d(w_8, w_4)$ . If  $d(w_1, w_4) \leq d(w_2, w_4)$ , then  $d(w_1, w_4) < d_{\min}$  since  $w_2w_4$  is light; if  $d(w_1, w_4) \leq d(w_8, w_4)$ , then  $d(w_1, w_4) \leq d_{\max}$  since  $w_8w_4$  is an edge in  $W_8$ ; both options contradict the fact that  $w_1w_4$  is heavy.

Case 3:  $d(w_8, w_1) = \max\{d(w_1, w_2), d(w_2, w_8), d(w_8, w_1)\}$ . Given the symmetry of the configuration in Figure 7, this case can be verified as in Case 2 if we rotate the vertex labels around the rim one unit counterclockwise and then reverse their order clockwise.

Since contradictions were reached in all possible cases, the lemma holds.  $\square$

### 3. Closing remarks

We proved that  $W_8$  is a PCG, but  $W_n$  for  $n \geq 9$  are not PCGs, settling an open problem proposed in [Calamoneri and Sinimeri 2016]. The difficulty in showing

$W_8$  is a PCG stemmed from the many degrees of freedom one has in constructing potential witness trees — as both the tree’s structure and its edge weights must be specified, the collection of candidate witness trees is both very large and highly varied. A natural direction for future work would be to ask whether some subfamilies of trees could be conclusively ruled out as witness trees. Our results followed from a series of lemmas concerning light and heavy edges. While considerably distanced from the properties of any underlying witness tree, this layer of abstraction is nonetheless extremely useful. We have presented here a collection of general tools concerning configurations of heavy and/or light edges, but this set is by no means exhaustive — indeed, Lemma 2.7 hints at other families of forbidden subgraphs. We hope to see expanded results, both in terms of composition and complexity of such configurations, in the months and years to come.

### Acknowledgements

The authors would like to thank Sarah Spence Adams for handling administrative requirements regarding student research credits, and Sophia Nielsen and Robert Siegel for participating in brainstorming sessions through the second half of this research. The authors are also in debt to the Involve Editorial Board and to the reviewer for their careful consideration of our work. In addition, Denise Sakai Troxell would like to thank Babson College for its support through the Babson Research Scholar award.

### References

- [Baiocchi et al. 2017] P. Baiocchi, T. Calamoneri, A. Monti, and R. Petreschi, “Some classes of graphs that are not PCGs”, preprint, 2017. [arXiv](#)
- [Baiocchi et al. 2018] P. Baiocchi, T. Calamoneri, A. Monti, and R. Petreschi, “Graphs that are not pairwise compatible: a new proof technique”, pp. 39–51 in *Combinatorial algorithms* (Singapore, 2018), edited by C. Iliopoulos et al., Lecture Notes in Comput. Sci. **10979**, Springer, 2018. [Zbl](#)
- [Bibelnieks and Dearing 1993] E. Bibelnieks and P. M. Dearing, “Neighborhood subtree tolerance graphs”, *Discrete Appl. Math.* **43**:1 (1993), 13–26. [MR](#) [Zbl](#)
- [Brandstädt and Le 2006] A. Brandstädt and V. B. Le, “Structure and linear time recognition of 3-leaf powers”, *Inform. Process. Lett.* **98**:4 (2006), 133–138. [MR](#) [Zbl](#)
- [Brandstädt et al. 2008] A. Brandstädt, V. B. Le, and R. Sritharan, “Structure and linear-time recognition of 4-leaf powers”, *ACM Trans. Algorithms* **5**:1 (2008), art. id. 11. [MR](#) [Zbl](#)
- [Calamoneri and Petreschi 2014] T. Calamoneri and R. Petreschi, “On pairwise compatibility graphs having Dilworth number two”, *Theoret. Comput. Sci.* **524** (2014), 34–40. [MR](#) [Zbl](#)
- [Calamoneri and Sinaimeri 2016] T. Calamoneri and B. Sinaimeri, “Pairwise compatibility graphs: a survey”, *SIAM Rev.* **58**:3 (2016), 445–460. [MR](#) [Zbl](#)
- [Calamoneri et al. 2013a] T. Calamoneri, D. Frascaria, and B. Sinaimeri, “All graphs with at most seven vertices are pairwise compatibility graphs”, *Comput. J.* **56**:7 (2013), 882–886.
- [Calamoneri et al. 2013b] T. Calamoneri, R. Petreschi, and B. Sinaimeri, “On the pairwise compatibility property of some superclasses of threshold graphs”, *Discrete Math. Algorithms Appl.* **5**:2 (2013), art. id. 1360002. [MR](#) [Zbl](#)

- [Darwin 1859] C. Darwin, *On the origin of species*, Murray, London, 1859.
- [Dom et al. 2004] M. Dom, J. Guo, F. Hüffner, and R. Niedermeier, “Error compensation in leaf root problems”, pp. 389–401 in *Algorithms and computation* (Hong Kong, 2004), edited by R. Fleischer and G. Trippen, Lecture Notes in Comput. Sci. **3341**, Springer, 2004. MR Zbl
- [Dom et al. 2005] M. Dom, J. Guo, F. Hüffner, and R. Niedermeier, “Extending the tractability border for closest leaf powers”, pp. 397–408 in *Graph-theoretic concepts in computer science* (Metz, France, 2005), edited by D. Kratsch, Lecture Notes in Comput. Sci. **3787**, Springer, 2005. MR Zbl
- [Durocher et al. 2015] S. Durocher, D. Mondal, and M. S. Rahman, “On graphs that are not PCGs”, *Theoret. Comput. Sci.* **571** (2015), 78–87. MR Zbl
- [Farber 1983] M. Farber, “Characterizations of strongly chordal graphs”, *Discrete Math.* **43**:2-3 (1983), 173–189. MR Zbl
- [Hossain et al. 2017] M. I. Hossain, S. A. Salma, M. S. Rahman, and D. Mondal, “A necessary condition and a sufficient condition for pairwise compatibility graphs”, *J. Graph Algorithms Appl.* **21**:3 (2017), 341–352. MR Zbl
- [Kearney et al. 2003] P. Kearney, J. I. Munro, and D. Phillips, “Efficient generation of uniform samples from phylogenetic trees”, pp. 177–189 in *Algorithms in bioinformatics* (Budapest, 2003), edited by G. Benson and R. D. M. Page, Lecture Notes in Comput. Sci. **2812**, Springer, 2003.
- [Mehnaz and Rahman 2013] S. Mehnaz and M. S. Rahman, “Pairwise compatibility graphs revisited”, art. id. 447 in *International Conference on Informatics, Electronics and Vision* (Dhaka, Bangladesh, 2013), IEEE, Piscataway, NJ, 2013.
- [Phillips 2002] D. Phillips, *Uniform sampling from phylogenetic trees*, master’s thesis, University of Waterloo, 2002, available at <https://tinyurl.com/phylomast>.
- [Rautenbach 2006] D. Rautenbach, “Some remarks about leaf roots”, *Discrete Math.* **306**:13 (2006), 1456–1461. MR Zbl
- [Salma et al. 2013] S. A. Salma, M. S. Rahman, and M. I. Hossain, “Triangle-free outerplanar 3-graphs are pairwise compatibility graphs”, *J. Graph Algorithms Appl.* **17**:2 (2013), 81–102. MR Zbl
- [Yanhaona et al. 2009] M. N. Yanhaona, K. S. M. Tozammel Hossain, and M. S. Rahman, “Pairwise compatibility graphs”, *J. Appl. Math. Comput.* **30**:1-2 (2009), 479–503. MR Zbl
- [Yanhaona et al. 2010] M. N. Yanhaona, M. S. Bayzid, and M. S. Rahman, “Discovering pairwise compatibility graphs”, *Discrete Math. Algorithms Appl.* **2**:4 (2010), 607–623. MR Zbl

Received: 2018-09-27      Revised: 2019-01-28      Accepted: 2019-01-30

matthew.beaudouin-lafon@students.olin.edu

*Franklin W. Olin College of Engineering, Needham, MA,  
United States*

serena.chen@students.olin.edu

*Franklin W. Olin College of Engineering, Needham, MA,  
United States*

nkarst@babson.edu

*Mathematics and Sciences Division, Babson College,  
Babson Park, MA, United States*

troxell@babson.edu

*Mathematics and Sciences Division, Babson College,  
Babson Park, MA, United States*

xzheng3@gsmt.p.babson.edu

*Johns Hopkins University, Baltimore, MD, United States*

# The financial value of knowing the distribution of stock prices in discrete market models

Ayelet Amiran, Fabrice Baudoin, Skylyn Brock, Berend Coster,  
Ryan Craver, Ugonna Ezeaka, Phaniel Mariano and Mary Wishart

(Communicated by Jonathon Peterson)

An explicit formula is derived for the value of weak information in a discrete-time model that works for a wide range of utility functions, including the logarithmic utility and power utility. We assume a complete market with a finite number of assets and a finite number of possible outcomes. Explicit calculations are performed for a binomial model with two assets.

## 1. Introduction

Suppose an investor knows the distribution of the prices of the stocks in the market at a future time and this investor wants to optimize her or his expected utility from wealth at that future time. Our basic question is: *What is the financial value of this information?*

Much of the research into utility optimization and the financial value of weak information has been looked at previously in a continuous time setting [Baudoin 2003; Baudoin and Nguyen-Ngoc 2004]. The purpose of this paper is to investigate how to optimize a stock portfolio given weak information in a discrete-time setting. It should be stressed that the results we obtain are new and cannot be obtained as a consequence of the results in [Baudoin 2003; Baudoin and Nguyen-Ngoc 2004].

We will assume that the market is complete. We will also assume that there are no transactions costs. For a definition of complete markets, see [Björk 2009]. The main tool we use in finding the optimal expected utility given the weak information on future stock prices is the martingale method; see [Shreve 2004]. The reader might recognize that the problem treated here is related to robust utility maximization problems, as discussed in [Gilboa and Schmeidler 1989] and later works in mathematical finance by H. Föllmer, A. Gundel and S. Weber.

---

MSC2010: 91G10.

**Keywords:** anticipation, mathematical finance, financial value of weak information, portfolio optimization, discrete market models, insider trading.

As with classical results in this field, we will be looking at the expected utility as opposed to the expected wealth. This is an important difference to note since utility functions allow us to include an individual's attitude towards risk.

## 2. Utility functions

There are many different utility functions used in mathematics and economics to measure an individual's happiness or satisfaction. We denote our utility functions by  $U$ . We require that a utility function is strictly concave, strictly increasing, and continuously differentiable. We assume as in [Baudoin 2003] that

$$\lim_{x \rightarrow 0} U'(x) = +\infty \quad \text{and} \quad \lim_{x \rightarrow \infty} U'(x) = 0. \quad (1)$$

These conditions are sufficient for a utility function to exhibit risk aversion, to satisfy the law of diminishing marginal utility, and to guarantee that an increase in wealth results in an increase in utility. Further, when discussing the risk aversion of our utility functions, we use the absolute and relative risk aversion functions; see [Meyer and Meyer 2005]. We will be looking specifically at three different types of utility functions:

(i) Log utility:  $U(x) = \ln(x)$ ,  $x > 0$ . The log utility function has a constant relative risk aversion of 1. This implies the individual will always take on a constant proportion of risk with respect to their wealth.

(ii) Power utility:  $U(x) = x^\gamma / \gamma$  for  $-\infty < \gamma < 0$  and  $0 < \gamma < 1$  and  $x > 0$ . The power utility function also has a constant relative risk aversion, but the constant value is  $1 - \gamma$ . Thus, the power utility function is less risk-averse compared to the log utility function for  $0 < \gamma < 1$ . In this case, the constant  $\gamma$  reflects the relative risk aversion with the individual becoming more risk-averse as  $\gamma$  approaches 0. If  $-\infty < \gamma < 0$ , the individual is more risk-averse than an individual whose preferences can be described by the logarithmic utility function. As  $\gamma$  approaches  $-\infty$ , the individual becomes more and more risk-averse.

(iii) Exponential utility:  $U(x) = -e^{-\alpha x}$  for  $\alpha > 0$  and  $x \in \mathbb{R}$ . The exponential utility function has a constant absolute risk aversion of 1. Thus, the individual with an exponential utility function will assume a constant amount of risk rather than a constant proportion of risk with respect to their wealth. Notice that the exponential utility function does not satisfy the condition (1), but it is still an interesting function to note, and our results still hold true for this function.

## 3. Modeling the financial value of weak information on discrete-time complete markets with a discrete state space

**Setup.** Suppose we have a market with  $d$  financial assets, and a sample space  $\Omega_1 = \{\omega_1, \dots, \omega_M\}$  of possible outcomes of all the asset prices after one time

period. For all probability measures  $\mathbb{P}$ , we always assume  $\mathbb{P}(\omega_j) > 0$  for all  $j \in \{1, \dots, M\}$ . This is not a restriction since if  $\mathbb{P}(\omega_j) = 0$ , then we exclude  $\omega_j$  from  $\Omega_1$ . Let  $N$  be our final time period, and let  $\vec{S}_n \in \mathbb{R}^d$  denote the asset prices at time  $n$  where  $n \in \{0, 1, \dots, N\}$ . Further, let the random variable  $V_n$  denote the value of the portfolio at time  $n$ . Denote the initial wealth of the investor  $V_0$  by  $v$ . Without loss of generality we can assume one of the assets is a risk-free asset. We define  $r$  to be the rate of return of the risk-free asset. We will denote by  $\mathcal{M}$  the set of equivalent<sup>1</sup> probability measures under which discounted stock prices are martingales. Furthermore, we will assume our market is free from arbitrage. Thus, we can assume that the set  $\mathcal{M}$  is nonempty. For a complete market,  $\mathcal{M}$  is a singleton, say  $\mathcal{M} = \{\tilde{\mathbb{P}}\}$ , where  $\tilde{\mathbb{P}}$  is the unique probability measure under which discounted stock prices are martingales; see [Björk 2009] for more details about arbitrage, completeness, and equivalent martingale measures. We denote by  $\Psi^v$  the set of self-financing portfolios given initial wealth  $v$ . The probability measure  $\tilde{\mathbb{P}}$  basically represents the “knowledge” of the uninformed investor. Notice that by Jensen’s inequality this is the same as having no information at all, since it is optimal to invest in the risk-free asset only.

**3.1. Weak anticipation.** Now suppose we have some weak anticipation (weak information) regarding the prices of assets at our final time period. That is to say, we know the distribution of  $\vec{S}_N$ . We will denote this distribution by  $\nu$ . Let  $\Omega$  denote the path space of the ( $M$ -dimensional) stock price process  $\{\vec{S}_n\}_{1 \leq n \leq N}$ . Further, let  $\mathcal{A}$  be the (finite) set of possible asset prices at time  $N$ . Note  $|\mathcal{A}| \leq M^N$ .

**Definition.** The probability measure  $\mathbb{P}^\nu$  defined by

$$\mathbb{P}^\nu(\omega) := \sum_{\vec{x} \in \mathcal{A}} \tilde{\mathbb{P}}(\omega \mid \vec{S}_N = \vec{x}) \nu(\vec{S}_N = \vec{x})$$

is called the minimal probability measure associated with the weak information  $\nu$ , where  $\tilde{\mathbb{P}} \in \mathcal{M}$  is an (remember  $\mathcal{M}$  is a singleton in a complete market) equivalent martingale measure.

In the sense of the following proposition,  $\mathbb{P}^\nu$  is minimal in the set of probability measures  $\mathbb{Q}$  equivalent to  $\mathbb{P}$  such that  $\mathbb{Q}(\vec{S}_N = \vec{x}) = \nu(\vec{S}_N = \vec{x})$  for all  $\vec{x} \in \mathcal{A}$ . We denote this set by  $\mathcal{E}^\nu$ .

**Proposition 3.1.** *Let  $\phi$  be a convex function. Then*

$$\min_{\mathbb{Q} \in \mathcal{E}^\nu} \tilde{\mathbb{E}} \left[ \phi \left( \frac{d\mathbb{Q}}{d\tilde{\mathbb{P}}} \right) \right] = \tilde{\mathbb{E}} \left[ \phi \left( \frac{d\mathbb{P}^\nu}{d\tilde{\mathbb{P}}} \right) \right],$$

where  $d\mathbb{Q}/d\tilde{\mathbb{P}}$  denotes the Radon–Nikodym derivative of  $\mathbb{Q}$  with respect to  $\tilde{\mathbb{P}}$ .

<sup>1</sup>In our finite discrete sample space, by equivalent we simply mean, for all  $i \in \{1, 2, \dots, M\}$ ,  $\mathbb{Q}(\omega_i) > 0$ .

*Proof.* Let  $\vec{x} \in \mathcal{A}$  and  $\mathbb{Q} \in \mathcal{E}^\nu$  be given. Then,

$$\tilde{\mathbb{E}}\left[\frac{d\mathbb{Q}}{d\tilde{\mathbb{P}}} \mid \vec{S}_N = \vec{x}\right] = \frac{v(\vec{S}_N = \vec{x})}{\tilde{\mathbb{P}}(\vec{S}_N = \vec{x})}.$$

Let  $\phi$  be a convex function. Then from the conditional version of Jensen's inequality

$$\phi\left(\frac{v(\vec{S}_N = \vec{x})}{\tilde{\mathbb{P}}(\vec{S}_N = \vec{x})}\right) = \phi\left(\tilde{\mathbb{E}}\left[\frac{d\mathbb{Q}}{d\tilde{\mathbb{P}}} \mid \vec{S}_N = \vec{x}\right]\right) \leq \tilde{\mathbb{E}}\left[\phi\left(\frac{d\mathbb{Q}}{d\tilde{\mathbb{P}}}\right) \mid \vec{S}_N = \vec{x}\right].$$

Taking the expected value on both sides, we get

$$\tilde{\mathbb{E}}\left[\phi\left(\frac{v(S_N)}{\tilde{\mathbb{P}}(S_N)}\right)\right] = \tilde{\mathbb{E}}\left[\phi\left(\frac{d\mathbb{P}^\nu}{d\tilde{\mathbb{P}}}\right)\right] \leq \tilde{\mathbb{E}}\left[\phi\left(\frac{d\mathbb{Q}}{d\tilde{\mathbb{P}}}\right)\right],$$

and the result is proved.  $\square$

**3.2. Value of weak information.** Since an insider's anticipation has a different final time distribution than an uninformed investor's, it is natural to find a way to characterize the value of this information. Since we focused on maximizing our utility of wealth rather than the monetary value of wealth, we will define our value accordingly.

**Definition.** The *financial value of weak information* is the lowest expected utility that can be gained from anticipation. We write

$$u(v, v) = \min_{\mathbb{Q} \in \mathcal{E}^\nu} \max_{\psi \in \Psi^v} \mathbb{E}^\mathbb{Q}[U(V_N)].$$

Our main theorem is the following:

**Theorem 3.2.** *The financial value of weak information in a complete market is*

$$u(v, v) = \max_{\psi \in \Psi^v} \mathbb{E}^\psi[U(V_N)] = \mathbb{E}^\nu\left[U\left(I\left(\frac{\lambda(v)}{(1+r)^N} \frac{d\tilde{\mathbb{P}}}{d\mathbb{P}^\nu}\right)\right)\right],$$

where  $\lambda(v)$  is determined by

$$\tilde{\mathbb{E}}\left[\frac{1}{(1+r)^N} I\left(\frac{\lambda(v)}{(1+r)^N} \frac{d\tilde{\mathbb{P}}}{d\mathbb{P}^\nu}\right)\right] = v,$$

where  $\tilde{\mathbb{P}} \in \mathcal{M}$  is the unique probability measure under which the prices are martingales. Moreover, the optimal wealth at time  $n$ ,  $\widehat{V}_n$ , is given by

$$\widehat{V}_n = \frac{1}{(1+r)^{N-n}} \sum_{\omega \in \Omega} I\left(\frac{\lambda(v)}{(1+r)^N} \frac{d\tilde{\mathbb{P}}}{d\mathbb{P}^\nu}(\omega)\right) \tilde{\mathbb{P}}(\omega \mid \vec{S}_n) \quad \text{for } n \in \{0, 1, \dots, N\}.$$

At time  $n$ , the optimal amount to purchase of the  $i$ -th linearly independent asset is

$$\delta_n^i = \sum_{j=1}^M (D_{n+1}^{-1})_{i,j} \widehat{V}_{n+1}(\omega_j) \quad \text{for } n \in \{0, 1, \dots, N-1\},$$



where

$$D_{n+1} = \begin{bmatrix} S_{n+1}^1(\omega_1) & S_{n+1}^2(\omega_1) & \cdots & S_{n+1}^M(\omega_1) \\ S_{n+1}^1(\omega_2) & S_{n+1}^2(\omega_2) & \cdots & S_{n+1}^M(\omega_2) \\ \vdots & \vdots & & \vdots \\ S_{n+1}^1(\omega_M) & S_{n+1}^2(\omega_M) & \cdots & S_{n+1}^M(\omega_M) \end{bmatrix}$$

is the matrix of  $M$  linearly independent asset prices at time  $n+1$ ,  $(D_{n+1}^{-1})_{i,j}$  represents the element  $(i, j)$  of the matrix  $D_{n+1}^{-1}$ , and  $\widehat{V}_{n+1}$  comes from the above equation.

*Proof.* We will proceed by rewriting  $\max_{\psi \in \Psi^v} \mathbb{E}^{\mathbb{Q}}[U(V_N)]$ . In order to do this, we need the convex conjugate  $\widetilde{U}(y) := \max_{x>0} [U(x) - xy]$ ; see [Karatzas et al. 1991]. We form the Lagrangian for solving  $\max_{\psi \in \Psi^v} \mathbb{E}^{\mathbb{Q}}[U(V_N)]$  by

$$\mathcal{L}(\lambda) = \mathbb{E}^{\mathbb{Q}}[U(V_N)] + \lambda \left[ v - \mathbb{E}^{\mathbb{Q}} \left[ \frac{d\widetilde{\mathbb{P}}}{d\mathbb{Q}} \frac{V_N}{(1+r)^N} \right] \right].$$

Now using  $\widetilde{U}$ , substituting in for  $V_N$  from the martingale method (see the Appendix), and doing algebra, we can rewrite our Lagrangian as

$$\mathcal{L}(\lambda) = \lambda v + \widetilde{\mathbb{E}} \left[ \frac{d\mathbb{Q}}{d\widetilde{\mathbb{P}}} \widetilde{U} \left( \frac{\lambda}{(1+r)^N} \frac{d\widetilde{\mathbb{P}}}{d\mathbb{Q}} \right) \right].$$

Thus, we deduce

$$\begin{aligned} u(v, v) &= \min_{\mathbb{Q} \in \mathcal{E}^v} \min_{\lambda > 0} \left[ \lambda v + \widetilde{\mathbb{E}} \left[ \frac{d\mathbb{Q}}{d\widetilde{\mathbb{P}}} \widetilde{U} \left( \frac{\lambda}{(1+r)^N} \frac{d\widetilde{\mathbb{P}}}{d\mathbb{Q}} \right) \right] \right] \\ &= \min_{\lambda > 0} \left[ \lambda v + \min_{\mathbb{Q} \in \mathcal{E}^v} \widetilde{\mathbb{E}} \left[ \frac{d\mathbb{Q}}{d\widetilde{\mathbb{P}}} \widetilde{U} \left( \frac{\lambda}{(1+r)^N} \frac{d\widetilde{\mathbb{P}}}{d\mathbb{Q}} \right) \right] \right]. \end{aligned}$$

Since the convexity of  $\widetilde{U}$  implies the function mapping  $z \mapsto z\widetilde{U}(\lambda/((1+r)^N z))$  is convex, we can use Proposition 3.1 to get

$$u(v, v) = \min_{\lambda > 0} \left[ \lambda v + \widetilde{\mathbb{E}} \left[ \frac{d\mathbb{P}^v}{d\widetilde{\mathbb{P}}} \widetilde{U} \left( \frac{\lambda}{(1+r)^N} \frac{d\widetilde{\mathbb{P}}}{d\mathbb{P}^v} \right) \right] \right].$$

Taking the derivative now with respect to  $\lambda$  and setting it equal to 0, we find

$$v = \widetilde{\mathbb{E}} \left[ \frac{1}{(1+r)^N} I \left( \frac{\lambda^*(v)}{(1+r)^N} \frac{d\widetilde{\mathbb{P}}}{d\mathbb{P}^v} \right) \right],$$

where  $\lambda^*(v)$  is the minimizer. Now,

$$u(v, v) = \lambda^*(v)v + \widetilde{\mathbb{E}} \left[ \frac{d\mathbb{P}^v}{d\widetilde{\mathbb{P}}} \widetilde{U} \left( \frac{\lambda^*(v)}{(1+r)^N} \frac{d\widetilde{\mathbb{P}}}{d\mathbb{P}^v} \right) \right] = \mathbb{E}^v \left[ U \left( I \left( \frac{\lambda^*(v)}{(1+r)^N} \frac{d\widetilde{\mathbb{P}}}{d\mathbb{P}^v} \right) \right) \right].$$

Thus, we have shown the first part of the theorem. Now note that the discounted optimal wealth process  $\{\widehat{V}_n / (1+r)^n\}_{0 \leq n \leq N}$  is a martingale under  $\widetilde{\mathbb{P}}$  (see the Appendix). As a result,

$$\widehat{V}_n = \frac{1}{(1+r)^{N-n}} \widetilde{\mathbb{E}}[\widehat{V}_N \mid \vec{S}_n] = \frac{1}{(1+r)^{N-n}} \sum_{\omega \in \Omega} I\left(\frac{\lambda(v)}{(1+r)^N} \frac{d\widetilde{\mathbb{P}}}{d\mathbb{P}^v}(\omega)\right) \widetilde{\mathbb{P}}(\omega \mid \vec{S}_n)$$

for all  $n \in \{0, 1, \dots, N\}$ . Further, note that wealth is determined by your portfolio from the previous time period and the current prices. Thus,

$$\widehat{V}_{n+1} = D_{n+1} \vec{\delta}_n,$$

so we have

$$D_{n+1}^{-1} \widehat{V}_{n+1} = \vec{\delta}_n. \quad \square$$

**Remark.** We know from [Björk 2009] that the matrix of all asset prices in the complete market has rank  $M$ . Therefore, we can choose  $M$  linearly independent assets to invest in. Further, note that the optimal amount to purchase for each asset is only unique when  $M = d$ .

**Definition.** We define the *additional value of weak information* as the extra utility gained from investing with anticipation instead of just putting all of your wealth in the risk-free asset, which we define by

$$F(v, v) = u(v, v) - U(v(1+r)^N).$$

**Definition.** We also define the *ratio of added value to the total value* by

$$\pi(v, v) = \frac{F(v, v)}{u(v, v)} = 1 - \frac{U(v(1+r)^N)}{u(v, v)}.$$

As a consequence of Theorem 3.2 we obtain the following interpretation of the additional value of weak information for the log utility function.

**Corollary 3.3.** *The additional value of weak information for the log utility function is given by the relative entropy of  $v$  with respect to  $\widetilde{\mathbb{P}}_{\vec{S}_N}$ :*

$$F(v, v) = \mathbb{E}^v \left[ \ln \left( \frac{dv}{d\widetilde{\mathbb{P}}_{\vec{S}_N}} \right) \right].$$

*Proof.* We first solve for  $\lambda$ :

$$v = \widetilde{\mathbb{E}} \left[ \frac{1}{(1+r)^N} I \left( \frac{\lambda}{(1+r)^N} \frac{d\widetilde{\mathbb{P}}}{d\mathbb{P}^v} \right) \right] = \widetilde{\mathbb{E}} \left[ \frac{1}{(1+r)^N} \frac{(1+r)^N}{\lambda} \frac{d\mathbb{P}^v}{d\widetilde{\mathbb{P}}} \right] \Rightarrow \lambda = \frac{1}{v}.$$

Substituting for  $\lambda$  in our value of weak information equation, we thus have

$$\begin{aligned} u(v, v) &= \mathbb{E}^v \left[ U \left( I \left( \frac{\lambda}{(1+r)^N} \frac{d\tilde{\mathbb{P}}}{d\mathbb{P}^v} \right) \right) \right] \\ &= \mathbb{E}^v \left[ \ln \left( \frac{(1+r)^N}{1/v} \frac{d\mathbb{P}^v}{d\tilde{\mathbb{P}}} \right) \right] = \ln(v(1+r)^N) + \mathbb{E}^v \left[ \ln \left( \frac{d\mathbb{P}^v}{d\tilde{\mathbb{P}}} \right) \right]. \end{aligned}$$

This implies the additional value of weak information for the log utility is

$$F(v, v) = \mathbb{E}^v \left[ \ln \left( \frac{d\mathbb{P}^v}{d\tilde{\mathbb{P}}} \right) \right] = \mathbb{E}^v \left[ \ln \left( \frac{dv}{d\tilde{\mathbb{P}}_{\tilde{S}_N}} \right) \right],$$

where the last equality follows from the definition of  $\mathbb{P}^v$ .  $\square$

Just like the log utility, we can also find the financial value of weak information for the power utility.

**Corollary 3.4.** *The value of weak information for the power utility function is given by*

$$u(v, v) = \frac{v^\gamma (1+r)^{N\gamma}}{\gamma (\tilde{\mathbb{E}}[(d\tilde{\mathbb{P}}/d\mathbb{P}^v)^{1/(\gamma-1)}])^{\gamma-1}} \mathbb{E}^v \left[ \left( \frac{d\tilde{\mathbb{P}}}{d\mathbb{P}^v} \right)^{\gamma/(\gamma-1)} \right].$$

*Proof.* We now will solve for the value of  $\lambda$ :

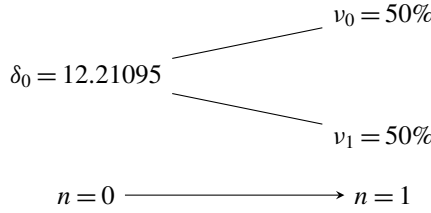
$$\tilde{\mathbb{E}} \left[ \frac{1}{(1+r)^N} \left( \frac{\lambda}{(1+r)^N} \frac{d\tilde{\mathbb{P}}}{d\mathbb{P}^v} \right)^{1/(\gamma-1)} \right] = v \quad \Rightarrow \quad \lambda = \left( \frac{v(1+r)^{N\gamma/(\gamma-1)}}{\tilde{\mathbb{E}}[(d\tilde{\mathbb{P}}/d\mathbb{P}^v)^{1/(\gamma-1)}]} \right)^{\gamma-1}.$$

Substituting in for  $\lambda$ , we get

$$\begin{aligned} u(v, v) &= \mathbb{E}^v \left[ U \left( I \left( \frac{\lambda}{(1+r)^N} \frac{d\tilde{\mathbb{P}}}{d\mathbb{P}^v} \right) \right) \right] \\ &= \mathbb{E}^v \left[ \frac{1}{\gamma} \left( \left( \frac{v(1+r)^{N\gamma/(\gamma-1)}}{\tilde{\mathbb{E}}[(d\tilde{\mathbb{P}}/d\mathbb{P}^v)^{1/(\gamma-1)}]} \right)^{\gamma-1} \frac{1}{(1+r)^N} \frac{d\tilde{\mathbb{P}}}{d\mathbb{P}^v} \right)^{\gamma/(\gamma-1)} \right] \\ &= \frac{v^\gamma (1+r)^{N\gamma}}{\gamma (\tilde{\mathbb{E}}[(d\tilde{\mathbb{P}}/d\mathbb{P}^v)^{1/(\gamma-1)}])^{\gamma-1}} \mathbb{E}^v \left[ \left( \frac{d\tilde{\mathbb{P}}}{d\mathbb{P}^v} \right)^{\gamma/(\gamma-1)} \right]. \end{aligned} \quad \square$$

#### 4. Complete markets: the binomial model

**Single-period binomial model.** We first will focus on a single-period binomial model with two assets: one risk-free asset with payoff  $1+r$ , and one risky asset with payoffs  $S_0(1+h)$  if the stock goes up, and  $S_0(1-k)$  if the stock goes down, where we assume  $S_0 > 0$  and  $k < 1$ . In order to have an arbitrage-free market, we



**Figure 1.** An example of a single-period binomial model using the log utility, where the parameter values are  $r = .032$ ,  $h = .09$ ,  $k = .019$ ,  $v = 200.0$ , and  $s = 20.0$ .

require  $h > r > -k$ . Since there is only one risky asset, we will denote the amount of units owned of the risky asset at time  $n$  by  $\delta_n$ .

Figure 1 shows a basic single-period binomial using the log utility. It represents the amount of stock you should buy initially,  $\delta_0$ . From here there are only two outcomes for our final time; the stock price will either go up or down.

**Example 1** (log utility). When looking at the log utility function, we begin by maximizing  $\mathbb{E}[U(V_N)]$  with respect to  $\delta$ . We then are able to obtain our equation for the optimal number of shares with respect to wealth,  $\hat{\delta}$ , in a single-period model:

$$\hat{\delta}_0 = \frac{v(1+r)(v_0(h-r) + v_1(-k-r))}{-s(h-r)(-k-r)}.$$

**Example 2** (power utility). As in the log utility case, we solve for our optimal number of shares with respect to wealth,  $\hat{\delta}_0$ , in a single-period model:

$$\hat{\delta}_0 = \frac{((v_0(h-r))^{1/(\gamma-1)} - (v_1(-k-r))^{1/(\gamma-1)})(1+r)v}{(v_1(-k-r))^{1/(\gamma-1)}s(-k-r) - (v_0(h-r))^{1/(\gamma-1)}s(h-r)}.$$

**Example 3** (exponential utility). Similarly to the previously examined utilities, we solve for the optimal number of shares with respect to wealth,  $\hat{\delta}$ , in a single-period model for the exponential utility:

$$\hat{\delta}_0 = \frac{\ln(v_0(h-r)) - \ln(-v_1(-k-r))}{s(h+k)}.$$

***N*-period binomial model.** In binomial models, everything can be explicitly computed. For instance, the following proposition gives the formula for the transition probabilities of the minimal probability  $\mathbb{P}^v$ . It is easy to establish by using the formula for conditional probabilities and straightforward combinatorial arguments. We note that  $\{S_n\}_{1 \leq n \leq N}$  is a Markov chain under the probability  $\mathbb{P}^v$ .

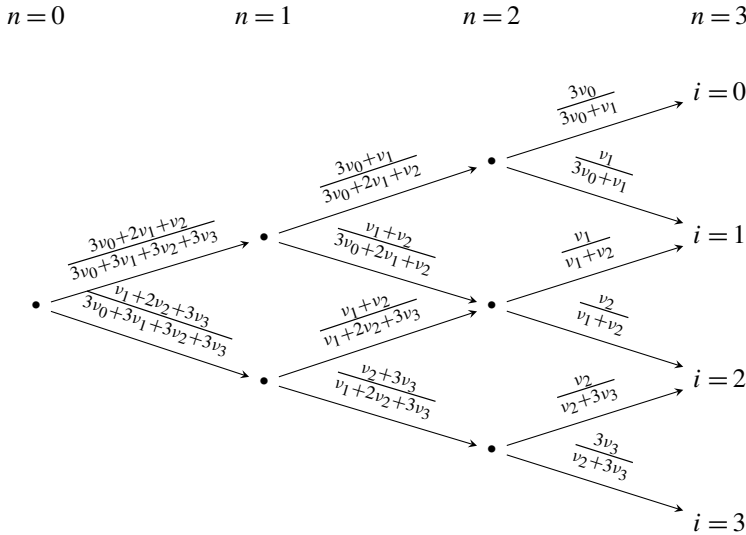
**Proposition 4.1.** Let  $l \in \{1, \dots, N-1\}$  and  $i \in \{0, \dots, N-l\}$ . Then

$$\begin{aligned} \mathbb{P}^v(S_{N-l+1} = (1+h)S_{N-l} \mid S_{N-l} = (1+h^{N-l-i})(1-k)^i S_0) \\ = \frac{\sum_{j=0}^{l-1} \binom{l-1}{j} (N-i-j) \cdots (N-i-(l-1))(i+1)(i+2) \cdots (i+j) v_{i+j}}{\sum_{j=0}^l \binom{l}{j} (N-i-j) \cdots (N-i-(l-1))(i+1)(i+2) \cdots (i+j) v_{i+j}} \end{aligned}$$

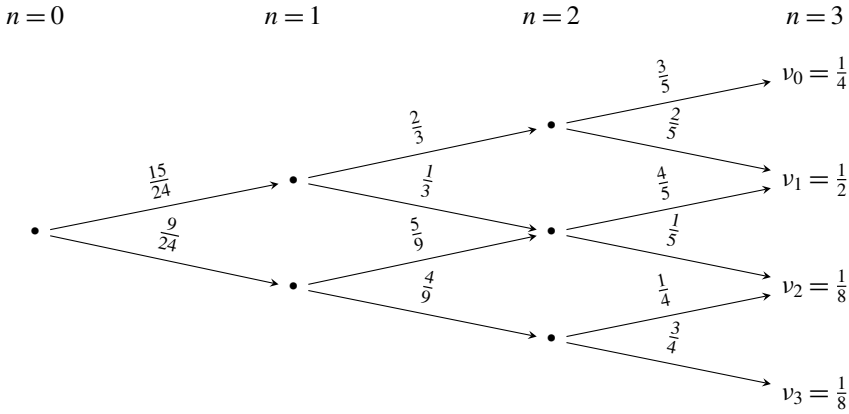
and

$$\begin{aligned} \mathbb{P}^v(S_{N-l+1} = (1-k)S_{N-l} \mid S_{N-l} = (1+h^{N-l-i})(1-k)^i S_0) \\ = \frac{\sum_{j=0}^{l-1} \binom{l-1}{j} (N-i-j-1) \cdots (N-i-(l-1))(i+1) \cdots (i+j+1) v_{i+j+1}}{\sum_{j=0}^l \binom{l}{j} (N-i-j) \cdots (N-i-(l-1))(i+1)(i+2) \cdots (i+j) v_{i+j}}. \end{aligned}$$

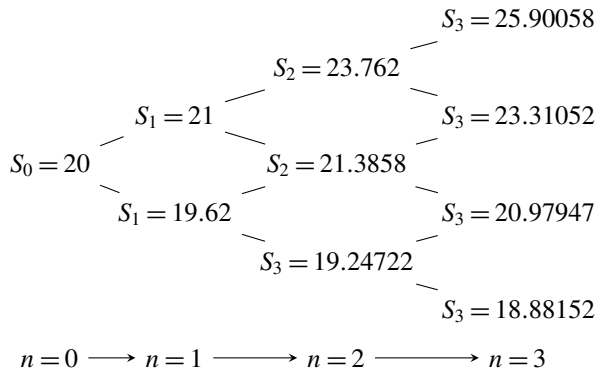
**Example 4** (log utility). Figure 5 shows an example of two different 3-period binomial trees with set values. The first tree shows the values of  $\delta$  at time  $n$  when the anticipation has a uniform distribution. The second tree, however, shows an optimistic anticipation example. One can see how the amount of stocks in which one should invest changes depending on the distribution of the anticipation. For example, one would want to buy more stocks in an optimistic model because there is a better chance of the stock increasing in price as time goes on than in the model where all of the probabilities are the same. Negative values of  $\delta$  correspond to short-selling the asset.



**Figure 2.**  $\mathbb{P}^v$  for a 3-period binomial model.



**Figure 3.**  $\mathbb{P}^\nu$  for a 3-period binomial model for a specific choice of  $\nu$ .



**Figure 4.** A 3-period binomial tree showing the values of  $S_n$ , where the parameters are  $r = .032$ ,  $h = .09$ ,  $k = .019$ ,  $v = 200.0$ , and  $s = 20.0$ .

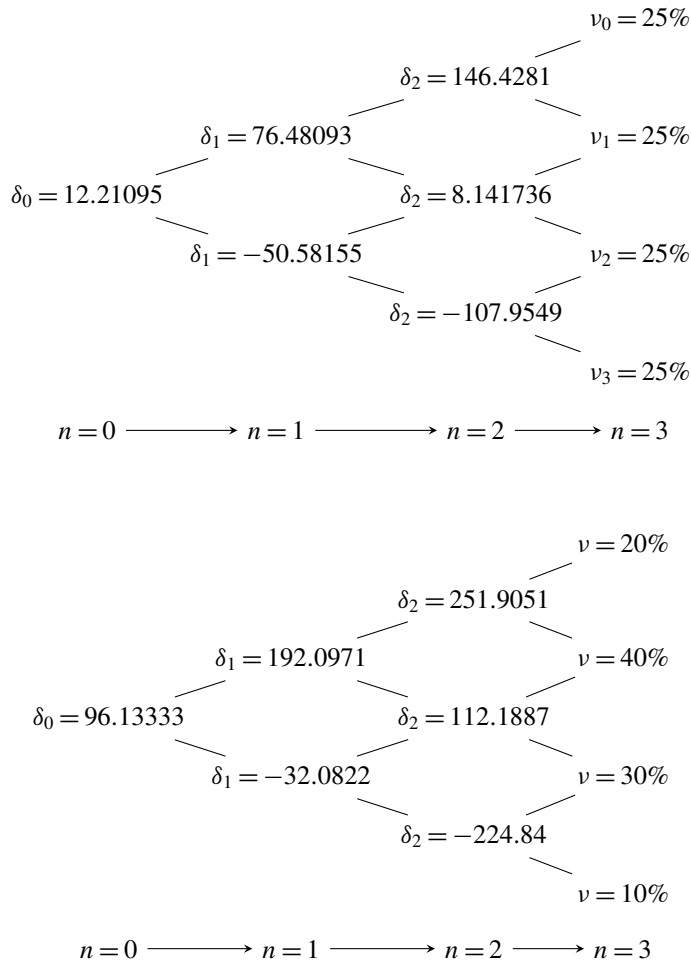
Recall from Corollary 3.3 the additional value of weak information for the log utility is

$$F(v, \nu) = \mathbb{E}^\nu \left[ \ln \left( \frac{d\mathbb{P}^\nu}{d\tilde{\mathbb{P}}} \right) \right],$$

and the proportion is

$$\pi(v, \nu) = \frac{\mathbb{E}^\nu [\ln(d\mathbb{P}^\nu / d\tilde{\mathbb{P}})]}{\ln(v(1+r)^N) + \mathbb{E}^\nu [\ln(d\mathbb{P}^\nu / d\tilde{\mathbb{P}})]}.$$

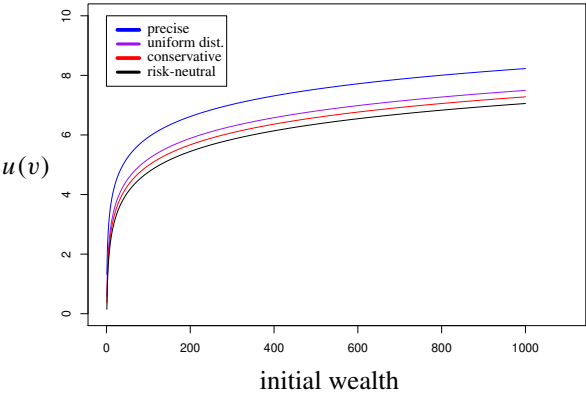
Note that  $F(v, \nu)$  is only a function of  $\nu$ , so for any fixed  $\nu$ , we have that  $F(v, \nu)$  is constant. Furthermore,  $\pi(v, \nu)$  is a decreasing function of  $v$  for any fixed  $\nu$ . As



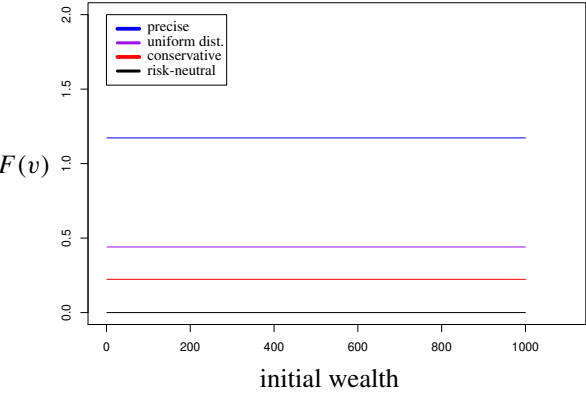
**Figure 5.** 3-period binomial trees showing the values of  $\delta$  for various anticipations of  $v$  using the log utility, where the parameters are  $r = .032$ ,  $h = .09$ ,  $k = .019$ ,  $v = 200.0$ , and  $s = 20.0$ .

a result, the wealthier you are, the less proportion of utility you are gaining as a result of anticipation. In a 5-period binomial model, with the four anticipations below, we can look at the above functions as functions of  $v$ :

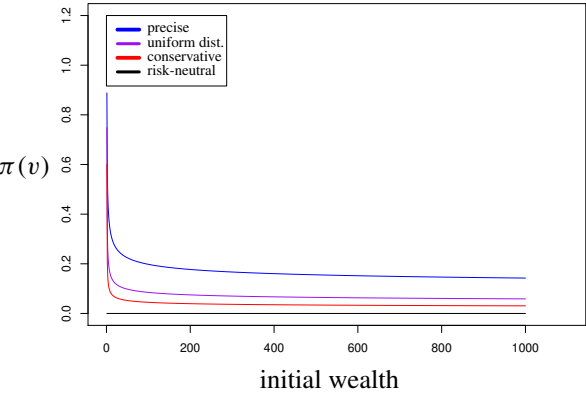
- Precise:  $\{0.01, 0.01, 0.01, 0.95, 0.01, 0.01\}$ .
- Uniform distribution:  $\{\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\}$ .
- Conservative:  $\{0.1, 0.2, 0.2, 0.2, 0.2, 0.1\}$ .
- Risk-neutral:  $v = \tilde{\mathbb{P}}$ .



**Figure 6.** Value of weak information, given  $r = 3\%$ ,  $h = 8\%$ ,  $k = 4\%$ , using the log utility. The legend labels the curves in order, top to bottom.

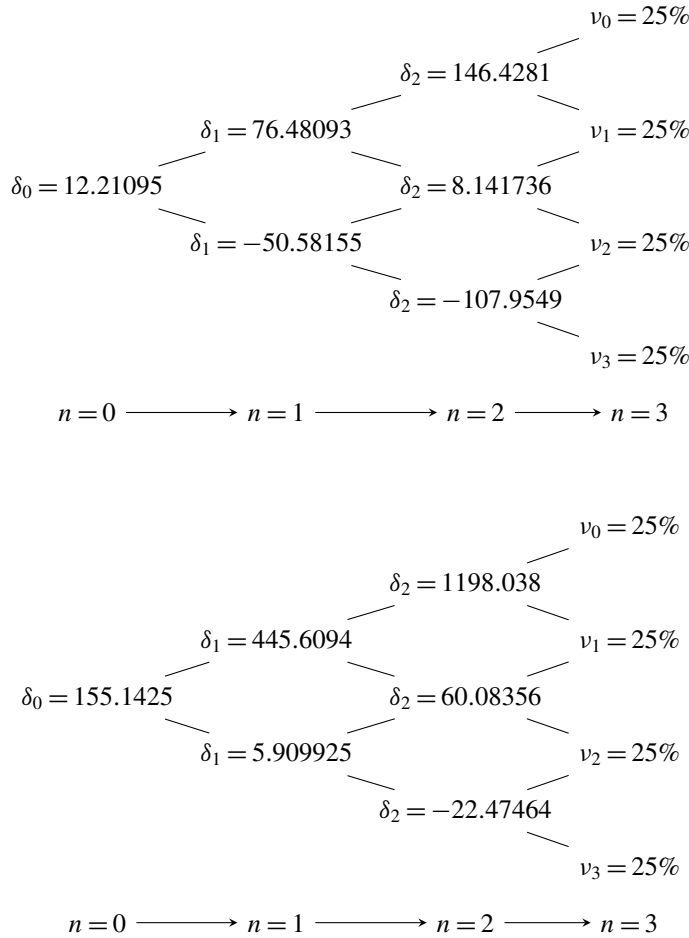


**Figure 7.** Additional value of weak information, given  $r = 3\%$ ,  $h = 8\%$ ,  $k = 4\%$ , using the log utility. Legend labels curves in order.



**Figure 8.** Proportion of value added, given  $r = 3\%$ ,  $h = 8\%$ ,  $k = 4\%$ , using the log utility. Legend labels curves in order.



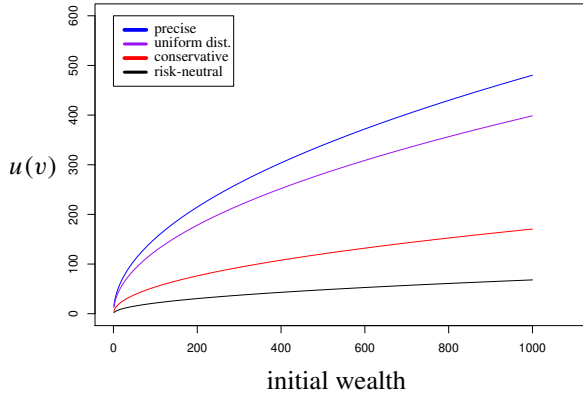


**Figure 9.** Two different 3-period binomial trees showing the values of  $\delta$  for equal anticipations of  $v$  using the log utility (top) and the power utility (bottom), where the constants are the same as Figure 5. In the power utility model,  $\gamma = .5$ .

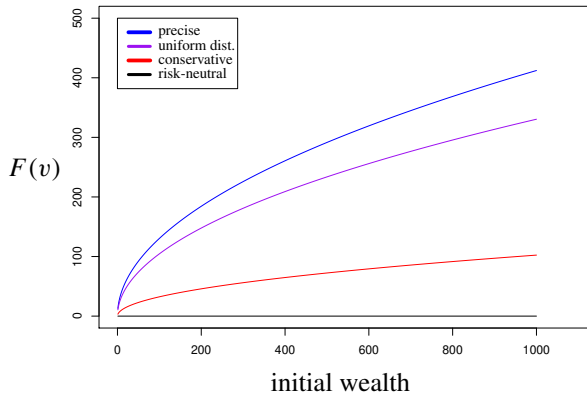
**Example 5** (power utility). Figure 9 shows the difference between the log and power utilities. As the log utility is a more relatively risk-averse utility function (for  $\gamma = 0.5$ ), the absolute value of  $\delta$  tends to be smaller when compared to the power utility function.

From Corollary 3.4 we have that the additional value for the power utility is

$$F(v, v) = \frac{v^\gamma (1+r)^{N\gamma}}{\gamma (\tilde{\mathbb{E}}[(d\tilde{\mathbb{P}}/d\mathbb{P}^v)^{1/(\gamma-1)}])^{\gamma-1}} \mathbb{E}^v \left[ \left( \frac{d\tilde{\mathbb{P}}}{d\mathbb{P}^v} \right)^{\gamma/(\gamma-1)} \right] - \frac{v^\gamma (1+r)^{N\gamma}}{\gamma},$$



**Figure 10.** Value of weak information, given  $r = 3\%$ ,  $h = 8\%$ ,  $k = 4\%$ , using the power utility. Legend labels curves in order.



**Figure 11.** Additional value of weak information, given  $r = 3\%$ ,  $h = 8\%$ ,  $k = 4\%$ , using the power utility. Legend labels curves in order.

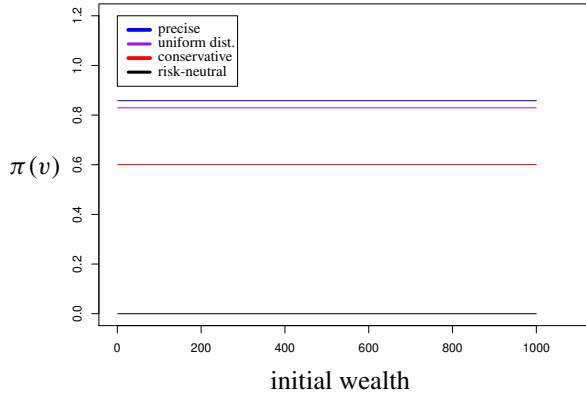
and the proportion is

$$\pi(v, v) = 1 - \frac{1}{\mathbb{E}^v[(d\tilde{\mathbb{P}}/d\mathbb{P}^v)^{\gamma/(\gamma-1)}] \cdot (\tilde{\mathbb{E}}[(d\tilde{\mathbb{P}}/d\mathbb{P}^v)^{1/(\gamma-1)}])^{1-\gamma}}.$$

For the power utility, we have the opposite relationship for a fixed  $v$  with the proportion remaining constant and the added value being an increasing function of initial wealth.

**Example 6** (exponential utility). We can also find the financial value of weak information for exponential utility.

$$\mathbb{E}^v[-e^{-a\hat{V}_N}] = e^{-v\alpha(1+r)^N - \sum_{i=0}^N \binom{N}{i} \tilde{p}^{N-i} \tilde{q}^i \ln((\binom{N}{i} \cdot \tilde{p}^{N-i} \tilde{q}^i / v_i))}.$$



**Figure 12.** Proportion of value added, given  $r = 3\%$ ,  $h = 8\%$ ,  $k = 4\%$ , using the power utility. Legend labels curves in order.

We begin by solving for  $\lambda$ .

$$\tilde{\mathbb{E}}\left[\frac{1}{(1+r)^N} I\left(\frac{\lambda}{(1+r)^N} \frac{d\tilde{\mathbb{P}}}{d\mathbb{P}^v}\right)\right] = v.$$

We use this equation and then plug in for  $I$ :

$$\tilde{\mathbb{E}}\left[\frac{1}{(1+r)^N} \frac{-1}{\alpha} \ln\left(\frac{\lambda}{\alpha(1+r)^N} \frac{d\tilde{\mathbb{P}}}{d\mathbb{P}^v}\right)\right] = v.$$

We then solve for  $\lambda$ :

$$\lambda = \alpha(1+r)^N e^{-v\alpha(1+r)^N - \mathbb{E}^v[\tilde{d}\tilde{\mathbb{P}}/d\mathbb{P}^v \ln(\tilde{d}\tilde{\mathbb{P}}/d\mathbb{P}^v)]}.$$

Finally we can plug our  $I$  and our  $\lambda$  into our equation for the financial value of weak information to solve for the value as it specifically relates to exponential utility:

$$\begin{aligned} u(v, v) &= \mathbb{E}^v\left[U\left(I\left(\frac{\lambda}{(1+r)^N} \frac{d\tilde{\mathbb{P}}}{d\mathbb{P}^v}\right)\right)\right] \\ &= \mathbb{E}^v\left[-e^{-a(-1/\alpha) \ln(\lambda/(\alpha(1+r)^N) \cdot (d\tilde{\mathbb{P}}/d\mathbb{P}^v))}\right] \\ &= e^{-v\alpha(1+r)^N - \sum_{i=0}^N \binom{N}{i} \tilde{p}^{N-i} \tilde{q}^i \ln\left(\binom{N}{i} \cdot \tilde{p}^{N-i} \tilde{q}^i / v_i\right)}. \end{aligned}$$

## Appendix

The following is with respect to the general discrete case in a complete market. As in Section 3, we denote by  $\Psi^v$  the set of self-financing portfolios given initial wealth  $v$ .

**Theorem A.1.** *The discounted wealth process is a martingale under the martingale measure  $\mathbb{Q}$ .*

*Proof.* See [Runggaldier 2005]. □

**Theorem A.2.** *Maximizing  $\mathbb{E}[U(V_N)]$  over the set of self-financing portfolios  $\Psi^v$  is equivalent to maximizing  $\mathbb{E}[U(V_N)]$  subject to  $\tilde{\mathbb{E}}[U(V_N)] = v$ , with  $\tilde{\mathbb{P}}$  being the unique equivalent martingale measure.*

*Proof.* See [Rásonyi and Stettner 2005, Lemma 4.9]. □

**Theorem A.3.**

$$\widehat{V}_N = I\left(\frac{\lambda}{(1+r)^N} \frac{d\tilde{\mathbb{P}}}{d\mathbb{Q}}\right).$$

*More specifically, optimal terminal wealth  $\widehat{V}_N$  is attained when  $\lambda$  satisfies*

$$v = \tilde{\mathbb{E}}\left[\frac{1}{(1+r)^N} I\left(\frac{\lambda}{(1+r)^N} \frac{d\tilde{\mathbb{P}}}{d\mathbb{Q}}\right)\right].$$

*Proof.* See [Runggaldier 2005, p. 16]. □

### Acknowledgments

This research was funded by the NSF grant DMS 1659643. The authors would like to thank Oleksii Mostovyi for several instructive discussions and comments on this work.

### References

- [Baudoin 2003] F. Baudoin, “Modeling anticipations on financial markets”, pp. 43–94 in *Paris-Princeton Lectures on Mathematical Finance, 2002*, edited by R. A. Carmona et al., Lecture Notes in Math. **1814**, Springer, 2003. MR Zbl
- [Baudoin and Nguyen-Ngoc 2004] F. Baudoin and L. Nguyen-Ngoc, “The financial value of a weak information on a financial market”, *Finance Stoch.* **8**:3 (2004), 415–435. MR Zbl
- [Björk 2009] T. Björk, *Arbitrage theory in continuous time*, 3rd ed., Oxford University Press, 2009.
- [Gilboa and Schmeidler 1989] I. Gilboa and D. Schmeidler, “Maxmin expected utility with nonunique prior”, *J. Math. Econom.* **18**:2 (1989), 141–153. MR Zbl
- [Karatzas et al. 1991] I. Karatzas, J. P. Lehoczky, S. E. Shreve, and G.-L. Xu, “Martingale and duality methods for utility maximization in an incomplete market”, *SIAM J. Control Optim.* **29**:3 (1991), 702–730. MR Zbl
- [Meyer and Meyer 2005] D. J. Meyer and J. Meyer, “Relative risk aversion: what do we know?”, *J. Risk Uncertainty* **31**:3 (2005), 243–262. Zbl
- [Rásonyi and Stettner 2005] M. Rásonyi and L. Stettner, “On utility maximization in discrete-time financial market models”, *Ann. Appl. Probab.* **15**:2 (2005), 1367–1395. MR Zbl
- [Runggaldier 2005] W. Runggaldier, “Portfolio optimization in discrete time”, preprint, 2005, available at [https://www.math.unipd.it/runggaldier/MPS\\_ru.pdf](https://www.math.unipd.it/runggaldier/MPS_ru.pdf).
- [Shreve 2004] S. E. Shreve, *Stochastic calculus for finance, I: The binomial asset pricing model*, Springer, 2004. MR Zbl

Received: 2018-11-08 Accepted: 2019-01-26

aamiran@umass.edu	<i>University of Massachusetts, Amherst, MA, United States</i>
fabrice.baudoin@uconn.edu	<i>University of Connecticut, Storrs, CT, United States</i>
snbrock@mavs.coloradomesa.edu	<i>Colorado Mesa University, Grand Junction, CO, United States</i>
berend.coster@uconn.edu	<i>University of Connecticut, Storrs, CT, United States</i>
rcraver@terpmail.umd.edu	<i>University of Maryland, College Park, MD, United States</i>
uezeaka@umass.edu	<i>University of Massachusetts, Amherst, MA, United States</i>
phanu9000@sbcglobal.net	<i>Purdue University, West Lafayette, IN, United States</i>
wishartmar@my.easternct.edu	<i>Eastern Connecticut State University, Willimantic, CT, United States</i>



## Guidelines for Authors

Submissions in all mathematical areas are encouraged. All manuscripts accepted for publication in *Involve* are considered publishable in quality journals in their respective fields, and include a minimum of one-third student authorship. Submissions should include substantial faculty input; faculty co-authorship is strongly encouraged. Authors may submit manuscripts in PDF format on-line at the Submission page at the Involve website.

**Originality.** Submission of a manuscript acknowledges that the manuscript is original and is not, in whole or in part, published or under consideration for publication elsewhere. It is understood also that the manuscript will not be submitted elsewhere while under consideration for publication in this journal.

**Language.** Articles in *Involve* are usually in English, but articles written in other languages are welcome.

**Required items.** A brief abstract of about 150 words or less must be included. It should be self-contained and not make any reference to the bibliography. If the article is not in English, two versions of the abstract must be included, one in the language of the article and one in English. Also required are keywords and subject classifications for the article, and, for each author, postal address, affiliation (if appropriate), and email address.

**Format.** Authors are encouraged to use L<sup>A</sup>T<sub>E</sub>X but submissions in other varieties of T<sub>E</sub>X, and exceptionally in other formats, are acceptable. Initial uploads should be in PDF format; after the refereeing process we will ask you to submit all source material.

**References.** Bibliographical references should be complete, including article titles and page ranges. All references in the bibliography should be cited in the text. The use of BibT<sub>E</sub>X is preferred but not required. Tags will be converted to the house format, however, for submission you may use the format of your choice. Links will be provided to all literature with known web locations and authors are encouraged to provide their own links in addition to those supplied in the editorial process.

**Figures.** Figures must be of publication quality. After acceptance, you will need to submit the original source files in vector graphics format for all diagrams in your manuscript: vector EPS or vector PDF files are the most useful.

Most drawing and graphing packages (Mathematica, Adobe Illustrator, MATLAB, etc.) allow the user to save files in one of these formats. Make sure that what you are saving is vector graphics and not a bitmap. If you need help, please write to [graphics@msp.org](mailto:graphics@msp.org) with details about how your graphics were generated.

**White space.** Forced line breaks or page breaks should not be inserted in the document. There is no point in your trying to optimize line and page breaks in the original manuscript. The manuscript will be reformatted to use the journal's preferred fonts and layout.

**Proofs.** Page proofs will be made available to authors (or to the designated corresponding author) at a Web site in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

# involve

2019      vol. 12      no. 5

Orbigraphs: a graph-theoretic analog to Riemannian orbifolds	721
KATHLEEN DALY, COLIN GAVIN, GABRIEL MONTES DE OCA, DIANA OCHOA, ELIZABETH STANHOPE AND SAM STEWART	
Sparse neural codes and convexity	737
R. AMZI JEFFS, MOHAMED OMAR, NATCHANON SUAYSOM, ALEINA WACHTEL AND NORA YOUNGS	
The number of rational points of hyperelliptic curves over subsets of finite fields	755
KRISTINA NELSON, JÓZSEF SOLYMOSI, FOSTER TOM AND CHING WONG	
Space-efficient knot mosaics for prime knots with mosaic number 6	767
AARON HEAP AND DOUGLAS KNOWLES	
Shabat polynomials and monodromy groups of trees uniquely determined by ramification type	791
NAIOMI CAMERON, MARY KEMP, SUSAN MASLAK, GABRIELLE MELAMED, RICHARD A. MOY, JONATHAN PHAM AND AUSTIN WEI	
On some edge Folkman numbers, small and large	813
JENNY M. KAUFMANN, HENRY J. WICKUS AND STANISŁAW P. RADZISZOWSKI	
Weighted persistent homology	823
GREGORY BELL, AUSTIN LAWSON, JOSHUA MARTIN, JAMES RUDZINSKI AND CLIFFORD SMYTH	
Leibniz algebras with low-dimensional maximal Lie quotients	839
WILLIAM J. COOK, JOHN HALL, VICKY W. KLIMA AND CARTER MURRAY	
Spectra of Kohn Laplacians on spheres	855
JOHN AHN, MOHIT BANSIL, GARRETT BROWN, EMILEE CARDIN AND YUNUS E. ZEYTUNCU	
Pairwise compatibility graphs: complete characterization for wheels	871
MATTHEW BEAUDOUIN-LAFON, SERENA CHEN, NATHANIEL KARST, DENISE SAKAI TROXELL AND XUDONG ZHENG	
The financial value of knowing the distribution of stock prices in discrete market models	883
AYELET AMIRAN, FABRICE BAUDOIN, SKYLYN BROCK, BEREND COSTER, RYAN CRAVER, UGONNA EZEAKA, PHANUEL MARIANO AND MARY WISHART	



1944-4176(2019)12:5;1-0