involve

a journal of mathematics

msp

2016

vol. 9, no. 2

involve

msp.org/involve

MANAGING EDITOR

Kenneth S. Berenhaut, Wake Forest University, USA, berenhks@wfu.edu

PRODUCTION

Silvio Levy, Scientific Editor

Cover: Alex Scorpan

See inside back cover or msp.org/involve for submission instructions. The subscription price for 2016 is US $160/year for the electronic version, and $215/year (+$35, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues from the last three years and changes of subscribers address should be sent to MSP.

Involve (ISSN 1944-4184 electronic, 1944-4176 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/

INVOLVE 9:2 (2016)
dx.doi.org/10.2140/involve.2016.9.181

On the independence and domination numbers of replacement product graphs

Jay Cummings and Christine A. Kelley

(Communicated by Joseph A. Gallian)

This paper examines invariants of the replacement product of two graphs in terms of the properties of the component graphs. In particular, we present results on the independence number, the domination number, and the total domination number of these graphs. The replacement product is a noncommutative graph operation that has been widely applied in many areas. One of its advantages over other graph products is its ability to produce sparse graphs. The results in this paper give insight into how to construct large, sparse graphs with optimal independence or domination numbers.

1. Introduction

It is natural to construct graphs from smaller component graphs, and as such, products of graphs have long been studied for both their theoretical interest and practical applicability. Standard products include the cartesian product, direct product, and strong product [Imrich and Klavžar 2000; Hammack et al. 2011]. As many modern applications require sparse graphs, newer products have been introduced. In particular, the replacement product is a noncommutative graph product of two regular component graphs that produces a regular graph whose degree depends only on the degree of the second component graph. Thus, the replacement product can be easily used to generate large, sparse graphs. In addition, it was shown that the expansion of the replacement product graph inherits the expansion properties of both component graphs [Reingold et al. 2002; Hoory et al. 2006]. The replacement product has been widely used in many areas including group theory, expander graphs, and graph-based coding schemes [Reingold et al. 2002; Hoory et al. 2006; Gamburd and Pak 2006; Kelley et al. 2008].

Invariants of graphs, including the independence and domination numbers of a graph, have also been widely studied. Many applications in computer science and

MSC2010: 05C10.
Keywords: minimized domination number, total domination number, maximized independence number, replacement product of a graph.

engineering require graphs with large independence numbers or small domination numbers. For example, in [Shannon 1956] it is shown that the independence number characterizes the largest number of bits that can be communicated without error in a particular communication problem. Studying the invariants of product graphs based on the invariants of the component graphs is an interesting problem, and in fact has led to many long-standing open problems in graph theory. Notable examples include Vizing's conjecture on the domination number of cartesian product graphs and Hedetniemi's conjecture on the chromatic number of direct product graphs (see, e.g., [Brešar et al. 2012; Hammack et al. 2011]). In [Alon and Orlitsky 1995], the independence numbers of graphs constructed using the n-fold AND product and the n-fold OR product are determined with respect to communicating multiple bits per channel use in a repeated communication model, generalizing the result in [Shannon 1956]. Similar applications that studied large independence number and large chromatic number in graph products are given in [Alon and Lubetzky 2006; Witsenhausen 1976]. Domination numbers have also been heavily studied and generalized (see, e.g., [Haynes et al. 1998a; 1998b; Chelvam and Chellathurai 2011]). The importance of the independence and domination numbers in applications and the advantages of the replacement product provide the motivation to study these invariants in replacement product graphs.

In this paper, we investigate the independence number, the domination number, and the total domination number of replacement product graphs in terms of their component graphs. One of our main results, Theorem 3.4, expresses the independence number of the replacement product of G and H in terms of the independence number of the second component graph, H. We also derive lower and upper bounds on the domination and total domination numbers for replacement product graphs. Another main result, Theorem 4.14, gives an upper bound on the total domination number for the replacement product of G and H in terms of the number of edges in a certain spanning subgraph of G.

The paper is organized as follows. We introduce some preliminary definitions and notation in Section 2. In Section 3, we determine the independence number of replacement product graphs. In Section 4, we present lower and upper bounds on the domination number and the total domination number of replacement product graphs. In addition, we include examples of families of graphs that meet the bounds.

2. Preliminaries

This paper studies properties of the replacement product $G \circledR H$ of two graphs G and H. We will assume in this work that G and H are finite simple connected graphs. We first recall some basic terminology and notation that will be used in this paper. We will use $V(G)$ and $E(G)$ to denote the vertex set and edge set of a

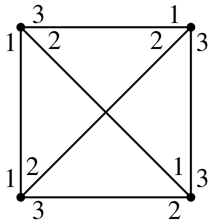


Figure 1. Rotation map.

graph G, respectively. Moreover, the minimum degree and the maximum degree of a vertex in G will be denoted by $\delta(G)$ and $\Delta(G)$, respectively. A *walk* is an alternating sequence of vertices and edges, beginning and ending with a vertex, where each vertex is incident to both the edge that precedes it and the edge that follows it in the sequence. The *length* of a walk is the number of edges in the walk. A *trail* is a walk where all edges are distinct. An *Eulerian trail* in G is a trail that contains each edge from G exactly once. A *closed Eulerian trail* is an Eulerian trail that begins and ends at the same vertex. A *path* is a walk where each vertex in the walk is distinct. A *Hamiltonian path* in G is a path that contains each vertex from G exactly once. The *distance* between vertices $u, v \in G$, denoted $d(u, v)$, is the length of the shortest path between vertices u and v. Finally, we will use $[n]$ to denote the set of integers $\{1, \dots, n\}$.

Definition 2.1. A *rotation map* on a graph G is a labeling of the edges of G where each edge gets two labels, one at each endpoint, and in addition, the edge labels around each vertex v in G are distinct and numbered using $1, 2, \dots, \deg(v)$.

For example, Figure 1 is an example of a rotation map on K_4.

We now introduce the replacement product of two graphs. This product is noncommutative and depends on the specific rotation map on the first component graph.

Definition 2.2. Let G be a b-regular graph with $|V(G)| = n$ and H be a k-regular graph with $|V(H)| = b$. Assign the vertices of G distinct labels in $[n]$ and assign the vertices of H distinct labels in $[b]$. Then given a rotation map on G, the replacement product $G \circledR H$ is a graph whose vertices are the ordered pairs (u, v) for $u \in [n]$ and $v \in [b]$. There is an edge between (u, v) and (w, l) in $G \circledR H$ if either (i) $u = w$ and there is an edge between vertex v and vertex l in H, or (ii) $u \neq w$ and there is an edge between u and w in G having label v at u and label l at w assigned by the rotation map on G.

The replacement product graph $G \circledR H$ as described in Definition 2.2 is a $(k+1)$-regular graph with nb vertices. Note that the degree of regularity of the product graph depends only on the degree of regularity of the second component graph H.

The graph $G \circledR H$ may be more easily seen in the following way. First, each vertex of G is replaced by a copy of the graph H; such a copy will be referred to as a *cloud*

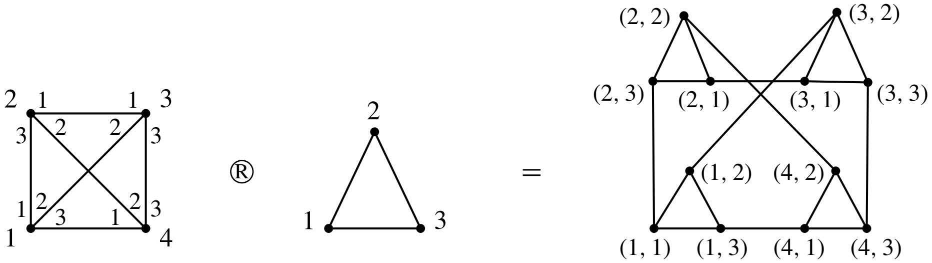


Figure 2. The replacement product $K_4 \circledR K_3$ with the specified rotation map on K_4.

and the cloud that replaces vertex i will be called the i-th cloud. Specifically, the i-th cloud is the subgraph induced by the set of vertices $\{(i, w) \in G \circledR H \mid w \in [b]\}$. Next, given any pair of distinct $i, j \in [n]$, there is exactly one edge between clouds i and j in $G \circledR H$ if and only if $(i, j) \in E(G)$. The vertices in the clouds that are connected by such an edge are determined by the rotation map on G. We will refer to edges that go between clouds as *intercloud edges*, and edges within clouds as *cloud edges*.

See Figure 2 for an example of the replacement graph product $K_4 \circledR K_3$.

3. Independence number of replacement product graphs

In this section we determine the independence number for replacement product graphs based on the independence number of the second component graph.

Definition 3.1. Let G be a graph. An *independent set* of G is a subset $S \subseteq V(G)$ such that no pair of distinct elements in S is adjacent. The *independence number* of G, denoted $\alpha(G)$, is the size of a largest independent set.

Due to the dependence of the replacement product on the rotation map, we introduce the following definition.

Definition 3.2. For graphs G and H, the *maximized independence number*, denoted $\hat{\alpha}(G \circledR H)$, is defined as the maximum possible independence number of $G \circledR H$ over all rotation maps m on G. That is,

$$\hat{\alpha}(G \circledR H) = \max_m\{\alpha(G^{(m)} \circledR H)\} = \max_m\{\max_S |S|\},$$

where $G^{(m)}$ is the graph G with rotation map m and S is an independent set of $G^{(m)} \circledR H$.

Note that in practice, one can often choose the rotation map for the replacement product graph. Indeed, this is typically done randomly. When the independence number of the graph is of interest, Definition 3.2 characterizes the best possible value one can obtain. The main result in this section shows explicitly how to design a rotation map that attains $\hat{\alpha}(G \circledR H)$. We first state the following known result, and include its proof for convenience.

Lemma 3.3 [Haynes et al. 1998b]. *For a k-regular graph G,*

$$\alpha(G) \leq \frac{|V(G)|}{2}.$$

Proof. Let S be an independent set of G with $|S| = m$. We now bound the total number of edges incident with S in G. Each of the $(|V(G)| - m)$ vertices in $V(G) - S$ may be adjacent to at most k members of S. So the total number of edges in G from these vertices to S is at most $(|V(G)| - m)k$. Each vertex in S has degree k, and by the independence of S, the m vertices in S are pairwise nonadjacent. Thus, the total number of edges incident with S is mk. Therefore,

$$mk \leq (|V(G)| - m)k,$$

from which we obtain

$$m \leq \frac{|V(G)|}{2}.$$

Since the statement holds for any independent set, it holds for a maximally sized independent set. □

Next we present the main result of this section, which determines the maximized independence number for a replacement product graph.

Theorem 3.4. *Let G be a b-regular graph with $|V(G)| = n$ and H a k-regular graph with $|V(H)| = b$. Then*

$$\hat{\alpha}(G \circledR H) = \alpha(H)|V(G)|.$$

Proof. First, it is easy to see that $\hat{\alpha}(G \circledR H) \leq \alpha(H)|V(G)|$, since otherwise, for some choice of a rotation map, there would exist a maximal independent set $\mathcal{S}$ of $G \circledR H$ such that some cloud contains more than $\alpha(H)$ members of $\mathcal{S}$. Since each cloud is an isomorphic copy of H, this gives a contradiction.

Now we show the reverse inequality by designing a specific rotation map that meets the bound. Label the vertices of G using $1, 2, \ldots, n$. Let I' be an independent set of H. By Lemma 3.3, $|I'| \leq b/2$. Label the vertices in each copy of H in $G \circledR H$ using the numbers $\{1, \ldots, b\}$ such that the vertices in I' receive the even numbers $2, 4, \ldots, 2|I'|$. Let (i, j) be the vertex in cloud i with label j.

We will show that there exists a rotation map on G with the property that for every vertex $(i, j) \in G \circledR H$, if j is even and (i, j) is adjacent to some (k, l) where $i \neq k$, then l must be odd. From this we will conclude that

$$I := \big\{(i, j) \in G \circledR H \mid j \in \{2, 4, \ldots, 2|I'|\}\big\}$$

is an independent set of size $\alpha(H)|V(G)|$.

We introduce the following algorithm which will be used to generate such a rotation map.

(1) Assign to each vertex $v \in V(G)$ a number T_v and a set S_v with initial values $T_v = 0$ and $S_v = [b]$. Set $\mathcal{V} := V(G)$.

(2) Choose a vertex $v \in \mathcal{V}$.

(3) Choose an unlabeled edge e incident with v. If there is an even number in S_v, then choose any such even number $a \in S_v$ and label the endpoint of e at v using a. Then set $T_v := T_v + 1$ and $S_v := S_v - a$. Otherwise label the endpoint of e at v using any odd number $a \in S_v$ and set $T_v := T_v - 1$ and $S_v := S_v - a$. Let u be the other vertex incident to e, and label the endpoint of e at u using any odd number $c \in S_u$. Then set $T_u = T_u - 1$ and $S_u := S_u - c$.

(4) If there is an unlabeled edge at u, set $v := u$ and go to Step 3.

(5) Let $\mathcal{U} = \{u \in V(G) \mid S_u = \varnothing\}$. Set $\mathcal{V} := \mathcal{V} - (\mathcal{U} \cap \mathcal{V})$. If $\mathcal{V} = \varnothing$, stop. Otherwise, go to Step 2.

Observe that T_v counts the number of even-labeled edges at v minus the number of odd-labeled edges at v. During the algorithm, T_v is never less than -1 because in Step 3, whenever a vertex receives an odd label at an edge, either another edge at that vertex receives an even label at the next step of the algorithm, or the vertex has all its edges labeled from 1 to b. Moreover, note that each edge receives its two endpoint labels consecutively. Thus, the vertex u in Step 3 always exists.

We now show that any rotation map generated by this algorithm satisfies the desired property by considering the parity of b, the regularity of G.

Case 1: Suppose b is even. Then there exists a closed Eulerian trail T in G. Then in Step 3 of the algorithm, instead of arbitrarily choosing the next edge to take after reaching a vertex, we choose an edge in order according to T. When the algorithm stops, $T_v = 0$ for all vertices $v \in V(G)$. Thus, every edge has one even label and one odd label at its endpoints, ensuring that the resulting rotation map on G has the asserted property.

Case 2: Suppose b is odd. For any vertex v at any stage of the algorithm, $T_v = -1$ if $S_v = \varnothing$, and when $S_v \neq \varnothing$, we have $T_v = -1$, 0 or 1. Therefore, since $[b]$ contains one more odd number than even number, when $S_u \neq \varnothing$, there is always an odd number to select for the edge in Step 3, and as a result, no edge will receive two even labels during the algorithm. Thus, the resulting rotation map on G has the asserted property.

Finally, let $G \circledR H$ be the replacement product graph in which the rotation map on G was obtained as described above. Then by construction, an edge from (i, j) to (k, ℓ) in $G \circledR H$ for j even and $i \neq k$ must have ℓ odd. Therefore

$$I := \{(i, j) \in G \circledR H \mid j \in \{2, 4, \ldots, 2|I'|\}\}$$

is an independent set and has size

$$|I| = |I'||V(G)| = \alpha(H)|V(G)|.$$

Thus, $\hat{\alpha}(G \circledR H) \geq \alpha(H)|V(G)|$, proving the equality. □

4. Domination numbers of replacement product graphs

In this section we present lower and upper bounds on two main types of domination numbers: the domination number and the total domination number. For more background on these parameters, see [Haynes et al. 1998a; 1998a].

Domination number.

Definition 4.1. A *dominating set* of a graph G is a subset $D \subseteq V(G)$ such that for every $v \in G \setminus D$, v is adjacent to some $v' \in D$. The *domination number* of G, denoted $\gamma(G)$, is the size of a smallest dominating set.

The domination number of a graph has been a parameter of great interest in applications such as communication and transportation networks. Again, due to the dependence of the replacement product on the rotation map, we introduce the following definition.

Definition 4.2. For graphs G and H, the *minimized domination number*, denoted $\hat{\gamma}(G \circledR H)$, is defined as the minimum possible domination number of $G \circledR H$ over all rotation maps m on G. That is,

$$\hat{\gamma}(G \circledR H) = \min_{m}\{\gamma(G^{(m)} \circledR H)\} = \min_{m}\{\min_{S} |S|\},$$

where $G^{(m)}$ is the graph G with rotation map m and S is a dominating set of $G^{(m)} \circledR H$.

We now give a lower bound on the domination number of a replacement product graph $G \circledR H$ in terms of the domination number of the second component graph, H.

Proposition 4.3. *Let G be a b-regular graph with $|V(G)| = n$ and H a k-regular graph with $|V(H)| = b$. Then*

$$\frac{n\gamma(H)}{2} \leq \hat{\gamma}(G \circledR H).$$

Moreover, if (i) $k = b - 2$, (ii) *n is even, and* (iii) *G contains a Hamiltonian cycle, then the bound is tight.*

Proof. Let G have any rotation map and let D be a dominating set of $G \circledR H$. Every vertex (i, j) in $G \circledR H$ is in one cloud (namely, the i-th cloud) and is adjacent to exactly one vertex in a different cloud. Note that there are at least $\gamma(H)$ elements of D in the vertex set of each cloud and its neighborhood, since otherwise there is a copy of H dominated by a vertex set of size strictly smaller than $\gamma(H)$. Thus, there are at least $\gamma(H)$ vertices in D dominating each cloud. Since there are n clouds and each vertex dominates vertices in two clouds, there must be at least $n\gamma(H)/2$ vertices in D. Thus, for any rotation map on G, we have $\gamma(G \circledR H) \geq n\gamma(H)/2$.

Now assume that the three additional properties (i)–(iii) hold as well. We will design a specific rotation map on G so that the lower bound is met. First, label the vertices of G in order from 1 to n according to a chosen Hamiltonian cycle $\mathcal{C}$. Then choose two nonadjacent vertices in H and label them 1 and 2, and label the rest of $V(H)$ using $3, \ldots, b$. Since $k = b-2$, each vertex is nonadjacent to exactly one other vertex, and the pair form a smallest dominating set in H. In particular, $\gamma(H) = 2$.

Now we construct our rotation map on G. For each $i \in [n-1]$, label the edge $(i, i+1) \in E(G)$ with a 1 at vertex i and a 2 at vertex $i+1$, and label edge $(n, 1)$ with a 1 at vertex n and a 2 at vertex 1. Then complete the rotation map in any way. Note that in $\mathcal{C}$, every edge is labeled by a 1 at one endpoint and a 2 at the other endpoint.

Now consider the product $G \circledR H$. We claim the set

$$D = \{(i, 1) \in V(G \circledR H) \mid i \in [n]\}$$

forms a dominating set of $G \circledR H$. For each i, every vertex in cloud i except vertex $(i, 2)$ is dominated by vertex $(i, 1)$, and vertex $(i, 2)$ is dominated by vertex $(i+1, 1)$. So D is a dominating set with size $|D| = n$. Therefore,

$$\hat{\gamma}(G \circledR H) \geq n = \frac{n\gamma(H)}{2}. \qquad \square$$

The next example gives a sequence of replacement product graphs that meet the bound in Proposition 4.3.

Example 4.4. Let m be any even integer and let K_n denote the complete graph on n vertices. Define $H_m := K_m - \mathcal{M}_1$ and $G_m := K_{m+2} - \mathcal{M}_2$, where $\mathcal{M}_1$ and $\mathcal{M}_2$ are perfect matchings of K_m and K_{m+2}, respectively. Then $k = m-2 = b-2$, $n = m+2$ is even, and G contains a Hamiltonian cycle. So

$$\{G_m \circledR H_m\}_{m \in 2\mathbb{Z}}$$

is a sequence of replacement product graphs meeting the bound in Proposition 4.3. More generally, in order to have a pair of graphs G, H that satisfy the three conditions, H must be isomorphic to H_m for some m, since no other regular graph has the property that $k = b-2$. However, G can be any graph of the form $C_{m+2k} \cup F$, where $k \in \mathbb{N}$, C_n denotes a cycle on n vertices, and F is an $(m-2)$-regular graph with $V(F) = V(C_{m+2k})$ and $E(F) \cap E(C_{m+2k}) = \varnothing$. $\square$

We have shown a lower bound on the minimized domination number of replacement product graphs, and a sequence of graphs that meet that bound. We now focus on deriving an upper bound on this parameter. For this, we will use the notion of the k-independence number of a graph, defined next.

Definition 4.5. For $k \in [|V(G)| - 1]$, a *k-independent set* of a graph G is a subset $S \subseteq V(G)$ such that S is an independent set and for every $v \in V(G) - S$, we have v is adjacent to at most k members from S. The largest cardinality of a *k-independent set* will be called the *k-independence number* and will be denoted $\alpha_k(G)$.

This parameter is related to the more familiar 2-packing number of a graph as defined below.

Definition 4.6. A *2-packing set* of a graph G is a subset $S \subseteq V(G)$ such that S is an independent set and for any pair of distinct $u, v \in S$, we have $d(u, v) \geq 3$, i.e., u and v have disjoint neighborhoods. Define the *2-packing number* of G, denoted $P_2(G)$, to be the largest cardinality of a 2-packing set of G.

The 2-packing number of G was introduced in [Meir and Moon 1975] and is a generalization of the independence number of G. Note that from the above definitions, the $(|V(G)|-1)$-independence number of a graph G is simply the independence number of G, and the 1-independence number of G is the 2-packing number of G.

We are now ready to present the upper bound on the minimized domination number.

Proposition 4.7. *Let G be a b-regular graph with $|V(G)| = n$ and H a k-regular graph with $|V(H)| = b$. Then*

$$\hat{\gamma}(G \circledR H) \leq (n - \alpha_{\gamma(H)}(G))\,\gamma(H).$$

Proof. There exist $\alpha_{\gamma(H)}(G)$ vertices in G that form a $\gamma(H)$-independent set S. Choose such a set and label these vertices $1, 2, \ldots, \alpha_{\gamma(H)}(G)$. Label the rest of the vertices of G using $\alpha_{\gamma(H)}(G) + 1, \ldots, n$. Choose a dominating set D' of H of size $\gamma(H)$ and label these vertices $1, 2, \ldots, \gamma(H)$. Label the rest of the vertices in H using $\gamma(H) + 1, \ldots, b$.

We now create a rotation map on G. Pick any $i \in [n] - [\alpha_{\gamma(H)}(G)]$ and let v_i be the vertex in G with label i. Label the edges at v_i by starting with the edges adjacent to members of S. Since S is a $\gamma(H)$-independent set, we can ensure that these edges get labels from the set $[\gamma(H)]$. Once this labeling has been done for every vertex from $V(G) - S$, complete the rotation map on G in any way.

We will show that the set

$$D := \big\{(i, j) \in V(G \circledR H) \,\big|\, i \in \{\alpha_{\gamma(H)}(G) + 1, \alpha_{\gamma(H)}(G) + 2, \ldots, n\}, \, j \in [\gamma(H)]\big\}.$$

is a dominating set in $G \circledR H$ with this rotation map.

Note that for each $i \in [n] - [\alpha_{\gamma(H)}(G)]$, cloud i is dominated by D, since every such cloud contains a copy of the dominating set D'. Furthermore, for each $i \in [\alpha_{\gamma(H)}(G)]$, every vertex in cloud i is adjacent via its intercloud edges to some cloud with label $j \in [n] - [\alpha_{\gamma(H)}(G)]$.

Moreover, by construction of the rotation map on G, we see that each vertex in cloud j, for $j \in [n] - [\alpha_{\gamma(H)}(G)]$, is adjacent to a vertex (a, b), for some $a \in [\alpha_{\gamma(H)}(G)]$ and $b \in [\gamma(H)]$, and hence an element of D. Thus, D is a dominating set and has size

$$|D| = \big(|V(G)| - \alpha_{\gamma(H)}(G)\big)\,\gamma(H),$$

giving the desired bound. □

The next example gives a sequence of graphs meeting the upper bound in Proposition 4.7.

Example 4.8. Let $G = K_{n+1}$ and $H = K_n$ for any $n \in \mathbb{N}$. Then $\gamma(H) = 1$ and $\alpha_1(G) = 1$. Then, given any rotation map on G, let D be the set of n vertices adjacent to a vertex in cloud 1 via an intercloud edge. Then D is a dominating set with size

$$|D| = n = ((n+1)-1)\cdot 1 = \big(|V(G)| - \alpha_{\gamma(H)}(G)\big)\,\gamma(H). \qquad \square$$

Total domination number. In this subsection we consider a related parameter, the total domination number of a graph, that has also been heavily studied in similar applications.

Definition 4.9. A *total dominating set* of a graph G is a subset $D \subseteq V(G)$ such that for every $v \in G$, v is adjacent to some $v' \in D$. The *total domination number* of G, denoted $\gamma_t(G)$, is the size of a smallest total dominating set.

Note that unlike in a dominating set of G, in a total dominating set of G a vertex does not dominate itself. Again, due to the dependence of the replacement product on the rotation map, we introduce the following definition.

Definition 4.10. For graphs G and H, the *minimized total domination number*, denoted $\hat{\gamma}_t(G \circledR H)$, is defined as the minimum possible total domination number of $G \circledR H$ over all rotation maps m on G. That is,

$$\hat{\gamma}_t(G \circledR H) = \min_m\{\gamma_t(G^{(m)} \circledR H)\} = \min_m\{\min_S |S|\},$$

where $G^{(m)}$ is the graph G with rotation map m and S is a total dominating set of $G^{(m)} \circledR H$.

In the rest of this section, we obtain lower and upper bounds on the total domination number of replacement product graphs. First, we state the following known result whose proof is straightforward.

Lemma 4.11 [Haynes et al. 1998b]. *If G is a k-regular graph then*

$$\gamma_t(G) \geq \frac{|V(G)|}{k}.$$

We next present a lower bound on the minimized total domination number of a replacement product graph, and the proof uses the notion of a k-factor. Recall that a *k-factor* of a graph G is a k-regular spanning subgraph of G.

Proposition 4.12. *Let G be a b-regular graph with $|V(G)| = n$ and H a k-regular graph with $|V(H)| = b$. Then,*

$$\hat{\gamma}_t(G \circledR H) \geq \frac{|V(G \circledR H)|}{k+1} = \frac{|V(G)||V(H)|}{k+1}.$$

Moreover, when G and H have the additional properties that (i) $b = \gamma(H)(k+1)$ *and* (ii) *G contains a $\gamma(H)$-factor, equality holds.*

Proof. The first statement follows from the $(k+1)$-regularity of $G \circledR H$ and Lemma 4.11. Assume that the additional properties (i) and (ii) hold for G and H. Let D' be a smallest dominating set in H, and let D be the set

$$D = \{(i, j) \in V(G \circledR H) \mid i \in [n], j \in D'\}.$$

Let G' be a $\gamma(H)$-factor of G. Design a rotation map on G by first labeling the edges of the subgraph G' at each vertex of G using the numbers $1, 2, \ldots, \gamma(H)$, and label the remaining edges at each vertex using $\gamma(H) + 1, \ldots, b$. Label the vertices in H by using the numbers $1, 2, \ldots, \gamma(H)$ for those in D' and the numbers $\gamma(H) + 1, \ldots, b$ for those not in D'.

Now consider the replacement product $G \circledR H$ with this rotation map, and as before let (i, j) denote the vertex in cloud i with label j. Consider an arbitrary intercloud edge, say from (i, j) to (m, l), where $i \neq m$. Then we see that, by construction, $j \in \{1, 2, \ldots, \gamma(H)\}$ if and only if $l \in \{1, 2, \ldots, \gamma(H)\}$. Moreover, since $(i, j) \in D$ if and only if $j \in \{1, 2, \ldots, \gamma(H)\}$, and every vertex is incident to exactly one intercloud edge, this also implies that every $v \in D$ is adjacent via an intercloud edge to some other $v' \in D$. This guarantees that D is not only a dominating set, but is in fact a total dominating set. Finally,

$$|D| = |V(G)|\gamma(H) = |V(G)|\frac{b}{k+1} = \frac{|V(G)||V(H)|}{k+1}. \qquad \square$$

In the next example, we construct a sequence of pairs G, H such that $G \circledR H$ meets the bound in Proposition 4.12 by showing that G and H satisfy the additional properties in the proposition.

Example 4.13. Let $G = K_{4m+1}$, the complete graph on $n = 4m + 1$ vertices. It is a known result from the theory of degree sequences that for any positive even integer m, the graph K_{4m+1} contains an m-factor [Chen 1988], and therefore G satisfies condition (ii). We now design H to be a 3-regular graph with $b = 4m$ vertices. Begin with m disjoint copies of $K_4 - e$, where e is any edge of K_4. Give each copy a distinct label from $[m]$. For each i, label the two vertices in the i-th copy that have degree two $(i, 1)$ and $(i, 2)$. Then connect vertices $(i, 1)$ to $(i+1, 2)$ for each $i \in [m-1]$ and connect $(m, 1)$ to $(1, 2)$. This yields the graph H. Since H is 3-regular,

$$\gamma(H) \geq \frac{b}{k+1} = \frac{4m}{4} = m.$$

Observe that each of the m copies of $K_4 - e$ can be dominated by exactly one vertex. Hence, $\gamma(H) = m$, satisfying the additional condition (i). Therefore, $\hat{\gamma}_t(G \circledR H)$ meets the bound. □

Our final result gives an upper bound on the minimized total domination number of replacement product graphs.

Theorem 4.14. *Let G be a b-regular graph with $|V(G)| = n$ and H a k-regular graph with $|V(H)| = b$. Let G' be a spanning subgraph of G for which $|E(G')|$ is minimal given that $\delta(G') \geq \gamma(H)$. Then*

$$\hat{\gamma}_t(G \circledR H) \leq 2|E(G')|.$$

Proof. First note that G' always exists since

$$\delta(G) = |V(H)| \geq \gamma(H).$$

This also shows that $|E(G)|$ is an upper bound for $|E(G')|$. Now let D' be a smallest dominating set in H and let D be the set

$$D = \{(i, j) \in V(G \circledR H) \mid i \in [n], j \in D'\}$$

in $G \circledR H$. Design a rotation map on G by first labeling the edges of the subgraph G' at each vertex v of G using the numbers $1, 2, \ldots, \deg_{G'}(v)$, and label the remaining edges at each vertex v using $\deg_{G'}(v) + 1, \ldots, b$. Label the vertices in H by using the numbers $1, 2, \ldots, \gamma(H)$ for those in D' and the numbers $\gamma(H) + 1, \ldots, b$ for those not in D'. Last, for each $v \in V(G)$, if $\deg_{G'}(v) > \gamma(H)$, then add the vertices $(L_v, \gamma(H)+1), (L_v, \gamma(H)+2), \ldots, (L_v, \deg_{G'}(v))$ to D, where L_v denotes the vertex label of v in G.

Now consider the product $G \circledR H$ with this rotation map on G. By construction of the rotation map, every $v \in D$ is adjacent to a vertex $v' \in D$ via an intercloud edge. This shows that D is a total dominating set. Finally,

$$\begin{aligned}
|D| &= \gamma(H)|V(G)| + \sum_{v \in V(G')} (\deg_{G'}(v) - \gamma(H)) \\
&= \gamma(H)|V(G)| + 2|E(G')| - |V(G')|\gamma(H) \\
&= \gamma(H)\big(|V(G)| - |V(G')|\big) + 2|E(G')| \\
&= 2|E(G')|.
\end{aligned}$$

Therefore with the specified rotation map,

$$\gamma_t(G \circledR H) \leq 2|E(G')|,$$

which implies that

$$\hat{\gamma}_t(G \circledR H) \leq 2|E(G')|. \qquad \square$$

The following example illustrates that the bound in Theorem 4.14 is sharp.

Example 4.15. Let G be a b-regular graph on n vertices that contains a 1-factor, and let $H = K_b$, where $b \geq 2$. Let G' be a 1-factor of G. Note that $\delta(G') = 1 \geq 1 = \gamma(H)$. Since a 1-factor has the fewest number of edges of any spanning subgraph of G, the graph G' satisfies the condition in Theorem 4.14. Fix a rotation map on G and let S be the set of all vertices in $G \circledR H$ that are incident to the intercloud edges corresponding to $E(G')$. Then each $v \in S$ is adjacent to some other vertex $v' \in S$, where $v' \neq v$. Moreover, since $E(G')$ is a 1-factor of G, there exists exactly one vertex from S in each cloud of $G \circledR H$. Since each cloud is a complete graph, every vertex in $G \circledR H$ is dominated by S. Therefore, S is a total dominating set and has size $|S| = 2|E(G')|$.

We now show that this is the smallest such total dominating set of $G \circledR H$. Assume that D is a smallest total dominating set of $G \circledR H$ for some arbitrary rotation map on G. Assume further that there exists a cloud $\mathcal{C}$ that does not contain a member of D. Then $\mathcal{C}$ must be dominated by b vertices in D, say $s_1, \ldots, s_b$, via intercloud edges. Moreover, each s_i is contained in a different cloud, say s_i is from cloud $\mathcal{C}_i$. Since each s_i must be adjacent to some other vertex in D using a cloud edge, there must be other members of D, say $t_1, \ldots, t_b$ such that $t_i \in \mathcal{C}_i$ for each i.

Thus, with the assumption that there exists such a cloud $\mathcal{C}$, we can deduce that at least b clouds must contain at least two vertices in D. Moreover, the vertices $t_1, \ldots, t_b$ from D collectively only dominate b additional vertices from $G \circledR H$ which were not already dominated by $s_1, \ldots, s_b$. So for each cloud that does not contain a member of D, b additional vertices are needed and at most one additional cloud can be dominated. Further note that if a cloud is not completely dominated via intercloud edges then there must exist a member of D within that cloud. Thus, since $b \geq 2$, we see that for D to be minimally sized, there cannot be a cloud that does not have a member of D contained within it. Therefore $|D| \geq n = 2|E(G')|$, which implies our conclusion. $\square$

Acknowledgements

The results in this paper were part of Cummings' undergraduate honors thesis, "On invariants of replacement product graphs", that was conducted under the supervision of Kelley. This work was supported in part by the Undergraduate Creative Activities and Research Experience (UCARE) program at UNL. We thank the anonymous reviewers for their helpful suggestions that improved the quality of this paper.

References

[Alon and Lubetzky 2006] N. Alon and E. Lubetzky, "The Shannon capacity of a graph and the independence numbers of its powers", *IEEE Trans. Inform. Theory* **52**:5 (2006), 2172–2176. MR 2007b:05149 Zbl 1247.05167

[Alon and Orlitsky 1995] N. Alon and A. Orlitsky, "Repeated communication and Ramsey graphs", *IEEE Trans. Inform. Theory* **41**:5 (1995), 1276–1289. MR 1366324 Zbl 0831.94003

[Brešar et al. 2012] B. Brešar, P. Dorbec, W. Goddard, B. L. Hartnell, M. A. Henning, S. Klavžar, and D. F. Rall, "Vizing's conjecture: a survey and recent results", *J. Graph Theory* **69**:1 (2012), 46–76. MR 2012k:05003 Zbl 1234.05173

[Chelvam and Chellathurai 2011] T. T. Chelvam and S. R. Chellathurai, *Recent trends in domination in graph theory: new domination parameters, bounds and links with other parameters*, Lambert Academic Pubishing, 2011.

[Chen 1988] Y. C. Chen, "A short proof of Kundu's k-factor theorem", *Discrete Math.* **71**:2 (1988), 177–179. MR 89i:05209 Zbl 0651.05054

[Gamburd and Pak 2006] A. Gamburd and I. Pak, "Expansion of product replacement graphs", *Combinatorica* **26**:4 (2006), 411–429. MR 2007f:05083 Zbl 1121.05114

[Hammack et al. 2011] R. Hammack, W. Imrich, and S. Klavžar, *Handbook of product graphs*, 2nd ed., CRC Press, Boca Raton, FL, 2011. MR 2012i:05001 Zbl 1283.05001

[Haynes et al. 1998a] T. W. Haynes, S. T. Hedetniemi, and P. J. Slater (editors), *Domination in graphs: advanced topics*, Monographs and Textbooks in Pure and Applied Mathematics **209**, Marcel Dekker, New York, 1998. MR 2000j:05091 Zbl 0883.00011

[Haynes et al. 1998b] T. W. Haynes, S. T. Hedetniemi, and P. J. Slater, *Fundamentals of domination in graphs*, Monographs and Textbooks in Pure and Applied Mathematics **208**, Marcel Dekker, New York, 1998. MR 2001a:05112 Zbl 0890.05002

[Hoory et al. 2006] S. Hoory, N. Linial, and A. Wigderson, "Expander graphs and their applications", *Bull. Amer. Math. Soc.* (*N.S.*) **43**:4 (2006), 439–561. MR 2007h:68055 Zbl 1147.68608

[Imrich and Klavžar 2000] W. Imrich and S. Klavžar, *Product graphs*, Wiley-Interscience, New York, 2000. MR 2001k:05001 Zbl 0963.05002

[Kelley et al. 2008] C. A. Kelley, D. Sridhara, and J. Rosenthal, "Zig-zag and replacement product graphs and LDPC codes", *Adv. Math. Commun.* **2**:4 (2008), 347–372. MR 2010f:94381 Zbl 1231.94101

[Meir and Moon 1975] A. Meir and J. W. Moon, "Relations between packing and covering numbers of a tree", *Pacific J. Math.* **61**:1 (1975), 225–233. MR 53 #5346 Zbl 0315.05102

[Reingold et al. 2002] O. Reingold, S. Vadhan, and A. Wigderson, "Entropy waves, the zig-zag graph product, and new constant-degree expanders", *Ann. of Math.* (2) **155**:1 (2002), 157–187. MR 2003c:05145 Zbl 1008.05101

[Shannon 1956] C. E. Shannon, "The zero error capacity of a noisy channel", *Institute of Radio Engineers, Transactions on Information Theory,* **IT-2**:September (1956), 8–19. MR 19,623b

[Witsenhausen 1976] H. S. Witsenhausen, "The zero-error side information problem and chromatic numbers", *IEEE Trans. Information Theory* **IT-22**:5 (1976), 592–593. MR 56 #15164 Zbl 0336.94015

Received: 2011-10-22 Revised: 2015-02-25 Accepted: 2015-02-26

jjcummings@math.ucsd.edu *Department of Mathematics, University of California, San Diego, La Jolla, CA 92093, United States*

ckelley2@math.unl.edu *Department of Mathematics, University of Nebraska–Lincoln, Lincoln, NE 68588, United States*

INVOLVE 9:2 (2016)
dx.doi.org/10.2140/involve.2016.9.195

An optional unrelated question RRT model

Jeong S. Sihm, Anu Chhabra and Sat N. Gupta

(Communicated by Kenneth S. Berenhaut)

We propose a modified unrelated question randomized response technique (RRT) model which allows respondents the option of answering a sensitive question directly without using the randomization device if they find the question nonsensitive. This situation has been handled before by Gupta, Tuck, Spears Gill, and Crowe using the split sample approach. In this work we avoid the split sample approach, which requires larger total sample size. Instead, we estimate the prevalence of the sensitive characteristic by using an optional unrelated question RRT model, but the corresponding sensitivity level is estimated from the same sample by using the traditional binary unrelated question RRT model of Greenberg, Abul-Ela, Simmons, and Horvitz. We compare the simulation results of this new model with those of the split-sample based optional unrelated question RRT model of Gupta et al. and the simple unrelated question RRT model of Greenberg et al. Computer simulations show that the new binary response and quantitative response models have the smallest variance among the three models when they have the same sample size.

1. Introduction

Social Desirability Bias (SDB) is the tendency in survey respondents to reply to a sensitive question in a socially desirable manner as opposed to replying truthfully. In order to encourage truthful answers, several techniques have been developed for sensitive survey questions. These include the indirect response technique of [Warner 1965] and the unrelated question technique of [Greenberg et al. 1969], both belonging to the family of randomized response techniques (RRT) models.

Several variations of the original RRT models, both binary response and quantitative response models, have been discussed by researchers, including Mangat and Singh [1990], Gupta [2001], Gupta et al. [2002; 2010; 2013a; 2013b], Christofides [2003], Mehta et al. [2012], and Sihm et al. [2015]. Among the many RRT procedures, we will focus here on the binary response unrelated question RRT model

MSC2010: 62-02, 62-04, 62D05.
Keywords: optional randomized response models, unrelated questions randomized response models, parameter estimation, simulation study.

of [Greenberg et al. 1969], as well as the quantitative response unrelated question RRT model of [Greenberg et al. 1971].

Gupta et al. [2013b] demonstrated that an optional unrelated question RRT model can lead to improvement in estimating the prevalence of the sensitive characteristic. They used the split sample approach because they estimated both the prevalence of the sensitive characteristic and sensitivity level of the question from the same set of responses. We will introduce in this paper new binary and quantitative optional unrelated question RRT models without using the split sample approach and estimate the prevalence of the sensitive characteristic and the sensitivity level of the question from two different sets of responses from the same sample. We will demonstrate by an extensive simulation study that the proposed models work better than the optional unrelated question RRT models of [Gupta et al. 2013b] and the traditional unrelated question RRT models of [Greenberg et al. 1969; 1971] for a fixed sample size.

2. Unrelated question RRT models

2.1. *Unrelated question models.*

2.1.1. *Binary response model.* This model was first introduced in [Greenberg et al. 1969]. In this model, you use a randomization device to ask a respondent the sensitive binary question with preassigned probability p_a and an innocuous question (whose prevalence is already known) with probability of $1 - p_a$.

Let π_a be the known prevalence of an unrelated characteristic and π be the unknown prevalence of the sensitive characteristic. Let P_y be the probability of a "yes" response from a respondent. Then P_y can be expressed as

$$P_y = p_a \pi + (1 - p_a)\pi_a. \tag{1}$$

Solving for π, we have

$$\pi = \frac{P_y - (1 - p_a)\pi_a}{p_a}.$$

This leads to the estimator of [Greenberg et al. 1969],

$$\hat{\pi}_g = \frac{\widehat{P}_y - (1 - p_a)\pi_a}{p_a}, \tag{2}$$

where $\widehat{P}_y$ is the proportion of "yes" responses in the survey. It is known that $\hat{\pi}_g$ is an unbiased estimator with its variance given by

$$\operatorname{Var}(\hat{\pi}_g) = \frac{P_y(1 - P_y)}{np_a^2}. \tag{3}$$

2.1.2. *Quantitative response model.* Very much like the binary response model, in this model the researcher will also ask a sensitive question with preassigned probability p_a and an innocuous question with probability $1 - p_a$.

Let μ_y and σ_y^2 respectively be the known mean and variance of an unrelated question and μ_x and σ_x^2 respectively be the unknown mean and variance of the sensitive question in the population. Let Z be the reported response from a respondent. Then Z can be expressed as

$$Z = \begin{cases} X & \text{with probability } p_a \text{ (sensitive question)}, \\ Y & \text{with probability } 1 - p_a \text{ (nonsensitive question)}, \end{cases}$$

with

$$E(Z) = \mu_z = p_a \mu_x + (1 - p_a)\mu_y, \tag{4}$$

$$\begin{aligned} \operatorname{Var}(Z) &= p_a E(X^2) + (1 - p_a)E(Y^2) - \mu_z^2 \\ &= p_a(\sigma_x^2 + \mu_x^2) + (1 - p_a)(\sigma_y^2 + \mu_y^2) - \mu_z^2. \end{aligned} \tag{5}$$

Solving (4) for μ_x, we have

$$\mu_x = \frac{\mu_z - (1 - p_a)\mu_y}{p_a}.$$

This leads to the estimator of [Greenberg et al. 1971],

$$\hat{\mu}_g = \frac{\bar{Z} - (1 - p_a)\mu_y}{p_a}, \tag{6}$$

where $\bar{Z}$ is the sample mean of the quantitative responses in the survey.

It is known that $\hat{\mu}_g$ is an unbiased estimator with its variance given by

$$\operatorname{Var}(\hat{\mu}_g) = \frac{1}{np_a^2} \operatorname{Var}(Z) = \frac{1}{np_a^2}\big(\sigma_y^2 + p_a(\sigma_x^2 - \sigma_y^2) + p_a(1 - p_a)(\mu_x - \mu_y)^2\big). \tag{7}$$

2.2. ***Optional unrelated question models.*** These models were proposed in [Gupta et al. 2013b] as a generalization of the original unrelated question models of [Greenberg et al. 1969; 1971] by giving respondents the option of responding to the sensitive question directly if they consider the question nonsensitive, while they can still give a scrambled response by using the model of [Greenberg et al. 1969] for a binary response and by using the model of [Greenberg et al. 1971] for a quantitative response if they feel the question is sensitive.

2.2.1. *Binary response model.* Let π_a be the known prevalence of an unrelated characteristic, π be the unknown prevalence of the sensitive characteristic, p be the preassigned probability of the respondent selecting the sensitive question, and ω be the unknown sensitivity level of the survey question in the population. Sensitivity level means the proportion of respondents in the population who would consider the question sensitive and subsequently opt to use a randomization device.

The probability of a "yes" response (P_y) in this model can be expressed as

$$P_y = (1-\omega)\pi + \omega(\pi p + (1-p)\pi_a). \tag{8}$$

Using two independent samples with sample sizes n_1 and n_2 respectively, and assuming that p_1 and p_2 are two different preassigned probabilities of the respondents selecting the sensitive question in the two samples, (8) can be written as

$$P_{y_1} - \pi = \omega(1-p_1)(\pi_a - \pi) \quad \text{and} \quad P_{y_2} - \pi = \omega(1-p_2)(\pi_a - \pi). \tag{9}$$

Solving for π, we have

$$\pi = \frac{\lambda P_{y_2} - P_{y_1}}{\lambda - 1} \quad (p_1, p_2 \neq 1,\ \ p_1 \neq p_2,\ \ \pi_a \neq \pi), \quad \text{where } \lambda = \frac{p_1 - 1}{p_2 - 1}. \tag{10}$$

Equation (10) leads to the unbiased estimator for π of [Gupta et al. 2013b],

$$\hat{\pi}_{gu} = \frac{\lambda \widehat{P}_{y_2} - \widehat{P}_{y_1}}{\lambda - 1}, \tag{11}$$

with its variance given by

$$\operatorname{Var}(\hat{\pi}_{gu}) = \frac{1}{(\lambda-1)^2}\left(\lambda^2 \frac{P_{y_2}(1-P_{y_2})}{n_2} + \frac{P_{y_1}(1-P_{y_1})}{n_1}\right). \tag{12}$$

Similarly from (9), we have

$$\omega = \frac{P_{y_1} - P_{y_2}}{(p_2 - p_1)\pi_a + (1-p_2)P_{y_1} - (1-p_1)P_{y_2}} \quad (p_1 \neq p_2,\ \pi_a \neq \pi), \tag{13}$$

which leads to an estimator of ω given by

$$\hat{\omega}_{gu^*} = \frac{\widehat{P}_{y_1} - \widehat{P}_{y_2}}{(p_2 - p_1)\pi_a + (1-p_2)\widehat{P}_{y_1} - (1-p_1)\widehat{P}_{y_2}}. \tag{14}$$

Gupta et al. [2013b] show that, up to first-order Taylor approximation, $\hat{\omega}_{gu^*}$ is an unbiased estimator for ω with its variance given by

$$\operatorname{Var}(\hat{\omega}_{gu^*}) = \frac{(p_2 - p_1)^2\left((\pi_a - P_{y_2})^2 \dfrac{P_{y_1}(1-P_{y_1})}{n_1} + (\pi_a - P_{y_1})^2 \dfrac{P_{y_2}(1-P_{y_2})}{n_2}\right)}{\left((p_2-p_1)\pi_a + (1-p_2)P_{y_1} - (1-p_1)P_{y_2}\right)^4}. \tag{15}$$

2.2.2. *Quantitative response model.* Let μ_y and σ_y^2 respectively be the known mean and variance of an innocuous question, μ_x and σ_x^2 respectively be the unknown mean and variance of the sensitive question in the population, p_1 and p_2 be the preassigned probabilities of answering the sensitive survey question in two subsamples, and ω be the unknown sensitivity level of the survey question in the population. Let Z_i be the reported response from a respondent in the i-th subsample $(i = 1, 2)$. Then Z_i can be expressed as

$$Z_i = \begin{cases} X_i & \text{with probability } (1-\omega)+\omega p_i \text{ (sensitive question)}, \\ Y_i & \text{with probability } \omega(1-p_i) \text{ (nonsensitive question)}, \end{cases}$$

with

$$E(Z_i) = \mu_{z_i} = ((1-\omega)+\omega p_i)\mu_x + \omega(1-p_i)\mu_y \quad (i = 1, 2), \tag{16}$$

$$\begin{aligned} \operatorname{Var}(Z_i) &= ((1-\omega)+\omega p_i)E(X_i^2) + \omega(1-p_i)E(Y_i^2) - \mu_{z_i}^2 \\ &= ((1-\omega)+\omega p_i)(\sigma_x^2+\mu_x^2) + \omega(1-p_i)(\sigma_y^2+\mu_y^2) - \mu_{z_i}^2. \end{aligned} \tag{17}$$

Equation (16) can be rearranged as

$$\mu_{z_i} - \mu_x = \omega(p_i - 1)(\mu_x - \mu_y) \quad (i = 1, 2). \tag{18}$$

Solving (18) for μ_x, we have

$$\mu_x = \frac{\mu_{z_1} - \lambda\mu_{z_2}}{1-\lambda} \quad (p_1, p_2 \neq 1,\ \ p_1 \neq p_2,\ \ \mu_x \neq \mu_y), \quad \text{where } \lambda = \frac{p_1 - 1}{p_2 - 1}. \tag{19}$$

This leads to the unbiased estimator for μ_x of [Gupta et al. 2013b],

$$\hat{\mu}_{gu} = \frac{\hat{\mu}_{z_1} - \lambda\hat{\mu}_{z_2}}{1-\lambda} = \frac{\bar{Z}_1 - \lambda\bar{Z}_2}{1-\lambda}, \tag{20}$$

with its variance given by

$$\operatorname{Var}(\hat{\mu}_{gu}) = \frac{1}{(1-\lambda)^2}\left(\frac{\operatorname{Var}(Z_1)}{n_1} + \lambda^2\frac{\operatorname{Var}(Z_2)}{n_2}\right). \tag{21}$$

Solving (18) for ω, we have

$$\omega = \frac{\mu_{z_1} - \mu_{z_2}}{(p_2 - p_1)\mu_y + (1-p_2)\mu_{z_1} - (1-p_1)\mu_{z_2}} \quad (p_1 \neq p_2,\ \mu_x \neq \mu_y), \tag{22}$$

which leads to

$$\hat{\omega}_{gu^{**}} = \frac{\bar{Z}_1 - \bar{Z}_2}{(p_2 - p_1)\pi_a + (1-p_2)\bar{Z}_1 - (1-p_1)\bar{Z}_2}. \tag{23}$$

Gupta et al. [2013b] show that, up to first-order Taylor approximation, $\hat{\omega}_{gu^{**}}$ is an unbiased estimator for ω with its variance given by

$$\operatorname{Var}(\hat{\omega}_{gu^{**}}) = \frac{(p_2-p_1)^2\left((\mu_y-\mu_{z_2})^2\frac{\operatorname{Var}(Z_1)}{n_1}+(\mu_{z_1}-\mu_y)^2\frac{\operatorname{Var}(Z_2)}{n_2}\right)}{\left((p_2-p_1)\pi_a+(1-p_2)\mu_{z_1}-(1-p_1)\mu_{z_2}\right)^4}. \quad (24)$$

3. The proposed model

The main motivation for this model is to avoid the split sample approach which requires unnecessarily larger total sample size. We do this by asking respondents two questions. Question 1 is the main research question, which the respondent answers using the optional model of [Gupta et al. 2013b]. Question 2 is the auxiliary question about whether or not the underlying research question is sensitive enough for the respondent to opt for a scrambled response. Respondents will answer Question 2 by using the original model of [Greenberg et al. 1969].

3.1. ***Binary response model.*** Let π_a be the known prevalence of an unrelated innocuous characteristic, π_b be the known prevalence of another unrelated innocuous characteristic, π be the unknown prevalence of the sensitive characteristic, p_a be the preassigned probability of the respondent selecting the sensitive question in answering Question 1, p_b be the preassigned probability of the respondent selecting the question about sensitivity in answering Question 2, and ω be the unknown sensitivity level of the survey question in the population.

Let P_{y_i} be the probability of a "yes" response from a respondent to Question i $(i=1,2)$. We have

$$P_{y_1} = (1-\omega)\pi+\omega(\pi p_a+(1-p_a)\pi_a), \quad (25)$$

$$P_{y_2} = p_b\omega+(1-p_b)\pi_b. \quad (26)$$

Solving (25) and (26) for π and ω respectively, we have

$$\pi = \frac{P_{y_1}-(1-p_a)\omega\pi_a}{1-(1-p_a)\omega} \quad \text{and} \quad \omega = \frac{P_{y_2}-(1-p_b)\pi_b}{p_b}, \quad (27)$$

which lead to the estimators

$$\hat{\pi}_p = \frac{\widehat{P}_{y_1}-(1-p_a)\hat{\omega}_{p^*}\pi_a}{1-(1-p_a)\hat{\omega}_{p^*}} \quad \text{and} \quad \hat{\omega}_{p^*} = \frac{\widehat{P}_{y_2}-(1-p_b)\pi_b}{p_b}, \quad (28)$$

where $\widehat{P}_{y_i}$ is the proportion of "yes" responses in the sample to Question i $(i=1,2)$.

Notice that $\hat{\omega}_{p^*}$ is an unbiased estimator with its variance given by

$$\operatorname{Var}(\hat{\omega}_{p^*}) = \frac{P_{y_2}(1-P_{y_2})}{np_b^2}. \quad (29)$$

After applying first-order Taylor expansion to $\hat{\pi}_p$, we have

$$\hat{\pi}_p \approx \hat{\pi}(P_{y_1}, \omega) + \frac{\partial \hat{\pi}(\widehat{P}_{y_1}, \hat{\omega}_{p^*})}{\partial \widehat{P}_{y_1}}\bigg|_{P_{y_1},\omega} (\widehat{P}_{y_1} - P_{y_1}) + \frac{\partial \hat{\pi}(\widehat{P}_{y_1}, \hat{\omega}_{p^*})}{\partial \hat{\omega}_{p^*}}\bigg|_{P_{y_1},\omega} (\hat{\omega}_{p^*} - \omega) \tag{30}$$

$$= \frac{P_{y_1} - \omega(1-p_a)\pi_a}{1-(1-p_a)\omega} + \frac{\widehat{P}_{y_1} - P_{y_1}}{1-(1-p_a)\omega} + \frac{(1-p_a)(P_{y_1}-\pi_a)(\hat{\omega}_{p^*}-\omega)}{(1-(1-p_a)\omega)^2}. \tag{31}$$

Up to first-order Taylor approximation, $\hat{\pi}_p$ is an unbiased estimator for π with its variance given by

$$\operatorname{Var}(\hat{\pi}_p) = \frac{1}{(1-(1-p_a)\omega)^2} \cdot \frac{P_{y_1}(1-P_{y_1})}{n} + \frac{(1-p_a)^2(P_{y_1}-\pi_a)^2}{(1-(1-p_a)\omega)^4} \cdot \frac{P_{y_2}(1-P_{y_2})}{np_b^2}. \tag{32}$$

3.2. ***Quantitative response model.*** Let μ_y and σ_y^2 respectively be the known mean and variance of an innocuous question, μ_x and σ_x^2 respectively be the unknown mean and variance of the sensitive question in the population, p_a be the preassigned probability of the respondent selecting the sensitive question in answering Question 1, and ω be the unknown sensitivity level of the survey question in the population. Let Z be the reported response to Question 1 from a respondent. Then Z can be expressed as

$$Z = \begin{cases} X & \text{with probability } (1-\omega) + \omega p_a \text{ (sensitive question)}, \\ Y & \text{with probability } \omega(1-p_a) \text{ (nonsensitive question)}, \end{cases}$$

with

$$E(Z) = \mu_z = ((1-\omega) + \omega p_a)\mu_x + \omega(1-p_a)\mu_y, \tag{33}$$

$$\begin{aligned} \operatorname{Var}(Z) &= ((1-\omega) + \omega p_a)E(X^2) + \omega(1-p_a)E(Y^2) - \mu_z^2 \\ &= ((1-\omega) + \omega p_a)(\sigma_x^2 + \mu_x^2) + \omega(1-p_a)(\sigma_y^2 + \mu_y^2) - \mu_z^2. \end{aligned} \tag{34}$$

Let π_b be the known prevalence of a binary innocuous characteristic for Question 2 and p_b be the preassigned probability of the respondent selecting the question about sensitivity in answering Question 2. We have the probability of a “yes” response to Question 2 given by

$$P_y = p_b\omega + (1-p_b)\pi_b. \tag{35}$$

Solving (33) and (35) for μ_x and ω respectively, we have

$$\mu_x = \frac{\mu_z - \mu_y(1-p_a)\omega}{1-(1-p_a)\omega} \quad \text{and} \quad \omega = \frac{P_y - (1-p_b)\pi_b}{p_b}, \tag{36}$$

which lead to the estimators

$$\hat{\mu}_p = \frac{\bar{Z} - \mu_y(1-p_a)\hat{\omega}_{p^{**}}}{1-(1-p_a)\hat{\omega}_{p^{**}}} \quad \text{and} \quad \hat{\omega}_{p^{**}} = \frac{\widehat{P}_y - (1-p_b)\pi_b}{p_b}, \tag{37}$$

where $\widehat{P}_y$ is the proportion of "yes" responses to Question 2 in the sample.

Notice that $\hat{\omega}_{p^{**}}$ is an unbiased estimator with the variance given by

$$\operatorname{Var}(\hat{\omega}_{p^{**}}) = \frac{P_y(1-P_y)}{np_b^2}. \tag{38}$$

After applying first-order Taylor expansion to $\hat{\mu}_p$, we have

$$\hat{\mu}_p \approx \hat{\mu}(\mu_z, \omega) + \frac{\partial \hat{\mu}(\hat{\mu}_z, \hat{\omega}_{p^{**}})}{\partial \hat{\mu}_z}\bigg|_{\mu_z,\omega} (\hat{\mu}_z - \mu_z) + \frac{\partial \hat{\mu}(\hat{\mu}_z, \hat{\omega}_{p^{**}})}{\partial \hat{\omega}_{p^{**}}}\bigg|_{\mu_z,\omega} (\hat{\omega}_{p^{**}} - \omega) \tag{39}$$

$$= \frac{\mu_z-\mu_y(1-p_a)\omega}{1-(1-p_a)\omega} + \frac{\hat{\mu}_z-\mu_z}{1-(1-p_a)\omega} + \frac{(1-p_a)(\mu_z-\mu_y)(\hat{\omega}_{p^{**}}-\omega)}{(1-(1-p_a)\omega)^2}$$

$$= \frac{\mu_z-\mu_y(1-p_a)\omega}{1-(1-p_a)\omega} + \frac{\bar{Z}-\mu_z}{1-(1-p_a)\omega} + \frac{(1-p_a)(\mu_z-\mu_y)(\hat{\omega}_{p^{**}}-\omega)}{(1-(1-p_a)\omega)^2}. \tag{40}$$

Up to first-order Taylor approximation, $\hat{\mu}_p$ is an unbiased estimator for μ_x with the variance given by

$$\operatorname{Var}(\hat{\mu}_p) = \frac{1}{(1-(1-p_a)\omega)^2} \cdot \frac{\operatorname{Var}(Z)}{n} + \frac{(1-p_a)^2(\mu_z-\mu_y)^2}{(1-(1-p_a)\omega)^4} \cdot \frac{P_y(1-P_y)}{np_b^2}. \tag{41}$$

4. Simulation results

In this section, simulation results are presented for our estimators, $\hat{\pi}_p$, $\hat{\omega}_{p^*}$, $\hat{\mu}_p$, and $\hat{\omega}_{p^{**}}$. We compare simulation results of the proposed models with the results from other models. Specifically, we compare $\hat{\pi}_p$ with $\hat{\pi}_g$ and $\hat{\pi}_{gu}$, and $\hat{\omega}_{p^*}$ with $\hat{\omega}_{gu^*}$ for binary response. Likewise, for the quantitative response models, we compared $\hat{\mu}_p$ with $\hat{\mu}_g$ and $\hat{\mu}_{gu}$, and $\hat{\omega}_{p^{**}}$ with $\hat{\omega}_{gu^{**}}$.

All the simulations were conducted by using the R programming language (http://www.R-project.org). For the binary response models, two parameters, π and ω, were allowed to vary while all the other variables were fixed. We used 10000 iterations with $p_a = 0.85$, $\pi_a = 0.7$, $p_b = 0.5$, $\pi_b = 0.1$, $p_1 = 0.85$, and $p_2 = 0.15$. For the quantitative response models, two parameters, μ_x and ω, were allowed to vary while all the other variables were fixed. Again we used 10000 iterations with $p_a = 0.85$, $\mu_y = 7.0$, $p_b = 0.6$, $\pi_b = 0.1$, $p_1 = 0.85$, and $p_2 = 0.15$. For the distributions of X and Y, we used the Poisson distributions with the parameters μ_x and μ_y respectively, as done in [Gupta et al. 2013b]. Also notice that all models have the same total sample size ($n = 1000$). For the optional unrelated question RRT models with split samples, the optimal sample ratios were used according to the formulas in [Gupta et al. 2013b].

4.1. ***Simulation of $\hat{\pi}$ and $\hat{\omega}$ for binary case.*** Table 1 shows that the theoretical $\operatorname{Var}(\hat{\pi}_p)$ values are always the smallest in comparison with the theoretical $\operatorname{Var}(\hat{\pi}_g)$

		$\pi = 1.0$	$\pi = 2.0$	$\pi = 3.0$	$\pi = 4.0$	$\pi = 5.0$
New model	EMean($\hat{\pi}_p$)	0.099923	0.199995	0.300015	0.400170	0.500034
	EVar($\hat{\pi}_p$)	0.000101	0.000169	0.000220	0.000243	0.000256
	TVar($\hat{\pi}_p$)	0.000103	0.000172	0.000220	0.000249	0.000258
	EMean($\hat{\omega}_{p*}$)	0.100110	0.099778	0.099747	0.100091	0.099960
	EVar($\hat{\omega}_{p*}$)	0.000360	0.000355	0.000366	0.000365	0.000355
	TVar($\hat{\omega}_{p*}$)	0.000360	0.000360	0.000360	0.000360	0.000360
Simple unrelated	EMean($\hat{\pi}_g$)	0.099875	0.200037	0.300155	0.400311	0.500007
	EVar($\hat{\pi}_g$)	0.000209	0.000273	0.000316	0.000336	0.000342
	TVar($\hat{\pi}_g$)	0.000213	0.000276	0.000319	0.000342	0.000345
Optional unrelated with optimal split	EMean($\hat{\pi}_{gu}$)	0.100290	0.199921	0.300257	0.399756	0.499761
	EVar($\hat{\pi}_{gu}$)	0.000203	0.000345	0.000445	0.000492	0.000495
	TVar($\hat{\pi}_{gu}$)	0.000207	0.000341	0.000436	0.000493	0.000510
	EMean($\hat{\omega}_{gu*}$)	0.097581	0.098096	0.093820	0.090546	0.073563
	EVar($\hat{\omega}_{gu*}$)	0.004716	0.010569	0.021178	0.041900	0.102535
	TVar($\hat{\omega}_{gu*}$)	0.004727	0.010600	0.020670	0.041002	0.094970
	Optimal n_1	831	843	847	849	850
	Optimal n_2	169	157	153	151	150
New model	EMean($\hat{\pi}_p$)	0.099817	0.200152	0.299968	0.399772	0.500247
	EVar($\hat{\pi}_p$)	0.000122	0.000188	0.000240	0.000264	0.000276
	TVar($\hat{\pi}_p$)	0.000127	0.000194	0.000240	0.000267	0.000275
	EMean($\hat{\omega}_{p*}$)	0.300150	0.299862	0.300151	0.300172	0.300387
	EVar($\hat{\omega}_{p*}$)	0.000659	0.000628	0.000623	0.000641	0.000641
	TVar($\hat{\omega}_{p*}$)	0.000640	0.000640	0.000640	0.000640	0.000640
Simple unrelated	EMean($\hat{\pi}_g$)	0.099657	0.200152	0.299961	0.399758	0.500268
	EVar($\hat{\pi}_g$)	0.000210	0.000274	0.000320	0.000338	0.000344
	TVar($\hat{\pi}_g$)	0.000213	0.000276	0.000319	0.000342	0.000345
Optional unrelated with optimal split	EMean($\hat{\pi}_{gu}$)	0.100291	0.199864	0.299693	0.400126	0.500136
	EVar($\hat{\pi}_{gu}$)	0.000247	0.000366	0.000437	0.000496	0.000508
	TVar($\hat{\pi}_{gu}$)	0.000247	0.000367	0.000450	0.000497	0.000509
	EMean($\hat{\omega}_{gu*}$)	0.298015	0.299102	0.295495	0.288993	0.278986
	EVar($\hat{\omega}_{gu*}$)	0.005779	0.010766	0.019549	0.037961	0.090777
	TVar($\hat{\omega}_{gu*}$)	0.005654	0.010818	0.019634	0.037539	0.085578
	Optimal n_1	813	834	843	848	851
	Optimal n_2	187	166	157	152	149

Table 1. Simulation results of binary models: trials = 10000, $p_a = 0.85$, $\pi_a = 0.7$, $p_b = 0.5$, $\pi_b = 0.1$, $p_1 = 0.85$, $p_2 = 0.15$, $n = 1000$. Continued on next two pages.

		$\pi = 1.0$	$\pi = 2.0$	$\pi = 3.0$	$\pi = 4.0$	$\pi = 5.0$
New model	EMean($\hat{\pi}_p$)	0.100065	0.199972	0.299964	0.400173	0.500032
	EVar($\hat{\pi}_p$)	0.000150	0.000219	0.000258	0.000289	0.000294
	TVar($\hat{\pi}_p$)	0.000153	0.000217	0.000262	0.000287	0.000293
	EMean($\hat{\omega}_{p^*}$)	0.500294	0.499501	0.500029	0.499672	0.499882
	EVar($\hat{\omega}_{p^*}$)	0.000846	0.000840	0.000819	0.000840	0.000824
	TVar($\hat{\omega}_{p^*}$)	0.000840	0.000840	0.000840	0.000840	0.000840
Simple unrelated	EMean($\hat{\pi}_g$)	0.100058	0.199923	0.299879	0.400215	0.499955
	EVar($\hat{\pi}_g$)	0.000214	0.000281	0.000316	0.000341	0.000348
	TVar($\hat{\pi}_g$)	0.000213	0.000276	0.000319	0.000342	0.000345
Optional unrelated with optimal split	EMean($\hat{\pi}_{gu}$)	0.099788	0.200111	0.300070	0.400095	0.500009
	EVar($\hat{\pi}_{gu}$)	0.000282	0.000377	0.000458	0.000506	0.000522
	TVar($\hat{\pi}_{gu}$)	0.000281	0.000387	0.000460	0.000500	0.000508
	EMean($\hat{\omega}_{gu^*}$)	0.499378	0.495986	0.496737	0.491159	0.483003
	EVar($\hat{\omega}_{gu^*}$)	0.006075	0.010543	0.018355	0.035501	0.081250
	TVar($\hat{\omega}_{gu^*}$)	0.006043	0.010546	0.018314	0.034180	0.076555
	Optimal n_1	807	830	842	849	852
	Optimal n_2	193	170	158	151	148
New model	EMean($\hat{\pi}_p$)	0.099826	0.200188	0.300180	0.399989	0.500289
	EVar($\hat{\pi}_p$)	0.000176	0.000238	0.000284	0.000307	0.000314
	TVar($\hat{\pi}_p$)	0.000180	0.000242	0.000285	0.000309	0.000313
	EMean($\hat{\omega}_{p^*}$)	0.700494	0.700062	0.699990	0.700182	0.700184
	EVar($\hat{\omega}_{p^*}$)	0.000951	0.000958	0.000996	0.000954	0.000958
	TVar($\hat{\omega}_{p^*}$)	0.000960	0.000960	0.000960	0.000960	0.000960
Simple unrelated	EMean($\hat{\pi}_g$)	0.099812	0.200224	0.300152	0.399974	0.500196
	EVar($\hat{\pi}_g$)	0.000216	0.000276	0.000318	0.000342	0.000348
	TVar($\hat{\pi}_g$)	0.000213	0.000276	0.000319	0.000342	0.000345
Optional unrelated with optimal split	EMean($\hat{\pi}_{gu}$)	0.099964	0.199881	0.299798	0.399785	0.499802
	EVar($\hat{\pi}_{gu}$)	0.000309	0.000399	0.000476	0.000501	0.000508
	TVar($\hat{\pi}_{gu}$)	0.000308	0.000403	0.000466	0.000500	0.000505
	EMean($\hat{\omega}_{gu^*}$)	0.698348	0.697556	0.698043	0.695540	0.686860
	EVar($\hat{\omega}_{gu^*}$)	0.006058	0.009892	0.017114	0.031608	0.072975
	TVar($\hat{\omega}_{gu^*}$)	0.006026	0.009977	0.016837	0.030600	0.068462
	Optimal n_1	808	831	844	850	854
	Optimal n_2	192	169	156	150	146

Table 1 (continued).

		$\pi = 1.0$	$\pi = 2.0$	$\pi = 3.0$	$\pi = 4.0$	$\pi = 5.0$
New model	EMean$(\hat{\pi}_p)$	0.099938	0.199859	0.300231	0.399827	0.499915
	EVar$(\hat{\pi}_p)$	0.000210	0.000268	0.000303	0.000332	0.000325
	TVar$(\hat{\pi}_p)$	0.000209	0.000269	0.000310	0.000332	0.000334
	EMean$(\hat{\omega}_{p^*})$	0.899699	0.900364	0.900596	0.900153	0.899261
	EVar$(\hat{\omega}_{p^*})$	0.000993	0.001002	0.000974	0.000995	0.000982
	TVar$(\hat{\omega}_{p^*})$	0.001000	0.001000	0.001000	0.001000	0.001000
Simple unrelated	EMean$(\hat{\pi}_g)$	0.099951	0.199827	0.300278	0.399858	0.499884
	EVar$(\hat{\pi}_g)$	0.000217	0.000278	0.000310	0.000343	0.000336
	TVar$(\hat{\pi}_g)$	0.000213	0.000276	0.000319	0.000342	0.000345
Optional unrelated with optimal split	EMean$(\hat{\pi}_{gu})$	0.099964	0.200086	0.299842	0.400039	0.499913
	EVar$(\hat{\pi}_{gu})$	0.000333	0.000414	0.000474	0.000506	0.000496
	TVar$(\hat{\pi}_{gu})$	0.000329	0.000414	0.000470	0.000499	0.000502
	EMean$(\hat{\omega}_{gu^*})$	0.898232	0.898642	0.897893	0.888967	0.886717
	EVar$(\hat{\omega}_{gu^*})$	0.005918	0.009252	0.015552	0.027088	0.065652
	TVar$(\hat{\omega}_{gu^*})$	0.005709	0.009163	0.015084	0.027281	0.060890
	Optimal n_1	815	836	847	853	856
	Optimal n_2	185	164	153	147	144

Table 1 (end).

values of the traditional unrelated question RRT model and the theoretical Var$(\hat{\pi}_{gu})$ values of the split sample optional unrelated question RRT model. Similarly, Var$(\hat{\omega}_{p^*})$ is always smaller than Var$(\hat{\omega}_{gu^*})$.

In Table 1, the variances of the proposed models consistently have the smallest value. For each model, the theoretical and empirical variances match very well. First-order Taylor approximation was used to calculate Var$(\hat{\pi}_p)$ and Var$(\hat{\omega}_{gu^*})$.

Notice that Var$(\hat{\pi}_{gu}) >$ Var$(\hat{\pi}_g)$, except for one case in Table 1. In this case, different models are used, the former a 2-parameter model and the latter a 1-parameter model.

4.2. ***Simulation of $\hat{\mu}_x$ and $\hat{\omega}$ for quantitative case.*** Table 2 shows that the Var$(\hat{\mu}_p)$ values are always the smallest in comparison with the Var$(\hat{\mu}_g)$ values of the traditional unrelated question RRT model and the Var$(\hat{\mu}_{gu})$ values of the split sample optional unrelated question RRT model. Similarly, Var$(\hat{\omega}_{p^{**}})$ is always smaller than Var$(\hat{\omega}_{gu^{**}})$.

In Table 2, the variances of the proposed models are consistently the smallest. For each model, the theoretical and empirical variances match very well. First-order Taylor approximation was used to calculate the theoretical values of Var$(\hat{\mu}_p)$ and Var$(\hat{\omega}_{gu^{**}})$.

		$\pi = 1.0$	$\pi = 2.0$	$\pi = 3.0$	$\pi = 4.0$	$\pi = 5.0$
New model	EMean($\hat{\mu}_p$)	1.000132	1.999822	2.999639	3.999132	4.999228
	EVar($\hat{\mu}_p$)	0.001693	0.002556	0.003328	0.004309	0.005253
	TVar($\hat{\mu}_p$)	0.001880	0.002664	0.003490	0.004358	0.005268
	EMean($\hat{\omega}_{p^{**}}$)	0.099985	0.100201	0.100164	0.099987	0.099953
	EVar($\hat{\omega}_{p^{**}}$)	0.000249	0.000245	0.000252	0.000252	0.000252
	TVar($\hat{\omega}_{p^{**}}$)	0.000250	0.000250	0.000250	0.000250	0.000250
Simple unrelated	EMean($\hat{\mu}_g$)	0.998261	2.000599	3.000724	3.999652	4.998115
	EVar($\hat{\mu}_g$)	0.009062	0.008097	0.007823	0.007660	0.008019
	TVar($\hat{\mu}_g$)	0.008983	0.008218	0.007806	0.007747	0.008042
Optional unrelated with optimal split	EMean($\hat{\mu}_{gu}$)	1.000980	2.001495	2.998708	3.999821	4.999394
	EVar($\hat{\mu}_{gu}$)	0.004008	0.005554	0.007014	0.008719	0.010369
	TVar($\hat{\mu}_{gu}$)	0.003965	0.005506	0.007094	0.008756	0.010504
	EMean($\hat{\omega}_{gu^{**}}$)	0.099302	0.098822	0.100371	0.098259	0.096624
	EVar($\hat{\omega}_{gu^{**}}$)	0.001197	0.002033	0.003726	0.007750	0.019916
	TVar($\hat{\omega}_{gu^{**}}$)	0.001162	0.002018	0.003724	0.007716	0.020077
	Optimal n_1	777	809	828	839	845
	Optimal n_2	223	191	172	161	155
New model	EMean($\hat{\mu}_p$)	0.999257	2.000800	2.999809	3.997820	5.000697
	EVar($\hat{\mu}_p$)	0.003161	0.003674	0.004222	0.005034	0.005707
	TVar($\hat{\mu}_p$)	0.003512	0.003912	0.004429	0.005064	0.005817
	EMean($\hat{\omega}_{p^{**}}$)	0.300102	0.300129	0.300195	0.300102	0.299864
	EVar($\hat{\omega}_{p^{**}}$)	0.000475	0.000476	0.000487	0.000491	0.000469
	TVar($\hat{\omega}_{p^{**}}$)	0.000477	0.000477	0.000477	0.000477	0.000477
Simple unrelated	EMean($\hat{\mu}_g$)	0.999620	2.001652	2.999560	3.998422	5.000657
	EVar($\hat{\mu}_g$)	0.008981	0.008030	0.007813	0.007787	0.007886
	TVar($\hat{\mu}_g$)	0.008983	0.008218	0.007806	0.007747	0.008042
Optional unrelated with optimal split	EMean($\hat{\mu}_{gu}$)	1.000295	2.000283	2.999919	3.999949	5.000366
	EVar($\hat{\mu}_{gu}$)	0.007276	0.008010	0.008841	0.009987	0.010842
	TVar($\hat{\mu}_{gu}$)	0.007258	0.007911	0.008746	0.009780	0.011036
	EMean($\hat{\omega}_{gu^{**}}$)	0.299975	0.298695	0.299377	0.299276	0.294327
	EVar($\hat{\omega}_{gu^{**}}$)	0.002125	0.002947	0.004738	0.008544	0.020036
	TVar($\hat{\omega}_{gu^{**}}$)	0.002111	0.002957	0.004594	0.008385	0.019787
	Optimal n_1	757	784	807	826	838
	Optimal n_2	243	216	193	174	162

Table 2. Simulation results of quantitative models: trials = 10000, $p_a = 0.85$, $\mu_y = 7.0$, $p_b = 0.6$, $\pi_b = 0.1$, $p_1 = 0.85$, $p_2 = 0.15$, $n = 1000$. Continued on the next two pages.

		$\pi = 1.0$	$\pi = 2.0$	$\pi = 3.0$	$\pi = 4.0$	$\pi = 5.0$
New model	EMean($\hat{\mu}_p$)	1.000104	2.000087	3.000718	4.001133	5.000078
	EVar($\hat{\mu}_p$)	0.004778	0.004885	0.005055	0.005810	0.006419
	TVar($\hat{\mu}_p$)	0.005204	0.005213	0.005416	0.005815	0.006409
	EMean($\hat{\omega}_{p^{**}}$)	0.499925	0.500087	0.500174	0.500064	0.499728
	EVar($\hat{\omega}_{p^{**}}$)	0.000626	0.000628	0.000623	0.000612	0.000613
	TVar($\hat{\omega}_{p^{**}}$)	0.000623	0.000623	0.000623	0.000623	0.000623
Simple unrelated	EMean($\hat{\mu}_g$)	1.001130	1.999331	3.000900	4.001092	4.999751
	EVar($\hat{\mu}_g$)	0.008974	0.008272	0.007543	0.007785	0.008200
	TVar($\hat{\mu}_g$)	0.008983	0.008218	0.007806	0.007747	0.008042
Optional unrelated with optimal split	EMean($\hat{\mu}_{gu}$)	0.999785	2.000865	2.998736	3.999293	4.999645
	EVar($\hat{\mu}_{gu}$)	0.010030	0.009858	0.010053	0.010752	0.011474
	TVar($\hat{\mu}_{gu}$)	0.010021	0.009904	0.010105	0.010627	0.011484
	EMean($\hat{\omega}_{gu^{**}}$)	0.499831	0.498203	0.499094	0.497134	0.496861
	EVar($\hat{\omega}_{gu^{**}}$)	0.002595	0.003521	0.004917	0.008461	0.019331
	TVar($\hat{\omega}_{gu^{**}}$)	0.002614	0.003421	0.004978	0.008500	0.019160
	Optimal n_1	762	782	802	820	835
	Optimal n_2	238	218	198	180	165
New model	EMean($\hat{\mu}_p$)	1.000434	1.999794	3.000376	3.999070	4.998875
	EVar($\hat{\mu}_p$)	0.006498	0.006200	0.006230	0.006572	0.007062
	TVar($\hat{\mu}_p$)	0.006956	0.006570	0.006457	0.006617	0.007051
	EMean($\hat{\omega}_{p^{**}}$)	0.700136	0.700524	0.699319	0.700041	0.700131
	EVar($\hat{\omega}_{p^{**}}$)	0.000695	0.000686	0.000703	0.000678	0.000668
	TVar($\hat{\omega}_{p^{**}}$)	0.000690	0.000690	0.000690	0.000690	0.000690
Simple unrelated	EMean($\hat{\mu}_g$)	0.999969	2.000120	3.000306	3.999057	4.998973
	EVar($\hat{\mu}_g$)	0.009033	0.008287	0.007767	0.007932	0.008049
	TVar($\hat{\mu}_g$)	0.008983	0.008218	0.007806	0.007747	0.008042
Optional unrelated with optimal split	EMean($\hat{\mu}_{gu}$)	0.999089	2.000346	2.999364	4.000602	4.998395
	EVar($\hat{\mu}_{gu}$)	0.012147	0.011724	0.011007	0.011255	0.011901
	TVar($\hat{\mu}_{gu}$)	0.012241	0.011503	0.011193	0.011309	0.011855
	EMean($\hat{\omega}_{gu^{**}}$)	0.700095	0.699526	0.700043	0.698377	0.697909
	EVar($\hat{\omega}_{gu^{**}}$)	0.002767	0.003456	0.004917	0.008352	0.018523
	TVar($\hat{\omega}_{gu^{**}}$)	0.002757	0.003508	0.004967	0.008260	0.018173
	Optimal n_1	777	790	805	820	834
	Optimal n_2	223	210	195	180	166

Table 2 (continued).

		$\pi = 1.0$	$\pi = 2.0$	$\pi = 3.0$	$\pi = 4.0$	$\pi = 5.0$
New model	EMean($\hat{\mu}_p$)	0.997806	1.999640	2.999296	4.000093	5.000968
	EVar($\hat{\mu}_p$)	0.008691	0.007787	0.007430	0.007428	0.007794
	TVar($\hat{\mu}_p$)	0.008770	0.007986	0.007554	0.007475	0.007749
	EMean($\hat{\omega}_{p^{**}}$)	0.900024	0.900222	0.899738	0.899885	0.900270
	EVar($\hat{\omega}_{p^{**}}$)	0.000684	0.000673	0.000662	0.000664	0.000681
	TVar($\hat{\omega}_{p^{**}}$)	0.000677	0.000677	0.000677	0.000677	0.000677
Simple unrelated	EMean($\hat{\mu}_g$)	0.998399	2.000097	2.999036	4.000156	5.001079
	EVar($\hat{\mu}_g$)	0.009059	0.008152	0.007844	0.007811	0.008153
	TVar($\hat{\mu}_g$)	0.008983	0.008218	0.007806	0.007747	0.008042
Optional unrelated with optimal split	EMean($\hat{\mu}_{gu}$)	1.000834	1.997784	2.999923	3.999134	4.999262
	EVar($\hat{\mu}_{gu}$)	0.013713	0.012646	0.011814	0.011912	0.011936
	TVar($\hat{\mu}_{gu}$)	0.013854	0.012677	0.012003	0.011828	0.012148
	EMean($\hat{\omega}_{gu^{**}}$)	0.898653	0.900103	0.898165	0.898256	0.897476
	EVar($\hat{\omega}_{gu^{**}}$)	0.002611	0.003239	0.004693	0.007889	0.016584
	TVar($\hat{\omega}_{gu^{**}}$)	0.002582	0.003297	0.004657	0.007755	0.016864
	Optimal n_1	800	807	815	825	834
	Optimal n_2	200	193	185	175	166

Table 2 (end).

5. Conclusion

We propose a modification of the optional unrelated question models of [Gupta et al. 2013b] for binary and quantitative responses by avoiding the split sample approach, which requires a larger total sample size. Rather than trying to estimate prevalence and sensitivity simultaneously, we estimate them independently by asking two questions. The simulation study shows that the proposed models always achieve smaller variances of the estimators for both binary and quantitative response cases. As shown in the tables, Var($\hat{\pi}_p$), Var($\hat{\omega}_{p^*}$), Var($\hat{\mu}_p$), and Var($\hat{\omega}_{p^{**}}$) are respectively smaller than Var($\hat{\pi}_g$) and Var($\hat{\pi}_{gu}$), Var($\hat{\omega}_{gu^*}$), Var($\hat{\mu}_g$) and Var($\hat{\mu}_{gu}$), and Var($\hat{\omega}_{gu^{**}}$).

References

[Christofides 2003] T. C. Christofides, "A generalized randomized response technique", *Metrika* **57**:2 (2003), 195–200. MR 2004b:62018

[Greenberg et al. 1969] B. G. Greenberg, A.-L. A. Abul-Ela, W. R. Simmons, and D. G. Horvitz, "The unrelated question randomized response model: Theoretical framework", *J. Amer. Statist. Assoc.* **64** (1969), 520–539. MR 40 #982

[Greenberg et al. 1971] B. G. Greenberg, R. R. Kuebler, J. R. Abernathy, and

D. G. Horvitz, "Application of the randomized response technique in obtaining quantitative data", *J. Amer. Statist. Assoc.* **66**:334 (1971), 243–250.

[Gupta 2001] S. Gupta, "Quantifying the sensitivity level of binary response personal interview survey questions", *J. Combin. Inform. System Sci.* **26**:1-4 (2001), 79–86. MR 2049000 Zbl 1219.62015

[Gupta et al. 2002] S. Gupta, B. Gupta, and S. Singh, "Estimation of sensitivity level of personal interview survey questions", *J. Statist. Plann. Inference* **100**:2 (2002), 239–247. MR 1877192 Zbl 0985.62010

[Gupta et al. 2010] S. Gupta, J. Shabbir, and S. Sehra, "Mean and sensitivity estimation in optional randomized response models", *J. Statist. Plann. Inference* **140**:10 (2010), 2870–2874. MR 2011h:62027 Zbl 1191.62009

[Gupta et al. 2013a] S. Gupta, S. Mehta, J. Shabbir, and B. K. Dass, "Generalized scrambling in quantitative optional randomized response models", *Comm. Statist. Theory Methods* **42**:22 (2013), 4034–4042. MR 3170981 Zbl 06250934

[Gupta et al. 2013b] S. Gupta, A. Tuck, T. Spears Gill, and M. Crowe, "Optional unrelated-question randomized response models", *Involve* **6**:4 (2013), 483–492. MR 3115981 Zbl 06227512

[Mangat and Singh 1990] N. S. Mangat and R. Singh, "An alternative randomized response procedure", *Biometrika* **77**:2 (1990), 439–442. MR 1064823 Zbl 0713.62011

[Mehta et al. 2012] S. Mehta, B. K. Dass, J. Shabbir, and S. N. Gupta, "A three-stage optional randomized response model", *J. Stat. Theory Pract.* **6**:3 (2012), 417–427.

[Sihm and Gupta 2015] J. S. Sihm and S. N. Gupta, "A two-stage binary optional randomized response model", *Commun. Stat. Simul. Comp.* **44**:9 (2015), 2278–2296.

[Warner 1965] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias", *J. Amer. Statist. Assoc.* **60**:309 (1965), 63–69.

Received: 2013-11-21 Accepted: 2014-03-20

j_sihm@uncg.edu
Department of Mathematics and Statistics,
The University of North Carolina at Greensboro, 317
College Ave, Greensboro, NC 27412, United States

a.chhabra02@gmail.com
Department of Mathematics, University of Delhi,
Delhi 110052, India

sngupta@uncg.edu
Department of Mathematics and Statistics,
University of North Carolina at Greensboro, 317
College Ave, Greensboro, NC 27412, United States

INVOLVE 9:2 (2016)
dx.doi.org/10.2140/involve.2016.9.211

On counting limited outdegree grid digraphs and greatest increase grid digraphs

Joshua Chester, Linnea Edlin, Jonah Galeota-Sprung, Bradley Isom, Alexander Moore, Virginia Perkins, A. Malcolm Campbell, Todd T. Eckdahl, Laurie J. Heyer and Jeffrey L. Poet

(Communicated by Ronald Gould)

In this paper, we introduce two special classes of digraphs. A limited outdegree grid (LOG) directed graph is a digraph derived from an $n \times n$ grid graph by removing some edges and replacing some edges with arcs such that no vertex has outdegree greater than 1. A greatest increase grid (GIG) directed graph is a LOG digraph whose vertices can be labeled with distinct labels such that each arc represents the direction of greatest increase in the underlying grid graph. We enumerate both GIG and LOG digraphs for the 3×3 case.

1. Introduction

Some search algorithms, such as hill climbing [Russell and Norvig 2010], use local information to seek a global maximum of a function of two variables, $f(x, y)$. At every point in an $n \times n$ lattice, the algorithm determines the direction of greatest increase in f, and moves to the adjacent lattice point in that direction. We can think of this algorithm as discrete gradient ascent. In what follows, we make the simplifying assumptions that the function values are the integers $1, 2, \ldots, n^2$ and directions are restricted to horizontal and vertical on a square grid. For example, consider the function values 1 through 9 on the 3×3 lattice shown in Figure 1(a). The direction of greatest increase from each lattice point is shown in Figure 1(b) as an arrow to the appropriate adjacent point. Note that there are no arrows originating at local maxima on this lattice.

MSC2010: 05C20, 05C30.

Keywords: graph, directed graph, greatest increase grid graph, limited outdegree grid graph, discrete gradient ascent, enumeration.

We gratefully acknowledge the funding support of NSF grants MCB-1120558 and MCB-1120578 awarded to the Missouri Western State University and Davidson College campuses, respectively, the James G. Martin Program in Genomics at Davidson College, the support of our administrators on both campuses, and the input from our other biology and mathematics undergraduates.

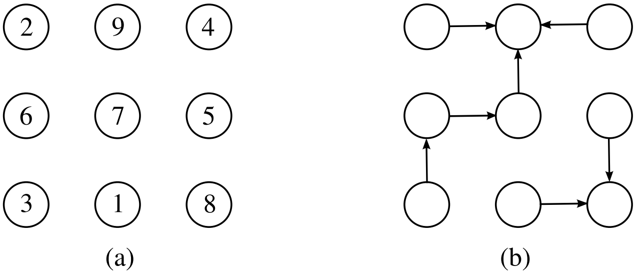


Figure 1. (a) A sample function $f(x, y)$ mapping a 3×3 grid onto $1, 2, \ldots, 9$. (b) The direction of greatest increase in $f(x, y)$ from each grid point.

The lattice points and arrows in Figure 1 are easily described in the language of graph theory [Tutte 2001]. In particular, Figure 1(b) is derived from a grid graph by replacing some edges with a single arc, and eliminating other edges entirely, so there is at most one arc originating at each vertex. We call these limited outdegree grid (LOG) digraphs. If the vertices in a LOG digraph can be labeled with the integers $1, 2, \ldots, n^2$ such that each arc is in the direction of greatest increase from that vertex, we call the graph a greatest increase grid (GIG) digraph. The directed graph in Figure 1(b) is clearly a GIG digraph since it was derived from a labeling of the vertices of a lattice. Other LOG digraphs, such as that shown in Figure 2, are not GIG digraphs. Additionally, GIG digraphs can also be viewed as a type of proximity graph [Bose et al. 2012].

Graph labeling problems, that is, questions that ask if integers can be assigned to the vertices or edges (or both) of a graph subject to given conditions, have been studied for over 50 years. Gallian [2015] has compiled a dynamic survey of the known results of graph labeling problems, and many graph labeling problems are accessible to undergraduate students, such as that in [Poet et al. 2005].

In this paper, we describe two approaches to enumerating the 3×3 GIG digraphs, and a method for enumerating 3×3 LOG digraphs. Finally, we suggest two procedures for deciding if a given LOG digraph is a GIG digraph.

2. Enumerating LOG and GIG digraphs

In counting the number of distinct LOG and GIG digraphs, there are two complicating factors. First, a LOG digraph can be isomorphic to as many as 7 others: those obtained by 90-, 180-, and 270-degree clockwise rotations, and those obtained by reflecting each of these through a horizontal line. These motions are described by the dihedral group on the square. However, a LOG digraph with reflexive or rotational symmetry will have fewer than 8 LOG digraphs in its isomorphism class. Figure 2 illustrates the 8 LOG digraphs in one isomorphism class.

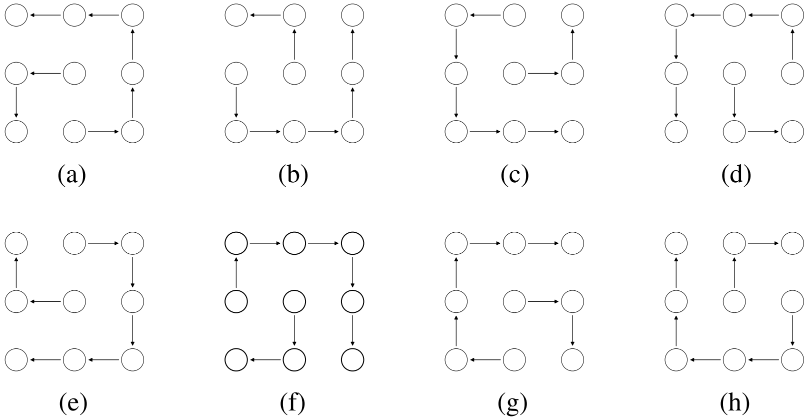


Figure 2. (a) A LOG digraph that is not also a GIG digraph, and its (b) 90-degree, (c) 180-degree, and (d) 270-degree rotations. The LOG digraphs in (e)–(h) are obtained by reflecting those in (a)–(d) through a horizontal line.

The second complicating factor is that a particular GIG digraph can be labeled in more than one way, and the number of ways is dependent upon the underlying LOG digraph. For example, Figure 3 shows three labelings of one particular GIG digraph. The variability described in these two observations prohibits us from being able to find the number of nonisomorphic LOG or GIG digraphs by computing the total number of directed graphs with a certain property and dividing by an easily computable constant. Our research group of undergraduates was split across two campuses, Missouri Western State University and Davidson College. Students from the two campuses took different approaches to enumerating GIG digraphs.

2.1. ***Counting approach 1:*** *construct one candidate LOG digraph from each isomorphism class, and test each one to see if it can be labeled.* On the Missouri Western campus, we approached the problem by considering the list of nonisomorphic candidate LOG digraphs, and then asking if each of these could be labeled.

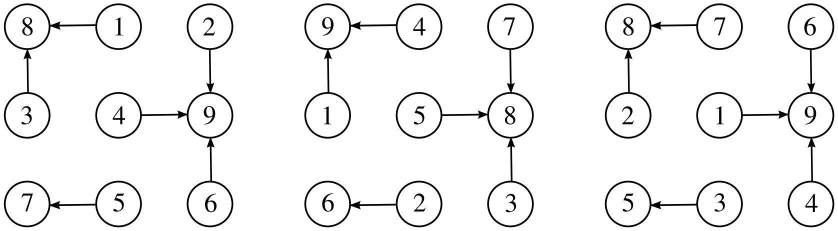


Figure 3. Three possible labelings of the same GIG digraph.

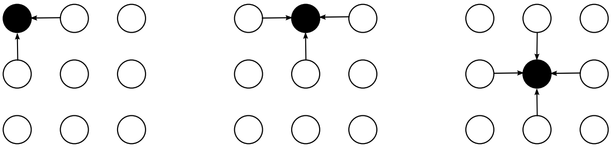


Figure 4. Three possible locations (black vertex) of a complete sink corresponding to the label 9, upper left corner, upper middle, and center.

First, we observe that every GIG digraph must contain at least one vertex that is a complete sink, that is, a vertex with indegree equal to the number of adjacent vertices in the underlying grid graph, corresponding to the label 9. Furthermore, because we want to consider only one candidate LOG digraph in each isomorphism class, we need only consider the label 9 in one of three positions: the upper left corner, the upper middle, and the center. Any valid candidate LOG digraph can be put into correspondence with (at least) one of these three by an appropriate rotation. Hence, the candidate LOG digraphs can be put into three piles (A, B, and C) according to the location of the complete sink, labeled as 9, shown in Figure 4.

These three piles can further be subdivided according to the location of the label 8. For example, if the label 9 is in the upper left, then there are five locations (see Figure 5) that could be labeled with 8 since we want to account for a reflection about the main diagonal. We refer to these configurations as A1, . . . , A5. If the label 9 is in the upper middle, then the label 8 can go in one of the five positions shown in Figure 6 as B1, . . . , B5, taking into account the possible reflection through

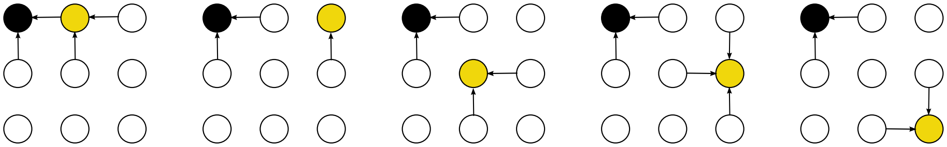


Figure 5. The five possible configurations, A1, . . . , A5, for placing the label 8 (gold vertex), given that the label 9 (black vertex) is in the top left corner.

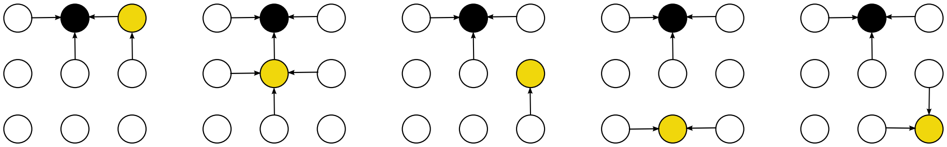


Figure 6. The five possible configurations, B1, . . . , B5, for placing the label 8 (gold vertex), given that the label 9 (black vertex) is in the upper middle.

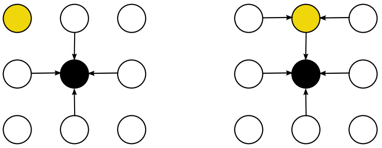


Figure 7. The two possible configurations, C1 and C2, for placing the label 8 (gold vertex) in the upper left and upper center, given that the label 9 (black vertex) is in the center.

the center vertical line. Finally, if the label 9 is in the center, there are only two places (up to rotation) we need to consider for the label 8: the upper left (C1) and the upper middle (C2) as shown in Figure 7. Observe that the digraphs A4 and B5 are isomorphic by a flip through the diagonal that runs from lower left to upper right so we eliminate B5, leaving 11 subsets of candidate LOG digraphs.

With each of these "skeletons" in place, it is relatively straightforward to consider all completions to a GIG digraph by exhaustion. As an example, for subset C1 in Figure 7, we need only consider what arcs might originate from the other three corner vertices. In each case, there are three possibilities: there could be a vertical arc, a horizontal arc, or neither. This leads to a family of 27 candidate LOG digraphs. However, by again taking symmetry into account, this number can further be reduced to the 11 candidates in Figure 8.

Similar arguments can be made to construct the other subsets. While each of our eleven subsets (A1, . . . , A5, B1, . . . , B4, C1, C2) was complete with regard to its

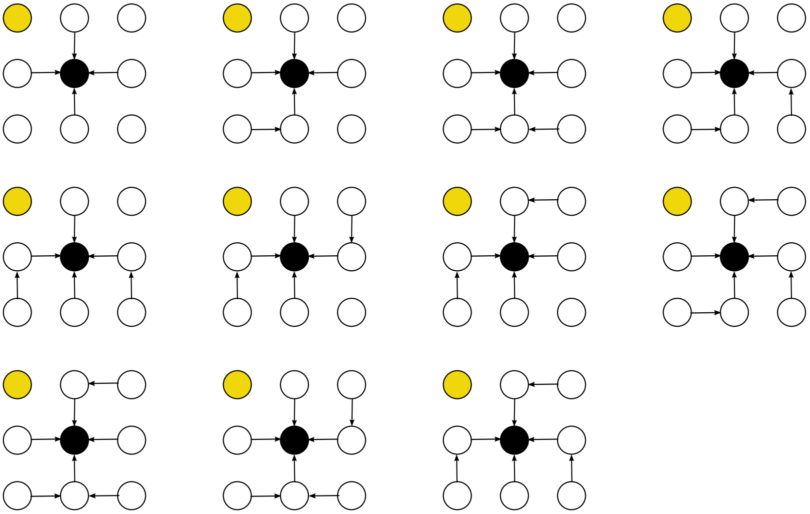


Figure 8. The 11 candidate configurations resulting from enumerating possible completions of configuration C1 in Figure 7.

construction, we knew there was the potential for overlap between sets. Through extensive cross-checking, we were able to eliminate these redundancies. Finally, for each of these potential GIG graphs, we either supplied a labeling of the vertices or provided a justification for why such a labeling was not possible. Our final product was a complete list of the 246 nonisomorphic GIG digraphs.

2.2. *Counting approach 2:* *construct all LOG and GIG digraphs, and identify and discard isomorphic copies.* On the Davidson campus, we produced digraphs on a 3×3 grid using the open-source mathematical software Sage [Stein et al. 2012], and filtered the results for the desired subsets of LOG and GIG digraphs. In this approach, we first needed a convenient data structure for storing and manipulating the graphs. Because a 3×3 grid graph has 12 edges (6 horizontal and 6 vertical), we can represent a 3×3 LOG digraph with a 12×1 arc indicator vector $\vec{a}$. Specifically, we let $a_i = -1$ if arc i points down or to the left, $a_i = 1$ if arc i points up or to the right, and $a_i = 0$ if no arc is present at the ith location. The locations are ordered as shown in Figure 9(a), numbering arcs clockwise around the perimeter of the grid, and then clockwise around the interior of the grid. For example, the LOG digraph in Figure 9(b) is represented by

$$\vec{a} = [1, -1, 0, -1, 1, 0, 1, 0, 1, 0, 0, 1].$$

Using the arc indicator representation, we began by producing all $3^{12} = 531,441$ possible arc indicator vectors, and discarding those that did not correspond to LOG digraphs. Specifically, we removed those vectors that produce an outdegree greater than 1 from any vertex. However, many of the remaining 36,250 LOG digraphs were isomorphic to each other. The isomorphism class of a given LOG digraph is easily obtained through multiplication by rotation and reflection matrices. For example, the equation below illustrates a 90-degree clockwise rotation of the LOG digraph in Figure 9(b):

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ 0 \\ -1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ -1 \\ 1 \\ 0 \\ -1 \\ -1 \\ 0 \\ -1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \qquad (1)$$

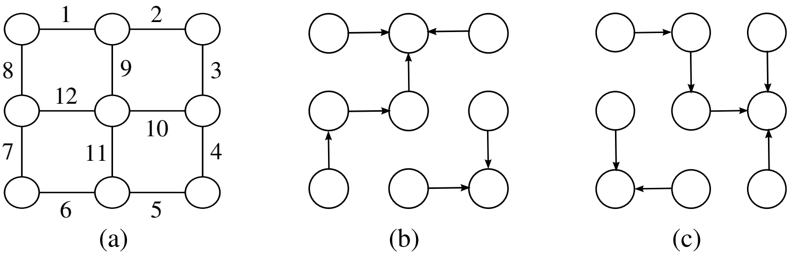


Figure 9. (a) The order in which the arc indicator vector represents arcs in a LOG digraph. (b) The LOG digraph from Figure 1(b). (c) The result of applying a 90-degree clockwise rotation to the LOG digraph in (b).

Note that although the ordering of arcs in the indicator vector was arbitrary, we chose the order illustrated in Figure 9(a) because this ordering produces nice patterns in the rotation and reflection matrices. Discarding isomorphic copies from the list of 36,250 distinct LOG digraphs produced 4,616 isomorphism classes of LOG digraphs. Note that this set includes not only the candidate LOG digraphs from the first approach, but also many LOG digraphs that do not contain a complete sink.

We produced the set of all GIG digraphs, and a unique representative of each isomorphism class, in a similar brute force manner. First, we considered all permutation of the integers 1 through 9, and removed those that were the reverse of another permutation in the set. This reduction was an easy way to filter out those labelings whose GIG digraphs were isomorphic under a 180-degree clockwise rotation. We produced all possible GIG digraphs by mapping these $9!/2$ permutations to the 3×3 grid, and drawing an arc in the direction of greatest increase from each vertex. We obtained 1,853 distinct labeled GIG digraphs with varying numbers of labelings corresponding to each one. Discarding isomorphic copies from the list of 1,853 GIG digraphs produced 246 isomorphism classes of GIG digraphs, the same number obtained through the first approach described in Section 2.1. One advantage of this brute-force computational approach to enumerating LOG and GIG digraphs is that we could easily collect various statistics about the graphs as they were produced. For example, the number of LOG and GIG digraphs with each possible number of arcs is summarized in Table 1.

number of arcs	0	1	2	3	4	5	6	7	8	9
GIG digraphs	0	0	0	0	6	23	86	98	33	0
LOG digraphs	1	4	36	174	570	1,128	1,378	949	335	41

Table 1. The number of GIG and LOG digraphs with each possible number of arcs.

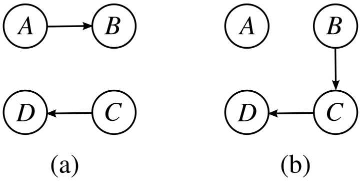


Figure 10. (a) Forbidden subgraph with arcs in opposite directions. (b) Forbidden subgraph with vertex of outdegree 0 and path of length 2.

2.3. ***Determining whether a LOG digraph is a GIG digraph.*** As stated earlier, there are 4,616 nonisomorphic LOG digraphs and 246 of these are GIG digraphs. What follows is a classification of the 4,370 LOG digraphs that are not labelable as GIG digraphs. To show the nonexistence of a labeling for these LOG digraphs, we filter our results with four filters and handle the remaining nine exceptions with ad hoc arguments.

First, observe that for a LOG digraph to be a GIG digraph, it must contain (at least) one vertex that is a complete sink. That is, a 3×3 GIG digraph will contain one of the following: a corner vertex of indegree 2, a side vertex of indegree 3, or a central vertex of indegree 4. The necessity of such a vertex is clear when one considers that the label 9 must appear as a label on a GIG digraph and will be the direction of greatest increase from each of its adjacent vertices. Of the 4,370 unlabelable LOG digraphs, only 614 have a complete sink.

Second, in a GIG digraph, there cannot exist two adjacent vertices (in the underlying grid graph) with outdegree 0. Any vertex of outdegree 0 has the greatest label in its neighborhood, and two adjacent vertices are each in the neighborhood of the other implying that $A < B$ and $B < A$.

Third, a GIG digraph cannot contain a 2×2 subgrid with two arcs on opposite sides of that subgrid pointing in opposite directions. In such a grid, if the vertices are labeled clockwise as A, B, C, and D with arcs AB and CD (as shown in Figure 10(a)), we observe that B and D are each in the neighborhood of A and the arc AB implies that $B > D$. The vertices B and D are also each in the neighborhood of C and the arc CD implies that $D > B$, a contradiction. Thus, such a subgraph cannot occur in a GIG digraph.

These two forbidden conditions are easy to spot in LOG digraphs. Of the 614 unlabelable LOG digraphs with at least one complete sink, all but 74 are eliminated by these two criteria. As our fourth and final filter, we next consider another 2×2 forbidden subgraph.

Suppose a GIG digraph contains a 2×2 subgrid with a vertex of outdegree 0, which we label A, and a path of length 2 on the other three vertices which we label to yield arcs BC and CD, as shown in Figure 10(b). Note that we can assume

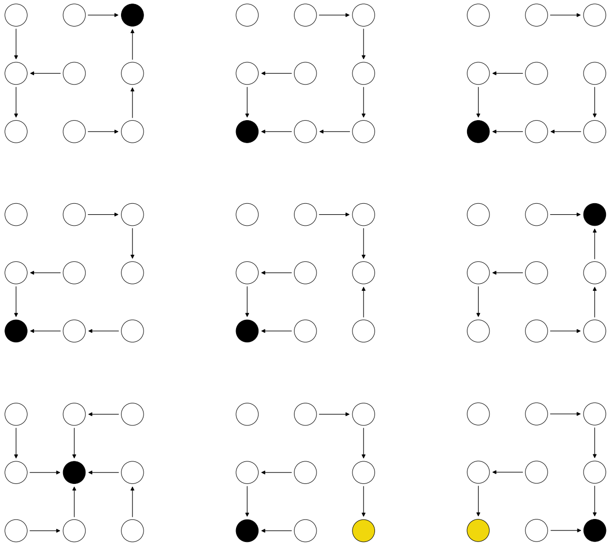


Figure 11. The nine unlabelable LOG digraphs that contain a complete sink, but do not contain one of the two forbidden induced subgraphs.

that D is not of outdegree 0 or the GIG digraph would have adjacent vertices with outdegree 0, which is forbidden. Since A has outdegree 0 and D is adjacent to A in the grid graph, $A > D$. Since A and C are both adjacent to B and the GIG digraph contains arc BC, we have $C > A$. Along any directed path in a GIG digraph, the labels must increase. Hence $D > C$. This gives a contradiction: $A < D$ and $D > A$.

Of the 74 remaining unlabelable GIG digraphs, 65 contain the forbidden subgraph in Figure 10(b), leaving only the 9 graphs in Figure 11. Of these 9 exceptional graphs, the first 7 can be eliminated from consideration as possible GIG digraphs by observing that in addition to a GIG digraph having a complete sink (so that the label 9 can be placed), it must also have either a second complete sink or a near complete sink, so that the label 8 can be placed. A near complete sink is a vertex that is (i) distance 2 from the complete sink in the underlying grid graph and (ii) all vertices adjacent to this vertex in the underlying grid graph and not adjacent to the complete sink terminate at this vertex.

Finally we consider the last two of our exceptional graphs, each of which have one complete sink and one near complete sink, these must be labeled with 9 and 8, respectively. It is easy to see that the label 7 cannot be placed on any of the remaining vertices without creating a contradiction. The vertex of label 7 must be the terminal vertex of an arc for every adjacent vertex that is not also adjacent to the complete sink or the near complete sink, but this does not hold for any of the 7 remaining vertices.

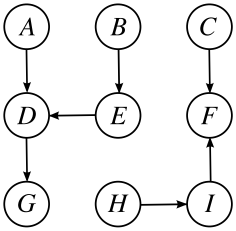


Figure 12. A LOG digraph that is not a GIG digraph.

We have, therefore, shown the nonexistence of a labeling scheme for 4,370 nonisomorphic LOG digraphs and have demonstrated a labeling (not shown here) for each of the 246 nonisomorphic GIG digraphs.

The second approach for determining if a LOG digraph is a GIG digraph relies on the properties of a GIG digraph, specifically the strict inequalities that each arc (or lack thereof) confers on the labels of the vertices. For example, consider the LOG digraph on vertices A–I shown in Figure 12. Suppose this is a GIG digraph. Then arc ED implies $D > F$, arc IF implies $F > I$, arc HI implies $I > G$, and arc DG implies $G > D$. Hence $D > D$, a contradiction. Therefore, this LOG digraph cannot be labeled as a GIG digraph. Note that we could have used other criteria to draw this conclusion, as this LOG digraph does not contain a complete sink. The inequality consistency checking method works for every 3×3 LOG digraph, as we confirmed with a SAGE program.

Many questions about LOG and GIG digraphs remain open. An obvious question is how the numbers of each type of graph, and the numbers of isomorphism classes, grow with increasing grid size. However, applying the techniques described here make extensions of this problem, even to a 4×4 grid, a monumental (and tedious) task, even for a computer. In future research, we hope to investigate new techniques to generalize our results, with the ultimate goal of enumerating $m \times n$ LOG and GIG digraphs. Another potential direction would be to search for efficient characterizations of forbidden subgraphs as the size of the $n \times n$ grid increases. We hope to prove such sets of forbidden subgraphs are both necessary and sufficient by some nonexhaustive method.

References

[Bose et al. 2012] P. Bose, V. Dujmović, F. Hurtado, J. Iacono, S. Langerman, H. Meijer, V. Sacristán, M. Saumell, and D. R. Wood, "Proximity graphs: E, δ, Δ, χ and ω", *Internat. J. Comput. Geom. Appl.* **22**:5 (2012), 439–469. MR 3028530 Zbl 1267.05072

[Gallian 2015] J. A. Gallian, "A dynamic survey of graph labeling", *Electron. J. Combin. 5* **5** (2015), Dynamic Survey 6, pp. 389.

[Poet et al. 2005] J. L. Poet, V. Onkoba, D. Daffron, H. Goforth, and C. Thomas, "On super edge-magic labelings of unions of star graphs", *J. Combin. Math. Combin. Comput.* **53** (2005), 49–63. MR 2137836 Zbl 1071.05066

[Russell and Norvig 2010] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed., Pearson Education, Upper Saddle River, NJ, 2010.

[Stein et al. 2012] W. A. Stein et al., *Sage mathematics software*, Version 5.0, Sage Development Team, 2012, available at http://www.sagemath.org.

[Tutte 2001] W. T. Tutte, *Graph theory*, Encyclopedia of Mathematics and its Applications **21**, Cambridge University Press, Cambridge, 2001. MR 1813436 Zbl 0964.05001

Received: 2,014-03-10 Revised: 2,015-04-28 Accepted: 2,015-05-02

chestermathematics@gmail.com *Department of Mathematics, University of Oklahoma,*
Norman, OK 73069, United States

ledlin@missouriwestern.edu *Department of Computer Science, Mathematics, and Physics,*
Missouri Western State University, Saint Joseph, MO 64507,
United States

sprungj@gmail.com *Department of Mathematics, Davidson College,*
Davidson, NC 28035, United States

bisom1@missouriwestern.edu *Department of Mathematics, University of Kansas,*
Lawrence, KS 66045, United States

moorealexanderk@gmail.com *Department of Chemistry, University of Illinois at*
Urbana–Champaign, Urbana, IL 61801, United States

vipe5210@gmail.com *Department of Computer Science, Mathematics, and Physics,*
Missouri Western State University, Saint Joseph, MO 64507,
United States

macampbell@davidson.edu *Department of Biology, Davidson College,*
Davidson, NC 28035, United States

eckdahl@missouriwestern.edu *Department of Biology, Missouri Western State University,*
Saint Joseph, MO 64507, United States

laheyer@davidson.edu *Department of Mathematics, Davidson College,*
Davidson, NC 28035, United States

poet@missouriwestern.edu *Department of Computer Science, Mathematics, and Physics,*
Missouri Western State University, Saint Joseph, MO 64507,
United States

INVOLVE 9:2 (2016)
dx.doi.org/10.2140/involve.2016.9.223

Polygonal dissections and reversions of series

Alison Schuetz and Gwyn Whieldon

(Communicated by Kenneth S. Berenhaut)

The Catalan numbers C_k were first studied by Euler, in the context of enumerating triangulations of polygons P_{k+2}. Among the many generalizations of this sequence, the Fuss–Catalan numbers $C_k^{(d)}$ count enumerations of dissections of polygons $P_{k(d-1)+2}$ into $(d+1)$-gons. In this paper, we provide a formula enumerating polygonal dissections of $(n+2)$-gons, classified by partitions λ of $[n]$. We connect these counts a_λ to reverse series arising from iterated polynomials. Generalizing this further, we show that the coefficients of the reverse series of polynomials $x = z - \sum_{j=0}^{\infty} b_j z^{j+1}$ enumerate colored polygonal dissections.

1. Catalan numbers and polygonal partitions

The Catalan numbers

$$C_k = \frac{1}{k+1}\binom{2k}{k} \quad \text{for } n \geq 0$$

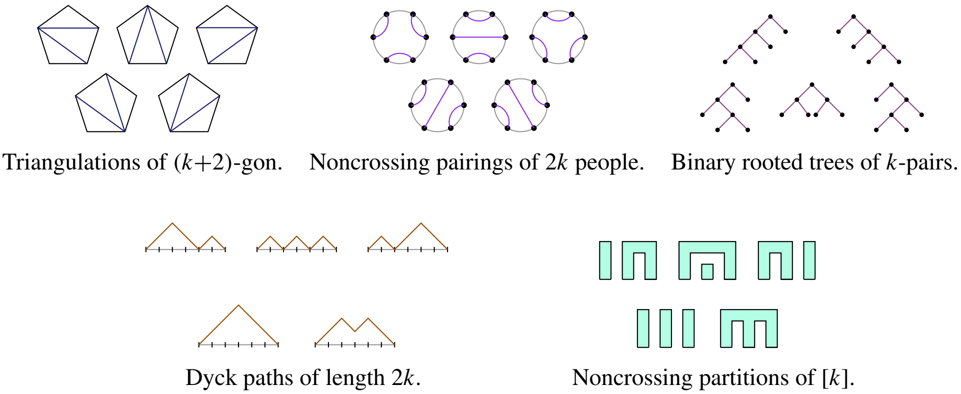
are the answer to myriad counting problems (see [Stanley 2012; 2013; Bajunaid et al. 2005]). For example, they count the number of triangulations of a $(k+2)$-gon, the number of noncrossing handshake-pairings of $2k$ people seated at a round table, the number of binary rooted trees with k internal nodes, the number of Dyck paths of length $2k$, and the number of noncrossing partitions of k; see Figure 1.

In this paper, we will alternate between the recursive definition of the Catalan numbers and the closed formula for C_k.

Theorem 1.1 (Catalan recursion [Stanley 2012]). *Let $\{C_0, C_1, C_2, \ldots\}$ be a sequence with $C_0 = 1$ and $C_{k+1} = \sum_{i=0}^{k} C_i \cdot C_{k-i}$. Then C_k are the Catalan numbers,*

$$C_k = \frac{1}{k+1}\binom{2k}{k}.$$

MSC2010: 05A15, 05E99.
Keywords: Catalan, Fuss–Catalan, series reversion, Lagrange inversion, polygon partitions.
Both authors were supported by the Summer Research Institute at Hood College. Whieldon also received supported from a Board of Associates Grant from Hood College.



Triangulations of $(k+2)$-gon. Noncrossing pairings of $2k$ people. Binary rooted trees of k-pairs.

Dyck paths of length $2k$. Noncrossing partitions of $[k]$.

Figure 1. Examples of sets counted by Catalan number $C_3 = 5$.

There is a similar recursive formula for the Fuss–Catalan numbers

$$C_k^{(d)} = \frac{1}{k(d-1)+1}\binom{dk}{k},$$

which specializes to the Catalan numbers when $d = 2$.

Theorem 1.2 (Fuss–Catalan recursion [Klarner 1970]). *Let $\{C_0^{(d)}, C_1^{(d)}, C_2^{(d)}, \dots\}$ be a sequence with $C_0^{(d)} = 1$ and*

$$C_{k+1}^{(d)} = \sum_{k_1+k_2+\cdots+k_d=n} C_{k_1}^{(d)} C_{k_2}^{(d)} \cdots C_{k_d}^{(d)}.$$

Then $C_k^{(d)}$ are the generalized Catalan (*or Fuss–Catalan*) *numbers*

$$C_k^{(d)} = \frac{1}{k(d-1)+1}\binom{dk}{k}.$$

There is a well-known bijection between triangulations of $(k+2)$-gons P_{k+2} and binary rooted trees with k internal nodes (see [Przytycki and Sikora 2000] for a history of this problem). There is also a bijection between partitions of a $(k(d-1)+2)$-gon $P_{k(d-1)+2}$ into $(d+1)$-gons and d-ary trees with k internal nodes (see [Hilton and Pedersen 1991]). The wording of Theorem 0.2 from [loc. cit.] has been changed slightly to reflect the notation used in this note.

Theorem 1.3 [Hilton and Pedersen 1991, Theorem 0.2]. *Let P_k^d denote the number of ways of subdividing a convex polygon into k disjoint $(d+1)$-gons by means of nonintersecting diagonals, $k \geq 1$, and let A_k^d denote the number of d-ary trees with k-internal nodes, $k \geq 1$. Then $P_k^d = A_k^d = C_k^{(d)}$ for all $d \geq 2$, $k \geq 1$.*

Proof. That the number of d-ary rooted trees with k internal nodes can be counted by the Fuss–Catalan numbers can easily be shown inductively via the generalized Catalan recursion formula. For $k = 1$, there are precisely $A_k^d = 1$ such trees. For

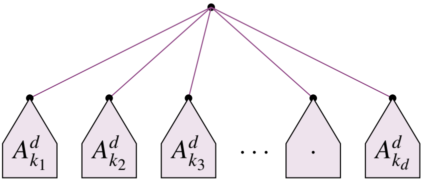


Figure 2. Recursive construction of A^d_{k+1}.

$k \geq 1$, each tree in the set A^d_{k+1} consists of an internal node with d branches and some rooted d-ary tree (possibly empty) attached to each branch (see Figure 2).

As there are k remaining internal nodes to partition amongst the branches, we have $k_1 + \cdots + k_d = k$, and our d-ary rooted trees with $(k+1)$ internal nodes must satisfy the recursion formula

$$A^d_{k+1} = \sum_{k_1+k_2+\cdots+k_d=k} A^d_{k_1} A^d_{k_2} \cdots A^d_{k_d}.$$

By Theorem 1.2, we have that the number of d-ary rooted trees is $A^d_k = C^{(d)}_k$.

There is a bijection (illustrated in Figure 3) between d-ary rooted trees with k internal nodes and subdivisions of convex polygons into k disjoint $(d{+}1)$-gons via diagonals. Choose one edge to correspond to the root; then draw the d branches from that root to the d other edges of the $(d{+}1)$-gon in the dissection. Continue this, treating each vertex lying on a diagonal as the new root of a subtree.

This process is easy to reverse (constructing a unique polygonal dissection from a rooted tree); i.e., given a polygonal dissection into $(d{+}1)$-gons, we have a unique d-ary rooted tree, and vice versa. This provides our desired bijection, and $P^d_k = A^d_k = C^{(d)}_k$. □

There have been several papers enumerating general polygonal dissections of n-gons with k nonintersecting diagonals (see [Motzkin 1948; Read 1978; McCammond 2006]). We consider here a case that does not currently appear in the literature: enumerating polygonal dissections of convex n-gons P_n, where each

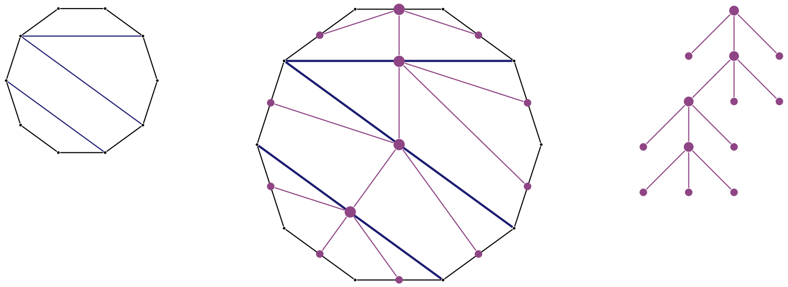


Figure 3. Bijection between $(d+1)$ dissections and rooted d-ary trees.

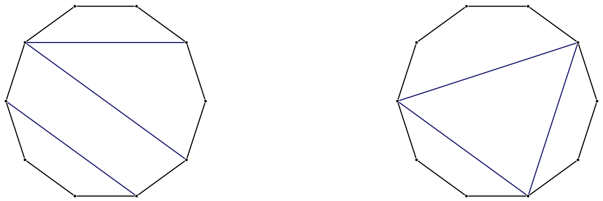


Figure 4. A 3-dissection (left) of type $\lambda = 2+2+2+2$ and a (2, 3, 4)-dissection (right) of type $\lambda = 3+2+2+1$ of a 10-gon.

piece of the dissection must be a $(d+1)$-gon, where $d \in \{d_1, d_2, \dots, d_k\}$ for fixed, distinct integers $d_i \geq 2$ (see Figure 4 for examples). To standardize terminology and indices for *polygonal dissections*, we include precise definitions here.

Definition 1.4 (polygonal dissections). A *polygonal dissection* of a convex n-gon is the union of the polygon and any nonintersecting subset of its diagonals. A *d-dissection* (respectively, a *$(d_1, d_2, \dots, d_r)$-dissection*) is a polygonal dissection such that the regions formed by the dissection are all convex $(d+1)$-gons (respectively, each region is a (d_i+1)-gon for some $d_i \in \{d_1, d_2, \dots, d_r\}$).

Definition 1.5 (type of a polygonal dissection). Let λ be a partition of n with k_j parts of size j. We say a dissection of an $(n+2)$-gon consisting of k_j $(j+2)$-gons is a *polygonal dissection of type* λ, and denote the set of all such polygonal dissections as $P_{\lambda,n}$.

Polygonal dissections of an $(n(d-1)+1)$-gon into $(d+1)$-gons are in bijection with d-ary rooted trees with $n+1$ internal nodes, as shown in Theorem 1.3. Similarly, polygonal dissections of type λ are in bijection with rooted plane trees with a particular downdegree sequence.

Definition 1.6 (rooted trees and downdegree sequences). A *rooted plane tree* is a tree T with a distinguished vertex called the *root*. The *downdegree sequence* $\boldsymbol{r} = (r_0, r_1, r_2, \dots, r_n)$ of a rooted tree counts the number of vertices r_j with j neighbors further away from the root than the vertex itself. See Figure 5 for an example.

Theorem 1.7 [Stanley 2012]. *Let $P_{\lambda,n}$ be the number of all polygonal dissections of type λ, where λ is a partition of n with k_j parts of size j and n total parts. Let $T_{\boldsymbol{r},m}$ be the number of rooted plane trees on $m+1$ vertices with downdegree sequence $\boldsymbol{r} = (r_0, r_1, r_2, \dots, r_m)$. Then $P_{\lambda,n} = T_{\boldsymbol{r},m}$ for $\boldsymbol{r} = (n+1, 0, k_1, k_2, \dots, k_n)$ and $m = n+k$.*

Proof. Let λ be a partition of n as above, and fix a polygonal dissection of type $\lambda \in P_{\lambda,n}$. We construct our tree in $T_{\boldsymbol{r},n+k}$ recursively as follows:

Choose an edge of the $(n+2)$-gon and place a root v there. This will be an edge of *some* $(j+2)$-gon in the dissection with $k_j \geq 1$ in λ. Place a vertex on each of

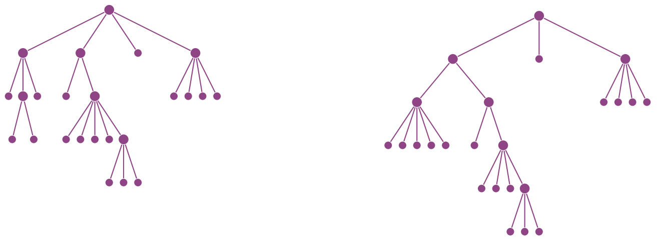


Figure 5. Rooted planar trees with downdegree sequence $(17, 0, 2, 2, 2, 1, 0, 0, \dots, 0)$.

the $j+1$ other sides of the $(j+2)$-gon, and connect each of these vertices to the original root vertex. The root note will have $(j+1)$ neighbors further from the root, contributing one to the value of r_{j+1} in the downdegree sequence of the rooted tree.

Repeat this process for each of the edges in the dissection now connected to v. If a node is connected to an edge in the boundary of the $(n+2)$-gon, it will contribute one to the value of r_0 in the downdegree sequence (as it will be a leaf of the rooted tree, and has no further neighbors). If a node is *not* on the boundary of the $(n+2)$-gon, treat it as the new root node of a subtree, and repeat the first step.

For each $k_j \geq 1$ in λ, we will have $k_j = r_{j+1}$ vertices with downdegree j, and as each boundary edge of our $(n+2)$-gon (excepting our first root node) will have a leaf vertex placed on it, we must have $r_0 = (n+2) - 1 = n+1$. The tree constructed has k internal vertices, one for the root and $k-1$ for the diagonals, and $n+1$ leaves, so we have $n+k+1$ total vertices in our tree. Note that no vertices in this tree will have downdegree 1, so $r_1 = 0$. So our constructed tree is in $T_{\boldsymbol{r},m}$ for $\boldsymbol{r} = (n+1, 0, k_1, k_2, \dots, k_n)$ and $m = n+k$.

Note also that this bijection can easily be reversed: Given a rooted tree in $T_{\boldsymbol{r},n+k}$ with k internal nodes each with some downdegree $j \geq 2$, we may place the internal

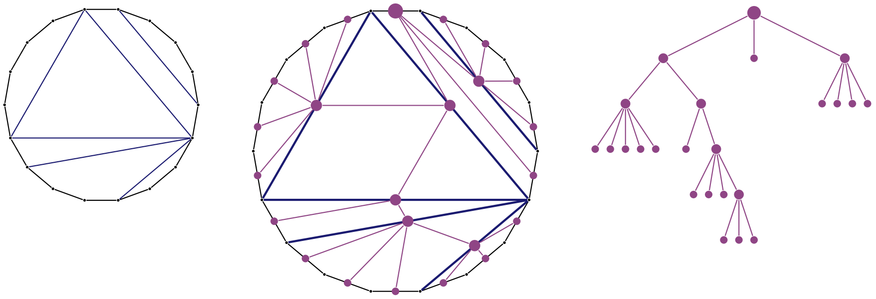


Figure 6. Bijection between a dissection of an 18-gon of type $\lambda = 1+1+2+2+3+3+4 = 16$ and a rooted planar tree with downdegree sequence $d = (17, 0, 2, 2, 2, 1, 0, 0, \dots, 0)$.

node and its neighbors on the edges of a $(j+1)$-gon. Glue a pair of these polygons together along an edge if they share a vertex, and shift the resulting shape so that all edges that are unmatched form the boundary of a convex polygon. Of the internal nodes, only the original root node will appear on the boundary of this polygon. As there were $n+k+1$ original vertices and k internal vertices, there must be $n+1$ vertices on the boundary apart from the root node. So we have constructed a polygonal dissection of an $(n+2)$-gon with k_j $(j+2)$-gons, and our bijection is complete. □

For an illustration of one such bijection between a polygonal dissection and a rooted planar tree, see Figure 6.

Using results of Kreweras [1972] and Armstrong and Eu [2008], the count for the number of rooted trees with a fixed downdegree sequence is known:

Theorem 1.8 [Rhoades 2011, Theorem 1.1]. *Let $n \geq 1$, $\boldsymbol{v} = (1, 2, \ldots, n)$ and $\boldsymbol{r} = (r_1, \ldots, r_n)$ such that $\boldsymbol{v} \cdot \boldsymbol{r} = n$. Set $\boldsymbol{r}! = r_1! r_2! \cdots r_n!$ and $|\boldsymbol{r}| = \sum r_j$. Then the number of rooted plane trees with $n+1$ vertices and downdegree sequence $(n - |\boldsymbol{r}| + 1, r_1, r_2, \ldots, r_n)$ is*

$$A_{\boldsymbol{r}}(\boldsymbol{v}) = \frac{1}{1+n} \frac{(1+n)_{|\boldsymbol{r}|}}{\boldsymbol{r}!},$$

where $(y)_k = y(y-1) \cdots (y-k+1) = y!/(y-k)!$ is the falling factorial.

For a list of several other related classes of connected Catalan-type objects, enumerated by type and counted by the same formula, see the recent paper [Rhoades 2011]. From Theorem 1.7, we have added a new class of objects (polygonal dissections of type λ) to their list. We will make use of the count provided by this bijection to prove our main theorem, showing the connection between the coefficients of reverse series of certain types and polygonal dissections.

Theorem 1.9 (polygonal partitions). *The polynomial $x = z - \sum_{i=1}^{r} z^{d_i}$ with $2 \leq d_1 < d_2 < \cdots < d_r$ has reverse series $z = \sum_{n=0}^{\infty} a_n x^{n+1}$, where a_n counts the number of $(d_1, d_2, \ldots, d_r)$-dissections of a convex $(n+2)$-gon.*

Before delving into the connections between series reversions and polygonal dissections, we examine in Section 2 the initial problem that led us to consider reversions of series of the form $x = z - z^d$.

2. Iterated Mandelbrot polynomials and reversions of series

The work in this paper began initially as a study of the coefficients of iterated Mandelbrot (and d-multibrot) polynomials.

Definition 2.1 (Mandelbrot and d-multibrot polynomials). For variables $x, z \in \mathbb{C}$, we define the Mandelbrot polynomial f_x by $f_x(z) = z^2 + x$. For $d \geq 2$, we define the d-multibrot polynomial f_d to be the map $f_{d,x}(z) = z^d + x$.

Of particular interest in complex dynamics is the orbit of 0 under f_x or $f_{d,x}$. We were interested in a formula for the coefficients of the power series in x of the infinitely iterated d-multibrot polynomial

$$f_d^{(\infty)}(x) = \lim_{n\to\infty} f_d^{(n)}(x),$$

where $f_d^{(n)}(x)$ is defined recursively by the formula $f_d^{(0)}(x) = 0$ and

$$f_d^{(n)}(x) = (f_d^{(n-1)}(x))^d + x \quad \text{for } n \geq 1.$$

Note that if the subscript d is omitted, we assume that $d = 2$.

The power series obtained by considering the limit of the iteration of zero under the d-multibrot polynomial $f_d^{(\infty)}(x) = \sum_k a_k x^{k+1}$ must satisfy

$$f_d^{(\infty)}(x) = (f_d^{(\infty)}(x))^d + x.$$

Setting $z = f_d^{(\infty)}(x)$, we see that calculating the coefficients of x in $f_d^{(\infty)}(x)$ is equivalent to computing the series reversion of the polynomial $x = z - z^d$. With this in mind, here we introduce a version of the Lagrange inversion formula to explicitly calculate the coefficients of $f_d^{(\infty)}(x) = \sum_{k=0}^{\infty} a_k x^{k+1}$.

Theorem 2.2 (Lagrange inversion formula [Muller 1985]). *Let x be a* (*convergent*) *power series*

$$x = z\left(1 - \sum_{n=1}^{\infty} b_n z^n\right),$$

with reverse series

$$z = x\left(1 + \sum_{n=1}^{\infty} a_n x^n\right).$$

Then the coefficients a_n are given in terms of the b_n by

$$a_n = \frac{1}{n+1} \sum_{\lambda} \binom{n+k}{k} \binom{k}{k_1, k_2, \ldots, k_n} \prod_{j=1}^{n} b_j^{k_j},$$

where the sum is taken across all partitions λ of n into k_j parts of size j and k total parts, *i.e.*, *across all nonnegative integer n-tuples $\{k_1, k_2, \ldots, k_n\}$ such that*

$$\sum_{j=1}^{n} k_j = k,$$

$$\sum_{j=1}^{n} k_j \cdot j = n.$$

As an immediate application of the Lagrange inversion formula, we produce the series reversion of the polynomial $x = z - z^d$:

Theorem 2.3. *The polynomial $z = z^d + x$ has inverse series solution*

$$z = \sum_{k=0}^{\infty} C_k^{(d)} x^{k(d-1)+1}.$$

Proof. This result is fairly immediate from noting that only for $j = d-1$ are the b_j nonzero (specifically $b_{d-1} = 1$). So all parts in partitions λ contributing to the sum are of size $(d-1)$, and nonzero a_n must be of form $n = k(d-1)$. From the Lagrange inversion formula in Theorem 2.2, these coefficients must then be

$$\begin{aligned} a_n &= \frac{1}{n+1}\binom{n+k}{k} \\ &= \frac{1}{k(d-1)+1}\binom{k(d-1)+k}{k} \\ &= \frac{1}{k(d-1)+1}\binom{kd}{k} \\ &= C_k^{(d)}. \end{aligned}$$

So the only nonzero terms in our series reversion are of the form

$$a_n x^{n+1} = a_{k(d-1)} x^{k(d-1)+1} = C_k^{(d)} x^{k(d-1)+1},$$

and we have our inverse series for $x = z - z^d$. □

As a corollary, we have that the coefficients of the infinitely iterated d-multibrot polynomials are given by the Fuss–Catalan numbers $C_k^{(d)}$. While this result was found and proved independently by the authors, the following statement appears to be fairly well known for $d = 2, 3$ (see the OEIS at A001764). The formula holds in general for all $d \geq 2$.

Corollary 2.4 (coefficients of infinitely iterated d-multibrot polynomials). *Let $f_d^{(n)}(x)$ be defined recursively by the formula $f_d^{(0)}(x) = 0$ and*

$$f_d^{(n)}(x) = (f_d^{(n-1)}(x))^d + x \quad \textit{for } n \geq 1,$$

and set $f_d^{(\infty)}(x) = \lim_{n\to\infty}(f_d^{(n)}(x))$. Then

$$f_d^{(\infty)}(x) = \sum_{k=0}^{\infty} C_k^{(d)} x^{k(d-1)+1}.$$

Proof. This is immediate from Theorem 2.3 and the fact that $z = f_d^{(\infty)}(x) = \sum_k a_k x^{k+1}$ must satisfy

$$f_d^{(\infty)}(x) = (f_d^{(\infty)}(x))^d + x,$$

or $z = z^d + x$. Note that this corollary could also be proved fairly directly via induction and the general recursion formula for the Fuss–Catalan series found in Theorem 1.2. □

As a further interesting note from this, the sum of the series formula for z found in the Mandelbrot case gives a formula for the two fixed points of the (filled) Julia set $\mathcal{J}_x$ (the set of all points $z \in \mathbb{C}$ such that the orbit of 0 remains bounded under iterations by $f_x(z) = z^2 + x$) for a fixed $x \in \mathbb{C}$. See [Milnor 2006] for more details on dynamical systems and their fixed points.

Remark 2.5 (fixed points of filled Julia sets $\mathcal{J}_x$). The series reversion of $z = z^2 + x$ is

$$\begin{aligned} z &= \sum_{k=0}^{\infty} C_k x^{k+1} \\ &= x \sum_{k=0}^{\infty} C_k x^k \\ &= \frac{2x}{1+\sqrt{1-4x}}. \end{aligned}$$

For a fixed $x \in \mathbb{C}$, the two complex values taken on by $2x/(1+\sqrt{1-4x})$ each correspond to a separate fixed point of the Mandelbrot map, one each for the stable and unstable fixed points of $f_x(z) = z^2 + x$ in $\mathcal{J}_x$.

3. Iterations of general polynomials and polygonal dissections

To return to polygonal partitions and their connections to the reversions of series, we note that Theorem 2.3 gives us immediately that d-dissections of polygons are counted by the coefficients of the series inverse of $x = z - z^d$.

Corollary 3.1. *The coefficients a_k of the series inversion $z = \sum_{k=0}^{\infty} a_k x^{k+1}$ of the polynomial $z = z^d + x$ count the number of $(d+1)$-gon polygonal partitions of a $(k+2)$-gon.*

Proof. From Theorem 2.3, we have that the coefficients of the series reversion of $x = z - z^d$ are the Fuss–Catalan numbers

$$C_n^{(d)} = \frac{1}{(d-1)n+1}\binom{nd}{d}.$$

The corollary is immediate from Theorem 1.3, as the Fuss–Catalan numbers enumerate d-partitions of $(n+2)$-gons. □

This will be a special case of our main theorem:

Theorem 1.9 (polygonal partitions). *The polynomial* $x = z - \sum_{i=1}^{r} z^{d_i}$ *with* $2 \le d_1 < d_2 < \cdots < d_r$ *has reverse series* $z = \sum_{n=0}^{\infty} a_n x^{n+1}$, *where* a_n *counts the number of* $(d_1, d_2, \ldots, d_r)$*-dissections of a convex* $(n+2)$*-gon.*

We begin our proof with a lemma counting the number of polygonal dissections of a fixed type (with a fixed number of each type of $(d+1)$-gon appearing in the dissection).

Lemma 3.2. *Fix integers* $2 \le d_1 < d_2 < \cdots < d_r$, *an integer* $n \ge 0$, *and a partition* λ *of* n *with parts of sizes* $j \in \{d_1 - 1, d_2 - 1, \ldots, d_r - 1\}$. *Let* k_j *for* $1 \le j \le r$ *be the number of times that* j *appears in* λ, *and let* k *be the total parts in* λ, *i.e.*,

$$n = \sum_{j=1}^{r} (d_j - 1) k_j,$$

$$k = \sum_{j=1}^{r} k_j.$$

Then the number of all polygonal dissections of type λ *of an* $(n+2)$*-gon is given by*

$$a_\lambda = \frac{1}{n+1} \binom{n+k}{k} \binom{k}{k_1, k_2, \ldots, k_r}.$$

Proof. From Theorem 1.7, we know that the number of polygonal dissections of type λ above is in bijection with the set of rooted planar trees with downdegree sequence $\boldsymbol{r} = (n+1, 0, k_1, k_2, \ldots, k_n)$ and $n+k+1$ vertices. From Theorem 1.8, we know that the count of such planar trees is

$$\begin{aligned} A_{\boldsymbol{r}}(\boldsymbol{v}) &= \frac{1}{1+n+k} \frac{(1+n+k)_k}{0!\, k_1!\, k_2! \cdots k_n!} \\ &= \frac{1}{n+k+1} \frac{(n+k+1)!}{k_1!\, k_2!\, \cdots k_n!\, (n+1)!} \\ &= \frac{1}{n+1} \frac{(n+k)!}{k_1!\, k_2! \cdots k_n!\, n!} \\ &= \frac{1}{n+1} \frac{(n+k)!}{k!\, n!} \frac{k!}{k_1!\, k_2! \cdots k_n!} \\ &= \frac{1}{n+1} \binom{n+k}{k} \binom{k}{k_1, k_2, \ldots, k_n}. \end{aligned}$$

This proves the count a_λ given in the statement of the theorem. □

With this in hand, we return to the proof of the main theorem.

Proof of Theorem 1.9. Given an $(n+2)$-gon, any fixed partition λ of n into positive integer parts of sizes chosen from the set $\{d_1 - 1, d_2 - 1, \ldots, d_r - 1\}$ corresponds to some fixed *type* of polygonal $(d_1, d_2, \dots, d_r)$-dissection. From Lemma 3.2, we know that there are

$$a_\lambda = \frac{1}{n+1}\binom{n+k}{k}\binom{k}{k_{d_1-1}, k_{d_2-1}, \dots, k_{d_r-1}}$$

such dissections, where k_{d_j-1} parts of size $d_j - 1$ appear in partition λ. Note that we have changed from k_j to k_{d_j-1} to better match the notation used in the statement the Lagrange inversion theorem.

Examining our polynomial

$$x = z - \sum_{i=1}^{r} z^{d_i} = z\biggl(1 - \sum_{i=1}^{r} z^{d_i-1}\biggr),$$

we see that in the notation of the Lagrange inversion formula given in Theorem 2.2, the only nonzero b_j are those with $j = d_i - 1$ for some $1 \le i \le r$. So the coefficients a_n of the reverse series $z = \sum_{i=0}^{\infty} a_n x^{n+1}$ are of the form

$$a_n = \frac{1}{n+1}\sum_{\lambda}\binom{n+k}{k}\binom{k}{k_1, k_2, \dots, k_n},$$

where the sum is taken across partitions λ of the form

$$n = k_{d_1-1}(d_1 - 1) + k_{d_2-1}(d_2 - 1) + \cdots + k_{d_r-1}(d_r - 1).$$

Note that

$$\begin{aligned} a_n &= \frac{1}{n+1}\sum_{\lambda}\binom{n+k}{k}\binom{k}{k_1, k_2, \dots, k_n} \\ &= \sum_{\lambda}\frac{1}{n+1}\binom{n+k}{k}\binom{k}{k_{d_1-1}, k_{d_2-1}, \dots, k_{d_r-1}} \\ &= \sum_{\lambda} a_\lambda, \end{aligned}$$

and our coefficients a_n can be calculated by summing over all possible types of dissections in λ, made from parts of size $(d+1)$, with $d \in \{d_1, d_2, \dots, d_r\}$. This completes the proof. □

Example 3.3. Consider the polynomial $f(z) = z^3 + z^2 + x$. The coefficients of the infinitely iterated polynomial are given by the series reversion of $z = z^3 + z^2 + x$, or $x = z - z^3 - z^2$:

$$z = x + x^2 + 3x^3 + 10x^4 + 38x^5 + 154x^6 + 654x^7 + \cdots.$$

These coefficients a_n count the number of dissections of an $(n+2)$-gon into triangles (3-gons) and quadrilaterals (4-gons).

As there are no 2-gons, there is one way to cover the empty object with triangles or squares, so the coefficient of x is $a_0 = 1$. For $n = 1, 2, 3$, see Figure 7.

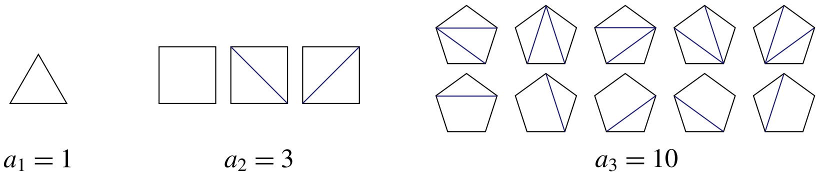


Figure 7. (2, 3)-dissections of n-gons for $n = 1, 2, 3$.

Extending this slightly, we have a power series whose reverse series has coefficients counting all dissections of an $(n+2)$-gon by noncrossing diagonals.

Theorem 3.4 (super-Catalan numbers and series reversions). *The power series $x = z - \sum_{j=1}^{\infty} z^j$ has reverse series $z = \sum_{k=0}^{\infty} s_n x^{n+1}$, where s_n counts the all possible subsets of noncrossing diagonals of a convex $(n+2)$-gon. The coefficient s_n is given by*

$$s_n = \frac{1}{n+1} \sum_{\lambda} \binom{n+k}{k} \binom{k}{k_1, k_2, \ldots, k_n},$$

where the sum is taken across all partitions λ of n with k_j parts of size j and k total parts.

Proof. From the Lagrange inversion theorem, we know that the coefficient s_n in the reverse power series of $x = z - \sum_{j=1}^{\infty} z^j$ must be of the form given in the statement of the theorem. All partitions λ of n contributing to the sum must have parts at *most* n, so the s_n above must be the same as the coefficient a_n for the reverse series of the polynomial $x = z - \sum_{j=1}^{n+1} z^j$. From Theorem 1.9, we know that the coefficients a_n of the reversion of the polynomial with nonzero terms $z^2, z^3, \ldots, z^{n+1}$ enumerate $(2, 3, \ldots, n+1)$-dissections of an $(n+2)$-gon — a set which includes all possible polygonal dissections. □

The set of all polygonal dissections of an $(n+2)$-gon is counted by the super-Catalan numbers s_n (also called the Schröder–Hipparchus numbers). (See [Fan et al. 2005] for an extensive list of other families of sets counted by s_n.) While several other formulas for the super-Catalan numbers are known, Theorem 3.4 gives a nice decomposition of s_n, summed across structures indexed by partitions λ of n.

4. Generalizations to colored dissections

The coefficients of slightly more general series reversions can be immediately interpreted using the formula in Theorem 1.9.

Definition 4.1. A *colored polygonal dissection* is a polygonal dissection where each $(d+1)$-gon appearing in the dissection can be assigned b_d possible colors for $d \geq 2$.

Theorem 4.2 (colored polygonal partitions). *The polynomial $x=z-\sum_{i=1}^{r} b_{d_i} z^{d_i}$ with $d_1>d_2>\cdots>d_r\geq 2$ and $b_{d_i}\geq 1$ for all $1\leq i\leq r$ has reverse series $z=\sum_{k=0}^{\infty} a_n x^{n+1}$, where a_n counts the number of colored polygonal $(d_1, d_2, \dots, d_r)$-dissections of a convex $(n+2)$-gon.*

Proof. From Lemma 3.2, we know that the number of $(d_1, d_2, \dots, d_r)$-partitions of an n-gon with precisely k_j of the (d_j+1)-gons appearing in the dissection for $1\leq j\leq r$ is given by

$$a_\lambda = \frac{1}{n+1}\binom{n+k}{k}\binom{k}{k_{d_1-1}, k_{d_2-1}, \dots, k_{d_r-1}}.$$

If each (d_i+1)-gon can be assigned one of b_{d_i} colors, then there are

$$a_\lambda^* = \frac{1}{n+1}\binom{n+k}{k}\binom{k}{k_{d_1-1}, k_{d_2-1}, \dots, k_{d_r-1}}\prod_{i=1}^{r} b_{d_i}^{k_{d_i-1}}$$

such colored dissections, as we have b_{d_i} choices for each of k_{d_i-1} of the (d_i+1)-gons appearing in a given dissection.

As in the proof of Theorem 1.9, we have that the coefficients of the inverse series of the polynomial $x=z-\sum_{i=1}^{r} b_{d_i} z^{d_i}$ must be

$$\begin{aligned} a_n &= \frac{1}{n+1}\sum_{\lambda}\binom{n+k}{k}\binom{k}{k_1, k_2, \dots, k_n}\prod_{j=1}^{n} b_{j+1}^{k_j} \\ &= \sum_{\lambda}\frac{1}{n+1}\binom{n+k}{k}\binom{k}{k_{d_1-1}, k_{d_2-1}, \dots, k_{d_r-1}}\prod_{i=1}^{r} b_{d_r}^{k_{d_r-1}} \\ &= \sum_{\lambda} a_\lambda^*. \end{aligned}$$

□

5. Further questions

This paper provides a complete combinatorial interpretation of series reversions of polynomials of the form

$$z=b_1 z^{d_1}+b_2 z^{d_2}+\cdots+b_r z^{d_r}+x$$

for positive integers b_j. As future work, we would be curious to see combinatorial approaches to the following question:

Question 5.1. In general, given a pair of polynomials $g(z)$ and $h(x)$ with integer coefficients, is there a family of sets of objects $\mathcal{A}_{g,h}$ counted by the coefficients of the reversion of the power series $z=g(z)+h(x)$?

This question is answered here for $h(x)=x$ and $g(z)$ with *positive* integer coefficients and all terms of degree at least two, but remains open in other cases.

Series other than the generating functions of Catalan-type objects may appear as series inversions using similar iterative techniques, and we would be interested in seeing other classes of objects enumerated by such coefficients.

References

[Armstrong and Eu 2008] D. Armstrong and S.-P. Eu, "Nonhomogeneous parking functions and noncrossing partitions", *Electron. J. Combin.* **15**:1 (2008), R146. MR 2009k:05007 Zbl 1163.05302

[Bajunaid et al. 2005] I. Bajunaid, J. M. Cohen, F. Colonna, and D. Singman, "Function series, Catalan numbers, and random walks on trees", *Amer. Math. Monthly* **112**:9 (2005), 765–785. MR 2006h:11023 Zbl 1168.60012

[Fan et al. 2005] A. N. Fan, T. Mansour, and S. X. M. Pang, "Elements of sets enumerated by super-Catalan numbers", preprint, 2005, available at http://math.haifa.ac.il/toufik/enumerative/supercat.pdf.

[Hilton and Pedersen 1991] P. Hilton and J. Pedersen, "Catalan numbers, their generalization, and their uses", *Math. Intelligencer* **13**:2 (1991), 64–75. MR 93d:05006 Zbl 0767.05010

[Klarner 1970] D. A. Klarner, "Correspondences between plane trees and binary sequences", *J. Combinatorial Theory* **9** (1970), 401–411. MR 45 #1773 Zbl 0205.54702

[Kreweras 1972] G. Kreweras, "Sur les partitions non croisées d'un cycle", *Discrete Math.* **1**:4 (1972), 333–350. MR 46 #8852 Zbl 0231.05014

[McCammond 2006] J. McCammond, "Noncrossing partitions in surprising locations", *Amer. Math. Monthly* **113**:7 (2006), 598–610. MR 2007c:05015 Zbl 1179.05015

[Milnor 2006] J. Milnor, *Dynamics in one complex variable*, 3rd ed., Annals of Mathematics Studies **160**, Princeton University Press, 2006. MR 2006g:37070 Zbl 1085.30002

[Motzkin 1948] T. Motzkin, "Relations between hypersurface cross ratios, and a combinatorial formula for partitions of a polygon, for permanent preponderance, and for non-associative products", *Bull. Amer. Math. Soc.* **54** (1948), 352–360. MR 9,489d Zbl 0032.24607

[Muller 1985] J. W. Muller, "Some observations on the reversion of series", Rapport BIPM-85/1, Bureau International des Poids et Mesures, 1985, available at http://www.bipm.org/utils/common/pdf/rapportBIPM/1985/01.pdf.

[Przytycki and Sikora 2000] J. H. Przytycki and A. S. Sikora, "Polygon dissections and Euler, Fuss, Kirkman, and Cayley numbers", *J. Combin. Theory Ser. A* **92**:1 (2000), 68–76. MR 2001g:05005 Zbl 0959.05004

[Read 1978] R. C. Read, "On general dissections of a polygon", *Aequationes Math.* **18**:3 (1978), 370–388. MR 80e:05069 Zbl 0396.05028

[Rhoades 2011] B. Rhoades, "Enumeration of connected Catalan objects by type", *European J. Combin.* **32**:2 (2011), 330–338. MR 2012e:05035 Zbl 1227.05048

[Stanley 2012] R. P. Stanley, *Enumerative combinatorics, I*, 2nd ed., Cambridge Studies in Advanced Mathematics **49**, Cambridge University Press, 2012. MR 2868112 Zbl 1247.05003

[Stanley 2013] R. P. Stanley, "Catalan addendum", 2013, available at http://www-math.mit.edu/~rstan/ec/catadd.pdf.

Received: 2014-09-02 Revised: 2015-02-06 Accepted: 2015-02-06

ags9@hood.edu
Department of Mathematics, Hood College,
401 Rosemont Avenue Frederick, MD 21701, United States

whieldon@hood.edu
Department of Mathematics, Hood College,
401 Rosemont Avenue, Frederick, MD 21701, United States

INVOLVE 9:2 (2016)
dx.doi.org/10.2140/involve.2016.9.237

Factor posets of frames and dual frames in finite dimensions

Kileen Berry, Martin S. Copenhaver, Eric Evert, Yeon Hyang Kim, Troy Klingler, Sivaram K. Narayan and Son T. Nghiem

(Communicated by David Royal Larson)

We consider frames in a finite-dimensional Hilbert space, where frames are exactly the spanning sets of the vector space. A factor poset of a frame is defined to be a collection of subsets of I, the index set of our vectors, ordered by inclusion so that nonempty $J \subseteq I$ is in the factor poset if and only if $\{f_i\}_{i \in J}$ is a tight frame. We first study when a poset $P \subseteq 2^I$ is a factor poset of a frame and then relate the two topics by discussing the connections between the factor posets of frames and their duals. Additionally we discuss duals with regard to ℓ^p-minimization.

1. Introduction

A frame for a finite-dimensional Hilbert space is a possibly redundant spanning set. The concept of frames was introduced by Duffin and Schaeffer [1952]. Daubechies [1992] popularized the use of frames. Many of the modern signal processing algorithms used in mobile phones or digital televisions are developed using the concept of frames. Redundancy in frames plays a pivotal role in the construction of stable signal representations and in mitigating the effect of losses in transmission of signals through communication channels [Goyal et al. 2001; 1998]. A tight frame is a special case of a frame, which has a reconstruction formula similar to that of an orthonormal basis. Because of this simple formulation of reconstruction, tight frames are employed in a variety of applications such as sampling, signal processing, filtering, smoothing, denoising, compression, and image processing.

A factor poset $\mathbb{F}_F$ of a frame $F = \{f_i\}_{i \in I}$ is the collection of subsets $J \subseteq I$ such that $\{f_j\}_{j \in J}$ is a tight frame for a finite-dimensional Hilbert space $\mathcal{H}^n$. We find necessary conditions for a given poset to be a factor poset of a frame. We show that a factor poset is determined entirely by its empty cover (the sets $J \in \mathbb{F}_F$ that have no proper subset in $\mathbb{F}_F$). Moreover, we show that if P is the factor poset of

MSC2010: 42C15, 05B20, 15A03.
Keywords: frames, tight frames, factor poset, ℓ_p-norm.
Research supported by NSF-REU Grant DMS 11-56890.

a frame $F \subseteq \mathbb{R}^2$ then it is also the factor poset of another frame $G \subseteq \mathbb{R}^2$ whose vectors are multiples of the standard orthonormal basis vectors e_1 and e_2.

We study the relationship among the factor posets of dual frame pairs. Also we study when the dual frame could be tight and when a dual frame can be scaled to be a tight frame. Finally we consider the group structure among all duals of a frame. It is known that a dual is a canonical dual frame if and only if the ℓ^2-sum of the frame coefficients is a minimizer among ℓ^2-sums of frame coefficients of all dual frames. We find new inequalities among ℓ^p-sums of these frame coefficients when $p = 1$ and $p > 2$.

2. Preliminaries

Throughout this paper $\mathcal{H}^n$ denotes either $\mathbb{R}^n$ or $\mathbb{C}^n$. A sequence $F = \{f_i\}_{i=1}^k \subseteq \mathcal{H}^n$ is called a frame for $\mathcal{H}^n$ with frame bounds $A, B > 0$ if for any $f \in \mathcal{H}^n$,

$$A\|f\|^2 \le \sum_{i=1}^{k} |\langle f, f_i\rangle|^2 \le B\|f\|^2. \tag{1}$$

When $A = B = \lambda$, we say that F is a λ-tight frame. For a sequence $F = \{f_i\}_{i=1}^k \subseteq \mathcal{H}^n$, define the analysis operator θ_F from $\mathcal{H}^n$ to $\mathcal{H}^k$ by

$$\theta_F x = \sum_{i=1}^{k} \langle x, f_i\rangle e_i,$$

where $\{e_i\}_{i=1}^k$ is an orthonormal basis for $\mathcal{H}^k$. The adjoint of θ_F, denoted by $\theta_F^* : \mathcal{H}^k \to \mathcal{H}^n$, is defined by $\theta_F^*(e_i) = f_i$ and is called the synthesis operator. The frame operator $\sigma_F : \mathcal{H}^n \to \mathcal{H}^n$ associated to F is defined by $\sigma_F = \theta_F^*\theta_F$, is a positive definite, self-adjoint, invertible operator and all of its eigenvalues belong to the interval $[A, B]$.

Given a frame F, another frame $G = \{g_i\}_{i=1}^k \subseteq \mathcal{H}^n$ is said to be a dual frame of F if the following reconstruction formula holds:

$$f = \sum_{i=1}^{k} \langle f, f_i\rangle g_i \quad \text{for all } f \in \mathcal{H}^n.$$

The canonical dual frame $\widetilde{F}$ associated with $F = \{f_i\}_{i=1}^k$ is given by $\widetilde{F} = \{\sigma_F^{-1} f_i\}_{i=1}^k$.

Definition 2.1. For any vector

$$f = \begin{bmatrix} f(1) \\ \vdots \\ f(n) \end{bmatrix} \in \mathbb{R}^n,$$

we define the diagram vector of f, denoted $\tilde{f}$, by

$$\tilde{f} = \frac{1}{\sqrt{n-1}} \begin{bmatrix} f^2(1) - f^2(2) \\ \vdots \\ f^2(n-1) - f^2(n) \\ \sqrt{2n} f(1) f(2) \\ \vdots \\ \sqrt{2n} f(n-1) f(n) \end{bmatrix} \in \mathbb{R}^{n(n-1)},$$

where the difference of squares $f^2(i) - f^2(j)$ and the product $f(i)f(j)$ occur exactly once for $i < j$, with $i = 1, \dots, n-1$.

Definition 2.2. For any vector $f \in \mathbb{C}^n$, we define the diagram vector $\tilde{f}$ of f to be

$$\tilde{f} = \frac{1}{\sqrt{n-1}} \begin{bmatrix} |f(1)|^2 - |f(2)|^2 \\ \vdots \\ |f(n-1)|^2 - |f(n)|^2 \\ \sqrt{n}\, f(1)\overline{f(2)} \\ \sqrt{n}\, \overline{f(1)} f(2) \\ \vdots \\ \sqrt{n}\, f(n-1)\overline{f(n)} \\ \sqrt{n}\, \overline{f(n-1)} f(n) \end{bmatrix} \in \mathbb{C}^{3n(n-1)/2},$$

where the difference of the form $|f(i)|^2 - |f(j)|^2$ occurs exactly once for $i < j$, with $i = 1, 2, \dots, n-1$, and the product of the form $f(i)\overline{f(j)}$ occurs exactly once for $i \neq j$.

Using these definitions, a characterization of tight frames in $\mathcal{H}^n$ is given in [Copenhaver et al. 2014].

Theorem 2.3 [Copenhaver et al. 2014]. *Let $\{f_i\}_{i \in I}$ be a sequence of vectors in $\mathcal{H}^n$, not all of which are zero. Then $\{f_i\}_{i \in I}$ is a tight frame if and only if $\sum_{i \in I} \tilde{f}_i = 0$. Moreover, for any $f, g \in \mathcal{H}^n$, we have $(n-1)\langle \tilde{f}, \tilde{g} \rangle = n|\langle f, g \rangle|^2 - \|f\|^2 \|g\|^2$.*

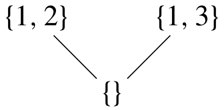
3. Factor posets

In [Lemvig et al. 2014], a tight frame $F = \{f_i\}_{i \in I}$ in $\mathcal{H}^n$ is said to be *prime* if no proper subset of F is a tight frame for $\mathcal{H}^n$. One of the main results in [loc. cit.] is that for $k \geq n$, every tight frame of k vectors in $\mathcal{H}^n$ is a finite union of prime tight frames called *prime factors* of F. Thus to study the structure of prime factors, we use a well-known combinatorial object, the poset. A nonempty set P with a partial ordering is called a partially ordered set, or poset. A poset can be represented by a Hasse diagram. We define a poset related to frames as follows:

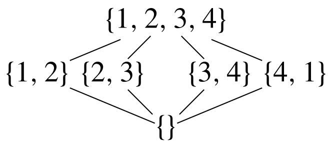
Definition 3.1. Let $F = \{f_i\}_{i \in I} \subseteq \mathcal{H}^n \setminus \{0\}$ be a finite frame, where $I = \{1, \dots, k\}$. The *factor poset* of F, denoted $\mathbb{F}_F$, is defined to be a collection of subsets of I

ordered by set inclusion so that nonempty $J \subseteq I$ is in $\mathbb{F}_F$ if and only if $\{f_j\}_{j\in J}$ is a tight frame for $\mathcal{H}^n$. We always assume $\varnothing \in \mathbb{F}_F$.

Example 3.2. Let $F=\{e_1, e_2, e_2\}\subseteq\mathbb{R}^2$ and $I=\{1,2,3\}$. Then $\mathbb{F}_F=\{\varnothing, \{1,2\}, \{1,3\}\}$ and the Hasse diagram is



Example 3.3. Let $F=\{e_1, e_2, -e_1, -e_2\}\subseteq\mathbb{R}^2$ and $I=\{1, 2, 3, 4\}$. Then the Hasse diagram of $\mathbb{F}_F$ is



The next lemma gives us three equivalent conditions for when the union of elements of $\mathbb{F}_F$ is an element of $\mathbb{F}_F$.

Proposition 3.4. *Let $F=\{f_i\}_{i\in I}\subseteq\mathcal{H}^n\setminus\{0\}$ be a tight frame. Suppose $\mathbb{F}_F$ is the factor poset and let C, $D\in\mathbb{F}_F$. Then the following are equivalent*:

(i) $C\cup D\in\mathbb{F}_F$.

(ii) $C\cap D\in\mathbb{F}_F$.

(iii) $C\triangle D\in\mathbb{F}_F$.

(iv) $C\setminus D\in\mathbb{F}_F$.

Proof. By the inclusion-exclusion principle, it is easy to verify that for diagram vectors of F, the following hold:

(a) $\sum_{\ell\in C\cup D}\tilde{f}_\ell=\sum_{\ell\in C}\tilde{f}_\ell+\sum_{\ell\in D}\tilde{f}_\ell-\sum_{\ell\in C\cap D}\tilde{f}_\ell$.

(b) $\sum_{\ell\in C\cup D}\tilde{f}_\ell=\sum_{\ell\in C\setminus D}\tilde{f}_\ell+\sum_{\ell\in D\setminus C}\tilde{f}_\ell+\sum_{\ell\in C\cap D}\tilde{f}_\ell$.

(c) $\sum_{\ell\in C}\tilde{f}_\ell=\sum_{\ell\in C\setminus D}\tilde{f}_\ell+\sum_{\ell\in C\cap D}\tilde{f}_\ell$.

Since C, $D\in\mathbb{F}_F$, using Theorem 2.3 we have $\sum_{\ell\in C}\tilde{f}_\ell=\sum_{\ell\in D}\tilde{f}_\ell=0$. Hence, from (a) we see that (i) $\iff$ (ii). By the definition of symmetric difference $C\triangle D$, the implication (i) $\iff$ (ii) and (b) above, it follows that (i) $\Longrightarrow$ (iii). Conversely, if (iii) holds, then from (b) we have $\sum_{\ell\in C\cup D}\tilde{f}_\ell=\sum_{\ell\in C\cap D}\tilde{f}_\ell$. But from (a), when C, $D\in\mathbb{F}_F$ we have $\sum_{\ell\in C\cup D}\tilde{f}_\ell=-\sum_{\ell\in C\cap D}\tilde{f}_\ell$. Hence (iii) $\Longrightarrow$ (ii). Thus (i)–(iii) are equivalent. Using (c) above, we conclude that (iv) $\iff$ (ii). □

The above proposition gives some necessary conditions for a given poset to be a factor poset of a frame.

Proposition 3.5. *Let $F = \{f_i\}_{i \in I} \subseteq \mathcal{H}^n \setminus \{0\}$ be a finite frame with corresponding factor poset $\mathbb{F}_F$. Then for any $m \in \mathbb{N}$ with $m \geq |I| = k$, there exists a frame sequence $G = \{g_j\}_{j \in J} \subseteq \mathcal{H}^n \setminus \{0\}$, where $J = \{1, \dots, m\}$, such that $\mathbb{F}_F = \mathbb{F}_G$.*

Proof. We show that there exists some $g_{k+1} \in \mathcal{H}^n \setminus \{0\}$ so that $G = F \cup \{g_{k+1}\}$ satisfies $\mathbb{F}_F = \mathbb{F}_G$. Consider

$$T = \Big\{ -\sum_{\ell \in L} \tilde{f}_\ell : \varnothing \subsetneq L \subseteq I \Big\}.$$

This is a finite collection of vectors in $\mathbb{R}^{n(n-1)}$ or $\mathbb{C}^{3n(n-1)/2}$. Now select $g_{k+1} \in \mathcal{H}^n \setminus \{0\}$ so that $\tilde{g}_{k+1} \notin T$. This completes the proof. □

Definition 3.6. For a frame $F = \{f_i\}_{i \in I} \subseteq \mathcal{H}^n$ and its factor poset $\mathbb{F}_F$, we define the empty cover of $\mathbb{F}_F$, denoted $\mathrm{EC}(\mathbb{F}_F)$, to be the set of $J \in \mathbb{F}_F$ which cover $\varnothing \in \mathbb{F}_F$; that is,

$$\mathrm{EC}(\mathbb{F}_F) = \big\{ J \in \mathbb{F}_F : J \neq \varnothing \text{ and } \nexists J' \in \mathbb{F}_F \text{ with } \varnothing \subsetneq J' \subsetneq J \big\}.$$

Example 3.7. Let $F = \{e_1, e_2, -e_1, -e_2\} \subseteq \mathbb{R}^2$. As seen from Example 3.3,

$$\mathrm{EC}(\mathbb{F}_F) = \big\{ \{1, 2\}, \{2, 3\}, \{3, 4\}, \{4, 1\} \big\}.$$

We now show that a factor poset is entirely determined by its empty cover.

Proposition 3.8. *Let $F = \{f_i\}_{i \in I} \subseteq \mathcal{H}^n$ be a finite frame. If $\mathbb{F}_F$ is the factor poset of F, then for any nonempty J in $\mathbb{F}_F \setminus \mathrm{EC}(\mathbb{F}_F)$, there exists $J_1, J_2 \in \mathbb{F}_F$ with $J_1 \subsetneq J$ and $J_2 \subsetneq J$ so that $J_1 \cap J_2 = \varnothing$ and $J_1 \sqcup J_2 = J$.*

Proof. Let $J \in \mathbb{F}_F \setminus \mathrm{EC}(\mathbb{F}_F)$ be a nonempty set. There must exist some nonempty $J_1 \in \mathbb{F}_F$ so that $J_1 \subsetneq J$, otherwise $J \in \mathrm{EC}(\mathbb{F}_F)$. Using Proposition 3.4, it is easy to see that $J_2 := J \setminus J_1 \in \mathbb{F}_F$. By the assumption on J and J_1, we see that J_2 is nonempty. Hence $J = J_1 \sqcup J_2$. □

Corollary 3.9. *Let $F = \{f_i\}_{i \in I}, G = \{g_i\}_{i \in I}$ be finite frames in $\mathcal{H}^n$ with factor posets $\mathbb{F}_F$ and $\mathbb{F}_G$, respectively. Then $\mathrm{EC}(\mathbb{F}_F) = \mathrm{EC}(\mathbb{F}_G)$ if and only if $\mathbb{F}_F = \mathbb{F}_G$.*

Proof. It is obvious that if $\mathbb{F}_F = \mathbb{F}_G$ then the empty covers are equal, so we restrict our attention to the other direction. It suffices to show that the factor poset of the frame is entirely determined by its empty cover. Let $T = \mathrm{EC}(\mathbb{F}_F) \cup \{\varnothing\}$ for a frame $F = \{f_i\}_{i \in I} \subseteq \mathcal{H}^n$ with $\mathbb{F}_F$ as its factor poset. For every $J_1, J_2 \in T$, if $J_1 \cap J_2 \in T$ then append $J_1 \cup J_2$ to T. Repeat this process until no more unions can be added. This process must terminate after finitely many iterations since I is finite. Clearly the new collection of sets, which we again denote by T, is contained in $\mathbb{F}_F$. From Proposition 3.8, the reverse inclusion holds. Therefore the factor poset $\mathbb{F}_F$ is determined by $\mathrm{EC}(\mathbb{F}_F) \cup \{\varnothing\}$. The desired result follows. □

As a consequence of Corollary 3.9, we get an alternate proof of the following result from [Lemvig et al. 2014].

Corollary 3.10. *Every tight frame $F = \{f_i\}_{i\in I} \subseteq \mathcal{H}^n \setminus \{0\}$ can be written as a union of prime tight frames.*

The proof of Corollary 3.10 follows from observing that if $J \in \mathrm{EC}(\mathbb{F}_F)$ then $\{f_j\}_{j\in J}$ is a prime tight frame. An important case of factor posets occurs when we consider a tight frame $F = \{f_i\}_{i\in I} \subseteq \mathcal{H}^n \setminus \{0\}$. Note that when F is tight, we have that $\varnothing, I \in \mathbb{F}_F$.

Definition 3.11. Suppose $F = \{f_i\}_{i\in I} \subseteq \mathcal{H}^n \setminus \{0\}$ is a frame. Let $\chi(F)$ denote the sequence indexed by I, where $\chi(F)(i)$ is the number of times i occurs in $\mathrm{EC}(\mathbb{F}_F)$ for each $i \in I$. We call $\chi(F)$ the characteristic of F. If $\chi(F)(i) = m$ for all $i \in I$ then F is said to have uniform characteristic.

Example 3.12. Let $F = \{e_1, e_1, e_2\} \subseteq \mathbb{R}^2$. Then $\mathrm{EC}(\mathbb{F}_F) = \{\{1, 3\}, \{2, 3\}\}$ and $\chi(F) = (1, 1, 2)$. Hence $\chi(F)$ need not be uniform.

Proposition 3.13. *If $F = \{f_i\}_{i\in I} \subseteq \mathcal{H}^n \setminus \{0\}$ has positive uniform characteristic, then F is a tight frame.*

Proof. Suppose that F has uniform characteristic $m > 0$. Let $T_1, \dots, T_h$ be the elements of $\mathrm{EC}(\mathbb{F}_F)$. Then $\sum_{i\in T_q} \tilde{f}_i = 0$ for $q = 1, \dots, h$. So $\sum_{q=1}^{h} \sum_{i\in T_q} \tilde{f}_i = 0$. Since $j \in I$ occurs in $\mathrm{EC}(\mathbb{F}_F)$ m times, it follows that $\tilde{f}_j$ occurs m times in the sum $\sum_q \sum_{i\in T_q} \tilde{f}_i = 0$. Hence

$$\sum_{j\in I} \tilde{f}_j = \frac{1}{m}\biggl(\sum_q \sum_{i\in T_q} \tilde{f}_i\biggr) = 0.$$

Hence F is a tight frame. □

Remark 3.14. The condition in Proposition 3.13 is sufficient but not necessary. Consider the frame $F = \{e_1, e_1, e_2, e_2, e_1 + e_2, e_1 - e_2\}$, which is a tight frame, but $\chi(F) = (2, 2, 2, 2, 1, 1)$.

The following theorem states that given a factor poset P of a frame in $\mathbb{R}^2$, we can find another frame with vectors parallel to e_1 and e_2 (the standard orthonormal basis vectors) whose factor poset is also P.

Theorem 3.15. *Let $F = \{f_i\}_{i\in I} \subseteq \mathbb{R}^2 \setminus \{0\}$ be a finite frame with $I = \{1, \dots, k\}$. Then there exists a frame $G = \{g_i\}_{i\in I}$ whose vectors are scaled multiples of the standard orthonormal basis vectors e_1 and e_2 such that $\mathbb{F}_F = \mathbb{F}_G$.*

Proof. Let $\{J_\ell : 1 \le \ell \le 2^k\}$ be an enumeration of 2^I and let

$$2^I \setminus \mathbb{F}_F = \biggl\{ J_{\ell_r} : \sum_{i\in J_{\ell_r}} \tilde{f}_i \neq 0 \biggr\}.$$

Consider a projection P of rank 1 on $\mathbb{R}^2$ such that range$(P) \neq \big(\text{span}\{\sum_{i\in J_\ell} \tilde{f}_i\}\big)^\perp$ for any J_ℓ. Let $\widetilde{V} = \{P(\tilde{f}_i) : 1 \leq i \leq k\}$ and V be the set of vectors of cardinality k in $\mathbb{R}^2$ whose diagram vectors are equal to the set $\widetilde{V}$.

We now claim that $\sum_{i\in J_\ell} \tilde{f}_i = 0$ if and only if $\sum_{i\in J_\ell} P(\tilde{f}_i) = 0$. The forward implication is clear. Now assume $\sum_{i\in J_\ell} P(\tilde{f}_i) = 0$. Then $\sum_{i\in J_\ell} \tilde{f}_i \in \ker(P)$. By the choice of P, we have that $\ker(P) \cap \big(\text{span}\{\sum_{i\in J_\ell} \tilde{f}_i\}\big) = \{0\}$ for all $J_\ell \in 2^I$. Therefore, $\sum_{i\in J_\ell} \tilde{f}_i \in \ker(P)$ if and only if $\sum_{i\in J_\ell} \tilde{f}_i = 0$. This proves the claim.

Now assume that $\mathbb{F}_F$ contains something other than the empty set. Then F has a tight subframe. Hence there exists some $J' \in 2^I$ such that $\sum_{i\in J'} \tilde{f}_i = 0$. This implies $\sum_{i\in J'} P(\tilde{f}_i) = 0$. Because $\tilde{f}_i \neq 0$ for each $i \in J'$, we know $P(\tilde{f}_i) \neq 0$. By assumption, range$(P) = \text{span}\{v\}$, where v is a unit vector. Then there exist nonzero scalars $\{\alpha_i\}_{i\in J'}$ such that $\alpha_i v = \tilde{f}_i$ for each $i \in J'$. Since $0 = \sum_{j\in J'} P(\tilde{f}_j) = \sum_{j\in J} \alpha_j v$ and $\alpha_j \neq 0$, we have $s, t \in J'$ such that $\text{sgn}(\alpha_s) = -\text{sgn}(\alpha_t)$. Since $\tilde{f}_s = \alpha_s v$ and $\tilde{f}_t = \alpha_t v$, we must have corresponding vectors in V that are nonzero and orthogonal. Since any two nonzero orthogonal vectors span $\mathbb{R}^2$, the vectors in V must span $\mathbb{R}^2$ and hence form a frame.

Suppose $J_\ell \in \mathbb{F}_F$. Then $\sum_{i\in J_\ell} \tilde{f}_i = 0$. From the claim, $\sum_{i\in J_\ell} \tilde{f}_i = 0$ if and only if $\sum_{i\in J_\ell} P(\tilde{f}_i) = 0$. Hence $J_\ell \in \mathbb{F}_V$. The reverse direction is similar, and thus $\mathbb{F}_F = \mathbb{F}_V$.

Since rank$(P) = 1$, there exists a unitary operator U such that $Uv = e_1$. Hence

$$UP(\tilde{f}_i) = \begin{bmatrix} \lambda_i \\ 0 \end{bmatrix}$$

for some $\lambda_i \in \mathbb{R}$. Define g as

$$g_i := \begin{cases} [\sqrt{\lambda_i} \;\; 0]^T & \lambda_i \geq 0, \\ [0 \;\; \sqrt{-\lambda_i}]^T & \lambda_i < 0. \end{cases}$$

Let $G = \{g_i\}_{i\in I}$. It easily follows that $UP(\tilde{f}_i) = \tilde{g}_i$. Moreover $\sum_{i\in J_\ell} P(\tilde{f}_i) = 0$ if and only if $\sum_{i\in J_\ell} UP(\tilde{f}_i) = 0$. Therefore $\mathbb{F}_G = \mathbb{F}_V = \mathbb{F}_F$. □

Remark 3.16. Based on the above Theorem 3.15, we propose the following inverse factor poset problem: given a poset $P \subseteq 2^I$, does there exist a frame $F \subseteq \mathbb{R}^n$ such that $\mathbb{F}_F = P$?

4. Dual frames

For a given frame F, we define the set $\mathcal{W}_F$ as

$$\mathcal{W}_F := \left\{ W = \begin{bmatrix} \leftarrow & w_1 & \rightarrow \\ & \vdots & \\ \leftarrow & w_n & \rightarrow \end{bmatrix} : \bar{w}_i \in \ker(\theta_F^*) \right\}.$$

Then, by the result in [Li 1995; Christensen et al. 2012], we have the following observation.

Observation 4.1. Let F be a frame for $\mathcal{H}^n$. Then any dual frame to a frame F can be expressed as columns of the matrix

$$\sigma_F^{-1}\theta_F^* + W \tag{2}$$

for some $W \in \mathcal{W}_F$.

For a given frame F, let $\mathcal{G} = \{\sigma_F^{-1}\theta_F^* + W : W \in \mathcal{W}_{\mathcal{F}}\}$ be the set of all matrices whose columns form duals of F. Define the operation $\oplus : \mathcal{G} \times \mathcal{G} \to \mathcal{G}$ by

$$(\sigma_F^{-1}\theta_F^* + W_1) \oplus (\sigma_F^{-1}\theta_F^* + W_2) := \sigma_F^{-1}\theta_F^* + W_1 + W_2.$$

Proposition 4.2. *Let F be a frame and let $\mathcal{G} = \{\sigma_F^{-1}\theta_F^* + W : W \in \mathcal{W}_{\mathcal{F}}\}$. Then $(\mathcal{G}, \oplus)$ defines an abelian group.*

Proof. For any $W_1, W_2 \in \mathcal{W}$, since $\ker(\theta_F^*)$ is a vector space, $W_1 + W_2 \in \mathcal{W}$, which implies that $\mathcal{G}$ is closed under $\oplus$. Associativity and commutativity follow from associativity and commutativity of matrix addition, and the identity is given by $\sigma_F^{-1}\theta_F^*$. Each element $\sigma_F^{-1}\theta_F^* + W \in \mathcal{G}$ has an inverse $\sigma_F^{-1}\theta_F^* - W$. □

Proposition 4.3. *Let F be a tight frame. Suppose that $G \in \{\sigma_F^{-1}\theta_F^* + W : W \in \mathcal{W}_{\mathcal{F}}\}$ is a matrix whose columns form a tight frame. Then the subgroup generated by G is contained in the set of matrices whose columns form tight duals of F.*

Proof. Let $\sigma_F = \lambda I_n$. Then $G = \frac{1}{\lambda}\theta_F^* + W$ for some $W \in \mathcal{W}_F$, where $WW^* = \alpha I_n$ for some $\alpha \in \mathbb{R}$. If $H = \frac{1}{\lambda}\theta_F^* + mW$ for some $m \in \mathbb{N}$, then $H^*H = \left(\frac{1}{\lambda} + m^2\alpha\right)I_n$, which completes the proof. □

The following example shows that in general, it is not true that the set of matrices in $\mathcal{G}$ whose columns form a tight frame is a subgroup of $(\mathcal{G}, \oplus)$.

Example 4.4. Let F be the frame where the synthesis operator θ_F^* is given by

$$\theta_F^* = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

Then the set of all matrices whose columns form a dual of F is

$$\mathcal{G} = \left\{ \begin{bmatrix} 1 & 0 & a & b \\ 0 & 1 & c & d \end{bmatrix} : a, b, c, d \in \mathbb{R} \right\}.$$

We consider the two matrices

$$A = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}.$$

Then $A, B \in \mathcal{G}$, and the columns of A and B form a tight dual of F. However, the columns of

$$A \oplus B = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix}$$

do not form a tight dual of F.

We study the relationship between the factor posets for a tight frame and its canonical dual. Recall that an isomorphism on posets $(P_1, \leq_1), (P_2, \leq_2)$ is a bijection $\phi : P_1 \to P_2$ so that $\phi(a) \leq_2 \phi(b)$ if and only if $a \leq_1 b$ for all $a, b \in P_1$. We define a stronger notion of order isomorphism. We let S_m denote the symmetric group on m elements.

Definition 4.5. We say that two factor posets $\mathbb{F}_F$ and $\mathbb{F}_G$ corresponding to frames $F = \{f_i\}_{i \in I}$ and $G = \{g_j\}_{j \in J}$ are *strongly isomorphic* if there exists some $m \in \mathbb{N}$ and some $\eta \in S_m$ such that $\eta(\mathbb{F}_F) = \eta(\mathbb{F}_G)$, where $\eta(\mathbb{F}_F) = \{\eta(J') : J' \in \mathbb{F}_F\}$ and $\eta(J') = \{\eta(j) : j \in J'\}$.

If F_J is a tight subframe of a λ-tight frame $F = \{f_i\}_{i \in I}$ for some $J \subseteq I$, then $\sum_{i \in J} \tilde{f}_i = 0$ so that we have

$$\sum_{i \in J} \sigma_F^{-1} \tilde{f}_i = \sum_{i \in J} \frac{1}{\lambda^2} \tilde{f}_i = 0.$$

This implies that $\{\sigma_F^{-1} f_i\}_{i \in J}$ is a tight frame. Thus we have the following result.

Proposition 4.6. *A tight frame F and its associated canonical dual frame $\{\sigma_F^{-1} f_i\}_{i=1}^k$ have the same factor posets.*

This result does not hold true for nontight frames and their canonical duals. For example, the factor poset of the following frame F and its canonical dual are not strongly isomorphic:

$$F = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} \frac{3989\sqrt{15912321}}{100} \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ \frac{3989}{10} \end{bmatrix} \right\}.$$

Proposition 4.7. *There exist a frame F such that no dual G of F has a factor poset structure that is strongly isomorphic to $\mathbb{F}_F$.*

Proof. Let $F = \{e_1, e_1, e_2\}$ be a frame for $\mathbb{R}^2$, where $\{e_1, e_2\}$ is the standard orthonormal basis of $\mathbb{R}^2$. Let G be an arbitrary dual of F; then by Observation 4.1,

$$\theta_G^* = \begin{bmatrix} \frac{1}{2} + a & \frac{1}{2} - a & 0 \\ b & -b & 1 \end{bmatrix}$$

for some $a, b \in \mathbb{R}$. We consider the poset $\mathbb{F}_F = \{\varnothing, \{1, 3\}, \{2, 3\}\}$. We know that two vectors in $\mathbb{R}^2$ form a tight frame if and only the vectors are orthogonal and are of equal norm. If F and G have strongly isomorphic poset structures, then two of the vectors of G must be orthogonal to the remaining vector in G, and all three vectors have equal norms. This is impossible. □

From the dual expression given in (2), we obtain a characterization of tight duals of a tight frame. The following result is remarked in [Krahmer et al. 2013]; the proof given here is different.

Theorem 4.8. *Let F be a λ-tight frame with k frame elements for $\mathcal{H}^n$. If $k < 2n$, then the canonical dual is the only tight dual of F. If $k \geq 2n$ then F has an alternate dual that is tight.*

Proof. Let G be a dual frame of F. Since $\theta_G^* = \frac{1}{\lambda}\theta_F^* + W$ for some $W \in \mathcal{W}$, we have that $\theta_G^*\theta_G = \frac{1}{\lambda}I_n + WW^*$. This implies that G is tight if and only if $WW^* = \alpha I_n$ for some $\alpha \in \mathbb{R}$. If $k < 2n$, since $\dim(\ker(\theta_F^*)) < n$, we have $\alpha = 0$. This implies that if $k < 2n$, then the canonical dual is the only tight dual of F. Let $k \geq 2n$ and let $\{\bar{w}_j\}_{j=1}^{k-n}$ be an orthonormal basis for $\ker(\theta^*)$. Then for any $\alpha \in \mathcal{H}\backslash\{0\}$, we consider

$$W = \alpha\begin{bmatrix} \leftarrow & w_1 & \rightarrow \\ & \vdots & \\ \leftarrow & w_n & \rightarrow \end{bmatrix}.$$

Since $W \in \mathcal{W}$, we have $\bigl(\frac{1}{\lambda}\theta_F^* + W\bigr)\bigl(\frac{1}{\lambda}\theta_F^* + W\bigr)^* = \bigl(\frac{1}{\lambda} + |\alpha|^2\bigr)I_n$, which implies that the columns of $\frac{1}{\lambda}\theta_F^* + W$ form a tight dual frame of F. □

Remark 4.9. We close this section with a simple and well-known construction for dual frames. Suppose $F = \{f_i\}_{i\in I}$ is a frame for $\mathcal{H}^n$ and $H = \{f_i\}_{j\in J}$ is a subframe of F. If $K = \{g_i\}_{i\in J}$ is a dual of H, then $G = \{g_i\}_{i\in I}$, where

$$g_i = \begin{cases} g_i & \text{if } i \in J, \\ 0 & \text{if } i \in I\backslash J \end{cases}$$

is a dual of F. Because any frame has a basis subframe, we have that if $F = \{f_i\}_{i=1}^k$ is a frame for $\mathcal{H}^n$, then there exists a dual of F consisting of n basis vectors for $\mathcal{H}^n$ and $(k-n)$ zero vectors. Likewise, if a frame in $\mathcal{H}^n$ has a tight subframe, then it has a tight dual.

5. ℓ^p-norm of the frame coefficients

It is well known that the ℓ^2-norm of the frame coefficients with respect to the canonical dual is smaller that the ℓ^2-norm of the frame coefficients with respect to any other dual. Moreover, this ℓ^2-minimization characterizes the canonical dual of a frame.

Proposition 5.1 [Han et al. 2007]. *Let $\{f_i\}_{i=1}^k$ be a frame for $\mathcal{H}^n$ and let $\{g_i\}_{i=1}^k$ be a dual frame of $\{f_i\}_{i=1}^k$. Then $\{g_i\}_{i=1}^k$ is the canonical dual if and only if*

$$\sum_{i=1}^{k} |\langle f, g_i\rangle|^2 \leq \sum_{i=1}^{k} |\langle f, h_i\rangle|^2 \quad \textit{for all } f \in \mathcal{H}^n$$

for all frames $\{h_i\}_{i=1}^k$ which are duals of $\{f_i\}_{i=1}^k$.

Using Newton's generalized binomial theorem and Hölder's inequality, for any two sequences $x = \{x_i\}_{i=1}^k$ and $y = \{y_i\}_{i=1}^k$ and $p \in (1, \infty)$, we have

$$\|x\|_p \leq \|x\|_1 \leq k^{1-1/p}\|x\|_p, \tag{3}$$

where $\|x\|_p = \left(\sum_{i=1}^k |x_i|^p\right)^{1/p}$. The right-side inequality in (3) with $p = 2$ and Proposition 5.1 gives us the following result:

Proposition 5.2. *Let $\{f_i\}_{i=1}^k$ be a frame for $\mathcal{H}^n$ and let $\{h_i\}_{i=1}^k$ be a dual frame of $\{f_i\}_{i=1}^k$. If $\{g_i\}_{i=1}^k$ is the canonical dual, then*

$$\sum_{i=1}^k |\langle f, g_i\rangle| \le \sqrt{k} \sum_{i=1}^k |\langle f, h_i\rangle| \quad \textit{for all } f \in \mathcal{H}^n.$$

From the inequalities (3) and Proposition 5.2, for any $p \in (1, \infty)$, we obtain

$$\sum_{i=1}^k |\langle f, g_i\rangle|^p \le k^{3p/2-1} \sum_{i=1}^k |\langle f, h_i\rangle|^p \quad \text{for all } f \in \mathcal{H}^n.$$

If $p > 2$, we have a better estimation.

Theorem 5.3. *Let $\{f_i\}_{i=1}^k$ be a frame for $\mathcal{H}^n$ and let $\{h_i\}_{i=1}^k$ be a dual frame of $\{f_i\}_{i=1}^k$. If $\{g_i\}_{i=1}^k$ is the canonical dual, then for any $p \in (2, \infty)$, we have*

$$\sum_{i=1}^k |\langle f, g_i\rangle|^p \le k^{p/2-1} \sum_{i=1}^k |\langle f, h_i\rangle|^p \quad \textit{for all } f \in \mathcal{H}^n.$$

Proof. First observe that the right-side inequality with $p/2$ implies that

$$\sum_{i=1}^k |\langle f, h_i\rangle|^p = \sum_{i=1}^k (|\langle f, h_i\rangle|^2)^{p/2} \ge k^{1-p/2} \left(\sum_{i=1}^k |\langle f, h_i\rangle|^2\right)^{p/2}.$$

By Proposition 5.1 and the left-side inequality with $p/2$, we have that

$$\sum_{i=1}^k |\langle f, h_i\rangle|^p \ge k^{1-p/2} \left(\sum_{i=1}^k |\langle f, g_i\rangle|^2\right)^{p/2} \ge k^{1-p/2} \sum_{i=1}^k |\langle f, g_i\rangle|^p. \qquad \square$$

Acknowledgment

This research work was done when K. Berry, E. Evert, and S. Nghiem participated in the Central Michigan University NSF-REU program in the summer of 2012.

References

[Christensen et al. 2012] O. Christensen, A. M. Powell, and X. C. Xiao, "A note on finite dual frame pairs", *Proc. Amer. Math. Soc.* **140**:11 (2012), 3921–3930. MR 2944732 Zbl 06204264

[Copenhaver et al. 2014] M. S. Copenhaver, Y. H. Kim, C. Logan, K. Mayfield, S. K. Narayan, M. J. Petro, and J. Sheperd, "Diagram vectors and tight frame scaling in finite dimensions", *Oper. Matrices* **8**:1 (2014), 73–88. MR 3202927 Zbl 06295119

[Daubechies 1992] I. Daubechies, *Ten lectures on wavelets*, CBMS-NSF Regional Conference Series in Applied Mathematics **61**, SIAM, Philadelphia, PA, 1992. MR 93e:42045 Zbl 0776.42018

[Duffin and Schaeffer 1952] R. J. Duffin and A. C. Schaeffer, "A class of nonharmonic Fourier series", *Trans. Amer. Math. Soc.* **72** (1952), 341–366. MR 13,839a Zbl 0049.32401

[Goyal et al. 1998] V. K. Goyal, J. Kovačević, and M. Vetterli, "Multiple description transform coding: Robustness to erasures using tight frame expansions", pp. 408 in *Proc. IEEE Int. Symp. Inform. Th.* (Cambridge, MA, 1998), IEEE, Piscataway, NJ, 1998.

[Goyal et al. 2001] V. K. Goyal, J. Kovačević, and J. A. Kelner, "Quantized frame expansions with erasures", *Appl. Comput. Harmon. Anal.* **10**:3 (2001), 203–233. MR 2002h:94012 Zbl 0992.94009

[Han et al. 2007] D. Han, K. Kornelson, D. Larson, and E. Weber, *Frames for undergraduates*, Student Mathematical Library **40**, Amer. Math. Soc., Providence, RI, 2007. MR 2010e:42044 Zbl 1143.42001

[Krahmer et al. 2013] F. Krahmer, G. Kutyniok, and J. Lemvig, "Sparsity and spectral properties of dual frames", *Linear Algebra Appl.* **439**:4 (2013), 982–998. MR 3061749 Zbl 1280.42026

[Lemvig et al. 2014] J. Lemvig, C. Miller, and K. A. Okoudjou, "Prime tight frames", *Adv. Comput. Math.* **40**:2 (2014), 315–334. MR 3194708 Zbl 1306.42050

[Li 1995] S. Li, "On general frame decompositions", *Numer. Funct. Anal. Optim.* **16**:9-10 (1995), 1181–1191. MR 97b:42055 Zbl 0849.42023

Received: 2014-09-26 Revised: 2015-02-28 Accepted: 2015-03-27

berry@math.utk.edu *Department of Mathematics, University of Tennessee, Knoxville, TN 37996, United States*

mcopen@mit.edu *Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA 02139, United States*

eevert@ucsd.edu *Department of Mathematics, University of California, San Diego, La Jolla, CA 92093, United States*

kim4y@cmich.edu *Department of Mathematics, Central Michigan University, Mount Pleasant, MI 48859, United States*

kling1te@cmich.edu *Department of Mathematics, Central Michigan University, Mount Pleasant, MI 48859, United States*

naray1sk@cmich.edu *Department of Mathematics, Central Michigan University, Mount Pleasant, MI 48859, United States*

sontnghiem@gmail.com *Berea College, Berea, KY 40403, United States*

INVOLVE 9:2 (2016)
dx.doi.org/10.2140/involve.2016.9.249

A variation on the game SET

David Clark, George Fisk and Nurullah Goren

(Communicated by Kenneth S. Berenhaut)

SET is a very popular card game with strong mathematical structure. In this paper, we describe “anti-SET”, a variation on SET in which we reverse the objective of the game by trying to avoid drawing “*sets*”. In anti-SET, two players take turns selecting cards from the SET deck into their hands. The first player to hold a *set* loses the game.

By examining the geometric structure behind SET, we determine a winning strategy for the first player. We extend this winning strategy to all nontrivial affine geometries over $\mathbb{F}_3$, of which SET is only one example. Thus we find a winning strategy for an infinite class of games and prove this winning strategy in geometric terms. We also describe a strategy for the second player which allows her to lengthen the game. This strategy demonstrates a connection between strategies in anti-SET and maximal caps in affine geometries.

1. Introduction

The card game SET is very popular among mathematics students. In addition to being an enjoyable pastime, it has a large amount of mathematical structure, including links to finite geometry, linear algebra, and combinatorics. This paper will focus on a particular variation of the game and the mathematics relevant to that variation. For much more information about the mathematics of SET, as well as positional games that are similar to the game in this paper, see [Davis and Maclagan 2003; Carroll and Dougherty 2004] and the citations contained therein.

SET consists of a deck of cards. Each card is printed with several figures which have four attributes: number, color, filling, and shape. For example, the card in Figure 1 would be described as “two green striped ovals”. The complete list of attributes is given in Table 1.

There are four attributes with three values each, and every possible combination appears exactly once. Thus there are $3^4 = 81$ cards in a complete SET deck.

The game requires players to find a *set*: three cards such that, for each attribute, all three cards are the same, or all three are different. Phrased differently, a *set*

MSC2010: primary 97A20, 51EXX; secondary 51E15, 51E22.
Keywords: SET (game), combinatorics, finite geometry, cap.

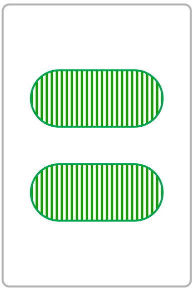


Figure 1. A SET card.

consists of three cards for which no attribute has two cards with one value, and another card with a different value. An example of a *set* is given in Figure 2: the number of the cards is all the same (1), but the colors are all different. The shading is all the same (solid), and the shapes are all different. A nonexample appears in Figure 3: two cards have solid shading, while the other is open. There are several other reasons why these cards are not a *set* as well. Note that, although the cards have three different colors, this *alone* is not enough to make a *set*.

Throughout this paper, we will use the notation SET to refer to the game, *set* to refer to a collection of cards as defined above, and "set" (without any special styling) to refer to the mathematical object consisting of an unordered collection of objects without repeated elements.

In the original game of SET, twelve cards are laid out at a time. Players compete to identify *sets* first, winning by collecting more *sets* than their opponents.

In this paper, we study a variation on SET which turns the usual goal upside down. Our game, *anti*-SET, is a two-player game played with a generalized SET deck in which each card has d different attributes (traditional SET has $d = 4$). This situation corresponds to a d-dimensional affine geometry over $\mathbb{F}_3$, which will be described later. The players, who we will call Xavier and Olivia, begin with the entire SET deck laid out in front of them. Xavier and Olivia then take turns selecting cards from these cards and take them into their hands. The first player to have a *set* in his or her hand *loses* the game. Thus, players are faced both with the challenge of trying

attribute	values		
number	1	2	3
color	red	green	purple
filling	open	striped	solid
shape	oval	diamond	squiggle

Table 1. Attributes of a SET card.

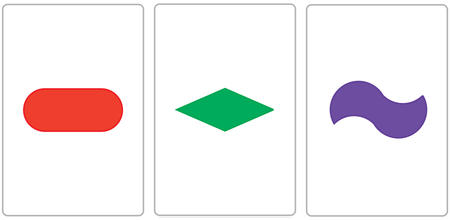


Figure 2. A *set*.

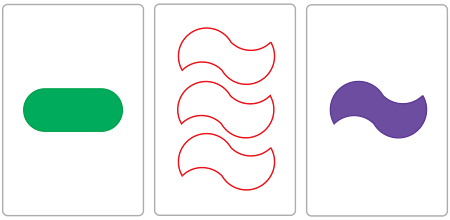


Figure 3. Not a *set*.

to avoid taking *sets* themselves but also trying to force the other player to take a *set*. As each player collects more cards in their hand, it becomes increasingly difficult to not take a *set*, as there are many more combinations of cards that can be made.

This game was inspired by a result of Pellegrino [1970]. Translated into the language of SET (which did not exist at the time of Pellegrino's writing), we have the result[1]:

Proposition 1 [Pellegrino 1970]. *Every set of* 21 SET *cards contains a* set.

Thus, anti-SET will always end once one player takes their 21st card, if not sooner. We initially created the game of anti-SET to explore the consequences of Pellegrino's result in the context of a game.

In the following sections, we will analyze this game, provide a winning strategy for the first player that applies to *all* nontrivial generalized SET decks (that is, nontrivial affine geometries over $\mathbb{F}_3$), and examine the maximum and minimum number of turns required to win. Along the way, we will demonstrate some unexpected links between Pellegrino's result and the losing player's strategy.

2. Example of gameplay

Before we give precise mathematical background for SET, we present an extended example of gameplay for anti-SET. For simplicity, we use a reduced version of

[1] We acknowledge that the *three* different uses of the word "set" in this result may make the reader's head spin.

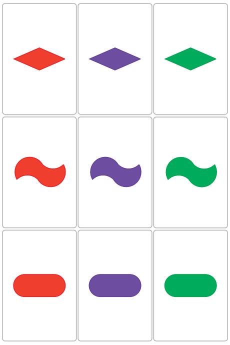


Figure 4. The nine-card reduced anti-SET deck.

anti-SET as played with the nine SET cards which are solid and have only one symbol per card. Later, we will justify this simplification geometrically and examine how it forms an important foundation for studying general anti-SET.

Let Xavier be the first player. He may choose any of the cards shown in Figure 4. We will mark Xavier's hand of cards with an "X" and Olivia's with an "O".

The moves are denoted as follows:

X_1**:** Xavier first arbitrarily chooses the red diamond.

O_1**:** Olivia, recognizing that every pair of cards determines a unique *set*, arbitrarily chooses the purple diamond.

The players hands at this point are represented in Figure 5(a).

X_2**:** Xavier chooses the green diamond, knowing that it is part of a *set* (the three cards in the top row) from which Olivia already owns one card. Thus, he avoids at least one *set*.

O_2**:** Olivia chooses the purple squiggle, again knowing that any pair of cards contains a *set* and thus all of her remaining choices are equally bad.

The players hands at this point are represented in Figure 5(b).

X_3**:** Xavier again chooses a card, the green oval, which he knows is part of at least one *set* which he cannot obtain.

O_3**:** Olivia chooses the red squiggle, leaving Xavier with at least one possibility (the green squiggle) which could complete a *set*.

The players hands at this point are represented in Figure 5(c).

X_4**:** Finally, Xavier chooses the red oval. This leaves Olivia with two options, both of which complete a *set*. Thus, Xavier will win. (See Figure 6.)

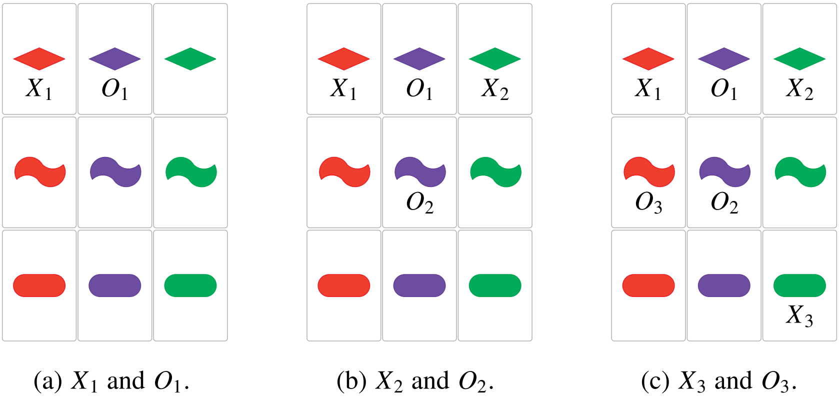


Figure 5. The first few steps of the anti-SET game.

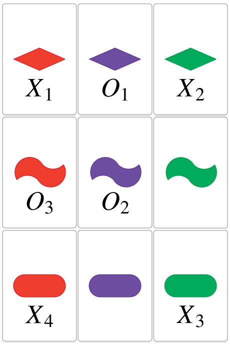


Figure 6. X_4 and Olivia's remaining options.

This nine-card example demonstrates the general flow of the game. In order to make valid conclusions about the game on a larger scale, we first need to describe the game mathematically, which we will do in the next section.

We also note that the board and style of play is similar to a backwards tic-tac-toe game, with players trying to avoid getting three in a row. Indeed, SET as played with the nine cards in this example can be thought of as playing tic-tac-toe on a torus, a concept which is explored in depth in [Carroll and Dougherty 2004]. Our names "Xavier" and "Olivia", and the idea of marking their cards with Xs and Os, were inspired by this interpretation.

3. Background

In this section, we will define the notation and concepts which will be used throughout the rest of the paper. Let X_n be the n-th card Xavier picks, and let

entry	0	1	2
n (number)	3	1	2
c (color)	red	green	purple
f (filling)	open	stripe	solid
s (shape)	squiggle	oval	diamond

Table 2. Correspondence between vectors (n, c, f, s) in $\mathbb{F}_3^4$ and characteristics of SET cards.

$\mathcal{X}(n) = \{X_1, X_2, \dots, X_n\}$ denote the collection of Xavier's first n cards. Similarly, let O_n denote the n-th card Olivia picks, and let $\mathcal{O}(n) = \{O_1, O_2, \dots, O_n\}$ denote the collection of Olivia's first n cards. Note that $\mathcal{X}(n) \subset \mathcal{X}(n+1)$ and $\mathcal{O}(n) \subset \mathcal{O}(n+1)$.

Xavier is the first player. The game proceeds with all cards in a SET deck available to both players. The players alternately take cards into their hands in the order $X_1, O_1, X_2, O_2, \dots$ until either $\mathcal{X}(n)$ or $\mathcal{O}(n)$ contains a *set*. The corresponding player loses on his or her n-th turn. We call a pair of choices (X_n, O_n) a *round* of anti-SET.

The mathematical structure of SET is an example of an *affine geometry*. For our purposes, we will define affine geometries from a coordinatized (vector-based) perspective, as described in [Beth et al. 1986; Dembowski 1997]. It is possible to do this from a purely axiomatic viewpoint as well (see [Dembowski 1997]). For more details about affine geometry in the context of SET, see [Carroll and Dougherty 2004].

The affine geometry $AG(d, q)$ is an incidence structure whose points are d-dimensional vectors with entries in $\mathbb{F}_q$. That is, the points are the elements of $\mathbb{F}_q^d$. The k-dimensional subspaces of $AG(d, q)$, referred to as k-flats, are the k-dimensional linear subspaces of $\mathbb{F}_q^d$ together with their cosets. We note that for a given k-dimensional linear subspace L, the coset of L by the vector $\vec{h} \in \mathbb{F}_q^d$ is defined as $L + \vec{h} = \{\vec{x} + \vec{h} : \vec{x} \in L\}$.

The cards of SET correspond to the points of $AG(4, 3)$, and the *sets* are the 1-flats (usually called *lines*). More specifically, the points are all vectors of the form (x_1, x_2, x_3, x_4), where $x_i \in \{0, 1, 2\}$, with all arithmetic done modulo 3. The 1-flats correspond to the 1-dimensional subspaces of $\mathbb{F}_3^4$ and their cosets. Each such 1-flat contains three points, corresponding to the three cards in a *set*.

To give a more concrete interpretation of SET in this context, we note that each card corresponds to a unique vector, with each coordinate corresponding to a characteristic of the cards. We arbitrarily identify the coordinates with characteristics of the SET cards as shown in Table 2. There are many equivalent ways to map between attributes of SET cards and the entries of $\mathbb{F}_3$.

Example 2. Consider the *set* in Figure 7. Using the correspondence from Table 2, these three cards, in order, form the vectors $(0, 0, 0, 0)$, $(2, 1, 1, 1)$, and $(1, 2, 2, 2)$.

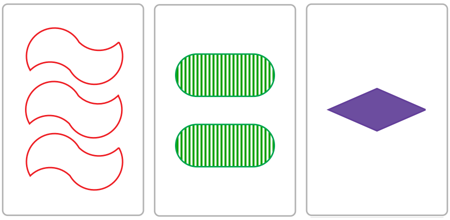


Figure 7. A *set*.

We will make extensive use of the following result about affine geometries:

Proposition 3 (affine collinearity rule). *Three points $\vec{a}$, $\vec{b}$ and $\vec{c}$ in $AG(d,3)$ form a line if and only if $\vec{a}+\vec{b}+\vec{c}=\vec{0}$.*

Proof. A set ℓ in $AG(d,3)$ is a line if and only if ℓ is a 1-dimensional subspace of $\mathbb{F}_3^d$ or a coset thereof. Thus ℓ is a line if and only if there exist a nonzero vector $\vec{x}$ and a vector $\vec{h}$, both in $\mathbb{F}_3^d$, such that $\ell=\{\vec{h}, \vec{x}+\vec{h}, 2\vec{x}+\vec{h}\}$. (Note that $\vec{h}=\vec{0}$ is possible.) In particular, all lines in $AG(d,3)$ contain three points. Then the sum of the elements in ℓ is $3\vec{h}+3\vec{x}=\vec{0}$, since we are working in $\mathbb{F}_3$.

Conversely, suppose $\ell=\{\vec{a}, \vec{b}, \vec{c}\}$ such that $\vec{a}+\vec{b}+\vec{c}=\vec{0}$. Then

$$\vec{0}+(\vec{b}-\vec{a})+(\vec{c}-\vec{a})=\vec{0}-3\vec{a}=\vec{0}$$

as well. Thus $\vec{c}-\vec{a}=2(\vec{b}-\vec{a})$, and so $m=\{\vec{0}, \vec{b}-\vec{a}, 2(\vec{b}-\vec{a})\}$ is a linear subspace of $\mathbb{F}_3^d$. Thus $\ell=m+\vec{a}$ is a line. □

In the context of SET, three cards $\{A, B, C\}$ form a *set* if and only if their corresponding vectors $\vec{a}$, $\vec{b}$ and $\vec{c}$ (respectively) satisfy $\vec{a}+\vec{b}+\vec{c}=\vec{0}$. To see this, consider three vectors whose associated cards form a *set*. The collection of three values in a given coordinate is limited to the following possibilities: $\{0, 0, 0\}$, $\{1, 1, 1\}$, $\{2, 2, 2\}$, or $\{0, 1, 2\}$. These collections of values constitute all possibilities for “all the same” or “all different”. The sum of the numbers in each of these collections is 0 (mod 3). Furthermore, no other collection of three values sums to 0 (mod 3).

We will also use the following well-known proposition:

Proposition 4. *In $AG(d,q)$, every pair of points appears in exactly one line.*

This can be seen as follows: A line is a 1-dimensional subspace or a coset of such a subspace. Let x and y be distinct points in $AG(d,q)$. If $x=\alpha y$ for some $\alpha\in\mathbb{F}_q$, then x and y appear together only in the 1-dimensional linear subspace defined by x. If x and y are not scalar multiples, then 0 and $y-x$ appear together only in the 1-dimensional linear subspace ℓ defined by $y-x$, and therefore x and y appear together only in the coset $\ell+x$.

This corresponds to the well-known fact that every pair of SET cards is part of a unique *set*. Algebraically, given two points $\vec{a}$ and $\vec{b}$, there exists a unique vector $\vec{c} \in \mathbb{F}_q^d$ such that $\vec{a}+\vec{b}+\vec{c}=\vec{0}$.

Because affine lines include both linear subspaces of $\mathbb{F}_q^d$ and their cosets, affine geometries naturally include parallel lines. All cosets of a given line ℓ are parallel to ℓ, and together this collection of cosets partitions the points of the geometry. Thus, for any *set* $\{A, B, C\}$ in the traditional 81 card SET game, there are $81/3 = 27$ *sets* (including $\{A, B, C\}$ itself) which are parallel to the original *set*. These 27 *sets* contain all 81 cards in the SET deck.

Example 5. In Example 2, we saw a *set* consisting of the vectors

$$S = \{(0, 0, 0, 0), (2, 1, 1, 1), (1, 2, 2, 2)\}.$$

The coset $S + (1, 0, 1, 2)$ is

$$S + (1, 0, 1, 2) = \{(1, 0, 1, 2), (0, 1, 2, 0), (2, 2, 0, 1)\},$$

which can be verified to be a *set* sharing no points with S.

In addition to points and lines, affine geometries contain other substructures with geometric interpretation. Of particular interest to us is the *affine plane* $AG(2, q)$, which can be viewed as a 2-dimensional subspace of a larger affine geometry. Affine planes are very well studied. In the case of SET, the set of vectors obtained by fixing any two coordinates of the vectors in $\mathbb{F}_3^4$ is isomorphic to an affine plane. With only two coordinates "free" to change, a plane contains $3^2 = 9$ points.

Example 6. The nine cards in Figure 4 form an affine plane. Here, the coordinate corresponding to "number" is fixed at 1, and the coordinate corresponding to "filling" is fixed at 2 (solid). Thus there are two free coordinates, giving a 2-dimensional plane.

An affine plane is spanned by two nonparallel lines. In the affine plane $AG(2, 3)$, each line is part of a *parallel class* of three parallel lines.

Notice that there are twelve lines (that is, *sets*) in $AG(2, 3)$. As represented in Example 6, there are three horizontal lines, three vertical lines, and then three lines in each diagonal direction. (For example, the *set* containing the red squiggle, purple diamond, and green oval is one of these diagonal *sets*.)

The last substructure of special interest to us is a *hyperplane*, a $(d-1)$-dimensional subspace within $AG(d, q)$. Equivalently, a hyperplane is a subspace of maximal size, or of codimension 1. In SET, a hyperplane corresponds to a set of 27 cards with a single attribute fixed.

The remainder of this paper will primarily use geometric language when discussing SET. In particular, we will use "point" and "line" to refer to cards and *sets*, respectively, except when interpreting our results in terms of the original SET game.

We note that the results of this paper apply to $AG(d,3)$ for all $d \geq 2$. That is, they apply not only to SET (which lives in $AG(d,3)$) but also to "general" SET as played in $AG(d,3)$. For example, a version of SET could be created in which each card has five attributes: the four usual ones, plus a scratch-and-sniff scent attribute with three different values. The game of anti-SET could be played with these $3^5 = 243$ cards without change. When $d = 1$, the geometry $AG(1,3)$ consists of a single line, in which no win or loss of anti-SET is possible.

4. Results

In this section, we prove that Xavier has a winning strategy in anti-SET, as played on any affine geometry $AG(d,3)$, $d \geq 2$. We first reformulate anti-SET in purely geometric terms:

Anti-SET is a two-player game played on $AG(d,3)$. The players, Xavier and Olivia, take turns (beginning with Xavier) selecting points from the geometry. The first player to have a line contained entirely in his or her hand *loses* the game.

Theorem 7 (winning strategy). *Suppose Xavier and Olivia play anti-*SET *using the points in $AG(d,3)$, $d \geq 2$. Moves X_1 and O_1 may be chosen arbitrarily. After those moves, Xavier will always win by following this strategy: for each move $n \geq 2$, Xavier chooses X_n to be the unique third point on the unique line containing X_1 and O_{n-1}.*

Xavier's strategy depends on him following Olivia's moves. The first two moves are arbitrary, after which Xavier begins to follow Olivia by completing lines which are not completely contained in either player's hands. The worked example in Section 2 implements exactly this strategy on a nine-card affine plane.

We note that we require $d \geq 2$ only because $d = 1$ is a degenerate case: $AG(1,3)$ consists of three points on a single line. Thus, every game ends in a tie, as neither player can fully collect the line. However, the condition that $q = 3$ is essential. Our strategy is highly dependent on the fact that each line contains exactly three points, a fact that is lost for $q \neq 3$.

The following lemmas are necessary to establish the correctness of this strategy.

Lemma 8 (Xavier can play). *If Xavier consistently follows the strategy in Theorem 7, then Xavier can always choose the required point.*

Proof. Consider the n-th round of the game. In the previous round, Olivia selected point O_{n-1}, and now Xavier wishes to choose as X_n the unique point C completing the line ℓ containing points $\{X_1, O_{n-1}\}$. Note that point C exists and is unique by Proposition 4. If C is unavailable, it must be in either $\mathcal{O}(n-2)$ (because Olivia's move O_{n-1} was not C) or $\mathcal{X}(n-1)$.

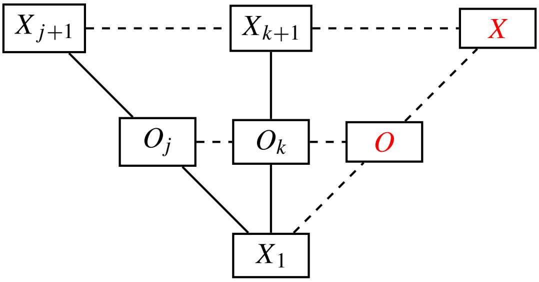


Figure 8. Diagram for proof of Lemma 9.

If Olivia previously chose C, then by following the strategy Xavier would have immediately chosen the other point on ℓ. If Xavier previously chose C, then he must have done so immediately after Olivia chose the other point on ℓ.

In either case, *all* points on ℓ appear in $\mathcal{O}(n-2) \cup \mathcal{X}(n-1)$, and thus it was impossible for Olivia to choose any point on ℓ as O_{n-1}. Therefore we obtain a contradiction, and C must be available for Xavier to choose. □

Lemma 9 (Xavier cannot lose). *If Xavier consistently follows the strategy in Theorem 7, then Xavier cannot lose.*

Proof. Without loss of generality, assume that at least two rounds have occurred. In round $j \geq 1$, Olivia chooses O_j. In round $k > j$, Olivia chooses O_k. Following the winning strategy, Xavier chooses X_{j+1} and X_{k+1}, respectively. Thus $\{X_1, O_j, X_{j+1}\}$ and $\{X_1, O_k, X_{k+1}\}$ are lines. This is represented geometrically by solid lines connecting the points in Figure 8. Algebraically, $X_1+O_j+X_{j+1} = \vec{0}$ and $X_1+O_k+X_{k+1} = \vec{0}$.

Suppose that at some future round m, while following the winning strategy, Xavier chooses point X which completes a line $\{X_{j+1}, X_{k+1}, X\} \subseteq \mathcal{X}(m)$, causing him to lose. Thus $X_{j+1} + X_{k+1} + X = \vec{0}$.

Xavier chose X in response to some move O by Olivia. Thus X is the unique third point on the line containing $\{X_1, O\}$. Therefore $X_1 + O + X = \vec{0}$. Beginning with this fact and applying algebra, we have

$$\begin{aligned}
0 &= X_1+O+X \\
&= X_1+O+(-X_{j+1}-X_{k+1}) && (\{X_{j+1}, X_{k+1}, X\} \text{ is a line}) \\
&= X_1+O+(X_1+O_j)+(X_1+O_k) && (\{X_1, O_j, X_{j+1}\}, \{X_1, O_k, X_{k+1}\} \text{ are lines}) \\
&= 3X_1+O+O_j+O_k \\
&= O+O_j+O_k && (3X_1 \equiv 0 \pmod 3)).
\end{aligned}$$

Therefore $\{O, O_j, O_k\}$ is a line. Because Olivia chose O before Xavier was forced to chose X, Olivia would have immediately lost with a line in $\mathcal{O}(m-1)$.

Thus, it is impossible for Xavier to have a line in $\mathcal{X}(m)$, since Olivia would immediately lose before he could choose to complete such a line. Therefore, Xavier cannot lose when following the strategy. □

Lemma 10 (there are no ties). *If Xavier consistently follows the strategy in Theorem 7, then the game cannot end in a tie.*

Proof. We first note that there are no ties in any nine-card plane $AG(2, 3)$. That is, it is impossible to partition the nine points into two sets, neither of which contains a line. In particular, any set of at least five points from a nine-card plane must contain a line. This may be demonstrated by brute force, or with an elegant counting argument such as that in [Carroll and Dougherty 2004].

After Xavier's third turn, the set of points selected is $S = \{X_1, O_1, X_2, O_2, X_3\}$. Note that, by following the strategy, S contains two nonparallel lines: $\{X_1, O_1, X_2\}$ and $\{X_1, O_2, X_3\}$. These two lines span an affine plane P.

The game may proceed in two ways:

(1) Olivia may choose to only select points in P. There are no ties in P and by Lemma 9, Xavier cannot lose. Thus Olivia must eventually lose.

(2) Olivia may choose to select some point outside of P. If Olivia does not lose, she will eventually run out of points outside of P, and therefore must choose a point from within P. As argued above, Olivia must then lose. Note that the points in P remain available for Olivia to choose, because Xavier will only choose a point in P if Olivia also chooses a point in P. This is because no line of $AG(d, 3)$ contains two points in a plane and one point outside of a plane.

Either way, Olivia loses. □

Together, these lemmas provide a proof of Theorem 7:

Proof of Theorem 7. By Lemma 8, Xavier can follow the strategy. By Lemma 9, Xavier can never lose when following the strategy. Finally, by Lemma 10, the game cannot end in a tie. Therefore, Xavier (the first player) wins. □

5. Length of the game

Now that we know that Xavier will always win, a reasonable question is "how many moves are required for Xavier to win?" Without assuming rational play, a game could be as short as three rounds: Olivia could choose three cards which form a *set* and lose after move O_3. But assuming rational play, Olivia can survive much longer.

In this section, we seek to answer the question "how long can Olivia force the game to continue?" Because there is some room for ambiguity, we provide the following precise definition:

d	1	2	3	4	5	6
$m_2(AG(d, 3))$	2	4	9	20	45	112

Table 3. Sizes of maximal caps for some small affine geometries.

Definition 11. The *length* of a game of anti-SET is the lowest index n such that $\mathcal{O}(n)$ contains a line.

Thus, for example, the game played in Section 2 has length 4. Note that, because Olivia plays second, length can be interpreted as the number of complete rounds played before the game ends.

We also require the concept of a *cap*:

Definition 12. A *cap* in $AG(d, q)$ is a set of points which contains no lines. A *maximal cap* is a cap with the largest possible size for a given set of parameters d, q, and its size is denoted $m_2(AG(d, q))$.

Example 13. Every set of five points in $AG(2, 3)$ contains a line. Consider the result of the sample game from Section 2, shown in Figure 6. The set of four points marked X contains no line and therefore forms a maximal cap in $AG(2, 3)$. The three marked O form a cap which is not maximal. Thus $m_2(AG(2, 3)) = 4$.

A long-standing question in finite geometry is to determine the size of a maximal cap. While a variety of bounds are known, no exact formula is known in general. For $q = 3$, some currently known values for $m_2(AG(d, 3))$ are summarized in Table 3. For more information, see [Potechin 2008] and references therein.

In the language of affine geometry, Proposition 1 can be restated:

Proposition 1 [Pellegrino 1970]. *In $AG(4, 3)$, we have $m_2(AG(4, 3)) = 20$.*

In other words, every set of 21 SET cards must contain a *set*.

Theorem 15. *The maximum possible length of a game of anti-*SET *played on $AG(d, 3)$ is $m_2(AG(d, 3))$.*

Proof. Let $m = m_2(AG(d, 3))$. Xavier is always the first to have k points in hand for any k, and thus $\mathcal{X}(m+1)$ (if the game lasts so long) must contain a *set*. However, by Theorem 7, Xavier cannot lose. Thus, Olivia's previous move, O_m, must have ended with $\mathcal{O}(m)$ containing a *set*. Thus the length of the game is at most m. □

As a corollary, the length of anti-SET played on $AG(4, 3)$ is at most 20. Computational simulations for small d suggest that Olivia can always achieve this bound, but we are unable to prove this.

Next, we determine a lower bound on the length of the game. We do this by demonstrating a strategy for Olivia which guarantees the game to last for a certain number of moves.

Lemma 16. *Let S_1, S_2, S_3 be three parallel hyperplanes in $AG(d, 3)$. Then any line which intersects S_1 and S_2 also intersects S_3.*

Proof. This is a direct consequence of the structure of the underlying vector space. Note that S_1, S_2, S_3 partition the points of $AG(d, 3)$, and also note that a line $\ell = \{x, y, z\}$ contains exactly three points. Because ℓ and each S_i are cosets of a linear subspace, $S_i \cap \ell$ must be a linear subspace (or coset) as well. In $\mathbb{F}_3^d$, each such subspace contains 3^k points for some $k \geq 0$. Thus ℓ must intersect each hyperplane in zero, one, or three points. If ℓ intersects both S_1 and S_2 in at least one point, then ℓ cannot intersect either in all three points. Thus ℓ intersects each of S_1 and S_2 in exactly one point, and so its third point must be in the remaining point set, S_3. □

Theorem 17. *Suppose Xavier and Olivia play anti-*SET *on $AG(d, 3)$, $d \geq 1$. Then Olivia can force the game to have length at least $2 + \sum_{i=1}^{d-1} m_2(AG(i, 3))$.*

Proof. We proceed by induction. As a basis, consider anti-SET played on $AG(2, 3)$. This is a nine-point plane. We saw in Section 2 that Olivia may extend the game to four rounds simply by not choosing her third point to be on the line defined by the first two. Furthermore, $2 + m_2(AG(1, 3)) = 4$ since a cap in $AG(1, 3)$ consists of any two points on the only line. (Recall that the fourth round ends with Olivia choosing her fourth card, which must complete a line in $\mathcal{O}(4)$.)

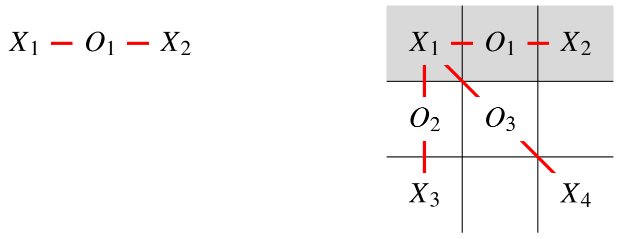
Assume, for anti-SET played in $AG(d-1, 3)$, that Olivia has a strategy which makes the length of the game $2 + \sum_{i=1}^{d-2} m_2(AG(i, 3))$. Then she can play on $AG(d-1, 3)$ for $1 + \sum_{i=1}^{d-2} m_2(AG(i, 3))$ rounds *without* losing. Let S_1 be a copy of $AG(d-1, 3)$ embedded as a hyperplane in $AG(d, 3)$, and let S_1, S_2, S_3 be the three hyperplanes parallel to S_1 in $AG(d, 3)$. Olivia proceeds as follows:

(1) Inductively, Olivia plays for $1 + \sum_{i=1}^{d-2} m_2(AG(i, 3))$ moves entirely in S_1 without losing. Note that Xavier's moves also fall entirely in S_1.

(2) Olivia then chooses the $m_2(AG(d-1, 3))$ points of a maximal cap entirely in S_2. Note that Olivia is free to choose these points, because Xavier's moves must now fall entirely in S_3 by Lemma 16.

This strategy describes Olivia's moves for $n = 1 + \sum_{i=1}^{d-1} m_2(AG(i, 3))$ rounds. Olivia never completes a line in $\mathcal{O}(n)$ by following this strategy. By our inductive assumption, no line exists within the subset of her moves falling in S_1. Because Olivia chooses the points of a cap in S_2, no line exists within her points in S_2. Finally, no line in $\mathcal{O}(n)$ can exist with one point in S_1 and another in S_2: By Lemma 16, the third point of such a line would be in S_3, but Olivia chooses no points in S_3.

Thus, Olivia does not lose by following the above strategy, and therefore Olivia can play for at least $2 + \sum_{i=1}^{n-1} m_2(AG(i, 3))$ rounds. □



(a) Play in $AG(1, 3)$. (b) Play expanded to $AG(2, 3)$.

Figure 9. Visualization of Olivia's strategy from Theorem 17.

Intuitively, Olivia's strategy works as follows: Olivia "fills up" a line with a cap, jumping up to a plane which she also fills with a cap. She continues jumping up to the next structure until she eventually runs out of room.

Example 18. Theorem 17 is demonstrated in Figures 9 and 10. In Figure 9(a), play begins on a line (that is, $AG(1, 3)$). In Figure 9(b), the line expands to a full plane (that is, $AG(2, 3)$). Note that Olivia's play occurs only in the second row, which is one coset of the original line. Her two plays form a cap on this line. Similarly, Xavier's plays all occur in the third row, another coset of the original line.

In Figure 10, play expands to cover the cosets of the plane. Figure 10 shows the original plane from Figure 9, now considered to be a subspace S_1. The other two planes in this figure are the cosets S_2 and S_3 of S_1. Note that Olivia plays only in coset S_2, and that her plays form a cap in $AG(2, 3)$. Xavier's plays are forced into coset S_3.

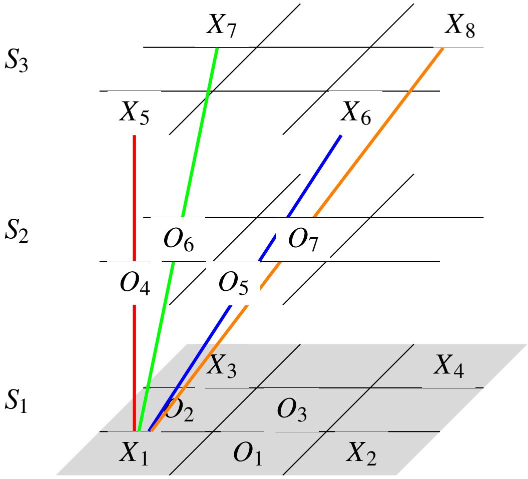


Figure 10. Continued visualization of Olivia's strategy from Theorem 17.

6. Open problems

Naturally, a variation on SET, such as anti-SET, leaves many open questions. The game SET has been widely studied, and several of the open problems below are based on generalizations and extensions already proposed for SET.

A more general category of games to which anti-SET belongs could be named "anti-tic-tac-toe on a design." A t-(v, k, λ) *design* (or t-design) is a set of v points $\mathcal{P}$ together with a collection $\mathcal{B}$ of k-subsets of the points, called blocks, such that every t-subset of $\mathcal{P}$ appears in exactly λ blocks. The points and lines of $AG(d, 3)$ form an example of an *affine geometry design*. See [Beth et al. 1986] for more details. To play anti-tic-tac-toe on a given t-(v, k, λ) design, two players alternate selecting points of the design. The first player to select all points in any block of the design loses. Thus the general question is "is there a winning strategy for anti-tic-tac-toe on a design?"

Specific instances of this general game will likely prove to be more tractable. For example:

- Play on nonternary affine geometries, which are also examples of affine geometry designs. The winning strategy described in this paper depends heavily on working in $AG(d, 3)$. Is it possible to have $q > 3$? The largest difference is that lines now have more than three points, opening the possibility that Olivia plays on a point which completes a line, leaving Xavier unable to "follow" Olivia.
- Play on a projective geometry. It is possible to play anti-SET on a projective geometry $PG(d, q)$? The set of points and k-dimensional subspaces of $PG(d, q)$ form a *projective geometry design*. (For information about "projective SET", see [Davis and Maclagan 2003].)
- Play on Steiner triple systems. This is a name given to the category of 2-$(v, 3, 1)$ designs. In this category, every pair of points determines a unique line, and every line has three points. This includes two key geometric features that figures in the strategy for anti-SET.

Other open problems involve changing the parameters of play for anti-SET:

- Play with three or more players. This must considerably change the strategy. Under the winning strategy described here, it would be possible for one player to "block" another player's necessary move.
- Recovering from an error. If Xavier does not follow the winning strategy, when is it possible for Olivia to win? Is it possible for Xavier to recover from this error, and if so, under what conditions?

Finally, we believe that Theorem 17 can be improved:

- Determine a strategy for Olivia which always forces a game length of $m_2(AG(d, 3))$ rounds, thus improving on Theorem 17.

Acknowledgments

The authors would like to thank the Mathematics Center for Educational Programs (MathCEP) at the University of Minnesota for supporting this research. This work was completed as part of the University of Minnesota Talented Youth in Mathematics (UMTYMP) undergraduate research program in summer 2013.

We would also like to thank the anonymous referee for very helpful comments and suggestions.

References

[Beth et al. 1986] T. Beth, D. Jungnickel, and H. Lenz, *Design theory*, Cambridge University Press, 1986. MR 88b:05021 Zbl 0602.05001

[Carroll and Dougherty 2004] M. T. Carroll and S. T. Dougherty, "Tic-tac-toe on a finite plane", *Math. Mag.* **77**:4 (2004), 260–274. MR 2087313 Zbl 1213.05023

[Davis and Maclagan 2003] B. L. Davis and D. Maclagan, "The card game SET", *Math. Intelligencer* **25**:3 (2003), 33–40. MR 2004i:91042 Zbl 1109.91013

[Dembowski 1997] P. Dembowski, *Finite geometries*, Classics in Mathematics **44**, Springer, Berlin, 1997. MR 97i:51005 Zbl 0865.51004

[Pellegrino 1970] G. Pellegrino, "Sul massimo ordine delle calotte in $S_{4,3}$", *Matematiche* (*Catania*) **25** (1970), 149–157. MR 51 #207

[Potechin 2008] A. Potechin, "Maximal caps in AG(6, 3)", *Des. Codes Cryptogr.* **46**:3 (2008), 243–259. MR 2008m:51030 Zbl 1187.51010

Received: 2014-10-13 Revised: 2015-01-29 Accepted: 2015-02-06

clarkdav@gvsu.edu *Department of Mathematics, Grand Valley State University,*
1 Campus Drive, Allendale, MI 49401, United States

geomfisk@gmail.com *Department of Mathematics, University of Minnesota,*
Minneapolis, MN 55455, United States

nurrygoren@gmail.com *Department of Mathematics, Pomona College,*
Claremont, CA 91711, United States

INVOLVE 9:2 (2016)
dx.doi.org/10.2140/involve.2016.9.265

The kernel of the matrix $[ij \pmod n]$ when n is prime

Maria I. Bueno, Susana Furtado, Jennifer Karkoska,
Kyanne Mayfield, Robert Samalis and Adam Telatovich

(Communicated by Kenneth S. Berenhaut)

In this paper, we consider the $n\times n$ matrix whose (i, j)-th entry is $ij \pmod n$ and compute its rank and a basis for its kernel (viewed as a matrix over the real numbers) when n is prime. We also give a conjecture on the rank of this matrix when n is not prime and give a set of vectors in its kernel, which is a basis if the conjecture is true. Finally, we include an application of this problem to number theory.

1. Introduction

When learning modular arithmetic, it is a natural exercise to consider the multiplication table modulo an integer n. This table can be seen as an $n \times n$ matrix whose entries are positive integers. A question in linear algebra, which is interesting by itself, is to determine the rank or, even better, a basis for the kernel, of this matrix over the real numbers.

In this paper, we denote by C_n the $n \times n$ matrix given by

$$C_n(i, j) = ij \pmod n, \quad i, j = 1, \dots, n, \tag{1}$$

where $C_n(i, j)$ denotes the (i, j)-th entry of C_n.

Using techniques from matrix analysis and analytic number theory, we find the rank and a basis for the kernel of C_n when n is prime. When n is composite, we give a conjecture on the rank of C_n and a set of vectors in the kernel of C_n that is a basis of the kernel if the conjecture is true.

Since the last row and column of C_n are both zero, the matrix H_n, obtained from C_n by deleting that row and that column, has the same rank as C_n. Moreover,

MSC2010: 15A03, 11M06, 11M20.

Keywords: rank of a matrix, kernel of a matrix, bisymmetric matrix.

Furtado completed this work within the activities of Centro de Estruturas Lineares e Combinatorias da Universidade de Lisboa. Bueno, Karkoska, Mayfield, Samalis, and Telatovich were supported by NSF grant DMS-0852065 for the REU program at UCSB.

it is easy to find a basis for the kernel of C_n from the kernel of H_n. Therefore, most of the paper will be focused on studying the kernel of H_n.

As an example, for $n = 5$, we have

$$H_5 = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 1 & 3 \\ 3 & 1 & 4 & 2 \\ 4 & 3 & 2 & 1 \end{bmatrix}.$$

The paper is organized as follows. In Section 2, we use a matrix theory approach to study the $(n-1) \times (n-1)$ matrix H_n. In particular, we give a block-diagonal matrix similar to H_n (Lemma 7) and use it to give a set of vectors in the kernel of H_n. This result allows us to obtain nontrivial lower and upper bounds for the rank of H_n for general n (Corollary 12). A conjecture for the exact value of this rank is also presented (Conjecture 16). In Section 3, we obtain the main result of the paper (Theorem 42) which describes the rank of the $n \times n$ matrix C_n when n is prime and gives a basis for its kernel. The proof of the rank result is done using techniques from character theory and analytic number theory. In Section 4, we present an application to number theory that motivated our work.

2. The kernel of the matrix H_n

We now present some properties of the matrix H_n for general n and use them to study the kernel of H_n. We first introduce some notation and recall some definitions.

We denote by $M_{n,m}$ the set of $n \times m$ matrices with entries in $\mathbb{R}$. We abbreviate $M_{n,n}$ to M_n.

We denote by R the exchange matrix (also called the flip-transpose of the identity matrix I) of appropriate size, that is,

$$R := \begin{bmatrix} 0 & \dots & 1 \\ \vdots & \iddots & \vdots \\ 1 & \dots & 0 \end{bmatrix}.$$

Note that $R^2 = I$.

Definition 1. Let $A \in M_n$.

- The matrix A is called *symmetric* if $A = A^T$.
- The matrix A is called *persymmetric* if $A = RA^T R$.
- The matrix A is called *centrosymmetric* if $A = RAR$.
- The matrix A is called *bisymmetric* (or *symmetric centrosymmetric*, or *doubly symmetric*) if it is symmetric and centrosymmetric.

Remark 2. If $A \in M_n$ is persymmetric, then RA is symmetric. Also, if A is symmetric and centrosymmetric (resp. persymmetric), then A is persymmetric (resp. centrosymmetric).

Note that $A \in M_n$ is bisymmetric if

$$A(i,j) = A(j,i) \quad \text{and} \quad A(i,j) = A(n+1-i, n+1-j), \quad i,j = 1,\dots,n.$$

This means that being bisymmetric is equivalent to being symmetric with respect to the main diagonal and being symmetric with respect to the antidiagonal. A look at H_5 shows that this matrix is bisymmetric.

Lemma 3. *Let $n \in \mathbb{N}$. The matrix $H_n \in M_{n-1}$ is bisymmetric.*

Proof. The matrix H_n is symmetric since $H_n(i,j) = ij \pmod n = H_n(j,i)$. Additionally, H_n is centrosymmetric since

$$(n-i)(n-j) \pmod n = ij \pmod n,$$

which implies that $H_n(i,j) = H_n(n-i, n-j)$. □

The following result follows from some well-known properties of bisymmetric matrices [Cantoni and Butler 1976, Lemma 2].

Lemma 4. *If n is odd, then H_n has the form*

$$H_n = \begin{bmatrix} A & RBR \\ B & RAR \end{bmatrix} \tag{2}$$

for some symmetric $A \in M_{(n-1)/2}$ and persymmetric $B \in M_{(n-1)/2}$.

If n is even, then H_n has the form

$$H_n = \begin{bmatrix} A & x & RBR \\ x^T & q & x^T R \\ B & Rx & RAR \end{bmatrix} \tag{3}$$

for some $q \in \mathbb{C}$, $x \in M_{(n-2)/2,1}$, symmetric $A \in M_{(n-2)/2}$ and persymmetric $B \in M_{(n-2)/2}$.

Next we give an explicit expression for the number q and the vector x in the block representation of H_n given in Lemma 4 when n is even.

Lemma 5. *If n is even, then the number q in (3) is given by*

$$\begin{cases} 0 & \text{if } n \equiv 0 \pmod 4, \\ \dfrac{n}{2} & \text{if } n \not\equiv 0 \pmod 4. \end{cases}$$

Proof. We have

$$q = H_n\left(\frac{n}{2}, \frac{n}{2}\right) = \frac{n}{2} \cdot \frac{n}{2} \pmod n.$$

If $n \equiv 0 \pmod 4$, then $n = 4k$ for some positive integer k. Thus,

$$\frac{n}{2} \cdot \frac{n}{2} \pmod n = kn \pmod n = 0.$$

If $n \not\equiv 0 \pmod 4$, then, since n is even, we know that $n = 4k + 2$ for some positive integer k, and

$$\frac{n}{2} \cdot \frac{n}{2} \pmod n = kn + 2k + 1 \pmod n = 2k + 1 = \frac{n}{2}. \qquad \square$$

Lemma 6. *If n is even, then the column vector x in (3) is given by*

$$x(i) = \begin{cases} \frac{n}{2} & \textit{if } i \textit{ is odd}, \\ 0 & \textit{if } i \textit{ is even}, \end{cases} \qquad i = 1, 2, \ldots, \frac{n-2}{2},$$

where $x(i)$ denotes the i-th component of x.

Proof. Note that x is located in the $(n/2)$-th column of H_n. Thus, $x(i) = H_n(i, n/2)$ for $i = 1, 2, \ldots, (n-2)/2$. If $i = 2k$ for some positive integer k, then

$$H_n\Big(i, \frac{n}{2}\Big) = kn \pmod n = 0.$$

Now, if $i = 2k + 1$ for some positive integer k, then

$$H_n\Big(i, \frac{n}{2}\Big) = kn + \frac{n}{2} \pmod n = \frac{n}{2}. \qquad \square$$

Taking into account Lemma 4, we next obtain a symmetric block-diagonal matrix similar to H_n for all n. This result also follows from [Cantoni and Butler 1976, Lemma 3]. Observe that $A - RB$ and $A + RB$, where A and B are as in Lemma 4, are symmetric matrices since RB is symmetric by Remark 2.

Lemma 7. (1) *Suppose that n is odd and let H_n be expressed as in (2). Then,*

$$K H_n K^{-1} = \begin{bmatrix} A - RB & 0 \\ 0 & A + RB \end{bmatrix}, \quad \textit{where } K = \begin{bmatrix} I & -R \\ I & R \end{bmatrix}.$$

(2) *Suppose that n is even and let H_n be expressed as in (3). Then,*

$$K H_n K^{-1} = \begin{bmatrix} A - RB & 0 & 0 \\ 0 & A + RB & \sqrt{2}x \\ 0 & \sqrt{2}x^T & q \end{bmatrix}, \quad \textit{where } K = \begin{bmatrix} I & 0 & -R \\ I & 0 & R \\ 0 & \sqrt{2} & 0 \end{bmatrix}.$$

As a consequence of the previous result, the study of the kernel of the bisymmetric matrix H_n can be reduced to the study of the kernel of the diagonal blocks of the block-diagonal matrix similar to H_n given in Lemma 7. In fact, when n is odd, if $\{u_1, \ldots, u_j\}$ is a basis for the kernel of $A - RB$ and $\{u_{j+1}, \ldots, u_{j+k}\}$ is a basis for the kernel of $A + RB$, then $\{K^{-1}w_1, \ldots, K^{-1}w_{j+k}\}$ is a basis for the kernel of H_n, where $w_i = [u_i \;\; 0]^T \in M_{n-1,1}$ for $i \le j$, and $w_i = [0 \;\; u_i]^T \in M_{n-1,1}$ for $i > j$.

Analogously, when n is even, if $\{u_1, \dots, u_j\}$ is a basis for the kernel of $A - RB$ and $\{u_{j+1}, \dots, u_{j+k}\}$ is a basis for the kernel of

$$\begin{bmatrix} A+RB & \sqrt{2}x \\ \sqrt{2}x^T & q \end{bmatrix}, \tag{4}$$

then $\{K^{-1}w_1, \dots, K^{-1}w_{j+k}\}$ is a basis for the kernel of $H_{n,}$ where each w_i is defined as before. Note that, if $n = 2$, the matrix $A - RB$ is empty.

In what follows, we denote by $\mathbb{A} + \mathbb{RB}$ the symmetric matrix $A + RB$ if n is odd and

$$\begin{bmatrix} A+RB & 2x \\ 2x^T & 2q \end{bmatrix} \tag{5}$$

if n is even. Clearly, $\mathbb{A} + \mathbb{RB} \in M_{\lfloor n/2 \rfloor}$. Note that v is in the kernel of the matrix (4) if and only if

$$\begin{bmatrix} I_{(n-2)/2} & 0 \\ 0 & \sqrt{2}/2 \end{bmatrix} v$$

is in the kernel of the matrix (5). In particular, the matrices (4) and (5) have the same rank.

Next we give an explicit expression for the symmetric matrix $\mathbb{A} + \mathbb{RB}$.

Lemma 8. *The matrix $\mathbb{A} + \mathbb{RB} \in M_{\lfloor n/2 \rfloor}$ is given by*

$$(\mathbb{A} + \mathbb{RB})(i, j) = \begin{cases} 0 & \textit{if } n \textit{ divides } ij, \\ n & \textit{otherwise}, \end{cases} \qquad i, j = 1, \dots, \left\lfloor \frac{n}{2} \right\rfloor.$$

Proof. Recall that $A, B \in M_{\lfloor (n-1)/2 \rfloor}$. Suppose that $1 \le i, j \le \lfloor (n-1)/2 \rfloor$. We have

$$A(i, j) = H_n(i, j)$$

and

$$RB(i, j) = B\left(\left\lfloor \frac{n+1}{2} \right\rfloor - i, j\right) = H_n(n - i, j).$$

Thus, for $1 \le i, j \le \lfloor (n-1)/2 \rfloor$,

$$\begin{aligned} (\mathbb{A} + \mathbb{RB})(i, j) &= H_n(i, j) + H_n(n - i, j) \\ &= ij \pmod n + (n - i)j \pmod n \\ &= ij \pmod n + (-ij) \pmod n, \end{aligned}$$

which implies the claim for the entry in position (i, j). If n is odd, the proof is complete. Now suppose that n is even. By Lemma 5,

$$(\mathbb{A} + \mathbb{RB})\left(\frac{n}{2}, \frac{n}{2}\right) = 2q = \begin{cases} 0 & \text{if } n \equiv 0 \pmod 4, \\ n & \text{if } n \not\equiv 0 \pmod 4. \end{cases}$$

Since n divides $(n/2)^2$ if and only if $n \equiv 0 \pmod 4$, the result follows for $(i, j) = (n/2, n/2)$.

Now we consider the case $j = n/2$, where $1 \le i \le n/2 - 1$. By Lemma 6,

$$(\mathbb{A} + \mathbb{R}\mathbb{B})\Big(i, \frac{n}{2}\Big) = 2x(i) = \begin{cases} n & \text{if } i \text{ is odd,} \\ 0 & \text{if } i \text{ is even.} \end{cases}$$

Since n divides $in/2$ if and only if i is even, the result follows for the entries in positions $(i, n/2)$. Taking into account that $\mathbb{A} + \mathbb{R}\mathbb{B}$ is symmetric, the result also follows for the entries in positions $(n/2, j)$, where $1 \le j \le n/2 - 1$. □

Next we compute the rank of $\mathbb{A} + \mathbb{R}\mathbb{B}$ in terms of the proper divisors of n. We call *a proper divisor of* n, where n is a positive integer, a positive divisor of n different from n. Note that any proper divisor of n is less than or equal to $\lfloor n/2 \rfloor$.

Lemma 9. *Let n be a positive integer and k be the number of proper divisors of n. Then,* $\operatorname{rank}(\mathbb{A} + \mathbb{R}\mathbb{B}) = k$.

Proof. Let $i \in \{1, \dots, \lfloor n/2 \rfloor\}$. If $\gcd(i, n) = 1$, then n is not a divisor of ij for all $j = 1, \dots, \lfloor n/2 \rfloor$. By Lemma 8, $(\mathbb{A} + \mathbb{R}\mathbb{B})(i, j) = n$ for all $j = 1, 2, \dots, \lfloor n/2 \rfloor$.

If $\gcd(i, n) \neq 1$ and i has order m in $\mathbb{Z}_n$ (that is, m is the smallest possible integer such that $mi \equiv 0 \pmod n$), then, by Lemma 8, $(\mathbb{A} + \mathbb{R}\mathbb{B})(i, j) = 0$ if and only if $j = ms$ for some positive integer s. Moreover, the nonzero entries in the i-th row are equal to n. Thus, from the comments above, we conclude that there are at most k distinct rows in $\mathbb{A} + \mathbb{R}\mathbb{B}$, corresponding to the k proper divisors of n. Moreover, one of these rows has all entries equal to n, while the remaining have the first zero entry in distinct columns and have all the nonzero entries equal to n. Note that distinct proper divisors have distinct orders. By elementary row operations, it can be seen that these k rows are linearly independent, which proves the result. □

Remark 10. When n is prime, Lemma 9 implies that $\operatorname{rank}(\mathbb{A} + \mathbb{R}\mathbb{B}) = 1$.

Another immediate consequence of Lemma 9 is given in the next corollary.

Corollary 11. *Let n be a positive integer and k be the number of proper divisors of n. Then,*

$$\dim(\ker(\mathbb{A} + \mathbb{R}\mathbb{B})) = \Big\lfloor \frac{n}{2} \Big\rfloor - k.$$

Since, from Lemma 7,

$$\operatorname{rank}(H_n) = \operatorname{rank}(A - RB) + \operatorname{rank}(\mathbb{A} + \mathbb{R}\mathbb{B}) \quad \text{and} \quad \operatorname{rank}(A - RB) \le \Big\lfloor \frac{n-1}{2} \Big\rfloor,$$

from Lemma 9 we get the next result.

Corollary 12. *Let n be a positive integer and let k be the number of proper divisors of n. Then,*

$$k \le \operatorname{rank}(H_n) \le \Big\lfloor \frac{n-1}{2} \Big\rfloor + k.$$

Next we compute a basis for the kernel of $\mathbb{A}+\mathbb{R}\mathbb{B}$ when $n > 2$. Note that when $n = 2$, the kernel of $\mathbb{A}+\mathbb{R}\mathbb{B}$ only contains the zero vector by Corollary 11. We start with a technical lemma.

Lemma 13. *Let n be a positive integer. For each $j \in \{1, 2, \dots, \lfloor n/2 \rfloor\}$, let $d_j = \gcd(j, n)$. Then, for $1 \le i \le \lfloor n/2 \rfloor$, we have $(\mathbb{A}+\mathbb{R}\mathbb{B})(i, j) = 0$ if and only if $(\mathbb{A}+\mathbb{R}\mathbb{B})(i, d_j) = 0$.*

Proof. Note that, from Lemma 8, the statement $(\mathbb{A}+\mathbb{R}\mathbb{B})(i, j) = 0$ if and only if $(\mathbb{A}+\mathbb{R}\mathbb{B})(i, d_j) = 0$ is equivalent to n divides ij if and only if n divides id_j.

Suppose that n divides ij. Then, there exists a positive integer k such that $nk = ij$. Since $\gcd(j, n) = d_j$, we have $d_j = jx + ny$ for some $x, y \in \mathbb{Z}$, $x \neq 0$. Thus,

$$nk = i\left(\frac{d_j - ny}{x}\right),$$

which implies $n(xk + iy) = id_j$ and, therefore, n divides id_j.

Suppose now that n divides id_j. Since d_j divides j, we have id_j divides ij and, therefore, n divides ij. □

We denote by e_i the vector of appropriate size whose entries are 0 except the entry in position i which is 1.

Theorem 14. *Let $n > 2$. The set of vectors $u_j := e_j - e_{d_j} \in M_{\lfloor n/2 \rfloor, 1}$, with $j \in \{1, \dots, \lfloor n/2 \rfloor\}$, where j is not a divisor of n and $d_j = \gcd(j, n)$, forms a basis for $\ker(\mathbb{A}+\mathbb{R}\mathbb{B})$.*

Proof. First we show that the vectors u_j are in the kernel of $\mathbb{A}+\mathbb{R}\mathbb{B}$. Note that, by the definition of u_j, the i-th entry of the vector $(\mathbb{A}+\mathbb{R}\mathbb{B})u_j$ is $(\mathbb{A}+\mathbb{R}\mathbb{B})(i, j) - (\mathbb{A}+\mathbb{R}\mathbb{B})(i, d_j)$. By Lemma 8, each entry of $\mathbb{A}+\mathbb{R}\mathbb{B}$ is either n or 0 and, by Lemma 13, $(\mathbb{A}+\mathbb{R}\mathbb{B})(i, j) = 0$ if and only if $(\mathbb{A}+\mathbb{R}\mathbb{B})(i, d_j) = 0$. This implies that $(\mathbb{A}+\mathbb{R}\mathbb{B})u_j = 0$ for all u_j, as desired.

Next, we show that the vectors u_j form a linearly independent set. Let U be the matrix whose columns are the vectors u_j and let J be the set of integers in $\{1, \dots, \lfloor n/2 \rfloor\}$ that are not divisors of n. Notice that if $j_1, j_2 \in J$, then $j_1 \neq d_{j_2}$ since, if $j_1 = d_{j_2} = \gcd(j_2, n)$, then j_1 would divide n. This implies that the submatrix of U formed by the rows indexed by J is a row permutation of the identity matrix of size $|J|$, which shows that U has full rank.

We have obtained a set of $|J|$ linearly independent vectors in the kernel of $\mathbb{A}+\mathbb{R}\mathbb{B}$. Since the largest proper divisor of n is less than or equal to $\lfloor n/2 \rfloor$, we have $|J| = \lfloor n/2 \rfloor - k$, where k is the number of proper divisors of n. By Corollary 11, the result follows. □

Example 15. Let $n = 24$. Then $\dim(\ker(\mathbb{A}+\mathbb{R}\mathbb{B})) = 5$ and the set J defined in the proof of Theorem 14 is given by $\{5, 7, 9, 10, 11\}$. A basis for $\ker(\mathbb{A}+\mathbb{R}\mathbb{B})$ is given by the vectors

$$\begin{bmatrix}-1\\0\\0\\0\\1\\0\\0\\0\\0\\0\\0\\0\end{bmatrix},\begin{bmatrix}-1\\0\\0\\0\\0\\0\\1\\0\\0\\0\\0\\0\end{bmatrix},\begin{bmatrix}0\\0\\-1\\0\\0\\0\\0\\0\\1\\0\\0\\0\end{bmatrix},\begin{bmatrix}0\\-1\\0\\0\\0\\0\\0\\0\\0\\1\\0\\0\end{bmatrix},\begin{bmatrix}-1\\0\\0\\0\\0\\0\\0\\0\\0\\0\\1\\0\end{bmatrix}.$$

Though we could not find appropriate techniques from matrix theory to show it, numerical experiments in Matlab, in which the rank of H_n was computed for any n from 2 to 1000, suggest the following conjecture. Recall that $\operatorname{rank}(C_n) = \operatorname{rank}(H_n)$, where C_n is the matrix defined in (1).

Conjecture 16. *Let n be a positive integer and let k be the number of proper divisors of n. Then,*

$$\operatorname{rank}(C_n) = \operatorname{rank}(H_n) = \left\lfloor \frac{n-1}{2} \right\rfloor + k.$$

Clearly, the conjecture holds when $n = 2$. In the next section we prove the conjecture when n is prime. The result when n is not prime remains open.

Remark 17. Because of Lemmas 7 and 9, it follows that, if Conjecture 16 is true and $n > 2$, then $A - RB$ is a nonsingular matrix. Note that, if $n = 2$, the matrix $A - RB$ is empty and $\mathbb{A} + \mathbb{R}\mathbb{B}$ is nonsingular as well.

3. The rank of the matrix H_n when n is prime

In this section we compute the rank of the matrix H_n, when n is prime, using techniques from character theory and analytic number theory.

We start with some basic concepts and lemmas that will be used to obtain the main result.

Definition 18 (character [Apostol 1976, Section 6.5]). Let G be a group and let $\mathbb{C}$ denote the set of complex numbers. A function $f : G \to \mathbb{C}$ is called a character of G if

(i) f is a group homomorphism of G, that is, $f(g_1 g_2) = f(g_1) f(g_2)$ for all $g_1, g_2 \in G$; and

(ii) $f(g) \neq 0$ for some $g \in G$.

The set of characters of a finite group G is also a group with respect to the group operation of pointwise multiplication defined by $(f_1 \cdot f_2)(g) = f_1(g) f_2(g)$ [Apostol 1976, Section 6.6]. This group is denoted by $\hat{G}$. The identity element of $\hat{G}$ is the character f_I given by $f_I(g) = 1$ for all $g \in G$. The inverse of a character f is $\bar{f}$ given by $\bar{f}(g) = \overline{f(g)}$ for all $g \in G$, where $\overline{f(g)}$ is the complex conjugate of $f(g)$. The identity element of $\hat{G}$ is called the *principal character of* G, while the other characters are called *nonprincipal characters* of G. Note that any character of G maps the identity element of G to 1.

According to the next result, if f is a character of a finite group G, the range of a character of G lies on the unit circle. We recall that if G is a finite group with identity element e, then the *exponent of* G is the least positive integer k such that $g^k = e$ for all $g \in G$.

Proposition 19 [Apostol 1976, Theorem 6.7]. *Let G be a finite group with identity element e and let $f \in \hat{G}$. Then, $f(e) = 1$ and each function value $f(g)$ is an m-th root of unity, where m is the exponent of G.*

One may think that the set of characters of a group could potentially contain many functions. The next theorem gives the exact number of characters when the group is finite and abelian.

Proposition 20 [Apostol 1976, Theorem 6.8]. *If G is a finite abelian group, then $|\hat{G}| = |G|$.*

In particular, if G is a finite cyclic group of order n (in which case the exponent of G equals the order of G) and g is a generator of G, then the n characters of G are determined by sending g to the different n-th roots of unity in $\mathbb{C}$.

Example 21. Let G be the additive group $\mathbb{Z}_4$. Then, there exist four characters f_1, f_2, f_3, f_4 of G and each character value is in the set $\{1, -1, i, -i\}$, the 4th roots of unity. Suppose that f_1 is the principal character and f_2, f_3, f_4 are defined by $f_2(1) = -1,\ f_3(1) = i$ and $f_4(1) = -i$. Note that, since 1 is a generator of G and characters are group homomorphisms, f_2, f_3, f_4 are well-defined. We give the range of the characters of G through a 4×4 matrix A whose entry $A(i, j)$ is given by $f_i(g_j)$, where $g_1 = 0,\ g_2 = 1,\ g_3 = 2$, and $g_4 = 3$:

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & i & -1 & -i \\ 1 & -i & -1 & i \end{bmatrix}.$$

The following concept will be key in the proof of our main results.

Definition 22 (group matrix [Chan et al. 1998]). Let G be a finite group of order n. Fix an enumeration $\{g_1, \dots, g_n\}$ of the elements of G. For every complex-valued function α on G, the matrix A_α given by $A_\alpha(i,j)=\alpha(g_ig_j^{-1})$ is called *a group matrix associated to* α.

Example 23. Let G be the additive group $\mathbb{Z}_4$ and let f_2 be the character defined in Example 21. Then, the following matrix is a group matrix associated to f_2:

$$A_{f_2}=\begin{bmatrix} 1 & -i & i & -1\\ -1 & 1 & -i & i\\ i & -1 & 1 & -i\\ -i & i & -1 & 1\end{bmatrix}.$$

In what follows, we let p denote a prime number. Next we show that the rank of H_p can be computed by finding the rank of a group matrix. In particular, the next lemma states that the matrix H_p can be obtained by permuting some columns of a group matrix associated with a real-valued function on the multiplicative group $\mathbb{Z}_p^\times$ consisting of the units of $\mathbb{Z}_p$.

Lemma 24. *Let p be a prime number. Let $\alpha:\mathbb{Z}_p^\times\to\mathbb{N}$ be given by $\alpha(\bar{m})=m$, where $\bar{m}$ denotes the equivalence class mod p of $m\in\{1,2,\dots,p-1\}$. Then, H_p is a column permutation of the group matrix A_α associated to α.*

Proof. First recall that, since p is a prime number, the group $\mathbb{Z}_p^\times$ is a cyclic group under multiplication. Let $\bar{g}$, where $g\in\{1,2,\dots,p-1\}$, be a generator for $\mathbb{Z}_p^\times$ and consider the enumeration of $\mathbb{Z}_p^\times$ given by $\{\overline{g^{\sigma(1)}},\overline{g^{\sigma(2)}},\dots,\overline{g^{\sigma(p-1)}}\}$, where σ is a permutation of $\{1,2,\dots,p-1\}$ such that $g^{\sigma(i)}=i$. Then,

$$A_\alpha(i,j)=\alpha(\overline{g^{\sigma(i)}}\,\overline{g^{-\sigma(j)}})=ij^{-1}\pmod p.$$

Let π be the permutation of $\{1,2,\dots,p-1\}$ such that $\pi(j)=j^{-1}$. Now consider the matrix $\widetilde{A_\alpha}$ obtained from A_α by permuting its columns as follows: column j of $\widetilde{A_\alpha}$ is column $\pi(j)=j^{-1}$ of A_α. Then, $\widetilde{A_\alpha}=H_p$ is obtained by permuting the columns of A_α and the result follows. □

The previous lemma implies that $\operatorname{rank}(H_p)=\operatorname{rank}(A_\alpha)$.

We next characterize the eigenvalues of a group matrix of a finite abelian group, associated to an injective function, and show that it is diagonalizable, implying that its rank is the number of its nonzero eigenvalues. For this purpose, we present the next lemma which gives the spectrum of a group matrix associated to an integer-valued injective function in terms of the values of the characters of G at an element of the group ring $\mathbb{Z}[G]$, when G is a finite abelian group. Note that any character in the character group of G can be extended by linearity to a complex-valued function on $\mathbb{Z}[G]$.

Lemma 25 [Chan et al. 1998; Jungnickel 1993, Theorem 7.7.4]. *Let G be a finite abelian group and α an injective function from G to $\mathbb{N}$. Let $a=\sum_{g\in G}\alpha(g)g\in\mathbb{Z}[G]$. Then, the group matrix A_α associated to α is diagonalizable and its spectrum is the set $\{f(a):f\in\hat{G}\}$.*

Since A_α is diagonalizable, we can compute the rank of A_α by counting the number of eigenvalues distinct from zero. Thus, $\operatorname{rank}(A_\alpha) = |\{f \in \hat{G} : f(a) \neq 0\}|$.

Remark 26. Taking into account Lemmas 24 and 25, in order to compute $\operatorname{rank}(H_p)$, it is enough to determine the number of characters f in the character group of $\mathbb{Z}_p^\times$ such that $\sum_{i=1}^{p-1} i f(\bar{i}) \neq 0$.

Here, it becomes convenient to work with the so-called Dirichlet characters whose definition we give below.

Definition 27 (Dirichlet character [Apostol 1976, Section 6.8]). Let $n \in \mathbb{N}$ and f be any character of $\mathbb{Z}_n^\times$. The function $\chi : \mathbb{N} \to \mathbb{C}$ given by

$$\chi(m) = \begin{cases} f(\bar{m}) & \text{if } n \text{ and } m \text{ are relatively prime,} \\ 0 & \text{if } n \text{ and } m \text{ are not relatively prime} \end{cases}$$

is called the *Dirichlet character modulo n induced by f*. The Dirichlet character induced by the principal character is called the *principal Dirichlet character modulo n*. A Dirichlet character modulo n that is not the principal character is called *nonprincipal*.

It is easy to see that Dirichlet characters modulo n are completely multiplicative and periodic with period n [Apostol 1976, Theorem 6.15]; that is, if χ is a Dirichlet character, then

- $\chi(x+n) = \chi(x)$ for all $x \in \mathbb{N}$;
- $\chi(xy) = \chi(x)\chi(y)$ for all $x, y \in \mathbb{N}$.

Note that the number of Dirichlet characters modulo n equals the order of $\mathbb{Z}_n^\times$ since, by Proposition 20, the number of characters of a finite abelian group equals its cardinality.

Example 28. The following table displays the Dirichlet characters for $n = 5$. We obtain four functions since $\mathbb{Z}_5$ contains 4 units. We only give the values of the functions on the set $\{1, \dots, 5\}$ since these Dirichlet characters are periodic functions of period 5:

x	1	2	3	4	5
$\chi_1(x)$	1	1	1	1	0
$\chi_2(x)$	1	-1	-1	1	0
$\chi_3(x)$	1	i	$-i$	-1	0
$\chi_4(x)$	1	$-i$	i	-1	0

Definition 29 (primitive Dirichlet character [Apostol 1976, Section 8.7]). If χ is a Dirichlet character modulo n, we say that χ is primitive if for every proper divisor d of n, there exists an integer a such that $a \equiv 1 \pmod{d}$, $\gcd(a, n) = 1$, and $\chi(a) \neq 1$.

Example 30. Consider the Dirichlet characters modulo 5, given in Example 28. The only proper divisor of 5 is 1. Note that χ_1 is not primitive since $\chi_1(a) = 1$ whenever $\gcd(a, n) = 1$. However, the rest of the Dirichlet characters are primitive since $\chi_i(2) \neq 1$ for $i = 2, 3, 4$.

The observations in the previous example can be generalized as follows.

Lemma 31 [Apostol 1976, Theorems 8.13 and 8.14]. *The principal Dirichlet character modulo n is not primitive. Moreover, if n is prime, all nonprincipal Dirichlet characters modulo n are primitive.*

Definition 32 (admissible Dirichlet character). Let χ be a Dirichlet character modulo n. We say that χ is *admissible* if

$$\sum_{i=1}^{n-1} i\chi(i) \neq 0.$$

Note, the principal Dirichlet character modulo p is admissible since $\sum_{i=1}^{p-1} i \neq 0$. Taking into account Remark 26, we obtain the following.

Remark 33. If p is prime, the rank of H_p is equal to the number of admissible Dirichlet characters modulo p.

In order to see which Dirichlet characters are admissible, we need some well-known results from the theory of Dirichlet L-functions.

Definition 34 (Dirichlet L-function [Apostol 1976, Sections 11 and 12]). Let χ be a Dirichlet character modulo n and $s \in \mathbb{C}$ with real part greater than 1. The associated Dirichlet L-series is the absolutely convergent series given by

$$L(s, \chi) = \sum_{i=1}^{\infty} \frac{\chi(i)}{i^s}.$$

If χ is nonprincipal, $L(s, \chi)$ is a complex-valued function in s that can be analytically extended to an entire function on the whole complex plane [Apostol 1976, Theorem 12.5]. This function is called a Dirichlet L-function and is also denoted by $L(s, \chi)$.

The following is a well-known result in analytic number theory.

Lemma 35 [Apostol 1976, Thm. 12.20]. *If χ is a nonprincipal Dirichlet character modulo n, then*

$$L(0, \chi) = -\frac{1}{n} \sum_{i=1}^{n-1} i\chi(i).$$

Remark 36. The admissible Dirichlet characters modulo p, where p is prime, are exactly the principal Dirichlet character and the nonprincipal Dirichlet characters such that $L(0, \chi) \neq 0$.

In order to determine when $L(0, \chi) \neq 0$, we introduce the functional equation for Dirichlet L-functions.

Let $\bar{\chi}$ denote the complex conjugate of the Dirichlet character χ.

Lemma 37 (functional equation [Apostol 1976, Theorem 12.11]). *Let χ be a primitive Dirichlet character modulo n. Then, for all $s \in \mathbb{C}$, we have*

$$L(1-s, \chi) = \frac{n^{s-1}\Gamma(s)}{(2\pi)^s}\big(e^{-\pi i s/2} + \chi(-1)e^{\pi i s/2}\big)G(1, \chi)L(s, \bar{\chi}),$$

where $\Gamma(s)$ is the Gamma function and $G(1, \chi) = \sum_{r=1}^{n} \chi(r)e^{2\pi i r/n}$ is the Gauss sum associated with χ.

The following are well-known results in analytic number theory.

Lemma 38 [Apostol 1976, Theorem 8.15]. *Let χ be a primitive Dirichlet character modulo n. Then, $G(1, \chi) \neq 0$.*

Lemma 39 [Apostol 1976, Section 7.3]. *Let χ be a nonprincipal Dirichlet character modulo n. Then, $L(1, \chi)$ is finite and nonzero.*

The next result gives necessary and sufficient conditions for a Dirichlet character modulo p to be admissible.

Lemma 40. *Let $p > 2$ be a prime number and consider the primitive $(p-1)$-th root of unity $w = e^{2\pi i/(p-1)}$. Let $\bar{g}$ be a generator of $\mathbb{Z}_p^\times$ and let f_k be the character of $\mathbb{Z}_p^\times$ defined by $f_k(\bar{g}) := w^{k-1}$, with $k = 1, \dots, p-1$. Let $\chi_1, \dots, \chi_{p-1}$ be the Dirichlet characters modulo p induced by $f_1, \dots, f_k$, respectively. Then, for $k = 2, \dots, p-1$, we have χ_k is admissible if and only if k is even.*

Proof. Since $\bar{g}$ is a generator of $\mathbb{Z}_p^\times$, we have $g^{p-1} \equiv 1 \pmod p$ and $g^s \neq 1 \pmod p$ for $s = 1, \dots, p-2$. Thus, $g^{(p-1)/2} \equiv -1 \pmod p$. So, for $k = 2, \dots, p-1$, we have $\chi_k(-1) = \chi_k(g^{(p-1)/2}) = (w^{(p-1)/2})^{k-1} = (-1)^{k-1}$. Therefore, $\chi_k(-1) = -1$ if k is even and $\chi_k(-1) = 1$ if k is odd. Since p is prime and χ_k is nonprincipal, χ_k is primitive by Lemma 31. By Lemma 37,

$$L(0, \chi_k) = \frac{1}{2\pi}\big(-i + \chi_k(-1)i\big)G(1, \chi_k)L(1, \overline{\chi_k}).$$

Note that if χ_k is a nonprincipal Dirichlet character, then $\bar{\chi}$ is also nonprincipal. Taking into account Lemmas 38 and 39, it follows that, if k is even, $L(0, \chi_k) \neq 0$; if k is odd, $L(0, \chi_k) = 0$. Now the result follows from Remark 36. □

We can now give the rank of the matrix H_p when $p > 2$ is a prime number.

Lemma 41. *Let $p > 2$ be a prime number. Then,* $\operatorname{rank}(H_p) = (p+1)/2$.

Proof. By Lemma 40, we have that the nonprincipal Dirichlet characters modulo p $\chi_2, \chi_4, \dots, \chi_{p-1}$ are admissible, while $\chi_3, \chi_5, \dots, \chi_{p-2}$ are not admissible. Since, by Remark 36, χ_1 is admissible, the result follows taking into account Remark 33. □

Observe that, by Lemma 41, we have that Conjecture 16 is true when $n > 2$ is prime. Then, by Remark 17, Lemma 7, and Theorem 14, we can obtain a basis for the kernel of H_p when $p > 2$ is prime (note that when $p = 2$, the kernel of H_p is $\{0\}$). From this basis for the kernel of H_p, we can easily obtain a basis for the kernel of C_p, the $p \times p$ matrix whose (i, j)-th entry is $ij \pmod{p}$.

Theorem 42. *Let $p > 2$ be a prime number and $C_p \in M_p$ be defined by $C_p(i, j) = ij \pmod{p}$. Let K be as in Lemma 7. Let $u_j := e_j - e_1 \in M_{(p-1)/2,1}$ and $w_j = [0_{(p-1)/2}, u_j, 0]^T \in M_{p,1}$, with $j = 2, \dots, (p-1)/2$. Then, the set of vectors $\{K^{-1}w_2, \dots, K^{-1}w_{(p-1)/2}, e_p\}$ is a basis for the kernel of C_p. In particular,* $\operatorname{rank}(C_p) = (p+1)/2$.

4. Application

We now present a number theoretic application of the problem we have considered in this paper. This application, which motivated our work, appears in the context of the study of Stickelberger relations on class groups of group rings.

Let G be a finite abelian group and let n be the order of G. Fix a primitive n-th root of unity z. Then, for each $g \in G$ and $f \in \hat{G}$, there is a unique integer r, with $1 \le r \le n$, such that $f(g) = z^r$. We therefore can define the function

$$\langle \cdot, \cdot \rangle : G \times \hat{G} \to \mathbb{Q}/\mathbb{Z}$$

given by

$$\langle g, f \rangle = \left\{ \frac{r}{n} \right\},$$

where $\{r/n\}$ denotes the fractional part of r/n.

Note that the group rings $\mathbb{Q}[G]$ and $\mathbb{Q}[\hat{G}]$ are $\mathbb{Q}$-vector spaces with dimension $|G| = |\hat{G}|$, and G and $\hat{G}$ are bases for $\mathbb{Q}[G]$ and $\mathbb{Q}[\hat{G}]$, respectively. Thus, we may extend the function above via linearity to

$$\langle \cdot, \cdot \rangle : \mathbb{Q}[G] \times \mathbb{Q}[\hat{G}] \to \mathbb{Q}$$

defined by

$$\left\langle \sum_{g \in G} c_g g, \sum_{f \in \hat{G}} c_f f \right\rangle = \sum_{g \in G} \sum_{f \in \hat{G}} c_g c_f \langle g, f \rangle,$$

where $c_g, c_f \in \mathbb{Q}$. Now consider the function h: $\mathbb{Q}[\hat{G}] \to \mathbb{Q}[G]$ given by

$$h(a) = \sum_{g \in G} \langle g, a \rangle g \quad \text{for any } a \in \mathbb{Q}[\hat{G}]. \tag{6}$$

We may view h as a linear map between two $\mathbb{Q}$-vector spaces of dimension $|G|$. An interesting problem, which motivated our work, is the study of the kernel of h.

When the group G is cyclic (and, therefore, isomorphic to $\mathbb{Z}_n$ for some n), we can determine explicitly the matrix representation of h as the following lemma states.

Lemma 43. *Let G be the additive group $\mathbb{Z}_n$ and g be a generator of G. Let $\hat{G} = \{f_1, f_2, \ldots, f_n\}$, where $f_i(g) = z^{i-1}$ and z is a primitive n-th root of unity. Then, the matrix representation R_n of h in the bases $\beta_1 = \{f_2, f_3, \ldots, f_n, f_1\}$ and $\beta_2 = \{g, g^2, \ldots, g^{n-1}, e\}$ is given by*

$$R_n(i, j) = \left\{\frac{ij}{n}\right\} = \frac{ij \pmod{n}}{n}, \quad i, j = 1, 2, \ldots, n.$$

Proof. For $i, j = 1, \ldots, n-1$, since $f_{j+1}(g^i) = z^{ij}$, the (i, j)-th entry of R_n is given by

$$\langle g^i, f_{j+1}\rangle = \left\{\frac{ij}{n}\right\} = \frac{ij \pmod{n}}{n}.$$

Since $f_j(e) = 1 = z^0$, we have $\langle e, f_j\rangle = 0$ for $j = 1, \ldots, n$, which implies that the last row of R_n is zero. Since $f_1(g^i) = 1 = z^0$, we have $\langle g^i, f_1\rangle = 0$ for $i = 1, \ldots, n$, and, then, the last column of R_n is zero. Thus, the claim follows. □

Note that

$$R_n = \frac{1}{n} C_n = \frac{1}{n}\begin{bmatrix} H_n & 0 \\ 0 & 0 \end{bmatrix}.$$

Finally, we observe that, although the function h given in (6) is defined between $\mathbb{Q}$-vector spaces, the determination of the kernel and rank of the matrix representation of h can be done by considering it as a matrix over the real numbers.

Acknowledgement

We would like to thank the anonymous referee for the time spent reading our paper and for the suggestions to improve its presentation.

References

[Apostol 1976] T. M. Apostol, *Introduction to analytic number theory*, Springer, New York, 1976. MR 55 #7892 Zbl 0335.10001

[Cantoni and Butler 1976] A. Cantoni and P. Butler, "Eigenvalues and eigenvectors of symmetric centrosymmetric matrices", *Linear Algebra and Appl.* **13**:3 (1976), 275–288. MR 53 #476 Zbl 0326.15007

[Chan et al. 1998] W.-K. Chan, Y.-C. Chan, and M.-K. Siu, "Minimal rank of abelian group matrices", *Linear and Multilinear Algebra* **44**:3 (1998), 277–285. MR 99j:15002 Zbl 0960.15003

[Jungnickel 1993] D. Jungnickel, *Finite fields: Structure and arithmetics*, Bibliographisches Institut, Mannheim, 1993. MR 94g:11109 Zbl 0779.11058

Received: 2014-10-14 Revised: 2015-04-13 Accepted: 2015-04-16

mbueno@math.ucsb.edu
Mathematics Department and College of Creative Studies,
The University of California, Santa Barbara,
Santa Barbara, CA 93106-3080, United States

sbf@fep.up.pt
CEAFEL, Faculdade de Economia, Universidade do Porto,
4200-464 Porto, Portugal

karkoj@rpi.edu
Applied Mathematics Department, Rensselaer Polytechnic
Institute, Troy, NY 12180, United States

mayfield13@up.edu
Fast Enterprises LLC, Madison, WI 53704, United States

robsamalis@gmail.com
Department of Mathematics, University of Georgia,
Athens, GA 30602, United States

aet156@psu.edu
Department of Mathematics, Pennsylvania State University,
State College, PA 16803, United States

mathematical sciences publishers

msp

INVOLVE 9:2 (2016)
dx.doi.org/10.2140/involve.2016.9.281

Harnack's inequality for second order linear ordinary differential inequalities

Ahmed Mohammed and Hannah Turner

(Communicated by Johnny Henderson)

We prove a Harnack-type inequality for nonnegative solutions of second order ordinary differential inequalities. Maximum principles are the main tools used, and to make the paper self-contained, we provide alternative proofs to those available in the literature.

1. Introduction

The aim of this paper is to present a self-contained discussion of the Harnack and Harnack-type inequalities for nonnegative solutions of second order linear ordinary differential inequalities of the form

$$Lu \le f(x), \quad x \in I := (A, B), \tag{1-1}$$

where, for $u \in C^2(I)$,

$$Lu := u''(x) + p(x)u'(x) + q(x)u. \tag{1-2}$$

Here and in the sequel, the notation $C^2(I)$ stands for the class of twice continuously differentiable real-valued functions on the open interval I. Likewise, we write $C(I)$ for the class of continuous real-valued functions on I. Throughout, we will assume, without further mention, that $p, q, f \in C(I)$. In this case, (1-2) can be rewritten as

$$Lu = \frac{1}{r(x)}(r(x)u')' + q(x)u, \quad \text{where } r(x) := \exp\left(\int^x p(t)\,dt\right). \tag{1-3}$$

Let $\mathcal{H}$ be a class of nonnegative and locally bounded functions in the open interval $I = (A, B)$. We say that Harnack's inequality holds for the class $\mathcal{H}$ if and only if given any closed interval $[a, b] \subseteq I$, there is a positive constant C such that

$$\sup_{x\in[a,b]} u(x) \le C \inf_{x\in[a,b]} u(x) \quad \text{for all } u \in \mathcal{H}. \tag{1-4}$$

MSC2010: 34C11.
Keywords: Harnack's inequality, maximum principles, ordinary differential inequalities.

The important point here is that C is independent of $u \in \mathcal{H}$. The class $\mathcal{H}$ is usually a collection of nonnegative (or nonpositive) solutions of some differential equations.

This type of inequality is named after Carl Gustav Axel von Harnack (1851–1888) who first derived the inequality for nonnegative harmonic functions in the plane. The inequality became a very important tool in the study of solutions to second order linear and nonlinear elliptic partial differential equations. We refer the interested reader to the article [Kassmann 2007] for a detailed account on some history and theoretical developments of this fascinating inequality, as well as an extensive bibliography of articles and monographs related to Harnack's inequality. We direct the reader to the paper [Berhanu and Mohammed 2005] for a simple application of Harnack's inequality to the ordinary differential equation $Lu = f$. The same paper also provides an example that shows the explicit dependence of the constant C in (1-4) on the differences $a - A$ and $B - b$.

2. On maximum principles

To develop a version of Harnack's inequality for nonnegative solutions of (1-1), we need several results on maximum principles which can be found in [Protter and Weinberger 1984]. To make the paper self-contained and for the readers' convenience, we provide alternative proofs to these maximum principles under the assumption that p and q are continuous on I.

We first introduce an auxiliary function that will be used in our proof of a basic theorem on maximum principles. We use the notation $J := (\alpha, \beta)$ for $\alpha < \beta$. Consider the following auxiliary function, with $\sigma > 0$ to be chosen:

$$z(x) = \sigma(x-\alpha) - e^{\sigma(x-\alpha)}. \tag{2-1}$$

We observe that

$$z(\alpha) = -1 \quad \text{and} \quad z'(\alpha) = 0.$$

Direct computation shows

$$Lz = -\sigma^2 e^{\sigma(x-\alpha)}\left(1 + \frac{p(x)}{\sigma}(1 - e^{-\sigma(x-\alpha)}) + \frac{q(x)}{\sigma}\left(\frac{1}{\sigma} - (x-\alpha)e^{-\sigma(x-\alpha)}\right)\right).$$

If p and q are bounded on $[\alpha, \beta]$, we see that

$$\lim_{\sigma\to\infty}\left(\frac{p(x)}{\sigma}(1 - e^{-\sigma(x-\alpha)}) + \frac{q(x)}{\sigma}\left(\frac{1}{\sigma} - (x-\alpha)e^{-\sigma(x-\alpha)}\right)\right) = 0,$$

uniformly on $[\alpha, \beta]$. Therefore, in this case we can choose $\sigma > 0$ large enough such that

$$Lz \le -c\sigma^2 e^{\sigma(x-\alpha)} \quad \text{in } J$$

for some constant $c > 0$.

Most of the theorems on maximum principles will be easy consequences of the following basic and useful result.

Theorem 2.1. *Let $p, q \in C(\bar{J})$ and $q \le 0$ in J. Let $u \in C^2(J) \cap C(\bar{J})$ be a solution of $Lu \le 0$ in J. Suppose u has a nonpositive minimum at $x_0 \in \{\alpha, \beta\}$. If u is differentiable at x_0, and $u'(x_0) = 0$, then u is a constant in J.*

Proof. We consider the case $x_0 = \alpha$ first. Suppose u has a nonpositive minimum $u(\alpha)$ at α. Furthermore, assume that u is differentiable at α, and $u'(\alpha) = 0$. We consider the auxiliary function z in (2-1) with $\sigma > 0$ such that $Lz \le 0$ in $\bar{J}$. We note that $z(\alpha) = -1$ and $z'(\alpha) = 0$. We fix $\varepsilon > 0$, and set $w := u + \varepsilon z$. We note that $Lw = Lu + \varepsilon Lz = Lu \le 0$. On recalling that $u(\alpha) \le 0$, we have $w(\alpha) = u(\alpha) - \varepsilon < 0$, and $w'(\alpha) = 0$. By continuity of w on $[\alpha, \beta]$, we see that $w(x) < 0$ on $[\alpha, \tau)$ for some $\tau > 0$. Let

$$\eta := \sup\{\rho \in [\alpha, \beta] : w(s) < 0 \ \forall\, 0 \le s < \rho\}.$$

Then we note that

$$\begin{aligned}(r(x)w')' &= (r(x)w')' + r(x)q(x)w - r(x)q(x)w(x)\\ &= r(x)Lw - r(x)q(x)w\\ &\le -r(x)q(x)w(x) \le 0, \quad \alpha < x < \eta. \end{aligned} \tag{2-2}$$

Thus rw' is decreasing on $[\alpha, \eta]$ so that $r(x)w'(x) \le r(\alpha)w'(\alpha) = 0$ on $[\alpha, \eta]$. In particular, this implies that w is decreasing on $[\alpha, \eta]$. Hence $w(x) \le w(\alpha) < 0$ for all $\alpha \le x \le \eta$. This and the continuity of w on $[\alpha, \beta]$ would contradict the definition of η if $\eta < \beta$. Therefore we must have $\eta = \beta$, so that w is decreasing on $[\alpha, \beta]$. In particular, we have

$$u(x) + \varepsilon z(x) \le u(\alpha) + \varepsilon z(\alpha), \quad \alpha \le x \le \beta.$$

Letting $\varepsilon \to 0$, we find that $u(x) \le u(\alpha)$ on $[\alpha, \beta]$. This, together with the fact that $u(x) \ge u(\alpha)$, shows that $u(x) = u(\alpha)$ on $[\alpha, \beta]$.

Now suppose u has a nonpositive minimum at β and $u'(\beta) = 0$. Let $w(x) = u(2\beta - x)$ for $x \in I := [\beta, 2\beta - \alpha]$. Then clearly $w \in C^2(I) \cap C(\bar{I})$, and moreover, w is differentiable at β with $w'(\beta) = -u'(\beta) = 0$. Furthermore, w satisfies the inequality

$$\tilde{L}w = w'' + \tilde{p}(x)w' + \tilde{q}(x)w \le 0, \quad x \in I,$$

where

$$\tilde{p}(x) = -p(2\beta - x) \quad \text{and} \quad \tilde{q}(x) = q(2\beta - x), \quad x \in I.$$

Finally we also note that w has a nonpositive minimum at β. Therefore, by the above result, we must have $w(y) = w(\beta)$ for all $y \in [\beta, 2\beta - \alpha]$. Thus for any $x \in [\alpha, \beta]$, we have $w(\beta) = w(2\beta - x) = u(x)$, that is, $u(x) = u(\beta)$, as was to be shown. □

As consequences of Theorem 2.1, we have the following immediate and useful theorems on maximum principles.

Theorem 2.2. *Let $p, q \in C(\bar{J})$ and $q \leq 0$ in J. Suppose u satisfies the differential inequality $Lu \leq 0$ in an interval J. If u assumes a nonpositive minimum value at an interior point x_0 of J, then $u(x) \equiv u(x_0)$.*

Proof. Suppose u attains its nonpositive minimum at $x_0 \in J = (\alpha, \beta)$. Then $u'(x_0) = 0$. Consider the intervals $[\alpha, x_0]$ and $[x_0, \beta]$. By Theorem 2.1, we see that $u(x) = u(x_0)$ for all $x \in [\alpha, x_0]$ and $u(x_0) = u(x)$ for all $x \in [x_0, \beta]$. That is, $u(x) = u(x_0)$ for all $x \in [\alpha, \beta]$. □

Theorem 2.3. *Let $p, q \in C(\bar{J})$ and $q \leq 0$ in J. Suppose $u \in C^2(J) \cap C(\bar{J})$ satisfies the differential inequality $Lu \leq 0$ in an interval $J := (\alpha, \beta)$. If u assumes a nonpositive minimum value at $x_0 \in \{\alpha, \beta\}$ and u is differentiable at x_0, then $u'(x_0) > 0$ if $x_0 = \alpha$, and $u'(x_0) < 0$ if $x_0 = \beta$ unless u is a constant on J.*

Proof. Suppose u satisfies $Lu \leq 0$ in J, and u has a nonpositive minimum at $x_0 \in \{\alpha, \beta\}$. By hypothesis, u is differentiable at x_0. Let us take the case $x_0 = \alpha$. Then clearly $u'(\alpha) \geq 0$. If $u'(\alpha) = 0$, then by Theorem 2.1, we conclude u is a constant. Therefore, if u is nonconstant, we must have $u'(\alpha) > 0$. If $x_0 = \beta$, here again we have $u'(\beta) \leq 0$. If u is nonconstant, then again by Theorem 2.1, we must have $u'(\beta) < 0$. □

Theorem 2.4. *Let $p, q \in C(\bar{J})$ and $q \leq 0$ in J. Suppose $u \in C^2(J) \cap C(\bar{J})$ satisfies the differential inequality $Lu \leq 0$ in an interval $J := (\alpha, \beta)$. Suppose $u(\gamma) \leq 0$ for some $\gamma \in \bar{J}$. In case $\gamma \in \{\alpha, \beta\}$, we assume that u is differentiable at γ:*

(i) *If $u'(\gamma) \leq 0$, then $u(x) \leq 0$ for all $x \in [\gamma, \beta]$.*

(ii) *If $u'(\gamma) \geq 0$, then $u(x) \leq 0$ for all $x \in [\alpha, \gamma]$.*

(iii) *If $u'(\gamma) = 0$, then $u(x) \leq 0$ for all $x \in \bar{J}$.*

Proof. Suppose $u'(\gamma) \leq 0$. We assume that $\gamma < \beta$, for otherwise there is nothing to prove. Suppose that $u(c) > 0$ for some $c \in (\gamma, \beta]$. Since $u(\gamma) \leq 0$, and $u(c) > 0$, we note that u has a nonpositive minimum on $[\gamma, c]$ at some $\gamma \leq d < c$. If $\gamma < d < c$, then $u'(d) = 0$ and we invoke Theorem 2.2 to conclude that u is a constant in $[\gamma, c]$. If $d = \gamma$, then the assumption $u'(\gamma) \leq 0$ and Theorem 2.3 lead us to conclude that u is a constant on $[\gamma, c]$. In any case, we see that $u(c) > 0$ for some $c \in (\gamma, \beta]$ implies that u is a constant on $[\gamma, c]$. But then $u(c) = u(\gamma) \leq 0$, which contradicts the assumption that $u(c) > 0$. This proves statement (i).

To prove (ii), let us assume that $u'(\gamma) \geq 0$, and that $\gamma > \alpha$. Assume that $u(c) > 0$ for some $c \in [\alpha, \gamma)$. Since $u(\gamma) \leq 0$, as in the previous case we note that u takes a nonpositive minimum on $[c, \gamma]$ at some $c < d \leq \gamma$. If $c < d < \gamma$, then $u'(d) = 0$, and by Theorem 2.2, we see that u is a constant on $[c, \gamma]$. If, on the other hand, $d = \gamma$, then since $u'(\gamma) \geq 0$, we conclude that u is a constant on $[c, \gamma]$ by Theorem 2.3. In either case, we conclude that u is a constant on $[c, \gamma]$. But this implies that $u(c) = u(\gamma) \leq 0$, which again contradicts the assumption that $u(c) > 0$. Therefore statement (ii) holds as well.

Finally statement (iii) follows from statements (i) and (ii). □

3. The Harnack and Harnack-type inequalities

We start with following existence and uniqueness theorem for solutions of $Lu = f$ that satisfy initial conditions. This theorem is usually taught in a first course on ordinary differential equations in undergraduate curriculum (see [Boyce and DiPrima 1965] for instance), and will be needed in our proof of Harnack's inequality.

Theorem E (existence and uniqueness). *Suppose $p, q, f \in C(I)$. Let $x_0 \in I$ and let c_0 and c_1 be arbitrary real constants. Then there exists a unique solution $u \in C^2(I)$ of equation $Lu = f$ such that $u(x_0) = c_0$ and $u'(x_0) = c_1$.*

We now begin our considerations of Harnack's inequality with respect to the class of nonnegative solutions of the differential inequality

$$Lu \leq 0 \quad \text{in } I := (A, B). \tag{3-1}$$

To proceed further, we fix some notations, some of which are fairly standard. For any function $h : (A, B) \to \mathbb{R}$, we write

$$h^+(x) := \max\{h(x), 0\} \quad \text{and} \quad h^-(x) := \max\{-h(x), 0\}, \quad x \in (A, B).$$

Note that we have

$$h = h^+ - h^-.$$

In the sequel, we will also use the following notation repeatedly.

$$L_0 u := u'' + p(x)u' - q^-(x)u, \quad x \in I.$$

Remark 3.1. We first make note of the following:

(1) If u is a nonnegative solution of (3-1), then u is a solution of $L_0 u \leq 0$ in I.

(2) If u is a nonnegative solution of (3-1) with $u \not\equiv 0$ on I, then $u > 0$ in I, for if $u(x_0) = 0$ for some $x_0 \in I$, then $u'(x_0) = 0$. Since $L_0 u \leq 0$ in I, we invoke Theorem 2.2 and conclude that $u(x) \equiv 0$ in I.

We start with the following theorem on Harnack's inequality for nonnegative solutions of (3-1).

Theorem 3.2. *Given $[a, b] \subseteq I$, there is a positive constant C that depends on the coefficients p, q and the constants A, B, a and b only, such that*

$$\max_{a \leq x \leq b} u(x) \leq C \min_{a \leq x \leq b} u(x) \tag{3-2}$$

for any nonnegative solution u of (3-1).

We break down the proof into two lemmas, each of which may be of independent interest. The proof closely follows the method in [Berhanu and Mohammed 2005].

Lemma 3.3. *Given $[a, b] \subseteq I$, there are constants C_a and C_b that depend on the coefficients p, q and the constants A, B, a and b only, such that*

$$\frac{u'(a)}{u(a)} \le C_a \quad \textit{and} \quad \frac{u'(b)}{u(b)} \ge C_b \tag{3-3}$$

for all positive solutions u of (3-1).

Proof. Let w_1 and w_2 be solutions of

$$L_0 w := w'' + p(x)w' - q^-(x)w = 0, \quad A < x < B,$$

such that $w_1(a) = 1$, $w_1'(a) = 0$, and $w_2(a) = 0$, $w_2'(a) = 1$.

We define

$$v(x) := u(x) - u(a)w_1(x) - u'(a)w_2(x), \quad A < x < B.$$

Then recalling that $L_0 u \le 0$, and $L_0 w_1 = 0 = L_0 w_2$ in (A, B), we see that $L_0 v \le 0$ in (A, B). Moreover, we have $v(a) = 0$ and $v'(a) = 0$. By Theorem 2.4, we conclude that $v \le 0$ on (A, B). Thus

$$0 \le u(x) \le u(a)w_1(x) + u'(a)w_2(x), \quad A < x < B,$$

whence

$$\frac{u'(a)}{u(a)} w_2(x) + w_1(x) \ge 0, \quad A < x < B. \tag{3-4}$$

Since $w_2'(a) = 1$, we note that there is a small interval centered at a on which w_2 is increasing. So we fix a^* with $\alpha < a^* < a$ such that $w_2(a^*) < 0$. Therefore, on taking $x = a^*$ in (3-4), we conclude that

$$\frac{u'(a)}{u(a)} \le -\frac{w_1(a^*)}{w_2(a^*)} = C_a. \tag{3-5}$$

Next we establish the second estimate in (3-3). This is very similar to the previous case, and hence we will be brief. Let z_1 and z_2 be solutions of

$$L_0 z_1 = 0, \quad z_1(b) = 1, \ z_1'(b) = 0 \qquad \text{and} \qquad L_0 z_2 = 0, \quad z_2(b) = 0, \ z_2'(b) = 1.$$

Let us consider the function

$$v(x) := u(x) - u(b)z_1(x) - u'(b)z_2(x), \quad A < x < B.$$

Then $Lv \le 0$ in (A, B) and $v(b) = 0$, $v'(b) = 0$. Arguing as before, we can show that $v \le 0$ on (A, B), from which we conclude

$$\frac{u'(b)}{u(b)} z_2(x) + z_1(x) \ge 0, \quad A < x < B.$$

Since z_2 is increasing in some interval centered at b, we can find $b < b^* < B$ such that $z_2(b^*) > 0$. Thus we find that

$$\frac{u'(b)}{u(b)} \geq -\frac{z_1(b^*)}{z_2(b^*)} = C_b. \tag{3-6}$$

This completes the proof of the lemma. □

Lemma 3.4. *Given $[a, b] \subseteq I$, there is a positive constant C, depending on the coefficients p, q and the constants A, B, a and b only, such that*

$$|u'(x)| \leq Cu(x), \quad a \leq x \leq b \tag{3-7}$$

for all nonnegative solutions u of (3-1) in I.

Proof. Let u be a nonnegative solution of (3-1) in I with $u \not\equiv 0$ so that $u > 0$ in I. Direct computation shows that

$$\begin{aligned}\Big(\frac{u'}{u}\Big)' &= \frac{u''}{u} - \Big(\frac{u'}{u}\Big)^2 \\ &\leq \frac{1}{u}\big(-p(x)u' - q(x)u\big) \\ &= -p(x)\Big(\frac{u'}{u}\Big) - q(x).\end{aligned}$$

Therefore,

$$\Big(\frac{u'}{u}\Big)' + p(x)\Big(\frac{u'}{u}\Big) \leq q^-(x).$$

This leads to

$$\bigg(\exp\bigg(\int_a^x p(t)\,dt\bigg)\frac{u'}{u}\bigg)' \leq q^- \exp\bigg(\int_a^x p(t)\,dt\bigg).$$

This gives

$$\bigg(r(x)\frac{u'}{u} - \int_a^x r(t)\,q^-(t)\,dt\bigg)' \leq 0, \quad x \in (a, b),$$

where we have set

$$r(x) := \exp\bigg(\int_a^x p(t)\,dt\bigg).$$

Thus for any $a \leq x \leq b$, we have

$$r(b)\frac{u'(b)}{u(b)} - \int_a^b r(t)\,q^-(t)\,dt \leq r(x)\frac{u'}{u} - \int_a^x r(t)\,q^-(t)\,dt \leq r(a)\frac{u'(a)}{u(a)}.$$

In conclusion, we have

$$r(b)\frac{u'(b)}{u(b)} - Q(b) \leq r(x)\frac{u'}{u} \leq \frac{u'(a)}{u(a)} + Q(b), \quad x \in (a, b), \tag{3-8}$$

where $Q(b)$ denotes the constant

$$Q(b) := \int_a^b r(t)q^-(t)\,dt.$$

Using Lemma 3.3 in (3-8), we obtain

$$r(x)\left|\frac{u'(x)}{u(x)}\right| \le C_0, \quad x \in [a,b],$$

for some positive constant C_0, independent of u. Since

$$\frac{1}{r(x)} = \exp\biggl(-\int_a^x p(t)\,dt\biggr) \le \exp\bigl(\|p\|_\infty(b-a)\bigr), \quad x \in [a,b],$$

we conclude that

$$\left|\frac{u'(x)}{u(x)}\right| \le C, \quad x \in [a,b], \tag{3-9}$$

for a constant $C > 0$ that is independent of u. □

Proof of Theorem 3.2. For any $x, y \in [a,b]$, we see that

$$\begin{aligned}\log\biggl(\frac{u(x)}{u(y)}\biggr) &= \int_y^x \frac{d}{dt}\log u(t)\,dt\\ &= \int_y^x \frac{u'(t)}{u(t)}\,dt.\end{aligned}$$

Therefore,

$$\frac{u(x)}{u(y)} = \exp\biggl(\int_y^x \frac{u'(t)}{u(t)}\,dt\biggr).$$

It follows from this and Lemma 3.4 that

$$\exp(-C|x-y|) \le \frac{u(x)}{u(y)} \le \exp(C|x-y|), \quad x, y \in [a,b].$$

Therefore, we finally see that

$$\exp(-C(b-a)) \le \frac{u(x)}{u(y)} \le \exp(C(b-a)), \quad x, y \in [a,b],$$

which leads to the inequality stated in (3-2). □

Remark 3.5. The differential inequality (3-1) with the inequality reversed doesn't satisfy Harnack's inequality as can be seen from the following simple example. Fix $x_0 \in [a,b]$. Then $u_k(x) = e^{k(x-x_0)}$ satisfies the inequality $u'' \ge 0$ in $\mathbb{R}$, and note that

$$e^{k(b-x_0)} \le \sup_{[a,b]} u_k(x) \le C \inf_{[a,b]} u_k(x) \le C u_k(x_0) = C.$$

But there is no single positive constant C, independent of u_k and hence k, such that

$$e^{k(b-x_0)} \le C.$$

Next we study a Harnack-type inequality for nonnegative solutions of nonhomogeneous equations.

We will start by deriving a Harnack-type inequality for nonnegative solutions of the following equation, assuming that $f \geq 0$ on I:

$$L_0 u = f \quad \text{in } I := (A, B). \tag{3-10}$$

However, it should be noted that Harnack's inequality (3-2) does not hold for nonnegative solutions of (3-10) for general f. This is to be expected as nonnegative solutions of (3-10) are not necessarily positive in (A, B). In fact, the following simple example shows that the inequality (3-2) cannot hold even for positive solutions of (3-10).

Example 3.6. Consider the equation $u'' = 1$ in the interval $(A, B) := (-2, 2)$. For any positive integer k,

$$u_k = \frac{1}{2}\left(x - \frac{1}{k}\right)^2 + \frac{1}{k}$$

is a solution of $u_k'' = 1$, and $u_k > 0$ in $(-2, 2)$ for all k. Suppose there is a constant $C > 0$ such that

$$\max_{[-1,1]} u \leq C \min_{[-1,1]} u \quad \forall u > 0, u'' = 1. \tag{3-11}$$

Then note that

$$u_k(1) = \frac{1}{2}\left(1 - \frac{1}{k}\right)^2 + \frac{1}{k} \quad \text{and} \quad u_k\left(\frac{1}{k}\right) = \frac{1}{k}.$$

If (3-11) were to hold, then

$$\frac{1}{2}\left(1 - \frac{1}{k}\right)^2 + \frac{1}{k} \leq C\left(\frac{1}{k}\right) \quad \forall k = 1, 2, \ldots.$$

Letting $k \to \infty$, we arrive at a contradiction.

We now state the following theorem on a Harnack-type inequality for solutions of (3-10). For the remainder of our discussion, we will use the following notations.

$$\alpha := \tfrac{1}{2}(a + A) \quad \text{and} \quad \beta := \tfrac{1}{2}(b + B).$$

We will also find it convenient to use the notation $\|g\|_\infty$ to denote the following number for any g bounded on an interval I:

$$\|g\|_{I,\infty} := \sup_{x \in I} |g(x)|,$$

or simply $\|g\|_\infty$ if I is clear from the context.

Theorem 3.7. *Suppose $f \geq 0$ in I. Given $[a, b] \subseteq I$, there is a positive constant C that depends on the coefficients p, q and the constants A, B, a and b such that*

$$\max_{a \leq x \leq b} u(x) \leq C\left(\min_{a \leq x \leq b} u(x) + \int_{\alpha}^{\beta} f(x)\, dx\right) \tag{3-12}$$

for all nonnegative solutions u of (3-10).

Proof. We prove the theorem in three steps. Suppose $u \geq 0$ in (A, B) is a solution of (3-10).

Step 1. Let $u(x_0) = \min\{u(x) : x \in [\alpha, \beta]\}$. By Theorem E, we pick $z_* \in C^2(I) \cap C(\bar{I})$ such that

$$L_0 z_* = f, \quad z_*(x_0) = 0 = z_*'(x_0). \tag{3-13}$$

By Theorem 2.4(iii), we note that $z_* \geq 0$ in (A, B). We claim that $u \geq z_*$ in $[\alpha, \beta]$. To see this, we start by observing that

$$L_0(u - z_*) = 0 \quad \text{in } (A, B) \qquad \text{and} \qquad (u - z_*)(x_0) \geq 0.$$

Suppose first that $\alpha < x_0 < \beta$. Then $u'(x_0) = 0$, and therefore $(u - z_*)'(x_0) = 0$. Consequently, by Theorem 2.4(iii), we conclude that $u - z_* \geq 0$ in $[\alpha, \beta]$, as desired. Suppose $x_0 = \alpha$. Then $u'(x_0) = u'(\alpha) \geq 0$, so that $(u - z_*)'(x_0) \geq 0$. By Theorem 2.4(i), we conclude that $u - z_* \geq 0$ in $[x_0, \beta] = [\alpha, \beta]$. Finally, suppose that $x_0 = \beta$. Then $u'(x_0) = u'(\beta) \leq 0$, so that $(u - z_*)'(x_0) \leq 0$. Again, by Theorem 2.4(ii), we conclude that $u - z_* \geq 0$ in $[\alpha, x_0] = [\alpha, \beta]$. Thus, in all cases, we have shown that $u \geq z_*$ in $[\alpha, \beta]$ as claimed.

Step 2. Let $u(\zeta) := \min\{u(x) : a \leq x \leq b\}$. Since $u - z_*$ is a nonnegative solution of $L_0 w = 0$ in (α, β), we invoke Theorem 3.2 to obtain a positive constant C that depends on p, q^- and the constants A, B, a and b only such that the following chain of inequalities hold:

$$\begin{aligned}
\max_{[a,b]} u &= \max_{[a,b]}(z_* + u - z_*) \\
&\leq \max_{[a,b]} z_* + \max_{[a,b]}(u - z_*) \\
&\leq \max_{[a,b]} z_* + C \min_{[a,b]}(u - z_*) \quad (\text{by Theorem 3.2}) \\
&\leq C(u - z_*)(\zeta) + \max_{[a,b]} z_* \\
&\leq Cu(\zeta) + \max_{[a,b]} z_* \quad (\text{recall that } u(\zeta) = \min_{[a,b]} u) \\
&= C \min_{[a,b]} u + \max_{[a,b]} z_*.
\end{aligned} \tag{3-14}$$

Step 3. We now estimate z_* on $[a, b]$. Recall the notation $\|g\|_\infty := \max_{x\in[\alpha,\beta]} |g(x)|$ for any function $g \in C([\alpha, \beta])$. We recall that

$$f = L_0 z_* = \frac{1}{r(x)}(r(x)z_*')' - q^-(x)z_*, \quad x \in I,$$

where

$$r(x) = \exp\biggl(\int_a^x p(s)\,ds\biggr).$$

For $x \in (x_0, b)$, we have

$$z_*(x) = \int_{x_0}^x \frac{1}{r(t)} \int_{x_0}^t r(s)\bigl(q^-(s)z_*(s) + f(s)\bigr)\,ds\,dt.$$

Therefore, for $x \in (x_0, b)$,

$$\begin{aligned} z_*(x) &\le \exp\bigl((b-a)\|p\|_\infty\bigr) \int_{x_0}^x \int_{x_0}^t \bigl(q^-(s)z_*(s) + f(s)\bigr)\,ds\,dt \\ &\le (b-a)\exp\bigl((b-a)\|p\|_\infty\bigr) \int_{x_0}^x \bigl(q^-(t)z_*(t) + f(t)\bigr)\,dt \\ &\le P_0 \int_\alpha^\beta f(t)\,dt + P_0\|q^-\|_\infty \int_{x_0}^x z_*(t)\,dt, \end{aligned}$$

where $P_0 := (b-a)\exp\bigl((b-a)\|p\|_\infty\bigr)$.

Denoting the right-hand side of the last inequality by $\vartheta(x)$ for $x_0 < x < b$, and on noting that $z_*(x) \le \vartheta(x)$ on (x_0, b), we find

$$\begin{aligned} \vartheta'(x) &= P_0\|q^-\|_\infty z_*(x) \\ &\le P_0\|q^-\|_\infty \vartheta(x) \quad (\text{since } z_*(x) \le \vartheta(x)), \quad x \in (x_0, b), \end{aligned}$$

so that

$$\frac{\vartheta'(x)}{\vartheta(x)} \le P_0\|q^-\|_\infty, \quad x \in (x_0, b).$$

Integrating on (x_0, x), we find that

$$z_*(x) \le \vartheta(x) \le P_0 \exp\bigl(P_0\|q^-\|_\infty(b-a)\bigr) \int_\alpha^\beta f(t)\,dt. \tag{3-15}$$

The same inequality holds if $a < x < x_0$.

Using (3-15) in (3-14) leads to the desired inequality (3-12). □

Finally we are ready to state and prove the following Harnack-type inequality for nonnegative solutions of the differential inequality (1-1) with the nonhomogeneous term f in $C(I)$, without any sign restrictions.

Theorem 3.8. *Given $[a,b]\subseteq I$, there is a positive constant C, that depends on the coefficients p, q and the constants A, B, a and b only such that the Harnack-type inequality (3-12), with f replaced by f^+, holds for all nonnegative solutions of (1-1).*

Proof. Let u be a nonnegative solution of (1-1) in $(A,\,B)$. Let $u(x_0)=\min_{[a,b]} u$, and consider the solution z of

$$L_0 z = f^+ \quad \text{in } (A,\,B) \quad \text{and} \quad z(x_0)=u(x_0),\; z'(x_0)=u'(x_0).$$

Then $L_0(u-z)=L_0u-L_0z\le f-f^+\le 0$, and $(u-z)(x_0)=0$ and $(u-z)'(x_0)=0$. By Theorem 2.4(iii), we conclude that $u-z\le 0$ in $(A,\,B)$, so that $0\le u\le z$ in $(A,\,B)$. Thus

$$\begin{aligned}\max_{x\in[a,b]} u(x) &\le \max_{x\in[a,b]} z(x)\\ &\le C\biggl(\min_{x\in[a,b]} z(x)+\int_\alpha^\beta f^+(x)\,dx\biggr) \quad \text{(by Theorem 3.7)}\\ &\le C\biggl(z(x_0)+\int_\alpha^\beta f^+(x)\,dx\biggr)\\ &= C\biggl(u(x_0)+\int_\alpha^\beta f^+(x)\,dx\biggr)\\ &= C\biggl(\min_{x\in[a,b]} u(x)+\int_\alpha^\beta f^+(x)\,dx\biggr).\end{aligned}$$

This is the desired result. □

References

[Berhanu and Mohammed 2005] S. Berhanu and A. Mohammed, "A Harnack inequality for ordinary differential equations", *Amer. Math. Monthly* **112**:1 (2005), 32–41. MR 2005i:34039 Zbl 1127.34001

[Boyce and DiPrima 1965] W. E. Boyce and R. C. DiPrima, *Elementary differential equations and boundary value problems*, Wiley, New York, 1965. MR 31 #3651 Zbl 0128.30601

[Kassmann 2007] M. Kassmann, "Harnack inequalities: an introduction", *Bound. Value Probl.* (2007), Art. ID 81415. MR 2007j:35001 Zbl 1144.35002

[Protter and Weinberger 1984] M. H. Protter and H. F. Weinberger, *Maximum principles in differential equations*, Springer, New York, 1984. MR 86f:35034 Zbl 0549.35002

Received: 2014-10-28 Revised: 2015-02-26 Accepted: 2015-03-04

amohammed@bsu.edu
Department of Mathematical Sciences,
Ball State University, Muncie, IN 47306, United States

hannahturner@math.utexas.edu
Department of Mathematical Sciences,
Ball State University, Muncie, IN 47306, United States

INVOLVE 9:2 (2016)
dx.doi.org/10.2140/involve.2016.9.293

The isoperimetric and Kazhdan constants associated to a Paley graph

Kevin Cramer, Mike Krebs, Nicole Shabazi,
Anthony Shaheen and Edward Voskanian

(Communicated by Kenneth S. Berenhaut)

In this paper, we investigate the isoperimetric constant (or expansion constant) of a Paley graph, and the Kazhdan constant of the group and generating set associated with a Paley graph.

We give two new upper bounds for the isoperimetric constant $h(X_p)$ for the Paley graph X_p. These bounds improve previously known eigenvalue bounds on $h(X_p)$. Along with a known eigenvalue lower bound for $h(X_p)$, they provide a narrow strip in which $h(X_p)$ must live. More precisely, we show that $(p-\sqrt{p})/4 \le h(X_p) \le (p-1)/4$, which implies that $\lim_{p\to\infty} h(X_p)/p = 1/4$.

In addition, we show that the Kazhdan constant associated with the integers modulo p and the generating set for the Paley graph X_p approaches 2 as p tends to infinity, which is the best possible limit that the Kazhdan constant can be.

1. Introduction

Paley graphs are interesting because they allow one to use graph-theoretic tools to study the theory of quadratic residues. They also have interesting properties that make them useful in graph theory. For example, they are strongly regular, self-complementary, and their eigenvalues are essentially Gauss sums.

Let p be an odd prime with $p \equiv 1 \pmod 4$. The Paley graph X_p is constructed as follows. The vertices of X_p consist of the integers modulo p, which we denote by $\mathbb{Z}_p$. Two vertices x and y from $\mathbb{Z}_p$ are adjacent if and only if $x-y$ is an element of $\Gamma_p = \{\gamma^2 \mid \gamma \in \mathbb{Z}_p \text{ and } \gamma \neq 0\}$. It is well known that -1 is in Γ_p since $p \equiv 1 \pmod 4$. Hence the above definition is well-defined; that is, $x-y$ is in Γ_p if and only if $y-x$ is in Γ_p.

For example, if $p = 13$, then $\Gamma_{13} = \{1, 3, 4, 9, 10, 12\}$. Then 1 and 10 are adjacent since $10-1=9$, which is in Γ_{13}. A picture of X_{13} is given in Figure 1.

MSC2010: 05C99.
Keywords: isoperimetric constant, expansion constant, Paley graph, Kazhdan constant.
This research was partially supported by NSF grant DMS-1247679.

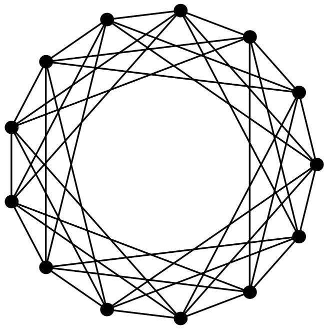


Figure 1. The Paley graph on $\mathbb{Z}_{13}$.

A great reference for Paley graphs is [Elsawy 2009]. Note that one can define a Paley graph on a finite field of size p^n. However, we are sticking with $n = 1$ in this paper.

We want to get approximations on two constants associated with the Paley graph: the isoperimetric constant and the Kazhdan constant. We first introduce the isoperimetric constant.

Let X be a graph with vertex set V. Let F be a subset of V. The boundary of F, denoted by ∂F, consists of the edges of X with one end in F and the other end in $V \setminus F$. The isoperimetric constant of X is defined to be

$$h(X) = \min\left\{ \frac{|\partial F|}{|F|} \;\middle|\; F \subseteq V \text{ and } |F| \leq \frac{|V|}{2} \right\}.$$

In layman's terms, the isoperimetric constant of a graph X gives a rough estimate for how "good" a graph is as a communications network. It has been heavily studied by both computer scientists and mathematicians. One main topic of investigation in this area is that of expander families. A family of finite regular graphs, each with the same degree, whose order is unbounded, is said to be an expander family if there is a uniform positive lower bound for $h(X)$ for all X in the family. Recently it has been shown that every family of finite nonabelian simple groups yields an expander family via the Cayley graph construction. This was proven for all families except Suzuki groups by Kassabov, Lubotzky, and Nikolov [Kassabov et al. 2006], with the final case of Suzuki groups proven by Breuillard, Green, and Tao [Breuillard et al. 2011].

In general it is a difficult combinatorial problem to get an exact value for the isoperimetric constant of a graph. Some examples where the isoperimetric constant of a graph family is known are as follows. The isoperimetric constant for cycle graphs of order n is equal to $4/n$ when n is even and $4/(n-1)$ when n is odd. The isoperimetric constant of a complete graph of order n is equal to $n/2$ when n is even and $(n+1)/2$ when n is odd. See [Krebs and Shaheen 2011] for proofs. Rosenhouse

[2002] shows that $h(X_n) = 4/n$, where X_n is the Cayley graph constructed using the dihedral group D_{2n} with generators r, r^{-1}, and s. Lanphier and Rosenhouse [2004] derive approximations on the isoperimetric constants of Platonic graphs.

Instead of calculating $h(X)$ exactly, one must usually be satisfied with approximations. One way to approximate $h(X)$ is to use the eigenvalues of X. The eigenvalues of X are especially useful in finding a lower bound on $h(X)$.

Let $\lambda_1(X)$ be the second largest eigenvalue of a d-regular graph. A well-known inequality is

$$\frac{d-\lambda_1(X)}{2} \le h(X) \le \sqrt{2d(d-\lambda_1(X))} \tag{1}$$

(see [Krebs and Shaheen 2011, p. 31]). One also has a tighter upper bound on $h(X)$ given by Mohar [1989]. It is

$$h(X) \le \sqrt{(d+\lambda_1(X))(d-\lambda_1(X))}. \tag{2}$$

Let us see what (1) and (2) tell us about Paley graphs. Since Paley graphs are strongly regular graphs, one can find a quadratic polynomial that the adjacency matrix satisfies. This leads one to the eigenvalues of a Paley graph. (For the details, see [Gross et al. 2014, pp. 684–685]). The eigenvalues of X_p are $(p-1)/2$ with multiplicity 1, $\sqrt{p}/2 - 1/2$ with multiplicity $(p-1)/2$ and $-\sqrt{p}/2 - 1/2$ with multiplicity $(p-1)/2$. Thus $\lambda_1(X_p) = \sqrt{p}/2 - 1/2$. Plugging this into (1) and using the fact that X_p is $(p-1)/2$ regular, we get that

$$\frac{p-\sqrt{p}}{4} \le h(X_p) \le \sqrt{\frac{p^2 - p\sqrt{p} - p + \sqrt{p}}{2}}. \tag{3}$$

Using the Mohar bound, given in (2), one gets

$$h(X_p) \le \frac{\sqrt{p^2 - 3p + 2\sqrt{p}}}{2}, \tag{4}$$

which reduces the upper bound in (3) by a factor of $\sqrt{2}$ as p tends to infinity.

The lower bound in (3) seems optimal for Paley graphs. However, the two upper bounds given above are far from optimal. In this paper, we will give two new upper bounds. One we call the α-bound, which is the average of the first half of the elements of Γ_p, and the other is the simpler bound of $(p-1)/4$. Both of these bounds give much better upper bounds than the eigenvalue bounds given above in (3) and (4). Consider Table 1. Note how close the eigenvalue lower bound is to both the α-bound and $(p-1)/4$, and how much better the two new upper bounds are. While $h(X_p)$ is still not known exactly, we have found a very narrow band in which it must exist. For example, we have $2{,}168{,}090 \le h(X_p) \le 2{,}168{,}277$ when $p = 8{,}675{,}309$.

prime p	13	577	40,961	8,675,309
eigenvalue lower bound from (3)	2.35	138.24	10,189	2,168,090
α-bound (new upper bound)	2.67	139.29	10,201	2,168,277
$(p-1)/4$ (new upper bound)	3	144	10,240	2,168,827
eigenvalue upper bound from (4)	5.86	287.77	20,479	4,337,654
eigenvalue upper bound from (3)	7.51	399.07	28,891	6,133,328

Table 1. Lower and upper bounds for $h(X_p)$.

Summarizing the above, we have our main result for the isoperimetric constant of a Paley graph.

Theorem 1. *Let p be an odd prime with $p \equiv 1 \pmod 4$. Then*

$$\frac{p-\sqrt{p}}{4} \leq h(X_p) \leq \frac{p-1}{4}.$$

Note that inequalities (3) and (4) show that

$$\frac{1}{4} \leq \liminf_{p\to\infty} \frac{h(X_p)}{p} \leq \limsup_{p\to\infty} \frac{h(X_p)}{p} \leq \frac{1}{2}.$$

Theorem 1, however, shows more precisely that

$$\lim_{p\to\infty} \frac{h(X_p)}{p} = \frac{1}{4}.$$

Before moving on, we would like to note that we do not know if the α-bound is always smaller than $(p-1)/4$, but from calculations it appears to be so.

The second result of this paper concerns the Kazhdan constant of the pair $(\mathbb{Z}_p, \Gamma_p)$ associated with the Paley graph X_p. We begin by giving the general definition of a Kazhdan constant for any finite group. The definition greatly simplifies when the group is the integers modulo p. The reader who has never encountered representation theory may skim the next paragraph to get the idea with no loss of understanding.

Let G be a finite group, and let Γ be a nonempty subset of G. Let ρ be a unitary representation of G acting on some vector space V_ρ. We define

$$\kappa(G, \Gamma, \rho) = \min_{\substack{\|v\|=1 \\ v\in V_\rho}} \max_{\gamma\in\Gamma} \|\rho(\gamma)v - v\|.$$

The Kazhdan constant of the pair (G, Γ) is defined to be

$$\kappa(G, \Gamma) = \min_{\rho}\{\kappa(G, \Gamma, \rho)\},$$

where the minimum is over all irreducible, nontrivial, unitary representations ρ of G. Question: why is one interested in computing such a constant? One answer is because, when Γ is a symmetric subset of G, we know that $\kappa(G, \Gamma)$ is related to the isoperimetric constant of the Caley graph built from G and Γ. More specifically, suppose that Γ is a symmetric subset of the group G. That is, $\gamma \in \Gamma$ if and only if $\gamma^{-1} \in \Gamma$. Then one can build the Caley graph $X = \mathrm{Cay}(G, \Gamma)$, where the vertices of X are the elements of G and $x, y \in G$ are adjacent if and only if $y^{-1}x \in \Gamma$. (Note that if $G = \mathbb{Z}_p$, then $\Gamma = \Gamma_p$ gives the Paley graph.) Here X is a regular graph of degree $d = |\Gamma|$. In this case, we have the relationship $h(X) \geq \kappa(G, \Gamma)^2/4d$. Hence, by finding lower bounds on $\kappa(G, \Gamma)$, one can find lower bounds on $h(X)$. For more information on the above discussion, see [Krebs and Shaheen 2011, Chapter 8].

We would like to note that it is difficult to calculate $\kappa(G, \Gamma)$ in general. There are very few results in this area. As an example, Bacher and de la Harpe [1994] calculate $\kappa(\mathbb{Z}_n, \Gamma)$ for several very specific sets Γ, such as $\kappa(D_{2n}, \{r, s\})$, where D_{2n} is the dihedral group and r and s are its generators, and $\kappa(S_n, \Gamma_n)$, where $\Gamma_n = \{(1, 2), \dots, (n-1, n)\}$.

We are interested in approximating the Kazhdan constant of the pair $(\mathbb{Z}_p, \Gamma_p)$. When $G = \mathbb{Z}_p$, the Kazhdan constant simplifies considerably. To simplify our notation, we set $\xi_p = e^{2\pi i/p}$. The irreducible, nontrivial, unitary representations of $\mathbb{Z}_p$ are given by the maps $\rho_a(\gamma) : \mathbb{C} \to \mathbb{C}$, where $\rho_a(\gamma)z = \xi_p^{a\gamma} z$, $V_a = \mathbb{C}$, and $a = 1, 2, \dots, p-1$. Hence,

$$\begin{aligned}\kappa(\mathbb{Z}_p, \Gamma_p) &= \min_{1 \leq a \leq p-1} \min_{\substack{\|v\|=1 \\ v \in \mathbb{C}}} \max_{\gamma \in \Gamma} \|\xi_p^{a\gamma} v - v\| \\ &= \min_{1 \leq a \leq p-1} \max_{\gamma \in \Gamma} \|\xi_p^{a\gamma} - 1\|.\end{aligned}$$

In words, $\kappa(\mathbb{Z}_p, \Gamma_p)$ is calculated by considering each a and finding the γ for which $\xi_p^{a\gamma}$ is the maximal distance away from 1, and then one finds the minimum of these maximums. Or another way of saying it is that for any $1 \leq a \leq p-1$, there exists a $\gamma \in \Gamma$ such that $\|\xi_p^{a\gamma} - 1\| \geq \kappa(\mathbb{Z}_p, \Gamma_p)$.

Let $\mathbb{Z}_p^\times = \mathbb{Z}_p \setminus \{0\}$ and $\overline{\Gamma}_p = \mathbb{Z}_p \setminus (\Gamma_p \cup \{0\})$. If $a \in \mathbb{Z}_p^\times$ then it is easy to show that $a\Gamma_p = \Gamma_p$ if $a \in \Gamma_p$; otherwise, $a\Gamma_p = \overline{\Gamma}_p$ if $a \in \overline{\Gamma}_p$. (To see this note that Γ_p is a subgroup of $\mathbb{Z}_p^\times$ under multiplication and there are only two cosets: Γ_p and $\overline{\Gamma}_p$.) Hence

$$\begin{aligned}\kappa(\mathbb{Z}_p, \Gamma_p) &= \min_{1 \leq a \leq p-1} \max_{\gamma \in \Gamma_p} \|\xi_p^{a\gamma} - 1\| \\ &= \min\Bigl\{\max_{\gamma \in \Gamma_p} \|\xi_p^{\gamma} - 1\|, \max_{\gamma \in \overline{\Gamma}_p} \|\xi_p^{\gamma} - 1\|\Bigr\}.\end{aligned}$$

Thus to calculate $\kappa(\mathbb{Z}_p, \Gamma_p)$, one must find the square $\gamma_1 \in \mathbb{Z}_p$, where $\xi_p^{\gamma_1}$ is as far away from 1 as possible, and the nonsquare $\gamma_2 \in \mathbb{Z}_p$, where $\xi_p^{\gamma_2}$ is as far away from 1 as possible. Then one calculates the minimum of those two distances. For example, when $p = 17$, we have that $\Gamma_{17} = \{1, 2, 4, 8, 9, 13, 15, 16\}$ and $\overline{\Gamma}_{17} = \{3, 5, 6, 7, 10, 11, 12, 14\}$; see Figure 2. We have labeled the elements ξ^γ by squares when γ is in Γ_{17} and by circles when γ is in $\overline{\Gamma}_{17}$. The element ξ^γ where

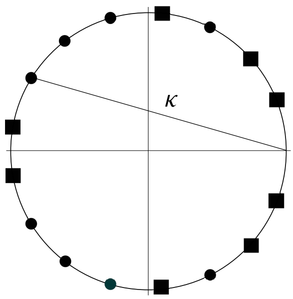


Figure 2. The Kazhdan constant $\kappa = \kappa(\mathbb{Z}_{17}, \Gamma_{17})$.

γ is a square that is furthest from 1 is ξ^8. The element ξ^γ where γ is a nonsquare that is furthest from 1 is ξ^7. Therefore, $\kappa(\mathbb{Z}_{17}, \Gamma_{17}) = \|\xi_{17}^7 - 1\|$.

For specific congruency classes of primes, one can use arguments involving the Legendre symbol to explicitly calculate $\kappa(\mathbb{Z}_p, \Gamma_p)$. For example, arguments from [Voskanian 2013] show that if p is a prime with $p \equiv 17 \pmod{24}$, then

$$\kappa(\mathbb{Z}_p, \Gamma_p) = \|e^{\pi i(1-3/p)} - 1\|.$$

And when $p \equiv 97 \pmod{120}$,

$$\kappa(\mathbb{Z}_p, \Gamma_p) = \|e^{\pi i(1-5/p)} - 1\|.$$

However, it seems that one cannot generalize these arguments to give a formula for $\kappa(\mathbb{Z}_p, \Gamma_p)$ for all $p \equiv 1 \pmod 4$.

Notice that $0 < \kappa(\mathbb{Z}_p, \Gamma_p) < 2$. We will not be able to explicitly calculate $\kappa(\mathbb{Z}_p, \Gamma_p)$; however, we will show the following theorem, which is our main result on the Kazhdan constant of a Paley graph.

Theorem 2. *We have that*

$$\lim_{p\to\infty} \kappa(\mathbb{Z}_p, \Gamma_p) = 2$$

as p goes over the primes which are congruent to 1 *modulo* 4.

2. The isoperimetric constant of a Paley graph

We now give the proofs of the new upper bounds for the isoperimetric constant of X_p that were discussed in the introduction to this paper. We begin with the α-bound and then proceed to the $(p-1)/4$ bound. Note that if $F \subseteq \mathbb{Z}_p$ with $0 < |F| \leq \mathbb{Z}_p/2$ then $h(X_p) \leq |\partial F|/|F|$. This is the technique that we will use in both proofs. That is, we will pick a specific F that will give an upper bound for $h(X_p)$.

2.1. *The α-bound.*

The proof of the α-bound relies on a table that we call the adjacency table for X_p. The adjacency table for X_p is obtained by constructing the group addition table for $\mathbb{Z}_p$ (under the usual addition modulo p) with all the rows corresponding to any $\delta \notin \Gamma_p$ omitted.

For each $\alpha \in \Gamma_p$, we write the additive inverse of α as α^{-1}. Note that $\alpha^{-1} = p - \alpha$, and $|\Gamma_p| = (p-1)/2$; hence we can arrange the elements of Γ_p in increasing order and we will write

$$\Gamma_p = \{\alpha_1, \alpha_2, \ldots, \alpha_k, \alpha_k^{-1}, \ldots, \alpha_2^{-1}, \alpha_1^{-1}\},$$

where $k = (p-1)/4$. Since 1 is the smallest element of Γ_p, we will always have $\alpha_1 = 1$ and $\alpha_1^{-1} = p - 1$. Incorporating these considerations into our construction, we arrive at the following adjacency table:

0	1	2	$\cdots$	$p-1$
1	2	3	$\cdots$	0
α_2	α_2+1	α_2+2	$\cdots$	α_2-1
α_3	α_3+1	α_3+2	$\cdots$	α_3-1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
α_k	α_k+1	α_k+2	$\cdots$	α_k-1
α_k^{-1}	$\alpha_k^{-1}+1$	$\alpha_k^{-1}+2$	$\cdots$	$\alpha_k^{-1}-1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
α_3^{-1}	$\alpha_3^{-1}+1$	$\alpha_3^{-1}+2$	$\cdots$	$\alpha_3^{-1}-1$
α_2^{-1}	$\alpha_2^{-1}+1$	$\alpha_2^{-1}+2$	$\cdots$	$\alpha_2^{-1}-1$
$p-1$	0	1	$\cdots$	$p-2$

For example, when $p = 13$ we have that $\Gamma_{13} = \{1, 3, 4, 9, 10, 12\}$, which gives the following table:

0	1	2	3	4	5	6	7	8	9	10	11	12
1	2	3	4	5	6	7	8	9	10	11	12	0
3	4	5	6	7	8	9	10	11	12	0	1	2
4	5	6	7	8	9	10	11	12	0	1	2	3
9	10	11	12	0	1	2	3	4	5	6	7	8
10	11	12	0	1	2	3	4	5	6	7	8	9
12	0	1	2	3	4	5	6	7	8	9	10	11

To get the α-bound, we will be considering the set $F = \{0, 1, 2, \ldots, (p-3)/2\}$. The following lemma and propositions will be useful when we tally the edges in ∂F row-wise.

Lemma 3. *Let* $\Gamma_p = \{\alpha_1, \alpha_2, \ldots, \alpha_k, \alpha_k^{-1}, \ldots, \alpha_2^{-1}, \alpha_1^{-1}\}$. *Then* $1 \le \alpha_i \le (p-1)/2$ *for all* $i = 1, \ldots, k$.

Proof. We know $\alpha_1 = 1$ is the smallest element in Γ_p, so we have $1 \le \alpha_i$. Now suppose that, for some i, we have $\alpha_i > (p-1)/2$. Since α_i is an integer and p

is odd, the smallest α_i can be is $(p+1)/2$. Thus, we see that $\alpha_i \geq (p+1)/2$. Therefore, it follows that

$$\alpha_i^{-1} = p - \alpha_i \leq p - \frac{p+1}{2} = \frac{p-1}{2}.$$

In this case, we have $\alpha_i \geq (p+1)/2$ and $\alpha_i^{-1} \leq (p-1)/2$. In particular, $\alpha_i^{-1} < \alpha_i$. Since this contradicts the ordering of Γ_p, we see that $\alpha_i \leq (p-1)/2$ for all $i = 1, 2, \ldots, k$. □

Proposition 4. *Let $F = \{0, 1, 2, \ldots, (p-3)/2\}$ be a subset of vertices in X_p. Then row α_i of the adjacency table for X_p contributes exactly α_i edges to the boundary set ∂F.*

Proof. By our choice of F, we only need to scan the entries in row α_i from column 0 to column $(p-3)/2$, and any entry we encounter contributes an edge to ∂F if and only if it is greater than $(p-3)/2$. Since $|F| = (p-1)/2$, there are a total of $(p-1)/2$ columns headed by elements of F, and thus $(p-1)/2$ entries to consider. Also, we recall that for any entry γ in the table, the entry in the same row, one column to the right, is $\gamma + 1$.

Let us tally the contributions made to ∂F by row α_i of the adjacency table. Starting at column 0, we scan row α_i until we arrive at the entry $(p-3)/2$ in some column β. In this case, all the entries encountered so far are less than or equal to $(p-3)/2$, and thus contribute no edges to ∂F. Since $(p-3)/2$ is the entry in row α_i, column β, we have $(p-3)/2 = \alpha_i + \beta$. Thus, $\beta = (p-3)/2 - \alpha_i$. Scanning from column 0 to column β, we have encountered $\beta + 1$ entries. This means there are

$$\frac{p-1}{2} - (\beta + 1) = \frac{p-1}{2} - \left(\frac{p-3}{2} - \alpha_i + 1\right) = \alpha_i$$

entries remaining to consider in the columns headed by the entries from F. These entries increase in unit increments from $(p-3)/2+1$ to $(p-3)/2+\alpha_i$. By Lemma 3, we have that $1 \leq \alpha_i \leq (p-1)/2$. This implies that $(p-3)/2 + \alpha_i \leq p-2$. This means the sequence of remaining entries never reaches p to revert to 0 modulo p. That is, each of the remaining α_i entries is strictly larger than $(p-3)/2$, and thus contributes an edge to ∂F. So row α_i contributes exactly α_i edges to ∂F. □

Proposition 5. *Let $F = \{0, 1, 2, \ldots, (p-3)/2\}$ be a subset of vertices in X_p. Then rows α_i and α_i^{-1} of the adjacency table each contribute the same number of edges to ∂F.*

Proof. By our choice of F, we only need to scan the entries in row α_i^{-1} from column 0 to column $(p-3)/2$, and any entry we encounter contributes an edge to ∂F if and only if it is greater than $(p-3)/2$. This gives a total of $(p-1)/2$ columns to scan through and, thus, $(p-1)/2$ entries to consider.

Noting that the entries increase in unit increments as we scan from left to right, we begin with the entry α_i^{-1} in column 0 and scan to the right until we reach the entry $p-1$ in column β. By Lemma 3, we have $1 \le \alpha_i \le (p-1)/2$, so it follows that $(p+1)/2 \le \alpha_i^{-1} \le p-1$. Thus, we see that every entry encountered so far, of which there are $\beta+1$, is greater than $(p-3)/2$ and contributes an edge to ∂F. Since $p-1$ resides in row α_i^{-1}, column β, we have $p-1 = \alpha_i^{-1} + \beta$, from which it follows that $\beta + 1 = p - \alpha_i^{-1} = \alpha_i$. That is, thus far we have encountered α_i entries in row α_i^{-1} contributing edges to ∂F.

Now, if $\alpha_i = (p-1)/2$, then we must have already scanned through all the necessary columns. This means there are no more entries to consider, and row α_i^{-1} contributes exactly α_i edges to ∂F.

If $1 \le \alpha_i \le (p-3)/2$, then there are $(p-1)/2 - \alpha_i$ entries, ranging from $(p-1)+1 = 0$ in column $\beta+1$ to

$$(p-1) + \left(\frac{p-1}{2} - \alpha_i\right) = \frac{p-3}{2} - \alpha_i$$

in column $(p-3)/2$, remaining to consider. However, since α_i is at least 1, $(p-3)/2-\alpha_i$ is no greater than $(p-5)/2$. So we see that the remaining entries range from 0 to at most $(p-5)/2$, which means that none of them contribute edges to ∂F.

We have shown that, for all possible values of α_i, row α_i^{-1} contributes exactly α_i edges to ∂F. But that is how many edges row α_i contributes. Thus, we see rows α_i and α_i^{-1} each contribute the same number of edges to ∂F. □

Proposition 6. *Let $F = \{0, 1, 2, \ldots, (p-3)/2\}$ be a subset of vertices in X_p and Γ_p be arranged in increasing order. Then*

$$|\partial F| = 2\sum_{i=1}^{k} \alpha_i,$$

where α_i is the i-th element of Γ_p and $k = (p-1)/4$.

Proof. Recall that when arranged in increasing order, we have labeled

$$\Gamma_p = \{\alpha_1, \alpha_2, \ldots, \alpha_k, \alpha_k^{-1}, \ldots, \alpha_2^{-1}, \alpha_1^{-1}\}$$

and that column 0 of the adjacency table is populated in increasing order from top to bottom by the elements of Γ_p. Since there are $2k$ elements in Γ_p and $|\Gamma_p| = (p-1)/2$, it follows that $k = (p-1)/4$.

Since $F = \{0, 1, 2, \ldots, (p-3)/2\}$, if we use the adjacency table to tally the edges in ∂F row-wise, by Proposition 5, rows α_i and α_i^{-1} each contribute exactly α_i edges to ∂F. Thus, we see rows α_1 through α_k contribute a total of $\sum_{i=1}^{k} \alpha_i$ edges to ∂F; as do rows α_k^{-1} through α_1^{-1}.

Since there are no other rows to consider, we see there are exactly $2\sum_{i=1}^{k} \alpha_i$ edges in ∂F, as required. □

Proposition 7. *The isoperimetric constant of a Paley graph satisfies the bound*

$$h(X_p) \leq \frac{1}{k} \sum_{i=1}^{k} \alpha_i,$$

where $\Gamma_p = \{\alpha_1, \alpha_2, \ldots, \alpha_k, \alpha_k^{-1}, \ldots, \alpha_2^{-1}, \alpha_1^{-1}\}$ *and* $k = (p-1)/4$ *as above.*

Proof. Let $F = \{0, 1, 2, \ldots, (p-3)/2\}$. By Proposition 6, this choice gives $|\partial F| = 2\sum_{i=1}^{k} \alpha_i$. Noting that $|F| = (p-1)/2$, we see that

$$\frac{|\partial F|}{|F|} = \frac{2\sum_{i=1}^{k} \alpha_i}{\frac{p-1}{2}} = \frac{1}{\frac{p-1}{4}} \sum_{i=1}^{k} \alpha_i = \frac{1}{k} \sum_{i=1}^{k} \alpha_i. \qquad \square$$

2.2. ***The ((p − 1)/4)-bound.*** We have the suspicion that the α-bound is smaller than $(p-1)/4$ for all primes p congruent to 1 modulo 4, though this has yet to be proven. In fact, early into our work, sample values for the α-bound supported this, and thus contributed to the plausibility for $(p-1)/4$ as an upper bound for $h(X_p)$. Whether or not the α-bound is smaller in general than the $((p-1)/4)$-bound, they appear to be very close.

We begin the proof for the $((p-1)/4)$-bound by introducing a key subset of vertices from the graph X_p. As above, let $\overline{\Gamma}_p = \mathbb{Z}_p \setminus (\Gamma_p \cup \{0\})$. That is, $\overline{\Gamma}_p$ consists of the nonsquares in $\mathbb{Z}_p$. We will prove that $h(X_p) \leq (p-1)/4$ by showing that

$$\frac{|\partial(\overline{\Gamma}_p)|}{|\overline{\Gamma}_p|} = \frac{p-1}{4}. \tag{5}$$

Noting that $\overline{\Gamma}_p$ is the set of all nonzero nonsquares in $\mathbb{Z}_p$, two results follow that will contribute towards our goal: no element of $\overline{\Gamma}_p$ is adjacent to 0, and $\{\Gamma_p, \{0\}, \overline{\Gamma}_p\}$ is a partition of the vertices of X_p. From these two results, we can distill that $\partial(\overline{\Gamma}_p)$ contains only edges going between $\overline{\Gamma}_p$ and Γ_p. Therefore, we will determine $|\partial(\overline{\Gamma}_p)|$ by figuring out how many of the edges incident to vertices in Γ_p remain once the edges going between either an element of Γ_p and 0 or two elements of Γ_p are accounted for. We have that

$$\begin{aligned} |\Gamma_p| \cdot \frac{p-1}{2} &= (\text{\# of edges going between } \Gamma_p \text{ and } \overline{\Gamma}_p) \\ &\quad + (\text{\# of edges going between } \Gamma_p \text{ and } 0) \\ &\quad + 2 \cdot (\text{\# of edges going between vertices in } \Gamma_p). \end{aligned}$$

The first term on the right side is $|\partial(\overline{\Gamma}_p)|$. Also, $|\Gamma_p| = (p-1)/2$, and every vertex in Γ_p is adjacent to 0, so the first factor on the left-hand side and the second term on the right-hand side are both $(p-1)/2$. Substituting these values and solving for $|\partial(\overline{\Gamma}_p)|$, we get

$$|\partial(\overline{\Gamma}_p)| = \frac{p-1}{2} \cdot \frac{p-1}{2} - \frac{p-1}{2} - 2 \cdot (\text{\# of edges between vertices in } \Gamma_p). \tag{6}$$

We now set our sights on determining the number of adjacencies between vertices in Γ_p. The way to do this is to count walks of length 3 that start and end at 0 within X_p. We use the following theorem to do this.

Theorem 8 [Stanley 2013]. *Let G be a graph, $\lambda_1, \lambda_2, \dots, \lambda_n$ be the eigenvalues of G, and N_k be the number of closed walks in G of length k. Then*

$$N_k = \sum_{i=1}^{n} \lambda_i^k.$$

We noted in the introduction that the eigenvalues of the Paley graph X_p are $(p-1)/2$, with multiplicity 1; $(\sqrt{p}-1)/2$, with multiplicity $(p-1)/2$; and $(-\sqrt{p}-1)/2$, with multiplicity $(p-1)/2$. So the number of closed walks of length 3 in X_p is given by

$$\left(\frac{p-1}{2}\right)^3 + \left(\frac{p-1}{2}\right)\left(\frac{\sqrt{p}-1}{2}\right)^3 + \left(\frac{p-1}{2}\right)\left(\frac{-\sqrt{p}-1}{2}\right)^3 = \frac{p}{8}(p-5)(p-1).$$

Paley graphs are Caley graphs, which have a nice property: vertex transitivity. More specifically, if x_1 and x_2 are both vertices of X_p, then the number of closed walks of length k beginning at x_1 is equal to the number of closed walks of length k beginning at x_2. (One can see this by shifting the walks from x_1 to x_2 by adding $-x_1+x_2$ to all the vertices of the closed walk starting at x_1, and vice versa.). Since there are p vertices in X_p, we see that the number of closed walks of length 3 beginning at any one vertex of X_p is $\frac{1}{8}(p-5)(p-1)$. In particular, this is how many such walks begin at 0.

Again, noting that 0 is adjacent to each element of Γ_p (and only to elements of Γ_p), it follows that if δ and β are nonzero elements of $\mathbb{Z}_p$, then $(0, \delta^2, \beta^2, 0)$ is a closed walk of length 3 beginning at zero if and only if $(0, \beta^2, \delta^2, 0)$ is a closed walk of length 3 beginning at zero if and only if δ^2 and β^2 are adjacent vertices of Γ_p. When viewed in this fashion, we see the number of closed walks of length 3 beginning at 0 double counts adjacencies between vertices in Γ_p. That is,

$$\tfrac{1}{8}(p-5)(p-1) = 2 \cdot (\text{\# of edges between vertices in } \Gamma_p).$$

It follows immediately from (6) that

$$\begin{aligned} |\partial(\overline{\Gamma}_p)| &= \frac{p-1}{2} \cdot \frac{p-1}{2} - \frac{p-1}{2} - \tfrac{1}{8}(p-5)(p-1) \\ &= \frac{p-1}{2} \cdot \frac{p-1}{4}. \end{aligned}$$

Dividing the above result by $|\overline{\Gamma}_p| = (p-1)/2$ gives us (5). Hence, $h(X_p) \le (p-1)/4$.

3. The Kazhdan constant of the pair associated with a Paley graph

In this section, we prove Theorem 2. Recall that $\overline{\Gamma}_p = \mathbb{Z}_p \setminus (\Gamma_p \cup \{0\})$, $\xi_p = e^{2\pi i/p}$ and

$$\kappa(\mathbb{Z}_p, \Gamma_p) = \min\bigl\{\max_{\gamma\in\Gamma_p} \|\xi^\gamma - 1\|, \max_{\gamma\in\overline{\Gamma}_p} \|\xi^\gamma - 1\|\bigr\}.$$

To attack the problem of approximating the Kazhdan constant of a Paley graph, we need to use facts about squares and nonsquares in $\mathbb{Z}_p$. For this we need the Legendre symbol. Recall that the Legendre symbol is defined as

$$\left(\frac{a}{p}\right) = \begin{cases} 0 & \text{if } p \text{ divides } a, \\ 1 & \text{if } a \text{ is a square modulo } p, \\ -1 & \text{if } a \text{ is a nonsquare modulo } p. \end{cases}$$

One can show that $\left(\frac{ab}{p}\right) = \left(\frac{a}{p}\right)\left(\frac{b}{p}\right)$. Also, if $x \equiv y \pmod p$, then $\left(\frac{x}{p}\right) = \left(\frac{y}{p}\right)$. It can also be shown that $\left(\frac{-1}{p}\right) = 1$ if and only if $p \equiv 1 \pmod 4$. Likewise $\left(\frac{2}{p}\right) = 1$ if and only if $p \equiv \pm 1 \pmod 8$.These results can be found in any standard book on number theory. For example, see [Niven et al. 1991].

We now show that $\lim_{p\to\infty} \kappa(\mathbb{Z}_p, \Gamma_p) = 2$, where the limit is over all primes with $p \equiv 1 \pmod 4$. We break this into two cases: when $p \equiv 1 \pmod 8$ and when $p \equiv 5 \pmod 8$.

Let $\epsilon > 0$ be an arbitrary small number.

Suppose that $p \equiv 5 \pmod 8$. In this case we have that

$$1 = \left(\frac{-1}{p}\right) = \left(\frac{(p-1)/2}{p}\right)\left(\frac{2}{p}\right) = -\left(\frac{(p-1)/2}{p}\right).$$

Hence $(p-1)\cdot 2^{-1}$ is in $\overline{\Gamma}_p$. Let N_1 be an integer such that if $p > N_1$ and $p \equiv 5 \pmod 8$, then

$$\|\xi_p^{(p-1)/2} - 1\| > 2 - \epsilon.$$

This gives us a nonsquare $(p-1)\cdot 2^{-1}$ of $\mathbb{Z}_p$, where $\xi_p^{(p-1)/2}$ is close to -1 in the complex plane. Now let α be a real number such that $1/2 < \alpha < 1$ and $\|e^{i\alpha\pi} - 1\| > 2 - \epsilon$. Consider the interval $\bigl[\sqrt{\alpha p/2}, \sqrt{p/2}\bigr]$. Note that $\lim_{p\to\infty}\bigl(\sqrt{p/2} - \sqrt{\alpha p/2}\bigr) = \infty$. Hence, there is a positive integer N_2 such that if $p > N_2$, then there exists an integer x such that $\sqrt{\alpha p/2} < x < \sqrt{p/2}$, which is equivalent to $\alpha\pi < (2\pi x^2)/p < \pi$. Hence if $p > N_2$ then there exists a square $\gamma \in \Gamma_p$ such that

$$\|\xi_p^\gamma - 1\| > \|e^{\alpha\pi i} - 1\| > 2 - \epsilon.$$

Combining the above, we have that if p is a prime with $p \equiv 5 \pmod 8$ and $p > \max\{N_1, N_2\}$ then $\kappa(\mathbb{Z}_p, \Gamma_p) > 2 - \epsilon$.

Now suppose that $p \equiv 1 \pmod 8$. In this case we have that

$$1 = \left(\frac{-1}{p}\right) = \left(\frac{(p-1)/2}{p}\right)\left(\frac{2}{p}\right) = \left(\frac{(p-1)/2}{p}\right).$$

Therefore $(p-1)\cdot 2^{-1}$ is in Γ_p. Let N_3 be an integer such that if $p > N_3$ and $p \equiv 1 \pmod 8$, then $\|\xi_p^{(p-1)/2} - 1\| > 2 - \epsilon$. Let $0 < j_p < p$ be the smallest nonsquare in $\overline{\Gamma}_p$. Note that j_p must be odd since if it was even, then

$$\left(\frac{j_p/2}{p}\right) = \left(\frac{2}{p}\right)\left(\frac{j_p/2}{p}\right) = \left(\frac{j_p}{p}\right) = -1.$$

This would imply that $j_p \cdot 2^{-1}$ is a smaller nonsquare than j_p in $\overline{\Gamma}_p$, which is not true. We also have that

$$\left(\frac{(p-j_p)/2}{p}\right) = \left(\frac{(p-j_p)/2}{p}\right)\left(\frac{2}{p}\right) = \left(\frac{-1}{p}\right)\left(\frac{j_p}{p}\right) = \left(\frac{j_p}{p}\right) = -1.$$

Thus, $(p-j_p)\cdot 2^{-1}$ is a nonsquare. We now introduce a lemma which is taken from [Pollack and Treviño 2014]. This lemma will give us a nice bound on j_p.

Lemma 9. $$0 < j_p < \tfrac{1}{2} + \sqrt{p}.$$

Proof. Note that $p < j_p\lceil p/j_p\rceil < p + j_p$. Hence the least nonnegative residue of $j_p\lceil p/j_p\rceil$ modulo p lies in the interval $(0, j_p)$. Therefore, $j_p\lceil p/j_p\rceil$ is a square modulo p. Since j_p is a nonsquare modulo p, we must have that $(j_p\lceil p/j_p\rceil)/j_p = \lceil p/j_p\rceil$ is a nonsquare. By the minimality of j_p, we have that $j_p \le \lceil p/j_p\rceil \le 1 + p/j_p$. Therefore, $j_p^2 - j_p < p$ and hence $j_p^2 - j_p + 1 \le p$. This implies that $(j_p - 1/2)^2 < j_p^2 - j_p + 1 \le p$. So, $j_p < 1/2 + \sqrt{p}$. □

By Lemma 9, we have that

$$\frac{p}{2} > \frac{p-j_p}{2} > \frac{p}{2} - \frac{\sqrt{p}}{2} - \frac{1}{4}.$$

Hence

$$\|\xi_p^{(p-j_p)/2} - 1\| > \|\xi_p^{p/2-\sqrt{p}/2-1/4} - 1\| = \|e^{\pi i - \pi i/\sqrt{p} - \pi i/2p} - 1\|.$$

Let N_4 be a positive integer such that if $p > N_4$ and $p \equiv 1 \pmod 8$, then

$$\|\xi^{(p-j_p)/2} - 1\| > 2 - \epsilon.$$

Thus, if p is a prime with $p \equiv 1 \pmod 8$ and $p > \max\{N_3, N_4\}$, then $\kappa(\mathbb{Z}_p, \Gamma_p) > 2 - \epsilon$.

Combining all of the above results, we have that Theorem 2 has been proved.

References

[Bacher and de la Harpe 1994] R. Bacher and P. de la Harpe, "Exact values of Kazhdan constants for some finite groups", *J. Algebra* **163**:2 (1994), 495–515. MR 95b:20018 Zbl 0842.20013

[Breuillard et al. 2011] E. Breuillard, B. Green, and T. Tao, "Suzuki groups as expanders", *Groups Geom. Dyn.* **5**:2 (2011), 281–299. MR 2012c:20066 Zbl 1247.20017

[Elsawy 2009] A. N. Elsawy, *Paley graphs and their generalizations*, Master's thesis, Heinrich Heine University, Düsseldorf, 2009. arXiv 1203.1818

[Gross et al. 2014] J. L. Gross, J. Yellen, and P. Zhang (editors), *Handbook of graph theory*, 2nd ed., CRC Press, Boca Raton, FL, 2014. MR 3185588 Zbl 1278.05001

[Kassabov et al. 2006] M. Kassabov, A. Lubotzky, and N. Nikolov, "Finite simple groups as expanders", *Proc. Natl. Acad. Sci. USA* **103**:16 (2006), 6116–6119. MR 2007d:20025 Zbl 1161.20010

[Krebs and Shaheen 2011] M. Krebs and A. Shaheen, *Expander families and Cayley graphs: a beginner's guide*, Oxford University Press, 2011. MR 3137611 Zbl 1238.05221

[Lanphier and Rosenhouse 2004] D. Lanphier and J. Rosenhouse, "Cheeger constants of Platonic graphs", *Discrete Math.* **277**:1-3 (2004), 101–113. MR 2005c:05102 Zbl 1033.05055

[Mohar 1989] B. Mohar, "Isoperimetric numbers of graphs", *J. Combin. Theory* (*B*) **47**:3 (1989), 274–291. MR 90m:05087 Zbl 0719.05042

[Niven et al. 1991] I. Niven, H. L. Montgomery, and H. S. Zuckerman, *An introduction to the theory of numbers*, 5th ed., Wiley, New York, 1991. MR 91i:11001 Zbl 0742.11001

[Pollack and Treviño 2014] P. Pollack and E. Treviño, "The primes that Euclid forgot", *Amer. Math. Monthly* **121**:5 (2014), 433–437. MR 3193727 Zbl 06367526

[Rosenhouse 2002] J. Rosenhouse, "Isoperimetric numbers of Cayley graphs arising from generalized dihedral groups", *J. Combin. Math. Combin. Comput.* **42** (2002), 127–138. MR 2004d:05092 Zbl 1019.05030

[Stanley 2013] R. P. Stanley, *Algebraic combinatorics: walks, trees, tableaux, and more*, Springer, New York, 2013. MR 3097651 Zbl 1278.05002

[Voskanian 2013] E. Voskanian, *Obtaining lower bounds for the Kazhdan constant*, Master's thesis, California State University, Los Angeles, 2013.

Received: 2014-10-28 Revised: 2015-01-25 Accepted: 2015-02-02

kevinhcramer@gmail.com
Department of Mathematics, California State University,
Los Angeles, Los Angeles, CA 90032, United States

mkrebs@calstatela.edu
Department of Mathematics, California State University,
Los Angeles, Los Angeles, CA 90032, United States

linus108nicole@yahoo.com
Department of Mathematics, California State University,
Los Angeles, Los Angeles, CA 90032, United States

ashahee@calstatela.edu
Department of Mathematics, California State University,
Los Angeles, Los Angeles, 90032, United States

voskanian@math.ucr.edu
Department of Mathematics, University of California,
Riverside, Riverside, CA 92521, United States

INVOLVE 9:2 (2016)
dx.doi.org/10.2140/involve.2016.9.307

Mutual estimates for the dyadic reverse Hölder and Muckenhoupt constants for the dyadically doubling weights

Oleksandra V. Beznosova and Temitope Ode

(Communicated by Kenneth S. Berenhaut)

Muckenhoupt and reverse Hölder classes of weights play an important role in harmonic analysis, PDEs and quasiconformal mappings. In 1974, Coifman and Fefferman showed that a weight belongs to a Muckenhoupt class A_p for some $1 < p < \infty$ if and only if it belongs to a reverse Hölder class RH_q for some $1 < q < \infty$. In 2009, Vasyunin found the exact dependence between p, q and the corresponding characteristic of the weight using the Bellman function method. The result of Coifman and Fefferman works for the dyadic classes of weights under an additional assumption that the weights are dyadically doubling. We extend Vasyunin's result to the dyadic reverse Hölder and Muckenhoupt classes and obtain the dependence between p, q, the doubling constant and the corresponding characteristic of the weight. More precisely, given a dyadically doubling weight in RH_p^d on a given dyadic interval I, we find an upper estimate on the average of the function w^q (with $q < 0$) over the interval I. From the bound on this average, we can conclude, for example, that w belongs to the corresponding $A_{q_1}^d$-class or that w^p is in $A_{q_2}^d$ for some values of q_i. We obtain our results using the method of Bellman functions. The main novelty of this paper is how we use dyadic doubling in the Bellman function proof.

1. Definitions and main results

We will be dealing with a family of dyadic intervals on the real line,

$$D := \{[n2^{-k}, (n+1)2^{-k}] : n, k \in \mathbb{Z}\}.$$

For an interval J, let $D(J)$ stand for the family of all its dyadic subintervals, $D(J) := \{I \in D : I \subset J\}$ and let $D_n(J)$ stand for the family of all dyadic subintervals

MSC2010: 42B37.
Keywords: reverse Hölder, Muckenhoupt, weights, dyadic.

of J of length exactly $2^{-n}|J|$. For a locally integrable function f, let $\langle f\rangle_I$ stand for the average of f over the interval I,

$$\langle f\rangle_I := \frac{1}{|I|}\int_I f(x)\,dx,$$

where $|I|$ is the Lebesgue measure of I.

Let w be a weight; i.e., w is a locally integrable, almost everywhere nonnegative function which is not identically zero. Since we will be dealing mostly with averages, we define the dyadic doubling constant of the weight w to be

$$\mathrm{Db}^d(w) := \inf_{I\in D}\bigl\{C : \langle w\rangle_{I^*} \leqslant C\langle w\rangle_I\bigr\} = \tfrac{1}{2}\inf_{I\in D}\biggl\{C : \int_{I^*} w(x)\,dx \leqslant C\int_I w(x)\,dx\biggr\},$$

where I^* is the dyadic "parent" of the interval I, i.e., the smallest dyadic interval that strictly contains the interval I. If the dyadic doubling constant of the weight w is bounded by Q, we will say that $w \in \mathrm{Db}^{d,Q}$. Note also that any weight is positive almost everywhere; therefore the dyadic doubling constant defined this way is always greater than $\frac{1}{2}$.

Our main assumption is that a weight w belongs to the dyadic reverse Hölder class on the interval J with the corresponding constant bounded by δ:

$$w \in RH_p^{\delta,d}(J) \quad\Longleftrightarrow\quad [w]_{RH_p^{\delta,d}(J)} := \sup_{I\in D(J)} \bigl\{C : \langle w^p\rangle_I^{1/p} \leqslant C\langle w\rangle_I\bigr\} \leqslant \delta.$$

We define $A_q^{\delta,d}(J)$ to be the class of the dyadic Muckenhoupt weights on the interval J with the corresponding constant bounded by δ:

$$w \in A_q^{\delta,d}(J) \quad\Longleftrightarrow\quad [w]_{A_q^{\delta,d}(J)} := \sup_{I\in D(J)} \langle w\rangle_I \langle w^{-1/(q-1)}\rangle_I^{q-1} \leqslant \delta.$$

Given a dyadically doubling weight $w \in RH_p^{\delta,d}(J)$, our goal in this paper is to bound the averages involved in the definitions of $w \in A_{q_1}^d$ and $w^p \in A_{q_2}^d$,

$$\langle w\rangle_J \langle w^{-1/(q_1-1)}\rangle_J^{q_1-1} \quad\text{and}\quad \langle w^p\rangle_J \langle w^{-p/(q_2-1)}\rangle_J^{q_2-1}.$$

Note that the quantities $\langle w\rangle_J$ and $\langle w^p\rangle_J$ are involved in the definition of $RH_p^{\delta,d}(J)$; therefore for our goals, it is enough to bound $\langle w^q\rangle_J$ from above for $q < 0$.

It is a well-known fact that $w \in A_q^d$ for some $1 < q < \infty$ implies that w is a dyadically doubling weight; it is also known that in the dyadic case, the reverse Hölder classes RH_p^d contain weights that are not dyadically doubling (see [Buckley 1990]). In fact, if $w \in RH_p^{\delta,d}(J)$, nothing prevents w from being close or even equal to 0 on, say, the left half of J; the local $RH_p^d(J)$-constant can be defined for such weights. There is no way to define an $A_q^d(J)$-constant for such a weight, and even the quantity $\langle w^q\rangle_J$ is undefined for $q < 0$, which is the case considered in this paper. What prevents this from happening is the doubling assumption that does not allow $\langle w\rangle_J$

to be too far from $\langle w\rangle_{J^\pm}$, and therefore if w is equal to 0 on any dyadic subinterval of J then w has to be identically 0 on the whole interval J (which is not permitted).

We are ready to define the Bellman function for our problem: for $p>1$, $q<0$ and $Q>2$, let

$$\mathcal{B}(x_1,x_2;p,q,\delta,Q):=\sup_{w\in RH_p^{\delta,d}(J),\,\mathrm{Db}^d(w)\leqslant Q}\bigl\{\langle w^q\rangle_J : w \text{ is s.t. } \langle w\rangle_J=x_1,\langle w^p\rangle_J=x_2\bigr\}.$$

The parameters p, q, δ and Q will be fixed throughout the paper, so we will skip them and write $\mathcal{B}(x_1,x_2)$. Note also that by a rescaling argument, $\mathcal{B}$ does not depend on the interval J. The constant Q corresponds to the doubling constant of the weight w. We know that for any weight, we have $\mathrm{Db}^d(w)>\frac{1}{2}$. We take $Q>2$ for technical reasons (we need it in the proof), so one may think of Q as being the maximum of the doubling constant of the weight w and 2; that is, $Q:=\max\{\mathrm{Db}^d(w),2\}$.

Then for the given p, q, δ, and Q, we have that $\mathcal{B}$ is defined on the domain

$$U_\delta:=\bigl\{\vec{x}=(x_1,x_2)\,:\,\exists w\in RH_p^{\delta,d} \text{ s.t. } \mathrm{Db}^d(w)\leqslant Q \text{ and } x_1=\langle w\rangle_J,\, x_2=\langle w^p\rangle_J\bigr\}.$$

In order to state the main theorem, we need to define functions $u_p^\pm(t)$. Let $u_p^\pm(t)$ be two solutions (positive and negative) of the equation

$$(1-pu)^{1/p}(1-u)^{-1}=t,\quad 0\leqslant t\leqslant 1. \tag{1-1}$$

For $Q\geqslant 2$, we define $\varepsilon(p,\delta,Q)$ as follows. Let

$$H:=H(p,Q)=\frac{Q^p-1}{Q-1}\quad\text{and}\quad \varepsilon:=\frac{H}{p}\left(\frac{p-1}{H-1}\right)^{(p-1)/p}\delta.$$

Then we can define

$$s^\pm(\varepsilon):=u^\pm\left(\frac{1}{\varepsilon}\right)\quad\text{and}\quad r^\pm:=u^\pm\left(\frac{y^{1/p}}{\varepsilon x}\right).$$

Note that since $u^+(t)$ is a decreasing function and in our domain

$$\frac{1}{\varepsilon}\leqslant\frac{y^{1/p}}{\varepsilon x},$$

we have that $r^+\in[0,s^+]$. Similarly, since $u^-(t)$ is an increasing function, we have that $r^-\in[s^-,0]$.

Theorem 1.1 (main theorem). *Let $p>1$, $q<0$, $Q\geqslant 2$ and $\delta>1$; let $s^-:=s^-(\varepsilon)$ for $\varepsilon(p,\delta,Q)$ defined above. If $q\in(1/s^-,0)$ then*

$$\mathcal{B}(x_1,x_2;p,q,\delta)\leqslant x_1^q\frac{1-qr^-}{1-qs^-}\left(\frac{1-s^-}{1-r^-}\right)^q=x_2^{q/p}\frac{1-qr^-}{1-qs^-}\left(\frac{1-ps^-}{1-pr^-}\right)^{q/p}.$$

The proof of Theorem 1.1 can be found in Section 2.

Note that the result from [Vasyunin 2008] assumes that the reverse Hölder inequality for the weight w holds for any interval $I \subset J$, while our Theorem 1.1 only uses dyadic subintervals $I \in D(J)$ and the doubling constant. Therefore our result is more general (in the sense that if a weight is in the continuous reverse Hölder class, it has to be in the dyadic class and it has to be doubling, so our theorem applies). Unfortunately, we lose the sharpness. Note also that Theorem 1.1 is not a straight-forward extension of Vasyunin's result because it fails in the case when the weight w is not dyadically doubling. The latter is easy to see: in his thesis, Buckley gave examples of weights in RH_p-classes that are not dyadically doubling and therefore do not belong to any of the A_p^d.

Let us consider the following simple example. Let $w(x) = \chi_{J^+}(x)$. Then $w \in RH_p^{d,2^{1-1/p}}(J)$ for all $1 < p < \infty$. At the same time, it is clearly impossible to bound $\langle w^q \rangle_J$ for $q < 0$. Note that this weight is not dyadically doubling, so the doubling assumption in Theorem 1.1 is necessary, and we have to find a way to use doubling in the Bellman function argument. Most of Vasyunin's proof works in the dyadic setting; it is Lemma 4 in his paper that fails and does not have a full size dyadic analogue. We replace Lemma 4 using a technique from [Pereyra 2009] to incorporate the doubling property of the weight in the Bellman function proof.

As a consequence of Theorem 1.1, we obtain the following corollary.

Corollary 1.2 (RH_p vs. A_q). *Let w be a reverse Hölder dyadically doubling weight with $[w]_{RH_p^d} = \delta$ and $Q := \max\{\mathrm{Db}^d(w), 2\}$. Let $\varepsilon(p, \delta, Q)$ be defined as above. Let $s^- = s^-(\varepsilon)$. Then*:

(i) *For every $q_1 > 1 - s^-$, we have $w \in A_{q_1}^d$, and moreover,*

$$[w]_{A_{q_1}^d} \leqslant \left(\frac{q_1 - 1}{q_1 - 1 + s^-} \right)^{q_1 - 1}.$$

(ii) *For every $q_2 > 1 - ps^-$, we have $w^p \in A_{q_2}^d$, and moreover,*

$$[w^p]_{A_{q_2}^d} \leqslant \left(\frac{q_2 - 1}{q_2 - 1 + ps^-} \right)^{q_2 - 1}.$$

Above, $s^-(\varepsilon)$ is the negative solution of the equation $(1 - ps^-)^{1/p}(1 - s^-)^{-1} = 1/\varepsilon$.

A result similar to the second part of the above corollary was used in [Beznosova et al. 2014] (without a proof) for the sharp norms of t-Haar multiplier operators. The difference is that in [loc. cit.], the ε was taken to be $\varepsilon_1 = Q\delta$, which is an upper bound for our $\varepsilon(p, \delta, Q)$.

The proof of Corollary 1.2 is very simple. Note that since $r^- \in [s^-, 0]$, we have $1 - r^- \leqslant 1 - s^-$; therefore

$$\frac{1 - s^-}{1 - r^-} \geqslant 1.$$

So, since $q<0$, we have that

$$\left(\frac{1-s^-}{1-r^-}\right)^q \leqslant 1.$$

We also have that

$$\left(\frac{1-ps^-}{1-pr^-}\right)^{q/p} \leqslant 1$$

since p is positive. At the same time, since both q and r^- are negative, qr^- is positive and $1-qr^- \leqslant 1$. Therefore, for our choice of parameters, we have that

$$\langle w^q\rangle_J \leqslant \frac{\min\{\langle w\rangle_J^q,\, \langle w^p\rangle_J^{q/p}\}}{1-qs^-}.$$

Using this rough estimate in the definition of the corresponding Muckenhoupt constant, we get the desired bounds.

2. Proof of Theorem 1.1

In this section, we essentially follow the proof from [Vasyunin 2008]. Unfortunately, we cannot use the full proof from Vasyunin's paper since it relies on Lemma 4 from his paper, which fails in the dyadic case. We will sketch the proof, referring to Vasyunin's results whenever possible, and replace his Lemma 4 with our dyadically doubling analogue, Lemma 2.4.

We fix $p>1,\ ,q<0,\ Q\geqslant 2,\ \delta>1$ and let

$$\mathcal{B}(x_1,x_2;p,q,\delta,Q) := \sup_{w\in RH_p^{\delta,d}(J),\,\mathrm{Db}^d(w)\leqslant Q} \left\{\langle w^q\rangle_J : w \text{ is s.t. } \langle w\rangle_J = x_1,\, \langle w^p\rangle_J = x_2\right\}$$

and

$$B_{\max} = B_{\max}(x_1,x_2;\,p,q,\delta,\,Q) := x_1^q\,\frac{1-qr^-}{1-qs^-}\left(\frac{1-s^-}{1-r^-}\right)^q$$

be defined on the domains

$$U_\delta = \left\{\vec{x}=(x_1,x_2)\ :\ \exists w\in RH_p^{\delta,d} \text{ s.t. } \mathrm{Db}^d(w)\leqslant Q \text{ and } x_1=\langle w\rangle_J,\, x_2=\langle w^p\rangle_J\right\}$$

and

$$\Omega_\varepsilon := \left\{\vec{x}=(x_1,x_2)\ :\ x_i>0,\ x_1^p\leqslant x_2\leqslant \varepsilon^p x_1^p\right\}$$

respectively. Please note that U_δ and Ω_ε here are two domains defined in two different ways. In Lemma 2.4, we show that one is contained in the other and any line segment that connects points in U_δ that correspond to the same weight and dyadic interval has to lie inside the enlarged domain Ω_ε. This part is the main difference between the continuous and the dyadic case.

Note that

$$x_1^q \frac{1-qr^-}{1-qs^-}\left(\frac{1-s^-}{1-r^-}\right)^q = x_2^{q/p}\frac{1-qr^-}{1-qs^-}\left(\frac{1-ps^-}{1-pr^-}\right)^{q/p}$$

by the definitions of s^- and r^-.

Our goal is to show that $\mathcal{B} \leqslant B_{\max}$. We will prove it using the Bellman function method. The proof consists of the following parts, which we will now state in the form of lemmata.

Lemma 2.1. *If the function $B_{\max}$, defined above, is concave on the domain Ω_δ, i.e.,*

$$B_{\max}\left(\frac{x^- + x^+}{2}\right) \geqslant \frac{B_{\max}(x^-) + B_{\max}(x^+)}{2} \tag{2-1}$$

for any x^+ and x^- such that there exists a weight $w \in RH_p^{\delta,d}$ with $\mathrm{Db}^d(w) \leqslant Q$, where $x^+ = (\langle w\rangle_{J^+}, \langle w^p\rangle_{J^+})$ and $x^- = (\langle w\rangle_{J^-}, \langle w^p\rangle_{J^-})$, then Theorem 1.1 holds.

Lemma 2.2. *The function $B_{\max}$ is locally concave on the domain Ω_ε; i.e., its Hessian matrix*

$$d^2 B_{\max} = \left\{\frac{\partial^2 B_{\max}}{\partial x_1 \partial x_2}\right\}$$

is not positive definite.

Lemma 2.3. *Let $x^o, x^+, x^- \in U_\delta$, where $x^o = \frac{1}{2}(x^+ + x^-)$ and the line segment connecting x^+ and x^- lies completely inside the larger domain Ω_ε. Suppose that the function $B_{\max}$ is locally convex on Ω_ε; i.e., on Ω_ε we have that the Hessian $d^2 B_{\max}$ is not positive definite. Then the inequality (2-1) holds.*

Lemma 2.4. *Let x^o, x^+, and x^- be three points in Ω_δ with the property that $x^o = \frac{1}{2}(x^+ + x^-)$ such that there is a weight $w \in RH_p^{\delta,d}$ with $\mathrm{Db}^d(w) \leqslant Q$ and a dyadic interval I such that*

$$x_1^o = \langle w\rangle_I, \quad x_2^o = \langle w^p\rangle_I,$$
$$x_1^\pm = \langle w\rangle_{I^\pm}, \quad x_2^\pm = \langle w^p\rangle_{I^\pm}.$$

Then the line segment connecting x^+ and x^- lies completely inside the larger domain Ω_ε.

Proof of Lemma 2.1. First, observe that if a weight w is constant on the interval J, say $w = c$, then $\langle w^q\rangle_J = \langle w\rangle_J^q = \langle w^p\rangle_J^{q/p}$; therefore in this case, $\mathcal{B} \leqslant B_{\max}$.

Now let w be a step function. Note that for any dyadic interval I, we have that $\langle w\rangle_I = \frac{1}{2}\big(\langle w\rangle_{I^+} + \langle w\rangle_{I^-}\big)$ and $\langle w^p\rangle_I = \frac{1}{2}\big(\langle w^p\rangle_{I^+} + \langle w^p\rangle_{I^-}\big)$. This, together with the concavity of $B_{\max}$, gives

$$\begin{aligned}
&|J|\,B_{\max}\big(\langle w\rangle_J, \langle w^p\rangle_J\big) \\
&\qquad\geqslant |J^-|\,B_{\max}\big(\langle w\rangle_{J^-}, \langle w^p\rangle_{J^-}\big) + |J^+|\,B_{\max}\big(\langle w\rangle_{J^+}, \langle w^p\rangle_{J^+}\big) \\
&\qquad\geqslant |J^{--}|\,B_{\max}\big(\langle w\rangle_{J^{--}}, \langle w^p\rangle_{J^{--}}\big) + |J^{-+}|\,B_{\max}\big(\langle w\rangle_{J^{-+}}, \langle w^p\rangle_{J^{-+}}\big) \\
&\qquad\quad + |J^{+-}|\,B_{\max}\big(\langle w\rangle_{J^{+-}}, \langle w^p\rangle_{J^{+-}}\big) + |J^{++}|\,B_{\max}\big(\langle w\rangle_{J^{++}}, \langle w^p\rangle_{J^{++}}\big) \\
&\qquad\geqslant \cdots \geqslant \sum_{I\in D_n(J)} |I|\,B_{\max}\big(\langle w\rangle_I, \langle w^p\rangle_I\big).
\end{aligned}$$

Now note that since w is a step function, it has at most finitely many jumps. Let the number of jumps be m. For n large enough, in the last formula we have that w is constant on $2^n - m$ subintervals $I \in D_n(J)$ (we will call these subintervals "good") and has jump discontinuities on the other m subintervals (we will call these subintervals "bad"). On good subintervals, w is constant, so for such intervals, we have that $|I|\,B_{\max}(\langle w\rangle_I, \langle w^p\rangle_I) \geqslant |I|\langle w^q\rangle_I$. For the bad intervals, we know that $B_{\max}$ is a continuous function and the set of points $\{x = (\langle w\rangle_I, \langle w^p\rangle_I) : I \in D(J)\}$ is a compact subset of Ω_ε, so $B_{\max}(\langle w\rangle_I, \langle w^p\rangle_I)$ for bad intervals $\{I_k\}_{k=1,\dots,m}$ are bounded by a uniform constant M. So the whole sum differs from $|J|\langle w^q\rangle_J$ by at most $M\sum_{I\text{ bad}}|I|$, which tends to 0 as $n \to \infty$.

This implies that $\langle w\rangle_J \leqslant B_{\max}(\langle w\rangle_J, \langle w^p\rangle_J)$ for all step functions w.

Next we extend this result to all weights w_m that are bounded from above and from below, say $m \leqslant w \leqslant M$. We take a sequence of step functions w_n that pointwise converge to w_m. By the Lebesgue dominated convergence theorem, Lemma 2.1 should hold for w_m.

The result of [Reznikov et al. 2010] extends our argument to an arbitrary weight w, which completes the proof of the Lemma 2.1. □

Proof of Lemma 2.2. We want to show that the matrix of second derivatives of $B_{\max}$ is not positive definite. We will just refer to [Vasyunin 2008], where it is shown in a more general case. □

Proof of Lemma 2.3. For the fixed points x^o, x^+ and x^- in the domain $U_\delta \subset \Omega_\varepsilon$ with $x^o = \frac{1}{2}(x^- + x^+)$ and such that the line segment connecting x^+ and x^- lies inside the domain Ω_ε, we introduce the function $b(t) := B(x_t)$, where $x_t := \frac{1}{2}(1+t)x^+ + \frac{1}{2}(1-t)x^-$. Note that defined this way, $B(x^o) = b(0)$, while $B(x^+) = b(1)$ and $B(x^-) = b(-1)$. Note also that

$$b''(t) = \begin{bmatrix} \frac{dx}{dt} & \frac{dy}{dt} \end{bmatrix} d^2 B_{\max} \begin{bmatrix} \frac{dx}{dt} \\ \frac{dy}{dt} \end{bmatrix}.$$

So, since $-d^2B_{\max}$ is not negative definite, $-b''(t) \geqslant 0$ for all $-1 \leqslant t \leqslant 1$.

On the other hand,

$$B_{\max}(x^o)-\frac{B_{\max}(x^+)+B_{\max}(x^-)}{2}=b(0)-\frac{b(1)+b(-1)}{2}=-\frac{1}{2}\int_{-1}^{1}(1-|t|)b''(t)\,dt.$$

The second part of the above formula is a simple calculus exercise of integrating by parts twice.

Clearly, since $-b''(t)$ is nonnegative,

$$B_{\max}(x^o)-\frac{B_{\max}(x^+)+B_{\max}(x^-)}{2}\geqslant 0,$$

which completes the proof of Lemma 2.3. □

Proof of Lemma 2.4. Let x^o, x^+ and x^- be three points in

$$U_\delta:=\left\{\vec{x}=(x_1,x_2)\,:\,\exists w\in RH_p^{\delta,d}\cap \mathrm{Db}^{Q,d}\text{ s.t. }x_1=\langle w\rangle_J,\,x_2=\langle w^p\rangle_J\right\}$$

that correspond to the same weight w and interval I; i.e., there is a weight $w\in RH_p^{\delta,d}$ with $\mathrm{Db}^d(w)\leqslant Q$ and a dyadic interval I such that

$$x_1^o=\langle w\rangle_I,\qquad x_2^o=\langle w^p\rangle_I,$$
$$x_1^\pm=\langle w\rangle_{I^\pm},\quad x_2^\pm=\langle w^p\rangle_{I^\pm}.$$

Note that the reverse Hölder property for the weight w implies that $x_1^p\leqslant x_2\leqslant\delta^p x_1^p$ for all three points x^o, x^+ and x^-, and the fact that w is almost everywhere positive implies that $x_1,x_2>0$. At the same time, the fact that w is dyadically doubling with a doubling constant at most Q implies that

$$x_1^o\leqslant Qx_1^\pm,\quad x_1^\pm\leqslant 2x_1^o,\quad\text{and}\quad x_1^\mp\leqslant(Q-1)x_1^\pm.$$

Without loss of generality, we will assume that $x_1^-<x_1^+$. Then we know that $x_1^o\leqslant Qx_1^-$, $x_1^+\leqslant 2x_1^o$ and $x_1^+\leqslant(Q-1)x_1^-$.

Therefore

$$U_\delta\subset\Omega_\delta:=\left\{\vec{x}=(x_1,x_2)\,:\,x_1^p\leqslant x_2\leqslant\delta^p x_1^p\right\}\subset\Omega_\varepsilon,$$

and the points x^o, x^+ and $x^-\in U_\delta$ are such that

$$x^o=\tfrac{1}{2}(x^++x^-),\quad x_1^-<x_1^o<x_1^+,$$
$$x_1^o\leqslant Qx_1^-,\quad x_1^+\leqslant 2x_1^o,\quad x_1^+\leqslant(Q-1)x_1^-.$$

We need to show that the line interval connecting x^+ and x^- lies inside the domain Ω_ε.

First observe that the worst case scenario is when the central point x^o and one of the endpoints lie on the upper boundary of U_δ, $x_2=\delta^p x_1^p$, while the other endpoint lies on the lower boundary of U_δ, $x_2=x_1^p$. There are two possibilities, so let us consider the two cases separately.

Case 1: x^o and x^- are on the upper boundary and x^+ is on the lower boundary. This means that

$$x^o = (x_1^o, \delta^p (x_1^o)^p), \quad x^- = (x_1^-, \delta^p (x_1^-)^p), \quad x^+ = (x_1^+, (x_1^+)^p).$$

We need to minimize the function $f(x) = x_2^{1/p} x_1^{-1}$ over the line that passes through the points x^o, x^+ and x^-. We are not going to use all of the conditions on our points. To simplify the problem, we will drop the condition that the point x^+ is on the lower boundary. We will only be using the points x^o and x^- and we will use the fact that $x_1^o \leqslant Q x_1^-$.

Again, in the worst case, which may be unattainable, $x_1^o = Q x_1^-$. The line through the points $x^- = (x_1^-, \delta^p (x_1^-)^p)$ and $x^o = (Qx^-, Q^p \delta^p (x_1^-)^p)$ has slope

$$\frac{\delta^p (x_1^-)^p (Q^p - 1)}{Q - 1}.$$

Therefore the equation is

$$x_2 - \delta^p (x_1^-)^p - \delta^p (x_1^-)^{p-1} \frac{Q^p - 1}{Q - 1} (x_1 - x_1^-) = 0.$$

So we need to solve the optimization problem

$$\begin{cases} f(x) = x_2^{1/p} x_1^{-1} \to \max, \\ x_2 - \delta^p (x_1^-)^p - \delta^p (x_1^-)^{p-1} \dfrac{Q^p - 1}{Q - 1} (x_1 - x_1^-) = 0. \end{cases}$$

The problem can be solved, for example, using method of Lagrange multipliers. If we let $H := (Q^p - 1)/(Q - 1)$ then

$$f_{\max} = \left(\frac{p - 1}{H - 1} \right)^{(p-1)/p} \frac{H}{p} \delta,$$

which is exactly our choice of ε.

Case 2: x^o and x^+ are on the upper boundary and x^- is on the lower boundary. In this case, we will drop the condition that x^- is on the lower boundary. Since the coordinates of our points are positive,

$$x_1^o = \frac{x_1^+ + x_1^-}{2} \geqslant \frac{x_1^+}{2},$$

so $x_1^+ \leqslant 2 x_1^o$. Therefore this case is similar to Case 1 with $Q = 2$. Since $Q \geqslant 2$, this case is covered as well. This is the only place where we use that $Q \geqslant 2$.

This completes the proof of Lemma 2.4 and Theorem 1.1. □

References

[Beznosova et al. 2014] O. Beznosova, J. C. Moraes, and M. C. Pereyra, "Sharp bounds for t-Haar multipliers on L^2", pp. 45–64 in *Harmonic analysis and partial differential equations* (El Escorial, 2012), edited by P. Cifuentes et al., Contemp. Math. **612**, Amer. Math. Soc., Providence, RI, 2014. MR 3204856 Zbl 1304.42095

[Buckley 1990] S. M. Buckley, *Harmonic analysis on weighted spaces*, Ph.D. thesis, University of Chicago, 1990, available at http://search.proquest.com/docview/275671193.

[Pereyra 2009] M. C. Pereyra, "Haar multipliers meet Bellman functions", *Rev. Mat. Iberoam.* **25**:3 (2009), 799–840. MR 2010m:42016 Zbl 1198.42007

[Reznikov et al. 2010] A. Reznikov, V. Vasyunin, and A. Volberg, "An observation: cut-off of the weight w does not increase the A_{p_1,p_2}-norm of w", preprint, 2010. arXiv 1008.3635v1

[Vasyunin 2008] V. I. Vasyunin, "Mutual estimates for L^p-norms and the Bellman function", *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov.* (*POMI*) **355**:36 (2008), 81–138. In Russian; translated in *J. of Math. Sci.***156**:5 (2009), 766–798. MR 2012b:42040 Zbl 1184.26009

Received: 2014-11-05 Revised: 2015-03-30 Accepted: 2015-04-16

ovbeznosova@ua.edu
Department of Mathematics, Box 870350, University of Alabama, Tuscaloosa, AL 35487-0350, United States

temitope_ode@baylor.edu
Baylor University, Waco, TX 76798, United States

INVOLVE 9:2 (2016)
dx.doi.org/10.2140/involve.2016.9.317

Radio number for fourth power paths

Min-Lin Lo and Linda Victoria Alegria

(Communicated by Jerrold Griggs)

Let G be a connected graph. For any two vertices u and v, let $d(u, v)$ denote the distance between u and v in G. The maximum distance between any pair of vertices of G is called the diameter of G and denoted by $\operatorname{diam}(G)$. A *radio labeling* (or multilevel distance labeling) of G is a function f that assigns to each vertex a label from the set $\{0, 1, 2, \dots\}$ such that the following holds for any vertices u and v: $|f(u) - f(v)| \geq \operatorname{diam}(G) - d(u, v) + 1$. The *span* of f is defined as $\max_{u,v \in V(G)}\{|f(u) - f(v)|\}$. The *radio number* of G is the minimum span over all radio labelings of G. The *fourth power* of G is a graph constructed from G by adding edges between vertices of distance four or less apart in G. In this paper, we completely determine the radio number for the fourth power of any path, except when its order is congruent to $1 \pmod 8$.

1. Introduction

Motivated by the *channel assignment problem* [Hale 1980] of dividing the radio broadcasting spectrum among radio stations in such a way that the interference caused by their proximity is minimized, radio labeling was introduced by Chartrand et al. [2001] to model the problem of finding the optimal distribution of channels using the smallest necessary range of frequencies.

Let G be a connected graph. For any two vertices u and v of G, the *distance* between u and v is the length of a shortest u-v path in G and is denoted by $d_G(u, v)$ or simply $d(u, v)$ if the graph G under consideration is clear. The *diameter* of G, denoted by $\operatorname{diam}(G)$, is the greatest distance between any two vertices of G. A *radio labeling* (or *multilevel distance labeling* [Liu 2008; Liu and Zhu 2005]) of a connected graph G is a function $f : V(G) \to \{0, 1, 2, 3, \dots\}$ with the property that

$$|f(u) - f(v)| \geq \operatorname{diam}(G) + 1 - d(u, v)$$

for every two distinct vertices u and v of G. The *span* of f is defined as

$$\max_{u,v \in V(G)}\{|f(u) - f(v)|\}.$$

MSC2010: 05C78.
Keywords: channel assignment problem, multilevel distance labeling, radio number, radio labeling.

The radio number of G, denoted by $\operatorname{rn}(G)$, is defined as

$$\min\{\text{span of } f : f \text{ is a radio labeling of } G\}.$$

A radio labeling for G with span equal to $\operatorname{rn}(G)$ is called an *optimal radio labeling*.

Finding the radio number for a graph is an interesting yet challenging task. So far the value is known only for very limited families of graphs. The radio numbers for paths and cycles were investigated in [Chartrand et al. 2001; Chartrand, Erwin and Zhang 2005; Zhang 2002]and were completely solved by Liu and Zhu [2005]. The radio number for trees was investigated in [Liu 2008].

The r-th power of a graph G, denoted by G^r, is the graph constructed from G by adding edges between vertices of distance r or less apart in G. The radio number for the square of a path on n vertices, denoted by P_n^2, was completely determined by Liu and Xie [2009], who also partially solved the problem for the square of a cycle on n vertices, denoted by C_n^2 [2004]. Motivated by [Liu and Xie 2009], Lo [2010] and Sooryanarayana et al. [2010] determined $\operatorname{rn}(P_n^3)$.

This paper will follow the structure in [Liu and Xie 2009] closely to determine the radio number of the fourth power of paths (or simply, fourth power paths). It is our hope that this paper will be helpful for those readers who wish to pursue finding the radio number for P_n^5, P_n^6, and eventually P_n^r for any positive integer r.

Theorem 1. *Let P_n^4 be a fourth power path on n vertices where $n \geq 6$ and let $k = \operatorname{diam}(P_n^4) = \lceil \frac{1}{4}(n-1) \rceil$. Then*

$$\operatorname{rn}(P_n^4) = \begin{cases} 2k^2+1 & \text{if } n \equiv 0, 3, 6, \text{ or } 7 \pmod 8 \text{ or } n = 9, \\ 2k^2+2 & \text{if } n \equiv 4 \text{ or } 5 \pmod 8, \\ 2k^2 & \text{if } n \equiv 2 \pmod 8. \end{cases}$$

If $n \equiv 1 \pmod 8$ and $n \geq 17$ (where n is of the form $8q+1$), then

$$2k^2 + 2 \leq \operatorname{rn}(P_{8q+1}^4) \leq 2k^2 + q.$$

2. General properties and notation

The diameter of P_n^4 is $\lceil \frac{1}{4}(n-1) \rceil$, based on the definition of P_n^4. Figure 1 shows P_8^4.

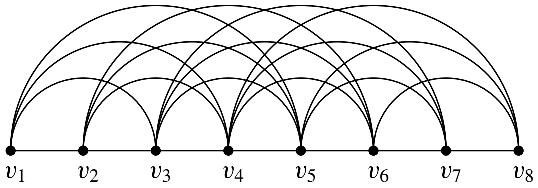


Figure 1. A fourth power path on 8 vertices, denoted by P_8^4.

Proposition 2. *For any $u, v \in V(P_n^4)$, we have*

$$d(u,v) = \left\lceil \tfrac{1}{4} d_{P_n}(u,v) \right\rceil.$$

A *center* of P_n is defined as a "middle" vertex of P_n. An odd path P_{2m+1} has only one center v_{m+1}, while an even path P_{2m} has two centers v_m and v_{m+1}. For each vertex $u \in V(P_n)$, the *level* of u, denoted by $L(u)$ is the smallest distance in P_n from u to a center of P_n. If we denote the levels of a sequence of vertices A by $L(A)$, we have

$$n=2m+1 \quad \Rightarrow \quad L(v_1, v_2, \dots, v_{2m+1}) = (m, m-1, \dots, 2, 1, 0, 1, 2, \dots, m-1, m),$$
$$n=2m \quad \Rightarrow \quad L(v_1, v_2, \dots, v_{2m}) = (m-1, \dots, 2, 1, 0, 0, 1, 2, \dots, m-1).$$

Define the *left-vertices* and *right-vertices* as follows:

If $n = 2m+1$, then the left-vertices and right-vertices respectively are

$$\{v_1, v_2, \dots, v_m, v_{m+1}\} \quad \text{and} \quad \{v_{m+1}, v_{m+2}, \dots, v_{2m}, v_{2m+1}\}.$$

In this case, the center v_{m+1} is both a left-vertex and a right-vertex.

If $n = 2m$, then the left-vertices and right-vertices respectively are

$$\{v_1, v_2, \dots, v_m\} \quad \text{and} \quad \{v_{m+1}, v_{m+2}, \dots, v_{2m}\}.$$

If two vertices are both right-vertices or left-vertices, then we say that they are on the *same side*; otherwise, they are on *opposite sides*.

Lemma 3. *If n is odd, then for any $u, v \in V(P_n^4)$, we have*

$$d(u,v) = \begin{cases} \left\lceil \frac{1}{4}(L(u)+L(v)) \right\rceil & \text{if } u \text{ and } v \text{ are on opposite sides,} \\ \left\lceil \frac{1}{4}|L(u)-L(v)| \right\rceil & \text{if } u \text{ and } v \text{ are on the same side.} \end{cases}$$

If n is even, then for any $u, v \in V(P_n^4)$, we have

$$d(u,v) = \begin{cases} \left\lceil \frac{1}{4}(L(u)+L(v)+1) \right\rceil & \text{if } u \text{ and } v \text{ are on opposite sides,} \\ \left\lceil \frac{1}{4}|L(u)-L(v)| \right\rceil & \text{if } u \text{ and } v \text{ are on the same side.} \end{cases}$$

In the proof of Lemma 7 below, the following proposition will be used frequently:

Proposition 4. *For any d_1, d_2 in $\mathbb{N}$, we have*

$$\left\lceil \frac{d_1+d_2}{r} \right\rceil = \begin{cases} \lceil d_1/r \rceil + \lceil d_2/r \rceil - 1 & \text{if } (d_1, d_2) \equiv (l, m) \pmod r, \text{where } l \neq 0, m \neq 0, \\ & \text{and } 2 \le (d_1+d_2) \pmod r \le r, \\ \lceil d_1/r \rceil + \lceil d_2/r \rceil & \text{otherwise,} \end{cases}$$

$$\left\lceil \frac{d_1-d_2}{r} \right\rceil = \begin{cases} \lceil d_1/r \rceil - \lceil d_2/r \rceil + 1 & \text{if } (d_1, d_2) \equiv (0, m) \pmod r, \text{where } m \neq 0, \\ & \text{or } (d_1, d_2) \equiv (l, m) \pmod r, \text{where } l \neq 0, m \neq 0, \\ & \text{and } 1 \le (d_1-d_2) \pmod r \le (r-2), \\ \lceil d_1/r \rceil - \lceil d_2/r \rceil & \text{otherwise.} \end{cases}$$

It is important for the reader to understand the notation used in the labeling of P_n^4 so we will define a few terms and notation first.

Let $M, N \in \mathbb{N}$. We define a *block* (M, N) to be a pattern to follow when consecutively labeling a certain group of vertices in P_n^r. Take an (M, N)-block for example: The first vertex labeled, x_i, will have $L(x_i) \equiv M \pmod r$. The next vertex labeled, x_{i+1}, will have $L(x_{i+1}) \equiv N \pmod r$. The following vertex labeled, x_{i+2}, will have $L(x_{i+2}) \equiv M \pmod r$. Continue in this fashion until we end at a vertex of level congruent to $N \pmod r$. We may also choose to specify what side the vertex is on by writing $(\mathrm{L}M, \mathrm{R}N)$. This would mean that the first vertex labeled, x_i, would be a left-vertex with $L(x_i) \equiv M \pmod r$, and x_{i+1} would be a right-vertex with $L(x_{i+1}) \equiv N \pmod r$, so on and so forth.

We say that a *disconnection* occurs when $L(x_i)+L(x_{i+1})$ is not congruent to said specified value modulo r that maximizes the distance between two consecutively labeled vertices. This specific value changes depending upon the parity of n for P_n^4.

A *labeling pattern* is a specific arrangement of blocks. Note that the same block may appear multiple times in a labeling pattern; however, the number of vertices in each "identical" block may be different. For any labeling pattern, P_n^4 will be said to have an "even" pairing if, for each (M, N)-block in the labeling pattern, the number of vertices with level congruent to $M \pmod r$ on one side equals the number of vertices with level congruent to $N \pmod r$ on the other side. Otherwise, P_n^4 will be said to have "extra" vertices.

3. Lower bound of $\mathrm{rn}(P_n^4)$ when n is even

Lemma 5. *Let P_n^4 be a fourth power path on n vertices, where $n \geq 6$, and let $k = \mathrm{diam}(P_n^4) = \lceil \frac{1}{4}(n-1) \rceil$. If n is even, then*

$$\mathrm{rn}(P_n^4) \geq \begin{cases} 2k^2+1 & \text{if } n \equiv 0 \text{ or } 6 \pmod 8, \\ 2k^2 & \text{if } n \equiv 2 \pmod 8, \\ 2k^2+2 & \text{if } n \equiv 4 \pmod 8. \end{cases}$$

Proof. Let f be a radio labeling for P_n^4. Rearrange $V(P_n^4) = \{x_1, x_2, \ldots, x_n\}$ so that $0 = f(x_1) < f(x_2) < f(x_3) < \cdots < f(x_n)$. Note that $f(x_n)$ is the span of f. By definition, $f(x_{i+1}) - f(x_i) \geq k+1-d(x_i, x_{i+1})$ for $1 \leq i \leq n-1$. Summing up these $n-1$ inequalities, we have

$$f(x_n) \geq (n-1)(k+1) - \sum_{i=1}^{n-1} d(x_i, x_{i+1}). \tag{3-1}$$

Thus to minimize $f(x_n)$, it suffices to maximize $\sum_{i=1}^{n-1} d(x_i, x_{i+1})$. Since n is even,

$$\sum_{i=1}^{n-1} d(x_i, x_{i+1}) \leq \sum_{i=1}^{n-1} \left\lceil \tfrac{1}{4}(L(x_i)+L(x_{i+1})+1) \right\rceil.$$

Observe, from the above inequality we have:

(1) For each i, the equality for $d(x_i, x_{i+1}) \le \lceil \frac{1}{4}(L(x_i)+L(x_{i+1})+1)\rceil$ holds when x_i and x_{i+1} are on opposite sides, or when they are on the same side but one of them is a center and the other vertex is of level not congruent to 0 (mod 4).

(2) In the summation $\sum_{i=1}^{n-1} \lceil \frac{1}{4}(L(x_i)+L(x_{i+1})+1)\rceil$, each vertex of P_n^4 occurs exactly twice, except for x_1 and x_n, which both occur only once.

By direct calculation, we have

$$\left\lceil \tfrac{1}{4}(L(u)+L(v)+1)\right\rceil = \begin{cases} \frac{1}{4}(L(u)+L(v)+4) & \text{if } L(u)+L(v) \equiv 0 \pmod 4, \\ \frac{1}{4}(L(u)+L(v)+4)-\frac{1}{4} & \text{if } L(u)+L(v) \equiv 1 \pmod 4, \\ \frac{1}{4}(L(u)+L(v)+4)-\frac{2}{4} & \text{if } L(u)+L(v) \equiv 2 \pmod 4, \\ \frac{1}{4}(L(u)+L(v)+4)-\frac{3}{4} & \text{if } L(u)+L(v) \equiv 3 \pmod 4. \end{cases}$$

Therefore,

$$\left\lceil \tfrac{1}{4}(L(x_i)+L(x_{i+1})+1)\right\rceil \le \tfrac{1}{4}(L(x_i)+L(x_{i+1})+4),$$

and the equality holds only if $L(x_i)+L(x_{i+1}) \equiv 0 \pmod 4$. Combining this with (1) above, there exist at most $n-4$ of the i such that $d(x_i, x_{i+1}) = \frac{1}{4}(L(x_i)+L(x_{i+1})+4)$; that is, there are at least three disconnections in the labeling. Note that when $L(x_i)+L(x_{i+1}) \equiv 1, 2,$ or $3 \pmod 4$, we say that there is a disconnection between x_i and x_{i+1} of the *best type*, *second best type*, or the *worst type*, respectively. Moreover, among all the vertices, only the centers are of level zero. Hence, $L(x_1)+L(x_n) \ge 0+0 = 0$. We conclude that

$$\begin{aligned} \sum_{i=1}^{n-1} d(x_i, x_{i+1}) &\le \left(\sum_{i=1}^{n-1} \tfrac{1}{4}(L(x_i)+L(x_{i+1})+4)\right) - \tfrac{1}{4} - \tfrac{1}{4} - \tfrac{1}{4} \\ &= \tfrac{1}{4}\left(\left(2\sum_{i=1}^{n} L(x_i)\right) - L(x_1) - L(x_n)\right) + (n-1) - \tfrac{3}{4} \\ &\le \tfrac{1}{4}\left(\left(2\sum_{i=1}^{n} L(x_i)\right) - 0 - 0\right) + (n-1) - \tfrac{3}{4} \\ &= \tfrac{1}{2}\bigl(2\bigl(0+1+2+\cdots+\bigl(\tfrac{1}{2}n-1\bigr)\bigr)\bigr) + n - \tfrac{7}{4} \\ &= \tfrac{1}{8}n^2 + \tfrac{3}{4}n - \tfrac{7}{4}. \end{aligned}$$

By direct calculation for (3-1) and considering that $\mathrm{rn}(P_n^4)$ is an integer, we have

$$\mathrm{rn}(P_n^4) \ge \begin{cases} \lceil 2k^2+\frac{3}{4}\rceil = 2k^2+1 & \text{if } n \equiv 0 \pmod 8 \ \text{(i.e., } n = 4k \text{ and } k \text{ is even)}, \\ \lceil 2k^2-\frac{1}{4}\rceil = 2k^2 & \text{if } n \equiv 2 \pmod 8 \ \text{(i.e., } n = 4k-2 \text{ and } k \text{ is odd)}, \\ \lceil 2k^2+\frac{3}{4}\rceil = 2k^2+1 & \text{if } n \equiv 4 \pmod 8 \ \text{(i.e., } n = 4k \text{ and } k \text{ is odd)}, \\ \lceil 2k^2-\frac{1}{4}\rceil = 2k^2 & \text{if } n \equiv 6 \pmod 8 \ \text{(i.e., } n = 4k-2 \text{ and } k \text{ is even)}. \end{cases}$$

Further investigation for a sharper lower bound of $\mathrm{rn}(P_n^4)$ when $n \equiv 4$ or $6 \pmod 8$ is needed. There are three cases to consider based on the number of disconnections that occur in the labeling pattern.

Case 1: There are at least five disconnections. Then we have,

$$\sum_{i=1}^{n-1} d(x_i, x_{i+1}) \le \left(\sum_{i=1}^{n-1} \tfrac{1}{4}(L(x_i) + L(x_{i+1}) + 4) \right) - \tfrac{\mathbf{5}}{\mathbf{4}} \le \tfrac{1}{8}n^2 + \tfrac{3}{4}n - \tfrac{9}{4}.$$

Hence, by direct calculation for (3-1) we have

$$\mathrm{rn}(P_n^4) \ge \begin{cases} \left\lceil \left(2k^2 + \frac{3}{4}\right) + \frac{\mathbf{2}}{\mathbf{4}} \right\rceil = 2k^2 + 2 & \text{if } n \equiv 4 \pmod 8 \ \text{(i.e., } n = 4k \text{ and } k \text{ is odd),} \\ \left\lceil \left(2k^2 - \frac{1}{4}\right) + \frac{\mathbf{2}}{\mathbf{4}} \right\rceil = 2k^2 + 1 & \text{if } n \equiv 6 \pmod 8 \ \text{(i.e., } n = 4k-2 \text{ and } k \text{ is even).} \end{cases}$$

Case 2: There are exactly four disconnections. This case will be broken down into two subcases based on $L(x_1) + L(x_n)$.

Case 2.1: $L(x_1) + L(x_n) \ge 1$. Therefore,

$$\sum_{i=1}^{n-1} d(x_i, x_{i+1}) \le \left(\sum_{i=1}^{n-1} \tfrac{1}{4}(L(x_i) + L(x_{i+1}) + 4) \right) - \tfrac{\mathbf{4}}{4} \le \tfrac{1}{8}n^2 + \tfrac{3}{4}n - \tfrac{9}{4}.$$

Case 2.2: $L(x_1) + L(x_n) = 0$.

Claim. *In this case, at least two of the disconnections that occur cannot be of the best type.*

Proof of claim. For $n \equiv 4$ or $6 \pmod 8$, we have the following types of blocks as well as extra vertices (without loss of generality, we start each block with a left-vertex):

$$(\mathrm{L0}, \mathrm{R0}), \quad (\mathrm{L1}, \mathrm{R3}), \quad (\mathrm{L2}, \mathrm{R2}), \quad (\mathrm{L3}, \mathrm{R1}) \quad \mathrm{L1}, \quad \mathrm{R1}.$$

We wish to have exactly four disconnections and we also want $L(x_1) + L(x_n) = 0 + 0 = 0$ under this case. Therefore we must use two (L0, R0)-blocks. Thus our new blocks become (blocks are boxed for easy identification of disconnections that occur in the labeling pattern):

$$\boxed{(\mathrm{L0}, \mathrm{R0})}, \quad \boxed{(\mathrm{L1}, \mathrm{R3}) - \mathrm{L1}}, \quad \boxed{(\mathrm{L2}, \mathrm{R2})}, \quad \boxed{\mathrm{R1} - (\mathrm{L3}, \mathrm{R1})}, \quad \boxed{(\mathrm{L0}, \mathrm{R0})}.$$

Since we want $L(x_1) + L(x_n) = 0 + 0 = 0$, our labeling pattern must start and end with the (L0, R0)-blocks. Special attention is given to the "end-1" vertices, namely, the first and the last vertices of the two block patterns (L1, R3) − L1 and R1 − (L3, R1) from above. All disconnections in the labeling pattern will occur at these four end-1 vertices. The best type of disconnection would occur if an end-1 vertex was followed or preceded by a vertex whose level was congruent to $0 \pmod 4$. However, there are only two such vertices available. Therefore, at least two of the four end-1 vertices cannot have disconnections of the best type. □

By direct calculation, our claim, and the assumption that $L(x_1)+L(x_n)=0$, we have,

$$\sum_{i=1}^{n-1} d(x_i,x_{i+1}) \le \left(\sum_{i=1}^{n-1} \tfrac{1}{4}(L(x_i)+L(x_{i+1})+4)\right) - \tfrac{\mathbf{6}}{4} = \tfrac{1}{8}n^2+\tfrac{3}{4}n-\tfrac{10}{4}.$$

Hence, by direct calculation for (3-1) for the two subcases, the same bounds as in the conclusion of Case 1 are obtained.

Case 3: There are exactly three disconnections.

Claim. *In this case, at least one of the disconnections in the labeling pattern will not be of the best type.*

Proof of claim. Similar to Case 2.2, to ensure that there are only three disconnections, our new blocks must be

$$\boxed{\text{(L0, R0)}},\quad \boxed{\text{(L1, R3)} - \text{L1}},\quad \boxed{\text{(L2, R2)}},\quad \boxed{\text{R1} - \text{(L3, R1)}}.$$

Thus, out of the three disconnections that occur, at least two of them will occur at the end-1 vertices. Furthermore, out of the disconnections that occur at the end-1 vertices, at least one of them will not be of the best type, unless two (L0, R0)-blocks are used, which would increase the number of disconnections. □

By calculation, our claim, and noting that $L(x_1)+L(x_n)\ge 1$ under this case, we have,

$$\sum_{i=1}^{n-1} d(x_i,x_{i+1}) \le \left(\sum_{i=1}^{n-1} \tfrac{1}{4}(L(x_i)+L(x_{i+1})+4)\right) - \tfrac{\mathbf{4}}{4} \le \tfrac{1}{8}n^2+\tfrac{3}{4}n-\tfrac{9}{4}.$$

Direct calculation for (3-1) in this case also leads to the same bounds as in the conclusion of Case 1. □

4. Lower bound of $\mathrm{rn}(P_n^4)$ when n is odd

Lemma 6. *Let P_n^4 be a fourth power path on n vertices, where $n \ge 6$, and let $k = \operatorname{diam}(P_n^4) = \lceil \frac{1}{4}(n-1)\rceil$. If n is odd, then*

$$\mathrm{rn}(P_n^4) \ge \begin{cases} 2k^2+2 & \textit{if } n \equiv 1 \pmod 8 \textit{ and } n \ge 17 \textit{ or } n \equiv 5 \pmod 8,\\ 2k^2+1 & \textit{if } n \equiv 3 \textit{ or } 7 \pmod 8 \textit{ or } n=9.\end{cases}$$

Proof. We retain the same notation and employ the same method used in the proof of Lemma 5. Since n is odd,

$$\sum_{i=1}^{n-1} d(x_i,x_{i+1}) \le \sum_{i=1}^{n-1} \left\lceil \tfrac{1}{4}(L(x_i)+L(x_{i+1}))\right\rceil.$$

Observe, from the above inequality we have:

(1) For each i, the equality for $d(x_i, x_{i+1}) \le \left\lceil \frac{1}{4}(L(x_i) + L(x_{i+1}))\right\rceil$ holds only when x_i and x_{i+1} are on opposite sides, unless one of them is a center.

(2) In the summation $\sum_{i=1}^{n-1} \left\lceil \frac{1}{4}(L(x_i)+L(x_{i+1}))\right\rceil$, each vertex of P_n^4 occurs exactly twice, except x_1 and x_n, which each occurs only once.

By direct calculation, we have

$$\left\lceil \tfrac{1}{4}(L(u)+L(v))\right\rceil = \begin{cases} \frac{1}{4}(L(u)+L(v)+3)-\frac{3}{4} & \text{if } L(u)+L(v) \equiv 0 \pmod 4,\\ \frac{1}{4}(L(u)+L(v)+3) & \text{if } L(u)+L(v) \equiv 1 \pmod 4,\\ \frac{1}{4}(L(u)+L(v)+3)-\frac{1}{4} & \text{if } L(u)+L(v) \equiv 2 \pmod 4,\\ \frac{1}{4}(L(u)+L(v)+3)-\frac{2}{4} & \text{if } L(u)+L(v) \equiv 3 \pmod 4. \end{cases}$$

Therefore

$$\left\lceil \tfrac{1}{4}(L(x_i)+L(x_{i+1}))\right\rceil \le \tfrac{1}{4}(L(x_i)+L(x_{i+1})+3),$$

and the equality holds only if $L(x_i) + L(x_{i+1}) \equiv 1 \pmod 4$. Note that when $L(x_i)+L(x_{i+1}) \equiv 2, 3$, or $0 \pmod 4$, we say that there is a disconnection between x_i and x_{i+1} of the *best type*, *second best type*, or the *worst type*, respectively. Combining this with (1), there are two possible cases to consider based on the number of disconnections in the labeling pattern:

Case 1: There are at least three disconnections. In this case, since n is odd, there is only one center. Therefore, $L(x_1)+L(x_n) \ge 1$. Then,

$$\sum_{i=1}^{n-1} d(x_i, x_{i+1}) \le \left(\sum_{i=1}^{n-1} \tfrac{1}{4}(L(x_i)+L(x_{i+1})+3)\right) - \tfrac{\mathbf{3}}{\mathbf{4}} \le \tfrac{1}{8}n^2 + \tfrac{3}{4}n - \tfrac{15}{8}.$$

By direct calculation for (3-1), we have

$$\operatorname{rn}(P_n^4) \ge \begin{cases} 2k^2+1 & \text{if } n \equiv 1 \pmod 8 \ \text{(i.e., } n=4k+1 \text{ and } k \text{ is even)},\\ \left\lceil 2k^2+\frac{1}{2}\right\rceil = 2k^2+1 & \text{if } n \equiv 3 \pmod 8 \ \text{(i.e., } n=4k-1 \text{ and } k \text{ is odd)},\\ 2k^2+1 & \text{if } n \equiv 5 \pmod 8 \ \text{(i.e., } n=4k+1 \text{ and } k \text{ is odd)},\\ \left\lceil 2k^2+\frac{1}{2}\right\rceil = 2k^2+1 & \text{if } n \equiv 7 \pmod 8 \ \text{(i.e., } n=4k-1 \text{ and } k \text{ is even)}. \end{cases}$$

Case 2: There are exactly two disconnections. In this case, neither x_1 nor x_n is the center (denoted by C).

Case 2.1: $n \equiv 1 \pmod 8$. The labeling pattern must be a permutation of the boxed blocks

$$\boxed{\text{(L0, R1)} - \text{C} - \text{(L1, R0)}}, \quad \boxed{\text{(L2,R3)}}, \quad \boxed{\text{(L3,R2)}}.$$

Therefore, $L(x_1)+L(x_n) \ge 4$. By similar calculations to Case 1, we have

$$\sum_{i=1}^{n-1} d(x_i, x_{i+1}) \le \left(\sum_{i=1}^{n-1} \tfrac{1}{4}(L(x_i)+L(x_{i+1})+3)\right) - \tfrac{\mathbf{2}}{\mathbf{4}} \le \tfrac{1}{8}n^2 + \tfrac{3}{4}n - \tfrac{19}{8}.$$

By direct calculations, since $n = 4k+1$ and k is even, we have

$$\mathrm{rn}(P_n^4) \geq \left\lceil (2k^2+1) + \tfrac{2}{4} \right\rceil = 2k^2+2.$$

Case 2.2: $n \equiv 3, 5, \text{or } 7 \pmod 8$. Note that P_{8q+3}^4 and P_{8q+7}^4 both have an extra pair of vertices whose level is congruent to $1 \pmod 4$. Therefore, the labeling pattern must be a permutation of the boxed blocks

$$\boxed{\mathrm{R1} - (\mathrm{L0}, \mathrm{R1}) - \mathrm{C} - (\mathrm{L1}, \mathrm{R0}) - \mathrm{L1}}, \quad \boxed{(\mathrm{L2}, \mathrm{R3})}, \quad \boxed{(\mathrm{L3}, \mathrm{R2})}.$$

Now, P_{8q+5}^4 has two extra pairs of vertices whose levels are congruent to $1 \pmod 4$ and $2 \pmod 4$. The labeling pattern must be a permutation of the boxed blocks

$$\boxed{\mathrm{R1} - (\mathrm{L0}, \mathrm{R1}) - \mathrm{C} - (\mathrm{L1}, \mathrm{R0}) - \mathrm{L1}}, \quad \boxed{(\mathrm{L2}, \mathrm{R3}) - \mathrm{L2}}, \quad \boxed{\mathrm{R2} - (\mathrm{L3}, \mathrm{R2})}.$$

Therefore, for $n \equiv 3, 5, \text{or } 7 \pmod 8$, considering all possible permutations mentioned above, $L(x_1) + L(x_n) \geq 3$. Therefore,

$$\sum_{i=1}^{n-1} d(x_i, x_{i+1}) \leq \left(\sum_{i=1}^{n-1} \tfrac{1}{4}(L(x_i) + L(x_{i+1}) + 3) \right) - \tfrac{2}{4} \leq \tfrac{1}{8}n^2 + \tfrac{3}{4}n - \tfrac{17}{8}.$$

Thus, by direct calculation we have,

$$\mathrm{rn}(P_n^4) \geq \begin{cases} \left\lceil \left(2k^2+\frac12\right)+\frac14 \right\rceil = 2k^2+1 & \text{if } n \equiv 3 \pmod 8 \text{ (i.e., } n = 4k-1 \text{ and } k \text{ is odd)}, \\ \left\lceil (2k^2+1)+\frac14 \right\rceil = 2k^2+2 & \text{if } n \equiv 5 \pmod 8 \text{ (i.e., } n = 4k+1 \text{ and } k \text{ is odd)}, \\ \left\lceil \left(2k^2+\frac12\right)+\frac14 \right\rceil = 2k^2+1 & \text{if } n \equiv 7 \pmod 8 \text{ (i.e., } n = 4k-1 \text{ and } k \text{ is even)}. \end{cases}$$

Now assume $n \equiv 1 \pmod 8$ and $n \geq 17$; that is, $n = 4k+1$, k is even and $k \geq 4$. Assume to the contrary that $f(x_n) = 2k^2+1$. Then only Case 1 is possible and all of the following must hold:

(1) $\{x_1, x_n\} = \{v_{2k+1}, v_{2k+2}\}$ or $\{v_{2k+1}, v_{2k}\}$. That is, $\{x_1, x_n\}$ is of the form $\{x_1, x_n\} = \{\text{center, a vertex right next to center}\}$.

(2) $f(x_{i+1}) = f(x_i) + k + 1 - d(x_i, x_{i+1})$ for all i.

(3) For all $i \geq 1$, the two vertices x_i and x_{i+1} are on opposites sides unless one of them is the center.

(4) There exist three t-values, $1 \leq t \leq n-1$, such that $L(x_t) + L(x_{t+1}) \equiv 2 \pmod 4$ while $L(x_t) + L(x_{t+1}) \equiv 1 \pmod 4$ for all other $i \neq t$.

By (1) and by symmetry, we can assume that $x_1 = v_{2k+1}$; i.e., x_1 is the center. Excluding the center, there are $\frac{1}{2}k$ vertices whose level is congruent to $0 \pmod 4$, $1 \pmod 4$, $2 \pmod 4$, and $3 \pmod 4$ on each side, respectively. Since x_n is of level one, by (2), (3), and (4) we have:

(5) The labeling pattern must be the arrangement of boxed blocks

$$\boxed{\mathrm{C}-(1,0)}-\boxed{(2,3)}-\boxed{(3,2)}-\boxed{(0,1)}.$$

Claim. $\{v_1, v_n\} = \{x_{k+1}, x_{3k+2}\}$ *(i.e.,* $\{v_1, v_n\}$ *consists of the last vertex whose level is congruent to* $0 \pmod 4$ *in the* $(1, 0)$*-block and the first vertex whose level is congruent to* $0 \pmod 4$ *in the* $(0, 1)$*-block).*

Proof of claim. Suppose $v_1 \notin \{x_{k+1}, x_{3k+2}\}$. Then v_1 is inside one of the (0, 1)- or $(1, 0)$-blocks, since $L(v_1) = 2k \equiv 0 \pmod 4$. Let $v_1 = x_c$ for some c, where x_{c-1} and x_{c+1} are both vertices on the right side. Thus, $L(x_{c-1}) \equiv L(x_{c+1}) \equiv 1 \pmod 4$. Let $L(x_{c-1}) = y$ and $L(x_{c+1}) = z$. By (2),

$$f(x_c) - f(x_{c-1}) = \tfrac{1}{2}k + 1 - \left\lceil \tfrac{1}{4}y \right\rceil,$$
$$f(x_{c+1}) - f(x_c) = \tfrac{1}{2}k + 1 - \left\lceil \tfrac{1}{4}z \right\rceil.$$

Therefore,

$$f(x_{c+1}) - f(x_{c-1}) = k + 2 - \left\lceil \tfrac{1}{4}y \right\rceil - \left\lceil \tfrac{1}{4}z \right\rceil,$$

contradicting that

$$f(x_{c+1}) - f(x_{c-1}) \ge k + 1 - \left\lceil \tfrac{1}{4}|z - y| \right\rceil \quad (\text{as } y \equiv z \equiv 1 \pmod 4, \text{ so } y, z \neq 0).$$

Therefore $v_1 \in \{x_{k+1}, x_{3k+2}\}$. Similarly, we can show that $v_n \in \{x_{k+1}, x_{3k+2}\}$. □

By the claim, we may assume that $v_n = x_{k+1}$ and $v_1 = x_{3k+2}$ (the proof for the other case is symmetric). By (5), $L(x_k) = a \equiv 1 \pmod 4$ and $L(x_{k+2}) = b \equiv 2 \pmod 4$. By (2), (3), the fact that k is even, and our assumption that $L(x_{k+1}) = L(v_n) = L(v_{4k+1}) = 2k$, we have

$$f(x_{k+1}) - f(x_k) = \tfrac{1}{2}k + 1 - \left\lceil \tfrac{1}{4}a \right\rceil,$$
$$f(x_{k+2}) - f(x_{k+1}) = \tfrac{1}{2}k + 1 - \left\lceil \tfrac{1}{4}b \right\rceil,$$

and so,

$$f(x_{k+2}) - f(x_k) = k + 2 - \left\lceil \tfrac{1}{4}a \right\rceil - \left\lceil \tfrac{1}{4}b \right\rceil.$$

By definition and by Lemma 3,

$$f(x_{k+2}) - f(x_k) \ge k + 1 - \left\lceil \tfrac{1}{4}|a - b| \right\rceil.$$

Therefore, a must equal 1. Thus $L(x_k) = 1$, which means x_k is the level-one vertex on the left side, since $x_{k+1} = v_n$ is a right-vertex. Thus $x_k = v_{2k}$. Similarly, we can show that x_{3k+3} is of level one and on the right side. Thus, $x_{3k+3} = v_{2k+2}$.

Now, x_n is a right-vertex since $x_{3k+2} = v_1$ is a left-vertex, and so $x_n = v_{2k+2}$. This implies that $x_n = v_{2k+2} = x_{3k+3}$ and therefore $k = 2$, contradicting the assumption $k \ge 4$. Therefore $\mathrm{rn}(P_n^4) \ge 2k^2 + 2$ if $n \equiv 1 \pmod 8$ and $n \ge 17$.

Similar techniques can be applied for the case $n \equiv 5 \pmod 8$. Assume that $n \equiv 5 \pmod 8$ and $n \ge 21$; that is, $n = 4k + 1$, k is odd, and $k \ge 5$. Assume to the contrary that $f(x_n) = 2k^2 + 1$. Then only Case 1 is possible and the same requirements (1), (2), (3), and (4) for the case $n = 1 \pmod 8$ and $n \geq 17$ must hold.

By (1) and by symmetry, we can assume that $x_1 = v_{2k+1}$; i.e., x_1 is the center. Excluding the center, there are $\frac{1}{2}(k-1)$ vertices whose level is congruent to $0 \pmod 4$, $\frac{1}{2}(k+1)$ vertices whose level is congruent to $1 \pmod 4$, $\frac{1}{2}(k+1)$ vertices whose level is congruent to $2 \pmod 4$, and $\frac{1}{4}(k-1)$ vertices whose level is congruent to $3 \pmod 4$, on each side. By (1), (2), (3), and the second part of (4), the labeling pattern must be the arrangement of boxed blocks

$$\boxed{\mathrm{C}-(1-0-1)} - \boxed{(2-3-2)} - \boxed{(2-3-2)} - \boxed{(1-0-1)}.$$

However, in this arrangement the three t-values for which $L(x_t) + L(x_{t+1})$ is not congruent to $1 \pmod 4$ are not all congruent to $2 \pmod 4$, which contradicts the first part of (4). Therefore, $\mathrm{rn}(P_n^4) \geq 2k^2 + 2$. □

5. Upper bound and optimal radio labelings

To establish Theorem 1, it suffices to give radio labelings achieving the desired spans. To this end, we will use the next lemma, which provides us with an easy way to verify that a given labeling of P_n^r is indeed a radio labeling of P_n^r.

Lemma 7. *Let P_n^r be an r-th power path graph on n vertices, where $k = \mathrm{diam}(P_n^r) = \left\lceil \frac{1}{r}(n-1) \right\rceil$. Let $\{x_1, x_2, x_3, \dots, x_n\}$ be a permutation of $V(P_n^r)$ such that for any $1 \leq i \leq n-2$,*

$$\min\bigl\{d_{P_n}(x_i, x_{i+1}), d_{P_n}(x_{i+1}, x_{i+2})\bigr\} \leq \tfrac{1}{2}rk + 1$$

and $\max\{d_{P_n}(x_i, x_{i+1}), d_{P_n}(x_{i+1}, x_{i+2})\} \not\equiv 1 \pmod r$ if k is even and the equality in the above holds. Let f be a function, $f : V(P_n^r) \longrightarrow \{0, 1, 2, \dots\}$ with $f(x_1) = 0$ and $f(x_{i+1}) - f(x_i) = k + 1 - d(x_i, x_{i+1})$ for all $1 \leq i \leq n-1$. Then f is a radio labeling for P_n^r.

Before we present the proof of Lemma 7, note that Proposition 4 will be used frequently throughout the proof of Lemma 7 below. The construction of this proof is adapted from [Liu and Xie 2009].

Proof. Let f be a function satisfying the assumption. It suffices to prove that $f(x_j) - f(x_i) \geq k + 1 - d(x_i, x_j)$ for any $j \geq i + 2$. For $i = 1, 2, \dots, n-1$, set

$$f_i = f(x_{i+1}) - f(x_i).$$

For any $j \geq i + 2$, it follows that $f(x_j) - f(x_i) = f_i + f_{i+1} + f_{i+2} + \cdots + f_{j-1}$. We divide the proof into three cases:

Case 1: $j = i+2$. Assume $d(x_i, x_{i+1}) \geq d(x_{i+1}, x_{i+2})$ (the proof for $d(x_i, x_{i+1}) \leq d(x_{i+1}, x_{i+2})$ is similar). Then,

$$d(x_{i+1}, x_{i+2}) \leq \left\lceil \frac{\frac{1}{2}rk+1}{r} \right\rceil \leq \begin{cases} \frac{1}{2}(k+2) & \text{if } k \text{ is even,} \\ \frac{1}{2}(k+1) & \text{if } k \text{ is odd.} \end{cases}$$

Therefore, $d(x_{i+1}, x_{i+2}) \leq \frac{1}{2}(k+2)$. It suffices to consider the following subcases:

Case 1.1: x_i is between x_{i+1} and x_{i+2}. Then $d(x_i, x_{i+1}) \leq d(x_{i+1}, x_{i+2})$. Since we assume $d(x_i, x_{i+1}) \geq d(x_{i+1}, x_{i+2})$, we have $d(x_i, x_{i+1}) = d(x_{i+1}, x_{i+2}) \leq \frac{1}{2}(k+2)$ and $d_{P_n}(x_i, x_{i+2}) \leq (r-1)$, from which we have $d(x_i, x_{i+2}) = 1$. Hence,

$$\begin{aligned} f(x_{i+2}) - f(x_i) &= k+1-d(x_i, x_{i+1}) + k+1-d(x_{i+1}, x_{i+2}) \\ &\geq k+1-d(x_i, x_{i+2}). \end{aligned}$$

Case 1.2: x_{i+1} is between x_i and x_{i+2}. This implies

$$d(x_i, x_{i+2}) \geq d(x_i, x_{i+1}) + d(x_{i+1}, x_{i+2}) - 1.$$

Similar to the calculations above, we have $f(x_{i+2}) - f(x_i) \geq k+1-d(x_i, x_{i+2})$.

Case 1.3: x_{i+2} is between x_i and x_{i+1}. Assume k is odd or

$$\min\bigl\{d_{P_n}(x_i, x_{i+1}), d_{P_n}(x_{i+1}, x_{i+2})\bigr\} \leq \bigl(\tfrac{1}{2}rk+1\bigr)-1,$$

then we have $d(x_{i+1}, x_{i+2}) \leq \frac{1}{2}(k+1)$ and $d(x_i, x_{i+2}) \geq d(x_i, x_{i+1}) + d(x_{i+1}, x_{i+2})$. Hence, $f(x_{i+2}) - f(x_i) \geq k+1-d(x_i, x_{i+2})$. If k is even and

$$\min\bigl\{d_{P_n}(x_i, x_{i+1}), d_{P_n}(x_{i+1}, x_{i+2})\bigr\} = \tfrac{1}{2}rk+1,$$

then by our assumption, it must be that $d_{P_n}(x_{i+1}, x_{i+2}) = \frac{1}{2}rk+1 \equiv 1 \pmod{r}$ and $d_{P_n}(x_i, x_{i+1}) \not\equiv 1 \pmod{r}$. Thus we have,

$$d(x_i, x_{i+2}) = d(x_i, x_{i+1}) - d(x_{i+1}, x_{i+2}) + 1,$$

which implies

$$\begin{aligned} f(x_{i+2}) - f(x_i) &= 2k+2-\bigl(d(x_i, x_{i+2}) + d(x_{i+1}, x_{i+2}) - 1\bigr) - d(x_{i+1}, x_{i+2}) \\ &\geq k+1-d(x_i, x_{i+2}). \end{aligned}$$

Case 2: $j = i+3$.

Case 2.1: The sum of some pair of the distances $d(x_i, x_{i+1})$, $d(x_{i+1}, x_{i+2})$, and $d(x_{i+2}, x_{i+3})$ is at most $k+2$. Then,

$$\begin{aligned} f(x_{i+3}) - f(x_i) &\geq 3k+3-(k+2)-k \\ &> k+1-d(x_i, x_{i+3}). \end{aligned}$$

Case 2.2: The sum of any pair of the distances $d(x_i, x_{i+1})$, $d(x_{i+1}, x_{i+2})$, and $d(x_{i+2}, x_{i+3})$ is greater than $k+2$. If we then assume that $d(x_i, x_{i+1}) \geq d(x_{i+1}, x_{i+2})$ (the proof for $d(x_i, x_{i+1}) \leq d(x_{i+1}, x_{i+2})$ is similar), from the calculation in Case 1, we have $d(x_{i+1}, x_{i+2}) \leq \frac{1}{2}(k+2)$. By our hypothesis, it follows that $d(x_i, x_{i+1})$ and $d(x_{i+2}, x_{i+3})$ must both be greater than $\frac{1}{2}(k+2)$. This result, together with $\operatorname{diam}(P_n^r) = k$ and our assumption under this case, implies that x_i must appear before x_{i+2}, then x_{i+1}, then x_{i+3}, from left to right on the r-th power path (or x_{i+3} must appear before x_{i+1}, then x_{i+2}, then x_i). Therefore,

$$d(x_i, x_{i+3}) \geq d(x_i, x_{i+1}) + d(x_{i+2}, x_{i+3}) - d(x_{i+1}, x_{i+2}) - 1.$$

Therefore, we have

$$\begin{aligned} f(x_{i+3}) - f(x_i) &\geq 3k+3 - d(x_i, x_{i+3}) - 2d(x_{i+1}, x_{i+2}) - 1 \\ &\geq k+1 - d(x_i, x_{i+3}). \end{aligned}$$

Case 3: $j \geq i+4$. Since

$$\min\bigl\{d_{P_n}(x_i, x_{i+1}), d_{P_n}(x_{i+1}, x_{i+2})\bigr\} \leq \tfrac{1}{2}(k+2)$$

and $f_i \geq k+1-d(x_i, x_{i+1})$ for any i, we have $\max\{f_i, f_{i+1}\} \geq \frac{1}{2}k$ for any $1 \leq i \leq n-2$. Therefore,

$$\begin{aligned} f(x_j) - f(x_i) &\geq (f_i + f_{i+1}) + (f_{i+2} + f_{i+3}) \\ &\geq \bigl(\tfrac{1}{2}k+1\bigr) + \bigl(\tfrac{1}{2}k+1\bigr) > k+1 - d(x_i, x_j). \end{aligned} \qquad \square$$

When $\operatorname{diam}(P_n^r)$ is odd, we have the following “looser” condition for checking that a given labeling is indeed a radio labeling:

Lemma 8. *Let P_n^r be an r-th power path graph on n vertices, where $k = \operatorname{diam}(P_n^r) = \left\lceil \frac{1}{r}(n-1) \right\rceil$ is odd. Let $\{x_1, x_2, x_3, \ldots, x_n\}$ be a permutation of $V(P_n^r)$ such that for any $1 \leq i \leq n-2$,*

$$\min\bigl\{d_{P_n}(x_i, x_{i+1}), d_{P_n}(x_{i+1}, x_{i+2})\bigr\} \leq \tfrac{1}{2}r(k+1).$$

Let f be a function, $f : V(P_n^r) \longrightarrow \{0, 1, 2, \ldots\}$ with $f(x_1) = 0$ and $f(x_{i+1}) - f(x_i) = k+1-d(x_i, x_{i+1})$ for all $1 \leq i \leq n-1$. Then f is a radio labeling for P_n^r.

Proof. Assume $d(x_i, x_{i+1}) \geq d(x_{i+1}, x_{i+2})$ (the proof for $d(x_i, x_{i+1}) \leq d(x_{i+1}, x_{i+2})$ is similar). Then

$$d(x_{i+1}, x_{i+2}) \leq \left\lceil \frac{\frac{1}{2}r(k+1)}{r} \right\rceil = \tfrac{1}{2}k + 1 \leq \tfrac{1}{2}k + 2.$$

Note that this is the same conclusion we obtained in the beginning of the proof of Lemma 7. Therefore we can use exactly the same proof as above for the case when k is odd to prove this lemma. □

For each radio labeling f of P_n^4 given in the following, we shall first define a permutation (line-up) of the vertices $V(P_n^4) = \{x_1, x_2, x_3, \ldots, x_n\}$, then define f by $f(x_1) = 0$, and for all $1 \leq i \leq n-1$, $f(x_{i+1}) - f(x_i) = k+1-d(x_i, x_{i+1})$.

Case 1: $\mathrm{rn}(P^4_{8q+5}) \leq 2k^2+2$. Let $n = 8q+5$ for some $q \in \mathbb{N}$. Then $k = \mathrm{diam}(P^4_{8q+5}) = 2q+1$. We give a radio labeling with span $2k^2+2$. The line-up of $V(P^4_n) = \{x_1, x_2, \dots, x_n\}$ is given by the arrows in the display below. That is, x_1 is the center, x_2 is the left-vertex of P^4_n whose level is equal to $4q+1, \dots, x_n$ is the right-vertex of P^4_n whose level is equal to 2. The values above and below each arrow indicate the distances in P^4_n and P_n, respectively, between consecutively labeled vertices.

$$\begin{aligned}
&\mathrm{C}\xrightarrow[4q+1]{q+1}\mathrm{L}(4q+1)\xrightarrow[4q+5]{q+2}\mathrm{R4}\xrightarrow[4q+1]{q+1}\mathrm{L}(4q-3)\xrightarrow[4q+5]{q+2}\cdots\xrightarrow[4q+1]{q+1}\mathrm{L5}\xrightarrow[4q+5]{q+2}\mathrm{R}(4q)\xrightarrow[4q+1]{q+1}\mathrm{L1}\\
&\xrightarrow[4q+2]{q+1}\mathrm{R}(4q+1)\xrightarrow[4q+5]{q+2}\mathrm{L4}\xrightarrow[4q+1]{q+1}\mathrm{R}(4q-3)\xrightarrow[4q+5]{q+2}\mathrm{L8}\xrightarrow[4q+1]{q+1}\cdots\xrightarrow[4q+1]{q+1}\mathrm{R5}\xrightarrow[4q+5]{q+2}\mathrm{L}(4q)\xrightarrow[4q+1]{q+1}\mathrm{R1}\\
&\xrightarrow[4q+3]{q+1}\mathrm{L}(4q+2)\xrightarrow[4q+5]{q+2}\mathrm{R3}\xrightarrow[4q+1]{q+1}\mathrm{L}(4q-2)\xrightarrow[4q+5]{q+2}\mathrm{R7}\xrightarrow[4q+1]{q+1}\cdots\xrightarrow[4q+1]{q+1}\mathrm{L6}\xrightarrow[4q+5]{q+2}\mathrm{R}(4q-1)\xrightarrow[4q+1]{q+1}\mathrm{L2}\\
&\xrightarrow[4q+4]{q+1}\mathrm{R}(4q+2)\xrightarrow[4q+5]{q+2}\mathrm{L3}\xrightarrow[4q+1]{q+1}\mathrm{R}(4q-2)\xrightarrow[4q+5]{q+2}\mathrm{L7}\xrightarrow[4q+1]{q+1}\cdots\xrightarrow[4q+1]{q+1}\mathrm{R6}\xrightarrow[4q+5]{q+2}\mathrm{L}(4q-1)\xrightarrow[4q+1]{q+1}\mathrm{R2}.
\end{aligned}$$

By Lemma 8, f is a radio labeling for P^4_{8q+5}. Observe from the above display, there are two possible distances in P^4_{8q+5} between consecutively labeled vertices, namely, $q+1$ and $q+2$, with the number of occurrences $4q+4$ and $4q$, respectively. It follows by direct calculation that

$$f(x_{8q+5}) = (8q+4)(k+1) - \sum_{i=1}^{8q+4} d(x_i, x_{i+1}) = 2k^2+2.$$

Case 2: $\mathrm{rn}(P^4_{8q+4}) \leq 2k^2+2$. Let $n = 8q+4$ for some $q \in \mathbb{N}$. Then $k = \mathrm{diam}(P^4_{8q+4}) = 2q+1$. Let $G = P^4_{8q+5}$ and H be the subgraph of G induced by the vertices $\{v_1, v_2, \dots, v_{8q+4}\}$. Then $H \cong P^4_{8q+4}$, $\mathrm{diam}(H) = \mathrm{diam}(G) = 2q+1$, and $d_G(u, v) = d_H(u, v)$ for every $u, v \in V(H)$. Let f be a radio labeling for G, then $f|_H$ is also a radio labeling for H. By Case 1, $\mathrm{rn}(P^4_{8q+4}) \leq \mathrm{rn}(P^4_{8q+5}) \leq 2k^2+2$.

Case 3: $\mathrm{rn}(P^4_{8q+3}) \leq 2k^2+1$. Let $n = 8q+3$ for some $q \in \mathbb{N}$. Then $k = \mathrm{diam}(P^4_{8q+3}) = 2q+1$. Similar to Case 1, we line up the vertices according to the display below.

$$\begin{aligned}
&\mathrm{C}\xrightarrow[4q+1]{q+1}\mathrm{L}(4q+1)\xrightarrow[4q+5]{q+2}\mathrm{R4}\xrightarrow[4q+1]{q+1}\mathrm{L}(4q-3)\xrightarrow[4q+5]{q+2}\cdots\xrightarrow[4q+1]{q+1}\mathrm{L5}\xrightarrow[4q+5]{q+2}\mathrm{R}(4q)\xrightarrow[4q+1]{q+1}\mathrm{L1}\\
&\xrightarrow[4q+2]{q+1}\mathrm{R}(4q+1)\xrightarrow[4q+5]{q+2}\mathrm{L4}\xrightarrow[4q+1]{q+1}\mathrm{R}(4q-3)\xrightarrow[4q+5]{q+2}\mathrm{L8}\xrightarrow[4q+1]{q+1}\cdots\xrightarrow[4q+1]{q+1}\mathrm{R5}\xrightarrow[4q+5]{q+2}\mathrm{L}(4q)\xrightarrow[4q+1]{q+1}\mathrm{R1}\\
&\xrightarrow[4q-1]{q}\mathrm{L}(4q-2)\xrightarrow[4q+1]{q+1}\mathrm{R3}\xrightarrow[4q-3]{q}\mathrm{L}(4q-6)\xrightarrow[4q+1]{q+1}\mathrm{R7}\xrightarrow[4q-3]{q}\cdots\xrightarrow[4q-3]{q}\mathrm{L2}\xrightarrow[4q+1]{q+1}\mathrm{R}(4q-1)\\
&\xrightarrow[4q+2]{q+1}\mathrm{L3}\xrightarrow[4q+1]{q+1}\mathrm{R}(4q-2)\xrightarrow[4q+5]{q+2}\mathrm{L7}\xrightarrow[4q+1]{q+1}\mathrm{R}(4q-6)\xrightarrow[4q+5]{q+2}\cdots\xrightarrow[4q+5]{q+2}\mathrm{L}(4q-1)\xrightarrow[4q+1]{q+1}\mathrm{R2}.
\end{aligned}$$

By Lemma 7, f is a radio labeling for P^4_{8q+3}. If follows by direct calculation that

$$f(x_{8q+3}) = (8q+2)(k+1) - \sum_{i=1}^{8q+2} d(x_i, x_{i+1}) = 2k^2+1.$$

Case 4: $\mathrm{rn}(P^4_{8q+2}) \le 2k^2$. Let $n = 8q+2$ for some $q \in \mathbb{N}$. Then $k = \mathrm{diam}(P^4_{8q+2}) = 2q+1$. Similarly, we line up the vertices according to the display below.

$$\begin{aligned}
&\mathrm{R}0 \xrightarrow[4q+1]{q+1} \mathrm{L}(4q) \xrightarrow[4q+5]{q+2} \mathrm{R}4 \xrightarrow[4q+1]{q+1} \mathrm{L}(4q-4) \xrightarrow[4q+5]{q+2} \cdots \xrightarrow[4q+1]{q+1} \mathrm{L}4 \xrightarrow[4q+5]{q+2} \mathrm{R}(4q)\\
&\xrightarrow[4q+2]{q+1} \mathrm{L}1 \xrightarrow[4q+1]{q+1} \mathrm{R}(4q-1) \xrightarrow[4q+5]{q+2} \mathrm{L}5 \xrightarrow[4q+1]{q+1} \mathrm{R}(4q-5) \xrightarrow[4q+5]{q+2} \cdots \xrightarrow[4q+5]{q+2} \mathrm{L}(4q-3) \xrightarrow[4q+1]{q+1} \mathrm{R}3\\
&\xrightarrow[4q+2]{q+1} \mathrm{L}(4q-2) \xrightarrow[4q+1]{q+1} \mathrm{R}2 \xrightarrow[4q-3]{q} \mathrm{L}(4q-6) \xrightarrow[4q+1]{q+1} \mathrm{R}6 \xrightarrow[4q-3]{q} \cdots \xrightarrow[4q-3]{q} \mathrm{L}2 \xrightarrow[4q+1]{q+1} \mathrm{R}(4q-2)\\
&\xrightarrow[4q+2]{q+1} \mathrm{L}3 \xrightarrow[4q+1]{q+1} \mathrm{R}(4q-3) \xrightarrow[4q+5]{q+2} \mathrm{L}7 \xrightarrow[4q+1]{q+1} \mathrm{R}(4q-7) \xrightarrow[4q+5]{q+2} \cdots \xrightarrow[4q+5]{q+2} \mathrm{L}(4q-1) \xrightarrow[4q+1]{q+1} \mathrm{R}1 \xrightarrow[2]{1} \mathrm{L}0.
\end{aligned}$$

By Lemma 7, f is a radio labeling for P^4_{8q+2}. If follows by direct calculation that

$$f(x_{8q+2}) = (8q+1)(k+1) - \sum_{i=1}^{8q+1} d(x_i, x_{i+1}) = 2k^2.$$

Case 5: $\mathrm{rn}(P^4_{8q+1}) \le 2k^2 + q$. Let $n = 8q+1$ for some $q \in \mathbb{N}$. Then $k = \mathrm{diam}(P^4_{8q+1}) = 2q$. Similarly, we line up the vertices according to the display below.

$$\begin{aligned}
&\mathrm{C} \xrightarrow[4q-3]{q} \mathrm{L}(4q-3) \xrightarrow[4q+1]{q+1} \mathrm{R}4 \xrightarrow[4q-3]{q} \mathrm{L}(4q-7) \xrightarrow[4q+1]{q+1} \cdots \xrightarrow[4q-3]{q} \mathrm{L}1 \xrightarrow[4q+1]{q+1} \mathrm{R}(4q)\\
&\xrightarrow[8q-2]{2q} \mathrm{L}(4q-2) \xrightarrow[4q+1]{q+1} \mathrm{R}3 \xrightarrow[4q-3]{q} \mathrm{L}(4q-6) \xrightarrow[4q+1]{q+1} \mathrm{R}7 \xrightarrow[4q-3]{q} \cdots \xrightarrow[4q-3]{q} \mathrm{L}2 \xrightarrow[4q+1]{q+1} \mathrm{R}(4q-1)\\
&\xrightarrow[8q-2]{2q} \mathrm{L}(4q-1) \xrightarrow[4q+1]{q+1} \mathrm{R}2 \xrightarrow[4q-3]{q} \mathrm{L}(4q-5) \xrightarrow[4q+1]{q+1} \mathrm{R}6 \xrightarrow[4q-3]{q} \cdots \xrightarrow[4q-3]{q} \mathrm{L}3 \xrightarrow[4q+1]{q+1} \mathrm{R}(4q-2)\\
&\xrightarrow[8q-2]{2q} \mathrm{L}(4q) \xrightarrow[4q+1]{q+1} \mathrm{R}1 \xrightarrow[4q-3]{q} \mathrm{L}(4q-4) \xrightarrow[4q+1]{q+1} \mathrm{R}5 \xrightarrow[4q-3]{q} \cdots \xrightarrow[4q-3]{q} \mathrm{L}4 \xrightarrow[4q+1]{q+1} \mathrm{R}(4q-3).
\end{aligned}$$

By Lemma 7, f is a radio labeling for P^4_{8q+1}. It follows by direct calculation that

$$f(x_{8q+1}) = (8q)(k+1) - \sum_{i=1}^{8q} d(x_i, x_{i+1}) = 2k^2 + q.$$

Case 6: $\mathrm{rn}(P^4_{8q}) \le 2k^2 + 1$. Let $n = 8q$ for some $q \in \mathbb{N}$. Then $k = \mathrm{diam}(P^4_{8q}) = 2q$. Similarly, we line up the vertices according to the display below.

$$\begin{aligned}
&\mathrm{R}0 \xrightarrow[4q-3]{q} \mathrm{L}(4q-4) \xrightarrow[4q+1]{q+1} \mathrm{R}4 \xrightarrow[4q-3]{q} \mathrm{L}(4q-8) \xrightarrow[4q+1]{q+1} \cdots \xrightarrow[4q-3]{q} \mathrm{L}4 \xrightarrow[4q+1]{q+1} \mathrm{R}(4q-4)\\
&\xrightarrow[8q-6]{2q-1} \mathrm{L}(4q-3) \xrightarrow[4q+1]{q+1} \mathrm{R}3 \xrightarrow[4q-3]{q} \mathrm{L}(4q-7) \xrightarrow[4q+1]{q+1} \mathrm{R}7 \xrightarrow[4q-3]{q} \cdots \xrightarrow[4q-3]{q} \mathrm{L}1 \xrightarrow[4q+1]{q+1} \mathrm{R}(4q-1)\\
&\xrightarrow[8q-2]{2q} \mathrm{L}(4q-2) \xrightarrow[4q+1]{q+1} \mathrm{R}2 \xrightarrow[4q-3]{q} \mathrm{L}(4q-6) \xrightarrow[4q+1]{q+1} \mathrm{R}6 \xrightarrow[4q-3]{q} \cdots \xrightarrow[4q-3]{q} \mathrm{L}2 \xrightarrow[4q+1]{q+1} \mathrm{R}(4q-2)\\
&\xrightarrow[8q-2]{2q} \mathrm{L}(4q-1) \xrightarrow[4q+1]{q+1} \mathrm{R}1 \xrightarrow[4q-3]{q} \mathrm{L}(4q-5) \xrightarrow[4q+1]{q+1} \mathrm{R}5 \xrightarrow[4q-3]{q} \cdots \xrightarrow[4q-3]{q} \mathrm{L}3 \xrightarrow[4q+1]{q+1} \mathrm{R}(4q-3) \xrightarrow[4q-2]{q} \mathrm{L}0.
\end{aligned}$$

By Lemma 7, f is a radio labeling for P^4_{8q}. It follows by direct calculation that

$$f(x_{8q}) = (8q-1)(k+1) - \sum_{i=1}^{8q-1} d(x_i, x_{i+1}) = 2k^2 + 1.$$

Case 7: $\mathrm{rn}(P^4_{8q-2}) \le \mathrm{rn}(P^4_{8q-1}) \le 2k^2+1$. Since $k = \mathrm{diam}(P^4_{8q-2}) = \mathrm{diam}(P^4_{8q-1}) = \mathrm{diam}(P^4_{8q}) = 2q$, using the same subgraph argument as in Case 2, we have that $\mathrm{rn}(P^4_{8q-2}) \le \mathrm{rn}(P^4_{8q-1}) \le \mathrm{rn}(P^4_{8q}) \le 2k^2+1$.

Cases 1–7, together with Lemmas 5 and 6, complete the proof of Theorem 1.

References

[Chartrand, Erwin and Zhang 2005] G. Chartrand, D. Erwin, and P. Zhang, "A graph labeling problem suggested by FM channel restrictions", *Bull. Inst. Combin. Appl.* **43** (2005), 43–57. MR 2005h:05175 Zbl 1066.05125

[Chartrand et al. 2001] G. Chartrand, D. Erwin, F. Harary, and P. Zhang, "Radio labelings of graphs", *Bull. Inst. Combin. Appl.* **33** (2001), 77–85. MR 2003d:05185 Zbl 0989.05102

[Hale 1980] W. K. Hale, "Frequency assignment: theory and applications", *Proc. IEEE* **68**:12 (1980), 1497–1514.

[Liu 2008] D. D.-F. Liu, "Radio number for trees", *Discrete Math.* **308**:7 (2008), 1153–1164. MR 2009h:05180 Zbl 1133.05090

[Liu and Xie 2004] D. D.-F. Liu and M. Xie, "Radio number for square of cycles", *Congr. Numer.* **169** (2004), 101–125. MR 2005m:05198 Zbl 1064.05089

[Liu and Xie 2009] D. D.-F. Liu and M. Xie, "Radio number for square paths", *Ars Combin.* **90** (2009), 307–319. MR 2010b:05141 Zbl 1224.05451

[Liu and Zhu 2005] D. D.-F. Liu and X. Zhu, "Multilevel distance labelings for paths and cycles", *SIAM J. Discrete Math.* **19**:3 (2005), 610–621. MR 2006i:05142 Zbl 1095.05033

[Lo 2010] M.-L. Lo, "Radio number for cube paths", unpublished manuscript, 2010.

[Sooryanarayana et al. 2010] B. Sooryanarayana, M. Vishu Kumar, and K. Manjula, "Radio number of cube of a path", *Int. J. Math. Comb.* **1** (2010), 5–29. MR 2662413 Zbl 1203.05136

[Zhang 2002] P. Zhang, "Radio labelings of cycles", *Ars Combin.* **65** (2002), 21–32. MR 2003i:05117 Zbl 1071.05573

Received: 2014-11-24 Revised: 2015-04-12 Accepted: 2015-04-12

mlo@csusb.edu
Department of Mathematics, California State University,
San Bernardino, San Bernardino, CA 92407, United States

linda.alegria05@gmail.com
Department of Mathematics, California State University,
San Bernardino, San Bernardino, CA 92407, United States

INVOLVE 9:2 (2016)
dx.doi.org/10.2140/involve.2016.9.333

On closed graphs, II

David A. Cox and Andrew Erskine

(Communicated by Colin Adams)

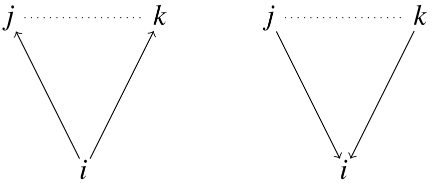
A graph is closed when its vertices have a labeling by $[n]$ with a certain property first discovered in the study of binomial edge ideals. In this article, we explore various aspects of closed graphs, including the number of closed labelings and clustering coefficients.

1. Introduction

Given a simple graph G with vertices $V(G)$ and edges $E(G)$, a *labeling* of G is a bijection $V(G) \simeq [n] = \{1, \dots, n\}$. Given a labeling, we assume $V(G) = [n]$.

Definition 1.1. A labeling of G is *closed* when $\{j, i\}, \{i, k\} \in E(G)$ with $j > i < k$ or $j < i > k$ implies $\{j, k\} \in E(G)$. We say that G is *closed* if it has a closed labeling.

A labeling of G gives a direction to each edge $\{i, j\} \in E(G)$ where the arrow points from i to j when $i < j$; that is, the arrow points to the bigger label. In this context, closed means that when two edges point away from a vertex or towards a vertex, the remaining vertices are connected by an edge, as shown below:



(1-1)

Closed graphs were first encountered in the study of binomial edge ideals defined in [Herzog et al. 2010; Ohtani 2011]. Properties of these ideals are explored in [Ene et al. 2011; Saeedi Madani and Kiani 2012] and their relation to closed graphs features in [Crupi and Rinaldo 2011; Ene et al. 2014; 2015; Ene and Zarojanu 2015].

It is natural to ask for a characterization of those graphs that have a closed labeling. One solution was given in [Crupi and Rinaldo 2011], which characterizes closed graphs using the clique complex of G. Another approach, taken in our

MSC2010: primary 05C75; secondary 05C25, 05C78.
Keywords: closed graph, clustering coefficient.

previous paper [Cox and Erskine 2015], shows that a connected graph is closed if and only if it is chordal, claw-free, and narrow (see [loc. cit., Definition 1.3] for the definition of narrow).

In this paper, we will use tools developed in [Cox and Erskine 2015] to study the combinatorial properties of closed graphs. Our main results include:

- Section 4: Theorem 4.3 counts the number of closed labelings of a closed graph.
- Section 5: Theorem 5.4 counts the number of closed graphs with fixed layer structure (see Section 2 for the definition of layer).
- Section 6: Theorem 6.3 gives a sharp lower bound for the clustering coefficient of a closed graph.

To prepare for these results, we will recall some relevant results and definitions in Section 2 and explore when a labeling remains closed after exchanging two labels in Section 3.

2. Notation and known results

We recall some notation and results from [Cox and Erskine 2015]. The *neighborhood* of $v \in V(G)$ is

$$N_G(v) = \{w \in V(G) \mid \{v, w\} \in E(G)\}.$$

When G is labeled and $i \in V(G) = [n]$, we have a disjoint union

$$N_G(i) = N_G^{>}(i) \cup N_G^{<}(i),$$

where

$$N_G^{>}(i) = \{j \in N_G(i) \mid j > i\} \quad \text{and} \quad N_G^{<}(i) = \{j \in N_G(i) \mid j < i\}.$$

Also, vertices $i, j \in [n]$ with $i \leq j$ give the *interval* $[i, j] = \{k \in [n] \mid i \leq k \leq j\}$.

Here is a characterization of when a labeling of a connected graph is closed.

Proposition 2.1 [Cox and Erskine 2015, Proposition 2.4]. *A labeling on a connected graph G is closed if and only if for all $i \in [n]$, the set $N_G^{>}(i)$ is a complete subgraph and is an interval.*

When a connected graph G has a labeling with $V(G) = [n]$, we can decompose G into layers as follows. The *N-th layer of G* is the set L_N of all vertices that are distance N from vertex 1; i.e.,

$$L_N = \{i \in [n] \mid i \text{ is distance } N \text{ from } 1\}.$$

Since G is connected, we have a disjoint union

$$[n] = L_0 \cup L_1 \cup \cdots \cup L_h, \tag{2-1}$$

where $h = \max\{N \mid L_N \neq \varnothing\}$. Here is a simple property of layers.

Lemma 2.2 [Cox and Erskine 2015, Lemma 2.6]. *Let G be labeled and connected. If $i \in L_N$ and $\{i, j\} \in E(G)$, then $j \in L_{N-1}, L_N$, or L_{N+1}.*

When G is closed and connected, the layers are especially nice.

Proposition 2.3 [Cox and Erskine 2015, Proposition 2.7]. *If G is connected with a closed labeling, then*:

(1) *Each layer L_N is complete.*

(2) *If $d = \max\{L_N\}$, then $L_{N+1} = N_G^{>}(d)$.*

The diameter of G is denoted $\operatorname{diam}(G)$, and a *longest shortest path* of G is a shortest path of length $\operatorname{diam}(G)$. These concepts relate to layers as follows.

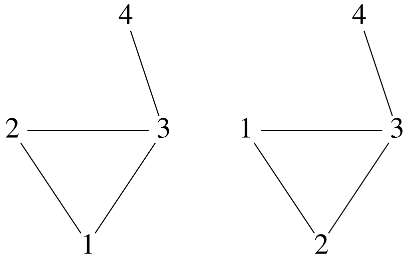
Proposition 2.4 [Cox and Erskine 2015, Proposition 2.8]. *If G is connected with a closed labeling, then*:

(1) $\operatorname{diam}(G)$ *is the integer h appearing in* (2-1).

(2) *If P is a longest shortest path of G, then one endpoint of P is in L_0 or L_1 and the other is in L_h, where $h = \operatorname{diam}(G)$.*

3. Exchangeable vertices

A closed graph with at least two vertices has at least two closed labelings, since the reversal of a closed labeling is clearly closed. But there may be other closed labelings, as shown by this simple example:



(3-1)

To explore what makes this example work, we need some definitions.

Definition 3.1. Let G be a graph.

(1) The *full neighborhood* of a vertex $v \in V(G)$ is $N_G^\star(v) = \{v\} \cup N_G(v)$.

(2) $v, w \in V(G)$ are *exchangeable*, written $v \sim w$, if $N_G^\star(v) = N_G^\star(w)$.

Vertices 1 and 2 are exchangeable in the left-hand graph of (3-1). Switching labels gives the right-hand graph, which is still closed. Here is the general result.

Proposition 3.2. *Let G have a closed labeling. If $i, j \in [n]$, where $i \neq j$, are exchangeable, then the labeling that switches i and j is also closed.*

Proof. Define $\phi : [n] \to [n]$ by $\phi(i) = j$, $\phi(j) = i$, and $\phi(k) = k$ for $k \in [n] \setminus \{i, j\}$. Pick $u, v, w \in V(G)$ with $\{u, v\}, \{v, w\} \in E(G)$, $u \neq w$, and $\phi(u) > \phi(v) < \phi(w)$ or $\phi(u) < \phi(v) > \phi(w)$. We need to prove that $\{u, w\} \in E(G)$.

If $\{i, j\} \cap \{u, v, w\} = \varnothing$, then $\{u, w\} \in E(G)$ since the original labeling is closed. Now suppose $\{i, j\} \cap \{u, v, w\} \neq \varnothing$ and $\phi(u) > \phi(v) < \phi(w)$. There are several cases to consider. First suppose that $i = v$. If $j \in \{u, w\}$, then without loss of generality we may assume $j = u$. Then

$$w \in N_G^\star(v) = N_G^\star(i) = N_G^\star(j) = N_G^\star(u)$$

implies $\{u, w\} \in E(G)$. If $j \notin \{u, w\}$, then $\phi(u) > \phi(i) < \phi(w)$ means that $u > j < w$. Then $\{u, w\} \in E(G)$ since the original labeling is closed and $j \sim i = v$.

The proof when $j = v$ is similar and is omitted. Then two cases remain:

- $i = u$ and $j \notin \{v, w\}$. Thus $\phi(u) > \phi(v) < \phi(w)$ means that $j > v < w$. Then $\{j, w\} \in E(G)$ since the original labeling is closed and $j \sim i = u$. Using $j \sim i = u$ again, we conclude that $\{u, w\} \in E(G)$.
- $i = u$ and $j = w$. Then $\phi(u) > \phi(v) < \phi(w)$ means $j > v < i$. Then $\{u, w\} = \{i, j\} \in E(G)$ since the original labeling is closed.

The proof when $\phi(u) < \phi(v) > \phi(w)$ is similar and is omitted. □

Exchangeability, denoted $v \sim w$, is an equivalence relation on $V(G)$ with equivalence classes

$$e(v) = \{w \in V(G) \mid w \sim v\} = \{w \in V(G) \mid N_G^\star(w) = N_G^\star(v)\}.$$

Equivalence classes are complete, since $v \sim w$ implies $v \in N_G^\star(v) = N_G^\star(w)$, so that $\{v, w\} \in E(G)$ whenever $v \neq w$.

Since permutations are generated by transpositions, Proposition 3.2 implies that when G has a closed labeling, every permutation of an equivalence class yields a new closed labeling.

When G is connected and closed, equivalence classes have the following structure.

Proposition 3.3. *If G is connected with a closed labeling and $i \in [n]$, then the equivalence class $e(i)$ is an interval.*

Proof. It suffices to show that if i and j are exchangeable and $i < k < j$, then $N_G^\star(k) = N_G^\star(i)$. First note that $\{i, k\} \in E(G)$ since $j \in N_G^>(i)$ and $N_G^>(i)$ is an interval by Proposition 2.1. Then $\{j, k\} \in E(G)$ since $i \sim j$.

Now take $m \in N_G^\star(k)$. We need to show $m \in N_G^\star(i)$. If $m = k$, this follows from the previous paragraph. If $\{m, k\} \in E(G)$, there are two possibilities:

- If $m < k$, then $m < k > i$, so $\{m, i\} \in E(G)$ since the labeling is closed.
- If $m > k$, then $m > k < j$, so either $m = j$ or $\{m, j\} \in E(G)$ since the labeling is closed.

Since $N^\star_G(i) = N^\star_G(j)$, both possibilities imply $m \in N^\star_G(i)$.

Conversely, take $m \in N^\star_G(i)$. If $m = i$, then $m \in N^\star_G(k)$ since $\{i, k\} \in E(G)$ by the first paragraph of the proof. If $\{m, i\} \in E(G)$, then $\{m, j\} \in E(G)$ since $i \sim j$. Again, there are two possibilities:

- If $m < i$, then $m < i < k < j$, so $\{m, k\} \in E(G)$ since $N^>_G(m)$ is an interval.
- If $m > i$, then $m > i < k$, so either $m = k$ or $\{m, k\} \in E(G)$ since the labeling is closed.

Thus $m \in N^\star_G(k)$ and the proof is complete. □

4. Counting closed labelings

Some graphs have no nontrivial exchangeable vertices.

Definition 4.1. A graph G is *collapsed* if all exchangeable vertices are equal, i.e., $N^\star_G(v) = N^\star_G(w)$ implies $v = w$.

Proposition 4.2. *Let G be a closed graph with at least three vertices. Then the following are equivalent*:

(1) *G has exactly two closed labelings.*

(2) *G is connected and collapsed.*

Proof. The proof of (1) ⇒ (2) is easy. If G is not connected, then G is a disjoint union $G = G_1 \cup G_2$, where G_i is closed. We may assume G_1 has at least two vertices, so G_1 has at least two labelings. Then we get at least four closed labelings of G: two where 1 is in G_1, and two where 1 is in G_2. Also, if G is not collapsed, then some equivalence class $e(i)$ has at least two elements. If $|e(i)| \geq 3$, then switching labels within $e(i)$ gives at least six closed labelings, and if $|e(i)| = 2$, then G has at least one more vertex, which makes it easy to see that G has at least four closed labelings.

The proof of (2) ⇒ (1) will take more work. First note that $\operatorname{diam}(G) = h \geq 2$. This follows because $h = 1$ would imply that G is complete, which is impossible since G is collapsed with at least 3 vertices, and $h = 0$ is impossible since G is connected with at least 3 vertices.

Fix a closed labeling with $V(G) = [n]$. This gives layers $L_0 = \{1\}, L_1, \dots, L_h$ associated with the labeling, and Proposition 2.4(2) implies that every longest shortest path has one endpoint in L_0 or L_1 and the other in L_h.

Let $\phi : [n] \to [n]$ be another closed labeling which we will call the ϕ-labeling. Pick $1' \in [n]$ such that $\phi(1') = 1$. Then some longest shortest path of G begins at $1'$. By the previous paragraph, $1' \in L_0 \cup L_1$ or $1' \in L_h$. Replacing ϕ with its reversal if necessary, we may assume that $1' \in L_0 \cup L_1$. We claim that ϕ is the identity function. This will prove the theorem.

We first show that $1' = 1$, i.e, $\phi(1) = 1$. Recall that $L_1 = N_G(1)$ and that L_1 is complete by Proposition 2.3(1). It follows that $N_G^\star(1) = L_0 \cup L_1$ is also complete. The same argument implies that $N_G^\star(1')$ is complete. Now suppose $1 \neq 1'$ and pick $m \in N_G^\star(1')$ different from 1. Then $\{1, m\} \in E(G)$ since $1 \in N_G^\star(1')$ and $N_G^\star(1')$ is complete. This implies $m \in L_1 = N_G(1)$, and then the inclusion $N_G^\star(1') \subseteq N_G^\star(1)$ follows easily. The opposite inclusion follows by interchanging the two labelings. Hence we have proved $N_G^\star(1') = N_G^\star(1)$. Since we are assuming $1 \neq 1'$, this contradicts the fact that G is collapsed. Hence we must have $1' = 1$, as claimed.

Now suppose that vertices $1, \dots, u-1 \in [n]$ have the same ϕ-label as in the original labeling, i.e., $\phi(j) = j$ for $1 \le j \le u-1$. Then pick $u' \in [n]$ such that $\phi(u') = u$. To prove that $u' = u$, i.e., $\phi(u) = u$, suppose that $u' \neq u$. Since ϕ is the identity on $1, \dots, u-1$ and $\phi(u') = u$, we have $u' > u$ and $\phi(u') < \phi(u)$.

We first show that $\{u, u'\} \in E(G)$. Since G is connected, Proposition 2.1 implies that every vertex is connected by an edge to its successor in any closed labeling. For the original labeling, this gives $\{u-1, u\} \in E(G)$, and for the ϕ-labeling, this gives $\{u-1, u'\} \in E(G)$ since $\phi(u-1) = u-1$ and $\phi(u') = u$. Proposition 2.1 implies that $N_G^>(u-1)$ (in the original labeling) is complete, and $\{u, u'\} \in E(G)$ follows.

We next prove that $N_G^\star(u) \subseteq N_G^\star(u')$. Pick $m \in N_G^\star(u)$. Then:

- If $m = u$, then $m \in N_G^\star(u')$ since $\{u, u'\} \in E(G)$.
- If $m > u$, then either $m = u'$, in which case $m \in N_G^\star(u')$ is obvious, or $m \neq u'$, in which case $m \in N_G^\star(u')$ since $m > u < u'$ implies $\{m, u'\} \in E(G)$ as the original labeling is closed.
- If $m < u$, then $m \in N_G^\star(u')$ since $\phi(m) = m < u < \phi(u) > \phi(u')$ implies $\{m, u'\} \in E(G)$ as the ϕ-labeling is closed.

This proves $N_G^\star(u) \subseteq N_G^\star(u')$. By symmetry, we get $N_G^\star(u') = N_G^\star(u)$, which contradicts $u' \neq u$ since G is collapsed. We conclude that $u' = u$, and then ϕ is the identity by induction on u. □

Now suppose that G is a connected graph with a closed labeling. Since each equivalence class is an interval by Proposition 3.3, we can order the equivalence classes

$$E_1 < E_2 < \cdots < E_r \tag{4-1}$$

so that if $i \in E_a$ and $j \in E_b$, then $i < j$ if and only if $a < b$. This induces an ordering on $V(G)/{\sim} = \{E_1, \dots, E_r\}$. Then define the graph $G/{\sim}$ with vertices

$$V(G/{\sim}) = V(G)/{\sim} = \{E_1, \dots, E_r\} \tag{4-2}$$

and edges

$$E(G/{\sim}) = \big\{\{E_a, E_b\} \mid \{i, j\} \in E(G) \text{ for some } i \in E_a, j \in E_b\big\}. \tag{4-3}$$

Since $i \sim i'$ and $j \sim j'$ imply that $\{i, j\} \in E(G)$ if and only if $\{i', j'\} \in E(G)$, we can replace "for some" with "for all" in (4-3).

Theorem 4.3. *Let G be connected with a closed labeling and exchangeable equivalence classes $E_1, \dots, E_r$. Then*:

(1) *The quotient graph $G/{\sim}$ defined in* (4-2) *and* (4-3) *is connected*, *collapsed*, *and closed with respect to the labeling* (4-1).

(2) *If $r > 1$, then G has precisely $2\prod_{a=1}^{r} |E_a|!$ closed labelings.*

Proof. For (1), we omit the straightforward proof that $G/{\sim}$ is connected and closed with respect to (4-1). To prove that $G/{\sim}$ is collapsed, we first observe that for vertices $u, v \in V(G)$,

$$u \in N_G^\star(v) \iff e(u) \in N_G^\star(e(v)). \tag{4-4}$$

We leave the simple proof to the reader. Now suppose that equivalence classes $e(v), e(w)$ satisfy $e(v) \sim e(w)$. Then by (4-4), we have

$$u \in N_G^\star(v) \iff e(u) \in N_G^\star(e(v)) \iff e(u) \in N_G^\star(e(w)) \iff u \in N_G^\star(w).$$

This proves that $N_G^\star(v) = N_G^\star(w)$. Then $v \sim w$, which implies $e(v) = e(w)$. It follows that $G/{\sim}$ is collapsed.

For (2), first note that $r > 1$ implies $r \geq 3$, for if there were only two equivalence classes E_1 and E_2, then since G is connected there must be $\{v, w\} \in E(G)$ with $v \in E_1$ and $w \in E_2$. The observation following (4-3) implies that $\{s, t\} \in E(G)$ for all $s \in E_1$ and $t \in E_2$. It follows easily that G is complete, which implies $r = 1$, a contradiction. Hence $r \geq 3$.

According to Proposition 4.2, $G/{\sim}$ has exactly two closed labelings since it has $r \geq 3$ vertices by the previous paragraph and is connected, closed, and collapsed by (1). It follows from (4-1) that any closed labeling of G induces one of these two closed labelings of $G/{\sim}$. Hence all closed labelings of G arise from the two ways of ordering the equivalence classes, together with how we order elements within each equivalence class. Proposition 3.2 and the remarks following the proposition imply that we can use any of the $|E|!$ orderings of the elements of an equivalence class E. Since different equivalence classes can be ordered independently of each other, we get the desired formula for the total number of closed orderings of G. □

5. Counting closed graphs

In Theorem 4.3, we fixed a connected graph and counted the number of closed labelings. Here we change the point of view, where we fix a labeling and count the number of connected graphs for which the given labeling is closed.

Here is how a layer of a connected closed graph connects to the next layer.

Definition 5.1. Let G be a connected graph with a closed labeling. Let the layers of G be $L_0 = \{1\}, L_1, \dots, L_h$, $h = \operatorname{diam}(G)$.

(1) Let $a_N = |L_N|$ for $N = 0, \dots, h$. Note that $a_0 = 1$.

(2) If $N < h$, write the vertices of L_N in order. For $1 \leq s \leq a_N$, let b_s be the number of edges of G connecting the s-th vertex of L_N to a vertex of L_{N+1}.

(3) The *sequence* of L_N is the sequence $S_N = (b_1, b_2, \dots, b_{a_N})$.

Here is some further notation we will need. First, let $m_N = \min\{L_N\}$. Propositions 2.1 and 2.3 imply that L_N is complete and is an interval. Thus $L_N = [m_N, m_N + a_N - 1]$, and the s-th vertex of L_N is $u_s = m_N + s - 1$.

We can now show that the sequence $S_N = (b_1, b_2, \dots, b_{a_N})$ determines precisely how L_N is connected to L_{N+1}.

Proposition 5.2. *Let G be connected with a closed labeling. If $u_s = m_N + s - 1 \in L_N$ is the s-th vertex of L_N and $b_s > 0$, then*

$$\{v \in L_{N+1} \mid \{u_s, v\} \in E(G)\} = [m_{N+1}, m_{N+1} + b_s - 1].$$

Thus b_s determines how u_s links to L_{N+1}.

Proof. Let $A = \{v \in L_{N+1} \mid \{u_s, v\} \in E(G)\}$. Note that every $v \in A$ satisfies $v > u_s$ by Proposition 2.3(2). It follows easily that

$$A = N_G^{>}(u_s) \cap L_{N+1}.$$

We know that L_{N+1} is an interval, and the same is true for $N_G^{>}(u_s)$ by Proposition 2.1. Hence A is an interval. However, if $v \in A$ and $v \neq m_{N+1}$, then $m_{N+1} < v > u_s$ and the fact that the labeling is closed imply $\{u_s, m_{N+1}\} \in E(G)$ since $\{m_{N+1}, v\} \in E(G)$ by the completeness of L_{N+1}. Hence $m_{N+1} \in A$, and from here, the proposition follows without difficulty. □

Here is an important property of the sequence S_N.

Proposition 5.3. *Let G be connected with a closed labeling. If $N < \operatorname{diam}(G)$, then the sequence $S_N = (b_1, b_2, \dots, b_{a_N})$ of the layer L_N has the following properties*:

(1) *The last element of S_N is a_{N+1}*; *i.e., $b_{a_N} = a_{N+1}$.*

(2) *S_N is increasing*; *i.e., $b_s \leq b_{s+1}$ for $s = 1, \dots, a_N - 1$.*

Proof. For (1), note that the last vertex of L_N connects to every vertex of L_{N+1} by Proposition 2.3(2). It follows that $b_{a_N} = |L_{N+1}| = a_{N+1}$.

For (2), let u_s be the s-th vertex of L_N, with $1 \leq s \leq a_N - 1$. If $b_s = 0$, then $b_s \leq b_{s+1}$ clearly holds. If $b_s > 0$, then u_s connects to $m_{N+1} + b_s - 1$ by Proposition 5.2, and it connects to u_{s+1} since L_N is complete. Then $m_{N+1} + b_s - 1 > u_s < u_{s+1}$

implies that u_{s+1} connects to $m_{N+1}+b_s-1$ since the labeling is closed. Using Proposition 5.2 again, we obtain

$$m_{N+1}+b_s-1 \in [m_{N+1}, m_{N+1}+b_{s+1}-1],$$

and $b_s \leq b_{s+1}$ follows. □

We now come to the main result of this section.

Theorem 5.4. *Fix n and an integer partition $n = a_0+a_1+\cdots+a_h$, with $a_0=1$ and $a_N \geq 1$ for $N=1,\dots,h$. Also set $\mathcal{L}_0=\{1\}$ and*

$$\mathcal{L}_N = [a_0+\cdots+a_{N-1}+1, a_0+\cdots+a_N] \tag{5-1}$$

for $N=1,\dots,h$, so that $|\mathcal{L}_N| = a_N$. Then the number of graphs G satisfying the conditions

(1) *$V(G)=[n]$,*

(2) *G is connected and closed with respect to the labeling $V(G)=[n]$, and*

(3) *the N-th layer of G is $\mathcal{L}_N$ for $N=0,\dots,h$*

is given by the product

$$\prod_{N=0}^{h-1}\binom{a_{N+1}+a_N-1}{a_N-1}.$$

Proof. Let G satisfy (1), (2) and (3). Each layer of G is complete, and every edge of G connects to the same layer or an adjacent layer by Lemma 2.2. Then Proposition 5.2 shows that the edges of G are uniquely determined by $S_0,\dots,S_{h-1}$.

By Proposition 5.3, each $S_N = (b_1, b_2, \dots, b_{a_N})$ is an increasing sequence of nonnegative integers of length a_N that ends at a_{N+1}. It is well known that the number of such sequences equals the binomial coefficient

$$\binom{a_{N+1}+a_N-1}{a_N-1}.$$

It follows that the product in the statement of the proposition is an upper bound for the number of graphs satisfying (1), (2) and (3).

To complete the proof, we need to show that every sequence counted by the product corresponds to a graph G satisfying (1), (2) and (3). First note that the minimal element of $\mathcal{L}_N$ is

$$m_N = a_0+\cdots+a_{N-1}+1$$

when $N>0$. Now suppose we have sequences $S_0,\dots,S_{h-1}$, where each $S_N = (b_1, b_2, \dots, b_{a_N})$ is an increasing sequence of nonnegative integers of length a_N that ends at a_{N+1}. This determines a graph G with $V(G)=[n]$ and the following edges:

(A) All possible edges connecting elements in the same level $\mathcal{L}_N$.

(B) For each $N = 0, \dots, h-1$, all edges $\{u_s, v\}$, where u_s is the s-th vertex of $\mathcal{L}_N$ and v is any vertex in the interval $[m_{N+1}, m_{N+1} + b_s - 1] \subseteq \mathcal{L}_{N+1}$ from Proposition 5.2.

Once we prove that G is closed and connected with $\mathcal{L}_N$ as its N-th layer, the theorem will be proved.

Since $b_{a_N} = a_{N+1}$, we see that for $N = 0, \dots, h-1$, the last element of $\mathcal{L}_N$ connects to all elements of $\mathcal{L}_{N+1}$. This enables us to construct a path from 1 to any $u \in \mathcal{L}_N$ for $N = 1, \dots, h$. It follows that G is connected and that all $u \in \mathcal{L}_N$ have distance at most N from vertex 1. Since every edge of G connects elements of $\mathcal{L}_M$ to $\mathcal{L}_M$, $\mathcal{L}_{M+1}$, or $\mathcal{L}_{M-1}$, any path connecting 1 to $u \in \mathcal{L}_N$ must have length at least N. It follows that $\mathcal{L}_N$ is indeed the N-th layer of G.

It remains to show that G is closed with respect to the natural labeling given by $V(G) = [n]$. A vertex of G is the s-th vertex u_s of $\mathcal{L}_N$ for some s and N. We will show that $N_G^>(u_s)$ satisfies Proposition 2.1. The formula (5-1) for $\mathcal{L}_N$ and the description of the edges of G given in (A) and (B) make it clear that

$$\begin{aligned} N_G^>(u_s) &= [u_{s+1}, a_0 + \cdots + a_N] \cup [m_{N+1}, m_{N+1} + b_s - 1] \\ &= [u_{s+1}, m_{N+1} + b_s - 1], \end{aligned}$$

where the second equality follows from $m_{N+1} = a_0 + \cdots + a_N + 1$. To show that $N_G^>(u_s)$ is complete, take distinct vertices $v, w \in N_G^>(u_s)$. If both lie in $\mathcal{L}_N$ or $\mathcal{L}_{N+1}$, then $\{v, w\} \in V(G)$ by (A). Otherwise, we may assume without loss of generality that $v = u_t$, where $t \geq s$, and $w \in [m_{N+1}, m_{N+1} + b_s - 1]$. Note that u_t links to every vertex in $[m_{N+1}, m_{N+1} + b_t - 1]$ by (B). We also have $b_s \leq b_t$ since S_N is increasing. It follows that $\{v, w\} = \{u_t, w\} \in E(G)$. Hence $N_G^>(u_s)$ is complete, so that G is closed by Proposition 2.1. □

6. Local clustering coefficients

In a social network, one can ask how often a friend of a friend is also a friend. Translated into graph theory, this asks how often a path of length two has an edge connecting the endpoints of the path. The illustration (1-1) from the Introduction indicates that this should be a frequent occurrence in a closed graph.

There are several ways to quantify the "friend of a friend" phenomenon. For our purposes, the most convenient is the *local clustering coefficient* of vertex v of a graph G, which is defined by

$$C_v = \begin{cases} \dfrac{\text{number of pairs of neighbors of } v \text{ connected by an edge}}{\text{number of pairs of neighbors of } v} & \text{if } \deg(v) \geq 2, \\ 0 & \text{if } \deg(v) \leq 1. \end{cases}$$

Local clustering coefficients are discussed in [Newman 2010, pp. 201–204].

Proposition 6.1. *Let v be a vertex of a closed graph G of degree $d = \deg(v) \geq 2$. Then the local clustering coefficient C_v satisfies the inequality*

$$C_v \geq \frac{1}{2} - \frac{1}{2(d-1)}.$$

Furthermore, $d \geq 3$ implies that $C_v \geq \frac{1}{3}$.

Proof. Pick a closed labeling of G and let $a = |N_G^{>}(v)|$ and $b = |N_G^{<}(v)|$. Then $a + b = |N_G(v)| = \deg(v) = d$. Since the labeling is closed, any pair of vertices in $N_G^{>}(v)$ or in $N_G^{<}(v)$ is connected by an edge. It follows that at least

$$\tfrac{1}{2}a(a-1) + \tfrac{1}{2}b(b-1)$$

pairs of neighbors of v are connected by an edge. Since the total number of such pairs is $\frac{1}{2}d(d-1)$ and $d = a + b$, we obtain

$$C_v \geq \frac{a(a-1)+b(b-1)}{d(d-1)} = \frac{a^2+b^2-d}{d(d-1)} \geq \frac{\frac{1}{2}d^2-d}{d(d-1)} = \frac{1}{2} - \frac{1}{2(d-1)}, \tag{6-1}$$

where we use $a^2 + b^2 - \frac{1}{2}d^2 = \frac{1}{2}(a-b)^2 \geq 0$. When $d \geq 4$, this inequality for C_v easily gives $C_v \geq \frac{1}{3}$. When $d = 3$, then $a + b = 3$, with $a, b \in \mathbb{Z}$, implies that $a^2 + b^2 \geq 5$, in which case the left half of (6-1) gives

$$C_v \geq \frac{5-3}{3(3-1)} = \frac{1}{3}. \qquad \square$$

A global version of the clustering coefficient defined by Watts and Strogatz is

$$C_{\mathrm{WS}} = \frac{1}{n} \sum_{v \in V(G)} C_v, \quad n = |V(G)|.$$

(See reference [323] of [Newman 2010]. A different global clustering coefficient is discussed in [loc. cit., pp. 199–204].) To estimate C_{WS} for a closed graph, we need the following lemma.

Lemma 6.2. *Let G be a connected closed graph.*

(1) *Set $h = \mathrm{diam}(G)$ and let c be the number of vertices $v \in G$ with $\deg(v) = 2$ and $C_v = 0$. Then $c \leq h - 1$.*

(2) *G has at most two leaves.*

Proof. For (1), fix a closed labeling for G with $V(G) = [n]$ and pick $v \in V(G)$ with $\deg(v) = 2$ and $C_v = 0$. We claim that v is in a layer of its own. To see why, let $v \in L_N$ and suppose there is $s \in L_N$ with $s \neq v$. Then $\{v, s\} \in E(G)$ since layers are complete by Proposition 2.3(1). Furthermore, $|L_N| \geq 2$, so $N > 0$. Then $\{s, d\}$, $\{v, d\} \in E(G)$ for $d = \max\{L_{N-1}\}$ by Proposition 2.3(2). Since $\deg(v) = 2$, we must have $N_G(v) = \{s, d\}$, and then $\{s, d\} \in E(G)$ contradicts $C_v = 0$. Thus $\{v\}$ is a layer when $\deg(v) = 2$ and $C_v = 0$.

Note that if $\{v\} = L_0$, then the two vertices in $N_G(v) = L_1$ would be linked by an edge. The same holds if $\{v\} = L_h$, for here the two vertices would be in L_{h-1} since L_h is the highest layer by Proposition 2.4(1). It follows that each of the c vertices with $\deg(v) = 2$ and $C_v = 0$ lies in a separate layer distinct from L_0 or L_h. Since there are only $h - 1$ intermediate layers, we must have $c \leq h - 1$.

For (2), assume G has leaves u, v, w and fix a closed labeling of G. We may assume $u < v < w$, and let u', v', w' be the unique vertices adjacent to u, v, w respectively. A shortest path from u to v is directed (see [Herzog et al. 2010] or Proposition 2.1 of [Cox and Erskine 2015]) and must pass through u' and v', hence $u < u' \leq v' < v$ since $u < v$. The same argument applied to v and w would imply $v < v' \leq w' < w$. Thus $v' < v$ and $v < v'$, so three leaves cannot exist. □

We can now estimate the clustering coefficient C_{WS} of a closed graph.

Theorem 6.3. *If G is connected and closed with $n > 1$ vertices and diameter h, then*

$$C_{\mathrm{WS}} \geq \frac{1}{3} - \frac{h+1}{3n}.$$

Proof. Since $n > 1$ and G is connected, all vertices of G have degree ≥ 1. Thus we can write $V(G)$ as the disjoint union

$$V(G) = \mathcal{A} \cup \mathcal{B} \cup \mathcal{C} \cup \mathcal{D},$$

where $\mathcal{A}$ consists of vertices of degree ≥ 3, $\mathcal{B}$ consists of vertices of degree 2 with $C_v = 1$, $\mathcal{C}$ consists of vertices of degree 2 with $C_v = 0$, and $\mathcal{D}$ consists of the leaves (which have $C_v = 0$). Since $C_v \geq \frac{1}{3}$ for $v \in \mathcal{A}$ by Proposition 6.1, we have

$$C_{\mathrm{WS}} \geq \tfrac{1}{n}\big(\tfrac{1}{3} \cdot |\mathcal{A}| + 1 \cdot |\mathcal{B}| + 0 \cdot |\mathcal{C}| + 0 \cdot |\mathcal{D}|\big) \geq \frac{|\mathcal{A}| + |\mathcal{B}|}{3n} = \frac{n - (|\mathcal{C}| + |\mathcal{D}|)}{3n}.$$

Then we are done since $|\mathcal{C}| \leq h - 1$ and $|\mathcal{D}| \leq 2$ by Lemma 6.2. □

By Theorem 6.3, the clustering coefficient C_{WS} is large when the diameter is small compared to the number of vertices. At the other extreme, both sides of the inequality in Theorem 6.3 are zero when G is a path graph.

Acknowledgements

We are grateful to Amherst College for the Post-Baccalaureate Summer Research Fellowship that supported the writing of this paper. Thanks also to Amy Wagaman for suggesting that we look at clustering coefficients.

References

[Cox and Erskine 2015] D. A. Cox and A. Erskine, “On closed graphs, I”, *Ars Combin.* **120** (2015), 259–274. MR 3363281

[Crupi and Rinaldo 2011] M. Crupi and G. Rinaldo, “Binomial edge ideals with quadratic Gröbner bases”, *Electron. J. Combin.* **18**:1 (2011), Paper 211. MR 2012k:13047 Zbl 1235.13024

[Ene and Zarojanu 2015] V. Ene and A. Zarojanu, “On the regularity of binomial edge ideals”, *Math. Nachr.* **288**:1 (2015), 19–24. MR 3310496 Zbl 1310.13021

[Ene et al. 2011] V. Ene, J. Herzog, and T. Hibi, “Cohen–Macaulay binomial edge ideals”, *Nagoya Math. J.* **204** (2011), 57–68. MR 2012j:13032 Zbl 1236.13011

[Ene et al. 2014] V. Ene, J. Herzog, and T. Hibi, “Koszul binomial edge ideals”, pp. 125–136 in *Bridging algebra, geometry, and topology*, edited by D. Ibadula and W. Veys, Springer Proc. Math. Stat. **96**, Springer, Cham, 2014. MR 3297112 Zbl 06515927

[Ene et al. 2015] V. Ene, J. Herzog, and T. Hibi, “Linear flags and Koszul filtrations”, *Kyoto J. Math.* **55**:3 (2015), 517–530. MR 3395974 Zbl 06489502

[Herzog et al. 2010] J. Herzog, T. Hibi, F. Hreinsdóttir, T. Kahle, and J. Rauh, “Binomial edge ideals and conditional independence statements”, *Adv. in Appl. Math.* **45**:3 (2010), 317–333. MR 2011j:13041 Zbl 1196.13018

[Newman 2010] M. E. J. Newman, *Networks: An introduction*, Oxford University Press, 2010. MR 2011h:05002 Zbl 1195.94003

[Ohtani 2011] M. Ohtani, “Graphs and ideals generated by some 2-minors”, *Comm. Algebra* **39**:3 (2011), 905–917. MR 2012e:13020 Zbl 1225.13028

[Saeedi Madani and Kiani 2012] S. Saeedi Madani and D. Kiani, “Binomial edge ideals of graphs”, *Electron. J. Combin.* **19**:2 (2012), Paper 44. MR 2946102 Zbl 1262.13012

Received: 2014-12-30 Accepted: 2015-04-05

dacox@amherst.edu
Department of Mathematics and Statistics, Amherst College,
Amherst, MA 01002-5000, United States

aperskine@gmail.com
Department of Mathematics and Statistics, Amherst College,
Amherst, MA 01002-5000, United States

INVOLVE 9:2 (2016)
dx.doi.org/10.2140/involve.2016.9.347

Klein links and related torus links

Enrique Alvarado, Steven Beres, Vesta Coufal,
Kaia Hlavacek, Joel Pereira and Brandon Reeves

(Communicated by Colin Adams)

In this paper, we present our constructions and results leading up to our discovery of a class of Klein links that are not equivalent to any torus links. In particular, we calculate the number and types of components in a $K_{p,q}$ Klein link and show that $K_{p,p} \equiv K_{p,p-1}$, $K_{p,2} \equiv T_{p-1,2}$, and $K_{2p,2p} \equiv T_{2p,p}$. Finally, we show that in contrast to the fact that every Klein knot is a torus knot, no Klein link $K_{p,p}$, where $p \geq 5$ is odd, is equivalent to a torus link.

1. Introduction

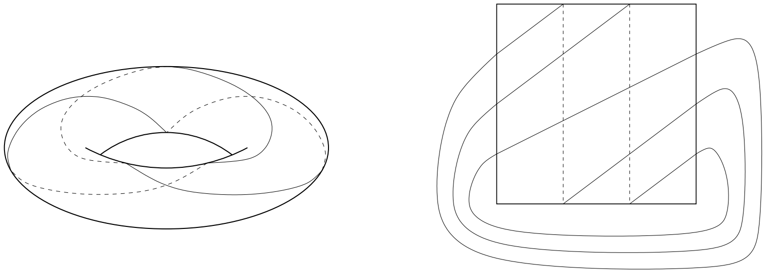
When we began thinking about Klein knots, we were told that they were uninteresting since all Klein knots are torus knots. We decided to see if we could prove that statement using elementary methods, and whether it was also true for Klein links. In this paper, we present our constructions and results leading up to our discovery of a class of Klein links that are not equivalent to any torus links. While results identical or similar to Theorems 2, 3, 4 and 5 are also proved in [Bowen et al. 2013; Bush et al. 2014; Freund and Smith-Polderman 2013; Shepherd et al. 2012] using braids, our approach uses different constructions and methods.

2. Constructions

Our construction of Klein knots and links is modeled after the standard construction of torus knots and links, as in [Adams 2004; Murasugi 2008]. Recall that for nonnegative integers p and q, the torus link $T_{p,q}$ is the link on the torus which crosses the longitude p times and crosses the meridian q times, with no crossing on the torus itself. We illustrate the construction of $T_{2,3}$ in Figure 1. Notice that we can translate the construction to a planar diagram as in Figure 1.

We will construct Klein knots and links in a similar way, being careful of certain issues. The first difficulty is that Klein bottles do not exist in three-dimensional

MSC2010: 57M25.
Keywords: knot theory, Klein links, torus links.
The authors received partial support from McDonald Work Awards .



(a) Constructing $T_{2,3}$ on a torus. (b) Planar diagram for $T_{2,3}$.

Figure 1. Torus knot $T_{2,3}$.

space, and knots are trivial in four-dimensional space. To get around this, we will work with punctured Klein bottles in three-dimensional space. The puncture occurs where the Klein bottle appears to (but does not) intersect itself. Warning: the knots and links we work with will be dependent on the relative position of the puncture. Mimicking the construction of $T_{p,q}$, the Klein link $K_{2,3}$ is illustrated in Figure 2. The corresponding planar diagram representation of $K_{2,3}$ is again modeled after the torus version, except that we need to account for the Möbius-band twist and be mindful of the puncture. By deforming the Klein bottle as in Figure 3, we see that the twist produces a pattern of additional crossings as in Figure 4, with the puncture occurring in the lower left corner. Note that $K_{p,0}$ is the p-component unlink.

In general, we construct $K_{p,q}$ on the planar diagram just as for $T_{p,q}$, except with the pattern of extra crossings. See Figure 5. We emphasize that the class of links that we are denoting by $K_{p,q}$ and the results in this paper are dependent on placing the puncture in the lower left corner. We do not consider Klein links with the puncture placed in different positions in this paper. Furthermore, deformations of our links are as links in space, not on the Klein bottle, and so the puncture does

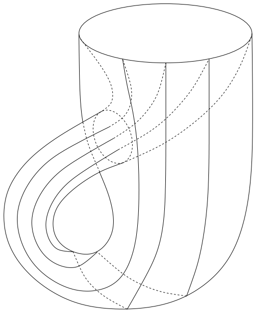
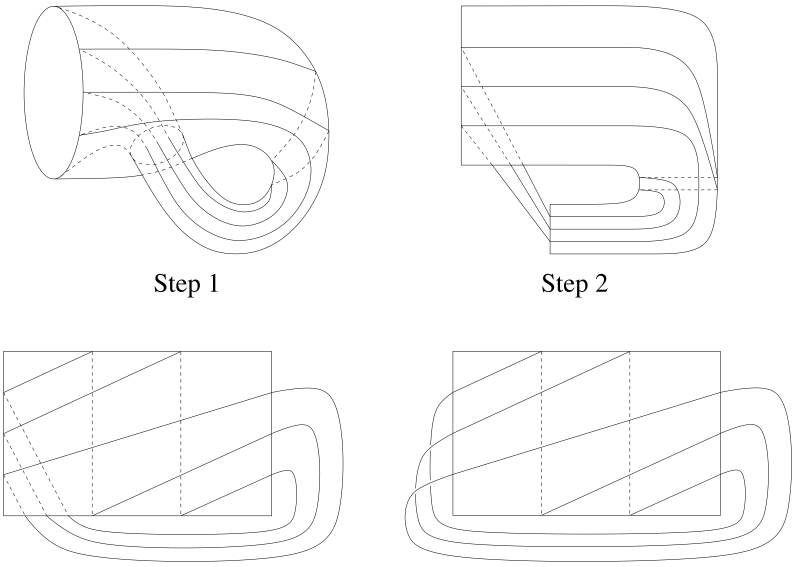


Figure 2. Klein link $K_{2,3}$.



Step 1

Step 2

Step 3

Step 4

Figure 3. Deformations of $K_{2,3}$.

not affect deformations. For this reason, and since our puncture is always in the lower left corner, we do not include it in our illustrations.

It is worth noting that, while the diagrams are configured a bit differently, our $K_{p,q}$ Klein links are the same as the $K(p,q)$ Klein links found in [Bowen et al. 2013; Bush et al. 2014; Freund and Smith-Polderman 2013; Shepherd et al. 2012]. Additionally, some of the same authors of the previously cited papers have done preliminary work in which they found explicit relationships between Klein links with different choices of puncture. There are certainly more questions to be answered in this regard.

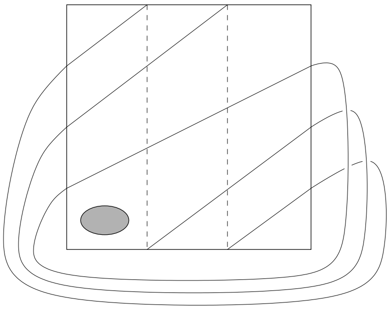


Figure 4. Planar diagram for $K_{2,3}$.

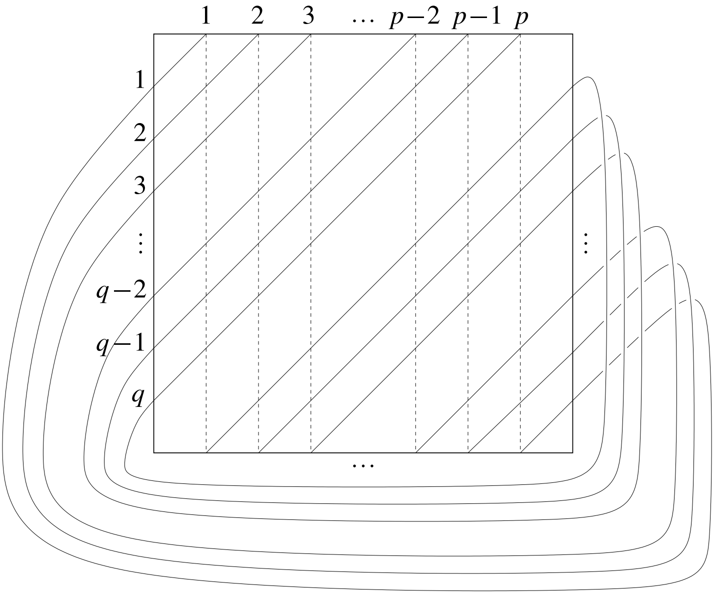


Figure 5. Planar diagram for $K_{p,q}$.

3. The wrapping function

The underlying key to many of our results is our "wrapping" function. Given a strand entering the left side of the rectangle in the planar diagram construction of $K_{p,q}$ (see Figure 6), the wrapping function determines where that particular strand re-enters the left side of the rectangle.

This allows us to count components, to characterize the types of components, and sometimes to tell that the components are unlinked. We have a similar wrapping

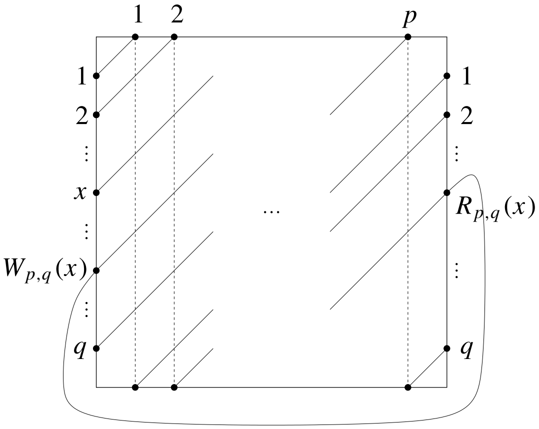


Figure 6. The wrapping function.

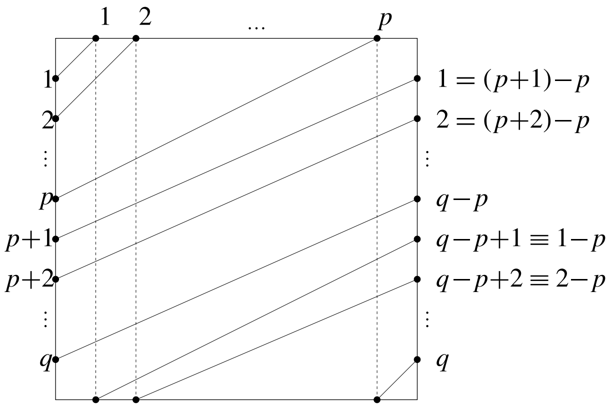


Figure 7. $R_{p,q}$ with $p \leq q$.

function and similar results for torus links, though we will concentrate only on our results for Klein links here.

Let $1 \leq x \leq q$, so that x is the position at which a particular strand passes through the left side of the planar diagram for $K_{p,q}$ as in Figure 6. Then the wrapping function is given by

$$W_{p,q}(x) = 1 - x + p \pmod{q}.$$

To see why this formula works, we will first back up a step and determine the position at which the strand entering the left side at x *exits the right side* of the planar diagram as shown in Figure 6; we denote this position by $R_{p,q}(x)$. In Figure 7, with $p \leq q$, we can see that

$$R_{p,q}(x) = x - p \pmod{q}.$$

In particular, notice that $R_{q,q}(x) = x - q = x \pmod{q}$, as one would expect. Next, if $p > q$, we divide p by q to get $p = nq + r$ for some $n, r \in \mathbb{N}$ with $1 \leq r \leq q$. By concatenating n copies of the planar diagram for $K_{q,q}$ and one copy of the diagram for $K_{r,q}$, we get that

$$\begin{aligned} R_{p,q}(x) &= R_{nq+r,q}(x) \\ &= R_{r,q} \circ R_{q,q} \circ R_{q,q} \circ \cdots \circ R_{q,q}(x) \\ &= R_{r,q}(x) \\ &= x - r \\ &= x - (p - nq) \\ &= x - p \pmod{q}. \end{aligned}$$

Finally, since strands exiting the right side of the planar diagram enter the left side in reverse order, we have that $W_{p,q}(x) = 1 - R_{p,q}(x) = 1 - x - p \pmod{q}$.

In our work, we will reference the following result about composing the wrapping function with itself.

Lemma 1. *For any $p, q \geq 0$, we have $W^2_{p,q}(x) = x$. Therefore, every component of $K_{p,q}$ wraps at most twice.*

Proof. Applying $W_{p,q}$ twice, we see that

$$\begin{aligned} W^2_{p,q}(x) = W_{p,q} \circ W_{p,q}(x) &= W_{p,q}(1 - x + p) \\ &= 1 - (1 - x + p) + p \\ &= x \pmod q. \end{aligned}$$ □

4. Results

First we compute the number of components in a $K_{p,q}$ link and determine the types of components, results that we use to prove that $K_{5,5}$ is not equivalent to any torus link.

Theorem 2 (number of components). *If $q = 0$, then $K_{p,q}$ has p components. If $q \neq 0$, then the number of components of $K_{p,q}$ is*

$$\begin{cases} (q+1)/2 & \text{if } q \text{ is odd,} \\ q/2 & \text{if } q \text{ and } p \text{ are both even,} \\ (q+2)/2 & \text{if } q \text{ is even and } p \text{ is odd.} \end{cases}$$

Proof. For $q = 0$, no components wrap around the rectangle. So there is a component for each point on the top. Thus, there are p components.

For $q > 0$, by Lemma 1, each component wraps at most twice. We find how many components wrap once. If there are t components that wrap once, then there are $(q - t)/2$ components that wrap exactly twice. So there will be $(q - t)/2 + t = (q + t)/2$ components in all.

To find the number of components of $K_{p,q}$ that wrap once, we solve the modular equation

$$\begin{aligned} W_{p,q}(x) = 1 - x + p = x \pmod q, \\ 2x - p - 1 = 0 \pmod q. \end{aligned} \tag{1}$$

In other words, q divides $2x - p - 1$.

Case 1: q odd. Since we are adding modular q, without loss of generality, we can assume $p < q$. Since $x \leq q$, we have

$$q < 2x - p < 2q - p \quad \Longrightarrow \quad q - 1 < 2x - p - 1 < 2q - p - 1.$$

Thus we have that $q-1 < 2x-p-1 < 2q$. So, if $2x-p-1 = 0 \pmod q$, then $2x-p-1 = q$. Then there is one component that wraps once. So, for q odd, $K_{p,q}$ has $(q+1)/2$ components.

Case 2: p and q even. Let $p = 2n$ and $q = 2r$. In this special case, (1) becomes

$$2(x-n)-1 = 2rk \quad \text{for some } k \in \mathbb{Z}. \tag{2}$$

Notice that the left-hand side of (2) is odd while the right-hand side is even. So $W_{p,q}(x) \neq x$ for any x, and thus no components wrap once. Thus, for $K_{p,q} = K_{2n,2r}$, there are $q/2 = r$ components.

Case 3: p odd, q even. Let $p = 2n+1$ for some integer n and $q = 2r$ for some integer r. Then, (1) becomes $2(x-n-1) = 0 \pmod q$. Now since $x \leq q$, we have

$$2(x-n-1) < 2(q-n-1) < 2q.$$

Thus the only possibilities for (1) to be true are $2(x-n-1) = 0$ or $2(x-n-1) = q$. So, there are exactly two components of $K_{p,q}$ that wrap once, namely when $x = n+1$ and $x = n+1+q/2$. Then, for $K_{p,q}$, there are $(q+2)/2$ components. □

The components of a link are knots. More generally, a link can be viewed as a collection of sublinks, possibly tangled together. For notational purposes, if a link L is made up of m copies of a sublink M and n copies of a sublink N, we will write $L \equiv m \cdot M \cup n \cdot N$. In the next theorem, we determine the types of knots that make up a Klein link.

Theorem 3 (types of components). *If $p < q$, then*

$$K_{p,q} \equiv K_{p,p} \cup K_{0,q-p},$$

where the sublink $K_{p,p}$ is disjoint (untangled) from $K_{0,q-p}$. Furthermore:

(1) *If $q = 2n+r$ with $n \in \mathbb{N}$ and $r = 0$ or $r = 1$, then*

$$K_{0,q} \equiv n \cdot K_{0,2} \cup r \cdot K_{0,1}.$$

(2) *If $p = 2n+r$ with $n \in \mathbb{N}$ and $r = 0$ or $r = 1$, then*

$$K_{p,p} \equiv n \cdot K_{2,2} \cup r \cdot K_{1,1}.$$

Proof. We begin by showing that if $p < q$, then $K_{p,q} \equiv K_{p,p} \cup K_{0,q-p}$. Let X_1 be the positions $1, 2, \ldots, p$ on the left side of the planar diagram for $K_{p,q}$, as shown in Figure 8. Similarly, let X_2 be the positions $p+1, \ldots, q$ on the left, Y_1 be the positions $1, \ldots, q-p$ on the right, and Y_2 be the positions $q-p+1, \ldots, q$ on the right.

Notice that $|X_1| = |Y_2| = p$ and $|X_2| = |Y_1| = q-p$.

By construction of the planar diagram for $K_{p,q}$ with $p < q$, strands from the p positions in X_1 pass through the p positions on the top of the diagram, then through

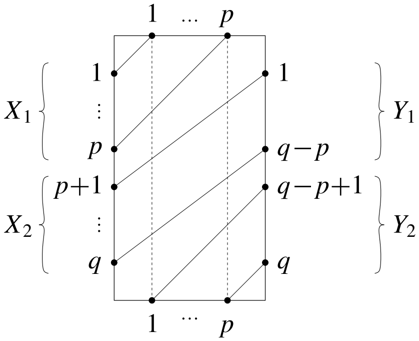


Figure 8. X_1, X_2, Y_1 and Y_2.

the p positions on the bottom of the diagram, and hence to the p positions in Y_2. Throughout, the order is preserved. In other words, $R_{p,q}|_{X_1} : X_1 \to Y_2$ is a bijection. Similarly, strands from the $q - p$ positions in X_2 cross the diagram directly to the $q - p$ positions in Y_1, preserving order, and $R_{p,q}|_{X_2} : X_2 \to Y_1$ is also a bijection.

Inside of the rectangle in the diagram, all strands from X_2 on the left cross over all strands from X_1 before passing through positions in Y_1. Outside of the rectangle, these same strands exit from positions in Y_1, cross over all strands exiting from Y_2, and re-enter through X_2 in reverse order. Thus $W_{p,q}|_{X_2} : X_2 \to X_2$, and the strands passing through positions in X_2 form a link L_2 that crosses over all other strands in $K_{p,q}$. Similarly, strands through positions in X_1 pass under strands from X_2 both inside and outside the rectangle in the diagram, $W_{p,q}|_{X_1} : X_1 \to X_1$, and these strands form another link L_1 completely disjoint from L_2. These two links are illustrated in Figure 9.

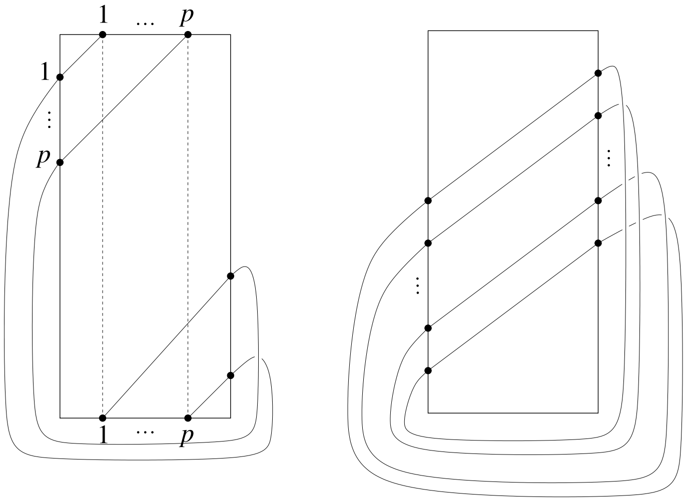


Figure 9. The links L_1 (left) and L_2 (right).

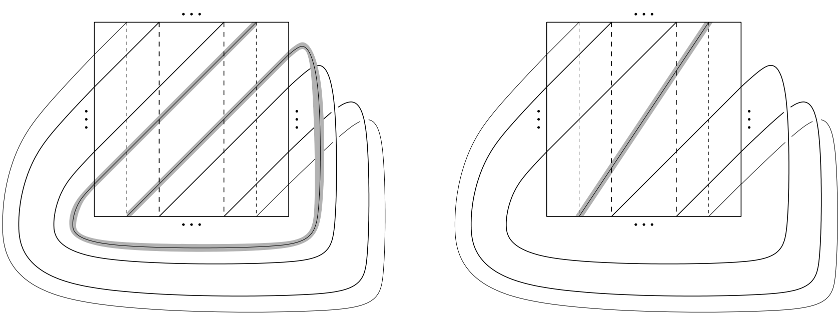


Figure 10. Planar diagrams of $K_{p,p}$ (left) and $K_{p,p-1}$ (right).

Viewed separately, these disjoint links are $L_1 = K_{p,p}$ and $L_2 = K_{0,q-p}$. Thus $K_{p,q} \equiv K_{p,p} \cup K_{0,q-p}$.

Since $K_{p,q}$ is composed of the two links $K_{p,p}$ and $K_{0,q-p}$, our next step is to characterize the components of Klein links of these types.

Consider $K_{0,q}$ for any value of $q \geq 1$. By Lemma 1, for all $0 < x \leq q$, we have $W^2_{0,q}(x) = x$, and hence each component of $K_{0,q}$ wraps horizontally around the rectangle in the planar diagram at most twice. It follows that each component is either a $K_{0,1}$ or a $K_{0,2}$. Now, to have a $K_{0,1}$ component, we must have some $0 < x \leq q$ such that $x = W_{0,q}(x) = 1 - x \pmod q$. This occurs exactly when q divides $x - (1-x) = 2x - 1$, that is, exactly when q is odd. In this case, $q = 2n+1$ for some positive integer n (and $r = 1$ in the statement of the theorem). Since $2x - 1 \leq 2q - 1 < 2q$, we see that $W_{0,q}(x) = x$ implies that $2x - 1 = q$. Thus, there is only one component of the form $K_{0,1}$ which passes through $x = (q+1)/2$. Since all other components wrap twice, there are n components of the form $K_{0,2}$, and $K_{0,q} \equiv n \cdot K_{0,2} \cup K_{0,1}$. On the other hand, suppose q is even with $q = 2n$ for some positive integer n (and $r = 0$). Then every component wraps twice, so $K_{0,q} \equiv n \cdot K_{0,2}$.

For $K_{p,p}$, since every component wraps at most twice, every component is of the form $K_{2,2}$ or $K_{1,1}$. Similar to the $K_{0,q}$ situation, there is at most one component of type $K_{1,1}$ and it exists if and only if p is odd. Therefore, if $p = 2n + r$ with $r = 0$ or 1, we have $K_{p,p} \equiv n \cdot K_{2,2} \cup r \cdot K_{1,1}$. □

Above we see that $K_{p,p}$ has components consisting entirely of the knots $K_{2,2}$ and $K_{1,1}$. It turns out that we can also view $K_{p,p}$ as the slightly simpler Klein link $K_{p,p-1}$.

Theorem 4 (Klein to Klein). *If $p \in \mathbb{N}$, then $K_{p,p} \equiv K_{p,p-1}$.*

Proof. By construction, the diagram of $K_{p,p}$ has a loop sitting on top of the rest of the link. This top loop, which is highlighted in Figure 10 (left), can be pulled tight (with one of the basic Reidemeister moves), turning the double diagonal strands into one diagonal strand. The resulting diagram, Figure 10 (right), is the link $K_{p,p-1}$. □

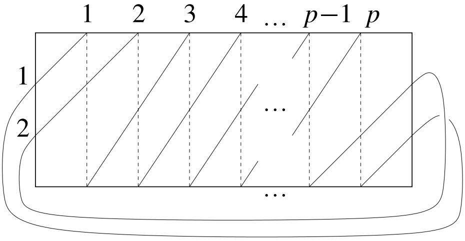


Figure 11. Planar diagram of $K_{p,2}$.

Now we are ready to compare Klein links and torus links. Recall that all Klein knots are torus knots. Similarly, certain classes of Klein links are torus links. In the next two theorems, we investigate Klein links of the forms $K_{p,2}$ and $K_{2p,2p}$.

Theorem 5 (Klein to torus, I). *If $p \in \mathbb{N}$, then $K_{p,2} \equiv T_{p-1,2}$.*

Proof. Consider the planar diagram of $K_{p,2}$ as in Figure 11.

Notice the crossing on the right of the planar diagram, in particular the strand that crosses underneath. If we follow this strand to the left, it wraps under the planar diagram. By unwrapping this strand and pulling it upward, the crossing is now gone. (This is a type-II Reidemeister move.) The resultant diagram is shown in Figure 12 (left).

We no longer have two nodes on the right side of the planar diagram. However, we may slide the strand so the strand exits the planar diagram from the right side as opposed to the top side. See Figure 12 (right).

So there are $p-1$ nodes along the top and two nodes along the side. Notice the strand that exits on the top node on the right enters through the top node on the left, and similarly the strand that exits on the bottom node on the right enters through the bottom node on the left. Thus, we obtain the planar diagram for $T_{p-1,2}$ and $K_{p,2} \equiv T_{p-1,2}$. □

Theorem 6 (Klein to torus, II). *If $p \in \mathbb{N}$, then $K_{2p,2p} \equiv T_{2p,p}$.*

Proof. A general $K_{2p,2p}$ has $2p$ strands entering or leaving each side of the rectangle in the planar diagram. Instead of manipulating each strand separately, we will collect

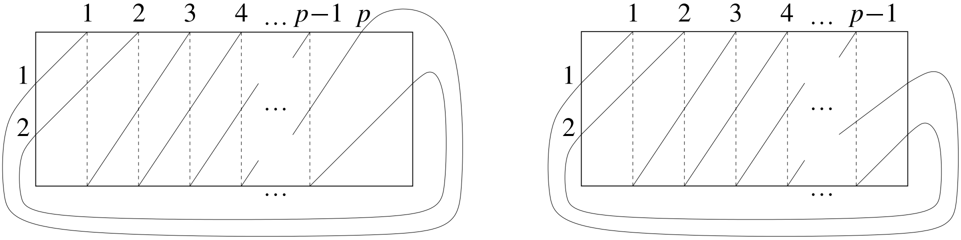


Figure 12. $K_{p,2}$ with crossing removed (left) and with two nodes on right restored (right).

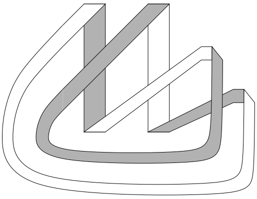


Figure 13. $K_{2p,2p}$ as a ribbon.

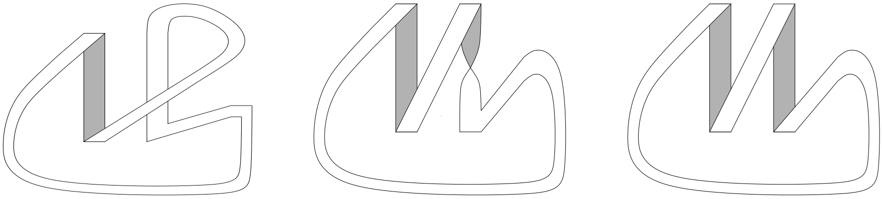
together the first p strands entering the left side of the rectangle as if they are on a ribbon, as illustrated in Figure 13.

Manipulating the ribbon moves the p strands together, resulting in an equivalent link. Our first steps are to flip up the inner loop of the ribbon, and unfold the lower right portion of the ribbon, resulting in Figure 14(a). In Figure 14(b), we turn the loop into a twist in the ribbon, and in Figure 14(c), we push the twist down to produce a fold. Returning to the p strands instead of the ribbon, we now have $T_{2p,p}$, as desired. □

The final class of Klein links that we consider are those of the form $K_{b,b}$ where $b \geq 5$ is odd. We will use linking numbers in our proof. To denote the linking number for an oriented link L of more than two components, we use the notation $\operatorname{lk}(L) = [l_1, l_2, \ldots, l_n]$, where $l_1, l_2, \ldots, l_n$ are the individual linking numbers of each two-component pair and are arranged in no particular order. This is not the total linking number found in [Bush et al. 2014; Murasugi 2008] which goes one step further by summing up the pair-wise linking numbers.

Theorem 7 (Klein not torus). *Let $b \geq 5$ be an odd integer. For every choice of $p, q \in \mathbb{N}$, we have $K_{b,b} \not\equiv T_{p,q}$. In other words, $K_{b,b}$ is not a torus link.*

Proof. By Theorem 2, $K_{b,b}$ has $c = (b+1)/2$ components, and by Theorem 3, one of the components is a copy of $K_{1,1}$ and all of the other components are copies of $K_{2,2}$. Note that both $K_{2,2}$ and $K_{1,1}$ are unknots.



(a) $K_{2p,2p}$ after flipping and unfolding.

(b) $K_{2p,2p}$ with twist.

(c) $K_{2p,2p}$ as $T_{2p,p}$.

Figure 14. Deforming $K_{2p,2p}$ into $T_{2p,p}$.

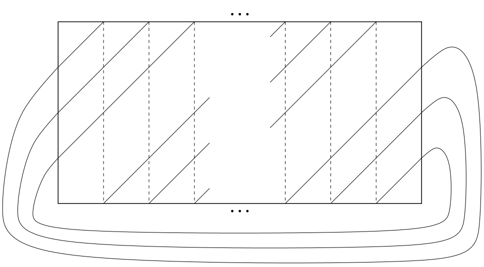


Figure 15. $T_{3m,3}$.

If $K_{b,b} \equiv T_{p,q}$, then $T_{p,q}$ must also have $c = (b+1)/2$ components. It is well known (see [Adams 2004; Murasugi 2008]) that $T_{p,q}$ has c components exactly when $\gcd(p, q) = c$. So we let $p = mc$ and $q = nc$, where $m, n \in \mathbb{N}$ with $\gcd(m, n) = 1$. Then, as determined in [Murasugi 2008], $T_{p,q} \equiv T_{mc,nc} \equiv c \cdot T_{m,n}$. Furthermore, we may assume that $m > n$ without loss of generality since $T_{p,q} \equiv T_{q,p}$ for all p, q (see [Adams 2004; Murasugi 2008]). As noted above, each individual component of $K_{b,b}$ is an unknot. From our (very convenient) knowledge of torus knots and [Murasugi 2008], a torus knot $T_{m,n}$ is equivalent to the unknot only when $n = 1$. So if $K_{b,b}$ is equivalent to some torus link, it must be the case that $K_{b,b} \equiv T_{mc,c}$.

In order to determine if $K_{b,b} \equiv T_{mc,c}$ for some m, we examine the linking numbers. From our standard planar diagram of $K_{b,b}$, regardless of orientation, we have that $\text{lk}(K_{b,b}) = [2, \ldots, 2, 1, \ldots, 1]$. Any pair of components consisting of the copy of $K_{1,1}$ and one of the copies of $K_{2,2}$ has four crossings within the rectangle in the planar diagram, all of the same type, and two crossings outside of the rectangle, all of the opposite type, resulting in a crossing number of $(4-2)/2 = 1$. Any pair consisting of two copies of $K_{2,2}$ has eight crossings within the rectangle and four opposite crossings outside of the rectangle, and hence a linking number of $(8-4)/2 = 2$.

Now consider the planar diagram for the general $T_{mc,c}$. As an example, $T_{3m,3}$ is shown in Figure 15. Orient each strand entering the left side of the rectangle in an upward direction. Consider the block on the left side of the diagram corresponding to the first c points along the top of the rectangle. In this block, each component crosses each of the other components exactly twice, and each crossing is left-handed. The same thing happens in each of the m blocks corresponding to groups of c points along the top of the rectangle. Thus the linking number between any two components in the $T_{mc,c}$ is $|-2m|/2 = m$ and consequently $\text{lk}(T_{mc,c}) = [m, m, \ldots, m] \neq [2, \ldots, 2, 1, \ldots, 1]$. So there is no m for which $K_{b,b} \equiv T_{mc,c}$. Hence, $K_{b,b} \not\equiv T_{p,q}$ for any choice of p, q. □

Since all Klein knots are also torus knots, we expected all Klein links to be torus links. So this last result was a pleasant surprise.

References

[Adams 2004] C. C. Adams, *The knot book: An elementary introduction to the mathematical theory of knots*, Amer. Math. Soc., Providence, RI, 2004. MR 2005b:57009 Zbl 1065.57003

[Bowen et al. 2013] J. Bowen, D. Freund, J. Ramsay, and S. Smith-Polderman, "Klein link multiplicity and recursion", preprint, 2013, available at http://discover.wooster.edu/jbowen/files/2013/10/Klein-Link-Multiplicity-and-Recursion.pdf.

[Bush et al. 2014] M. A. Bush, K. R. French, and J. R. H. Smith, "Total linking numbers of torus links and klein links", *Rose-Hulman Undergrad. Math J.* **15**:1 (2014), 73–92. MR 3216221

[Freund and Smith-Polderman 2013] D. Freund and S. Smith-Polderman, "Klein links and braids", *Rose-Hulman Undergrad. Math J.* **14**:1 (2013), 71–84. MR 3071244

[Murasugi 2008] K. Murasugi, *Knot theory & its applications*, Birkhäuser, Boston, 2008. MR 2347576 Zbl 1138.57001

[Shepherd et al. 2012] D. Shepherd, J. Smith, S. Smith-Polderman, J. Bowen, and J. Ramsay, "The classification of a subset of klein links", *Proceedings of the Midstates Conference for Undergraduate Research in Computer Science and Mathematics at Ohio Wesleyan University* (2012), 38–47.

Received: 2015-01-03 Revised: 2015-02-23 Accepted: 2015-02-26

ealvarado9611@gmail.com
Department of Mathematics, Washington State University, Pullman, WA 99164, United States

sberes@zagmail.gonzaga.edu
Department of Mathematics, Gonzaga University, Spokane, WA 99258, United States

coufal@gonzaga.edu
Department of Mathematics, Gonzaga University, Spokane, WA 99258, United States

khlavacek@zagmail.gonzaga.edu
Department of Mathematics, Gonzaga University, Spokane, WA 99258, United States

pereira@gonzaga.edu
Department of Mathematics, Gonzaga University, Spokane, WA 99258, United States

breeves@wisc.edu
Department of Economics, University of Wisconsin, Madison, WI 53706, United States

Guidelines for Authors

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 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 LaTeX but submissions in other varieties of TeX, 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 BibTeX 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, Corel Draw, 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 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

2016 vol. 9 no. 2