# involve

## a journal of mathematics

**msp**

# involve

msp.org/involve

## INVOLVE YOUR STUDENTS IN RESEARCH

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

### MANAGING EDITOR

Kenneth S. Berenhaut    Wake Forest University, USA

### BOARD OF EDITORS

### PRODUCTION

Silvio Levy, Scientific Editor

Cover: Alex Scorpan

### PUBLISHED BY

**mathematical sciences publishers**

nonprofit scientific publishing

http://msp.org/

© 2018 Mathematical Sciences Publishers

# On halving-edges graphs

Tanya Khovanova and Dai Yang

(Communicated by Kenneth S. Berenhaut)

In this paper we study halving-edges graphs corresponding to a set of halving lines. Particularly, we study the vertex degrees, path, cycles and cliques of such graphs. In doing so, we study a vertex-partition of said graph called chains which are equipped with interesting properties.

## 1. Introduction

Halving lines have been an interesting object of study for a long time. Let $n$ points be in general position in $\mathbb{R}^2$, where $n$ is even. A *halving line* is a line through two of the points that splits the remaining $n-2$ points into two sets of equal size. The minimum number of halving lines is $\frac{1}{2}n$. The maximum number of halving lines is unknown. The first bounds were found by Lovász [1971] and by Erdős et al. [1973]. The current asymptotic upper bound of $O(n^{4/3})$ was proven by Dey [1998].

We approach the subject of halving lines by studying the properties of the underlying graph. From our set of $n$ points, we define a *halving-edges graph* of $n$ vertices, where each point is a vertex and each pair of vertices is connected by an edge if and only if there is a halving line through the corresponding two points; see [Matoušek 2002].

In Section 2 we discuss some basic properties of halving-edges graphs including degrees and the number of connected components. We also prove that any graph can be an induced subgraph of a halving-edges graph. In Section 3 we show that a halving-edges graph with $n$ vertices can contain an $(n-1)$-path, and an $(n-3)$-cycle at most and provide a construction to show that the bound is exact. We give an example of a halving-edges graph containing a clique of size of at least $\sqrt{\frac{1}{2}n}$. We continue by studying chains, introduced by Dey [1998], in Section 4. The chain methods allow us to prove more properties of halving-edges graphs. In particular, we show that the largest clique cannot exceed a size of $\sqrt{2n}+1$.

## 2. Basic properties of the halving-edges graph

The following properties of halving edges graphs are well known.

**Lemma 2.1.** *A halving-edges graph does not have isolated vertices. It has at least three leaves.*

**Theorem 2.2.** *Each vertex of a halving-edges graph has an odd degree.*

We will use another related result in the future.

**Lemma 2.3.** *Given two halving lines VP and VQ sharing a vertex V, there exists another halving line VR such that R lies in the opposite angle of $\angle PVQ$. Equivalently, the vectors $\vec{VP}$, $\vec{VQ}$, $\vec{VR}$ do not all lie on a single half-plane.*

As each vertex has at least one halving line passing through it, the minimum number of halving lines is $\frac{1}{2}n$. This number is achieved when points form a convex $n$-gon. Any number of halving lines between the lower bound and the upper bound is achievable as the following theorem states [Khovanova and Yang 2012].

**Theorem 2.4.** *For a fixed n, if there exist two configurations with $k_1$ and $k_2$ halving lines respectively, then for all k such that $k_1 \leq k \leq k_2$, there exists a configuration with k halving lines.*

***Segmenterizing.*** We will use this construction a lot through this paper. Suppose we have a set of points. Any affine transformation does not change its halving-edges graph. Sometimes it is useful to picture that our points are squeezed into a long narrow rectangle. This way our points are almost on a segment. We call this procedure *segmenterizing*. Figure 1 shows three pictures. The first picture has six points, that we would squeeze towards the line $y = 0$. The second picture shows the configuration squeezed by a factor of 10, and if we make the factor arbitrarily large the points all lie very close to a segment, as shown in the last picture.

This procedure makes all the points very close to a single line segment and all the halving lines very close to this line. If we add a point not too close to this line, then it lies on the same side of all the halving lines. Moreover, it lies on the same side of all the lines connecting any two original points. Note that by the nature of affine transformations, we do not necessarily have to squeeze along the direction that is perpendicular to the segment.
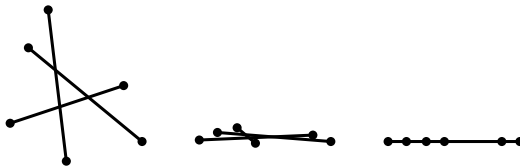


**Figure 1.** Segmenterizing.
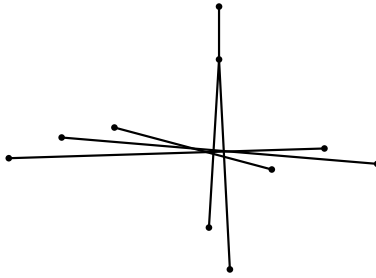
**Figure 2.** The cross construction.

***Degrees and connected components.*** In the proof of the following lemma we need a construction we call a *cross*. Given two sets of points with $n_1$ and $n_2$ points respectively whose halving-edges graphs are $G_1$ and $G_2$, the cross is the construction of $n_1 + n_2$ points on the plane whose halving-edges graph has two isolated components $G_1$ and $G_2$. We form the cross as follows. Segmenterize graphs $G_1$ and $G_2$ and intersect the resulting segments at middle lines, so that half of the points of each segment lie on one side of all halving lines that pass through the points of the other segment (see Figure 2).

**Lemma 2.5.** *Any odd degree between* 1 *and* $n - 1$ *can appear in a halving-edges graph of n vertices. Any number of connected components between* 1 *and* $\frac{1}{2}n$ *inclusive can appear in a halving-edges graph of n vertices.*

*Proof.* Consider a configuration with $2k$ vertices, where all but one of them are on a convex hull. The resulting halving graph is a star. We build a cross of this star graph and of a convex polygon with $n - 2k$ vertices. The cross has $\frac{1}{2}n - k + 1$ connected components. It has $n - 1$ leaves and one vertex of degree $2k - 1$.   □

***Degree sequence.*** The *degree sequence* of a graph is the nonincreasing sequence of its vertex degrees. The Erdős–Gallai [1960] theorem describes which sequences could be degree sequences of graphs.

**Theorem 2.6** (Erdős–Gallai theorem). *A nondecreasing sequence of n numbers $d_i$ is the degree sequence of a simple graph if and only if the total sum of degrees is even and*

$$\sum_{i=1}^{k} d_i \leq k(k-1) + \sum_{i=k+1}^{n} \min(d_i, k) \quad \text{for } k \in \{1, \ldots, n\}.$$

The following lemma is about vertices of large degrees in a halving graph.

**Lemma 2.7.** *At most one vertex can have degree* $n - 1$, *at most three vertices can have degree* $n - 3$. *If the halving-edges graph has a vertex of degree* $n - 1$, *then it is a star graph.*

The proof is straightforward [Khovanova and Yang 2012].

**Lemma 2.8.** *Any degree sequence consisting of only ones and threes, with at least 3 ones, is achievable by the halving-edges graph of some configuration.*

*Proof.* The degree sequence with 3 ones and everything else threes corresponds to the configuration in the path construction in Lemma 3.1. This configuration crossed with a matching graph can produce any odd number of ones with the rest being threes.

To achieve an even number of ones, we can use the following modified version of the path construction: replace the two vertices lying on the $y$-axis by two vertices that form a horizontal segment which makes the bottom side of the convex hull. Under this configuration, the four vertices on the convex hull have degree 1, and the remaining vertices have degree 3. □

## 3. Paths cycles, and cliques

*Paths.* Here we consider the size of non-self-intersecting paths in halving graphs. A path cannot have more than two leaves, so an easy upper bound for the largest path is $n - 1$ vertices. It turns out that this bound is exact.

**Lemma 3.1.** *For every n, there exists a halving-edges graph of size n having a path through n − 1 vertices.*

*Proof.* Figure 3 shows the path construction for a configuration with eight points. To avoid clutter, only relevant halving lines are shown by thin lines and thick lines show the path in the halving-edges graph. We generalize this construction to any $n$.

Consider $\frac{1}{2}(n - 2)$ points that lie on a concave function. We segmenterize these points onto a segment lying on the $x$-axis. Now we place one such segment onto a line $y = x$, to the right of the origin, and another segment on the line $y = -x$ to the left of the origin. We keep the segments oriented in such a way that a line that passes through any two neighboring points of a segment has the remaining $\frac{1}{2}(n - 2) - 2$ points of the segment below it. Now add two more points: $(0, -1)$ and $(0, -2)$. Thus, every line that passes through two neighboring points of a segment
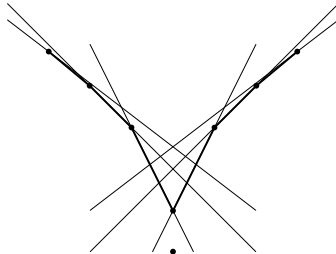


**Figure 3.** Path.
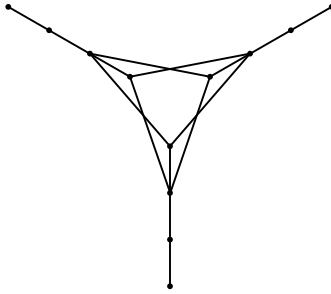
**Figure 4.** The Y-shape construction.

becomes a halving line. In addition, the point $(0, -1)$ forms halving lines with the rightmost point of the left segment and the leftmost point of the right segment.

The path goes through every point except $(0, -2)$, forming a V-shape. $\qquad\square$

*Cycles.* Here we consider the size of cycles in halving-edges graphs. Vertices on the convex hull cannot be part of a cycle, so an easy upper bound for the length of the largest cycle is $n - 3$. It turns out that this bound is asymptotically exact. But first we need to introduce the Y-shape construction.

Suppose we have three configurations $G_1$, $G_2$, and $G_3$ with $n$ points each and $k_1$, $k_2$, and $k_3$ halving lines correspondingly. The Y-shape construction allows us to build a new configuration with $3n$ points which has each of the three initial configurations as a subgraph and has a total of $k_1 + k_2 + k_3 + \frac{3}{2}n$ halving lines.

The construction works as follows. We segmenterize each set of points $G_i$. Then we draw three rays emanating from the origin, forming an angle of 120 degrees between each other, and place each segmenterized set of points along one of the rays; see Figure 4. This makes a Y-shape of $3n$ points, with $n$ points on each branch.

On an individual branch, all halving lines prior to segmenterization remain halving lines. In addition, we can find halving lines that go through two points on different branches of the Y-shape. There are a total of $\frac{3}{2}n$ such lines, so we have produced a configuration with $3n$ points and $k_1 + k_2 + k_3 + \frac{3}{2}n$ halving lines.

**Lemma 3.2.** *Suppose a configuration of points with two neighboring points on the convex hull, denoted by $A$ and $B$, is given. We can segmenterize in such a way and choose a direction on the segment so that $A$ becomes the first point of the segment and $B$ the $k$-th point, for $1 < k \leq n$, where $n$ is the total number of points.*

See the proof in [Khovanova and Yang 2012]. The *halving difference* of a line is the difference of the number of points on each side of the line. Sometimes we will produce a construction that does not disturb the halving difference of certain lines. That is to say, we add the same number of points on both sides of the line and the difference is preserved.
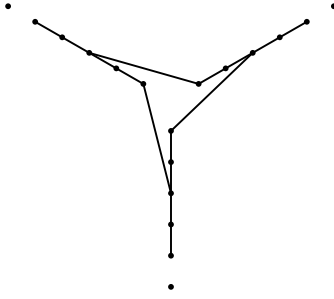
**Figure 5.** Cycle.

**Theorem 3.3.** *If $n$ is a multiple of* 6, *the maximum length of a cycle is exactly $n-3$.*

*Proof.* We can write $n = 3b$, where $b$ is even. Using Lemma 3.1, we can create a configuration of $b$ points with a path of length $b-1$. Note that the endpoints of the path (of the V-shape) are neighboring points on a convex hull. This allows us to use Lemma 3.2 to segmenterize this configuration so that the endpoints of the path occupy the positions 1 and $\frac{1}{2}b$.

Now we use three copies of this segment in the Y-shape construction. We orient segments in such a way that the point 1 is oriented closer to the center of the construction. The edges of the $(b-1)$-path inside each branch all remain edges, and we also have edges between the first points of the branches and $\frac{1}{2}b$ points connecting all these paths together. This creates a cycle of length $n-3$ as desired. In Figure 5 we demonstrate this cycle for 18 vertices. Note that each branch has six vertices and the outermost vertex of each branch does not belong to the cycle. □

***Induced subgraphs.*** We just showed that a halving-edges graph can contain a large path/cycle as a subgraph. If we restrict the graph to the vertices of the path/cycle that we constructed above, we can see that the graph has extra edges in addition to the path/cycle. To differentiate any subgraph from a subgraph that retains all the edges, the notion of induced subgraph is used.

A subgraph $H$ of graph $G$ is said to be an *induced* subgraph if any pair of vertices in $H$ is connected by an edge if and only if it is connected by an edge in $G$.

**Theorem 3.4.** *Any graph with $2k$ vertices and $e$ edges can be an induced subgraph of a halving-edges graph with at most $2k + 2ek - 4e + 2\binom{2k}{2}$ vertices.*

*Proof.* Notice that if the number of vertices is even, then every line has an even halving difference. We process the configuration line by line. Take a line. Suppose we want to make it a halving line. For this we need to add an even number of points on one side of the line without disturbing the halving difference of other lines. If it is a halving line and we want to make it a nonhalving line we can add 2 points on
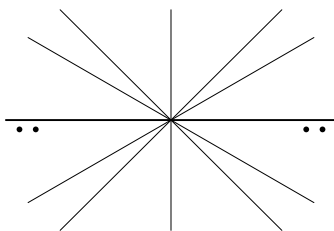
**Figure 6.** Zooming out and adding points.

one side. Let us draw all possible lines connecting the points and zoom out. From a big distance the point configuration will look like a bunch of lines intersecting at one point; see Figure 6. In Figure 6 the thick line is the line we are processing. Suppose the line needs an addition of four points below it. We add half of the points (two in our example) below the line far away on each side.

Each line that should be an edge in the new halving-edges graph requires an addition of at most $2k-2$ vertices. All of the future edges require at most $2ek-2e$ extra points. Other lines require at most 2 points each for a total of $2\left(\binom{2k}{2}-e\right)$. □

***Cliques.*** Halving-edges graphs with 4 vertices have cliques of size 2. We can have a clique of size 3 in a graph with six vertices. By Theorem 3.4 we can have a clique of size $n$ as an induced subgraph in a halving-edges graph of size $O(n^3)$. In this subsection we would like to improve the bound by using a construction similar to the construction of Theorem 3.4, where we process several lines at a time. Clustering lines together allows us to reduce the total number of extra points that we need.

**Theorem 3.5.** *The largest possible clique in a halving-edges graph of n vertices is at least $\Omega(\sqrt{n})$.*

*Proof.* Let $k$ be even. To produce a clique of size $k$, take a regular $k$-gon, and distort it a little bit using a projective transformation that makes one end of the $k$-gon slightly wider than the other end. This perturbs all the diagonals (and sides) of the $k$-gon that were once parallel, making them intersect somewhere far away from the polygon, but still remain nearly parallel. You can imagine the $k$-gon as drawn on the floor in a painting that respects the perspective properly. This way the lines that are parallel in the $k$-gon intersect on a point on the horizon line in the painting. We assume that the $k$-gon is in a general position, that is, no two vertices are connected by a line parallel to the horizon. Note that there are now $k$ sets of nearly parallel diagonals and sides, each set having either $\frac{1}{2}k$ or $\frac{1}{2}k-1$ lines.

We will now add $O(k^2)$ points to turn this $k$-gon into a $k$-clique. Consider a set of $\frac{1}{2}k$ or $\frac{1}{2}k-1$ nearly parallel lines. We will process each cluster of lines separately. In Figure 7 we depict one cluster of near parallel sides and diagonals. We rotated the picture so that it fits better in the page, and now the imaginary horizon line is a
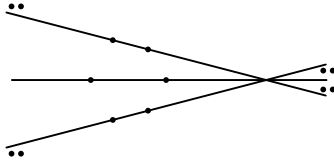
**Figure 7.** The cluster of nearly parallel lines.

vertical line through the intersection points on the right. On the half-plane beyond
the horizon line add two points between every pair of consecutive lines. This way
each line in the cluster becomes a halving line. In addition, we want every cluster
to be independent. That means we want to add more points so that the halving
difference of every line that is not in the cluster does not change.

We just added to the right of all other lines that are not in the cluster either $k-2$
or $k-4$ points. We need to add the same number of points to the left of all other
lines as not to disturb the halving difference we just created in this cluster. The
extra points you can see on the picture are put into two equal groups on the left
above and below the current cluster.

This process requires a total of $2k-4$ or $2k-8$ new points, but turns all of our
nearly parallel lines into halving lines without disturbing the other diagonals and
sides. We do this a total of $k$ times for each set of nearly parallel lines, and we have
constructed a halving-edges graph with a $k$-clique by adding $2k^2-6k$ points.

Given $n$, we have shown how to construct a halving-edges graph with a clique
of size at least $\sqrt{\frac{1}{2}n}$ with no more than $n$ vertices. We can pad this graph to any
number of vertices by crossing it with 2-paths.                                   □

We will discuss the upper bound on the size of the clique later.

Any graph can be a subgraph of a clique, so an arbitrary graph with $k$ vertices
can always be found as a subgraph, not necessarily induced, of a halving-edges
graph with no more than $O(k^2)$ vertices.

## 4. Chains

We define the following algorithm to group the halving lines into sets that are called
*chains*, introduced by Dey [1998].

Choose an orientation to define as "up". The $\frac{1}{2}n$ leftmost vertices are called the
left half, and the rightmost vertices are called the right half. We assume that no two
points are vertically aligned, so that leftmost and rightmost are well defined. Start
with a vertex on the left half of the graph, and take a vertical line passing through
this vertex. Rotate this line clockwise until it either aligns itself with an edge, or
becomes vertical again. If it aligns itself with an edge in the halving-edges graph,
define this edge to be part of the chain, and continue rotating the line about the

rightmost vertex in the current chain. If the line becomes vertical, we terminate the process. The set of edges in our set is defined as the chain. Repeat on a different point on the left half of the halving-edges graph until every edge is part of a chain.

Note that the chains we get are determined by which direction we choose as "up". The following properties of chains follow immediately. Later properties on the list follow from the previous ones:

- A vertex on the left half of the halving-edges graph is a left endpoint of a chain.

- The process is reversible. We could start each chain from the right half and rotate the line counterclockwise instead, and obtain the same chains.

- A vertex on the right half of the halving-edges graph is a right endpoint of a chain.

- Every vertex is the endpoint of exactly one chain.

- The number of chains is exactly $\frac{1}{2}n$.

- The degrees of the vertices are odd. Indeed, each vertex has one chain ending at it and several passing through it.

- Every halving line is part of exactly one chain.

- The length of each chain is bounded by $\frac{1}{2}n$.

The following property bounds the number of vertices with a large degree.

**Lemma 4.1.** *For every integer $k$, a halving graph has at most $2k$ vertices with degree $n - 2k + 1$.*

*Proof.* The $i$-th vertex from the left in the left half plane can have at most $i - 1$ chains passing through it and is a start of exactly one chain. So its degree cannot be more than $2i - 1$. Hence, only $k$ rightmost vertices in the left plane and $k$ leftmost vertices in the right plane can have degree $n - 2k + 1$. $\square$

***The sums of degrees of two vertices.*** Now we use our knowledge about chains to refine our knowledge about degrees of the vertices of the halving-edges graph.

**Theorem 4.2.** *The degrees of two distinct vertices sum to at most $n$, if they are connected by an edge, and at most $n - 2$ otherwise.*

*Proof.* Denote the vertices in question as $P$ and $Q$. Rotate the geometric graph until segment $PQ$ is nearly vertical, so that there are no vertices between the horizontal projections of $P$ and $Q$. If $P$ and $Q$ do not belong to the same chain, then each of the $\frac{1}{2}n$ chains contributes at most 2 to the sum of degrees of $P$ and $Q$. We have to subtract 2 from this sum since $P$ and $Q$ are both endpoints of some chains. Thus the total sum does not exceed $n - 2$.

If $PQ$ is an edge, then it can add two more to the sum of degrees making it at most $n$.                                                                                                                                                   □

It immediately follows that the largest clique in the halving-edges graph cannot be bigger than $\frac{1}{2}n$. We can use chains to prove an upper bound on the size of the largest clique that is much closer to the lower bound. But before doing so we would like to introduce some definitions.

***The straddling span and the largest clique.*** Given a line that does not pass through any vertex of a given graph, we call edges that intersect it *straddling* edges. The maximum number of straddling edges that can be produced by a line is called the *straddling span* of the halving-edges geometric graph. Naturally, this notion applies to subgraphs as well. Let us consider some examples (see [Khovanova and Yang 2012]).

- The straddling span of a $k$-clique is at least $\left\lfloor \frac{1}{4}k^2 \right\rfloor$.
- The straddling span of an $(a, b)$-complete bipartite graph is at least $\frac{1}{2}ab$.

**Theorem 4.3.** *If a halving-edges geometric graph has straddling span $w$, then it has at least $\frac{1}{2}w$ vertices.*

*Proof.* Choose the up direction along the line that produces the straddling span. We claim that no two straddling edges belong to the same chain. Indeed, if two edges are straddling, then their projections onto the $x$-axis must overlap at the point that is the projection of the line that produces the straddling span. But it is clear that the projections along the $x$-axis of the edges of any given chain must be mutually nonoverlapping. Therefore, our graph contains at least one chain for every straddling edge. Since there are at least $w$ straddling edges, the number of chains must be at least the same and the number of vertices must be at least $\frac{1}{2}w$.          □

**Corollary 4.4.** *If a halving-edges graph contains a $k$-clique, then it has at least $\left\lfloor \frac{1}{2}k^2 \right\rfloor$ vertices. Consequently, the largest clique in the halving-edges graph with $n$ vertices cannot exceed $\sqrt{2n} + 1$ vertices.*

**Corollary 4.5.** *If a halving-edges graph contains an $(a, b)$-complete bipartite subgraph, then it has at least $ab$ vertices.*

Note that now both the lower bound and the upper bound for the largest clique are on the order of $\sqrt{n}$.

## 5. Acknowledgements

# References

[Dey 1998] T. K. Dey, "Improved bounds for planar $k$-sets and related problems", *Discrete Comput. Geom.* **19**:3 (1998), 373–382. MR Zbl

[Erdős and Gallai 1960] P. Erdős and T. Gallai, "Gráfok előírt fokszámú pontokkal", *Mat. Lapok* **11** (1960), 264–274.

[Erdős et al. 1973] P. Erdős, L. Lovász, A. Simmons, and E. G. Straus, "Dissection graphs of planar point sets", pp. 139–149 in *A survey of combinatorial theory* (Fort Collins, CO, 1971), edited by G.-C. Rota and S. S. Shrikhande, North-Holland, Amsterdam, 1973. MR Zbl

[Khovanova and Yang 2012] T. Khovanova and D. Yang, "Halving lines and their underlying graphs", preprint, 2012. arXiv

[Lovász 1971] L. Lovász, "On the number of halving lines", *Ann. Univ. Sci. Budapest. Eötvös Sect. Math.* **14** (1971), 107–108. MR

[Matoušek 2002] J. Matoušek, *Lectures on discrete geometry*, Graduate Texts in Mathematics **212**, Springer, 2002. MR Zbl

tanyakh@yahoo.com    *Department of Mathematics, MIT, 77 Massachusetts Ave, Cambridge, MA 20139, United States*

zephyredx@gmail.com    *Department of Mathematics, MIT, 575 S Rengstorff, Mountain View, CA 94040, United States*

# Knot mosaic tabulation

Hwa Jeong Lee, Lewis D. Ludwig, Joseph Paat and Amanda Peiffer

(Communicated by Kenneth S. Berenhaut)

In 2008, Lomonaco and Kauffman introduced a knot mosaic system to define a quantum knot system. A quantum knot is used to describe a physical quantum system such as the topology or status of vortexing that occurs in liquid helium II for example. Kuriya and Shehab proved that knot mosaic type is a complete invariant of tame knots. In this article, we consider the mosaic number of a knot, which is a natural and fundamental knot invariant defined in the knot mosaic system. We determine the mosaic number for all eight-crossing or fewer prime knots. This work is written at an introductory level to encourage other undergraduates to understand and explore this topic. No prior knowledge of knot theory is assumed or required.

## 1. Introduction

In this work we will determine the mosaic number of all 36 prime knots of eight crossings or fewer. Before we do this, we will give a short introduction to knot mosaics. Take a length of rope, tie a knot in it, glue the ends of the rope together and you have a *mathematical knot* — a closed loop in 3-space. A rope with its ends glued together without any knots is referred to as the *trivial knot*, or just an unknotted circle in 3-space. There are other ways to create mathematical knots aside from rope. For example, *stick knots* are created by gluing sticks end to end until a knot is formed (see [Adams 2004]). Lomonaco and Kauffman [2008] developed an additional structure for considering knots which they called *knot mosaics*. Kuriya and Shehab [2014] showed that this representation of knots was equivalent to tame knot theory, or knots with rope, implying that tame knots can be represented equivalently with knot mosaics. This means any knot that can be made with rope can be represented equivalently with a knot mosaic.
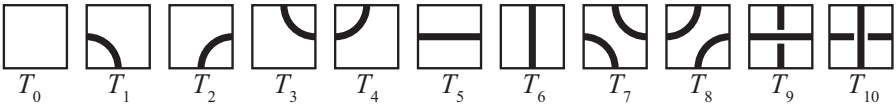
**Figure 1.** Tiles $T_0$–$T_{10}$.

A *knot mosaic* is the representation of a knot on an $n \times n$ grid composed of 11 tiles as depicted in Figure 1. A tile is said to be *suitably connected* if each of its connection points touches a connection point of a contiguous tile. Several examples of knot mosaics are depicted in Figure 2. It should be noted that in Figure 2, the first mosaic is a knot, the trefoil knot, the second mosaic is a link, the Hopf Link, and the third is the composition of two trefoil knots (remove a small arc from two trefoils then connect the four endpoints by two new arcs depicted in red, denoted by $3_1 \# 3_1$). A knot is made of one component (i.e., one piece of rope), and a link is made of one or more components (i.e., one or more pieces of rope). For this work, we will focus on knot mosaics of *prime knots*. A prime knot is a knot that cannot be depicted as the composition of two nontrivial knots. The trefoil knot is a prime knot.

When studying knots, a useful and interesting topic used to help distinguish two knots is *knot invariants*. A knot invariant is a quantity defined for each knot that is not changed by *ambient isotopy*, or continuous distortion, without cutting or gluing the knot. One such knot invariant is the *crossing number* of a knot. The crossing number is the fewest number of crossings in any diagram of the knot. For example, the crossing number of the trefoil is three, which can be seen in Figure 2. A *reduced* diagram of a knot is a projection of the knot in which none of the crossings can be reduced or removed. The fourth knot mosaic depicted in Figure 2 is an example of a nonreduced trefoil knot diagram. In this example, the crossing number of three is not realized because there are two extra crossings that can be easily removed.

An interesting knot invariant for knot mosaics is the *mosaic number*. The mosaic number of a knot $K$ is the smallest integer $n$ for which $K$ can be represented on an $n \times n$ mosaic board. We will denote the mosaic number of a knot $K$ as $m(K)$. For the trefoil, it is an easy exercise to show that the mosaic number for the trefoil



| trefoil | Hopf link | composition | nonreduced trefoil |
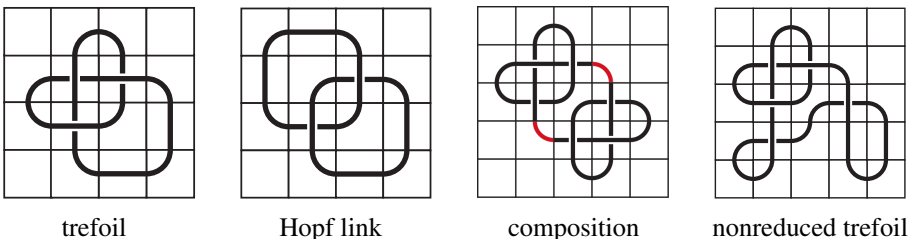
**Figure 2.** Examples of link mosaics.

**Figure 3.** Reidemeister Type I moves.

is four, or $m(3_1) = 4$. To see this, try making the trefoil on a $3 \times 3$ board and you will arrive at a contradiction.

Next, we introduce a technique that can be used to "clean up" a knot mosaic by removing unneeded crossing tiles. Kurt Reidemeister [1927] demonstrated that two knot diagrams belonging to the same knot, up to ambient isotopy, can be related by a sequence of three moves, now know as the Reidemeister moves. For our purposes, we will consider two of these moves on knot mosaics, the mosaic Reidemeister type I and type II moves as described by Lomonaco and Kauffman [2008]. For more about Reidemeister moves, the interested reader should see [Adams 2004].

The mosaic Reidemeister type I moves are shown in Figure 3, and the mosaic Reidemeister type II moves are shown in Figure 4.

Next we make several observations that will prove useful.

**Observation 1.1** [Hong et al. 2014, two-fold rule]. Once the inner tiles of a mosaic board are suitably placed, there are only two ways to complete the board so that it is suitably connected, resulting in a knot or a link.

For any given $n \times n$ mosaic board, we will refer to the collection of inner tiles as the *inner board*. For example, in Figure 8 the tiles $I_1$–$I_9$ would make the inner board for this mosaic board.

**Observation 1.2.** Assume $n$ is even. For a board of size $n$ with the inner board consisting of all crossing tiles, any resulting suitably connected mosaic will either be an $n - 2$ component link mosaic or $n - 3$ component nonreduced link mosaic; see Figure 6 for an example.

**Observation 1.3.** Assume $n$ is odd. For a board of size $n$ with the inner board consisting of all crossing tiles, any resulting suitably connected mosaic is a nonreduced knot mosaic.

Observation 1.3 can be generalized in the following way.

**Observation 1.4.** Let $M$ be a knot mosaic with two corner crossing tiles in a top row of the inner board. If the top boundary of the row has an odd number of
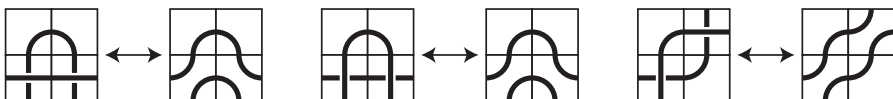


**Figure 4.** Reidemeister type II moves.

**Figure 5.** Two examples of knot mosaics with an odd number of connection points on the top boundary of the top row of the inner board; a connection point on the boundary is marked by red circle.

connection points, then a Reidemeister type I move can be applied to either corner of the row of $M$. This extends via rotation to the outer-most columns and the lower-most row of the inner board. See Figure 5 for examples of such knot mosaics.

Armed with this quick introduction to knot mosaics, we are ready to determine the mosaic number for all prime knots with a crossing number of eight or fewer in the next section. Before we proceed, it should be noted that there are many other questions to consider regarding knot mosaics besides finding the mosaic number. For example, what is the fewest number of nonblank tiles needed to create a specific knot? This could be known as the *tile number* of a knot. With this in mind, if we allow knot mosaics to be rectangular, can some knots have a smaller tile number if they are presented in a rectangular $m \times n$ configuration as opposed to a square configuration? We will conclude this article with a number of other open questions about knot mosaics that you can consider and try to solve.

## 2. Determining the mosaic number of small prime knots

In this section, we will determine the mosaic number for all prime knots of eight or fewer crossings. We refer to these as "small" prime knots. We will see that for some knots, the mosaic number is "obvious", while others take some considerable work. We begin with knots[1] of an obvious mosaic number as shown in Table 1.

Why do these knots have obvious mosaic numbers? As previously noted, the trefoil knot ($3_1$) cannot fit on a $3 \times 3$ mosaic board, as such a board would only allow one crossing tile when $3_1$ requires at least three crossings. Hence, the mosaic number for the trefoil is obvious. Similarly, the knots $5_1$, $5_2$, and $6_2$ have more than four crossings, so they cannot fit on a $4 \times 4$ mosaic board, which only allows at most four crossing tiles. In the Appendix, we have provided representations of these knots on $5 \times 5$ mosaic boards, thus determining the mosaic number for these knots. It should be noted that as the knots become larger, it is often difficult to determine whether a specific knot mosaic represents a given knot. To check that a knot mosaic

---

[1]At this point, we adopt the Alexander–Briggs notation for knots. The number represents the number of crossings, while the subscript represents the order in the table as developed by Alexander–Briggs and extended by Rolfsen. The knot $6_2$ is the second knot of six crossings in the Rolfsen knot table [Adams 2004].

unknot                King Solomon knot (link)
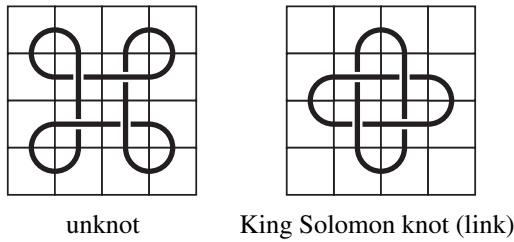
**Figure 6.** Possible four crossing mosaics on $4 \times 4$ board.

represents a specific knot, we used a software packaged called KnotScape [Hoste and Thistlethwaite 1999] developed by Professor Morwen Thistlethawite, which looks at the Dowker notation of a knot to determine the knot presented. While KnotScape cannot determine all knots, it can determine small prime knots. For more information, see [Adams 2004].

Next we consider knots whose mosaic number is "almost obvious". At first glance, one may think that the figure-eight knot ($4_1$) should have a mosaic number of four. Start with a $4 \times 4$ mosaic with the four inner tiles being crossing tiles. By Observation 1.1, this $4 \times 4$ mosaic can be completed in two ways, as seen in Figure 6. However, the knot $4_1$ is known as an *alternating* knot. An alternating knot is a knot with a projection that has crossings that alternate between over and under as one traverses around the knot in a fixed direction. So, if we were to try to place $4_1$ on a $4 \times 4$ mosaic, there would be four crossing tiles and they would have to alternate. Thus $4_1$ cannot be placed on a $4 \times 4$ board. In the Appendix we see a presentation of $4_1$ on a $5 \times 5$ mosaic board, hence $m(4_1) = 5$.

Another knot with an almost obvious mosaic number is $6_1$. Figure 7 first depicts a configuration of $6_1$ on a $6 \times 6$ mosaic board. However, by performing a move called a *flype* (see [Adams 2004]) we can fit $6_1$ on a $5 \times 5$ mosaic board. It should be noted that this mosaic representation of $6_1$ has seven crossings instead of six. Thus the mosaic number for $6_1$ is realized when the crossing is not. It turns out that $6_1$ is not the only knot with such a property. Ludwig, Evans, and Paat [Ludwig et al. 2013] created an infinite family of knots whose mosaic numbers were realized only when their crossing numbers were not.
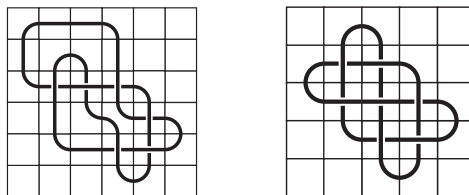


**Figure 7.** The knot $6_1$ as a 6-mosaic and a 5-mosaic.
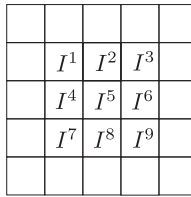
**Figure 8.** The $5 \times 5$ mosaic board with inner-tiles $I^1$–$I^9$.

At this point, we have determined the mosaic number for all six or fewer crossing knots except $6_3$. Surprisingly, we will see that $6_3$ cannot fit on a $5 \times 5$ board, even though such a board has nine possible positions to place crossing tiles.

**Theorem 2.1.** *The mosaic number of the knot $6_3$ is six; that is, $m(6_3) = 6$.*

*Proof.* Since $6_3$ has six crossings, we know that $m(6_3) \geq 5$. In the Appendix we see a representation of $6_3$ on a $6 \times 6$ mosaic board, so $m(6_3) \leq 6$. This means $m(6_3) = 5$ or $m(6_3) = 6$. We now argue $m(6_3) = 6$.

Assume to the contrary that $m(6_3) = 5$. By the definition of the mosaic number, this implies that there is some $5 \times 5$ mosaic $M$ that represents $6_3$. We will show via a case analysis that regardless of how the crossing tiles are arranged on $M$, the resulting knot is in fact *not* $6_3$. This will give us a contradiction, implying that $m(6_3) = 6$.

In order to help with the case analysis, we label the nine inner tiles of the $5 \times 5$ mosaic $I^1$–$I^9$, as depicted in Figure 8. The $6_3$ knot uses at least six crossing tiles. Since $6_3$ has a crossing number of six, by the pigeon hole principle at least one of the four corner inner tiles ($I^1$, $I^3$, $I^7$, or $I^9$) must be a crossing tile. By rotations, it is enough to consider the four cases that are depicted in Figure 9. Note that in Figure 9, gray tiles describe noncrossing tiles and white tiles could be crossing tiles (but they do not have to be).

**Case 1:** Suppose that $I^3$ is a crossing tile, while $I^1$, $I^7$, and $I^9$ are not. Thus $I^2$, $I^4$, $I^5$, $I^6$, and $I^8$ are all crossing tiles since $M$ has exactly six crossing tiles. Every



**Figure 9.** Four different cases for placing crossing tiles on the inner corner tiles for a $5 \times 5$ mosaic.

**Figure 10.** Case 1 results for $6_2$.

reduced projection of an alternating knot is alternating, and since $M$ represents the alternating knot $6_3$, the crossings on $M$ must alternate. A quick inspection shows that to suitably connect the inner tiles of $M$ we would need to ensure that

(i) $I^1$, $I^7$, and $I^9$ are not crossing tiles,

(ii) the crossings are alternating, and

(iii) there are no easily removed crossings.

However, this results in a mosaic that represents $6_2$, as seen in Figure 10. Therefore, $6_3$ cannot be constructed in this case.

**Case 2:** For the case when $M$ has two corner inner tiles that are crossings, we require two subcases.

Case 2(a): Suppose that $I^1$ and $I^3$ are crossing tiles, while $I^7$ and $I^9$ are not. If $I^2$ is a crossing tile then Observation 1.4 may be applied to the top inner row of $M$. Applying this observation will either change $I^1$ or $I^3$ to a noncrossing tile. Without loss of generalization, suppose that $I^1$ is changed. Notice that $M$ now satisfies Case 1, and from the previous analysis, $M$ does not represent $6_3$. Therefore, we may assume that $I^2$ is not a crossing tile.

Since $M$ has at least six crossing tiles, and $I^2$, $I^7$, and $I^9$ are not crossing tiles, the remaining 6 inner tiles must be crossing tiles. Then $I^2$ has four connection points. By Observation 1.4, either $I^1$ or $I^3$ can be changed to a noncrossing tile. Again $M$ falls into Case 1 and does not represent $6_3$.

Case 2(b): Suppose that $I^3$ and $I^7$ are crossing tiles while $I^1$ and $I^9$ are not.

**Claim 2.2.** In Case 2(b), if $M$ has six crossing tiles, then $M$ cannot be $6_3$.

*Proof of Claim.* Assume to the contrary that $M$ only has six crossing tiles. Then exactly one of $\{I^2, I^4, I^5, I^6, I^8\}$ is a noncrossing tile. Up to rotation and reflection, we only need to consider the following two situations:

(i) $I^3$, $I^4$, $I^5$, $I^6$, $I^7$ and $I^8$ are the only crossing tiles.

(ii) $I^2$, $I^3$, $I^4$, $I^6$, $I^7$ and $I^8$ are the only crossing tiles.

**Figure 11.** Possible configurations of six crossing tiles under Case 2(b).

There are two crossing arrangements up to mirror describing case (ii), as shown in the last two images in Figure 11.

Since $M$ has six crossings, the corner inner-tiles $I^3$ and $I^7$ cannot be changed from crossing tiles. With this in mind, suitably connecting the crossing tiles in Figure 11 so that a knot (and not a 2-component link) is created leads to $6_2$ or $3_1 \# 3_1$ (see Figure 12). This is a contradiction, proving our claim. □

**Claim 2.3.** In Case 2(b), if $M$ has seven crossing tiles, then $M$ cannot be $6_3$.

*Proof of Claim.* If the crossing tiles are alternating, then $M$ represents $7_4$, contradicting that $M$ is $6_3$. So assume that $M$ has seven crossings and is nonalternating.

Observe that if any of the pairs $\{I^2, I^3\}$, $\{I^3, I^6\}$, $\{I^4, I^7\}$, or $\{I^7, I^8\}$ are nonalternating, then a type II Reidemeister move is present, and $M$ can be reduced to five crossings. However, this contradicts that $M$ is $6_3$. So each pair $\{I^2, I^3\}$, $\{I^3, I^5\}$, $\{I^4, I^7\}$, and $\{I^7, I^8\}$ is alternating. Since $M$ is nonalternating, at least one of the pairs $\{I^2, I^5\}$, $\{I^6, I^5\}$, $\{I^4, I^5\}$, or $\{I^5, I^8\}$ is nonalternating. Without loss of generality, assume $\{I^2, I^5\}$ creates a pair of nonalternating crossings. Then, up to ambient isotopy, $M$ is $6_1$ or $3_1$, as seen in Figure 13. Therefore Case 2(b) does not result in $6_3$. □

**Case 3:** Suppose that $I^1$, $I^3$ and $I^9$ are crossing tiles, and $I^7$ is not. Since $M$ has at least six crossings, there must be at least three more crossing tiles on the board.

Observe that if either $I^2$ or $I^6$ are crossing tiles, then Observation 1.4 may be applied to the top inner row or the right inner column, respectively. As a result of this, one of the crossings on $I^1$, $I^3$, and $I^9$ can be changed to a noncrossing tile, leaving only two corner inner-tiles that are crossings. This reverts to Case 2
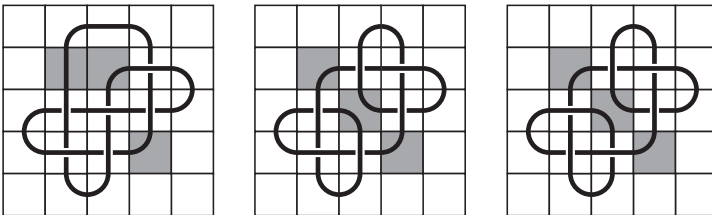


**Figure 12.** Knot mosaic configurations of $6_2$ and $3_1 \# 3_1$.
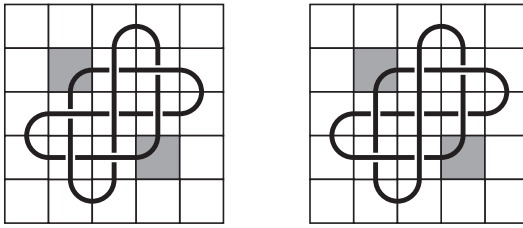
**Figure 13.** $6_1$ and $3_1$, respectively.

showing that $M$ would not represent $6_3$. Therefore, we may assume that neither $I^2$ nor $I^6$ are crossing tiles.

With $I^2$ and $I^6$ eliminated as crossing tiles, $I^1$, $I^3$, $I^4$, $I^5$, $I^8$, and $I^9$ must all be crossing tiles. Then by Observation 1.4, either $I^3$ or $I^9$ can be changed from a crossing tile to a noncrossing tile. This again reverts to Case 2 showing that $M$ would not represent $6_3$.

**Case 4:** Suppose that $I^1$, $I^3$, $I^7$, and $I^9$ are crossing tiles. Note that at least one of the tiles in the set $\{I^2, I^4, I^6, I^8\}$ must be a crossing tile. This means Observation 1.4 applies to some row or column of $M$, and $M$ can be reduced to Case 3. Hence $M$ does not represent $6_3$.

By the above four cases, we see that the $6_3$ cannot be placed on a $5 \times 5$ mosaic. Hence, by the figure for $6_3$ in the Appendix, we see that the mosaic number of $6_3$ is six; that is $m(6_3) = 6$. □

We next consider the seven-crossing knots. As seen above, the knot $7_4$ can be placed on a $5 \times 5$ mosaic board. We formalize this result in the following proposition as well as establish the mosaic number for the other seven-crossing knots.

**Theorem 2.4.** *The mosaic number of $7_4$ is five; that is, $m(7_4) = 5$. Moreover, $7_4$ is the only seven-crossing prime knot with mosaic number five; the remaining seven-crossing prime knots have mosaic number six.*

*Proof.* We have already seen via Observation 1.4 that at most seven crossing tiles can be placed on a $5 \times 5$ board without reduction via a Reidemeister type I move. Moreover, since all seven-crossing knots are alternating, there is only one way to place seven alternating crossing tiles up to mirror, reflection, and rotation as depicted in Figure 14. When this arrangement is suitably connected, the only knot resulting is $7_4$. Therefore $m(7_4) = 5$ and all other seven-crossing knots have mosaic number six as depicted in the Appendix. □

Next we consider the eight-crossing knots. Let $K$ be a knot of eight crossings. By the proof of Theorem 2.1, we know that $K$ cannot fit on a $5 \times 5$ mosaic board. This means $m(K) \geq 6$. Furthermore, there exists a knot mosaic of $K$ on a $6 \times 6$ board (see the Appendix). This implies $m(K) = 6$.
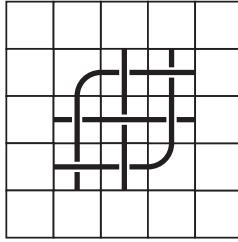
**Figure 14.** Seven alternating tiles on a $5 \times 5$ board.

Given the above arguments, we see that the mosaic number of the eight-crossing knots is greater than five. By the Appendix and use of KnotScape, we see that the mosaic number of all eight-crossing knots is six. We summarize our findings in Table 1. For each knot $K$ with at most eight crossings, the Appendix includes a mosaic of size $m(K)$ representing $K$.

In Table 1, the superscript symbols denote the following properties of the mosaic number:

† Obvious;

‡ from Observation 1.2, we have $m(K) \geq 5$;

♮ by Theorem 2.1; and

♯ by Theorem 2.4.

| $K$ | $m(K)$ |
|-----|--------|
| $0_1$ | $2^\dagger$ |
| $3_1$ | $4^\dagger$ |
| $4_1$ | $5^\ddagger$ |
| $5_1$ | $5^\dagger$ |
| $5_2$ | $5^\dagger$ |
| $6_1$ | $5^\dagger$ |

| $K$ | $m(K)$ |
|-----|--------|
| $6_2$ | $5^\dagger$ |
| $6_3$ | $6^\natural$ |
| $7_1$ | $6^\sharp$ |
| $7_2$ | $6^\sharp$ |
| $7_3$ | $6^\sharp$ |
| $7_4$ | $5^\sharp$ |

| $K$ | $m(K)$ |
|-----|--------|
| $7_5$ | $6^\sharp$ |
| $7_6$ | $6^\sharp$ |
| $7_7$ | $6^\sharp$ |
| $8_1$ | $6$ |
| $8_2$ | $6$ |
| $8_3$ | $6$ |

| $K$ | $m(K)$ |
|-----|--------|
| $8_4$ | $6$ |
| $8_5$ | $6$ |
| $8_6$ | $6$ |
| $8_7$ | $6$ |
| $8_8$ | $6$ |
| $8_9$ | $6$ |

| $K$ | $m(K)$ |
|-----|--------|
| $8_{10}$ | $6$ |
| $8_{11}$ | $6$ |
| $8_{12}$ | $6$ |
| $8_{13}$ | $6$ |
| $8_{14}$ | $6$ |
| $8_{15}$ | $6$ |

| $K$ | $m(K)$ |
|-----|--------|
| $8_{16}$ | $6$ |
| $8_{17}$ | $6$ |
| $8_{18}$ | $6$ |
| $8_{19}$ | $6$ |
| $8_{20}$ | $6$ |
| $8_{21}$ | $6$ |

**Table 1.** Mosaic number of knots with up to eight-crossing.

## 3. Further work

We conclude with a number of open questions that would make good projects for undergraduate research. To begin, often in mathematics when something is proved for the first time, the proof may not be very elegant. For example, Newton's "proofs" of various facts in calculus look much different than the proofs you would find in a typical calculus book today. Although Theorem 2.1 proved $m(6_3) = 6$, could the proof be shortened?

**Question 3.1.** Is there a more direct proof of $m(6_3) = 6$?

It is often interesting to see how various knot invariants compare. For example, in 2009, Ludwig, Paat, and Shapiro showed that the mosaic number and crossing number of a knot can be related in the following way:

$$\lceil \sqrt{c(k)} \rceil + 3 \leq m(k).$$

Moreover, Lee et al. [2014] showed

$$m(K) \leq c(k) + 1.$$

**Question 3.2.** Do tighter upper or lower bounds exist on the mosaic number of a knot using the crossing number of the knot?

Ludwig, Evans, and Paat [Ludwig et al. 2013] created an infinite family of knots whose mosaic number is realized when the crossing number is not. We have seen that $6_1$ is such a knot. The mosaic number for $6_1$ is five, but in that projection, the number of crossing tiles is seven. To realize the crossing number for $6_1$ it has to be projected on a $6 \times 6$ mosaic board. In general, Ludwig et al. created a family of knots whose mosaic number was realized on an $n \times n$ mosaic board with $n$ odd, $n \geq 5$ and whose crossing number was realized on an $(n+1) \times (n+1)$ mosaic board.

**Question 3.3.** Does there exist an infinite family of knots with mosaic number $n \times n$ whose crossing number is realized on an $(n+k) \times (n+k)$ for each $k \geq 2$?

While working with undergraduates, Adams et al. [1997] proved the surprising fact that the composition of $n$ trefoils has stick number exactly $2n + 4$. Is there a similar result for knot mosaics?

**Question 3.4.** What is the mosaic number of the composition of $n$ trefoil knots?

We have briefly touched on mosaic Reidemeister moves. It is often the case that to make these moves, one has to add a certain number of rows and columns to the mosaic board to provide enough room for the moves to occur (see Kuriya and Shehab [2014] for example). Is such an expansion always necessary?

**Question 3.5.** Let $M_1$ and $M_2$ be $n$-mosaics that represent the same knot. Is there a set of mosaic Reidemeister moves from $M_1$ to $M_2$ on a mosaic board of size $n \times n$?
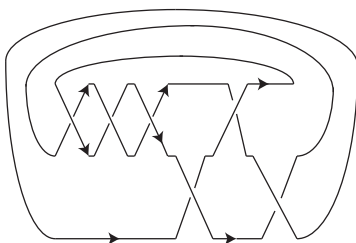
**Figure 15.** A braid representation of the knot $5_2$.

Another well-studied area of knot theory is braid diagrams. J. W. Alexander [1923] proved that every knot or link has a closed braid representation. From Figure 15, we see that braids appear "rectangular" in nature, which leads to the next question.

**Question 3.6.** If we allow mosaics to be rectangular, what is the smallest rectangular board on which we can place a knot?

**Definition 3.7.** Let $t(K)$ denote the tile number of a knot $K$, the fewest non-$T_0$ tiles needed to construct a given knot.

**Question 3.8.** What is the tile number of the knots of 10 or fewer crossings?

**Question 3.9.** Is there an infinite family of knots whose tile number can be determined?

Lastly, it should be noted that in the Appendix, the knots $8_3$, $8_6$, $8_9$, and $8_{11}$ are depicted with nine or more crossing tiles. Therefore, the crossing numbers of these knots are not realized in this representation.

**Question 3.10.** Does there exist a representation of $8_3$ (respectively $8_6$, $8_9$, or $8_{11}$) on a $6 \times 6$ board with only eight crossing tiles?

These are just a few of the questions about knot mosaics that one can consider. For those interested in studying knot mosaics using a tangible manipulative, visit Thingiverse.com to 3D print your very own knot mosaics!

## Appendix: Knots on 5 × 5 mosaic boards
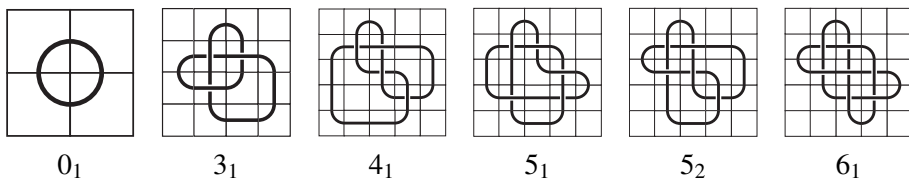


$0_1$        $3_1$        $4_1$        $5_1$        $5_2$        $6_1$
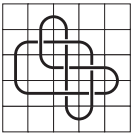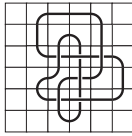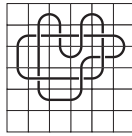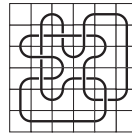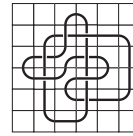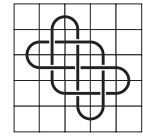
$6_2$ $6_3$ $7_1$ $7_2$ $7_3$ $7_4$
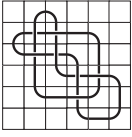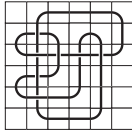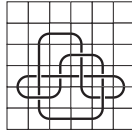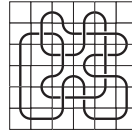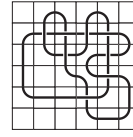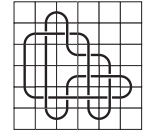
$7_5$ $7_6$ $7_7$ $8_1$ $8_2$ $8_3$
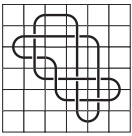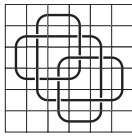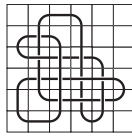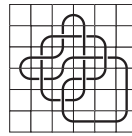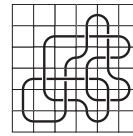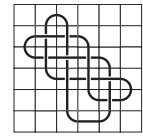
$8_4$ $8_5$ $8_6$ $8_7$ $8_8$ $8_9$
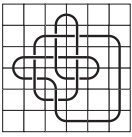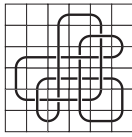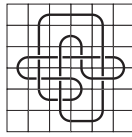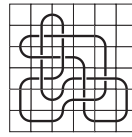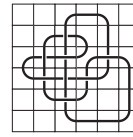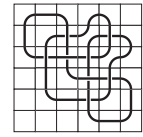
$8_{10}$ $8_{11}$ $8_{12}$ $8_{13}$ $8_{14}$ $8_{15}$
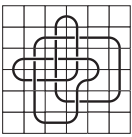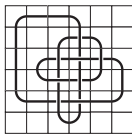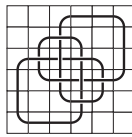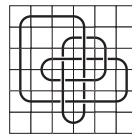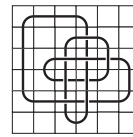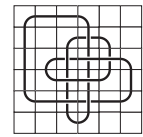
$8_{16}$ $8_{17}$ $8_{18}$ $8_{19}$ $8_{20}$ $8_{21}$

# References

[Adams 2004] C. C. Adams, *The knot book: an elementary introduction to the mathematical theory of knots*, American Mathematical Society, Providence, RI, 2004. MR Zbl

[Adams et al. 1997] C. C. Adams, B. M. Brennan, D. L. Greilsheimer, and A. K. Woo, "Stick numbers and composition of knots and links", *J. Knot Theory Ramifications* **6**:2 (1997), 149–161. MR Zbl

[Alexander 1923] J. W. Alexander, "A lemma on systems of knotted curves", *Proc. Natl. Acad. Sci. USA* **9**:3 (1923), 93–95. JFM

[Hong et al. 2014] K. Hong, H. Lee, H. J. Lee, and S. Oh, "Small knot mosaics and partition matrices", *J. Phys. A* **47**:43 (2014), art. id. 435201, 13 pp. MR Zbl

[Hoste and Thistlethwaite 1999] J. Hoste and M. Thistlethwaite, "Knotscape", software package, 1999, available at http://www.math.utk.edu/~morwen/knotscape.html.

[Kuriya and Shehab 2014] T. Kuriya and O. Shehab, "The Lomonaco–Kauffman conjecture", *J. Knot Theory Ramifications* **23**:1 (2014), art. id. 1450003, 20 pp. MR Zbl

[Lee et al. 2014] H. J. Lee, K. Hong, H. Lee, and S. Oh, "Mosaic number of knots", *J. Knot Theory Ramifications* **23**:13 (2014), art. id. 1450069, 8 pp. MR Zbl

[Lomonaco and Kauffman 2008] S. J. Lomonaco and L. H. Kauffman, "Quantum knots and mosaics", *Quantum Inf. Process.* **7**:2-3 (2008), 85–115. MR Zbl

[Ludwig et al. 2013] L. D. Ludwig, E. L. Evans, and J. S. Paat, "An infinite family of knots whose mosaic number is realized in non-reduced projections", *J. Knot Theory Ramifications* **22**:7 (2013), art. id. 1350036, 11 pp. MR Zbl

[Reidemeister 1927] K. Reidemeister, "Elementare Begründung der Knotentheorie", *Abh. Math. Sem. Univ. Hamburg* **5**:1 (1927), 24–32. MR Zbl

hjwith@dgist.ac.kr    School of Undergraduate Studies, College of Transdisciplinary Studies, Daegu Gyeongbuk Institute of Science and Technology, Daegu 42988, South Korea

ludwigl@denison.edu    Department of Mathematics and Computer Science, Denison University, 100 West College, Granville, OH 43023, United States

jpaat1@jhu.edu    Department of Applied Mathematics and Statistics, Johns Hopkins, 211-E Whitehead Hall, 3400 N. Charles St., Baltimore, MD 21218-2682, United States

peiffe_a1@denison.edu    Department of Math and Computer Science, Denison University, Granville, OH 43020, United States

Current address:    Program in Chemical Biology, University of Michigan, 930 N University Ave., Ann Arbor, MI 48109, United States

msp

# Extending hypothesis testing with persistent homology to three or more groups

### Christopher Cericola, Inga Johnson, Joshua Kiers, Mitchell Krock, Jordan Purdy and Johanna Torrence

(Communicated by Kenneth S. Berenhaut)

We extend the work of Robinson and Turner to use hypothesis testing with persistent homology to test for measurable differences in shape between the spaces of three or more groups. We conduct a large-scale simulation study to validate our proposed extension, considering various combinations of groups, sample sizes and measurement errors. For each such combination, the percentage of p-values below an $\alpha$-level of 0.05 is provided. Additionally, we apply our method to a cardiotocography data set and find statistically significant evidence of measurable differences in shape between the spaces corresponding to normal, suspect and pathologic health status groups.

## 1. Introduction

Consider a data set, obtained via random sampling, where each data point is a vector of $m$ quantitative variables and one categorical variable with $s$ levels. Ideally, several of the quantitative variables are real-valued. According to the levels of the categorical variable, we will group the data points into $s$ not necessarily distinct collections of points in $\mathbb{R}^m$, referred to as point clouds. For each group, we can view the corresponding point cloud as a representative subset of a space which consists of all such points in $\mathbb{R}^m$ with the respective level of the categorical variable. Of interest is whether or not these $s$ spaces have measurably different shapes? But what does shape even mean if $m$ is large?

Topology, in particular algebraic topology, is an area of mathematics that can be used to qualitatively measure the shape of a point cloud. For a given point cloud, we construct an infinite family of simplicial complexes that vary according to a real-valued distance parameter. Each complex in the family is an object that inherits a shape from the point cloud and the topological tool known as homology can be used

to detect this shape. Since any single complex within the infinite family corresponds to a choice of parameter value, we might ask which parameter value "best" captures the shape of the point cloud? Persistent homology is a study of the homological features that persist over long intervals of the distance parameter, thus sidestepping the search for a best choice parameter value. Hence, persistent homology can be used to determine if point clouds have different shape. While persistent homology allows comparisons of shapes across point clouds obtained from a sample of data points, can any resulting differences then be generalized to the corresponding spaces at large? The answer is yes, but as random sampling unavoidably introduces variability, a method is needed which can distinguish true differences in shape between the spaces from artificial differences in shape between the point clouds obtained via random sampling of data points. Statistical hypothesis testing is an inferential method often implemented to assess whether or not randomly sampled data provide sufficient evidence of a difference, with respect to some characteristic, between two or more populations (or, as we have been calling them, spaces). K. Turner and A. Robinson [2013] conducted such an assessment on $s = 2$ spaces using a specific type of hypothesis testing procedure known as a permutation test, where the characteristic of interest is shape, as measured via persistent homology. As this procedure requires multiple point clouds from both spaces, in practice the two point clouds obtained from the random sample of data points are further partitioned, via subsampling, into multiple "smaller", or less dense, point clouds. The assessment is then conducted using the persistent homology of these subsampled point clouds within the procedure. We extend this procedure to three or more spaces, $s \geq 3$.

The remainder of the paper is organized as follows. In Section 2 we provide definitions of the Vietoris–Rips complex of a point cloud, homology groups, persistent homology and persistence diagrams. In Section 3 we describe the permutation test of Robinson and Turner. In Section 4 we propose an extension of the permutation test for three or more spaces. In Section 5 we present the results of a large-scale simulation study, incorporating various measurement errors and sample sizes, that validate our proposed extension. Finally, in Section 6 we apply our extension to a cardiotocography data set and find significant evidence of differences in shape, as measured by persistent homology, between the spaces corresponding to normal, suspect and pathologic health groups.[1]

## 2. Persistent homology

Before defining the persistent homology of a point cloud, we associate to the point cloud a nested family of abstract simplicial complexes. A thorough explanation of

---

[1] Throughout we use "difference in shape" to mean shape as measured by persistent homology in a specified dimension.

simplicial complexes and abstract simplicial complexes is given in [Edelsbrunner and Harer 2010; Munkres 1984]. Here we motivate the definition of an abstract simplicial complex with a brief geometric introduction to simplicial complexes, followed by the definition of the Vietoris–Rips complex, which is the abstract simplicial complex used herein.

Geometrically, a 0-simplex is a point, a 1-simplex is a line segment, a 2-simplex is a triangular subset of a plane, a 3-simplex is a solid tetrahedron, and an $n$-simplex is the $n$-dimensional analogue of these convex sets. Observe that the boundary of an $n$-simplex, $\sigma$, is a collection of $(n-1)$-simplices; these boundary simplices are called faces of $\sigma$. A *simplicial complex* is a collection of simplices in $\mathbb{R}^d$ that satisfy certain subset and intersection properties specifying how simplices can be put together to create a larger structure. More precisely, a simplicial complex is a finite collection of simplices, $K$, such that (1) if $\sigma \in K$ and $\rho$ is a face of $\sigma$ then $\rho \in K$, and (2) given any two simplices $\sigma_1, \sigma_2 \in K$, either $\sigma_1 \cap \sigma_2$ is the empty set or a face of both $\sigma_1$ and $\sigma_2$. More generally, and without relying on geometry, an *abstract simplicial complex* is a finite collection of sets, $A$, such that if $\alpha \in A$ and $\beta \subseteq \alpha$, then $\beta \in A$. It is well known that a finite abstract simplicial complex can be geometrically realized as a simplicial complex in $\mathbb{R}^N$ for $N$ sufficiently large.

**2.1.** *The Vietoris–Rips complex.* The *Vietoris–Rips complex*, denoted VR$(D, r)$, is an abstract simplicial complex associated to a point cloud $D$ for a fixed radius value $r > 0$. The elements of $D$ form the 0-simplices or vertex set of VR$(D, r)$. A simplex of VR$(D, r)$ is a finite subset $\alpha$ of $D$ such that the diameter of $\alpha$ is less than $r$. A simplex $\alpha \subseteq D$ with $k$-elements is called a $(k-1)$-simplex of $D$. Thus, a 1-simplex corresponds to a two element set (viewed geometrically as the endpoints of a line segment), a 2-simplex corresponds to a three element set (viewed as the vertices of a triangle), and so on. Observe that if $\alpha$ is a $k$-simplex, then every subset of $\alpha$ is a simplex of $D$ as the diameter of a subset of $\alpha$ can be no larger than the diameter of $\alpha$. Hence the Vietoris–Rips complex satisfies the definition of an abstract simplicial complex. For readers that are new to topological data analysis, an example Vietoris–Rips complex is given in the Appendix.

We note that Vietoris–Rips complexes for increasing radius values are always a nested family of simplicial complexes associated to $D$; that is, the complexes satisfy

$$\text{VR}(D, r_1) \subseteq \text{VR}(D, r_2) \quad \text{whenever } r_1 \leq r_2.$$

This nested feature of the complexes along with the functorial nature of homology are what give rise the concept of persistence to be defined below.

Although the Vietoris–Rips complex is relatively straightforward to define and calculate, it can be computationally expensive when used with large point clouds. There are economical alternatives to the Vietoris–Rips complex, such as the lazy

witness complex introduced in [de Silva and Carlsson 2004]. Persistent homology can be applied using any nested family of complexes indexed by some parameter.

**2.2. *Homology.*** The homology of a simplicial complex $K$ is an algebraic measurement of how the $n$-simplices are attached to the $(n-1)$-simplices within $K$. Below we define some technical machinery (chains, boundary maps, and cycles) used to define homology groups.

The *p-chains* of a simplicial complex $K$, denoted $C_p(K)$, is the group of formal linear combinations of the $p$-simplices of $K$ with coefficients from $\mathbb{Z}_2$. (More general definitions of homology with ring coefficients can be found in the standard algebraic topology texts [Edelsbrunner and Harer 2010; Hatcher 2002].) Since $\mathbb{Z}_2$ is a field, the $p$-chains of $K$ are $\mathbb{Z}_2$-vector spaces with basis the $p$-simplices of $K$.

The *boundary map*, denoted $\delta_p$, identifies each $p$-chain with its boundary, a $(p-1)$-chain. Each boundary map, $\delta_p : C_p \to C_{p-1}$, is a homomorphism and in the case of $\mathbb{Z}_2$ coefficients, as considered here, these maps are linear transformations.

Notice that $\delta_p \circ \delta_{p+1}$ is the zero map as the boundary of a boundary is empty. This fundamental property of chain complexes ensures that the image of $\delta_{p+1}$ is a normal subgroup of the kernel of $\delta_p$. The collective sequence of boundary maps and chains, as shown below, is called a *chain complex*:

$$\cdots \xrightarrow{\delta_n} C_n(K) \xrightarrow{\delta_{n-1}} \cdots \xrightarrow{\delta_2} C_1(K) \xrightarrow{\delta_1} C_0(K) \xrightarrow{\delta_0} 0.$$

Homology groups are defined using both the kernel and image of each boundary map. The kernel of $\delta_p$ is the set of all $p$-chains whose boundary is empty. The elements of the kernel of $\delta_p$ are called *p-cycles* of $K$. The image of $\delta_{p+1}$ is the set of $p$-chains that are boundaries of a $(p+1)$-chain. The *p-th homology group of $K$*, denoted $H_p(K; \mathbb{Z}_2)$, is defined as the quotient group $\ker(\delta_p)/\operatorname{im}(\delta_{p+1})$.

As the parameter $r > 0$ increases, the Vietoris–Rips complex includes more simplices, thus the homology of the complex changes. The functorial property of homology and the inclusion map $i : \mathrm{VR}(D, r_1) \to \mathrm{VR}(D, r_2)$ whenever $r_1 \leq r_2$, give rise to induced maps between the homology of the complexes

$$i_* : H_*(\mathrm{VR}(X, r_1); \mathbb{Z}_2) \to H_*(\mathrm{VR}(X, r_2); \mathbb{Z}_2).$$

A nontrivial homology class $\alpha \in H_*(\mathrm{VR}(X, r_1); \mathbb{Z}_2)$ is said to be born at radius $r_b$ if $r_b$ is the least radius value for which $H_*(\mathrm{VR}(X, r_b); \mathbb{Z}_2)$ contains an element mapping onto $\alpha$ under the map

$$H_*(\mathrm{VR}(X, r_b); \mathbb{Z}_2) \to H_*(\mathrm{VR}(X, r_1); \mathbb{Z}_2).$$

The homology class $\alpha$ is said to die at radius value $r_d$ provided that $r_d$ is the least radius value for which the class $\alpha$ maps to zero in the mapping

$$H_*(\mathrm{VR}(X, r_1); \mathbb{Z}_2) \to H_*(\mathrm{VR}(X, r_d); \mathbb{Z}_2).$$
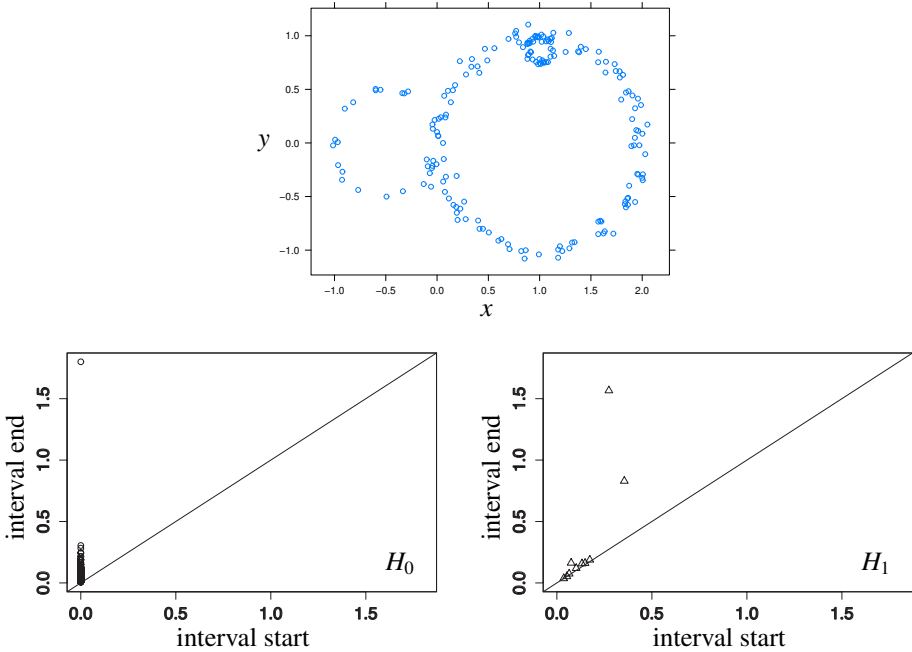
**Figure 1.** An example data set (top) and the corresponding persistence diagrams for the homological dimensions 0 and 1.

The topological feature that $\alpha$ represents is then said to have a birth and death "time" corresponding to the radius values $r_b$ and $r_d$. We say that the class $\alpha$ persists over the interval $[r_b, r_d]$. Persistent homology of a data set $D$ is a cataloguing of the homological classes of the abstract simplicial complexes $\mathrm{VR}(D, r)$ that persist for large intervals of radius values, $r$.

For a fixed $k$, the *persistence diagram for* $H_k(\mathrm{VR}(X, *); \mathbb{Z}_2)$ is a plot of points $(r_b, r_d)$ for each nonzero class $\alpha \in H_k(\mathrm{VR}(X, *); \mathbb{Z}_2)$.

Figure 1 contains an example data set that includes several 1-dimensional homological features of varying size and the corresponding persistence diagrams in dimensions 0 and 1.

Within the persistence diagram in Figure 1, we see two lone triangles at the points $p_1 = (0.35, 0.8)$ and $p_2 = (0.3, 1.55)$. The point $p_2$, with the early birth time, is the 1-dimensional homology class representing the larger circular feature on the right. The earlier birth time is due to the closer scattering of the data points about the larger circle. The point $p_1$, with the earlier death time, is the 1-dimensional homology class representing circle of smaller radius on the left. The early death time is due to the smaller radius of this circular feature. The persistence diagram in Figure 1 also contains several triangles near the diagonal which represent classes that only persist for a short while, and it includes a triangle at the point $(0.1, 0.15)$
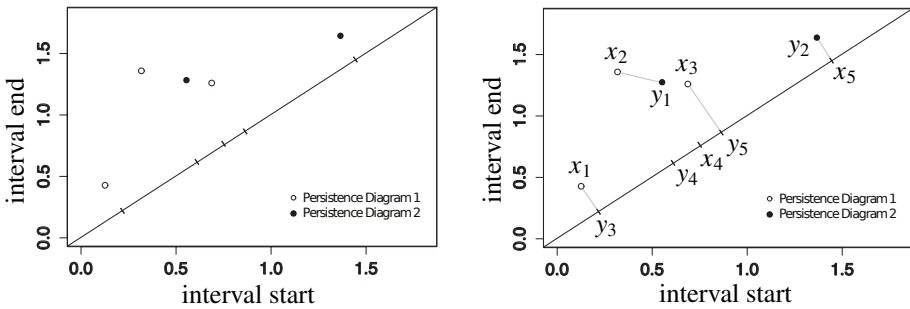
**Figure 2.** On the left, two superimposed persistence diagrams of the same homological dimension. On the right, the points $\{x_1, \ldots, x_5\}$, $\{y_1, \ldots, y_5\}$ and line segments indicating the optimal bijection. The diagram distance is the sum of the lengths of the line segments $\overline{x_1 y_3} + \overline{x_2 y_1} + \overline{x_3 y_5} + \overline{x_5 y_2}$. The segment $\overline{x_4 y_4}$ is not included as it is a segment between diagonal points.

representing the 1-dimensional homology class resulting from the tiny circle of points at the top of the larger circle. Notice that the 0-dimensional homology classes, which are plotted as small circles in the persistence diagram, all have birth time $r = 0$ as a result of each data point representing a unique 0-dimensional class at $r = 0$. As $r$ increases, the complex consists of fewer connected components until it is one connected component. The 0-dimensional persistence class plotted at the point $(0, 0.35)$ represents the joining of the last two components into a single component. In other words, for $r \geq 0.35$ the simplicial complex $\text{VR}(X, r)$ is one connected component. The 0-dimensional class plotted at $(0, 2)$ is merely the result of using a maximum radius value of $r = 2$ in the persistent homology calculation. This class indicates that the complex $\text{VR}(X, 2)$ is one connected component.

The discussion above defines a persistence diagram for a data set using the Vietoris–Rips complex. There are, however, several other routes that lead to the creation of a persistence diagram. The omnibus test described below can be applied to a collection of persistence diagrams obtained by any means.

**2.3. *A metric on persistence diagrams.*** We follow Robinson and Turner in selecting the metric on persistence diagrams that is analogous to the $L^2$ norm in the space of functions on a discrete space. Given two persistence diagrams $X$ and $Y$, let $x_1, x_2, \ldots, x_n \in X$ be a listing of the off-diagonal points of $X$ and $y_1, y_2, \ldots, y_m \in Y$ be the off-diagonal points of $Y$. Select points $x_{n+1}, \ldots, x_{n+m}$ and $y_{m+1}, \ldots, y_{m+n}$ along the diagonal so that $x_{n+k}$ is the point closest (in Euclidean distance) to $y_k$ and vice versa. Let $X' = \{x_1, \ldots, x_{n+m}\}$ and $Y' = \{y_1, \ldots, y_{n+m}\}$. We consider the set of all bijections $\phi : X' \to Y'$ such that (1) the off-diagonal point $x_k$ is paired either with an off-diagonal point of $Y$ or with $y_{m+k}$ and (2) the diagonal point $x_l$

is paired either with $y_{l-n}$ or with one of the diagonal points in $Y'$. For a specific bijection $\phi$, if both $x_k$ and $y_j$ are diagonal points, the *cost* of assigning $x_k$ to $y_j$, denoted $C(x_k, y_j)$, is 0, else the cost is the Euclidean distance between $x_k$ and $y_j$.

Define $d(X, Y)$, the *distance between the persistence diagrams $X$ and $Y$*, by

$$d(X, Y) = \left( \inf_{\phi: X' \to Y'} \sum_{x \in X'} C(x, \phi(x)) \right)^{\frac{1}{2}}.$$

A bijection between $X$ and $Y$ is called optimal if it achieves the infimum. The Hungarian algorithm [Kuhn 1955; Munkres 1957], also known as Munkres' assignment algorithm, presents a method for obtaining an optimal bijection in polynomial time. Figure 2 gives an example of two simple persistence diagrams and the bijection exhibiting their diagram distance.

## 3. Hypothesis testing and topological data analysis

When persistent homology is applied to point clouds obtained from a random sample of points from various spaces, an element of variability is unavoidably introduced. Point clouds obtained from different samples of the same space, if somewhat representative, are expected to have "small" differences in their respective persistence diagrams, while point clouds obtained from samples of different spaces are expected to have comparatively "large" differences in their persistence diagrams. However, when the true shape-related features of two spaces are unknown, and all that is available are the point clouds obtained from samples of each of these spaces, what qualifies as a small or large difference is unclear. A tool is needed which can determine whether or not the shapes of the underlying spaces are measurably different. Statistical hypothesis testing is a method that can be implemented in these situations to decide if there is sufficient evidence to classify the shapes of the spaces as measurably different. A thorough development of statistical hypothesis testing is available in many standard sources, including [Casella and Berger 2002; DeGroot and Schervish 2012].

**3.1. *Hypothesis testing via the joint loss function.*** Consider two spaces in $\mathbb{R}^m$, arbitrarily labeled $X_1$ and $X_2$, suspected of having measurably different shapes. Suppose $n_1$ point clouds are available from $X_1$ and $n_2$ point clouds are available from $X_2$, with their corresponding persistence diagrams in a fixed dimension denoted respectively by $X_{1,1}, X_{1,2}, \ldots, X_{1,n_1}$ and $X_{2,1}, X_{2,2}, \ldots, X_{2,n_2}$. Further suppose that each of these $n_1+n_2$ point clouds was obtained via random sampling from either $X_1$ or $X_2$. Note that in practice, for each space, a single point cloud will usually be obtained via a random sample of $X_i$ and then partitioned, via subsampling, into $n_i$ smaller, or less dense, point clouds. Within the statistical hypothesis testing paradigm, the null hypothesis asserts that the shapes of $X_1$ and $X_2$ are not measurably

different, while the alternative hypothesis asserts the opposite. The corresponding test statistic, proposed by Robinson and Turner [2013], is the *joint loss function*

$$\sigma_{\chi_2}^2 = \sum_{m=1}^{2} \frac{1}{2n_m(n_m - 1)} \sum_{i=1}^{n_m} \sum_{j=1}^{n_m} d(X_{m,i}, X_{m,j})^2,$$

where $d(\,\cdot\,,\,\cdot\,)$ is the persistence diagram distance metric described in Section 2.3.

The joint loss function is ultimately an aggregate measure of within-group variation. More specifically, $\sigma_{\chi_2}^2$ adds the variation in the $\binom{n_1}{2}$ persistence diagram distances from $X_1$ and the variation in the $\binom{n_2}{2}$ persistence diagram distances from $X_2$. Unfortunately, the sampling distribution of $\sigma_{\chi_2}^2$ is nontrivial to determine and is currently unknown, which renders the "standard" (i.e., distribution-based) hypothesis testing paradigm impossible. To circumvent this, Robinson and Turner propose implementing a permutation test, which in this context is free of any distributional assumptions. A thorough development of permutation tests, and the often corresponding approximate permutation test p-values, is available in [Higgins 2004; Ramsey and Schafer 2013].

To perform the permutation test, we assume that the null hypothesis is true, i.e., $X_1$ and $X_2$ are not measurably different in shape. Such an assumption effectively means that the observed labeling of the point clouds to either space $X_1$ or $X_2$ is just one of $\binom{n_1+n_2}{n_1}$ possible assignments, all of which are arbitrary and equally likely. For each of these possible assignments, the value of $\sigma_{\chi_2}^2$ is then computed. Collectively, these values yield the *permutation distribution* for $\sigma_{\chi_2}^2$, which is analogous to a sampling distribution in the standard hypothesis testing paradigm. Finally, analogous to a standard hypothesis testing p-value, the *permutation test p-value* is obtained by calculating the proportion of values in the permutation distribution which are less than or equal to the observed value of the joint loss function. In practice, the number of possible assignments may be unreasonably large, in which case the above procedure is subtly altered to produce an *approximate permutation test p-value*. In particular, rather than using the $\binom{n_1+n_2}{n_1}$ possible assignments of the $n_1 + n_2$ point clouds to the two spaces, numerous (e.g., 1000) randomly selected permutations (i.e., "shuffles") of the $n_1 + n_2$ point clouds are instead used where after each shuffle the first $n_1$ point clouds are labeled as "belonging" to space $X_1$ and the remaining $n_2$ point clouds are labeled as "belonging" to space $X_2$.

If the null hypothesis of the permutation test is actually false, then we would expect the permutation test p-value to be small since the observed labeling of point clouds would be the only assignment that did not mix point clouds from both spaces. When a permutation test p-value is less than the $\alpha$-level, an a priori established threshold (e.g., 0.05), the observed value of $\sigma_{\chi_2}^2$ is considered smaller than what can reasonably be explained by chance assignment of the point clouds to spaces $X_1$

and $X_2$. The null hypothesis would then be rejected and $X_1$ and $X_2$ classified as having measurably different shape.

It is important to note that if the point clouds were not obtained via random sampling of $X_1$ and $X_2$, then a permutation test only allows us to draw conclusions with respect to the point clouds. For instance, if the permutation test p-value is less than our threshold, then we can conclude that the shapes of the point clouds from $X_1$ and $X_2$ are measurably different; however, this conclusion cannot be generalized to $X_1$ and $X_2$ at large. As limited as such a conclusion may be, it is still informative to know that such differences exist among the point clouds, particularly when $m > 3$ and the corresponding point clouds cannot be visualized.

## 4. Extending hypothesis testing to three or more groups

While the methods of Section 3 are useful for determining whether or not two spaces are measurably different in a particular homological dimension, many practical applications involve more than two spaces. The cardiotocography data set considered in Section 6 is one such example. Given $s \geq 3$ spaces, suppose we have $n_1$ point clouds, obtained via random sampling, from space $X_1$, $n_2$ point clouds from space $X_2$, ..., and $n_s$ point clouds from space $X_s$. Analogous to before, note that in practice, for each space, a single point cloud will usually be obtained via a random sample of $X_i$ and then partitioned, via subsampling, into $n_i$ smaller, or less dense, point clouds. In this section we extend the methods of Section 3 to obtain a hypothesis testing procedure which can determine whether or not sufficient evidence of measurable differences in shape exists between the $s$ spaces.

**4.1.** *Hypotheses and justification.* To conduct such an inquiry, we follow through with the suggestion of Robinson and Turner and use an approach analogous to a standard one-way ANOVA procedure in which there are potentially two stages of hypothesis testing. An omnibus (i.e., "global") test is conducted at the first stage and if this test produces significant results, a number of post hoc (i.e., "local") tests are performed at the second stage to identify the source(s) of the global significance. A thorough development of the one-way ANOVA procedure is available in [Casella and Berger 2002; DeGroot and Schervish 2012; Ramsey and Schafer 2013]. As with the joint loss function in Section 3, the sampling distribution of the test statistic corresponding to the omnibus test, which is presented below in Section 4.2, is nontrivial to determine and currently unknown. Hence, we again use a permutation test to carry out the omnibus test, which we will henceforth refer to as the *omnibus permutation test*. The logic behind and mechanics of this test are developed below in Section 4.2.

The null hypothesis for the omnibus permutation test asserts that the shapes of $X_1$, $X_2$, ..., $X_s$ are not measurably different, while the alternative hypothesis asserts that the shapes of at least two of the $s$ spaces are measurably different. If

we fail to reject the null hypothesis of this omnibus permutation test, then we are done. However, if we reject the null hypothesis, then we know that at least two of the $s$ spaces have shapes that are measurably different, though we do not yet know which spaces. Hence, up to $\binom{s}{2}$ post hoc tests are performed, one for each possible pairing of two of the $s$ spaces. For each post hoc test, the null hypothesis asserts that the shapes of the two spaces are not measurably different, while the alternative hypothesis asserts that the shapes are measurably different. Thus, each post hoc test can be conducted via the methods described in Section 3.

Before describing the test statistic and corresponding details for the omnibus permutation test, note that the primary purpose of the test pertains to management of the familywise type I error rate. A type I error is the general term used to identify a hypothesis test decision in which the null hypothesis is incorrectly rejected. For any single hypothesis test, the pre-established $\alpha$-level is the probability of making a type I error. When multiple post hoc tests are performed, the familywise type I error rate refers to the probability of incorrectly rejecting at least one of the corresponding null hypotheses. Many methods exist for bounding the familywise type I error rate associated with multiple pairwise post hoc tests (e.g., Bonferroni), but such methods invariably require different and smaller $\alpha$-levels for each individual post hoc test. Hence, an insignificant omnibus permutation test result prevents the analyst from unnecessarily performing post hoc tests and needlessly managing the familywise type I error rate. Stated another way, if the null hypothesis of the omnibus permutation test is true, then all of the null hypotheses of the various post hoc tests are also true, and thus do not need to be performed, which eliminates any need to manage the familywise type I error rate. However, if an omnibus permutation test in which the null hypothesis is ultimately true is not performed, then $\binom{s}{2}$ post hoc tests are unnecessarily performed and the familywise type I error rate must needlessly be managed.

**4.2. *Omnibus permutation test specifics.*** Suppose, possibly after subsampling, that $n_1$ point clouds are available from $X_1$, $n_2$ point clouds from $X_2, \ldots,$ and $n_s$ point clouds from $X_s$, with their corresponding persistence diagrams in a fixed dimension denoted respectively by $X_{1,1}, X_{1,2}, \ldots, X_{1,n_1}, X_{2,1}, X_{2,2}, \ldots, X_{2,n_2}$, and $X_{s,1}, X_{s,2}, \ldots, X_{s,n_s}$. Analogous to the test statistic for the two-space permutation test presented in Section 3, the test statistic for the omnibus permutation test, for three or more spaces, is a function of the diagram distances for all $\binom{n_1}{2}$ pairings of persistence diagrams from $X_1$, all $\binom{n_2}{2}$ pairings of persistence diagrams from $X_2, \ldots,$ and all $\binom{n_s}{2}$ pairings of persistence diagrams from $X_s$. In particular, the *omnibus joint loss function* is defined as

$$\sigma_{\chi_s}^2 = \sum_{m=1}^{s} \frac{1}{2n_m(n_m - 1)} \sum_{i=1}^{n_m} \sum_{j=1}^{n_m} d(X_{m,i}, X_{m,j})^2,$$

where $d(\,\cdot\,,\cdot\,)$ is the persistence diagram distance metric described in Section 2.3. Analogous to $\sigma^2_{\chi_2}$, the function $\sigma^2_{\chi_s}$ is ultimately an aggregate measure of variability since the omnibus joint loss function adds the within-group variation of persistence diagram distances from each of the $s$ spaces. As previously mentioned, the sampling distribution of $\sigma^2_{\chi_s}$ is nontrivial to determine and currently unknown; hence, we turn to the omnibus permutation test.

The logic behind and the mechanics of this omnibus permutation test are analogous to the two-space permutation test described in Section 3. We assume that the null hypothesis is true, which effectively means that the observed assignment of the point clouds to the $s$ spaces is just one of $\prod_{i=1}^{s-1} \binom{\sum_{j=i}^{s} n_j}{n_i}$ possible assignments, all of which are arbitrary and equally likely. For each of these possible assignments, the value of $\sigma^2_{\chi_s}$ is then computed. Collectively, these values yield the permutation distribution for $\sigma^2_{\chi_s}$. Finally, the permutation test p-value is then obtained by calculating the proportion of values in the permutation distribution which are less than or equal to the observed value of $\sigma^2_{\chi_s}$. As in the two-space scenario of Section 3, in practice the number of possible assignments may be unreasonably large, in which case the above procedure is analogously altered to produce an *approximate permutation test p-value*. In particular, rather than using the $\prod_{i=1}^{s-1} \binom{\sum_{j=i}^{s} n_j}{n_i}$ possible assignments of the $n_1 + n_2 + \cdots + n_s$ point clouds to the $s$ spaces, numerous (e.g., 1000) randomly selected permutations (i.e., "shuffles") of the $n_1 + n_2 + \cdots + n_s$ point clouds are instead used where after each shuffle the first $n_1$ point clouds are labeled as "belonging" to space $X_1$, the next $n_2$ point clouds are labeled as "belonging" to space $X_2$, ..., and the remaining $n_s$ point clouds are labeled as "belonging" to space $X_s$.

Analogous to the two-space scenario of Section 3, if the null hypothesis of this omnibus permutation test is actually false, then we would expect the permutation test p-value to be small since the observed labeling of point clouds would be the only assignment that did not mix point clouds across the $s$ spaces. The permutation test p-value is then compared to the $\alpha$-level (e.g., 0.05). If the permutation test p-value is smaller than this threshold, then the observed value of $\sigma^2_{\chi_s}$ is considered smaller than what can reasonably be explained by chance assignment of the point clouds to the $s$ spaces. The null hypothesis would then be rejected and at least two of the $s$ spaces are declared as having measurably different shape. To then identify the source(s) of this difference, i.e., to determine which spaces have measurably different shape, a requisite number of post hoc tests are conducted via the two-space methods of Section 3.

## 5. Simulation study

To confirm the two-space permutation test introduced by Robinson and Turner [2013] and to validate our proposed generalization for three or more spaces, we conducted a large-scale simulation study. Throughout the study, shape was measured

via 1-dimensional persistent homology. Three different scenarios were considered and all three consisted of three spaces ($s = 3$). For each scenario, a trial consisted of obtaining 20 point clouds, via random sampling of points, from each of the three spaces and then calculating the approximate omnibus permutation test p-value. While the 20 point clouds from a particular space were ultimately drawn independently, they can be viewed as 20 disjoint subsamples of one larger, i.e., more dense, point cloud obtained via random sampling of points of the space. All approximate omnibus permutation test p-values were based on 100,000 randomly selected permutations of the 60 collective point clouds. In particular, for each permutation, the 60 point clouds were shuffled and then labeled such that the first 20 were in the first space, the next 20 were in the second space, and the final 20 were in the third space. In the third and final scenario, each of the three possible post hoc tests were additionally performed using the two-space permutation test described in Section 3. The corresponding approximate two-space permutation test p-values were based on 100,000 randomly selected permutations of the 40 collective point clouds. In particular, for each permutation, the 40 point clouds were shuffled and then labeled such that the first 20 were in the first space and the final 20 were in the second space. A total of 100 trials were performed for each scenario and the percentage of these 100 trials that produced approximate (omnibus/two-space) permutation test p-values less than or equal to 0.05 was calculated.

## 5.1. *Unbalanced unit circles.*

For the first scenario, each of the three spaces was the unit circle; hence, the omnibus permutation test null hypothesis that there is no measurable difference in shape between the three spaces is ultimately true. The number of sampled points making up a point cloud from each space, however, was not the same (i.e., the sample sizes are unbalanced). Each point cloud in the first space consisted of a random sample of size 18, whereas each point cloud in the second space consisted of a random sample of size 36 and each point cloud in the third space consisted of a random sample of size 54. For all three spaces, samples were obtained without allowing for measurement error; i.e., all sampled points were on their respective unit circle. Counterintuitively, 100% of the 100 trials performed produced approximate omnibus permutation test p-values less than or equal to 0.05. In fact, 100% of the trials produced approximate omnibus permutation test p-values less than 0.01. Thus, in every trial the null hypothesis would be rejected at the 5% level and we would conclude that the shapes of at least two of the three spaces are measurably different.

While such results may appear to suggest that the omnibus permutation test is ineffective, ultimately these results are an expected consequence of allowing different (i.e., unbalanced) sample sizes across point clouds. Relative to a point cloud obtained from a random sample of size 18 from the unit circle, a point cloud

obtained from a random sample of size 54 is likely to produce a persistence diagram (corresponding to homology dimension 1) containing a point that is measurably further from the diagonal. This point in the persistence diagram is expected as the circular feature within the point cloud will be born sooner and thus persist for a longer time interval. Hence, in order for the hypothesis testing methods described in Sections 3 and 4 to detect truly measurable differences in shape between the various spaces, every point cloud, both within a space and across spaces, must consist of the same number of randomly sampled data points. We will henceforth refer to this procedural necessity as balanced sampling. In practice, balanced sampling will usually be implemented at the subsampling level when the sampled data points of a space are partitioned, via subsampling, into multiple point clouds; this is demonstrated using the cardiotocography data set considered in Section 6.

**5.2. *Balanced samples from circles with varying radius.*** For the second scenario, the three spaces were circles with radii of 1, $\frac{1}{2}$ and $\frac{1}{3}$ units. Notice that these three spaces are topologically equivalent, though geometrically different, and there is in fact a measurable difference in shape among the three spaces as measured by persistent homology in dimension 1. Hence, the null hypothesis for the corresponding omnibus permutation test is ultimately false. Point clouds for each of the three circles consisted of random samples of size 24. As in the unbalanced unit circles scenario, all samples were obtained without allowing for measurement error; i.e., all sampled points were on their respective circle. Of the 100 trials performed, 100% of them produced approximate omnibus permutation test p-values less than or equal to 0.05. In fact, as in the unbalanced unit circles scenario, 100% of the trials produced approximate omnibus permutation test p-values less than 0.01. Hence, in every trial the null hypothesis would be rejected at the 5% level and we would conclude that the shapes of at least two of the three spaces are measurably different.

As the three spaces of this second scenario are all topologically equivalent, these results suggest that the omnibus permutation test is capable of recognizing when purely geometrical differences exist between the spaces. Stated another way, this second scenario suggests that the hypothesis testing methods described in Sections 3 and 4 are not scale invariant. This is not a surprising result. More specifically, as seen in the example data of Figure 1, a point cloud obtained from a sample of points from the circle with radius $\frac{1}{3}$ will result in birth and death times for comparatively smaller radii values than a point cloud obtained from a sample of points from the unit circle. This is an artifact of the distances between neighboring points in the point cloud from the circle with radius $\frac{1}{3}$ typically being smaller than those from the unit circle. While in practice it will usually be difficult to determine whether a significant hypothesis test is a result of topological or geometrical differences between the various spaces, it is informative nonetheless to find evidence of any measurable difference in shape.
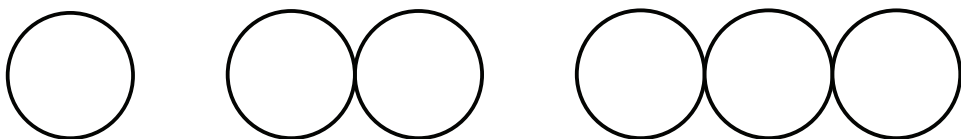
**Figure 3.** The three spaces of the first case of the balanced wedges simulation scenario. On the left is the unit circle, in the middle is the wedge of two unit circles, and on the right is the wedge of three unit circles.
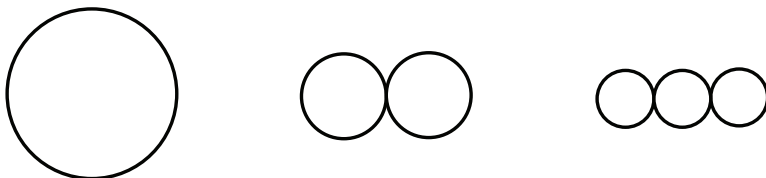


**Figure 4.** The three spaces of the second case of the balanced wedges simulation scenario. On the left is the unit circle, in the middle is the wedge of two circles of radius $\frac{1}{2}$, and on the right is the wedge of three circles of radius $\frac{1}{3}$.

**5.3. *Balanced wedges.*** The third and final scenario consisted of three distinct, but related cases in which only balanced sample sizes were considered. In the first case, the three spaces were the unit circle, the 2-wedge consisting of two unit circles, and the 3-wedge consisting of three unit circles. Hence, in this first case, the radius of every component circle is 1. An image of these three spaces is given in Figure 3. In the second case, the three spaces were the unit circle consisting of one circle of radius 1, the 2-wedge consisting of two circles of radius $\frac{1}{2}$, and the 3-wedge consisting of three circles of radius $\frac{1}{3}$. Hence, in this second case, the radii of the component circles within a space sum to 1. An image of these three spaces is given in Figure 4. In the third and final case, the three spaces were the unit circle, the unit circle with a single chord traversing the interior of the circle, and the unit circle with two nonintersecting chords traversing the interior of the circle. Hence, in this third case, the area of each of the three spaces is $\pi$ units. An image of these three spaces is given in Figure 5. Observe that across these three scenarios the representations of the three spaces are topologically equivalent, but geometrically different. We consider all three scenarios since persistence diagrams are unavoidably influenced by such differences.

Within each of the three cases, the null hypothesis of the omnibus permutation test is ultimately false. In other words, there are measurable differences in shape between the three spaces. The point clouds for each of the three spaces, in all
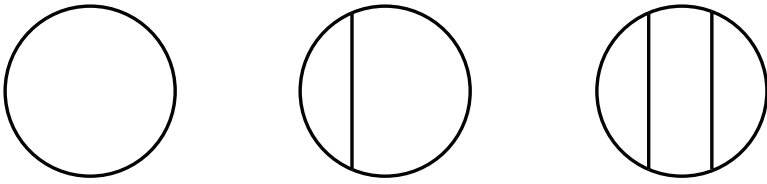
**Figure 5.** The three spaces of the third case of the balanced wedges simulation scenario. On the left is the unit circle, in the middle is the unit circle with a single chord, and on the right is the unit circle with two nonintersecting chords.

three cases, were obtained from random samples of the same size (i.e., balanced samples). Ten different sample sizes were considered: 6, 12, 18, 24, 30, 36, 42, 48, 54, and 60. Figure 6 provides examples of point clouds obtained from random samples of sizes 12 and 60, respectively, from each of the three spaces for the second case.

For each of these ten sample sizes, three distinct measurement errors were considered: 0 (i.e., no error), $\frac{1}{3}$, and $\frac{2}{3}$ units. For example, in the 2-wedge of the first case, measurement error was incorporated in the following manner. A random sample of points was obtained separately from each of the two unit circles of the 2-wedge. Each point on the circles was obtained by randomly selecting the angle of the point from a uniform distribution $U(0, 2\pi)$. Each point was then assigned a radius value of 1 and converted to Cartesian coordinates. Finally, for
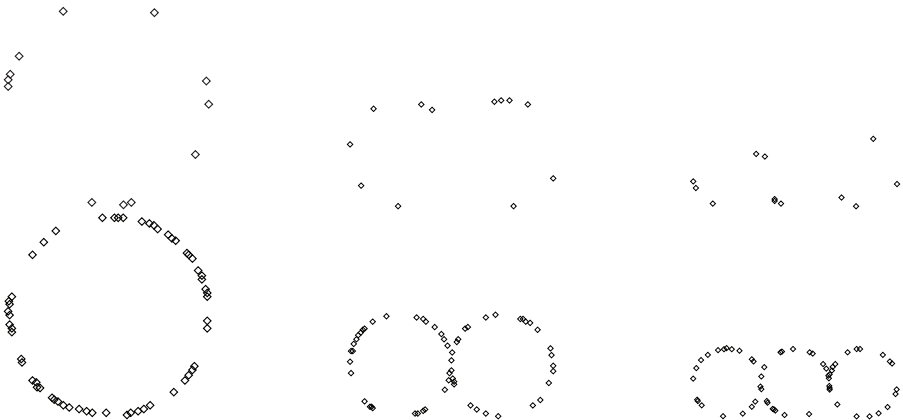


**Figure 6.** Point clouds obtained from random samples of each of the three spaces of the second case of the balanced wedges simulation scenario. The first row contains point clouds obtained from random samples of size 12. The second row contains point clouds obtained from random samples of size 60.

no error              error of $\frac{1}{3}$              error of $\frac{2}{3}$
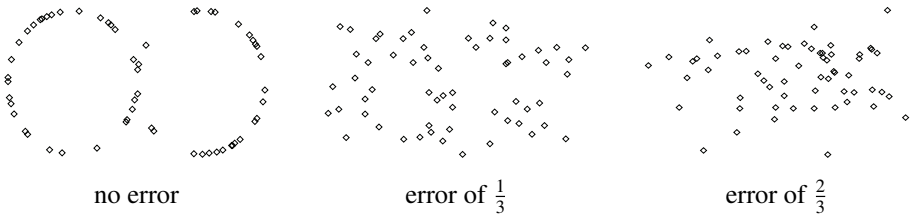
**Figure 7.** Point clouds obtained from random samples of size 60, under various measurement errors, from the 2-wedge of the first case of the balanced wedges simulation scenario.

each point, two errors were randomly sampled from a normal distribution $\mathcal{N}(0, \sigma)$, where $\sigma$ is the specified measurement error $\left(\text{e.g., } \frac{1}{3}\right)$, and respectively added to the Cartesian coordinates of the point. For each of the three measurement errors, Figure 7 exemplifies a point cloud obtained from a sample of size 60 from the 2-wedge. From these images it is clear that as the measurement error increases, the extent to which the point cloud resembles the 2-wedge dramatically decreases. Measurement errors for the other spaces of the second case, as well as for the other cases of the third scenario, was analogously incorporated.

For each of the 30 combinations of sample size and measurement error, the percentage of the 100 trials producing an approximate omnibus permutation test p-value less than or equal to 0.05 for case one is given in Table 1. Two trends are readily apparent from these results. First, as sample size increases for a fixed measurement error, the percentage of significant omnibus permutation test results almost uniformly increases. This is intuitive and desirable since we would expect measurable differences in shape between the three spaces to become more easily identifiable as sample size increases. Second, as measurement error increases for a fixed sample size, the percentage of significant omnibus permutation test results almost uniformly decreases. This too is intuitive and desirable since we would expect measurable differences in shape between the three spaces to become less easily identifiable as measurement error increases. Given these trends and the fact that there are so many entries in the table at or near 100%, these results suggest that the proposed omnibus permutation test successfully identified measurable differences in shape between at least two of these three spaces. The results for the second and third cases, depicted in Figures 4 and 5, are analogous to those above for the first case and, therefore, are omitted.

As the omnibus permutation test successfully identified measurable differences in shape between at least two of the three spaces, in all three cases, each of the three possible post hoc tests were then conducted. For each such post hoc test, the null hypothesis asserts that there is no measurable difference in shape between the two spaces, while the alternative hypothesis asserts the opposite. Hence, in all

| sample size | noise | | |
|:---:|:---:|:---:|:---:|
| | 0 | $\frac{1}{3}$ | $\frac{2}{3}$ |
| 6 | 6% | 9% | 1% |
| 12 | 95% | 57% | 18% |
| 18 | 100% | 65% | 41% |
| 24 | 100% | 96% | 41% |
| 30 | 100% | 100% | 85% |
| 36 | 100% | 100% | 98% |
| 42 | 100% | 100% | 100% |
| 48 | 100% | 100% | 100% |
| 54 | 100% | 100% | 100% |
| 60 | 100% | 100% | 100% |

**Table 1.** Balanced unit wedges — results of omnibus permutation tests. For each combination of sample size and measurement error, the percentage of approximate omnibus permutation test p-values (out of 100) yielding a value less than or equal to 0.05 is given. The three spaces are the unit circle, the 2-wedge and the 3-wedge.

three tests, for all three cases, the null hypothesis is ultimately false. As the results across the three cases were ultimately analogous, only the results for the first case are discussed below. In particular, for each of the 30 combinations of sample size and measurement error, the percentage of the 100 trials producing an approximate post hoc test p-value less than or equal to 0.05 is given in Table 2 for the circle versus the 2-wedge, in Table 3 for the circle versus the 3-wedge, and in Table 4 for the 2-wedge versus the 3-wedge.

The two trends that were apparent in the corresponding omnibus permutation tests for this simulation scenario are also readily apparent in all three of these post hoc tests. Specifically, as sample size increases for a fixed measurement error, the percentage of significant post hoc tests tends to increase. Similarly, as measurement error increases for a fixed sample size, the percentage of significant post hoc tests tends to decrease. A cell-by-cell comparison of the percentages among the three post hoc tests, however, reveals an additional interesting trend. The percentages for the post hoc test between the circle and the 3-wedge are almost uniformly larger than or equal to the corresponding percentages between the circle and the 2-wedge, which are in turn almost uniformly larger than or equal to the corresponding percentages between the 2-wedge and the 3-wedge. This too is mostly intuitive and desirable since, among the three spaces, the unit circle and the three wedge are the most different with respect to shape. We are uncertain why the post hoc test appears more adept at recognizing measurable differences

| sample size | noise | | |
|---|---|---|---|
| | 0 | $\frac{1}{3}$ | $\frac{2}{3}$ |
| 6 | 2% | 5% | 2% |
| 12 | 90% | 29% | 13% |
| 18 | 99% | 40% | 15% |
| 24 | 100% | 83% | 28% |
| 30 | 100% | 97% | 49% |
| 36 | 100% | 100% | 64% |
| 42 | 100% | 100% | 80% |
| 48 | 100% | 100% | 82% |
| 54 | 100% | 100% | 92% |
| 60 | 100% | 100% | 97% |

**Table 2.** Balanced wedges, first case — results of unit circle vs. 2-wedge post hoc tests. For each combination of sample size and measurement error, the percentage of approximate two-space permutation test p-values (out of 100) yielding a value less than or equal to 0.05 is given.

| sample size | noise | | |
|---|---|---|---|
| | 0 | $\frac{1}{3}$ | $\frac{2}{3}$ |
| 6 | 2% | 5% | 1% |
| 12 | 97% | 65% | 30% |
| 18 | 100% | 85% | 40% |
| 24 | 100% | 100% | 53% |
| 30 | 100% | 100% | 95% |
| 36 | 100% | 100% | 100% |
| 42 | 100% | 100% | 100% |
| 48 | 100% | 100% | 100% |
| 54 | 100% | 100% | 100% |
| 60 | 100% | 100% | 100% |

**Table 3.** Balanced wedges, first case — results of unit circle vs. 3-wedge post hoc tests. For each combination of sample size and measurement error, the percentage of approximate two-space permutation test p-values (out of 100) yielding a value less than or equal to 0.05 is given.

in shape between the circle and the 2-wedge rather than between the 2-wedge and the 3-wedge. Regardless, all three of these trends, when coupled with the volume of entries in all three tables which are at or near 100%, indicate that the proposed

| sample size | noise | | |
|---|---|---|---|
| | 0 | $\frac{1}{3}$ | $\frac{2}{3}$ |
| 6 | 0% | 1% | 1% |
| 12 | 4% | 17% | 13% |
| 18 | 62% | 16% | 18% |
| 24 | 86% | 33% | 14% |
| 30 | 93% | 42% | 20% |
| 36 | 87% | 66% | 26% |
| 42 | 95% | 67% | 43% |
| 48 | 99% | 87% | 65% |
| 54 | 100% | 93% | 66% |
| 60 | 100% | 98% | 84% |

**Table 4.** Balanced wedges, first case — results of 2-wedge vs. 3-wedge post hoc tests. For each combination of sample size and measurement error, the percentage of approximate two-space permutation test p-values (out of 100) yielding a value less than or equal to 0.05 is given.

post hoc tests successfully identified measurable differences in shape between each of the three possible pairings of these three spaces. Such findings additionally corroborate the legitimacy of the two-space permutation test.

**5.4. *Summary of findings.*** In summary, the major findings of the simulation study are three-fold. First and foremost, these simulations demonstrate that the proposed omnibus permutation testing procedure successfully identified measurable differences in shape between at least two of the three spaces. Second, these simulations confirm that the post hoc testing component successfully identified measurable differences in shape between any two spaces; such findings corroborate the legitimacy of the two-space permutation testing procedure. Third and finally, these simulations reveal that, for any number of spaces, balanced sampling is required in obtaining the point clouds utilized in the testing procedure.

## 6. Applications to real data sets

We apply our methods to the cardiotocography (CTG) data set that is freely available from the University of California at Irvine Machine Learning Repository at https://archive.ics.uci.edu/ml/datasets/cardiotocography. The CTG data set includes 23 variables for each of 2126 subjects. We apply our methods on a focused subset of four quantitative variables, including fetal heart rate baseline in beats per minute, number of accelerations per second, number of uterine contractions per second,

and number of light decelerations per second. These four quantitative variables are chosen because they are seemingly independent, and we want to consider no more than four such variables. The categorical variable of interest is health status, which has three levels: normal, suspect, and pathologic. The question of interest is whether or not the 4-dimensional space created by the quantitative variables has a measurably different shape across the three health status groups. To answer this question, we use the omnibus permutation testing procedure developed in Section 4.1, measuring shape via 1-dimensional persistent homology. Before this procedure can be performed, however, multiple point clouds from the three health status groups must be obtained via balanced subsampling of the subjects.

Of the 2126 sampled subjects, 1655 are of normal health status, 295 are of suspect health status, and 176 are of pathologic health status. Hence, from the sampled data points we obtain three 4-dimensional point clouds, one consisting of 1655 subjects from the normal health status group, another consisting of 295 subjects from the suspect health status group, and one other consisting of 176 subjects from the pathologic health status group. As our methods require balanced sampling across multiple point clouds from each of the groups, we partitioned, via subsampling, each given point cloud into smaller 4-dimensional point clouds consisting of 44 subjects each. Consequently, we obtained 37 points clouds from the normal health status group, 6 point clouds from the suspect health status group, and 4 point clouds from the pathologic health status group. As neither 1655 nor 295 are divisible by 44, we simply discarded the leftover 27 normal health status subjects and the 31 suspect health status subjects.

The omnibus permutation test was then performed using the persistence diagrams corresponding to the 47 subsampled point clouds. The null hypothesis asserted that there were no measurable differences in shape between the three spaces corresponding to the three health status groups. The resulting approximate permutation test p-value of 0.00005 was based on 100,000 randomly sampled permutations of the 47 point clouds. In particular, for each permutation, the 47 point clouds were shuffled and then labeled such that the first 37 were in the normal health status group, the next 6 were in the suspect health status group, and the last 4 were in the pathologic health status group. Given that the p-value is so small, we reject the null hypothesis and conclude that there are measurable differences in shape between at least two of the three spaces.

To determine the source(s) of the difference, we ultimately performed three post hoc tests, one for each possible pairing of the three health status groups. For each such test, the null hypothesis asserted that there were no measurable differences in shape between the two spaces of the respective health status groups. For the normal and suspect health status groups, the approximate permutation test p-value of 0.00009 was based on 100,000 randomly sampled permutations of the 43 point

clouds. In particular, for each permutation, the 43 point clouds were shuffled and then labeled such that the first 37 were in the normal health status group and the final 6 were in the suspect health status group. For the normal and pathologic health status groups, the approximate permutation test p-value of 0.0060 was based on 100,000 randomly sampled permutations of the 41 point clouds. In particular, for each permutation, the 41 point clouds were shuffled and then labeled such that the first 37 were in the normal health status group and the final 4 were in the pathologic health status group. Finally, for the suspect and pathologic health status groups, the approximate permutation test p-value of 0.3012 was based on 100,000 randomly sampled permutations of the 10 point clouds. In particular, for each permutation, the 10 point clouds were shuffled and then labeled such that the first 6 were in the suspect health status group and the final 4 were in the pathologic health status group. Note that while (exact) permutation test p-values could have straightforwardly been obtained for the post hoc tests involving normal versus pathologic (101,270 possible assignments) and suspect versus pathologic (210 possible assignments), such p-values could not have reasonably been obtained for the post hoc test involving normal versus suspect (6,096,454 possible assignments) or for the omnibus test ($1.087394 \times 10^{12}$ possible assignments); therefore, for the sake of consistency, approximate permutation test p-values were obtained in all instances. Based on these results, there is significant evidence of measurable differences in shape between the spaces corresponding to the normal and suspect health status groups, and between the normal and pathologic health status groups, but insignificant evidence of such differences between the suspect and pathologic health status groups.

## 7. Conclusion

For multiple point clouds obtained from (sub)sampled points of three or more spaces, we propose using an omnibus permutation test on the corresponding persistence diagrams to determine whether statistically significant evidence exists of measurable differences in shape between any of the respective spaces. If such differences do exist, we then propose using a number of post hoc (i.e., two-space) permutation tests to identify the specific pairwise differences. To validate this proposed procedure, we conducted a large-scale simulation study using point clouds obtained from samples of points from three spaces. Various combinations of spaces, sample sizes and measurement errors were considered in the simulation study and for each combination the percentage of p-values below an $\alpha$-level of 0.05 was provided. The results of the simulation study clearly suggest that the procedure works, but additionally reveal that the method is neither scale invariant nor insensitive to imbalanced sample sizes across point clouds. Finally, we applied our omnibus testing procedure to a cardiotocography data set and found statistically significant

evidence of measurable differences in shape between the spaces corresponding to normal, suspect and pathologic health status groups.

While the proposed omnibus testing procedure is applicable in any homological dimension, the simulation study and CTG application presented in this paper focus exclusively on homological dimension 1. Hence, to validate the effectiveness of the method in other homological dimensions, and to assess the consistency of the method across various dimensions, additional simulation studies can be performed.

## Appendix

For readers that are less familiar with simplicial complexes, homology, and persistence diagrams, we include here examples of each for a small accessible example. Consider the set, $D$, of five points in the plane as pictured in Figures 8 and 9. Each point in $D$ is a 0-simplex, each line segment drawn between points is a 1-simplex, and each shaded triangle a 2-simplex. As the parameter $r$ increases beyond $r = 4$ the Vietoris–Rips complex will contain additional 2-simplices, a 3-simplex at $r = 4.9$, and eventually a 4-simplex when $2r$ is equal to the diameter of $D$. Note that the abstract simplicial complex $VR(D, 4.9)$ in Figure 9 cannot be geometrically realized in $\mathbb{R}^2$ since it contains pairs of 2-simplices whose intersection is not a face of either simplex.

The complex $VR(D, 4)$, on the left in Figure 9, is labeled with an ordering assigned to its 0, 1, and 2-simplices: the five 0-simplices, $v_1, v_2, v_3, v_4, v_5$; six 1-simplices $e_1, e_2, e_3, e_4, e_5, e_6$; and one 2-simplex $f_1$.

With respect to this notation, the boundary of a chain is relatively easy to calculate. For example, $\delta_1(e_6 + e_1 + e_2) = v_5 + v_3$ and $\delta_2(f_1) = e_2 + e_3 + e_4$. More precisely, the chain complex of $VR(D, 4)$ is

$$0 \longrightarrow \mathbb{Z}_2 \xrightarrow{\delta_2} (\mathbb{Z}_2)^6 \xrightarrow{\delta_1} (\mathbb{Z}_2)^5 \xrightarrow{\delta_0} 0,$$
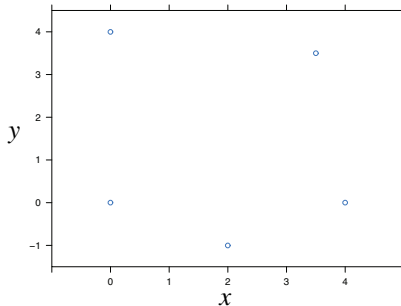


**Figure 8.** Five data points in the plane.
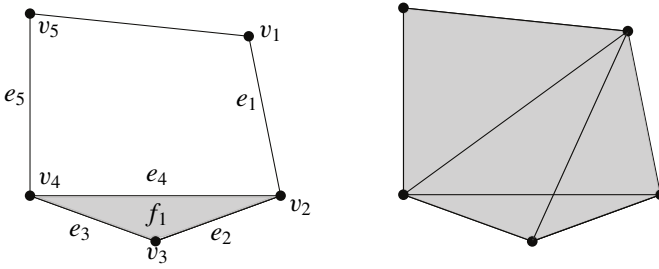
**Figure 9.** Representations of the abstract simplicial complexes VR($D$, 4) and VR($D$, 4.9) for the five point data set $D$.

with boundary maps given in matrix form by

$$\delta_2 = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \delta_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}, \quad \text{and} \quad \delta_0 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Intuitively, the $p$-th homology group measures equivalence classes of $p$-cycles of $K$ that are not "filled" by $(p+1)$-chains. In homological dimension $p = 1$ for the complex VR($D$, 4), an example of a 1-cycle that is not the boundary of a 2-cycle is $e_1 + e_2 + e_3 + e_5 + e_6$. Hence this 1-cycle is in a nonzero equivalence class of $H_1(\mathrm{VR}(D, 4); \mathbb{Z}_2)$. The 1-cycle $e_2 + e_3 + e_4$, however, is the boundary of the 2-cycle $f_1$ (this 1-cycle is "filled" by $f_1$), so this 1-cycle is equivalent to zero in the homology group. Hence, in dimension $p = 1$, the homology of VR($D$, 4) is measuring the circular hole that is seen in the complex.

To complete the homology calculation for the simplicial complex VR($D$, 4), we see that the kernel of $\delta_0$ is $(\mathbb{Z}_2)^5$ and the rank of $\delta_1$ is 4. Thus $H_0(\mathrm{VR}(D, 4); \mathbb{Z}_2) \cong \mathbb{Z}_2$. Similarly, the nullity of $\delta_1$ is 2 and the image of $\delta_2$ is 1-dimensional. This implies that $H_1(\mathrm{VR}(D, 4); \mathbb{Z}_2) \cong \mathbb{Z}_2$. We have $H_2(\mathrm{VR}(D, 4); \mathbb{Z}_2) \cong 0$, since the kernel of $\delta_2$ is 0. Because the complex contains no simplices in higher dimensions, $H_p(\mathrm{VR}(D, 4); \mathbb{Z}_2) = 0$ for all $p > 2$.

The calculation $H_0(\mathrm{VR}(D, 4); \mathbb{Z}_2) = \mathbb{Z}_2$ measures that VR($D$, 4) is a connected complex. The nontrivial group $H_1(\mathrm{VR}(D, 4); \mathbb{Z}_2) = \mathbb{Z}_2$ measures the existence of a 1-dimensional cycle that is not the boundary of a 2-simplex, namely $e_1 + e_2 + e_3 + e_5 + e_6$.

For the complex VR($D$, 4.9), on the right in Figure 9, the homology groups are $H_0(\mathrm{VR}(D, 4.9)) = \mathbb{Z}_2$ and $H_p(\mathrm{VR}(D, 4.9)) = 0$ for all $p \geq 1$. In this example, the
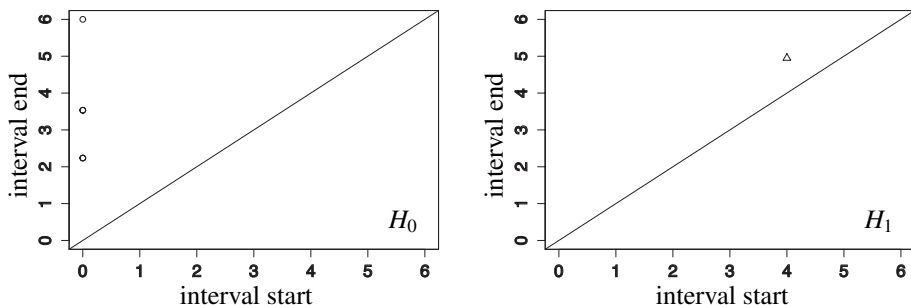
**Figure 10.** The persistence diagrams corresponding to the five-point data set in Figure 8 in the homological dimensions 0 and 1.

first homology group disappeared, or died, as $r$ increases from 4 to 4.9 as a result of the additional 2-simplices that span the 1-cycle $e_1 + e_2 + e_3 + e_5 + e_6$.

The persistence diagrams in Figure 10 display the $H_0$ and $H_1$ persistence diagrams for the five-point data set $D$ first seen in Figure 8. Note that all points in a persistence diagram are plotted above the line $y = x$, as a persistent homology class must be born before it can die.

In homological dimension 1 (the $H_1$ diagram), the small triangle plotted at the point $(4, 4.9)$ indicates that the five-point data set contains a 1-dimensional homology class that is born at radius 4 and dies at radius 4.9. In homological dimension 0 (the $H_0$ diagram), the circles plotted at the points $(0, 2.236)$ and $(0, 3.54)$ represent the connection of data points by 1-simplices at $r = 2.236$ and at $r = 3.54$ resulting in the death of a connected component when it is joined with another connected component by a 1-simplex. For $r > 3.54$ the five points are path connected via 1-simplices; thus this connected complex gives rise to a single 0-dimensional persistent homology class. This single class is plotted at $(0, 6)$ as a result of considering only $r$-values in the range $0 \le r \le 6$.

## References

[Casella and Berger 2002]  G. Casella and R. L. Berger, *Statistical inference*, 2nd ed., Brooks/Cole, Pacific Grove, CA, 2002.

[DeGroot and Schervish 2012]  M. H. DeGroot and M. J. Schervish, *Probability and statistics*, 4th ed., Addison-Wesley, Boston, 2012.

[Edelsbrunner and Harer 2010]  H. Edelsbrunner and J. L. Harer, *Computational topology*, Amer. Math. Soc., Providence, RI, 2010.  MR  Zbl

[Hatcher 2002]  A. Hatcher, *Algebraic topology*, Cambridge Univ. Press, 2002.  MR  Zbl

[Higgins 2004]  J. J. Higgins, *Introduction to modern nonparametric statistics*, Brooks/Cole, Pacific Grove, CA, 2004.

[Kuhn 1955]  H. W. Kuhn, "The Hungarian method for the assignment problem", *Naval Res. Logist. Quart.* **2** (1955), 83–97.  MR  Zbl

[Munkres 1957] J. R. Munkres, "Algorithms for the assignment and transportation problems", *J. Soc. Indust. Appl. Math.* **5**:1 (1957), 32–38. MR Zbl

[Munkres 1984] J. R. Munkres, *Elements of algebraic topology*, Addison-Wesley, Menlo Park, CA, 1984. MR Zbl

[Ramsey and Schafer 2013] F. L. Ramsey and D. W. Schafer, *The statistical sleuth: a course in methods of data analysis*, 3rd ed., Brooks/Cole, Boston, 2013. Zbl

[Robinson and Turner 2013] A. Robinson and K. Turner, "Hypothesis testing for topological data analysis", preprint, 2013. arXiv

[de Silva and Carlsson 2004] V. de Silva and G. Carlsson, "Topological estimation using witness complexes", pp. 157–166 in *SPBG'04 Symposium on Point–Based Graphics* (Zurich, 2004), edited by M. Gross et al., Eurographics Association, Geneva, 2004.

cceric1@lsu.edu                  Department of Mathematics, Louisiana State University,
                                 Baton Rouge, LA 70803, United States

ijohnson@willamette.edu          Department of Mathematics, Willamette University,
                                 Salem, OR 97301, United States

jokiers@live.unc.edu             Department of Mathematics, University of North Carolina
                                 at Chapel Hill, Chapel Hill, NC 27599, United States

mitchell.krock@colorado.edu      Department of Applied Mathematics, University of Colorado,
                                 Boulder, CO 80309, United States

jordan.e.purdy@gmail.com         Department of Mathematics, Willamette University,
                                 Salem, OR 97301, United States

jtorrence@uchicago.edu           Department of Computer Science, University of Chicago,
                                 Chicago, IL 60637, United States

# Merging peg solitaire on graphs

### John Engbers and Ryan Weber

(Communicated by Anant Godbole)

Peg solitaire has recently been generalized to graphs. Here, pegs start on all but one of the vertices in a graph. A move takes pegs on adjacent vertices $x$ and $y$, with $y$ also adjacent to a hole on vertex $z$, and jumps the peg on $x$ over the peg on $y$ to $z$, removing the peg on $y$. The goal of the game is to reduce the number of pegs to one.

We introduce the game *merging peg solitaire on graphs*, where a move takes pegs on vertices $x$ and $z$ (with a hole on $y$) and merges them to a single peg on $y$. When can a configuration on a graph, consisting of pegs on all vertices but one, be reduced to a configuration with only a single peg? We give results for a number of graph classes, including stars, paths, cycles, complete bipartite graphs, and some caterpillars.

## 1. Introduction

Peg solitaire on graphs has recently been introduced as a generalization of peg solitaire on geometric boards [Avis and Deza 2001; Beeler and Hoilman 2011]. Peg solitaire on graphs is played on a simple connected graph $G$ and begins with a starting configuration consisting of pegs in all vertices but one; the remaining vertex is said to have a *hole*. A move involves finding vertices $x$, $y$, and $z$ with $x$ and $y$ adjacent and $y$ and $z$ adjacent with pegs on $x$ and $y$ only, and jumping the peg from $x$ over $y$ and into $z$ (while removing the peg at $y$); see Figure 1.

If there is some starting configuration of pegs and some combination of moves that reduces the number of pegs to one, we say the graph is *solvable*; if the graph is solvable for every starting configuration then we say the graph is *freely solvable*.

Recently, several variations on peg solitaire were introduced. One variant, called fool's solitaire [Beeler and Rodriguez 2012] tries to maximize the number of pegs left in the game when no more moves can be made. A second variant, called reversible peg solitaire [Engbers and Stocker 2015], asks which graphs are solvable if both moves and reverse moves are allowed.
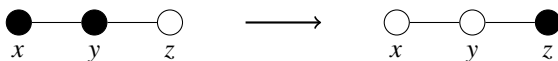
**Figure 1.** A move in peg solitaire on graphs.
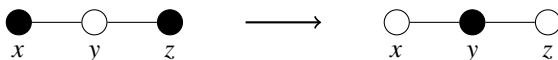


**Figure 2.** A move in merging peg solitaire on graphs.

In this paper, we introduce a new variation on peg solitaire, called *merging peg solitaire on graphs*, by using a different move. We again consider vertices $x$, $y$, and $z$ with $x$ and $y$ adjacent and $y$ and $z$ adjacent. However, now we start with pegs on vertices $x$ and $z$ only, and the new move merges those two pegs to a single peg on $y$; see Figure 2.

For a fixed simple connected graph $G$ and some initial configuration of pegs — occupying all but a single vertex — the goal of the game is to use this move to reduce the number of pegs to one. If this is possible for some initial configuration, we again say that the graph is *solvable*, and if it is possible for any initial configuration we say that the graph is *freely solvable*. The main question that we ask is the following. Given a fixed simple connected graph $G$, is $G$ solvable, and if so, is $G$ freely solvable?

Notice that the merging move is the only other symmetric way of reducing exactly two pegs in a path on 3 vertices, $P_3$, to exactly one peg where each vertex must change from peg to hole or vice versa. In this way, this new game may be viewed as a restricted version of Lights Out on graphs, a game where the *entire* closed neighborhood of a vertex flips all states (here pegs/holes). In this formulation, we are allowed to flip the states of all vertices in a $P_3$ subgraph if the endpoints of the $P_3$ have pegs and the center has a hole. For a survey of Lights Out, see, e.g., [Fleischer and Yu 2013].

The game is also similar to graph rubbling (see, e.g., [Belford and Sieben 2009] for an introduction to graph rubbling) in that the moves allowed are nearly identical, but the end goal of the game is quite different. Indeed, in graph rubbling, a number of pebbles (pegs) are placed on some vertices, and the allowable move removes two pebbles at vertices $v$ and $w$ adjacent to a vertex $u$ while an extra pebble is added at $u$. The goal of graph rubbling is to use the least number of pebbles $m$ so that any vertex is reachable from any pebble distribution of the $m$ pebbles. In addition to the goal of merging peg solitaire on graphs being different, our game also does not allow for multiple pebbles on the vertices (and so, in particular, forces $v \neq w$).

## 2. Preliminary results

In this section we describe some preliminary results for various classes of graphs. As usual, we let $P_n$ and $C_n$ denote the path and cycle on $n$ vertices, respectively. The complete bipartite graph with $V = X \cup Y$, where $|X| = m$ and $|Y| = n$, is denoted $K_{m,n}$; when $m = 1$ we refer to the complete bipartite graph as a *star*. A vertex of degree one is a *pendant* vertex. We begin with several useful lemmas.

**Lemma 2.1.** *Let $G$ be a graph and suppose that the only holes on the vertices of $G$ are on pendant vertices. Then there are no available moves.*

*Proof.* Any move requires two pegs on distinct vertices, both adjacent to the vertex with a hole. $\qquad\square$

The next results follow from Lemma 2.1.

**Lemma 2.2.** *Let $G$ be a graph. If $G$ has any pendant vertices, then $G$ is not freely solvable.*

**Corollary 2.3.** *Let $T$ be a tree. Then $T$ is not freely solvable.*

Next, we show that a star on at least 4 vertices is not solvable.

**Theorem 2.4.** *Fix $n > 2$. The star $K_{1,n}$ is not solvable.*

*Proof.* Let $G = K_{1,n}$. If the hole starts on a pendant vertex, then there are no available moves by Lemma 2.1. If the hole starts on the center, then a single move will leave exactly two holes on two pendant vertices. Again, by Lemma 2.1, there are no more available moves. Since $n > 2$, there are at least two pegs remaining. $\quad\square$

We already know that trees are not freely solvable. For the games of peg solitaire on graphs and reversible peg solitaire on graphs, not all paths are solvable [Beeler and Hoilman 2011; Engbers and Stocker 2015]; in particular, $P_5$ is not solvable in either of those two games. In contrast, for merging peg solitaire on graphs all paths are solvable.

**Theorem 2.5.** *If $n \geq 2$, the path $P_n$ is solvable, and furthermore if an initial configuration can be reduced to a single vertex, then the initial hole must start on a vertex adjacent to a pendant vertex.*

*Proof.* We induct on $n$, with the base case $n = 2$ clear. Let the vertices of the path be labeled $1, \ldots, n$. By Lemma 2.1, the hole cannot start on vertex 1 or on vertex $n$. If the hole starts on vertex 2, then one move creates holes on vertices 1 and 3 only. By considering the vertices $2, \ldots, n$ we have a path on $n - 1$ vertices with a hole second from one end. Therefore we are done by induction.

Suppose the hole is on vertex $i$ with $2 < i < n - 1$. After the first move, there are holes on vertices $i - 1$ and $i + 1$. Suppose next that the pegs on vertices $i$ and $i - 2$ merge to a peg on $i - 1$, leaving a configuration with holes on vertices $i - 2$, $i$, and
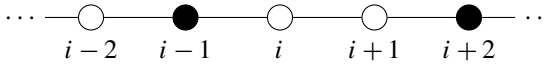
**Figure 3.** The configuration after the first two moves.

$i + 1$; see Figure 3. The only move available is to merge pegs into $i - 2$ and iterate this process, producing a graph with pegs on vertex 2 and on vertices $i + 1, \ldots, n$. By the assumption on $i$, at least two pegs remain.

The other possible second move produces a similar result, and so no set of moves can reduce the path to a configuration with a single peg unless the hole starts on a vertex adjacent to a pendant vertex. □

Part of the proof of Theorem 2.5 will be useful later, but in the following (slightly) generalized form. We start with a definition.

**Definition 2.6.** In a configuration of pegs on a graph $G$, an *empty bridge* is a pair of adjacent degree 2 vertices, joined by a cut-edge, both of which have holes.

**Lemma 2.7.** *Suppose that $G$ is a graph and some configuration of pegs and holes on the graph has an empty bridge and a nonzero number of pegs on either side of the empty bridge. Then $G$ is not solvable from that configuration.*

*Proof.* Suppose that the empty bridge consists of vertices $u$ and $v$. To solve the graph from this configuration, a peg must be moved to either vertex $u$ or vertex $v$, since there are a nonzero number of pegs on either side of the empty bridge. But any move that puts a peg on $u$ requires a prior peg on $v$, and any move that puts a peg on $v$ requires a prior peg on $u$. □

Since any graph containing a spanning solvable subgraph must also be solvable, we have the following result.

**Theorem 2.8.** *Let $n > 2$. The n-cycle $C_n$ is freely solvable.*

The cycle is *freely* solvable since given a hole on any vertex of the cycle we can choose a spanning path so that the hole is adjacent to a pendant vertex of the path.

**Corollary 2.9.** *If $G$ is Hamiltonian, then $G$ is freely solvable.*

Let us consider other graph classes. By Corollary 2.9, complete graphs are freely solvable. The behavior of nonstar complete bipartite graphs is more interesting.

**Theorem 2.10.** *Let $m, n \geq 2$ be integers. If $m - n$ is divisible by 3, then $K_{m,n}$ is freely solvable. If $m - n$ is not divisible by 3, then $K_{m,n}$ is solvable but not freely solvable.*

*Proof.* Notice that any move results in two pegs becoming holes on one partition class of the graph and a single hole becomes a peg on the other partition class. Therefore if there are $p$ pegs in the partition class of size $m$ and $q$ pegs in the

partition class of size $n$, then the quantity $f(p, q) := (p - q) \mod 3$ is preserved by a move. Notice that a configuration with only a single peg has $f(p, q) = 1$ or $f(p, q) = 2$.

This immediately implies several facts. If $f(m, n) = 1$, then a configuration with the hole on a vertex in the partition class of size $m$ cannot be reduced to a configuration with a single peg, and if $f(m, n) = 2$ then a configuration with the hole starting on a vertex in the partition class of size $n$ cannot be reduced to a configuration with a single peg.

Next, notice that given any $m, n \geq 2$ either $f(m - 1, n)$ or $f(m, n - 1)$ is nonzero. So suppose that $m, n \geq 2$ and either $f(m, n) = 0$, $f(m, n) = 1$ and the hole starts on a vertex in the partition class of size $n$, or $f(m, n) = 2$ and the hole starts on a vertex in the partition class of size $m$. We describe a collection of moves that, when iterated, produces a configuration with a single peg. A *partition move* is a sequence of moves that merges pegs from one partition class into the opposite partition class until either all of the holes on the latter partition class have been filled with pegs or the vertices on the former partition class are all holes (with possibly a single peg left, depending on parity). Each partition move decreases the total number of pegs on the vertices. Note that the iteration requires $m, n \geq 2$ so that partition moves can be made back and forth. This process will terminate when there is a single peg remaining (the terminating state can't have a single peg in each partition class by the assumptions on $m$ and $n$). If the initial configuration of pegs satisfies $f(p, q) = 1$ ($f(p, q) = 2$, resp.), then the final peg will be in the partition class of size $m$ ($n$, resp.). □

We also investigate what happens when an edge is added to a star and, more generally, when a matching is added to a star. These graphs were analyzed for peg solitaire on graphs in [Beeler and Hoilman 2012].

**Definition 2.11.** Given fixed nonnegative integers $B$ and $P$, the *windmill variant graph*, denoted $W(P, B)$, is the graph on $P + 2B + 1$ vertices obtained by taking a star $K_{1, P+2B}$ and adding a matching of size $B$ on the pendant vertices of the star. We will label the pendant vertices of $W(P, B)$ by $p_1, \ldots, p_P$ and the pendant vertices of $K_{1, P+2B}$ involved in the matching by $b_1, b_2, \ldots, b_{2B}$ so that $b_{2i-1} b_{2i}$ is an edge of $W(P, B)$ for $i = 1, \ldots, B$.

See Figure 4 for an example of a windmill variant graph. Note that if $B = 0$, then $W(P, 0) = K_{1, P}$ and if $P = 0$ then $W(0, B)$ is the *windmill graph*. The vertex corresponding to the center of $K_{1, P+2B}$ is called the *universal vertex $u$* which is adjacent to $B$ *blades* consisting of two vertices each. We now show that $W(P, B)$ is solvable unless $B = 0$, and $W(0, B)$ is freely solvable. We note that this differs from the results for peg solitaire, where $W(P, B)$ is solvable if and only if $P \leq 2B$
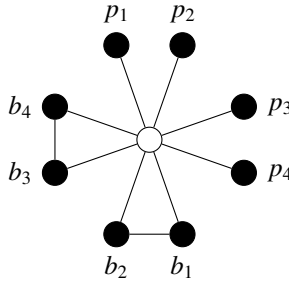
**Figure 4.** The windmill variant $W(4, 2)$.

and freely solvable if and only if $P \leq 2B - 1$ and $(P, B) \neq (0, 2)$ [Beeler and Hoilman 2012, Theorem 2.2].

**Theorem 2.12.** *Let $P$ and $B$ be nonnegative integers and let $W(P, B)$ be a windmill variant graph on at least 2 vertices. If $P = 0$, then $W(0, B)$ is freely solvable. If $P \neq 0$ and $B \geq 1$, then $W(P, B)$ is solvable but not freely solvable.*

*Proof.* Suppose first that $P = 0$ and the hole starts on the center $u$. If $B = 1$, then the result follows. For $B > 1$, we iteratively eliminate the pegs on distinct blades. We first merge the pegs on $b_{2B}$ and $b_1$ to a peg on $u$, and then merge the pegs on $u$ and $b_{2B-1}$ to a peg on $b_{2B}$. If $B = 2$, we merge the pegs on $b_{2B}$ and $b_2$ to $u$ and we're finished. If $B > 2$, we have $B - 2$ full blades and pegs on $b_2$ and $b_{2B}$. We merge $b_2$ and $b_4$ into $u$, and then $u$ and $b_3$ to $b_4$. Doing this last step $B - 2$ times leaves two pegs on distinct blades; we then merge them to $u$.

If $P = 0$ and the hole starts on a blade, say $b_2$, then we merge the pegs on $u$ and $b_1$ to a peg on $b_2$. If $B = 1$ we're done, so suppose $B > 1$. Now ignoring the blade $b_1 b_2$, we have a graph with $B - 1$ blades with the hole on $u$, which we can solve by the previous paragraph and end with the peg on $u$. We then merge the pegs on $u$ and $b_2$ to a peg on $b_1$.

Now suppose that $P \geq 1$. By Lemma 2.1, in this case $W(P, B)$ is not freely solvable. We show that if $B = 1$ and $P \geq 1$, then $W(P, 1)$ is solvable. Since for $B \geq 1$ and $P' = P + 2(B - 1)$, $W(P', 1)$ is a spanning subgraph of $W(P, B)$, this proves the result.

Start with the hole on $b_2$, and merge the pegs on $u$ and $b_2$ to a peg on $b_1$. Then merge the pegs on two pendant vertices to a peg on $u$, and subsequently merge the pegs on $u$ and $b_1$ to a peg on $b_2$. Iteratively merge the pegs on two pendant vertices to a peg on $u$ then merge the peg on $u$ with the peg on the blade to the hole on the blade. This process stops when there are 0 pegs or 1 peg remaining on the pendant vertices. If there are 0 pegs remaining, then we are done. If there is 1 peg remaining, then merge with the peg on the blade to a peg on $u$. $\qquad \square$

## 3. Double stars and caterpillars

Knowing whether or not a given tree is solvable would be extremely helpful in determining whether or not a connected graph is solvable or not; in particular, any connected graph with a solvable spanning tree would necessarily be solvable. Since stars are not solvable but paths are solvable, a natural first step in classifying the solvable trees is to describe when a caterpillar is solvable.

**Definition 3.1.** Let $n \geq 1$ be given, and let $p_1, \ldots, p_n$ be nonnegative integers. A *caterpillar* on $n + p_1 + \cdots + p_n$ vertices consists of a path on $n$ vertices so that the $i$-th vertex on the path has $p_i$ pendant vertices attached to it. We will denote this caterpillar by $P_n(p_1, \ldots, p_n)$.

See Figure 5 for an example of a caterpillar. Note also that $P_1(n)$ is isomorphic to the star $K_{1,n}$ and $P_n(0, \ldots, 0)$ is isomorphic to the path $P_n$.

We will prove that a large family of caterpillars are solvable and also fully classify the solvability of some special types of caterpillars. To do so, we start with a special type of caterpillar. A *double star* is a caterpillar of the form $P_2(m, n)$ — see Figure 6 — and the two vertices from the path are its *centers*.

**Theorem 3.2.** *Let* $m, n \geq 1$. *If* $|m - n| \leq 1$, *then the double star* $P_2(m, n)$ *is solvable. If* $|m - n| > 1$ *then the double star* $P_2(m, n)$ *is not solvable.*

*Also, if* $m = n$ *and the hole starts on center vertex* $u$, *then the final peg is on* $v$.

We note that in peg solitaire $P_2(m, n)$ (with $m \geq n$) is solvable if and only if $m \leq n + 1$ and $n \neq 1$ and freely solvable if and only if $m = n$ and $n \neq 1$ [Beeler and Hoilman 2012, Theorem 3.1].

*Proof.* We must start with the hole on one of the two center vertices $u$ or $v$; without loss of generality assume the hole starts on $u$, where $u$ has $m$ pendant vertices. If
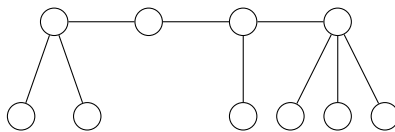


**Figure 5.** The caterpillar $P_4(2, 0, 1, 3)$.
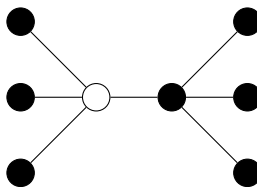


**Figure 6.** The graph $P_2(3, 3)$.

the pegs on two pendant vertices are merged to a peg on $u$, then by Lemma 2.1 no more moves are possible, and since there is a peg on $v$ this move will never produce a graph with a single peg remaining. Therefore the only move that allows for future moves is to merge the peg on $v$ and the peg on a pendant vertex of $u$ to a peg on $u$. We then repeat by merging the peg on $u$ and the peg on a pendant vertex of $v$ to a peg on $v$. Continuing in this way, we remove the same number of pegs from the pendant vertices of $u$ and $v$, so if $m = n$ this process terminates with a peg only on $v$.

If $m + 1 = n$, then after the first move we have the same number of pegs on the pendant vertices of $u$ and $v$ with the hole on $v$, and so the double star is solvable by the previous argument. If $m = n + 1$, then we start with the hole on $v$ and the previous argument shows that the graph is solvable.

Now suppose that $|m - n| \geq 2$. Notice that each move that allows for future move alternates reducing the number of pegs on pendant vertices of $u$ by 1 and the number of pegs on pendant vertices of $v$ by 1; without loss of generality assume the hole starts on $u$. If $m < n$, then removing the last peg on a pendant vertex of $u$ leaves pegs on $u$ and $n - m + 1$ pendant vertices of $v$. Then the final remaining move merges two of these pegs to a peg on $v$, and no further moves are possible. If $m > n$ and the hole starts on $u$, then removing the last peg on a pendant vertex of $v$ leaves pegs on $v$ and $n - m$ pendant vertices of $u$. Merging two of these pegs leaves $n - m$ pegs remaining with no further moves available.                    $\square$

Next, we see what happens to solvability when we subdivide the edge between the center vertices of a double star.

**Definition 3.3.** Fix an integer $k \geq 3$ and positive integers $m$ and $n$. A *path-$k$ double star* is the graph $P_k(m, 0, \ldots, 0, n)$.

See Figure 7 for an example of a path-3 double star. Recall that by Corollary 2.3 no tree is freely solvable. In what follows, we fully classify the solvability of path-$k$ double stars. We are unaware of any results in peg solitaire for path-$k$ double stars when $k > 2$.

**Theorem 3.4.** *Fix nonnegative integers $m$ and $n$ and let $P_3(m, 0, n)$ be a path-3 double star. Then $P_3(m, 0, n)$ is solvable if $\lfloor \frac{1}{2}(m - 1) \rfloor \leq n \leq 2m + 2$ and is not solvable otherwise.*

*Proof.* As before we cannot start with a hole on a pendant vertex; assume that the graph has nonpendant vertices $u$, $w$, and $v$ with $u$ having $m$ pendants attached to it and $v$ having $n$ pendants attached to it.

Suppose first that the hole starts on $u$. Merging two pendant pegs results in no further moves, so the only move is to merge pegs on a pendant vertex and $w$ to a peg on $u$, leaving one fewer peg on the pendants of $u$ and a hole on $w$. The initial

**Figure 7.** The graph $P_3(3, 0, 3)$.

move when the hole starts on $v$ is similar. It remains to analyze the situation where
a hole starts on $w$ only, and we note that if $P_3(a, 0, b)$ is solvable with initial hole
on $w$, then $P_3(a + 1, 0, b)$ is solvable with initial hole on $v$ and $P_3(a, 0, b + 1)$ is
solvable with initial hole on $w$.

Now, with a hole on $w$, the only available move is to merge the pegs on $u$ and $v$
to a peg on $w$, creating holes on $u$ and $v$. Focusing on the hole at $u$, either two pegs
on pendant vertices of $u$ can merge to a peg on $u$ or a peg on a pendant vertex of $u$
and the peg on $w$ can merge to a peg on $u$. Two similar moves are possible at $v$,
but these moves cannot be made independently. If two pendant pegs merge to $u$
and two pendant pegs merge to $v$, then no further moves can be made. So suppose
that $w$ and a pendant peg merge to $u$. Then the only available move merges two
pegs on pendants of $v$ to a peg on $v$. A similar result follows from merging $w$ and
a pendant peg to $v$.

This shows that if the holes are on $w$ and pendant vertices only, then the only sets
of moves that allow for future moves result in the removal of two pegs on pendant
vertices from $u$ ($v$, resp.), the removal of one peg on a pendant vertex from $v$ ($u$,
resp.), and a configuration where the only holes are on $w$ and pendant vertices again.

Next, it is useful to see which graphs $P_3(m, 0, n)$ are solvable with initial hole on
$w$ for small values of $m$ and $n$. Since we can effectively reduce one of $m$ and $n$ by
2 and the other by 1 (by viewing holes on pendant vertices as deleted vertices), we
only need to check the solvability of $P_3(m, 0, 0)$, $P_3(0, 0, n)$, and $P_3(1, 0, 1)$ with
initial hole on $w$. The only graphs that are solvable are $P_3(0, 0, 0)$, $P_3(1, 0, 0)$, and
$P_3(0, 0, 1)$, as $P_3(1, 0, 1)$ is not solvable by Theorem 2.5 and $P_3(m, 0, 0)$ for $m > 1$
is, after one move, essentially a star with $m + 1$ pendants and so is not solvable by
Theorem 2.4.

Suppose that we:

(1) complete $x$ sets of moves that remove 2 pegs on pendant vertices of $u$ and 1
     peg on a pendant vertex of $v$;

(2) complete $y$ sets of moves that remove 1 peg on a pendant vertex of $u$ and 2
     pegs on pendant vertices of $v$; and

(3) end with $P_3(0, 0, 0)$, $P_3(1, 0, 0)$, or $P_3(0, 0, 1)$ and a hole on $w$.

If the initial hole started on $w$, then $2x + y$ pegs were removed from the pendant vertices of $u$ and $x + 2y$ pegs were removed from the pendant vertices of $v$. If the initial hole started on $u$ ($v$, resp.) then $2x + y + 1$ ($2x + y$, resp.) pegs were removed from the pendant vertices of $u$ and $x + 2y$ ($x + 2y + 1$, resp.) pegs were removed from the pendant vertices of $v$. We analyze the possible values of $m$ and $n$ that are solvable by considering both where the hole starts and also which of $P_3(0, 0, 0)$, $P_3(1, 0, 0)$, or $P_3(0, 0, 1)$ remains.

For these fixed values of $x$ and $y$, if $P_3(0, 0, 0)$ remains, then we have $(m, n) = (2x + y, x + 2y)$, $(2x + y + 1, x + 2y)$, or $(2x + y, x + 2y + 1)$. If $P_3(0, 0, 1)$ remains, then $(m, n) = (2x + y, x + 2y + 1)$, $(2x + y + 1, x + 2y + 1)$, or $(2x + y, x + 2y + 2)$. If $P_3(1, 0, 0)$ remains, then $(m, n) = (2x + y + 1, x + 2y)$, $(2x + y + 2, x + 2y)$, or $(2x + y + 1, x + 2y + 1)$. By the above arguments, these are the only solvable values for $m$ and $n$.

Now, suppose $m > 0$ is fixed. What values of $n$ (as a function of $m$) are solvable? For $n$ to be maximized, we take $m = 2x + y$ and $n = x + 2y + 2$ where $x = 0$ and $y = m$. Then we have $n = 2m + 2$; therefore $n \leq 2m + 2$. Symmetrically we have $m \leq 2n + 2$, so $\lfloor \frac{1}{2}(m - 1) \rfloor \leq n$. To show that all values of $n$ in that range are possible, note that for a given $m$ there are values of $x$ and $y$ with $2x + y = m$. But for each $x$ and $y$ pair, we have, as possible values for $n$, $x + 2y$, $x + 2y + 1$, and $x + 2y + 2$. This shows that all values $\lfloor \frac{1}{2}(m + 1) \rfloor \leq n \leq 2m + 2$ are possible. But we can have $n = \lfloor \frac{1}{2}(m - 1) \rfloor$ by taking $m = 2x + y + 2$ or $m = 2x + y + 1$ (depending on parity) and $n = x + 2y$ where $x = \lfloor \frac{1}{2}(m - 1) \rfloor$ and $y = 0$.          □

**Theorem 3.5.** *Fix nonnegative integers $m$ and $n$. Then the graph $P_4(m, 0, 0, n)$ is solvable if*:

(1) *$m = n$, or*

(2) *$m$ is even and $n = m + 1, m + 2, m + 3$, or $m + 4$, or*

(3) *$n$ is even and $m = n + 1, n + 2, n + 3$, or $n + 4$,*

*and is not solvable otherwise.*

*Proof.* As before we cannot start with a hole on a pendant vertex; assume that the graph has nonpendant vertices $u$, $w_1$, $w_2$ and $v$ with $m$ pendant vertices on $u$, $n$ pendant vertices on $v$, and $u$ adjacent to $w_1$; see Figure 8.

Suppose first that the hole starts on $u$. Merging two pegs on pendant vertices results in no further possible moves, so the only move is to merge a peg on a pendant vertex and the peg on $w_1$ to a peg on $u$, leaving one fewer peg on the pendants of $u$ and a hole on $w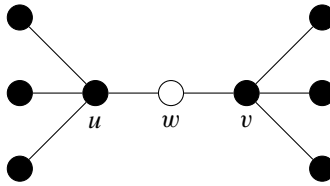_1$. When the hole starts on $v$ the analysis is similar. So we again consider the initial hole starting on $w_1$ (with a similar analysis of the hole at $w_2$ following immediately), and have that if $P_4(a, 0, 0, b)$ is solvable with initial hole
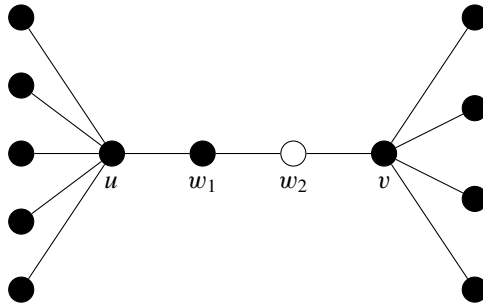
**Figure 8.** The graph $P_4(5, 0, 0, 4)$.

on $w_1$ ($w_2$, resp.), then $P_4(a + 1, 0, 0, b)$ ($P_4(a, 0, 0, b + 1)$, resp.) is solvable with initial hole on $u$ ($v$, resp.).

Suppose then that the only hole on a nonpendant vertex is on $w_1$. The available move is to merge pegs on $w_2$ and $u$ to a peg on $w_1$. If we then merge pegs on a pendant of $u$ and $w_1$ to a peg on $u$, we create an empty bridge which is not solvable by Lemma 2.7.

If we first merge the pegs on $w_1$ and $v$ to a peg on $w_2$, then we have holes on $u$, $w_1$, and $v$. To avoid creating an empty bridge, we must merge two pegs on pendant vertices of $u$ to a peg on $u$ and merge two pegs on pendant vertices of $v$ to a peg on $v$. This produces a hole on $w_1$ and removes two pegs on the pendant vertices of $u$ and two pegs from the pendant vertices of $v$.

Note that if we instead first merge two pegs on pendant vertices of $u$ to a peg on $u$, a similar analysis produces the same loss of two pegs from both sets of pendant vertices with a hole on $w_1$.

Again, we now analyze the small cases of $m$ and $n$; we see that $P_4(0, 0, 0, 0)$ is solvable with the hole on $w_1$ or $w_2$; $P_4(1, 0, 0, 0)$, $P_4(0, 0, 0, 2)$, and $P(0, 0, 0, 3)$ are solvable with the hole on $w_2$, and $P_4(0, 0, 0, 1)$, $P_4(2, 0, 0, 0)$, and $P_4(3, 0, 0, 0)$ are solvable with the hole on $w_1$, and by inspection no other graph $P_4(m, 0, 0, n)$ is solvable when one of $m$ or $n$ is 0 or 1 and the hole is on $w_1$ or $w_2$.

We now put all of this together. The graphs that are solvable with initial hole on $w_1$ are $P_4(2x, 0, 0, 2x)$, $P_4(2x, 0, 0, 2x + 1)$, $P_4(2x + 2, 0, 0, 2x)$ and $P_4(2x+3, 0, 0, 2x)$; the solvable graphs with initial hole on $w_2$ are $P_4(2x, 0, 0, 2x)$, $P_4(2x + 1, 0, 0, 2x)$, $P_4(2x, 0, 0, 2x + 2)$ and $P_4(2x, 0, 0, 2x + 3)$. This then shows that the graphs that are solvable with initial hole on $u$ are $P_4(2x + 1, 0, 0, 2x)$, $P_4(2x+1, 0, 0, 2x+1)$, $P_4(2x+3, 0, 0, 2x)$, and $P_4(2x+4, 0, 0, 2x)$; the graphs that are solvable with initial hole on $v$ are $P_4(2x, 0, 0, 2x + 1)$, $P_4(2x + 1, 0, 0, 2x + 1)$, $P_4(2x, 0, 0, 2x + 3)$ and $P_4(2x, 0, 0, 2x + 4)$. This gives the result. $\square$

We next show that the remaining nontrivial path-$k$ double stars are not solvable for positive integers $m$ and $n$.

**Theorem 3.6.** *Fix positive integers $m$, $n$ (where both $m$ and $n$ are not 1) and fix an integer $k \geq 5$. Then the graph $P_k(m, 0, \ldots, 0, n)$ is not solvable.*

*Proof.* Label the vertices of the path (in order) $u = w_0, w_1, w_2, \ldots, w_{k-2}$, and $v = w_{k-1}$. Assume that $u$ has $m$ pendants and $v$ has $n$ pendants.

If the hole starts on $w_i$ for some $i \in \{2, \ldots, k - 3\}$, then any two possible consecutive moves produces an empty bridge and thus a configuration that is not solvable by Lemma 2.7.

If the hole starts on $w_1$, then the first move produces holes on $u$ and $w_2$. If the next move merges two pegs on pendant vertices of $u$ to a peg on $u$, then we have a configuration with holes only on pendant vertices of $u$ and on $w_2$, which as above is not solvable. If the next move instead merges a peg on a pendant of $u$ with the peg on $w_1$, then we have a configuration with an empty bridge on vertices $w_1$ and $w_2$ and so the graph again is not solvable.

Lastly, suppose that the hole starts on $u$. Then the only move that allows for future moves merges the pegs on a pendant vertex of $u$ and $w_1$ to a peg on $u$. But the next move must merge the pegs on $u$ and $w_2$ to a peg on $w_1$.

From this configuration, if we merge two pegs on pendants of $u$ to $u$, then we are left with a configuration with holes only on pendant vertices of $u$ and $w_2$, which as above is not solvable. So we must merge the pegs on $w_1$ and $w_3$ to a peg on $w_2$.

If we then merge two pegs from the pendants of $u$ to $u$, then any subsequent move produces a configuration with an empty bridge and so the graph is not solvable. So the only other possible move is to merge the pegs on $w_2$ and $w_4$ to a peg on $w_3$. Now, if $m > 1$ we have a configuration with an empty bridge on vertices $w_1$ and $w_2$ and so is not solvable. If $m = 1$, then we can iterate this move through the path until finally we merge pegs on $w_{k-3}$ and $w_{k-1}$ to a peg on $w_{k-2}$, which leaves pegs on $w_{k-2}$ and the $n$ (where $n > 1$ as $m = 1$) pendant vertices of $v = w_{k-1}$. But the only possible move now merges two pegs from the neighbors of $v$ to $v$; since there are at least $n + 2$ pegs on the neighbors of $v$, this leaves at least two pegs remaining and no further moves. $\square$

We now provide a large class of caterpillars that are solvable by combining the double star and the path. We are unaware of any results in peg solitaire for this class of caterpillars.

**Theorem 3.7.** *Let $t_1, t_2, \ldots, t_{n-1}$ be nonnegative integers where $p_1 = t_1$, $p_n = t_{n-1}$, and $p_i = t_i + t_{i-1}$ for $2 \leq i \leq n - 1$. Then the caterpillar $P_n(p_1, p_2, \ldots, p_n)$ is solvable.*

We'll first provide the proof, and then give two specific examples of caterpillars that satisfy the conditions for $p_i$ in Theorem 3.7. We note that this theorem can also incorporate solvable path-$k$ double stars, but for reading ease we state this theorem
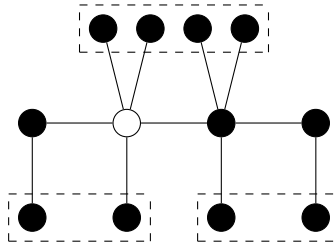
**Figure 9.** The solvable caterpillar $P_4(1, 3, 3, 1)$. Here $t_1 = 1$, $t_2 = 2$, and $t_3 = 1$.

and proof without adding path-3 double stars or path-4 double stars as intermediate steps. We leave the details of these changes to the reader.

*Proof.* Let $t_1, \ldots, t_{n-1}$ be any nonnegative integers, $p_1 = t_1$, $p_n = t_{n-1}$, and for $2 \leq i \leq n-1$ let $p_i = t_i + t_{i-1}$. We need to show that $P_n(p_1, \ldots, p_n)$ is solvable. Start with the hole on vertex 2 of the path, i.e., the vertex with $p_2$ pendant vertices. Then focus on the double star that has as its two centers the first two vertices of the path. By Theorem 3.2 we can eliminate pegs on $t_1$ pendant vertices from vertex 1 and vertex 2 in the path, leaving the hole on vertex 2. Then we merge pegs from vertex 1 and 3 (in the path) to vertex 2. We then focus on the double star that has as its two centers vertex 2 and vertex 3 in the path, noting that vertex 2 has a peg and vertex 3 has a hole. Again, by Theorem 3.2 we eliminate $t_2$ pendant vertices from each, leaving a peg on vertex 2 and a hole on vertex 3. Then we merge the pegs from vertex 2 and vertex 4 to vertex 3. We iteratively continue until we reach vertex $n-1$ and vertex $n$; eliminating $t_{n-1}$ vertices from each leaves a peg on vertex $n-1$ and a hole on vertex $n$. By construction, all pendant vertices have holes, and there is only one peg left on the path. This means that the caterpillar $P_n(p_1, \ldots, p_n)$ is solvable. $\square$

Notice that by solving for each $t_i$ we can find equivalent conditions on the values $p_i$: for each $i \in [1, n-1]$, $\sum_{j=1}^{i}(-1)^{i-j}p_j$ is nonnegative, and also $p_n = \sum_{j=1}^{n-1}(-1)^{i-j}p_j$.

Several interesting sequences that satisfy this condition include setting $p_i = \binom{n}{i}$ (see, e.g., Figure 9) and, for $n$ even, letting $p_i = c$ for some nonnegative integer $c$.

## 4. Related questions and future work

We end our discussion by giving several open problems that can serve as a basis for future investigations. The main open question is to classify all simple connected graphs according to whether they are freely solvable, solvable, or not solvable. A helpful step would be to classify all trees according to whether they are solvable or not. While this might prove difficult, even determining a nice characterization of solvable caterpillars would be interesting. Another possible direction toward

the main open question would be to determine which trees of a fixed diameter are solvable (see, e.g., [Walvoort 2013] for results related to peg solitaire on graphs with fixed diameter).

Another interesting question is the following. Let $G_{n,k}$ denote the set of all simple connected graphs on $n$ vertices with $k$ edges. Note that the only graph in $G_{n,n(n-1)/2}$ is solvable, while the star shows that not every graph in $G_{n,n-1}$ is solvable. For fixed $n$, what is the minimum value of $k$ so that every graph in $G_{n,k}$ is solvable?

Suppose that we wanted to leave the *maximum* number of pegs left so that no further moves can be made; i.e., we wanted to play *merging fool's solitaire on graphs* (for results for fool's solitaire on graphs, see, e.g., [Beeler and Rodriguez 2012; Loeb and Wise 2015]). For a given graph $G$, determine the maximum number of pegs that can be left when playing merging fool's solitaire on graphs.

## References

[Avis and Deza 2001] D. Avis and A. Deza, "On the solitaire cone and its relationship to multi-commodity flows", *Math. Program.* **90**:1 (2001), 27–57. MR Zbl

[Beeler and Hoilman 2011] R. A. Beeler and D. P. Hoilman, "Peg solitaire on graphs", *Discrete Math.* **311**:20 (2011), 2198–2202. MR Zbl

[Beeler and Hoilman 2012] R. A. Beeler and D. P. Hoilman, "Peg solitaire on the windmill and the double star graphs", *Australas. J. Combin.* **53** (2012), 127–134. MR Zbl

[Beeler and Rodriguez 2012] R. A. Beeler and T. K. Rodriguez, "Fool's solitaire on graphs", *Involve* **5**:4 (2012), 473–480. MR Zbl

[Belford and Sieben 2009] C. Belford and N. Sieben, "Rubbling and optimal rubbling of graphs", *Discrete Math.* **309**:10 (2009), 3436–3446. MR Zbl

[Engbers and Stocker 2015] J. Engbers and C. Stocker, "Reversible peg solitaire on graphs", *Discrete Math.* **338**:11 (2015), 2014–2019. MR Zbl

[Fleischer and Yu 2013] R. Fleischer and J. Yu, "A survey of the game 'Lights out!'", pp. 176–198 in *Space-efficient data structures, streams, and algorithms*, edited by A. Brodnik et al., Lecture Notes in Comput. Sci. **8066**, Springer, 2013. MR Zbl

[Loeb and Wise 2015] S. Loeb and J. Wise, "Fool's solitaire on joins and Cartesian products of graphs", *Discrete Math.* **338**:3 (2015), 66–71. MR Zbl

[Walvoort 2013] C. Walvoort, *Peg solitaire on trees with diameter four*, master's thesis, East Tennessee State University, 2013, available at http://dc.etsu.edu/etd/1113.

john.engbers@marquette.edu     *Department of Mathematics, Statistics and Computer Science, Marquette University, Milwaukee, WI 53201, United States*

rweber2006@aol.com     *Department of Mathematics, Statistics and Computer Science, Marquette University, Milwaukee, WI 53201, United States*

# Labeling crossed prisms with
# a condition at distance two

### Matthew Beaudouin-Lafon, Serena Chen, Nathaniel Karst, Jessica Oehrlein and Denise Sakai Troxell

(Communicated by Jerrold Griggs)

An *L(2,1)-labeling* of a graph is an assignment of nonnegative integers to its vertices such that adjacent vertices are assigned labels at least two apart, and vertices at distance two are assigned labels at least one apart. The $\lambda$-*number* of a graph is the minimum span of labels over all its L(2,1)-labelings. A *generalized Petersen graph* (GPG) of order $n$ consists of two disjoint cycles on $n$ vertices, called the *inner* and *outer cycles*, respectively, together with a perfect matching in which each matching edge connects a vertex in the inner cycle to a vertex in the outer cycle. A *prism* of order $n \geq 3$ is a GPG that is isomorphic to the Cartesian product of a path on two vertices and a cycle on $n$ vertices. A *crossed prism* is a GPG obtained from a prism by crossing two of its matching edges; that is, swapping the two inner cycle vertices on these edges. We show that the $\lambda$-number of a crossed prism is 5, 6, or 7 and provide complete characterizations of crossed prisms attaining each one of these $\lambda$-numbers.

## 1. Introduction

The labelings of graphs with a condition at distance two, also known as L(2,1)-labelings, have provided a fertile area of research for about a quarter of a century since their introduction in [Griggs and Yeh 1992]. These labelings were first used to model simplified instances of the channel assignment problem [Hale 1980] where geographically close transmitters in a communications network must receive frequency channels that are sufficiently far apart to avoid signal interference. The scholarly works on L(2,1)-labelings and their variations are numerous and touch upon a wide range of applied as well as purely theoretical aspects of such labelings. Notably, optimization questions concerning the minimum span of labels required by different types of graphs have consistently attracted a great deal of interest.

An L(2,1)-labeling of a graph $G$, or $k$-*labeling* for short, is a function $f : V(G) \to \{0, 1, \ldots, k\}$ such that $|f(u) - f(v)| \geq 2$ if $u$ and $v$ are adjacent vertices, and $|f(u) - f(v)| \geq 1$ if $u$ and $v$ are at distance 2. The minimum $k$ so that $G$ has a $k$-labeling is called the $\lambda$-*number* of $G$ and is denoted by $\lambda(G)$. Arguably, the appeal of this number has its roots in the long-standing conjecture stating that $\lambda(G) \leq \Delta^2(G)$ for $\Delta(G) \geq 2$, where $\Delta(G)$ denotes the maximum degree of $G$ [Griggs and Yeh 1992]. This conjecture, which is sometimes referred to as the $\Delta^2$-*conjecture*, holds for very large graphs (with $\Delta(G)$ larger than approximately $10^{69}$ [Havet et al. 2012]), for sufficiently small graphs (with at most $(\lfloor \Delta(G)/2 \rfloor + 1)(\Delta^2(G) - \Delta(G) + 1) - 1$ vertices [Franks 2015]), and for several particular classes of graphs. In addition, it has been possible to determine tighter bounds and even exact $\lambda$-numbers within some of these classes through interesting, nontrivial techniques, contributing to the incremental progress toward settling the $\Delta^2$-conjecture. An extensive annotated bibliography of related articles can be found in [Calamoneri 2011] and in its 2014 updated online version.

Determining exact $\lambda$-numbers can be a complex task even when considering seemingly basic graphs, such as the following generalizations of the classic Petersen graph (shown on the right of Figure 1 together with a 9-labeling).

**Definition 1.1.** A *generalized Petersen graph* (GPG) of order $n \geq 3$ consists of two disjoint cycles, called *outer* and *inner cycles*, so that each vertex on the outer (resp., inner) cycle is adjacent to exactly one vertex on the inner (resp., outer) cycle. More formally, a GPG has vertices $\{v_0, v_1, \ldots, v_{n-1}\} \cup \{w_0, w_1, \ldots, w_{n-1}\}$ with edges $\{v_i, v_{i+1}\}$ and $\{w_i, w_{i+1}\}$ for all $i = 0, 1, \ldots, n-1$, where subscript addition is taken modulo $n$, and each $v_i$ (resp., $w_i$), $i = 0, 1, \ldots, n-1$ is adjacent to exactly one $w_j$ (resp., $v_j$) for some $0 \leq j \leq n-1$. The cycle on vertices $\{v_0, v_1, \ldots, v_{n-1}\}$ (resp., $\{w_0, w_1, \ldots, w_{n-1}\}$) is the outer (resp., inner) cycle.

Observe that if $G$ is a GPG of order $n \geq 3$, then $G$ is 3-regular, so the $\Delta^2$-conjecture states that $\lambda(G) \leq 9$. This upper bound is tight if $G$ is the Petersen graph since it has diameter 2 and a 9-labeling [Griggs and Yeh 1992]. In contrast, if $G$ is anything other than the Petersen graph, then $\lambda(G) \geq 5$, $\lambda(G) \leq 7$ if $n \leq 6$, and $\lambda(G) \leq 8$ if $n \geq 7$ [Georges and Mauro 2002]. Therefore, the GPGs satisfy the $\Delta^2$-conjecture. As no GPG with $\lambda$-number exactly 8 is known, it has been conjectured that if $G$ is a GPG of order $n \geq 7$, then $\lambda(G) \leq 7$. This *GPG conjecture* has remained open since 2002 but has been verified for all GPGs of orders between 7 and 12 for which exact $\lambda$-numbers were completely determined [Adams et al. 2006; 2007; 2012; Huang et al. 2012]. In an attempt to expand the list of graphs satisfying the GPG conjecture, some articles have focused on the exact $\lambda$-numbers of infinite subclasses of GPGs that exhibit certain symmetric features. For instance, a *prism* (resp., an *n-star* for odd $n$) is a GPG of order $n \geq 3$ wherein the edges between
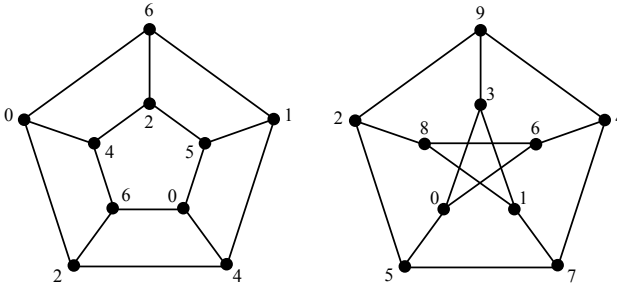
**Figure 1.** The prism of order 5 and the 5-star (Petersen graph) with respective L(2, 1)-labelings.

vertices on the outer and inner cycles are precisely $\{v_i, w_i\}$, $i = 0, 1, \ldots, n - 1$ (resp., $\{v_{(n-1)i/2}, w_i\}$ for $i = 0, 1, \ldots, n - 1$ where subscripts are taken modulo $n$), and the notation is as introduced in Definition 1.1. The prism of order 5 and the 5-star with respective L(2,1)-labelings are shown in Figure 1.

The $\lambda$-numbers of prisms have been completely determined in [Georges and Mauro 2002; Jha et al. 2000; Klavžar and Vesel 2003; Kuo and Yan 2004], and of $n$-stars in [Adams et al. 2007] using nontrivial techniques. Key to some of these were ingenious connections between the regularity and symmetry of these graphs used in [Georges and Mauro 2003; Adams et al. 2007] that would force impossible configurations of labels within 5-labelings for certain values of $n$. We were curious to see if the same strategies could be extended to other subclasses of GPGs where this symmetry would be slightly disturbed. This motivated our focus on GPGs obtained from prisms by "crossing" two edges connecting the outer cycle to the inner cycle:

**Definition 1.2.** Let $n$ and $d$ be integers so that $n \geq 3$ and $1 \leq d \leq n/2$. The *crossed prism XPr(n, d)* is a prism of order $n$ where the edges $\{v_0, w_0\}$ and $\{v_d, w_d\}$ are replaced by the crossed edges $\{v_0, w_d\}$ and $\{v_d, w_0\}$, with the notation as introduced in the definition of prisms. The *cross X(d)* is the graph isomorphic to the subgraph of $XPr(n, d)$ induced by the vertices $\{v_0, v_1, \ldots, v_d\} \cup \{w_0, w_1, \ldots, w_d\}$.

Figure 2 shows the crossed prism $XPr(5, i)$ with the cross $X(i)$ within the dashed oval for $i = 1, 2$, respectively.

It will be helpful to visualize the crossed prism $XPr(n, d)$ as copies of the two crosses $X(d)$ and $X(n - d)$ sharing the same crossed edges but otherwise disjoint. To illustrate, Figure 3 shows a 3-dimensional cylindrical representation of $XPr(9, 4)$ on the left and the crosses $X(4)$ and $X(5)$ on the top and bottom right, respectively (crossed edges in bold to facilitate their visualization within the graphs).

Let $f$ be an L(2,1)-labeling of $XPr(n, d)$. It will be often convenient to provide $f$ as a 2-by-$n$ matrix $A(n, d)$ where the entry on the $i$-th row, $j$-th column will be
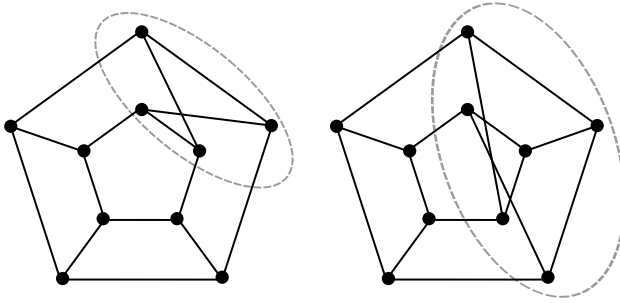
**Figure 2.** The crossed prism $XPr(5, i)$ with the cross $X(i)$ within the dashed oval for $i = 1, 2$, respectively.

the label $f(v_j)$ if $i = 0$, and $f(w_j)$ if $i = 1$, for $j = 0, 1, \ldots, n - 1$. Notice that the matrix $A(d)$ given by the first $d + 1$ columns of $A(n, d)$ is an L(2,1)-labeling of the cross $X(d)$. Similarly, the matrix $A(n - d)$ given by the last $n - d$ columns followed by the first column of $A(n, d)$ is an L(2,1)-labeling of the cross $X(n - d)$. These conventions are illustrated in Figure 4 with 6-labelings of $XPr(9, 4)$, $X(4)$, and $X(5)$ of Figure 3 given by the matrices $A(9, 4)$, $A(4)$, and $A(5)$, respectively.

The strategies used to find the $\lambda$-numbers of prisms leveraged the symmetries of these graphs, and even the minor breaks in symmetry introduced in crossed prisms prohibit the simple extension of these proof techniques into this new context. Nevertheless, we were able to use certain properties of crosses to determine the $\lambda$-numbers of all crossed prisms. In Section 2, we find the exact $\lambda$-number of
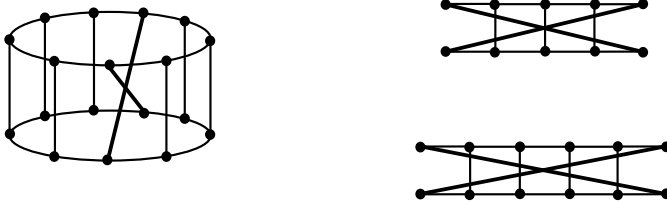


**Figure 3.** The crossed prism $XPr(9, 4)$, left, and the crosses $X(4)$, top right, and $X(5)$, bottom right.

$$A(9, 4) = \begin{bmatrix} 3 & 6 & 0 & 3 & 6 & 2 & 0 & 4 & 1 \\ 0 & 4 & 2 & 5 & 0 & 4 & 6 & 2 & 5 \end{bmatrix}$$

$$A(4) = \begin{bmatrix} 3 & 6 & 0 & 3 & 6 \\ 0 & 4 & 2 & 5 & 0 \end{bmatrix} \qquad A(5) = \begin{bmatrix} 6 & 2 & 0 & 4 & 1 & 3 \\ 0 & 4 & 6 & 2 & 5 & 0 \end{bmatrix}$$

**Figure 4.** The 6-labelings $A(9, 4)$, $A(4)$, and $A(5)$ of $XPr(9, 4)$, $X(4)$, and $X(5)$, respectively.

$X(d)$ for all $d \geq 1$, as well as exhibit all possible 5-labelings when $d \geq 2$ using an auxiliary directed graph where the vertices are particular 2-by-2 matrices with entries in $\{0, 1, \ldots, 5\}$. These results allow us to raise the general lower bound for the $\lambda$-number of a GPG from 5 to 6 if it contains a subgraph isomorphic to certain crosses, ultimately enabling us to verify the following result in Section 3.

**Theorem 1.3.** *Let $n$ and $d$ be integers so that $n \geq 3$ and $1 \leq d \leq n/2$. If $G$ is the crossed prism $XPr(n, d)$, then $\lambda(G) = 5$ when*

(a) *$d = 1$ and $n = 3$; or*

(b) *$d \equiv 0 \pmod{3}$ and $(n - d) \equiv 0 \pmod{3}$; or*

(c) *$d \equiv 1 \pmod{3}$ and $(n - d) \equiv 1 \pmod{3}$ with $d \geq 7$.*

*Furthermore, $\lambda(G) = 7$ when $d = 1$ and $n = 4$; otherwise $\lambda(G) = 6$.*

## 2. The $\lambda$-number of crosses

Now, we will completely determine the $\lambda$-number of crosses $X(d)$ with $d \geq 1$ in Theorem 2.4, the main result in this section. The following definitions will simplify the description of an auxiliary directed graph that will be helpful in the preliminary discussion. A sequence of nonnegative integers $x_1, x_2, \ldots, x_m$ *induces* a $k$-labeling of the path $P_m$ with vertices $u_1, u_2, \ldots, u_m$ and edges $\{u_i, u_{i+1}\}$ for $i = 1, 2, \ldots, m - 1$, if the assignment of $x_i$ to $u_i$ for $i = 1, 2, \ldots, m$ produces a $k$-labeling of $P_m$.

Let $D$ be the directed graph with vertex set containing the 2-by-2 matrices $M$ with entries in $\{0, 1, \ldots, 5\}$ such that:

- the sequence $M_{0,0}, M_{0,1}, M_{1,1}, M_{1,0}$ induces a 5-labeling of $P_4$, in which case $M$ is called a *left-vertex*; or

- the sequence $M_{0,1}, M_{0,0}, M_{1,0}, M_{1,1}$ induces a 5-labeling of $P_4$, in which case $M$ is called a *right-vertex*.

Notice that a vertex can be both a left- and right-vertex. Given a left-vertex $M$ and a right-vertex $N$ different from $M$, the directed edge set of $D$ contains:

- the *solid edge* $(M, N)$, if $M_{i,1} = N_{i,0}$ for $i = 0, 1$ (i.e., the last column of $M$ is equal to the first column of $N$), $M_{0,0} \neq N_{0,1}$, and $M_{1,0} \neq N_{1,1}$; and

- the *dashed edge* $(N, M)$, if there exists a directed path of solid edges from $M$ to $N$ of length at least 1 so that the two sequences $(N_{0,0}, N_{0,1}, M_{1,0}, M_{1,1})$ and $(N_{1,0}, N_{1,1}, M_{0,0}, M_{0,1})$ each induces a 5-labeling of $P_4$.

Observe that there is a natural one-to-one relationship between the set of 5-labelings of crosses $X(d)$ with $d \geq 2$ and the set of the *X-cycles* defined as directed cycles in $D$ containing exactly one dashed edge. More specifically, for a 5-labeling

$$A(7) = \begin{array}{|cccccccc|} \hline 0 & 2 & 5 & 0 & 2 & 4 & 1 & 5 \\ 0 & 4 & 1 & 3 & 5 & 0 & 3 & 5 \\ \hline \end{array}$$

$$\begin{array}{|cc|} \hline 0 & 2 \\ 0 & 4 \\ \hline \end{array} \rightarrow \begin{array}{|cc|} \hline 2 & 5 \\ 4 & 1 \\ \hline \end{array} \rightarrow \begin{array}{|cc|} \hline 5 & 0 \\ 1 & 3 \\ \hline \end{array} \rightarrow \begin{array}{|cc|} \hline 0 & 2 \\ 3 & 5 \\ \hline \end{array} \rightarrow \begin{array}{|cc|} \hline 2 & 4 \\ 5 & 0 \\ \hline \end{array} \rightarrow \begin{array}{|cc|} \hline 4 & 1 \\ 0 & 3 \\ \hline \end{array} \rightarrow \begin{array}{|cc|} \hline 1 & 5 \\ 3 & 5 \\ \hline \end{array}$$
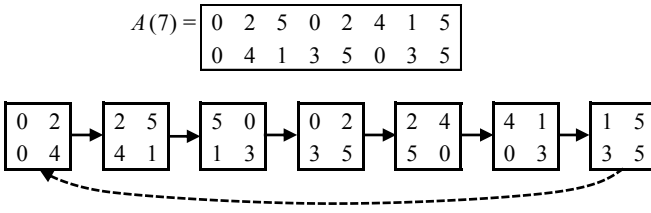
**Figure 5.** A 5-labeling $A(7)$ of $X(7)$ and corresponding $X$-cycle.

of the cross $X(d)$ represented by a 2-by-$(d + 1)$ matrix $A(d)$, consider for each $i = 0, 1, \ldots, d - 1$, the 2-by-2 submatrix $M(i)$ with the $i$-th and $(i + 1)$-th columns of $A(d)$. From the definition of $D$, it is straightforward to verify that the vertices $M(i)$ for $i = 0, 1, \ldots, d - 1$ induce an $X$-cycle and, moreover, this correspondence is one-to-one. We illustrate this correspondence in Figure 5 with a 5-labeling $A(7)$ of $X(7)$ and its associated $X$-cycle. In particular, solid and dashed edges are represented by solid and dashed arrows, respectively. The start and end vertices of the maximal directed path with solid edges within the $X$-cycle are left- and right-vertices, respectively. For the sake of simplicity, we will sometimes abuse the notation and use a 5-labeling in matrix form to refer to the corresponding $X$-cycle and vice-versa.

We define three operations on subgraphs $D^*$ of $D$ that will simplify the description of some of its properties:

- *dual* of $D^*$: replace entry $j$ of every vertex of $D^*$ with its *dual* $5 - j$.
- *flip* of $D^*$: swap the two rows of every vertex of $D^*$.
- *reverse* of $D^*$: swap the two columns of every vertex of $D^*$, and reverse the direction of every edge of $D^*$.

Notice that each of these operations coincides with its inverse and preserves the structure of $X$-cycles.

To generate the directed graph $D$, a computer program classified each of the $6^4$ matrices with entries in $\{0, 1, \ldots, 5\}$ as a left- and/or right-vertex of $D$ if possible, and discarded it otherwise. The algorithm then considered each pair of a left-vertex $M$ and right-vertex $N$ different from $M$, and added a solid edge $(M, N)$ and/or dashed edge $(N, M)$ if the pair satisfied the associated definition stated above. This algorithm relies on brute force — every vertex pair is considered individually — and could certainly be improved by cleverly integrating results about duals, flips, and reverses. Still, this algorithm is sound in the sense that every edge added satisfies either the solid or dashed edge definition, and complete, in the sense that every 2-by-2 matrix was considered from the outset and every pair of left- and right-vertices was tested for both solid and dashed connections.

$$A(7+3q) = \begin{array}{|cccccccccccc|} 0 & 2 & 5 & 0 & 2 & 4 & 0 & 2 & 4 & 1 & 5 \\ 0 & 4 & 1 & 3 & 5 & 1 & 3 & 5 & 0 & 3 & 5 \end{array}$$

$$A(3+3q) = \begin{array}{|ccccccc|} 2 & 4 & 0 & 2 & 4 & 0 & 2 \\ 5 & 1 & 3 & 5 & 1 & 3 & 5 \end{array} \quad \text{or} \quad \begin{array}{|ccccccc|} 4 & 0 & 2 & 4 & 0 & 2 & 4 \\ 1 & 3 & 5 & 1 & 3 & 5 & 1 \end{array}$$

$$\text{or} \quad \begin{array}{|ccccccc|} 0 & 2 & 4 & 0 & 2 & 4 & 0 \\ 3 & 5 & 1 & 3 & 5 & 1 & 3 \end{array}$$

$$A(3) = \begin{array}{|cccc|} 2 & 4 & 1 & 5 \\ 2 & 0 & 3 & 5 \end{array} \quad \text{or} \quad \begin{array}{|cccc|} 0 & 2 & 5 & 3 \\ 0 & 4 & 1 & 3 \end{array}$$

$$A(2) = \begin{array}{|ccc|} 0 & 2 & 4 \\ 1 & 5 & 3 \end{array} \quad \text{or} \quad \begin{array}{|ccc|} 2 & 0 & 4 \\ 1 & 3 & 5 \end{array} \quad \text{or} \quad \begin{array}{|ccc|} 0 & 2 & 4 \\ 0 & 5 & 3 \end{array}$$

$$\text{or} \quad \begin{array}{|ccc|} 2 & 0 & 5 \\ 1 & 3 & 5 \end{array} \quad \text{or} \quad \begin{array}{|ccc|} 2 & 4 & 1 \\ 3 & 0 & 5 \end{array} \quad \text{or} \quad \begin{array}{|ccc|} 0 & 5 & 2 \\ 4 & 1 & 3 \end{array}$$

**Figure 6.** $X$-cycles in component $D_1$ as 5-labelings $A(d)$ of crosses $X(d)$ with $d \geq 2$ (respective flips are not shown).

Excluding isolated vertices, the directed graph $D$ consists of four connected components $D_i$ for $i = 1, 2, 3, 4$, and their respective duals, where $D_1$, $D_2$, and $D_3$ are provided in the online supplement, and $D_4$ is the flip of $D_3$. The symmetry of crosses implies the following relationships among these components that can be verified by inspection:

(a) the flip of $D_1$ is $D_1$;

(b) the reverse of $D_1$ is the dual of $D_1$;

(c) the flip of $D_2$ is the dual of $D_2$;

(d) the reverse of $D_2$ is $D_2$;

(e) the reverse of $D_3$ (resp., $D_4$) is the dual of $D_4$ (resp., $D_3$).

In Lemmas 2.1 through 2.3, we exhibit all the $X$-cycles in components $D_i$ for $i = 1, 2, 3, 4$. These cycles together with their duals (which are the $X$-cycles in the dual of $D_i$ for $i = 1, 2, 3, 4$) are all possible 5-labelings $A(d)$ of crosses $X(d)$ with $d \geq 2$.

**Lemma 2.1.** *The $X$-cycles in component $D_1$ are given by the 5-labelings $A(d)$ of crosses $X(d)$ with $d \geq 2$ in Figure 6 and their respective flips (i.e., the flip of a matrix with two rows is obtained by swapping its rows). Each shaded block of three consecutive columns within a matrix can be replaced with $q \geq 0$ copies of itself, arranged consecutively as needed to reach the desired value of $d$ (this convention will be used from this point forward).*
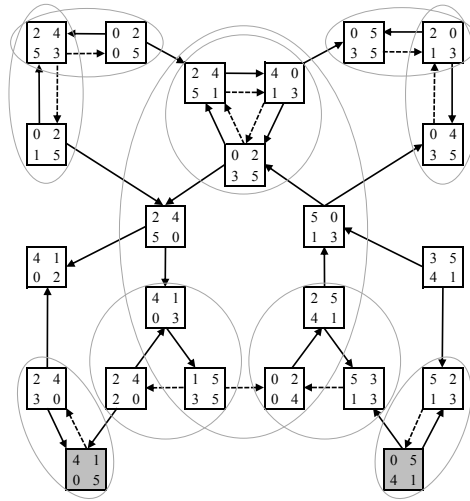
**Figure 7.** The directed subgraph $H$ of $D_1$ and its $X$-cycles (circled).

*Proof.* Let $H$ be the directed subgraph of $D_1$ in the online supplement induced by the two shaded vertices and all the vertices above them; $H$ is shown in Figure 7. By inspection, one can verify that the flip of $H$ is exactly the directed subgraph of $D_1$ induced by the two shaded vertices and all the vertices below them. Moreover, an $X$-cycle in component $D_1$ must be either completely within $H$ or completely within the flip of $H$. Therefore, the lemma follows by exhibiting all the $X$-cycles within $H$. They are circled in Figure 7 and their corresponding 5-labelings of crosses are given in Figure 6. □

**Lemma 2.2.** *The $X$-cycles in component $D_2$ are given by the 5-labelings $A(2)$ of crosses $X(2)$ in Figure 8.*

*Proof.* Since all solid edges in $D_2$ in the online supplement are directed from left to right and the only four dashed edges are directed from right to left, it is straightforward to verify that there are only four $X$-cycles of length 2 with corresponding 5-labelings given in Figure 8. □

**Lemma 2.3.** *The only $X$-cycle in component $D_3$ (resp., $D_4$) is given by the 5-labeling $A(3)$ (resp., flip of $A(3)$) of cross $X(3)$ in Figure 9.*

$$A(2) = \begin{array}{|ccc|} 3 & 1 & 5 \\ 2 & 4 & 0 \end{array} \quad \text{or} \quad \begin{array}{|ccc|} 1 & 5 & 2 \\ 4 & 0 & 3 \end{array} \quad \text{or} \quad \begin{array}{|ccc|} 2 & 5 & 1 \\ 3 & 0 & 4 \end{array} \quad \text{or} \quad \begin{array}{|ccc|} 5 & 1 & 3 \\ 0 & 4 & 2 \end{array}$$

**Figure 8.** $X$-cycles in component $D_2$ as 5-labelings $A(2)$ of crosses $X(2)$.

$$A(3) = \begin{array}{|cccc|} \hline 1 & 5 & 2 & 4 \\ 1 & 3 & 0 & 4 \\ \hline \end{array}$$

**Figure 9.** The only $X$-cycle in $D_3$ as a 5-labeling $A(3)$ of cross $X(3)$.

$$A(4) = \begin{array}{|ccccc|} \hline 3 & 6 & 0 & 3 & 6 \\ 0 & 4 & 2 & 5 & 0 \\ \hline \end{array} \qquad A(5+3q) = \begin{array}{|ccccccccc|} \hline 3 & 1 & 6 & 4 & 0 & 6 & 4 & 2 & 6 \\ 0 & 5 & 3 & 1 & 5 & 3 & 1 & 5 & 0 \\ \hline \end{array}$$

**Figure 10.** 6-labelings of crosses $X(d)$ for $d = 4$ and for [$d \equiv$ 2 (mod 3) and $d \geq 5$], respectively.

*Proof.* It is straightforward to verify that there is only one $X$-cycle in $D_3$ in the online supplement with corresponding 5-labeling given in Figure 9. Since $D_4$ is the flip of $D_3$, the flip of $A(3)$ corresponds to the only $X$-cycle in $D_4$.  □

We can finally state the main result of this section.

**Theorem 2.4.** *If $G$ is the cross $X(d)$ with $d \geq 1$, then $\lambda(G) = 4$ when $d = 1$, $\lambda(G) = 6$ when $d = 4$ or when [$d \equiv 2$ (mod 3) with $d \geq 5$], otherwise $\lambda(G) = 5$. In addition, the only possible 5-labelings of $G$ when $d \geq 2$ are the ones in Figure 6 and 9 with their respective flips, the ones in Figure 8, and all the respective duals (i.e., the dual of a matrix is obtained by replacing each entry $j$ with $5 - j$).*

*Proof.* The cross $X(1)$ is a cycle on four vertices which has $\lambda$-number 4 [Griggs and Yeh 1992]. The second sentence in the theorem's statement follows from the construction of $D$ and Lemmas 2.1 through 2.3. Hence, if $G$ has a 5-labeling, then $d = 2$, $d \equiv 0$ (mod 3), or [$d \equiv 1$ (mod 3) and $d \geq 7$] (refer to Figures 6, 8, and 9), thus $\lambda(G) = 5$ (recall from Section 1 that GPGs have $\lambda$-number at least 5). On the other hand, if $G$ does not have a 5-labeling, then $\lambda(G) \geq 6$ and either $d = 4$ or [$d \equiv 2$ (mod 3) and $d \geq 5$], thus $\lambda(G) = 6$ follows from the 6-labelings of $G$ in Figure 10.  □

We close this section by mentioning that the directed subgraph of $D$ induced by its vertices that are simultaneously left- and right-vertices (vertices within double-lined squares in the online supplement and their respective duals) was used in [Klavžar and Vesel 2003] to exhibit 5-labelings of prisms.

## 3. The λ-number of crossed prisms

Let $n$ and $d$ be integers so that $n \geq 3$ and $1 \leq d \leq n/2$. The main goal of this section is to find the $\lambda$-number of the crossed prism $XPr(n, d)$; that is, prove Theorem 1.3 of Section 1. In Lemma 3.1 we will discuss the case $d \geq 2$, and the case $d = 1$ will be examined in Lemma 3.2.

The following construction of $k$-labelings of $XPr(n, d)$ for $d \geq 2$ using $k$-labelings of the crosses $X(d)$ and $X(n-d)$ will be useful in the proof of Lemma 3.1. Consider a $k$-labeling of the cross $X(d)$ given as a 2-by-$(d+1)$ matrix $M$. In addition, consider

a $k$-labeling of the cross $X(n - d)$ given as a 2-by-$(n - d + 1)$ matrix $N$. We will
say that $M$ and $N$ *mesh* if the following two conditions are satisfied:

(i) $M_{i,d} = N_{i,0}$ for $i = 0, 1$ (i.e., the last column of $M$ is equal to the first column
of $N$), $M_{0,d-1} \neq N_{0,1}$, and $M_{1,d-1} \neq N_{1,1}$;

(ii) $N_{i,n-d} = M_{i,0}$ for $i = 0, 1$ (i.e., the last column of $N$ is equal to the first column
of $M$), $N_{0,n-d-1} \neq M_{0,1}$, and $N_{1,n-d-1} \neq M_{1,1}$.

Observe that if $M$ and $N$ mesh, then the matrix $mesh(M, N)$ obtained by combining
the first $d$ columns of $M$ immediately followed by the first $n - d$ columns of $N$
provides a $k$-labeling of the crossed prism $XPr(n, d)$. For example, $A(9, 4) =$
$mesh(A(4), A(5))$ as seen in Figure 4 is a 6-labeling of $XPr(9, 4)$.

**Lemma 3.1.** *Let $n$ and $d$ be integers so that $n \geq 3$ and $2 \leq d \leq n/2$. If $G$ is the
crossed prism $XPr(n, d)$, then $\lambda(G) = 5$ when*

(a) $d \equiv 0 \pmod 3$ *and* $(n - d) \equiv 0 \pmod 3$; *or*

(b) $d \equiv 1 \pmod 3$ *and* $(n - d) \equiv 1 \pmod 3$ *with* $d \geq 7$.

*Otherwise* $\lambda(G) = 6$.

*Proof.* Suppose (a) holds. Select the first of the corresponding three choices for
$A(3 + 3q)$ in Figure 6 (we could also select the second or the third choice instead).
Let $q_1$ and $q_2$ be integers so that $d = 3 + 3q_1$ and $n - d = 3 + 3q_2$. Hence $A(3 + 3q_1)$
and $A(3 + 3q_2)$ are 5-labelings of the crosses $X(d)$ and $X(n - d)$, respectively, and
these matrices mesh. From the observation right before the lemma, the matrix
$mesh(A(3 + 3q_1), A(3 + 3q_2))$ is a 5-labeling of $G$, hence $\lambda(G) \leq 5$. Recall from
Section 1 that GPGs have $\lambda$-number at least 5, therefore $\lambda(G) = 5$.

Suppose (b) holds. Select the $A(7 + 3q)$ in Figure 6. Let $q_1$ and $q_2$ be integers so
that $d = 7 + 3q_1$ and $n - d = 7 + 3q_2$ (note that $n - d \geq d \geq 7$). Hence $A(7 + 3q_1)$ and
the dual of $A(7 + 3q_2)$ are 5-labelings of the crosses $X(d)$ and $X(n - d)$, respectively,
and these matrices mesh. Similarly to the previous paragraph, we conclude that
$\lambda(G) = 5$.

Suppose for the remainder of the proof that neither (a) nor (b) is satisfied. We will
first show that $\lambda(G) \geq 6$. If $d = 4$ or $[d \equiv 2 \pmod 3$ with $d \geq 5]$, then the $\lambda$-number
of $X(d)$ is 6 by Theorem 2.4 and therefore $\lambda(G) \geq 6$ since $X(d)$ is a subgraph of $G$.
Likewise, we can replace $d$ with $n - d$ in the previous sentence and reach the same
conclusion. To verify the remaining cases, we suppose for contradiction that $G$ has
a 5-labeling given by a 2-by-$n$ matrix $A(n, d)$. The matrix $M$ given by the first
$d + 1$ columns of $A(n, d)$ is a 5-labeling of the cross $X(d)$, and the matrix $N$ given
by the last $n - d$ columns followed by the first column of $A(n, d)$ is a 5-labeling of
the cross $X(n - d)$. Thus $M$ and $N$ mesh and are instances of the set of matrices
described in Theorem 2.4. We will examine the following remaining cases and
reach a contradiction in all of them, which implies $\lambda(G) \geq 6$.

Case 1: [$d \equiv 0$ (mod 3) and $(n - d) \equiv 1$ (mod 3) with $n - d \geq 7$] or [$(n - d) \equiv 0$ (mod 3) and $d \equiv 1$ (mod 3) with $d \geq 7$]. Suppose [$d \equiv 0$ (mod 3) and $(n - d) \equiv 1$ (mod 3) with $n - d \geq 7$]. Note that $N$ or its dual must be an instance of $A(7 + 3q)$ in Figure 6 with their respective flips, so the first and last columns of $N$ are different and have entries in $\{0, 5\}$. Since $M$ and $N$ mesh, the first and last columns of $M$ must also be different and have entries in $\{0, 5\}$. Unfortunately, the same does not hold for any instance of $A(3 + 3q)$ and $A(3)$ in Figure 6 and 9 with their respective flips and all their respective duals, a contradiction. Similarly, we also reach a contradiction in the case [$(n - d) \equiv 0$ (mod 3) and $d \equiv 1$ (mod 3) with $d \geq 7$] by switching the roles of $M$ and $N$ in the discussion above.

Case 2: [$d = 2$ and $(n - d) \equiv 0$ (mod 3)] or [$d = 2$ and $(n - d) \equiv 1$ (mod 3) with $(n - d) \geq 7$]. Note that each instance $A(2)$ in Figure 6 with their respective flips, or in Figure 8, and all their respective duals, uses at least three different labels in the first and last columns combined. In contrast, each instance of $A(7 + 3q)$, $A(3 + 3q)$, and $A(3)$ in Figures 6 and 9 with their respective flips and all their respective duals, uses only two different labels in the first and last columns. So $M$ and $N$ cannot mesh, a contradiction.

Case 3: $d = 2$ and $(n - d) = 2$. We can verify by inspection that all pairs of instances of $A(2)$ in Figure 6 with their respective flips, or in Figure 8, and all their respective duals do not mesh (note that no component has directed cycles of length 4 containing only solid edges). So $M$ and $N$ cannot mesh, a final contradiction.

Finally, to prove that $\lambda(G) = 6$, it suffices to show that $\lambda(G) \leq 6$. Observe that $\{d, n - d\} = \{d_1, d_2\}$ for a combination of values $d_1$ and $d_2$ described by one of the rows of the table in the online supplement. This row exhibits two 6-labelings of the crosses $X(d_1)$ and $X(d_2)$, respectively, as two matrices that mesh. From the observation right before Lemma 3.1, we can conclude that $\lambda(G) \leq 6$.  □

**Lemma 3.2.** *If $G$ is the crossed prism $XPr(n, 1)$ with $n \geq 3$, then $\lambda(G) = 5$ if $n = 3$; $\lambda(G) = 7$ if $n = 4$; otherwise $\lambda(G) = 6$.*

*Proof.* Recall from Section 1 that an L(2,1)-labeling $f$ of $XPr(n, d)$ is given by a 2-by-$n$ matrix $A(n, d)$ where the entry on the $i$-th row, $j$-th column will be the label $f(v_j)$ if $i = 0$, and $f(w_j)$ if $i = 1$, for $j = 0, 1, \ldots, n - 1$, and the notation is as introduced in Definition 1.2 (the ends of crossed edges are in the 0-th and $d$-th columns). If $n = 3$, then the 5-labeling $A(3, 1)$ of $G$ in Figure 11 implies $\lambda(G) = 5$ (recall from Section 1 that GPGs have $\lambda$-number at least 5). If $n = 4$, then $G$ has diameter 2 and therefore $\lambda(G) \geq |V(G)| - 1 = 2n - 1 = 7$ [Griggs and Yeh 1992]; the 7-labeling $A(4, 1)$ of $G$ in Figure 11 implies $\lambda(G) = 7$.

Assume $n \geq 5$. We will show first that $\lambda(G) \geq 6$. Suppose for contradiction that $G$ has a 5-labeling given by a 2-by-$n$ matrix $A(n, 1)$. The matrix $M$ given by the first two columns of $A(n, 1)$ is a 5-labeling of the cross $X(1)$, and the matrix $N$ given

$$A(3, 1) = \begin{array}{|ccc|} \hline 0 & 2 & 4 \\ 5 & 3 & 1 \\ \hline \end{array} \qquad\qquad A(4, 1) = \begin{array}{|cccc|} \hline 0 & 2 & 5 & 3 \\ 4 & 6 & 1 & 7 \\ \hline \end{array}$$

**Figure 11.** A 5-labeling of $XPr(3, 1)$ and a 7-labeling of $XPr(4, 1)$, respectively.

$$A(5 + 3q, 1) = \begin{array}{|cccccccc|} \hline 0 & 4 & 6 & 1 & 3 & 6 & 1 & 3 \\ 1 & 5 & 2 & 4 & 0 & 2 & 4 & 6 \\ \hline \end{array}$$

$$A(6 + 3q, 1) = \begin{array}{|ccccccccc|} \hline 0 & 4 & 6 & 3 & 1 & 6 & 1 & 5 & 2 \\ 1 & 5 & 2 & 4 & 0 & 2 & 4 & 0 & 6 \\ \hline \end{array}$$

$$A(7 + 3q, 1) = \begin{array}{|cccccccccc|} \hline 0 & 4 & 6 & 3 & 1 & 6 & 0 & 2 & 4 & 6 \\ 1 & 5 & 2 & 0 & 5 & 2 & 4 & 6 & 0 & 3 \\ \hline \end{array}$$

**Figure 12.** The 6-labelings of $XPr(n, 1)$ for $n \geq 5$.

by the last $n - 1$ columns followed by the first column of $A(n, 1)$ is a 5-labeling of the cross $X(n - 1)$. Since $N$ has $n \geq 5$ columns, the $X$-cycle corresponding to $N$ must be in component $D_1$ or the dual of $D_1$ of the directed graph $D$ constructed in Section 2. We may assume without loss of generality that this $X$-cycle is in $D_1$. The cross $X(1)$ has diameter 2 so its four vertices must be assigned different labels, thus the first and last columns of $N$ must contain four different labels. Unfortunately, this is not the case for $A(7+3q)$ or $A(3+3q)$ in Figure 6 and their respective flips, implying that $n < 5$, a contradiction, and so $\lambda(G) \geq 6$ holds. The desired equality follows from the 6-labelings of $G$ provided in Figure 12. $\qquad \square$

Finally, Theorem 1.3 in Section 1 is a straightforward consequence of Lemmas 3.1 and 3.2.

## 4. Closing remarks

In this work, we made progress towards closing the GPG conjecture by showing that any crossed prism $G$ satisfies $\lambda(G) \leq 7$ and that, in fact, all but one $G$ satisfy $\lambda(G) \leq 6$. These crossed prisms were of particular interest, as they allowed us to examine how the controlled introduction of asymmetries to prisms would impact both the $\lambda$-number and overall proof strategies. The complications these breaks in symmetry introduced were nontrivial, and we ultimately determined $\lambda(G)$ by constructing and inspecting an auxiliary directed graph motivated by previous studies. We hope that these ideas help the community examine other families of graphs — for instance, prisms with more than one pair of crossed edges, perturbations of the $n$-stars — as we move closer to putting the general GPG conjecture to rest.

## Acknowledgements

## References

[Adams et al. 2006]  S. S. Adams, J. Cass, and D. S. Troxell, "An extension of the channel-assignment problem: $L(2, 1)$-labelings of generalized Petersen graphs", *IEEE Trans. Circuits Syst. I Regul. Pap.* **53**:5 (2006), 1101–1107. MR

[Adams et al. 2007]  S. S. Adams, J. Cass, M. Tesch, D. S. Troxell, and C. Wheeland, "The minimum span of $L(2, 1)$-labelings of certain generalized Petersen graphs", *Discrete Appl. Math.* **155**:10 (2007), 1314–1325. MR Zbl

[Adams et al. 2012]  S. S. Adams, P. Booth, H. Jaffe, D. S. Troxell, and S. L. Zinnen, "Exact $\lambda$-numbers of generalized Petersen graphs of certain higher-orders and on Möbius strips", *Discrete Appl. Math.* **160**:4–5 (2012), 436–447. MR Zbl

[Calamoneri 2011]  T. Calamoneri, "The $L(h, k)$-labelling problem: an updated survey and annotated bibliography", *Comput. J.* **54**:8 (2011), 1344–1371.

[Franks 2015]  C. Franks, "The $\Delta^2$ conjecture holds for graphs of small order", *Involve* **8**:4 (2015), 541–549. MR Zbl

[Georges and Mauro 2002]  J. P. Georges and D. W. Mauro, "On generalized Petersen graphs labeled with a condition at distance two", *Discrete Math.* **259**:1–3 (2002), 311–318. MR Zbl

[Georges and Mauro 2003]  J. P. Georges and D. W. Mauro, "On regular graphs optimally labeled with a condition at distance two", *SIAM J. Discrete Math.* **17**:2 (2003), 320–331. MR Zbl

[Griggs and Yeh 1992]  J. R. Griggs and R. K. Yeh, "Labelling graphs with a condition at distance 2", *SIAM J. Discrete Math.* **5**:4 (1992), 586–595. MR Zbl

[Hale 1980]  W. K. Hale, "Frequency assignment: theory and applications", *Proc. IEEE* **68**:12 (1980), 1497–1514.

[Havet et al. 2012]  F. Havet, B. Reed, and J.-S. Sereni, "Griggs and Yeh's conjecture and $L(p, 1)$-labelings", *SIAM J. Discrete Math.* **26**:1 (2012), 145–168. MR Zbl

[Huang et al. 2012]  Y.-Z. Huang, C.-Y. Chiang, L.-H. Huang, and H.-G. Yeh, "On $L(2, 1)$-labeling of generalized Petersen graphs", *J. Comb. Optim.* **24**:3 (2012), 266–279. MR Zbl

[Jha et al. 2000]  P. K. Jha, A. Narayanan, P. Sood, K. Sundaram, and V. Sunder, "On $L(2, 1)$-labeling of the Cartesian product of a cycle and a path", *Ars Combin.* **55** (2000), 81–89. MR Zbl

[Klavžar and Vesel 2003]  S. Klavžar and A. Vesel, "Computing graph invariants on rotagraphs using dynamic algorithm approach: the case of $(2, 1)$-colorings and independence numbers", *Discrete Appl. Math.* **129**:2–3 (2003), 449–460. MR Zbl

[Kuo and Yan 2004]  D. Kuo and J.-H. Yan, "On $L(2, 1)$-labelings of Cartesian products of paths and cycles", *Discrete Math.* **283**:1–3 (2004), 137–144. MR Zbl

matthew.beaudouin-lafon@students.olin.edu

*Franklin W. Olin College of Engineering, Needham, MA 02492, United States*

serena.chen@students.olin.edu      *Franklin W. Olin College of Engineering, Needham, MA 02492, United States*

nkarst@babson.edu      *Mathematics and Sciences Division, Babson College, Babson Park, MA 02457, United States*

jessica.oehrlein@columbia.edu      *Fu Foundation School of Engineering and Applied Sciences, Columbia University, New York, NY 10027, United States*

troxell@babson.edu      *Mathematics and Sciences Division, Babson College, Babson Park, MA 02457, United States*

msp

# Normal forms of
# endomorphism-valued power series

## Christopher Keane and Szilárd Szabó

(Communicated by Kenneth S. Berenhaut)

We show for $n, k \geq 1$, and an $n$-dimensional complex vector space $V$ that if an element $A \in \mathrm{End}(V)[[z]]$ has constant term similar to a Jordan block, then there exists a polynomial gauge transformation $g$ such that the first $k$ coefficients of $g A g^{-1}$ have a controlled normal form. Furthermore, we show that this normal form is unique by demonstrating explicit relationships between the first $nk$ coefficients of the Puiseux series expansion of the eigenvalues of $A$ and the entries of the first $k$ coefficients of $g A g^{-1}$.

## Introduction

From Galois theory, we know that polynomials of degree greater than 4 are not solvable by radicals. So finding the eigenvalues of a companion matrix of the form

$$\begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & 0 & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ \beta_{n-1} & \beta_{n-2} & \beta_{n-3} & \cdots & \beta_0 \end{pmatrix}$$

algebraically in terms of the $\beta_i$ is not possible. If, however, the $\beta_i$ have expansions $\beta_i(z)$ in terms of some other variable $z$ with $\beta_i(0) = 0$, we may then ask to find the coefficients in the series expansions of these eigenvalues in terms of these $\beta_i(z)$.

In this paper, we work with a formal power series $A \in \mathrm{End}(V)[[z]]$ whose constant term is a regular nilpotent endomorphism. We want to compute the coefficients of the Puiseux expansion of the eigenvalues of $A$, but since this is not possible algebraically we search for some normal form obtained via conjugating by an invertible transformation. Clearly, conjugating does not modify the eigenvalues

of $A$, and our aim is to conjugate $A(z)$ to a simple shape that allows us to compute explicit relationships between coefficients of the series expansion of the eigenvalues and the coefficients of the conjugate.

In [Ivanics et al. 2016], this problem arose in taking an endomorphism of a vector bundle with some fixed local behavior and searching for the base locus of its corresponding spectral curves. They work with the special case of rank-2 vector bundles $E$ and irregular Higgs fields $\theta(z)$, i.e., meromorphic sections of the endomorphism bundle of $E$ tensored by the canonical bundle. Specifically, the endomorphism $\theta$ is assumed to have a single pole of order 4 at $z = 0$ with leading-order term having nontrivial nilpotent part, and the authors show that its polar part may be brought to a simple form up to applying some holomorphic gauge transformations. The authors also note that the case of endomorphisms having two distinct eigenvalues is much simpler. Let us point out that the rank-2 cases can be tackled algebraically due to the existence of the quadratic formula, but that method breaks down in higher-rank cases for the Galois-theoretic reason alluded to above. Another observation is that up to a shift of the index of summation, it is equivalent to consider power series or Laurent series with a fixed finite pole order. Therefore, in this paper we content ourselves with working with power series, however the role of the pole order (the number of terms in the normal form to be controlled) is played by our parameter $k$.

We cover the general rank-$n$ case for endomorphism-valued power series where the leading-order term is a regular nilpotent endomorphism. That is, we maintain the assumptions of [Ivanics et al. 2016], aside from the pole of order 4 and the rank being equal to 2, extending their results to vector bundles of arbitrary rank and an arbitrary number of terms in the expansion of the endomorphism by presenting existence and uniqueness statements for the normal form of endomorphism-valued power series. This has the same consequence as in [Ivanics et al. 2016] concerning the base locus of generic irregular Higgs bundles with a regular nilpotent leading-order term.

This question is significantly more involved if the constant coefficient of $A$ is a regular matrix with more than one eigenvalue, and even more so if the constant coefficient of $A$ is not regular. The next step we would take to obtain future results would be to examine the case of the constant term of $A$ being regular with more than one eigenvalue.

## 1. Preliminaries: endomorphisms, gauge transformations, Puiseux series

In this section we describe what kinds of endomorphisms and gauge transformations we plan to examine.

**1A.** *Constraints on endomorphisms.* We begin by putting constraints on the endomorphisms we want to examine. We remark that the results in this paper hold over

any algebraically closed field of characteristic zero, but we will only be considering vector spaces over $\mathbb{C}$. Let $V$ be a vector space over $\mathbb{C}$ of dimension $n$. Suppose that $z$ is a complex variable, and let $A \in \operatorname{End}(V)[\![z]\!]$, that is, $A$ has the form

$$A(z) = \sum_{m=0}^{\infty} A_m z^m, \quad \text{with } A_m \in M_{n,n}(\mathbb{C}).$$

We observe $A_0 = A(0)$. We also place the following condition of regularity on $A_0$.

**Definition 1.1.** For a vector space $V$ over an algebraically closed field, an $n \times n$ matrix $A_0$ is *regular* if and only if its Jordan normal form is of the form

$$J_{d_1}(\lambda_1) \oplus \cdots \oplus J_{d_s}(\lambda_s),$$

with $i \neq j \implies \lambda_i \neq \lambda_j$, and where each $J_{d_i}(\lambda_i)$ is a Jordan block of size $d_i$ with corresponding eigenvalue $\lambda_i$.

More abstractly, this is equivalent to considering the space of complex $n \times n$ matrices as a Lie algebra and requiring that the centralizer of $A_0$ has minimal dimension. The importance of this will become clearer later with the discussion of the transformation applied to $A$.

**1B. *Constraints on gauge transformations.*** Consider $g \in \operatorname{Aut}(V)[\![z]\!]$, supposing that $g$ has a power series expansion

$$g(z) = \sum_{m=0}^{\infty} g_m z^m, \quad \text{with } g_m \in M_{n,n}(\mathbb{C}), \ g_0 \in \operatorname{GL}_n(\mathbb{C}).$$

We call $g$ an "analytic/formal gauge transformation" (according to whether the radius of convergence of the power-series is 0 or positive), and require that $g_0$ be invertible because we intend to conjugate $A$ by $g$. It is a well-known fact about rings of formal power series that an element is invertible if and only if its constant term is invertible. Since $g$ is a power series of matrices, this means we must have $g_0 \in \operatorname{GL}_n(\mathbb{C})$ for $g$ to be invertible.

We turn our attention to the conjugation of $A$ by $g$, and rename it $B$:

$$g(z)A(z)g^{-1}(z) = B(z) = \sum_{m=0}^{\infty} B_m z^m. \tag{1-1}$$

Our first goal is to design $g$ such that we may control any finite number of the matrix coefficients in the conjugation. Because eigenvalues are invariant under conjugation, transforming $A$ into $B$ will make computation of the eigenvalues of $A$ simpler. We obtain the following theorem, which will be restated later as Theorem 2.1.

**Theorem 1.2.** *Suppose $k, n \geq 1$, $V$ is an n-dimensional vector space over $\mathbb{C}$, and $A \in \mathrm{End}(V)[\![z]\!]$. If $A$ is such that $A_0$ is similar to a Jordan block with eigenvalue 0, then we may construct a polynomial gauge transformation $g$ such that $B_0$ is an upper triangular Jordan block of dimension n and the first k coefficients $B_1, \ldots, B_k$ of $gAg^{-1} = B$ are matrices with nonzero coefficients only in their n-th row.*

The series $B$ will be referred to as "the normal form" from now on. With the existence of this established, we move towards our second goal of determining explicit relationships between the eigenvalues of $A$ and the entries of the coefficients of $B$. Let us enumerate the possibly nonzero entries of $B_m$ from left to right as $b_{mn-n+1}, \ldots, b_{mn}$. We obtain the following result, which will be restated later as Theorem 3.2.

**Theorem 1.3.** *Let $B$ be the normal form of $A$ as described in Theorem 1.2, and suppose that the bottom left coefficient $b_1$ of $B_1$ determined by the normal form is nonzero. The eigenvalues of $A$ have a Puiseux expansion*

$$\zeta(z) = \sum_{m=1}^{\infty} a_m z^{m/n},$$

*and for fixed $s \geq 1$, the first s coefficients $a_1, \ldots, a_s$ of the Puiseux expansion explicitly determine and are determined by the first s entries $b_1, \ldots, b_s$ of the matrices making up the normal form $B$.*

In particular, this theorem tells us that for fixed $k$ the normal form $B$ of $A$ is uniquely determined. In all cases we assume $A(z) = \sum_{m=0}^{\infty} A_m z^m$ is such that $A_0$ is similar to a Jordan block. Thus we may define $g_0 \in \mathrm{GL}_n(\mathbb{C})$ such that

$$B_0 = g_0 A_0 g_0^{-1}$$

has the desired Jordan block form. This is a constant transformation, which is notable since the final $g$ will be a finite product of polynomials. Specifically, we will build $g$ as a product of $g_0$ introduced above and nonconstant factors $h_\ell$ of the form

$$h_\ell(z) = I_n + g_\ell z^\ell,$$

where $I_n$ is the $n \times n$ identity matrix and $1 \leq \ell \leq k \in \mathbb{Z}^+$. This is an important point, because it means that $g$ will be a polynomial, hence everywhere convergent, so applying them to $A$ will not affect the convergence radius of $A$. This means that the portion of our results concerning gauge transformations will apply to rings of power series where convergence is a relevant concern. Furthermore, since we only consider the terms of $A$ up to the $k$-th degree we will be applying $k$ of these $h_\ell$ transformations, so instead of computing an explicit form for $g^{-1}$, we will only need that $h_\ell^{-1}(z) =$

$I_n - g_\ell z^\ell + O(z^{\ell+1})$. Then conjugation of $A$ by one of the factors $h_\ell$ looks like

$$h_\ell(z)A(z)h_\ell^{-1}(z) = (I_n + g_\ell z^\ell)\left(\sum_{m=0}^{\infty} A_m z^m\right)(I_n - g_\ell z^\ell) + O(z^{\ell+1})$$

$$= \left(\sum_{m=0}^{\ell-1} A_m z^m\right) + (A_\ell - [A_0, g_\ell])z^\ell + O(z^{\ell+1}),$$

where $[A_0, g_\ell] = A_0 g_\ell - g_\ell A_0$ represents the commutator. In this manipulation we see that $g$ affects the $\ell$-th term of $A$ without changing the first $\ell - 1$ terms. This is important because we apply the transformations $I_n - g_\ell z^\ell$ iteratively for $1 \leq \ell \leq k$ for $\ell$ *increasing*, ultimately obtaining a polynomial transformation of the form

$$\begin{aligned} g(z) &= h_k(z)h_{k-1}(z)\ldots h_1(z)g_0 \\ &= (I_n + g_k z^k)(I_n + g_{k-1}z^{k-1})\ldots(I_n + g_1 z)g_0. \end{aligned} \tag{1-2}$$

Specifically, considering the map

$$\mathrm{ad}_{A_0} : M_{n,n}(\mathbb{C}) \to M_{n,n}(\mathbb{C}), \quad g_\ell \mapsto [A_0, g_\ell] = A_0 g_\ell - g_\ell A_0, \tag{1-3}$$

will tell us how to construct $g$ to generate a normal form for the conjugated series.

**1C.** *Factorization of the characteristic polynomial of A.* We consider the eigenvalues of endomorphisms in the variable $\zeta$. Let $A(z) = \sum_{m=0}^{\infty} A_m z^m$ be an element of $\mathrm{End}(V)[\![z]\!]$. We have that the characteristic polynomial of $A(z)$ has the form

$$\chi_{A(z)}(\zeta) = \chi_A(z, \zeta) = \det(\zeta I - A(z)) = \zeta^n + a_1(z)\zeta^{n-1} + \cdots + a_n(z), \tag{1-4}$$

with $a_1, \ldots, a_n \in \mathbb{C}[\![z]\!]$. We then recall the following particular case of a result attributed to Puiseux and Newton.

**Theorem 1.4** (Newton–Puiseux). *The characteristic polynomial* (1-4) *factors as*

$$\chi_A(w^n, \zeta) = \prod_{i=1}^{n}(\zeta - \zeta_i(w)), \quad with \quad \zeta_i \in \mathbb{C}[\![w]\!].$$

This version of the theorem is taken from [Abhyankar 1990, Lecture 12], except for identifying the ramification index as $n$ instead of some unspecified divisor of $n!$; this latter identification in turn follows from [Serre 1979, Chapter I, Proposition 17]. Indeed, according to the assumption $b_1 \neq 0$ the $z$-adic valuation of $a_n$ is 1, on the other hand the coefficients $a_1(0), \ldots, a_{n-1}(0)$ clearly vanish as $A_0$ is a nilpotent endomorphism. These conditions mean that $\chi_{A(z)}$ is an Eisenstein polynomial in $\zeta$, thus it is totally ramified, i.e., of ramification index $n$.

For us, the above theorem means that we may decompose the characteristic polynomial of $A$ into linear factors, with the roots being represented by Puiseux series. Furthermore, we will be able to obtain each root of the polynomial by

considering all of the conjugates (in the Galois-theory sense) of a single root by multiplying $w = z^{1/n}$ by some power of a primitive $n$-th root of unity $\omega$. Specifically, after a branch cut we may fix a choice $z^{1/n}$ of $n$-th root of $z$, and then all the roots of the characteristic polynomial are expressible in the form

$$\zeta_i(z) = \sum_{m=1}^{\infty} a_m (\omega^i z^{1/n})^m \tag{1-5}$$

for $i = 0, \ldots, n-1$. Different choices of $z^{1/n}$ only amount to a permutation of the $n$ roots $\zeta_i$.

## 2. Existence of the normal form

In this section we present the construction of a normal form for $A$ where the dimension of the ambient vector space $V$ is an arbitrary integer $n \geq 2$. Furthermore, we fix an arbitrary $k \in \mathbb{Z}^+$.

**Theorem 2.1.** *Take $V$ to be a vector space over $\mathbb{C}$ of dimension $n$, and suppose that $A(z) = \sum_{m=0}^{\infty} A_m z^m$ is an endomorphism of $V$ such that $A_0$ is similar to a Jordan matrix with a single eigenvalue. Then for fixed $k \geq 1$ we may construct a gauge transformation g of the form (1-2) such that the coefficient $B_0$ of $g A g^{-1}(z) = B(z) = \sum_{m=0}^{\infty} B_m z^m$ has the form*

$$B_0 = \begin{pmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & 0 & \cdots & 0 \end{pmatrix},$$

*and the subsequent coefficients have the form*

$$B_\ell = \begin{pmatrix} 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \\ b_{n(\ell-1)+1} & b_{n(\ell-1)+2} & \cdots & b_{n\ell-1} & b_{n\ell} \end{pmatrix}$$

*for $1 \leq \ell \leq k$.*

*Proof.* We want to find a way to conjugate $A$ into $B$ such that $A_0 = B_0$ and the subsequent $B_\ell$ for $1 \leq \ell \leq k$ have the indicated form. So we consider the map $\mathrm{ad}_{A_0} : V \to V$ for an arbitrary matrix $G$ given by $G \mapsto [A_0, G]$, with the bracket

representing the commutator of $A_0$ and $G$. To examine the image of this map, label the entries of $G$ in the usual way and expand:

$$\left[\begin{pmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & 0 & \cdots & 0 \end{pmatrix}, \begin{pmatrix} g_{11} & \cdots & g_{1n} \\ \vdots & \ddots & \vdots \\ g_{n1} & \cdots & g_{nn} \end{pmatrix}\right]$$

$$= \begin{pmatrix} g_{21} & g_{22} - g_{11} & g_{23} - g_{12} & \cdots & g_{2,n} - g_{1,n-1} \\ g_{31} & g_{32} - g_{21} & g_{33} - g_{22} & \cdots & g_{3,n} - g_{2,n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ g_{n,1} & g_{n,2} - g_{n-1,1} & g_{n,3} - g_{n-1,2} & \cdots & g_{n,n} - g_{n-1,n-1} \\ 0 & -g_{n,1} & -g_{n,2} & \cdots & -g_{n,n-1} \end{pmatrix}.$$

Name the above matrix $C$, and name the entries in the usual way. Then see that we may write each entry in the last row as

$$c_{n,t} = -\sum_{j=1}^{t-1} c_{n-j,t-j},$$

as $t$ ranges from 1 to $n$. That is, each entry in the last row is the negative of the sum of entries along the diagonal up and to the left of $c_{n,t}$. We set $c_{n,1} = 0$ by convention. Now although we considered the matrix $G$ to be arbitrary, we may pick the entries of $G$ so that we can make $A_\ell - [A_0, G]$ have a desired form. Specifically, the dependence of the last row of $C$ on the first $n-1$ rows ensures that we can eliminate the first $n-1$ rows of $A_\ell$. This almost certainly affects the last row of $A_\ell$, but this does not matter to us. Thus from the iterative process described at the end of Section 1B, we may find a polynomial of the form (1-2) that we may conjugate $A$ by to turn the $\ell$th coefficient of $B(z)$ into the form

$$B_\ell = \begin{pmatrix} 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \\ b_{n(\ell-1)+1} & b_{n(\ell-1)+2} & \cdots & b_{n\ell-1} & b_{n\ell} \end{pmatrix},$$

for $1 \leq \ell \leq k$. Turning $A_0$ into $B_0$ is much easier, since it is achieved by a constant transformation, and we are assuming that $A_0$ is similar to a matrix of the form $B_0$. This is the desired normal form for the first $k$ coefficients of $B$. □

## 3. Uniqueness of the normal form

In this section we again fix an arbitrary $k \in \mathbb{Z}^+$ and show that the coefficients $b_i$ for $1 \leq i \leq kn$ are uniquely determined by the shape of the normal form $B$ and the coefficients $a_1, \ldots, a_{kn}$ of the Puiseux expansion of the eigenvalues of $A$. We begin the search for relationships between the series of eigenvalues and the entries of the $B_\ell$ with a lemma. For the remainder of this section we now suppose that $A_0 = B_0$ is as in Theorem 2.1 and that the first $k$ coefficients of $B(z)$ may have nonzero entries only in the $n$-th row.

**Lemma 3.1.** *Let $t$ be an integer with $n > t \geq 1$, and $w_1, \ldots, w_{t+1} \in \mathbb{Z}$ be such that*

$$-n < w_1 < 0, \quad 0 < w_2, \ldots, w_{t+1} < n, \quad \sum_{\ell=1}^{t+1} w_\ell = 0.$$

*Define $\omega$ to be a primitive $n$-th root of unity. Then we have that*

$$\frac{1}{t!} \sum_{\substack{1 \leq s_1, \ldots, s_{t+1} \leq n \\ s_j \neq s_\ell, \forall \ell \neq j}} \omega^{w_1 s_1 + \cdots + w_{t+1} s_{t+1}} = (-1)^t n.$$

*Proof.* First, note the following basic identity regarding sums of powers of primitive $n$-th roots of unity: for any $w \in \mathbb{Z}$ such that $n \nmid w$ we have

$$\sum_{j=0}^{n-1} \omega^{jw} = \frac{\omega^{wn} - 1}{\omega^w - 1} = 0. \tag{3-1}$$

For our application below, let us point out that in the sum of the left-hand side the summation index $j$ may equally be chosen to range from 1 to $n$ without changing the value of the sum, because $\omega^{0w} = \omega^{nw}$. Then we proceed by induction on $t$. Starting with $t = 1$, we see that we must have $w_2 = -w_1$, since $w_1 < 0$, and $w_1 + w_2 = 0$. Then see that

$$\frac{1}{1!} \sum_{\substack{1 \leq s_1, s_2 \leq n \\ s_1 \neq s_2}} \omega^{w_1(s_1 - s_2)},$$

and relabeling $u = (s_1 - s_2) \bmod n$ gives

$$n \cdot \frac{1}{1!} \sum_{u=1}^{n-1} \omega^{w_1 u} = n \cdot \frac{1}{1!} \cdot (-1) = (-1)^1 \cdot n,$$

using (3-1) and observing each $u$ is obtained in $n$ possible ways. So the base case is proven.

Now suppose that the claim holds for $t - 1 \geq 1$. For $t$, we then have

$$\frac{1}{t!} \sum_{\substack{1 \leq s_1, \ldots, s_{t+1} \leq n \\ s_j \neq s_\ell, \forall \ell \neq j}} \omega^{w_1 s_1 + \cdots + w_{t+1} s_{t+1}} = \frac{1}{t!} \sum_{\substack{1 \leq s_2, \ldots, s_{t+1} \leq n \\ s_j \neq s_\ell, \forall \ell \neq j}} \omega^{w_2 s_2 + \cdots + w_{t+1} s_{t+1}} \left( \sum_{\substack{s_1 = 1 \\ s_1 \notin \{s_2, \ldots, s_{t+1}\}}}^{n} \omega^{w_1 s_1} \right)$$

and since $w_1 \not\equiv 0 \bmod n$, we may rewrite the inner sum using (3-1):

$$\frac{1}{t!} \sum_{\substack{1 \leq s_2, \ldots, s_{t+1} \leq n \\ s_j \neq s_\ell, \forall \ell \neq j}} \omega^{w_2 s_2 + \cdots + w_{t+1} s_{t+1}} \left( \sum_{\substack{s_1 = 1 \\ s_1 \notin \{s_2, \ldots, s_{t+1}\}}}^{n} \omega^{w_1 s_1} \right)$$

$$= \frac{1}{t!} \sum_{\substack{1 \leq s_2, \ldots, s_{t+1} \leq n \\ s_j \neq s_\ell, \forall \ell \neq j}} \omega^{w_2 s_2 + \cdots + w_{t+1} s_{t+1}} (-\omega^{w_1 s_2} - \cdots - \omega^{w_1 s_{t+1}})$$

$$= -\frac{1}{t!} \sum_{\substack{1 \leq s_1, \ldots, s_{t+1} \leq n \\ s_j \neq s_\ell, \forall \ell \neq j}} \omega^{(w_2 + w_1) s_2 + w_3 s_3 + \cdots + w_{t+1} s_{t+1}} - \cdots$$

$$- \frac{1}{t!} \sum_{\substack{1 \leq s_1, \ldots, s_{t+1} \leq n \\ s_j \neq s_\ell, \forall \ell \neq j}} \omega^{w_2 s_2 + (w_1 + w_3) s_3 + w_4 s_4 + \cdots + w_{t+1} s_{t+1}} - \cdots$$

$$- \frac{1}{t!} \sum_{\substack{1 \leq s_1, \ldots, s_{t+1} \leq n \\ s_j \neq s_\ell, \forall \ell \neq j}} \omega^{w_2 s_2 + w_3 s_3 + \cdots + w_t s_t + (w_1 + w_{t+1}) s_{t+1}}.$$

In each of the $t$ terms in the final sum, we may relabel the indices $w_1', w_2', \ldots, w_{t+1}'$ such that $w_1' = w_1 + w_\ell$ for $\ell = 1, \ldots, t + 1$. The remaining $w_j'$ are assigned lexicographically according to what is left; that is, if $w_1'$ takes the $\ell$-th spot in the list, then

$$w_2' = w_2, \quad w_3' = w_3, \ldots, w_{\ell-1}' = w_{\ell-1}, \quad w_\ell' = w_{\ell+1}, \ldots, w_{t-1}' = w_t, \quad w_t' = w_{t+1}.$$

These relabeled terms still satisfy $\sum_{u=1}^{t} w_s = 0$ since the original $w$ terms satisfy this relation. They also satisfy $-n < w_1' < 0$ and $0 < w_2', \ldots, w_t' < n$. This is clear for $w_j'$ with $j > 1$, and also holds for $w_1'$ since we have

$$w_1' = w_1 + w_\ell < \sum_{j=1}^{t+1} w_j = 0.$$

So we may apply the induction assumption to each of these sums to turn the last expression in the above manipulation to

$$-\frac{1}{t} \left( \frac{1}{(t-1)!} (-1)^{t-1} (t-1)! \cdot n + \frac{1}{(t-1)!} (-1)^{t-1} (t-1)! \cdot n \right.$$

$$\left. \cdots + \frac{1}{(t-1)!} (-1)^{t-1} (t-1)! \cdot n \right) = (-1)^t \cdot n. \quad \square$$

This lemma is crucial in determining the coefficients we're ultimately looking for. We now present the argument for the coefficient relationships of the rank-$n$ case.

Let $k \geq 1$, $A \in \text{End}(V)[\![z]\!]$ have $A_0$ similar to a Jordan block and have normal form $B$ as in Theorem 2.1 with $b_1 \neq 0$. Letting

$$\zeta(z) = \sum_{m=1}^{\infty} a_m z^{m/n}$$

denote the Puiseux expansion of the eigenvalues of $A$, our aim is to show that the coefficients $\{a_1, \ldots, a_s\}$ determine and are determined by $\{b_1, \ldots, b_s\}$ for arbitrary $1 \leq s \leq kn$. More precisely, writing

$$s = n\ell - t \tag{3-2}$$

for a unique $1 \leq \ell \leq k$ and $0 \leq t \leq n-1$, we have the following.

**Theorem 3.2.** *With the above assumptions, there exist polynomials $P_{s,n} \in \mathbb{C}[x_1, \ldots, x_{s-1}]$ only depending on $s$, $n$ such that we have*

$$b_s = (-1)^t n a_1^t a_s + P_{s,n}(a_1, \ldots, a_{s-1}).$$

*Conversely, there exist rational functions of the form $Q_{s,n} \in \mathbb{C}[x_1^{\pm 1}, \ldots, x_{s-1}]$ such that*

$$a_s = \frac{(-1)^s}{n} b_1^{-s/n} b_s + Q_{s,n}(b_1^{1/n}, \ldots, b_{s-1}).$$

*In particular, for any given $A \in \text{End}(V)[\![z]\!]$ and fixed $k$, the parameters $\{b_1, \ldots, b_{kn}\}$ appearing in Theorem 2.1 are uniquely determined.*

*Proof.* Let $\omega$ be a primitive $n$-th root of unity and recall our notation (1-5) for the eigenvalues of $A$. The key idea is to compare two different representations for the characteristic polynomial

$$\chi_{B(z)}(\zeta) = \chi_{A(z)}(\zeta).$$

Namely, up to order $k$ with respect to the variable $z$, the polynomial $\chi_{B(z)}$ can be read off directly from the form of the matrices $B_0, B_1, \ldots, B_k$ given in Theorem 2.1. On the other hand, as we have seen in Theorem 1.4 we may expand $\chi_{A(z)}$ into linear factors $(\zeta - \zeta_i(z))$. This provides us the identity

$$\zeta^n + \zeta^{n-1}\left(\sum_{\ell=1}^{k} b_{n\ell} z^\ell + O(z^{k+1})\right) + \cdots + \left(\sum_{\ell=1}^{k} b_{n\ell-(n-1)} z^\ell + O(z^{k+1})\right) \tag{3-3}$$

$$= \prod_{i=0}^{n-1}\left(\zeta - \zeta_i(z)\right)$$

$$= \left(\zeta - \sum_{m=1}^{\infty} a_m z^{m/n}\right)\left(\zeta - \sum_{m=1}^{\infty} a_m (\omega z^{1/n})^m\right)\cdots$$

$$\left(\zeta - \sum_{m=1}^{\infty} a_m (\omega^{n-1} z^{1/n})^m\right). \quad (3\text{-}4)$$

The generic term of (3-3) is

$$\zeta^{n-1-t}\left(\sum_{\ell=1}^{k} b_{n\ell-t} z^\ell + O(z^{k+1})\right).$$

We proceed now by comparing coefficients of (3-3) and (3-4), and to do this we apply induction on $s$.

Before starting the induction, we do some preliminary work in computing the coefficient in (3-4) of $\zeta^{n-1-t} z^\ell$, that is, the coefficient that corresponds to $b_{n\ell-t}$ in (3-3). We exclude the case where $\ell = 1$ and $t = n - 1$ (i.e., $b_1$), since this first nonzero term has simpler combinatorial structure than subsequent ones. We would like to have a general form for the subsequent terms.

To this end, we know that the coefficient of $\zeta^{n-1-t} z^\ell$ in (3-4) will be a complex linear combination of the products $a_{m_1} \ldots a_{m_{t+1}}$ such that $\sum_{i=1}^{t+1} m_i = n\ell$, with constants given in terms of a sum of powers of $\omega$. This is equivalent to noticing that the indices $m_i$ partition $n\ell$ into $t+1$ nonempty parts. To explain why there are $t+1$ parts, we first see that $n - 1 - t = n - (t + 1)$, and in the expansion (3-4), each term will have $n$ components. These components are formed by picking one term from each of the $n$ factors in (3-4), and are thus split into those that are just $\zeta$ and those that come from the $a_i$. In the particular case of $\zeta^{n-1-t}$ we can imagine that we use $n - 1 - t$ choices on $\zeta$, and the remaining $t + 1$ choices on various $a_{m_i}$. The correct coefficient in (3-4) to compare to $b_{n\ell-t}$ will be then those combinations of $a_{m_i}$ such that the indices $m_i$ sum to $n$ times the exponent of $z$ multiplying $b_{n\ell-t}$, that is, the $m_i$ sum to $n\ell$. We see that the parts must be nonempty since any $m_i = 0$ would give us a factor of $a_0 = 0$ in the product of all $a_{m_i}$, thus annihilating the product.

So we need to consider the set of all partitions of the integer $n\ell$ as a sum of $t + 1$ positive integers, say in decreasing order:

$$\mathcal{P}_{\ell,t} = \{m_1 \geq \ldots \geq m_{t+1} \geq 1 \mid m_1 + \cdots + m_{t+1} = n\ell\}.$$

With this notation, we can produce an initial expression for the general coefficient:

$$b_{n\ell-t} = \sum_{\mathcal{P}_{\ell,t}} a_{m_1} \ldots a_{m_{t+1}} \mu_{m_1,\ldots,m_{t+1}}, \quad (3\text{-}5)$$

where $\mu_{m_1,\ldots,m_{t+1}}$ denotes a yet undetermined linear combination of powers of $\omega$ with rational coefficients that depends on the partition $(m_1, \ldots, m_{t+1})$.

The expression in (3-5) can be refined by noticing that we care only about the partitions with $m_1 = n\ell - t = s$, since this will be the highest possible index for a given $s$ and given $t$, and the products coming from all partitions with $m_1 < n\ell - t$ will be absorbed in the polynomial $P_{s,n}(a_1^{1/n}, \ldots, a_{s-1})$. This assignment of $m_1$ then necessarily forces $m_2 = \cdots = m_{t+1} = 1$, since we still require that the partition contains $t + 1$ nonempty parts and that the $m_i$ sum to $n\ell$. Let us now introduce

$$\mathcal{P}_{\ell,t}^0 = \{(m_1, \ldots, m_{t+1}) \in \mathcal{P}_{\ell,t} \mid n\ell - t > m_1\}.$$

This $\mathcal{P}_{\ell,t}^0$ captures all of the partitions whose $m_1$ index we do not need to keep track of, allowing us to rewrite (3-5). In rewriting, we suppress the $m_i$ in the first term, instead presenting their actual values which we know to be $m_1 = n\ell - t, m_2 = \cdots = m_{t+1} = 1$:

$$a_1^t a_{n\ell-t} \mu_{n\ell-t,1,\ldots,1} + \sum_{\mathcal{P}_{\ell,t}^0} a_{m_1} \ldots a_{m_{t+1}} \mu_{m_1,\ldots,m_{t+1}}. \tag{3-6}$$

Again, as the indices $m_i$ of each term in the sum are all strictly less than $n\ell - t$, the second term in this formula only contributes to $P_{s,n}$, hence we only need to specify the constants $\mu_{n\ell-t,1,\ldots,1}$.

To gain a better understanding of the structure of the constant $\mu_{n\ell-t,1,\ldots,1}$ appearing in the above expression, we describe a way of visualizing each partition that will give more structure to the enumeration of the constant's summands. Consider the partition of $n\ell$ into parts $n\ell - t, 1, \ldots, 1$ with 1 appearing $t$ times. We align this partition with the combinatorial choice of picking a term out of each of the $n$ factors of (3-4) by considering the $m_i$ to be distributed among $n$ boxes, not necessarily in increasing order. We label the positions of these $m_i$ amongst the $n$ boxes by the labels $s_i$ for $i = 1, \ldots, t+1$, such that $s_i \neq s_j$ for $i \neq j$. Observe however that since $m_2 = \cdots = m_{t+1}$, any fixed set $\{s_2, \ldots, s_{t+1}\}$ of $t$ distinct positions in $\{1, \ldots, n\}$ and any further position $s_1 \notin \{s_2, \ldots, s_{t+1}\}$ give rise to a *single* term in (3-6) of the form $\omega^s a_1^t a_{n\ell-t}$ for some integer $s$ (to be specified below), independently of the order of $\{s_2, \ldots, s_{t+1}\}$. So we may (and from now on, will) assume that the positions $\{s_2, \ldots, s_{t+1}\}$ are in increasing order:

$$s_2 < \cdots < s_{t+1};$$

however, we have no restriction about the position of $s_1$ relative to the above increasing sequence. This gives us a way of picturing all possible configurations of the $m_i$. An example of one of these configurations is
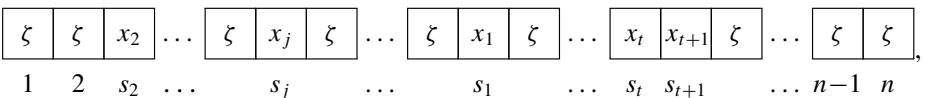
with $x_j = -a_{m_j}(\omega^{s_j-1}z)^{m_j}$ for all $1 \leq j \leq t+1$. We note that the $-1$ attached to each $s_i$ in the exponents occurs since the expansion in (3-4) is indexed from 0 to $n-1$, but we were considering the $s_i$ as elements of $\{1, \ldots, n\}$. This is a minor adjustment.

Computing $\mu_{m_1,\ldots,m_{t+1}}$ involves writing an expression for $\mu$ that reflects the fixing of $s_1$, the position of $m_1$, outside of the strict ordering of the other labels. We express this now, adopting the standard notation $[n] = \{1, \ldots, n\}$:

$$\mu_{n\ell-t,1,\ldots,1} = \sum_{\substack{s_2,\ldots,s_{t+1}\in\mathbb{Z}^+ \\ 1\leq s_2<\cdots<s_{t+1}\leq n}} \omega^{(s_2-1)}\cdots\omega^{(s_{t+1}-1)} \sum_{\substack{s_1\in[n]\setminus \\ \{s_2,\ldots,s_{t+1}\}}} \omega^{(s_1-1)(n\ell-t)}. \qquad (3\text{-}7)$$

Now we manipulate (3-7) as follows, recognizing that since $\omega$ is an $n$-th root of unity, we may work with any of the sums in the exponents modulo $n$:

$$\sum_{\substack{s_2,\ldots,s_{t+1}\in\mathbb{Z}^+ \\ 1\leq s_2<\cdots<s_{t+1}\leq n}} \sum_{\substack{s_1\in[n]\setminus \\ \{s_2,\ldots,s_{t+1}\}}} \omega^{s_2+\cdots+s_{t+1}-t+s_1\ell n-s_1t-\ell n+t}$$

$$= \sum_{\substack{s_2,\ldots,s_{t+1}\in\mathbb{Z}^+ \\ 1\leq s_2<\cdots<s_{t+1}\leq n}} \sum_{s_1\in[n]\setminus\{s_2,\ldots,s_{t+1}\}} \omega^{s_2+\cdots+s_{t+1}-s_1t}$$

$$= \sum_{\substack{s_2,\ldots,s_{t+1}\in\mathbb{Z}^+ \\ 1\leq s_2<\cdots<s_{t+1}\leq n}} \omega^{s_2+\cdots+s_{t+1}} \sum_{s_1\in[n]\setminus\{s_2,\ldots,s_{t+1}\}} \omega^{-s_1t}. \qquad (3\text{-}8)$$

We may recognize (3-8) as an ordered version of the sum examined by Lemma 3.1. Indeed, we have bounded weights that sum to zero and an exponent sum in $t+1$ terms, namely $w_1 = -t$, $w_2 = \cdots = w_{t+1} = 1$. In Lemma 3.1 we have $t+1$ unordered terms, but here we have $t$ ordered terms and one independent term. Multiplying (3-8) by $t!$ allows us to rewrite it without the ordering and allows us to apply the lemma, since we obtain sums over $t+1$ unordered terms. But then the lemma gives that dividing by $t!$ again allows us to compute the sum, and so the sum from the lemma and the sum in (3-8) are equivalent. So we find

$$\sum_{\substack{s_2,\ldots,s_{t+1}\in\mathbb{Z}^+ \\ 1\leq s_2<\cdots<s_{t+1}\leq n}} \omega^{s_2+\cdots+s_{t+1}} \sum_{\substack{s_1\in[n]\setminus \\ \{s_2,\ldots,s_{t+1}\}}} \omega^{-s_1t} = (-1)^t n.$$

We conclude that the leading-index term for $b_{n\ell-t}$ is $(-1)^t n a_1^t a_{n\ell-t}$.

Now we can start the induction on $s$, which will actually be a double induction, first on $\ell \in \{1, \ldots, k\}$ in increasing order then on $t \in \{0, \ldots, n-1\}$ in decreasing order; see (3-2). We determine $b_1$ by inspection, and apply the above argument for $b_2, \ldots, b_n$. So we have

$$b_1 = a_1^n, \quad b_2 = (-1)^{n-2}na_1^{n-2}a_2, \quad \ldots, \quad b_p = (-1)^{n-p}na_1^{n-p}a_p, \quad \ldots, \quad b_n = na_n.$$

We note that each of these $b_i$ relations matches that in the theorem statement, depending on $a_1$ and $a_i$. These relationships are certainly invertible in terms of the $a_i$:

$$a_1 = \sqrt[n]{b_1}, \quad a_2 = \frac{(-1)^{2-n}b_2}{nb_1^{1-2/n}}, \quad \ldots, \quad a_p = \frac{(-1)^{p-n}b_p}{nb_1^{1-p/n}}, \quad \ldots, \quad a_n = b_n/n.$$

We fix an $n$-th root of $b_1$ here so that everything is uniquely determined. Changing the choice of the root is equivalent to multiplying $a_1$ by a primitive $n$-th root of unity, which then affects all subsequent coefficients $a_k$ in the same way, eventually leading to a permutation of the roots $\zeta_j(z)$ in (3-4); thus, fixing an $n$-th root of $b_1$ is not a restrictive choice. Furthermore, we note that in one direction we have the desired polynomial relations, and in the other direction we have the desired rational relations. Thus, the statement holds for $\ell = 1$ and all $t$.

Then supposing that the claim holds for $2, \ldots, s-1$, we consider general $s$. From the earlier partition argument we also know that any terms $a_i$ in the full expression for $b_s$ that do not contain $a_{n\ell-t}$ will have indices at most $i \leq n\ell - t - 1 = s - 1$, so applying the induction hypothesis gives

$$b_{n\ell-t} = (-1)^t na_1^t a_{n\ell-t} + \mathrm{P}_{s,n}(a_1, \ldots, a_{s-1}),$$

since we have invertible relationships for the expressions contained in $\mathrm{P}_{s,n}(a_1, \ldots, a_{s-1})$. This new set of relationships will also be invertible since the only new term is $(-1)^t na_1^t a_{n\ell-t}$, which is a nonzero multiple of $a_{n\ell-t}$ since we are working over a field of characteristic zero with $a_1 \neq 0$. So $b_{n\ell-t}$ is determined explicitly by this expression, and vice versa. Thus we have shown that the claim holds for general $s$. $\square$

### References

[Abhyankar 1990] S. S. Abhyankar, *Algebraic geometry for scientists and engineers*, Mathematical Surveys and Monographs **35**, American Mathematical Society, Providence, RI, 1990. MR Zbl

[Ivanics et al. 2016] P. Ivanics, A. I. Stipsicz, and S. Szabó, "Two-dimensional moduli spaces of irregular Higgs bundles", preprint, 2016. arXiv

[Serre 1979] J.-P. Serre, *Local fields*, Graduate Texts in Mathematics **67**, Springer, New York, 1979. MR Zbl

chkeane@reed.edu    Department of Mathematics, Reed College, 3203 SE Woodstock Blvd, Portland, OR 97202, United States

szabosz@math.bme.hu    Department of Mathematics, Budapest University of Technology and Economics, Egry J. u. 1, H ep., Budapest, 1111, Hungary

# Continuous dependence and differentiating solutions of a second order boundary value problem with average value condition

Jeffrey W. Lyons, Samantha A. Major and Kaitlyn B. Seabrook

(Communicated by Martin J. Bohner)

Using a few conditions, continuous dependence, and a result regarding smoothness of initial conditions, we show that derivatives of solutions to the second order boundary value problem $y'' = f(x, y, y')$, $a < x < b$, satisfying $y(x_1) = y_1$, $1/(d-c)\int_c^d y(x)\,dx = y_2$, where $a < x_1 < c < d < b$ and $y_1, y_2 \in \mathbb{R}$ with respect to each of the boundary data $x_1$, $y_1$, $y_2$, $c$, $d$ solve the associated variational equation with interesting boundary conditions. Of note is the second boundary condition, which is an average value condition.

## 1. Introduction

Our concern is characterizing derivatives of solutions to the second order boundary value problem

$$y'' = f(x, y, y'), \quad a < x < b, \tag{1-1}$$

satisfying

$$y(x_1) = y_1, \quad \frac{1}{d-c}\int_c^d y(x)\,dx = y_2, \tag{1-2}$$

where $a < x_1 < c < d < b$, and $y_1, y_2 \in \mathbb{R}$ with respect to the boundary data. We make note of the average value condition.

The history and breadth of work on the subject of smoothness of conditions for various problems is quite rich and stretches back to the time of Peano as attributed by Hartman [1964]. Peano's result characterized the smoothness of initial conditions for initial value problems (IVPs). Subsequently, many researchers expanded the result to smoothness of boundary conditions for boundary value problems. The key to making the jump was utilizing a continuous dependence result for boundary

conditions. Once invoked, there were many articles published in the realm of boundary value problems for differential equations [Ehme 1993; Ehrke et al. 2007; Henderson 1987; Lyons 2011; Lyons and Miller 2015; Spencer 1975], difference equations [Benchohra et al. 2007; Datta 1998; Henderson and Jiang 2015; Hopkins et al. 2009; Lyons 2014a], and dynamic equations on time scales [Baxter et al. 2016; Lyons 2014b] with a host of interesting of boundary conditions.

Our main motivation for this paper is a recent result [Janson et al. 2014] in which the authors sought an analogue of Peano's theorem for a second order boundary value problem with an integral boundary condition. The novelty we contribute to the literature is employing an average value boundary condition, which, although similar, is very fascinating in its own right.

At first, the average value condition might seem unusual. However, the idea of an average value condition is quite useful when one is not concerned with what occurs at a specific point but instead the average over a range of points. For example, one may not need to specify the temperature at a certain time as long as the average temperature is fixed over a range of time. We point the reader to [Chua 2010] and the references therein for more discussion on average value conditions and more general functional conditions.

The remainder of the paper is organized as follows. In Section 2, we introduce the definition of a variational equation and place conditions upon the boundary value problem. Section 3 is comprised of interesting and crucial results for our research. We prove our main result and a corollary in Section 4.

## 2. Preliminaries

Throughout our work and previous research on the topic, a very important equation emerges which we now define.

**Definition 2.1.** Given a solution $y(x)$ of (1-1), we define the *variational equation along* $y(x)$ by

$$z'' = \frac{\partial f}{\partial u_1}(x, y(x), y'(x))z + \frac{\partial f}{\partial u_2}(x, y(x), y'(x))z', \qquad (2\text{-}1)$$

where $u_1$ and $u_2$ are the second and third components of $f$, respectively.

Next, we place five hypotheses upon the boundary value problem:

(i) $f(x, u_1, u_2) : (a, b) \times \mathbb{R}^2 \to \mathbb{R}$ is continuous.

(ii) For $i = 1, 2$, the map $\partial f / \partial u_i(x, u_1, u_2) : (a, b) \times \mathbb{R}^2 \to \mathbb{R}$ is continuous.

(iii) Solutions of IVPs for (1-1) extend to $(a, b)$.

(iv) Given $a < x_1 < c < d < b$, if $y(x_1) = z(x_1)$ and $1/(d-c)\int_c^d y(x)\,dx = 1/(d-c)\int_c^d z(x)\,dx$, where $y(x)$ and $z(x)$ are solutions of (1-1), then, on $(a, b)$, we have $y(x) \equiv z(x)$.

(v) Given $a < x_1 < c < d < b$ and a solution $y(x)$ of (1-1), if $u(x_1) = 0$ and $1/(d-c)\int_c^d u(x)\,dx = 0$, where $u(x)$ is a solution of (2-1) along $y(x)$, then, on $(a, b)$, we have $u(x) \equiv 0$.

Note that even though (i) and (ii) may seem to be very strict conditions, we remind the reader that since our aim is to compute derivatives of solutions to (1-1) and (1-2), they are not unusual. Condition (iii) is not necessary but instead allows us to suppress verbiage of finding an interval inside $(a, b)$ where the solution of the boundary value problem converges. Finally, conditions (iv) and (v) are required to ensure the uniqueness of the solution and variational equation.

## 3. Background theorems

We now introduce two theorems that play a key role in the proof of the main result. The first result is attributed to Peano and is in essence the type of result we seek for (1-1) and (1-2). We direct the reader to Hartman's book [1964] for more details.

**Theorem 3.1** (Peano's theorem). *Assume that, with respect to (1-1), conditions (i)–(iii) are satisfied. Let $x_0 \in (a, b)$ and $y(x) := y(x, x_0, c_1, c_2)$ denote the solution of (1-1) satisfying the initial conditions $y(x_0) = c_1$ and $y'(x_0) = c_2$. Then:*

(a) *For $i = 1, 2$, $\partial y/\partial c_i(x)$ exists on $(a, b)$, and $\alpha_i(x) := \partial y/\partial c_i(x)$ is the solution of the variational equation (2-1) along $y(x)$ satisfying the respective initial conditions*

$$\alpha_1(x_0) = 1, \quad \alpha_1'(x_0) = 0, \quad \alpha_2(x_0) = 0, \quad \alpha_2'(x_0) = 1.$$

(b) *$\partial y/\partial x_0(x)$ exists on $(a, b)$, and $\beta(x) := \partial y/\partial x_0(x)$ is the solution of the variational equation (2-1) along $y(x)$ satisfying the initial conditions*

$$\beta(x_0) = -y'(x_0), \quad \beta'(x_0) = -y''(x_0).$$

(c) *$\dfrac{\partial y}{\partial x_0}(x) = -y'(x_0)\dfrac{\partial y}{\partial c_1}(x) - y''(x_0)\dfrac{\partial y}{\partial c_2}(x).$*

The next result permits the leap from IVPs to boundary value problems. The proof requires mapping initial data to boundary data and an application of the Brouwer invariance of domain theorem. For a typical proof, we refer the reader to [Henderson et al. 2005].

**Theorem 3.2** (continuous dependence for boundary value problems). *Assume (i)–(iv) are satisfied with respect to (1-1). Let $y(x)$ be a solution of (1-1) on $(a, b)$, and let $a < \alpha < x_1 < c < d < \beta < b$ and $y_1, y_2 \in \mathbb{R}$ be given. Then, there exists a $\delta > 0$ such that, for $|x_1 - t_1| < \delta$, $|c - \xi| < \delta$, $|d - \Delta| < \delta$, $|y(x_1) - y_1| < \delta$, and $\left|1/(d-c)\int_c^d y(x)\,dx - y_2\right| < \delta$, there exists a unique solution $y_\delta(x)$ of (1-1) such that $y_\delta(t_1) = y_1$ and $1/(\Delta - \xi)\int_\xi^\Delta y_\delta(x)\,dx = y_2$ and, for $i = 1, 2$, $\{y_\delta^{(i)}(x)\}$ converges uniformly to $y^{(i)}(x)$ as $\delta \to 0$ on $[\alpha, \beta]$.*

## 4. Main result

In light of the information in the previous sections, we now present the main result. A reminder that the novel portion of our result is differentiation with respect to the terms in the average value condition, namely $c$ and $d$. We will only show the proof of part (d) as (c) is similar. In fact, each part (a)–(d) employs the same idea for a proof.

**Theorem 4.1.** *Assume conditions (i)–(v) are satisfied. Let $y(x)$ be a solution of (1-1) on $(a, b)$. Let $a < x_1 < c < d < b$ and $y_1, y_2 \in \mathbb{R}$ be given so that $y(x) = y(x, x_1, y_1, y_2, c, d)$, where*

$$y(x_1) = y_1, \quad \frac{1}{d-c} \int_c^d y(x) \, dx = y_2.$$

*Then*:

(a) *For $i = 1, 2$, $u_i(x) := \partial y / \partial y_i(x)$ exists on $(a, b)$ and is the solution of the variational equation (2-1) along $y(x)$ satisfying the respective boundary conditions*

$$u_1(x_1) = 1 \quad and \quad \frac{1}{d-c} \int_c^d u_1(x) \, dx = 0,$$

$$u_2(x_1) = 0 \quad and \quad \frac{1}{d-c} \int_c^d u_2(x) \, dx = 1.$$

(b) *$z_1(x) := \partial y / \partial x_1(x)$ exists on $(a, b)$ and is the solution of the variational equation (2-1) along $y(x)$ satisfying the respective boundary conditions*

$$z_1(x_1) = -y'(x_1) \quad and \quad \frac{1}{d-c} \int_c^d z_1(x) \, dx = 0.$$

(c) *$C(x) := \partial y / \partial c(x)$ exists on $(a, b)$ and is the solution of the variational equation (2-1) along $y(x)$ satisfying the boundary conditions*

$$C(x_1) = 0 \quad and \quad \frac{1}{d-c} \int_c^d C(x) \, dx = \frac{y(c) - y_2}{d - c}.$$

(d) *$D(x) := \partial y / \partial d(x)$ exists on $(a, b)$ and is the solution of the variational equation (2-1) along $y(x)$ satisfying the boundary conditions*

$$D(x_1) = 0 \quad and \quad \frac{1}{d-c} \int_c^d D(x) \, dx = \frac{y_2 - y(d)}{d - c}.$$

*Proof.* Since only $x$ and $d$ are not fixed, we denote $y(x, x_1, y_1, y_2, c, d)$ by $y(x, d)$.

Let $\delta > 0$ be as in Theorem 3.2, $0 < |h| < \delta$ be given, and define the difference quotient

$$D_h(x) = \frac{1}{h}[y(x, d + h) - y(x, d)].$$

Our goal is to show that the limit of $D_h$ exists, solves the variational equation, and satisfies the correct boundary conditions. First, we investigate the boundary conditions.

For every $h \neq 0$,

$$D_h(x_1) = \frac{1}{h}[y(x_1, d+h) - y(x_1, d)] = \frac{1}{h}[y_1 - y_1] = 0,$$

and by using the mean value theorem for integrals,

$$
\begin{aligned}
\frac{1}{d-c} \int_c^d D_h(x)\,dx &= \frac{1}{d-c} \int_c^d \frac{y(x,d+h)-y(x,d)}{h}\,dx \\
&= \frac{1}{d-c} \int_c^d \frac{y(x,d+h)}{h}\,dx - \frac{1}{d-c} \int_c^d \frac{y(x,d)}{h}\,dx \\
&= \frac{1}{d-c}\left[ \int_c^{d+h} \frac{y(x,d+h)}{h}\,dx + \int_{d+h}^d \frac{y(x,d+h)}{h}\,dx \right] - \frac{y_2}{h} \\
&= \frac{1}{d-c}\frac{(d+h)-c}{(d+h)-c} \int_c^{d+h} \frac{y(x,d+h)}{h}\,dx + \frac{y(e)(d-(d+h))}{h(d-c)} - \frac{y_2}{h} \\
&= \frac{((d+h)-c)y_2}{h(d-c)} - \frac{y(e)}{d-c} - \frac{y_2(d-c)}{h(d-c)} = \frac{y_2-y(e)}{d-c}
\end{aligned}
$$

for some $e$ between $d$ and $d+h$.

Next, we view $y(x)$ in terms of the solution of an IVP at $x_1$ so that we may employ Theorem 3.1.

To that end, let

$$\mu = y'(x_1, d) \quad \text{and} \quad v = v(h) = y'(x_1, d+h) - \mu.$$

Then, in terms of an IVP,

$$y(x) = u(x, x_1, y_1, \mu),$$

and we have

$$D_h(x) = \frac{1}{h}[u(x, x_1, y_1, \mu + v) - u(x, x_1, y_1, \mu)].$$

By Theorem 3.1 and the mean value theorem, we obtain

$$D_h(x) = \frac{1}{h}[\alpha_2(x, u(x, x_1, y_1, \mu + \bar{v}))(\mu + v - \mu)],$$

where $\alpha_2(x, u(\cdot))$ is the solution of (1-1) along $u(\cdot)$ satisfying

$$\alpha_2(x_1) = 0, \quad \alpha_2'(x_1) = 1.$$

Furthermore, $\mu + \bar{v}$ is between $\mu$ and $\mu + v$. Simplifying,

$$D_h(x) = \frac{v}{h}\alpha_2(x, u(x, x_1, y_1, \mu + \bar{v})).$$

Thus, to show $\lim_{h\to 0} D_h(x)$ exists, it suffices to show $\lim_{h\to 0} v/h$ exists. By condition (v), the fact that $\alpha_2(x, u(\,\cdot\,))$ is a nontrivial solution of (2-1) along $u(\,\cdot\,)$ and $\alpha_2(x_1, u(\,\cdot\,)) = 0$, we have

$$\frac{1}{d-c} \int_c^d \alpha_2(x, u(\,\cdot\,))\, dx \neq 0 \implies \int_c^d \alpha_2(x, u(\,\cdot\,))\, dx \neq 0.$$

Recall,

$$\frac{1}{d-c} \int_c^d D_h(x)\, dx = \frac{y_2 - y(e)}{d-c},$$

and so,

$$\frac{1}{d-c} \int_c^d \frac{v}{h} \alpha_2(x, u(x, x_1, y_1, \mu + \bar{v}))\, dx = \frac{y_2 - y(e)}{d-c}.$$

Hence, we obtain

$$\lim_{h\to 0} \frac{v}{h} = \frac{(y_2 - y(e))}{(d-c)} \frac{1}{1/(d-c) \int_c^d \alpha_2(x, u(\,\cdot\,))\, dx} = \frac{y_2 - y(e)}{\int_c^d \alpha_2(x, u(\,\cdot\,))\, dx} := U.$$

Now let

$$D(x) = \lim_{h\to 0} D_h(x),$$

and note by construction of $D_h(x)$,

$$D(x) = \frac{\partial y}{\partial d}(x).$$

Furthermore,

$$D(x) = \lim_{h\to 0} D_h(x) = U\alpha_2(x, y(x)),$$

which is a solution of the variational equation (2-1) along $y(x)$. In addition,

$$D(x_1) = \lim_{h\to 0} D_h(x_1) = \lim_{h\to 0} 0 = 0,$$

and

$$\frac{1}{d-c} \int_c^d D(x)\, dx = \lim_{h\to 0}\left[\frac{1}{d-c} \int_c^d D_h(x)\, dx\right] = \lim_{h\to 0} \frac{y_2 - y(e)}{d-c} = \frac{y_2 - y(d)}{d-c}. \quad \square$$

Finally, we present an analogue to (c) of Theorem 3.1 (Peano's theorem).

**Corollary 4.2.** *Under the assumptions of the previous theorem, we have*

(a) $z_1(x) = -y'(x_1)u_1(x),$

(b) $C(x) = -\dfrac{y_2 - y(c)}{y_2 - y(d)} D(x),$

(c) $C(x) = \dfrac{y(c) - y_2}{d-c} u_2(x).$

# References

[Baxter et al. 2016] L. H. Baxter, J. W. Lyons, and J. T. Neugebauer, "Differentiating solutions of a boundary value problem on a time scale", *Bull. Aust. Math. Soc.* **94**:1 (2016), 101–109. MR Zbl

[Benchohra et al. 2007] M. Benchohra, S. Hamani, J. Henderson, S. K. Ntouyas, and A. Ouahab, "Differentiation and differences for solutions of nonlocal boundary value problems for second order difference equations", *Int. J. Difference Equ.* **2**:1 (2007), 37–47. MR Zbl

[Chua 2010] S.-K. Chua, "Average value problems in ordinary differential equations", *J. Differential Equations* **249**:7 (2010), 1531–1548. MR Zbl

[Datta 1998] A. Datta, "Differences with respect to boundary points for right focal boundary conditions", *J. Differ. Equations Appl.* **4**:6 (1998), 571–578. MR Zbl

[Ehme 1993] J. A. Ehme, "Differentiation of solutions of boundary value problems with respect to nonlinear boundary conditions", *J. Differential Equations* **101**:1 (1993), 139–147. MR Zbl

[Ehrke et al. 2007] J. Ehrke, J. Henderson, C. Kunkel, and Q. Sheng, "Boundary data smoothness for solutions of nonlocal boundary value problems for second order differential equations", *J. Math. Anal. Appl.* **333**:1 (2007), 191–203. MR Zbl

[Hartman 1964] P. Hartman, *Ordinary differential equations*, Wiley, New York, 1964. MR Zbl

[Henderson 1987] J. Henderson, "Disconjugacy, disfocality, and differentiation with respect to boundary conditions", *J. Math. Anal. Appl.* **121**:1 (1987), 1–9. MR Zbl

[Henderson and Jiang 2015] J. Henderson and X. Jiang, "Differentiation with respect to parameters of solutions of nonlocal boundary value problems for difference equations", *Involve* **8**:4 (2015), 629–636. MR Zbl

[Henderson et al. 2005] J. Henderson, B. Karna, and C. C. Tisdell, "Existence of solutions for three-point boundary value problems for second order equations", *Proc. Amer. Math. Soc.* **133**:5 (2005), 1365–1369. MR Zbl

[Hopkins et al. 2009] B. Hopkins, E. Kim, J. Lyons, and K. Speer, "Boundary data smoothness for solutions of nonlocal boundary value problems for second order difference equations", *Comm. Appl. Nonlinear Anal.* **16**:2 (2009), 1–12. MR Zbl

[Janson et al. 2014] A. F. Janson, B. T. Juman, and J. W. Lyons, "The connection between variational equations and solutions of second order nonlocal integral boundary value problems", *Dynam. Systems Appl.* **23**:2-3 (2014), 493–503. MR Zbl

[Lyons 2011] J. W. Lyons, "Differentiation of solutions of nonlocal boundary value problems with respect to boundary data", *Electron. J. Qual. Theory Differ. Equ.* (2011), art. id. 51, 11 pp. MR Zbl

[Lyons 2014a] J. W. Lyons, "Disconjugacy, differences and differentiation for solutions of non-local boundary value problems for *n*th order difference equations", *J. Difference Equ. Appl.* **20**:2 (2014), 296–311. MR Zbl

[Lyons 2014b] J. W. Lyons, "On differentiation of solutions of boundary value problems for second order dynamic equations on a time scale", *Commun. Appl. Anal.* **18**:1-2 (2014), 215–224. Zbl

[Lyons and Miller 2015] J. W. Lyons and J. K. Miller, "The derivative of a solution to a second order parameter dependent boundary value problem with a nonlocal integral boundary condition", *J. Math. Stat. Sci.* **1**:2 (2015), 43–50.

[Spencer 1975] J. D. Spencer, "Relations between boundary value functions for a nonlinear differential equation and its variational equations", *Canad. Math. Bull.* **18**:2 (1975), 269–276. MR Zbl

jlyons@nova.edu                 *Department of Mathematics, Nova Southeastern University,*
                                *3301 College Ave, Fort Lauderdale, FL 33314, United States*

sm2791@nova.edu                 *Department of Mathematics, Nova Southeastern University,*
                                *3301 College Ave, Fort Lauderdale, FL 33314, United States*

ks1679@nova.edu                 *Department of Mathematics, Nova Southeastern University,*
                                *3301 College Ave, Fort Lauderdale, FL 33314, United States*

# On uniform large-scale volume growth for the Carnot–Carathéodory metric on unbounded model hypersurfaces in $\mathbb{C}^2$

Ethan Dlugie and Aaron Peterson

(Communicated by Michael Dorff)

We consider the rate of volume growth of large Carnot–Carathéodory metric balls on a class of unbounded model hypersurfaces in $\mathbb{C}^2$. When the hypersurface has a uniform global structure, we show that a metric ball of radius $\delta \gg 1$ either has volume on the order of $\delta^3$ or $\delta^4$. We also give necessary and sufficient conditions on the hypersurface to display either behavior.

## 1. Introduction

The study of holomorphic functions on pseudoconvex domains $\Omega \subseteq \mathbb{C}^n$ ($n \geq 2$) often reduces to studying the partial differential operator $\bar{\partial}$ on $\Omega$ given by $\bar{\partial}(f) = \sum f_{\bar{z}_j} \, d\bar{z}^j$. We can study the boundary values of holomorphic functions (on b$\Omega$) by studying the partial differential operator $\bar{\partial}_b$ induced on b$\Omega$ by $\bar{\partial}$. We locally express $\bar{\partial}_b$ in terms of differentiation with respect to $(n-1)$-antiholomorphic vector fields (the so-called Cauchy–Riemann, or CR, vector fields on b$\Omega$) that are tangent to b$\Omega$. Under mild nondegeneracy conditions on b$\Omega$ we can access a family of metrics on b$\Omega$ specifically adapted to the study of $\bar{\partial}$ and $\bar{\partial}_b$, in the sense that they capture important geometric aspects of b$\Omega$. One of these, the Carnot–Carathéodory (CC) metric $d(\boldsymbol{p}, \boldsymbol{q})$, measures the infimal length of paths on b$\Omega$ that not only connect the points $\boldsymbol{p}$ and $\boldsymbol{q}$, but are also almost-everywhere tangent to the real and imaginary parts of the CR vector fields; see [Street 2014] for an extensive history of this metric and its applications to the study of $\bar{\partial}$ and $\bar{\partial}_b$.

In this paper we consider the CC metric $d(\boldsymbol{p}, \boldsymbol{q})$ induced on the boundary of a model pseudoconvex domain $\Omega \subset \mathbb{C}^2$ by the real and imaginary parts of the CR vector field on b$\Omega$. In particular, we seek to understand the volume growth of the metric balls $B_d(\boldsymbol{p}, \delta)$ when $\Omega$ is of the form

$$\Omega = \{(z_1, z_2) \in \mathbb{C}^2 : \mathrm{Im}(z_2) > P(z_1)\},$$

where $P : \mathbb{C} \to \mathbb{R}$ is smooth, subharmonic, and nonharmonic. Under mild nonde-generacy conditions on $\Delta P$ is it known [Montanari and Morbidelli 2012; Nagel et al. 1985; 1988; 1989] that for $\delta \leq 1$ the metric ball $B_d(\boldsymbol{p}, \delta)$ is comparable to a "shorn" or "twisted" ellipsoid with radius $\delta$ in the directions spanned by the real and imaginary parts of the CR vector field and radius $\Lambda((z_1, z_2), \delta)$ in the $\mathrm{Re}(z_2)$-direction. If we equip $b\Omega$ with the Lebesgue measure $dm(z, t)$ that it receives via its identification with $\mathbb{C} \times \mathbb{R}$ given by $(z_1, z_2) \mapsto (z, t)$, where $z = z_1 = x + iy$ and $t = \mathrm{Re}(z_2)$, then this small CC metric ball has volume comparable to that of the twisted ellipsoid:

$$\mathrm{Vol}(B_d(\boldsymbol{p}, \delta)) \approx \delta^2 \Lambda(\boldsymbol{p}, \delta). \tag{1-1}$$

We build on the earlier work of the second author [Peterson 2014] which sought to understand the possible rate of growth of $\mathrm{Vol}(B_d(\boldsymbol{p}, \delta))$ for model domains $\Omega$ such that when $\delta$ is large, the Euclidean radius

$$\Lambda((z_1, z_2), \delta) = \sup\{|\mathrm{Re}(z_2' - z_2)| : d((z_1, z_2), (z_1, z_2')) < \delta\}$$

of $B_d((z_1, z_2), \delta)$ in the $\mathrm{Re}(z_2)$-direction is essentially independent of $(z_1, z_2)$. The quantity $\Lambda(\boldsymbol{p}, \delta)$ is called the *global structure* of $b\Omega$, and we make precise the $(z_1, z_2)$-independence condition described above with the following definition.

**Definition 1.1.** If there exists $\delta_0 > 0$, a function $f : [\delta_0, +\infty) \to [0, +\infty)$, and positive constants $0 < c < C < +\infty$ such that $cf(\delta) \leq \Lambda(\boldsymbol{p}, \delta) \leq Cf(\delta)$ for all $\delta \geq \delta_0$ and $\boldsymbol{p} \in b\Omega$, then we say that $(f(\delta), \delta_0)$ is a *uniform global structure* or UGS for $b\Omega$.

For such domains $\Omega$ we also have (1-1) when $\delta$ is large (see Remark 3.3), and therefore the volume growth of CC metric balls of any size is completely understood once we understand $\Lambda(\boldsymbol{p}, \delta)$ for large $\delta$.

**Example 1.2.** In [Nagel et al. 1988], it is shown that when $P(z_1)$ is a subharmonic, nonharmonic polynomial (and where $\Delta P$ has degree $m - 2$),

$$\Lambda((z_1, z_2), \delta) \approx \sum_{k=0}^{m-2} \left( \sum_{\alpha=0}^{k} \left| \frac{\partial^k \Delta P}{\partial z_1^\alpha \partial \bar{z}_1^{k-\alpha}}(z_1) \right| \right) \delta^{k+2}.$$

In particular, when $P(z_1) = |z_1|^2$ (so that $\Delta P(z_1) \equiv 4$) we have $\Lambda((z_1, z_2), \delta) \approx 4\delta^2$, and therefore $(\delta^2, 1)$ is a uniform global structure for $b\Omega$.

On the other hand, if $P(z_1) = |z_1|^4$, then $\Lambda((z_1, z_2), \delta) \approx |z_1|^2 \delta^2 + |z_1|\delta^3 + \delta^4 \approx (|z_1| + \delta)^2 \delta^2$, and therefore is not uniform in $z_1 \in \mathbb{C}$. This shows that $b\Omega$ has no uniform global structure. More generally, if $P$ is a subharmonic, nonharmonic polynomial, then $b\Omega$ does not have a uniform global structure when $\Delta P$ is not constant.

The following result from [Peterson 2014] controls the growth of uniform global structures.

**Theorem 1.3** [Peterson 2014, Theorem 1.2]. *If* b$\Omega$ *has a UGS* $(f(\delta), \delta_0)$, *then there are positive constants* $0 < c < C < +\infty$ *such that* $c\delta \le f(\delta) \le C\delta^2$ *for all* $\delta \ge \delta_0$.

So when b$\Omega$ has a UGS and $\delta \gg 1$, the global structure at any point grows at least linearly and at most quadratically in $\delta$. Examples are given in [Peterson 2014] where b$\Omega$ has a UGS linear in $\delta$ and quadratic in $\delta$. Our question is whether there exist examples where the UGS grows somewhere "between" linear and quadratic. For instance, are there examples for b$\Omega$ with UGS $(\delta^{3/2}, \delta_0)$ or $(\delta \log \delta, \delta_0)$?

**Example 1.4.** To see that this question is not trivial, fix $\alpha \in \left(0, \frac{2}{3}\right)$ and choose a subharmonic function $P : \mathbb{C} \to \mathbb{R}$ such that $\Delta P(z) = (1 + |z|^2)^{-\alpha/2}$. Using our techniques and those of [Peterson 2014] one can show that there exist constants $0 < c < C < +\infty$ such that for all $\delta > 0$,

$$c\delta^{2-\alpha} \le \Lambda((0,0), \delta) \le C\delta^{2-\alpha} \quad \text{and} \quad \Lambda((\delta^{3/2}, 0), \delta) \le C\delta^{2-3\alpha/2}.$$

Thus $\Lambda((0,0), \delta)$ grows at a rate comparable to $\delta^{2-\alpha}$, but $\Lambda((\delta^{3/2}, 0), \delta)$ grows no faster than $\delta^{2-3\alpha/2}$. This illustrates that it is possible for the global structure to grow (in $\delta$) at nonpolynomial rates, but $\left(\text{since } \alpha < \frac{3}{2}\alpha\right)$ not necessarily uniformly in the base point $(z_1, z_2)$.

Our first main theorem (proven in Section 4) answers our question negatively.

**Theorem 1.5.** *If* b$\Omega$ *has UGS* $(f(\delta), \delta_0)$, *then either* $(\delta^2, \delta^*)$ *or* $(\delta, \delta^*)$ *is a UGS for* b$\Omega$ *for some* $\delta^* > 0$.

We subsequently give necessary and sufficient conditions on b$\Omega$ for both linear (Theorem 5.1) and quadratic (Theorem 5.2) growth of the UGS, thereby completely describing the conditions under which any particular model domain has a uniform global structure.

The volume growth of CC metric balls in model domains $\Omega$ as above for large $\delta$ is only explicitly understood when $P$ is a subharmonic, nonharmonic polynomial [Nagel et al. 1988] or in the limited examples considered in [Peterson 2014] mentioned above. In some situations one can obtain upper bounds for the rate of volume growth (see [Chang and Chang 2014]), but one cannot hope for precise control of $\text{Vol}(B_d(\boldsymbol{p}, \delta))$ for general $P$. On the other hand, applications of volume growth estimates are many and varied; for example, one can use these estimates to identify spaces of homogeneous type [Coifman and Weiss 1977], study singular integral operators [Stein 1993], and even to decide whether or not the boundaries of two model domains are quasiconformally equivalent [Fässler et al. 2015; Heinonen and Koskela 1998].

Our paper is structured as follows: Section 2 gives relevant definitions and notation that will be used extensively throughout the paper and recalls past results. In Section 3 we gain some intuition about how a UGS behaves and prove a key and

explicit alternative characterization of the UGS. In Section 4 we prove Theorem 1.5, followed in Section 5 by necessary and sufficient conditions for a given model domain to possess a uniform global structure. Section 6 concludes the paper and offers future directions of study.

## 2. Preliminaries

With $\Omega$ as in the Introduction, the space of tangential CR vector fields on $b\Omega$ is spanned by

$$\overline{Z} = 2\frac{\partial}{\partial \overline{z}_1} - 4i\, P_{\overline{z}_1}(z_1)\frac{\partial}{\partial \overline{z}_2}.$$

We identify $b\Omega$ with $\mathbb{C} \times \mathbb{R}$ via the diffeomorphism $(z_1, z_2) \mapsto (z, t) \in \mathbb{C} \times \mathbb{R}$, where $z = z_1 = x + iy$ and $t = \mathrm{Re}(z_2)$. Under this transformation, $\overline{Z}$ becomes

$$\overline{Z} = 2\frac{\partial}{\partial \overline{z}} - 2i\, P_{\overline{z}}(z)\frac{\partial}{\partial t} = \left(\frac{\partial}{\partial x} + P_y(x, y)\frac{\partial}{\partial t}\right) - i\left(-\frac{\partial}{\partial y} + P_x(x, y)\frac{\partial}{\partial t}\right) \overset{\text{def}}{=} X - iY.$$

As stated in Introduction, we give $b\Omega$ the Lebesgue measure $dm(z, t)$ that it receives upon identification with $\mathbb{C} \times \mathbb{R}$. For the rest of the paper, we work on $\mathbb{C} \times \mathbb{R}$ instead of $b\Omega$ to simplify notation.

We define the CC metric $d : (\mathbb{C} \times \mathbb{R}) \times (\mathbb{C} \times \mathbb{R}) \to [0, +\infty)$ by

$$\begin{aligned}
d(\boldsymbol{p}, \boldsymbol{q}) = \inf\big\{\delta > 0 : \exists\, \gamma : [0, 1] &\to \mathbb{C} \times \mathbb{R},\ \gamma(0) = \boldsymbol{p},\ \gamma(1) = \boldsymbol{q}, \\
&\gamma'(s) = \delta\alpha(s)X(\gamma(s)) + \delta\beta(s)Y(\gamma(s))\ \text{a.e.}, \\
&\alpha, \beta \in \mathrm{FPWS}[0, 1],\ |\alpha(s)|^2 + |\beta(s)|^2 < 1\ \text{a.e.}\big\}. \quad (2\text{-}1)
\end{aligned}$$

Here FPWS[0, 1] (read "finite piecewise smooth") denotes the set of functions $f : [0, 1] \to \mathbb{R}$ which are smooth except at a finite number of points and whose derivatives extend continuously to those points from each side separately.

The global structure $\Lambda((z, t), \delta)$, the radius in the $t$-direction of the CC ball, is then defined as

$$\Lambda((z, t), \delta) \overset{\text{def}}{=} \sup\big\{|t' - t| : d((z, t), (z, t')) < \delta\big\}. \quad (2\text{-}2)$$

Note that the quantity (2-2) is actually *independent* of the $t$-coordinate because the solutions to the differential equation in (2-1) are translation invariant in $t$. To simplify notation, we will therefore write $\Lambda(z, \delta)$ instead of $\Lambda((z, t), \delta)$ for the remainder of the paper, treating $\Lambda$ as a function from $\mathbb{C} \times (0, +\infty) \mapsto [0, +\infty)$. The first observation of [Peterson 2014] is that definition (2-2) is in fact equivalent to the following statement in terms of curves in $\mathbb{C}$, independent of $t$:

$$\begin{aligned}
\Lambda(z, \delta) = \sup\big\{\textstyle\oint_\gamma P_y\, dx - P_x\, dy : \gamma : [0, 1] &\to \mathbb{C}, \\
&\gamma(0) = \gamma(1) = z,\ |\gamma'(s)| \le \delta\ \text{a.e.}, \\
&\gamma'(s) = \alpha(s) + i\beta(s),\ \alpha, \beta \in \mathrm{FPWS}[0, 1]\big\}. \quad (2\text{-}3)
\end{aligned}$$

We write $L(\gamma) = \int_a^b |\gamma'(s)|\,ds$ for the usual Euclidean length of a piecewise smooth curve $\gamma : [a, b] \to \mathbb{C}$. The following geometric definition from [Peterson 2014] will be essential to our understanding of global structures.

**Definition 2.1.** We say $A \subset \mathbb{C}$ is a *pen* if $A$ is open, connected, simply connected, and b$A$ can be parametrized by a continuous piecewise smooth curve $\gamma : [0, 1] \to \mathbb{C}$ with $\gamma'(s) = \alpha(s) + i\beta(s)$, where $\alpha, \beta \in \text{FPWS}[0, 1]$. We call $L(bA) = L(\gamma)$ the amount of *fencing* used to enclose $A$. For a fixed $z \in \mathbb{C}$ and $\delta > 0$, we say that a finite collection of pens $R = (R_1, \ldots, R_N)$ is a $(z, \delta)$-*stockyard* if

$$z \in \bigcup_{i=1}^N bR_i, \quad \sum_{i=1}^N L(bR_i) \leq \delta, \quad \text{and} \quad \bigcup_{i=1}^N bR_i \text{ is connected.}$$

**Remark 2.2.** We will often use the fact that given a pen $A$, we have $A \subseteq B(z, L(bA))$ for any point $z \in A$, where $B(z, \rho)$ denotes the open Euclidean disc in $\mathbb{C}$ of radius $\rho$ centered at $z$.

Thinking of global structures in terms of (2-3), [Peterson 2014] provides the following theorem.

**Theorem 2.3** [Peterson 2014, Theorem 1.1].

$$\Lambda(z, \delta) = \sup_{(z,\delta)-stockyards\ R} \sum_{R_i \in R} \int_{R_i} \Delta P(w)\,dm(w).$$

Here $dm(\cdot)$ denotes the Lebesgue measure on $\mathbb{C}$. The problem of calculating the global structure, an inherently three-dimensional problem, is therefore reduced to a question in two dimensions. Furthermore, notice that because $P$ was assumed to be subharmonic and nonharmonic, we can think of $\Delta P$ as a density function in the plane. In this context, integration over a pen measures the "mass" of the region covered by the pen, and integration over a stockyard is then the sum of the mass collected by the individual pens. The global structure $\Lambda(z, \delta)$ is then just the most mass one can collect with a stockyard touching $z$ constructed with at most $\delta$ amount of fencing.

We introduce the following simpler notation for use in our estimates. For two nonnegative quantities $A$ and $B$, we write $A \lesssim B$ (read "$A$ is controlled above by $B$") if there exists some constant $c > 0$, independent of all relevant quantities, such that $A \leq cB$. We say $A \gtrsim B$ (read "$A$ is controlled below by $B$") if $B \lesssim A$, and $A \approx B$ (read "$A$ is comparable to $B$") if both $A \lesssim B$ and $B \lesssim A$.

## 3. Alternate description of uniform global structures

When b$\Omega$ has a UGS $(f(\delta), \delta_0)$ and when $\delta \geq \delta_0$, we expect that for every point $z$ in the plane we can find a high density region whose distance from the point is

no more than $\delta$. We should then be able to construct a $(z, N\delta)$-stockyard for an appropriately fixed natural number $N$ which covers this region with one or more pens. Otherwise $\Lambda(z, \delta)$ would be uncontrollably small at certain points. We also expect that no point should be within $\delta$ of a region of exceedingly high density. Otherwise $\Lambda(z, \delta)$ would be uncontrollably large at certain points. Before we make this notion precise in Proposition 3.4 of this section, we need two lemmas.

A simple observation about one formula for a UGS is the following.

**Lemma 3.1.** *If* $b\Omega$ *has UGS* $(f(\delta), \delta_0)$*, then* $(\sup_{z \in \mathbb{C}} \Lambda(z, \delta), \delta_0)$ *is also a UGS for* $b\Omega$*.*

*Proof.* Fix some $z \in \mathbb{C}$. By the definition of UGS, there exist constants $c, C > 0$ independent of $z$ and $\delta$ such that

$$cf(\delta) \leq \Lambda(z, \delta) \leq Cf(\delta).$$

So $Cf(\delta)$ is an upper bound for $\{\Lambda(z, \delta) : z \in \mathbb{C}\}$, which gives $\sup_{z \in \mathbb{C}} \Lambda(z, \delta) \leq Cf(\delta)$ since the supremum is the least upper bound. Also $\sup_{z \in \mathbb{C}} \Lambda(z, \delta) \geq \Lambda(z, \delta) \geq cf(\delta)$. So then

$$\Lambda(z, \delta) \leq Cf(\delta) \leq \frac{C}{c} \sup_{z \in \mathbb{C}} \Lambda(z, \delta) \quad \text{and} \quad \Lambda(z, \delta) \geq cf(\delta) \geq \frac{c}{C} \sup_{z \in \mathbb{C}} \Lambda(z, \delta)$$

for all $\delta \geq \delta_0$. Therefore $(\sup_{z \in \mathbb{C}} \Lambda(z, \delta), \delta_0)$ is a UGS for $b\Omega$. $\qquad\square$

Lemma 3.1 makes it clear that we can take $f(\delta)$ to be a monotonically increasing function of $\delta$. We next show that $f(\delta)$ does not increase too quickly in the sense that if we double the amount of fencing available to construct stockyards, then the amount of mass one can collect should not grow exceedingly fast.

**Lemma 3.2.** *If* $b\Omega$ *has UGS* $(f(\delta), \delta_0)$ *then* $f(\delta) \approx f(2\delta)$ *for all* $\delta \geq \delta_0$*, with constants independent of* $\delta$*.*

*Proof.* By Lemma 3.1 we can without loss of generality take $f(\delta) = \sup_{z \in \mathbb{C}} \Lambda(z, \delta)$. For if $(g(\delta), \delta_0)$ is any other UGS for $b\Omega$ and we can prove the lemma for $f(\delta)$, then $g(\delta) \approx f(\delta) \approx f(2\delta) \approx g(2\delta)$. We prove first that $f(2\delta) \approx f(3\delta)$ for large $\delta$ and will show at the end of the proof that this is sufficient to establish the lemma.

Because $f(\delta)$ is a nondecreasing function, we trivially have $f(2\delta) \leq f(3\delta)$. We need only show then that $f(3\delta) \lesssim f(2\delta)$. To this end, fix $z_0 \in \mathbb{C}$ and $\delta \geq \frac{2}{3}\delta_0$, and let $R$ be any arbitrary $(z_0, 3\delta)$-stockyard. There is a FPWS curve $\gamma : [0, 1] \to \mathbb{C}$ with $\gamma(0) = \gamma(1) = z_0$, $L(\gamma) \leq 3\delta$, and

$$\sum_{R_i \in R} \int_{R_i} \Delta P(w) \, dm(w) = \oint_\gamma P_y \, dx - P_x \, dy.$$

We now produce seven continuous, piecewise smooth curves $\gamma_k : [0, 1] \to \mathbb{C}$, $k = 1, \ldots, 7$, with $L(\gamma_k) \leq 2\delta$ and $\gamma_k'(s) = \alpha_k(s) + i\beta_k(s)$ with $\alpha_k, \beta_k \in \text{FPWS}[0, 1]$

such that

$$\oint_\gamma P_y\,dx - P_x\,dy = \sum_{k=1}^{7} \oint_{\gamma_k} P_y\,dx - P_x\,dy.$$

Without loss of generality, suppose that $\gamma$ has constant speed so that

$$\int_0^{1/3} |\gamma'(s)|\,ds = \int_{1/3}^{2/3} |\gamma'(s)|\,ds = \int_{2/3}^1 |\gamma'(s)|\,ds \le \delta. \tag{3-1}$$

For convenience, we define $z_1 = \gamma\!\left(\tfrac{1}{3}\right)$, $z_2 = \gamma\!\left(\tfrac{2}{3}\right)$, and $z_3 = \gamma(1) = z_0$. We also denote by $\overrightarrow{z,w}$ the directed line segment from $z$ to $w$.

Now we have

$$\oint_\gamma P_y\,dx - P_x\,dy$$

$$= \int_{\gamma[0,1/3]} P_y\,dx - P_x\,dy + \int_{\gamma[1/3,2/3]} P_y\,dx - P_x\,dy + \int_{\gamma[2/3,1]} P_y\,dx - P_x\,dy$$

$$+ \int_{\overrightarrow{z_0,z_1}} P_y\,dx - P_x\,dy + \int_{\overrightarrow{z_1,z_2}} P_y\,dx - P_x\,dy + \int_{\overrightarrow{z_2,z_3}} P_y\,dx - P_x\,dy$$

$$+ \int_{\overrightarrow{z_1,z_0}} P_y\,dx - P_x\,dy + \int_{\overrightarrow{z_2,z_1}} P_y\,dx - P_x\,dy + \int_{\overrightarrow{z_3,z_2}} P_y\,dx - P_x\,dy$$

$$= \oint_{\gamma[0,1/3]+\overrightarrow{z_1,z_0}} P_y\,dx - P_x\,dy + \oint_{\gamma[1/3,2/3]+\overrightarrow{z_2,z_1}} P_y\,dx - P_x\,dy$$

$$+ \oint_{\gamma[2/3,1]+\overrightarrow{z_3,z_2}} P_y\,dx - P_x\,dy + \oint_{\overrightarrow{z_0,z_1}+\overrightarrow{z_1,z_2}+\overrightarrow{z_2,z_3}} P_y\,dx - P_x\,dy. \tag{3-2}$$

We consider the contours of integration in each integral.

We define $\gamma_i = \gamma\!\left[\tfrac{1}{3}(i-1), \tfrac{1}{3}i\right] + \overrightarrow{z_i, z_{i-1}}$ for $i = 1, 2, 3$. By (3-1), the length of each contour $\gamma\!\left[\tfrac{1}{3}(i-1), \tfrac{i}{3}\right]$ is no more than $\delta$. And as the straight line between the endpoints of these contours, each directed line segment $\overrightarrow{z_i, z_{i-1}}$ also has length no more than $\delta$. In other words, each $\gamma_i$ for $i = 1, 2, 3$ is a closed curve of length no more than $2\delta$.

The last integral in (3-2) is taken over a closed contour composed of three line segments, each of length no more than $\delta$. For each $j = 0, 1, 2$ define $b_j = \tfrac{1}{2}(z_j + z_{j+1})$ to be the bisector of segment $\overrightarrow{z_j, z_{j+1}}$, and for convenience define $b_{-1} = b_2$. We then define $\gamma_{j+4} = \overrightarrow{z_j, b_j} + \overrightarrow{b_j, b_{j-1}} + \overrightarrow{b_{j-1}, z_j}$ and define $\gamma_7 = \overrightarrow{b_0, b_1} + \overrightarrow{b_1, b_2} + \overrightarrow{b_2, b_0}$. Then we have

$$\oint_{\overrightarrow{z_0,z_1}+\overrightarrow{z_1,z_2}+\overrightarrow{z_2,z_3}} P_y\,dx - P_x\,dy = \sum_{k=4}^{7} \oint_{\gamma_k} P_y\,dx - P_x\,dy.$$

But by similar triangles,

$$\mathrm{L}(\gamma_k) = \tfrac{1}{2}\mathrm{L}(\overrightarrow{z_0, z_1} + \overrightarrow{z_1, z_2} + \overrightarrow{z_2, z_3}) \le \tfrac{3}{2}\delta$$

for each $k = 4, 5, 6, 7$. Combining these observations with (3-2) and (2-3), we have

$$\sum_{R_i \in R} \int_{R_i} \Delta P(w)\, dm(w) = \sum_{k=1}^{7} \oint_{\gamma_k} P_y\, dx - P_x\, dy$$

$$\le \sum_{k=1}^{7} \Lambda(\gamma_k(0), \mathrm{L}(\gamma_k)) \le 3f(2\delta) + 4f\left(\tfrac{3}{2}\delta\right) \le 7f(2\delta)$$

for all $(z_0, 3\delta)$-stockyards $R$. Therefore by Theorem 2.3 we see $\Lambda(z, 3\delta) \le 7f(2\delta)$ for all $z \in \mathbb{C}$; hence

$$f(3\delta) = \sup_{z \in \mathbb{C}} \Lambda(z, 3\delta) \le 7f(2\delta).$$

In summary, for all $\delta \ge \tfrac{2}{3}\delta_0$ we have

$$f(2\delta) \le f(3\delta) \le 7f(2\delta). \tag{3-3}$$

Now fix $\delta \ge \delta_0$. Because $f(\delta)$ is a nondecreasing function, we also trivially have $f(\delta) \le f(2\delta)$. But by monotonicity and (3-3) we see

$$f(2\delta) \le f\left(\tfrac{9}{4}\delta\right) \le 49f(\delta).$$

Therefore, $f(\delta) \approx f(2\delta)$ for all $\delta \ge \delta_0$. $\qquad\square$

**Remark 3.3.** Lemma 3.2 was used implicitly in [Peterson 2014] without proof or statement. The arguments of [Peterson 2014] show that for any fixed $z \in \mathbb{C}$,

$$\left\{(w, s) \in \mathbb{C} \times \mathbb{R} : |w - z| < \tfrac{1}{4}\delta,\ |s - t - T(z, w)| < \Lambda\left(z, \tfrac{1}{4}\delta\right)\right\} \subseteq B_d((z, t), \delta)$$

and

$$B_d((z, t), \delta) \subseteq \left\{(w, s) \in \mathbb{C} \times \mathbb{R} : |w - z| < 3\delta,\ |s - t - T(z, w)| < \Lambda(z, 3\delta)\right\},$$

where

$$T(z, w) = -2\mathrm{Im}\left(\int_0^1 (w - z) P_z(r(w - z) + z)\, dr\right)$$

is the "twist" of the CC ball. Lemma 3.2 then yields the formula

$$\mathrm{Vol}(B_d((z, t), \delta)) \approx \delta^2 \Lambda(z, \delta) \quad \text{for} \quad \delta \ge \delta_0$$

when $b\Omega$ has UGS $(f(\delta), \delta_0)$. This shows that we can think of $B_d((z, t), \delta)$ as a "twisted" ellipsoid in the case of large $\delta$, not just small $\delta$ as in (1-1).

We are now ready to make precise the intuition laid out in the beginning of this section.

**Proposition 3.4.** *If* $b\Omega$ *has UGS* $(f(\delta), \delta_0)$, *then*

$$\Lambda(z, \delta) \approx \sup_{\hat{z} \in B(z, \delta)} \sup_{0 < \hat{\delta} \leq \delta} \frac{\delta}{\hat{\delta}} \int_{B(\hat{z}, \hat{\delta})} \Delta P(w)\, dm(w)$$

*uniformly for* $z \in \mathbb{C}$ *and* $\delta \geq \delta_0$.

*Proof.* As in the proof of [Lemma 3.2], we assume without loss of generality that $f(\delta)$ is a nondecreasing function. For any choice of $\hat{z} \in B(z, \delta)$ and $0 < \hat{\delta} \leq \delta$, define a $(z, 4\pi\delta)$-stockyard $R = (R_0, R_1, \ldots, R_N)$ composed of one pen $R_0$ which is a circle touching $z$ and some point on $bB(\hat{z}, \hat{\delta})$ and $N = \lfloor \delta/\hat{\delta} \rfloor$ copies of $B(\hat{z}, \hat{\delta})$. Using the fact that $\lfloor \delta/\hat{\delta} \rfloor \geq \delta/(2\hat{\delta})$ because $\delta \geq \hat{\delta} > 0$, we have

$$\Lambda(z, \delta) \approx f(\delta) \approx f(16\delta) \geq f(4\pi\delta) \gtrsim \Lambda(z, 4\pi\delta)$$

$$\geq \sum_{R_i \in R} \int_{R_i} \Delta P(w)\, dm(w) \geq \left\lfloor \frac{\delta}{\hat{\delta}} \right\rfloor \int_{B(\hat{z}, \hat{\delta})} \Delta P(w)\, dm(w)$$

$$\geq \frac{\delta}{2\hat{\delta}} \int_{B(\hat{z}, \hat{\delta})} \Delta P(w)\, dm(w).$$

Therefore,

$$\Lambda(z, \delta) \gtrsim \sup_{\hat{z} \in B(z, \delta)} \sup_{0 < \hat{\delta} \leq \delta} \frac{\delta}{\hat{\delta}} \int_{B(\hat{z}, \hat{\delta})} \Delta P(w)\, dm(w).$$

Now let $R = (R_1, \ldots, R_M)$ be an arbitrary $(z, \delta)$-stockyard. For $i = 1, \ldots, M$, fix some point $z_i \in R_i$. Then, recalling [Remark 2.2], we have

$$\sum_{R_i \in R} \int_{R_i} \Delta P(w)\, dm(w) \leq \sum_{i=1}^{M} \int_{B(z_i, L(bR_i))} \Delta P(w)\, dm(w)$$

$$= \sum_{i=1}^{M} \frac{L(bR_i)}{\delta} \frac{\delta}{L(bR_i)} \int_{B(z_i, L(bR_i))} \Delta P(w)\, dm(w)$$

$$\leq \sum_{i=1}^{M} \left( \frac{L(bR_i)}{\delta} \right) \sup_{\hat{z} \in B(z, \delta)} \sup_{0 < \hat{\delta} \leq \delta} \frac{\delta}{\hat{\delta}} \int_{B(\hat{z}, \hat{\delta})} \Delta P(w)\, dm(w)$$

$$\leq \frac{\delta}{\delta} \sup_{\hat{z} \in B(z, \delta)} \sup_{0 < \hat{\delta} \leq \delta} \frac{\delta}{\hat{\delta}} \int_{B(\hat{z}, \hat{\delta})} \Delta P(w)\, dm(w)$$

$$= \sup_{\hat{z} \in B(z, \delta)} \sup_{0 < \hat{\delta} \leq \delta} \frac{\delta}{\hat{\delta}} \int_{B(\hat{z}, \hat{\delta})} \Delta P(w)\, dm(w).$$

Therefore

$$\Lambda(z, \delta) \leq \sup_{\hat{z} \in B(z, \delta)} \sup_{0 < \hat{\delta} \leq \delta} \frac{\delta}{\hat{\delta}} \int_{B(\hat{z}, \hat{\delta})} \Delta P(w)\, dm(w). \qquad \square$$

# 4. Proof of Theorem 1.5

Proposition 3.4 reveals very strong information about the density in the space around a point when there is a UGS. Armed with this knowledge, we are almost ready to prove Theorem 1.5. We begin by recalling and proving two lemmas, the first of which is a technical result from [Peterson 2014].

**Lemma 4.1** [Peterson 2014, Lemma 4.1]. *If* $b\Omega$ *has a UGS, then there are constants* $C_1, C_2 > 0$, *depending only on* $\Delta P$ *and* $\delta_0$, *such that*

(a) $\displaystyle \inf_{z \in \mathbb{C}} \sup_{\hat{z} \in B(z,\delta)} \sup_{0 < \hat{\delta} \leq \delta} (\hat{\delta} + \hat{\delta}^2)^{-1} \int_{B(\hat{z},\hat{\delta})} \Delta P(w) \, dm(w) \geq C_1$ *for all* $\delta \geq \delta_0$;

(b) $\displaystyle \sup_{z \in \mathbb{C}} \sup_{\delta > 0} (\delta + \delta^2)^{-1} \int_{B(z,\delta)} \Delta P(w) \, dm(w) \leq C_2$.

**Remark 4.2.** Note that increasing $\delta_0$ can only possibly increase $C_1$ and will not affect the constant $C_2$.

We also need a short geometric lemma.

**Lemma 4.3.** *Let* $0 < a \leq b$. *Then within any disc of radius* $b$ *in* $\mathbb{C}$, *one can pack at least* $b^2/(16a^2)$ *disjoint discs of radius* $a$.

*Proof.* Without loss of generality, assume the disc of radius $b$ is centered at the origin. Since $B(0, a) \subset B(0, b)$, we can always pack at least one disc of radius $a$ inside of $B(0, b)$. If $2a > \sqrt{2}b$, then we have at least one disc of radius $a$ inside of $B(0, b)$, and

$$1 > \frac{\sqrt{2}b}{2a} > \frac{b^2}{2a^2} > \frac{b^2}{16a^2}.$$

Note now that for all $x \geq 1$, we have $x = \lfloor x \rfloor + \alpha$ for some $\alpha \in [0, 1)$ so that

$$\lfloor x^2 \rfloor = \lfloor (\lfloor x \rfloor + \alpha)^2 \rfloor < \lfloor (\lfloor x \rfloor + \lfloor x \rfloor)^2 \rfloor = \lfloor 4\lfloor x \rfloor^2 \rfloor = 4\lfloor x \rfloor^2.$$

Assume that $2a \leq \sqrt{2}b$. The disc $B(0, b)$ contains a square of side length

$$\left\lfloor \frac{\sqrt{2}b}{2a} \right\rfloor 2a \leq \sqrt{2}b$$

centered at the origin. This square contains exactly $\lfloor \sqrt{2}b/(2a) \rfloor^2$ disjoint squares of side length $2a$, each of which contains a disc of radius $a$. So we again see that $B(0, b)$ contains at least

$$\left\lfloor \frac{\sqrt{2}b}{2a} \right\rfloor^2 > \frac{1}{4}\left\lfloor \frac{b^2}{2a^2} \right\rfloor \geq \frac{b^2}{16a^2}$$

discs of radius $a$. $\qquad\square$

We are now ready to prove Theorem 1.5.

*Proof of Theorem 1.5.* Proposition 3.4 shows that there is some constant $c > 0$ such that

$$\sup_{\hat{z} \in B(z,\delta)} \sup_{0 < \hat{\delta} \leq \delta} \frac{\delta}{\hat{\delta}} \int_{B(\hat{z},\hat{\delta})} \Delta P(w) \, dm(w) \geq cf(\delta)$$

for all $z \in \mathbb{C}$ and $\delta \geq \delta_0$. So for all $z \in \mathbb{C}$ and $\delta \geq \delta_0$, there exists $\hat{z} \in B(z,\delta)$ and $0 < \hat{\delta} \leq \delta$ such that

$$\frac{1}{\hat{\delta}} \int_{B(\hat{z},\hat{\delta})} \Delta P(w) \, dm(w) \geq \frac{c}{2} \frac{f(\delta)}{\delta}.$$

Now suppose $f(\delta) = \delta$ is not a UGS for $b\Omega$. That is, $\limsup_{\delta \to +\infty} f(\delta)/\delta = +\infty$. Then, taking $C_2 > 0$ as in Lemma 4.1, we can choose $\delta_1 > \max(1, \delta_0)$ such that $f(\delta_1)/\delta_1 > 4C_2/c$. Choose $\hat{\delta}$ associated to $\delta = \delta_1$ as above. If $\hat{\delta} \leq 1$, then by Lemma 4.1 we have

$$2C_2 < \frac{c}{2} \frac{f(\delta_1)}{\delta_1} \leq \frac{1}{\hat{\delta}} \int_{B(\hat{z},\hat{\delta})} \Delta P(w) \, dm(w) \leq \frac{2}{\hat{\delta} + \hat{\delta}^2} \int_{B(\hat{z},\hat{\delta})} \Delta P(w) \, dm(w) \leq 2C_2,$$

which is impossible. Therefore for all $z \in \mathbb{C}$, there exists $\hat{z} \in B(z,\delta_1)$ and $1 \leq \hat{\delta} \leq \delta_1$ such that

$$\int_{B(\hat{z},\hat{\delta})} \Delta P(w) \, dm(w) \geq \frac{c}{2} \frac{f(\delta_1)}{\delta_1} \hat{\delta} \geq 2C_2 > 0.$$

It follows that for all $z \in \mathbb{C}$,

$$\int_{B(z,2\delta_1)} \Delta P(w) \, dm(w) \geq \int_{B(\hat{z},\hat{\delta})} \Delta P(w) \, dm(w) \geq 2C_2.$$

By Lemma 4.3, for all $\delta \geq \delta_1$, we can pack $N > \delta^2/(16\delta_1^2)$ disjoint discs of radius $2\delta_1$ within a disc of radius $2\delta$. So for all $z \in \mathbb{C}$,

$$\int_{B(z,2\delta)} \Delta P(w) \, dm(w) \geq N 2C_2 > \frac{\delta^2}{16\delta_1^2} \cdot 2C_2 \approx (2\delta)^2.$$

Then for all $\delta \geq 2\delta_1$ and some $z_1 \in bB(z,\delta)$,

$$f(\delta) \approx f(2\pi\delta) \approx \Lambda(z_1, 2\pi\delta) \geq \int_{B(z,\delta)} \Delta P(w) \, dm(w) \gtrsim \delta^2.$$

But Theorem 1.3 implies $f(\delta) \lesssim \delta^2$ for all $\delta \geq 2\delta_1 \geq \delta_0$. Therefore setting $\delta^* = 2\delta_1$ we see that if $f(\delta) = \delta$ is not a UGS for $b\Omega$, then $(\delta^2, \delta^*)$ is a UGS for $b\Omega$. $\square$

So a UGS must grow in a linear or quadratic fashion. Linear growth means that for any point, the stockyards which pick up the most mass enclose a small, dense, nearby disc as many times as possible. Quadratic growth means a stockyard which picks up the most mass does so by taking a pen consisting of one large disc, collecting as much area as possible.

## 5. Identifying uniform global structures

So far, almost all of the results of this paper have taken as hypothesis that b$\Omega$ has a UGS and considered what that means for the global structure $\Lambda$. To look at an arbitrary model domain and determine if there is a UGS is a much more difficult question. But with Theorem 1.5, we see that we only need to provide conditions to identify uniform global structures where either $f(\delta) = \delta$ or $f(\delta) = \delta^2$. The following two theorems provide necessary and sufficient conditions for each case.

**Theorem 5.1.** *The hyperspace* b$\Omega$ *has UGS* $(\delta, \delta_0)$ *if and only if*

(a) $\displaystyle\int_{B(z,\delta)} \Delta P(w)\, dm(w) \lesssim \delta$ *for all* $z \in \mathbb{C}$ *and* $\delta > 0$, *and*

(b) *there exist constants* $\delta^* > M > 0$ *such that*

$$\inf_{z\in\mathbb{C}} \sup_{\hat{z}\in B(z,\delta^*)} \sup_{0<\hat{\delta}\leq M} \frac{1}{\hat{\delta}} \int_{B(\hat{z},\hat{\delta})} \Delta P(w)\, dm(w) \gtrsim 1.$$

*Proof.* Suppose $(\delta, \delta_0)$ is a UGS for b$\Omega$. For any $z \in \mathbb{C}$, fix some point $z_1$ with $|z_1 - z| = \delta$. If $2\pi\delta \geq \delta_0$ then

$$\int_{B(z,\delta)} \Delta P(w)\, dm(w) \leq \Lambda(z_1, 2\pi\delta) \approx 2\pi\delta \approx \delta.$$

If $0 < 2\pi\delta < \delta_0$, then taking a stockyard consisting of $\lfloor \delta_0/(2\pi\delta) \rfloor$ copies of $B(z, \delta)$ gives

$$\frac{\delta_0}{4\pi\delta} \int_{B(z,\delta)} \Delta P(w)\, dm(w) \leq \left\lfloor \frac{\delta_0}{2\pi\delta} \right\rfloor \int_{B(z,\delta)} \Delta P(w)\, dm(w) \leq \Lambda(z_1, \delta_0) \approx 1.$$

Therefore (a) holds.

Also, for any fixed $\delta^* \geq \delta_0 > 0$, Lemma 4.1 gives some constant $C_1 > 0$ such that

$$\inf_{z\in\mathbb{C}} \sup_{\hat{z}\in B(z,\delta^*)} \sup_{0<\hat{\delta}\leq\delta_0} \frac{1}{\hat{\delta}} \int_{B(\hat{z},\hat{\delta})} \Delta P(w)\, dm(w)$$

$$\geq \inf_{z\in\mathbb{C}} \sup_{\hat{z}\in B(z,\delta_0)} \sup_{0<\hat{\delta}\leq\delta_0} \frac{1}{\hat{\delta}} \int_{B(\hat{z},\hat{\delta})} \Delta P(w)\, dm(w)$$

$$\geq \inf_{z\in\mathbb{C}} \sup_{\hat{z}\in B(z,\delta_0)} \sup_{0<\hat{\delta}\leq\delta_0} \frac{1}{\hat{\delta}+\hat{\delta}^2} \int_{B(\hat{z},\hat{\delta})} \Delta P(w)\, dm(w) \geq C_1.$$

Therefore (b) holds (with $M = \delta_0$).

Now we suppose (a) and (b) hold. For any $\delta > 0$ and $z \in \mathbb{C}$, let $R = (R_1, \ldots, R_N)$ be an arbitrary $(z, \delta)$-stockyard. For each $i = 1, \ldots, N$, fix some point $z_i \in R_i$.

Then recalling Remark 2.2, (a) gives

$$\sum_{R_i \in R} \int_{R_i} \Delta P(w)\, dm(w) \leq \sum_{R_i \in R} \int_{B(z_i, \mathrm{L}(bR_i))} \Delta P(w)\, dm(w) \lesssim \sum_{R_i \in R} \mathrm{L}(bR_i) \leq \delta.$$

Therefore $\Lambda(z, \delta) \lesssim \delta$ uniformly for $z \in \mathbb{C}$ and $\delta > 0$.

For any $z \in \mathbb{C}$, fix a $\hat{z} \in B(z, \delta^*)$ and $0 < \hat{\delta} \leq M$ such that

$$\frac{1}{\hat{\delta}} \int_{B(\hat{z}, \hat{\delta})} \Delta P(w) dm(w) \gtrsim 1,$$

as given by (b). Then for all $\delta \geq 2\pi M \geq 2\pi \hat{\delta}$, there is a $(z, \pi\delta^* + \delta)$-stockyard $R$ which consists of one circular pen touching $z$ and some point on $bB(\hat{z}, \hat{\delta})$ and $\lfloor \delta/(2\pi\hat{\delta}) \rfloor$ copies of $B(\hat{z}, \hat{\delta})$. Then

$$\begin{aligned}
\Lambda(z, \pi\delta^* + \delta) &\geq \sum_{R_i \in R} \int_{R_i} \Delta P(w)\, dm(w) \\
&\geq \left\lfloor \frac{\delta}{2\pi\hat{\delta}} \right\rfloor \int_{B(\hat{z}, \hat{\delta})} \Delta P(w)\, dm(w) \geq \frac{\delta}{4\pi\hat{\delta}} \int_{B(\hat{z}, \hat{\delta})} \Delta P(w)\, dm(w) \\
&\gtrsim \delta = 2\pi M \frac{\delta}{2\pi M} \geq \frac{2\pi M}{2\pi M + \pi\delta^*}(\pi\delta^* + \delta),
\end{aligned}$$

where here we have used the fact that if $c \geq 0$ and $a \geq b > 0$, then $a/b \geq (a+c)/(b+c)$. Therefore $\Lambda(z, \delta) \approx \delta$ for all $\delta \geq \delta_0$ with $\delta_0 = \pi\delta^* + 2\pi M$. $\qquad\square$

**Theorem 5.2.** *The hypersurface* $b\Omega$ *has UGS* $(\delta^2, \delta_0)$ *if and only if there exists* $\delta^* > 0$ *such that, uniformly for* $z \in \mathbb{C}$,

(a) $\displaystyle\int_{B(z,\delta)} \Delta P(w)\, dm(w) \lesssim \delta$ *when* $\delta \leq \delta^*$, *and*

(b) $\displaystyle\int_{B(z,\delta)} \Delta P(w)\, dm(w) \approx \delta^2$ *when* $\delta \geq \delta^*$.

*Proof.* Suppose $(\delta^2, \delta_0)$ is a UGS for $b\Omega$. Then for any $z \in \mathbb{C}$ and some point $z_1$ with $|z_1 - z| = \delta$ we have

$$\int_{B(z,\delta)} \Delta P(w)\, dm(w) \leq \Lambda(z_1, 2\pi\delta) \approx (2\pi\delta)^2 \approx \delta^2$$

for all $\delta \geq \delta_0$.

Proposition 3.4 shows that there is some constant $c > 0$ such that

$$\sup_{\hat{z} \in B(z,\delta)} \sup_{0 < \hat{\delta} \leq \delta} \frac{\delta}{\hat{\delta}} \int_{B(\hat{z}, \hat{\delta})} \Delta P(w)\, dm(w) \geq c\delta^2$$

for all $z \in \mathbb{C}$ and $\delta \geq \delta_0$. So for all $z \in \mathbb{C}$ and $\delta \geq \delta_0$, there exists $\hat{z} \in B(z, \delta)$ and $0 < \hat{\delta} \leq \delta$ such that

$$\frac{1}{\hat{\delta}} \int_{B(\hat{z}, \hat{\delta})} \Delta P(w) \, dm(w) \geq \tfrac{1}{2} c \delta.$$

Taking $C_2 > 0$ as in Lemma 4.1, choose some $\delta_1 > \max(1, \delta_0, 4C_2/c)$. Choose $\hat{\delta}$ associated to $\delta = \delta_1$ as above. If $\hat{\delta} \leq 1$, then by Lemma 4.1 we have

$$2C_2 < \frac{c}{2}\delta_1 \leq \frac{1}{\hat{\delta}} \int_{B(\hat{z}, \hat{\delta})} \Delta P(w) \, dm(w) \leq \frac{2}{\hat{\delta} + \hat{\delta}^2} \int_{B(\hat{z}, \hat{\delta})} \Delta P(w) \, dm(w) \leq 2C_2,$$

which is impossible. Therefore for all $z \in \mathbb{C}$, there exists $\hat{z} \in B(z, \delta_1)$ and $1 \leq \hat{\delta} \leq \delta_1$ such that

$$\int_{B(\hat{z}, \hat{\delta})} \Delta P(w) \, dm(w) \geq \frac{c}{2}\delta_1 \hat{\delta} \geq 2C_2 > 0.$$

It follows that for all $z \in \mathbb{C}$,

$$\int_{B(z, 2\delta_1)} \Delta P(w) \, dm(w) \geq \int_{B(\hat{z}, \hat{\delta})} \Delta P(w) \, dm(w) \geq 2C_2.$$

By Lemma 4.3, for all $\delta \geq \delta_1$, we can pack $N > \delta^2/(16\delta_1^2)$ disjoint discs of radius $2\delta_1$ within a disc of radius $2\delta$. So for all $z \in \mathbb{C}$,

$$\int_{B(z, 2\delta)} \Delta P(w) \, dm(w) \geq N \int_{B(z, 2\delta_1)} \Delta P(w) \, dm(w) > \frac{\delta^2}{16\delta_1^2} \cdot 2C_2 \approx (2\delta)^2.$$

Therefore,

$$\int_{B(z, \delta)} \Delta P(w) \, dm(w) \approx \delta^2$$

for all $\delta \geq 2\delta_1 > \delta_0$. Setting $\delta^* = 2\delta_1$, we see (b) holds. Also, Lemma 4.1 yields

$$\int_{B(z, \delta)} \Delta P(w) \, dm(w) \leq C_2(\delta + \delta^2).$$

But if $\delta \leq \delta^*$ then

$$\int_{B(z, \delta)} \Delta P(w) \, dm(w) \leq C_2(\delta^* + 1)\delta \approx \delta,$$

so (b) holds.

Now we suppose (a) and (b) hold so that for $\delta \leq \delta^*$ we have

$$\int_{B(z, \delta)} \Delta P(w) \, dm(w) \leq a\delta$$

and for $\delta \geq \delta^*$ we have

$$\int_{B(z, \delta)} \Delta P(w) \, dm(w) \leq b\delta^2$$

for some constants $a, b > 0$ independent of $z \in \mathbb{C}$. For $\delta \geq 1$, let $R = (R_1, \ldots, R_N)$ be an arbitrary $(z, \delta)$-stockyard. Without loss of generality, we may relabel the pens so that $L(bR_i) \leq \delta^*$ for $i = 1, \ldots, L$ and $L(bR_i) \geq \delta^*$ for $i = L+1, \ldots, N$ for some integer $L \in \{0, \ldots, N\}$. For each $i = 1, \ldots, N$, fix some $z_i \in R_i$. Recalling Remark 2.2, we have

$$\sum_{R_i \in R} \int_{R_i} \Delta P(w) \, dm(w)$$

$$= \sum_{i=1}^{L} \int_{R_i} \Delta P(w) \, dm(w) + \sum_{i=L+1}^{N} \int_{R_i} \Delta P(w) \, dm(w)$$

$$\leq \sum_{i=1}^{L} \int_{B(z_i, L(bR_i))} \Delta P(w) \, dm(w) + \sum_{i=L+1}^{N} \int_{B(z_i, L(bR_i))} \Delta P(w) \, dm(w)$$

$$\leq a \sum_{i=1}^{L} L(bR_i) + b \sum_{i=L+1}^{N} L(bR_i)^2$$

$$\leq a \sum_{i=1}^{L} L(bR_i) + b \left( \sum_{i=L+1}^{N} L(bR_i) \right)^2 \leq a\delta + b\delta^2 \lesssim \delta^2.$$

So $\Lambda(z, \delta) \lesssim \delta^2$ for all $\delta \geq 1$.

Using (b), we may take a stockyard consisting of one large circular pen with radius $\delta \geq \delta^*$ and center $z_1$ satisfying $|z_1 - z| = \delta$ to see that

$$\Lambda(z, 2\pi\delta) \geq \int_{B(z_1, \delta)} \Delta P(w) \, dm(w) \approx \delta^2 \approx (2\pi\delta)^2.$$

Therefore $\Lambda(z, \delta) \approx \delta^2$ for all $\delta \geq \delta_0$ with $\delta_0 = \max(1, \delta^*/(2\pi))$. $\qquad\square$

## 6. Future directions

Although the results of this paper completely describe the nature of uniform global structures for the model domains we consider, several interesting avenues for further study present themselves when we weaken our hypotheses. One such direction would be to extend the results of this paper to higher dimensions. That is, is there an appropriate notion of stockyards in higher dimensions with which to analyze the global structure on the boundary of a model domain in $\mathbb{C}^n$? It is not clear how the Green's theorem argument used in [Peterson 2014] to prove Theorem 2.3 would generalize or even how (if at all) the notion of stockyards should generalize to higher dimensions.

One could also relax the conditions on $P$ which determine the boundary $b\Omega$. For example, do similar results hold assuming that $P$ is only once differentiable and that $\Delta P$ as a distribution is nonnegative? One could also allow $P$ to be a more

general function for which $\Omega$ is pseudoconvex, that is, take $P = P(z_1, \mathrm{Re}(z_2))$. In such a situation, the volume of CC balls with such a choice of $P$ would a priori depend on the $\mathrm{Re}(z_2)$-direction. Since the methods of this paper heavily exploited the $\mathrm{Re}(z_2)$-translation invariance of $\Omega$, it is unclear if these methods can be easily extended to handle the more general situation.

## Acknowledgements

## References

[Chang and Chang 2014] S.-C. Chang and T.-H. Chang, "On CR volume growth estimate in a complete pseudohermitian 3-manifold", *Int. J. Math.* **25**:4 (2014), art. id. 1450035, 22 pp. MR Zbl

[Coifman and Weiss 1977] R. R. Coifman and G. Weiss, "Extensions of Hardy spaces and their use in analysis", *Bull. Amer. Math. Soc.* **83**:4 (1977), 569–645. MR Zbl

[Fässler et al. 2015] K. Fässler, P. Koskela, and E. Le Donne, "Nonexistence of quasiconformal maps between certain metric measure spaces", *Int. Math. Res. Not.* **2015**:16 (2015), 6968–6987. MR Zbl

[Heinonen and Koskela 1998] J. Heinonen and P. Koskela, "Quasiconformal maps in metric spaces with controlled geometry", *Acta Math.* **181**:1 (1998), 1–61. MR Zbl

[Montanari and Morbidelli 2012] A. Montanari and D. Morbidelli, "Nonsmooth Hörmander vector fields and their control balls", *Trans. Amer. Math. Soc.* **364**:5 (2012), 2339–2375. MR Zbl

[Nagel et al. 1985] A. Nagel, E. M. Stein, and S. Wainger, "Balls and metrics defined by vector fields, I: Basic properties", *Acta Math.* **155**:1 (1985), 103–147. MR Zbl

[Nagel et al. 1988] A. Nagel, J.-P. Rosay, E. M. Stein, and S. Wainger, "Estimates for the Bergman and Szegő kernels in certain weakly pseudoconvex domains", *Bull. Amer. Math. Soc.* (*N.S.*) **18**:1 (1988), 55–59. MR Zbl

[Nagel et al. 1989] A. Nagel, J.-P. Rosay, E. M. Stein, and S. Wainger, "Estimates for the Bergman and Szegő kernels in $\mathbb{C}^2$", *Ann. of Math.* (2) **129**:1 (1989), 113–149. MR Zbl

[Peterson 2014] A. Peterson, "Carnot–Carathéodory metrics in unbounded subdomains of $\mathbb{C}^2$", *Arch. Math.* (*Basel*) **102**:5 (2014), 437–447. MR Zbl

[Stein 1993] E. M. Stein, *Harmonic analysis: real-variable methods, orthogonality, and oscillatory integrals*, Princeton Mathematical Series **43**, Princeton University Press, 1993. MR Zbl

[Street 2014] B. Street, *Multi-parameter singular integrals*, Annals of Mathematics Studies **189**, Princeton Univ. Press, 2014. MR Zbl

ethandlugie2018@u.northwestern.edu

Department of Mathematics, Northwestern University, 2033 Sheridan Road, Evanston, IL 60208, United States

aaron.peterson@northwestern.edu     Department of Mathematics, Northwestern University, 2033 Sheridan Road, Evanston, IL 60208, United States

# Variations of the Greenberg unrelated question binary model

## David P. Suarez and Sat Gupta

(Communicated by Kenneth S. Berenhaut)

We explore different variations of the Greenberg unrelated question RRT model for a binary response. In one of the variations, we allow multiple independent responses from each respondent. In another variation, we use inverse sampling. It turns out that both of these variations produce more efficient models, a fact validated by both theoretical comparisons as well as extensive computer simulations.

## 1. Introduction

Social desirability response bias (SDB) is a major concern in surveys involving sensitive topics. One method that could help circumvent SDB is the randomized response technique, introduced originally by Warner [1965] and then generalized by other researchers such as Greenberg et al. [1969; 1971], Warner [1971], Klein and Spady [1993], Gupta et al. [2002; 2013].

RRT models have been used extensively in field surveys. Abernathy et al. [1970] used RRT models to obtain estimates of induced abortion rates in urban North Carolina. From the open survey, it was noticed that female respondents would have hesitated to respond truthfully to the sensitive question of induced abortions. Striegel et al. [2006] used indirect questioning techniques to measure the prevalence of doping among elite athletes. In order to study the effect of higher education in favourable attitudes towards foreigners in Germany, Ostapczuck et al. [2009] used two survey methods: direct questioning and RRT. The results obtained by these two survey methods demonstrated great variation. Based upon the respondents who used RRT, the results obtained showed a sharp decline in the estimates for the proportion of xenophiles among both the less educated and highly educated. Gill et al. [2013] conducted a survey which used an RRT model to estimate the risky sexual behaviors among students at the University of North Carolina at Greensboro. The binary question of interest was "Have you been told by a healthcare professional

that you have a sexually transmitted disease?", whereas the quantitative question of interest was "How many sexual partners have you had in the last 12 months?" The survey was conducted using three methods: RRT method, direct face-to-face interviewing and anonymous check-box survey method. It was observed that the optional unrelated question RRT method's estimates were closer to the check-box survey method's estimates, and the lowest point estimate was obtained by face-to-face interview method, which is expected as it provided the lowest anonymity. More recently, Chhabra et al. [2016] used these models to estimate the prevalence of sexual abuse of female college students by either a friend or an acquaintance.

In this paper, we discuss some variations of the Greenberg et al. (1969) unrelated question RRT model. In one of the variations, we allow a respondent to provide multiple independent responses. In another variation, we use the inverse sampling technique.

## 2. Proposed models

**2.1.** *Using multiple independent responses in the Greenberg model.* Let us first recall the Greenberg et al. (1969) unrelated question RRT model, which we will henceforth refer to as the Greenberg model. Let $\pi_x$ be the unknown prevalence of a sensitive attribute $X$ in the population and $\pi_y$ be the known prevalence of a nonsensitive attribute $Y$. A randomization device offers respondents a choice between two questions, a sensitive question and an unrelated question with respective probabilities $p$ and $1 - p$. Let $p_y$ be the probability of a "yes" response. Then

$$p_y = \pi_x p + \pi_y (1 - p), \tag{1}$$

which leads to the estimator

$$\hat{\pi}_G = \frac{\hat{p}_y - \pi_y (1 - p)}{p}, \tag{2}$$

where $\hat{p}_y$ is the sample proportion of "yes" responses.

The mean of the estimator in (2) is given by

$$E(\hat{\pi}_G) = \pi_x,$$

which signifies that $\hat{\pi}_G$ is an unbiased estimator of $\pi_x$.

The variance of the estimator in (2) is given by

$$\mathrm{Var}(\hat{\pi}_G) = \frac{p_y (1 - p_y)}{np^2}. \tag{3}$$

Now suppose we allow $m$ independent responses from each respondent in a sample of size $n$. Let $T_i$ be the number of "yes" responses provided by the $i$-th

respondent. Then

$$T_i \sim \text{Binomial}(m, p_y) \quad \text{and} \quad E(T_i) = mp_y.$$

If $\bar{T} = \left(\sum T_i\right)/n$, then we know that $E(\bar{T}) = mp_y$.
Estimating $mp_y$ by $\bar{T}$, the estimator for $\pi_x$ in (2) can be refined to

$$\hat{\pi}_{GM} = \frac{\bar{T}/m - (1-p)\pi_y}{p}. \tag{4}$$

Note that

$$E(\hat{\pi}_{GM}) = \frac{E(\bar{T})/m - (1-p)\pi_y}{p} = \pi_x. \tag{5}$$

The variance of the estimator $\hat{\pi}_{GM}$ is given by

$$\text{Var}(\hat{\pi}_{GM}) = \frac{1}{m^2 p^2} \text{Var}(\bar{T}) = \frac{1}{m^2 p^2} \frac{mp_y(1-p_y)}{n} = \frac{p_y(1-p_y)}{nmp^2}. \tag{6}$$

**2.2. Inverse sampling: waiting for the first "yes" response.** Let each respondent continue to use the Greenberg model repeatedly until a "yes" response is recorded. Let $S_i$ be the total number of trials needed by the $i$-th respondent to get to the first "yes" response. Then,

$$S_i \sim \text{Geometric}(p_y),$$

with $E(S_i) = 1/p_y$ and $\text{Var}(S_i) = (1-p_y)/p_y^2$, where $p_y$ is defined in (1).

Also, let there be a sample of $n$ respondents and $\bar{S}$ be the sample mean of the $S_i$'s. Then $1/p_y$ can be estimated by $\bar{S}$ leading to $\hat{p}_y = 1/\bar{S}$ as an estimator of $p_y$. Using first-order Taylor's approximation of $1/S$, we can write

$$\frac{1}{\bar{S}} \approx \frac{1}{E(S)} + (\bar{S} - E(S))\left(\frac{-1}{(E(S))^2}\right), \tag{7}$$

where $E(S) = 1/p_y$.

With this approximation,

$$E\left(\frac{1}{\bar{S}}\right) \approx \frac{1}{E(S)} = p_y. \tag{8}$$

Then, using $1/\bar{S}$ as an estimator of $p_y$, the estimator in (2) becomes

$$\hat{\pi}_{GI} = \frac{1/\bar{S} - (1-p)\pi_y}{p}. \tag{9}$$

Note that

$$E(\hat{\pi}_{GI}) = \frac{E(1/\bar{S}) - (1-p)\pi_y}{p} \approx \frac{p_y - (1-p)\pi_y}{p} = \pi_x$$

since $E(1/\bar{S}) \approx p_y$, as argued in (8).

Thus, we see that $\hat{\pi}_{GI}$ is an unbiased estimator of $\pi_x$, up to first order of approximation.

From (9),

$$\text{Var}(\hat{\pi}_{GI}) = \frac{1}{p^2}\,\text{Var}\!\left(\frac{1}{\bar{S}}\right). \tag{10}$$

But

$$\text{Var}\!\left(\frac{1}{\bar{S}}\right) \approx \text{Var}\!\left(\frac{1}{E(\bar{S})} + (\bar{S} - E(\bar{S}))\!\left(-\frac{1}{(E(\bar{S}))^2}\right)\right) = \text{Var}(-\bar{S}p_y^2)$$

$$= p_y^4\,\text{Var}(\bar{S}) = p_y^4\,\frac{\text{Var}(S)}{n} = p_y^4\!\left(\frac{1-p_y}{np_y^2}\right) = \frac{p_y^2(1-p_y)}{n}. \tag{11}$$

Thus, we have

$$\text{Var}(\hat{\pi}_{GI}) \approx \frac{1}{p^2}\!\left(\frac{p_y^2(1-p_y)}{n}\right) = \frac{p_y^2(1-p_y)}{np^2}. \tag{12}$$

**2.3. Inverse sampling: waiting for k "yes" responses.** Let $S_i$ be the total number of trials needed to reach the $k$-th "yes" response. Then, we see that $S_i \sim$ Negative Binomial$(p_y, k)$ with

$$E(S_i) = \frac{k}{p_y} \tag{13}$$

and

$$\text{Var}(S_i) = \frac{k(1-p_y)}{p_y^2}. \tag{14}$$

Also, let there be a sample of $n$ respondents and $\bar{S}$ be the sample mean of the $n$ responses. Then

$$E(\bar{S}) = E(S_i) = \frac{k}{p_y}. \tag{15}$$

Therefore, $k/p_y$ can be estimated by $\bar{S}$ and $\hat{p}_y = k/\bar{S}$ can be used as an estimator of $p_y$. Using first-order Taylor's approximation,

$$\frac{1}{\bar{S}} = \frac{1}{E(S)} + (\bar{S} - E(S))\!\left(-\frac{1}{(E(S))^2}\right), \tag{16}$$

where $E(S) = k/p_y$.

Thus, our estimator for $\pi_x$ in (2) becomes

$$\hat{\pi}_{GI_k} = \frac{k/\bar{S} - \pi_y(1-p)}{p}. \tag{17}$$

From (17), we get the following for the mean of $\hat{\pi}_x$:

$$E(\hat{\pi}_{GI_k}) = \frac{kE(1/\bar{S}) - \pi_y(1-p)}{p}$$

$$\approx \frac{k(p_y/k) - \pi_y(1-p)}{p} = \frac{p_y - \pi_y(1-p)}{p} = \pi_x. \qquad (18)$$

Thus, $\hat{\pi}_{GI_k}$ is an unbiased estimator of $\pi_x$, up to first order of approximation.

From (17), we also get,

$$\text{Var}(\hat{\pi}_{GI_k}) = \frac{k^2}{p^2} \text{Var}\left(\frac{1}{\bar{S}}\right). \qquad (19)$$

But,

$$\text{Var}\left(\frac{1}{\bar{S}}\right) \approx \text{Var}\left(\frac{1}{E(\bar{S})} + (\bar{S} - E(\bar{S}))\left(-\frac{1}{(E(\bar{S}))^2}\right)\right) = \frac{p_y^4}{k^4} \text{Var}(\bar{S})$$

$$\approx \frac{p_y^4}{k^4}\left(\frac{k(1-p_y)/p_y^2}{n}\right) = \frac{p_y^4}{k^4}\left(\frac{k(1-p_y)}{p_y^2 n}\right) = \frac{p_y^2(1-p_y)}{k^3 n}. \qquad (20)$$

Thus, we have

$$\text{Var}(\hat{\pi}_{GI_k}) \approx \frac{k^2}{p^2}\left(\frac{p_y^2(1-p_y)}{k^3 n}\right) = \frac{p_y^2(1-p_y)}{knp^2}. \qquad (21)$$

## 3. Efficiency comparisons

In this section, we compare the efficiencies of the following Greenberg estimators:

$\hat{\pi}_G$ = standard estimator,

$\hat{\pi}_{GM}$ = estimator using $m$ independent responses,

$\hat{\pi}_{GI}$ = estimator using inverse sampling, waiting for the first "yes" response,

$\hat{\pi}_{GI_k}$ = estimator using inverse sampling, waiting for the $k$-th "yes" response.

Since

$$\text{Var}(\hat{\pi}_{GM}) = \frac{p_y(1-p_y)}{nmp^2} = \frac{1}{m}\left(\frac{p_y(1-p_y)}{np^2}\right) = \frac{1}{m}\text{Var}(\hat{\pi}_G), \qquad (22)$$

we have $\text{Var}(\hat{\pi}_{GM}) < \text{Var}(\hat{\pi}_G)$ for $m > 1$. Thus, the Greenberg multiple response model is more efficient than the single response model.

Since

$$\text{Var}(\hat{\pi}_{GI}) = \frac{p_y^2(1-p_y)}{np^2} = p_y\left(\frac{p_y(1-p_y)}{np^2}\right) = p_y \text{Var}(\hat{\pi}_G), \qquad (23)$$

we have $\text{Var}(\hat{\pi}_{GI}) < \text{Var}(\hat{\pi}_G)$ for $p_y < 1$. Thus, the inverse sampling model is more efficient than the Greenberg model.

Since

$$\text{Var}(\hat{\pi}_{GI}) = \frac{p_y^2(1 - p_y)}{np^2} = p_y\left(\frac{p_y(1 - p_y)}{np^2}\right) = mp_y\left(\frac{p_y(1 - p_y)}{nmp^2}\right)$$
$$= mp_y \, \text{Var}(\hat{\pi}_{GM}), \tag{24}$$

we have $\text{Var}(\hat{\pi}_{GI}) < \text{Var}(\hat{\pi}_{GM})$ for $mp_y < 1$. Thus, the inverse sampling model is more efficient than the Greenberg multiple response model when $m < 1/p_y$.

Since

$$\text{Var}(\hat{\pi}_{GI_k}) = \frac{p_y^2(1 - p_y)}{nkp^2} = \frac{1}{k}\left(\frac{p_y^2(1 - p_y)}{np^2}\right) = \frac{1}{k}\,\text{Var}(\hat{\pi}_{GI}), \tag{25}$$

we have $\text{Var}(\hat{\pi}_{GI_k}) < \text{Var}(\hat{\pi}_{GI})$ for $k > 1$. Thus, the inverse sampling model that waits for $k$ "yes" responses is more efficient than the inverse sampling model that waits for the first "yes" response.

We can summarize the above observations as follows:

$$\text{Var}(\hat{\pi}_{GI_k}) < \begin{cases} \text{Var}(\hat{\pi}_{GI}) & \text{if } k > 1, \\ \text{Var}(\hat{\pi}_{GM}) & \text{if } mp_y < 1, \\ \text{Var}(\hat{\pi}_G) & \text{if } m > 1. \end{cases} \tag{26}$$

## 4. Simulation results

All of the preceding theoretical formulas were tested empirically through computer simulations. Table 1 below presents simulation results that were obtained using SAS for a total of 10000 simulations with a sample size of 500, $\pi_x = 0.30$, $\pi_y = 0.7$ and $p = 0.85$. Note that the simulation results support the formulas for the means and variances of various estimators, even when first-order approximation is used.

## 5. Conclusion

Based on Table 1, we can see that the regular Greenberg model has higher variance (theoretical and empirical) than the modified Greenberg model with multiple responses as well as the models based on inverse sampling. Hence, the proposed variants of the Greenberg model are more efficient; although greater effort is needed in using these newer models. Given that the gain in efficiency with newer models is quite substantial, the newer models are worth trying. However, in practice, we need to keep $m$ and $k$ small, such as $m \leq 3$ and $k \leq 3$.

|   | $\hat{\pi}_G$ | $\widehat{\text{Var}}(\hat{\pi}_G)$ | $\text{Var}(\hat{\pi}_G)$ |
|---|---|---|---|
|   | 0.3001781 | 0.000638371 | 0.000637785 |
| $m$ | $\hat{\pi}_{GM}$ | $\widehat{\text{Var}}(\hat{\pi}_{GM})$ | $\text{Var}(\hat{\pi}_{GM})$ |
| 1 | 0.3001781 | 0.000638371 | 0.000637785 |
| 2 | 0.3002513 | 0.000318622 | 0.000318893 |
| 3 | 0.3000282 | 0.000214628 | 0.000212595 |
| 4 | 0.3000342 | 0.000156454 | 0.000159446 |
| 5 | 0.3000586 | 0.000126368 | 0.000127557 |
|   | $\hat{\pi}_{GI}$ | $\widehat{\text{Var}}(\hat{\pi}_{GI})$ | $\text{Var}(\hat{\pi}_{GI})$ |
|   | 0.3004394 | 0.000229236 | 0.000229603 |
| $k$ | $\hat{\pi}_{GI_k}$ | $\widehat{\text{Var}}(\hat{\pi}_{GI_k})$ | $\text{Var}(\hat{\pi}_{GI_k})$ |
| 1 | 0.3004394 | 0.000229236 | 0.000229603 |
| 2 | 0.3002714 | 0.000112837 | 0.000114801 |
| 3 | 0.3000640 | 0.000076382 | 0.000076534 |
| 4 | 0.3000161 | 0.000058789 | 0.000057401 |
| 5 | 0.3002176 | 0.000046028 | 0.000045921 |

**Table 1.** Estimators of $\pi_x$ with corresponding empirical and theoretical variances.

# References

[Abernathy et al. 1970] J. R. Abernathy, B. G. Greenberg, and D. G. Horvitz, "Estimates of induced abortion in urban North Carolina", *Demography* **7**:1 (1970), 19–29.

[Chhabra et al. 2016] A. Chhabra, B. K. Dass, and S. Gupta, "Estimating prevalence of sexual abuse by an acquaintance with an optional unrelated question RRT model", *North Carolina J. Math. Stat.* **2** (2016), 1–9.

[Gill et al. 2013] T. S. Gill, A. Tuck, S. Gupta, M. Crowe, and J. Figueroa, "A field test of optional unrelated question randomized response models: estimates of risky sexual behaviors", pp. 135–146 in *Topics from the 8th annual UNCG regional mathematics and statistics conference*, edited by J. Rychtář et al., Springer Proceedings in Mathematics and Statistics **64**, Springer, New York, 2013.

[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**:326 (1969), 520–539. MR

[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 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 Zbl

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

[Klein and Spady 1993] R. W. Klein and R. H. Spady, "An efficient semiparametric estimator for binary response models", *Econometrica* **61**:2 (1993), 387–421. MR Zbl

[Ostapczuk et al. 2009] M. Ostapczuk, J. Musch, and M. Moshagen, "A randomized-response investigation of the education effect in attitudes towards foreigners", *Eur. J. Soc. Psychol.* **39**:6 (2009), 920–931.

[Striegel et al. 2006] H. Striegel, P. Simon, J. Hansel, A. M. Niess, and R. Ulrich, "Doping and drug use in elite sports: an analysis using the randomized response technique", *Med. Sci. Sports Exerc.* **38**:5 (2006), art. id. S247.

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

[Warner 1971] S. L. Warner, "The linear randomized response model", *J. Amer. Statist. Assoc.* **66**:336 (1971), 884–888.

d_suarez@uncg.edu              *Department of Mathematics and Statistics, University of North Carolina at Greensboro, Greensboro, NC, United States*

sngupta@uncg.edu               *Department of Mathematics and Statistics, University of North Carolina at Greensboro, Greensboro, NC, United States*

# Generalized exponential sums
# and the power of computers

### Francis N. Castro, Oscar E. González and Luis A. Medina

#### (Communicated by Kenneth S. Berenhaut)

Today's era can be characterized by the rise of computer technology. Computers have been, to some extent, responsible for the explosion of the scientific knowledge that we have today. In mathematics, for instance, we have the four color theorem, which is regarded as the first celebrated result to be proved with the assistance of computers. In this article we generalize some fascinating binomial sums that arise in the study of Boolean functions. We study these generalizations from the point of view of integer sequences and bring them to the current computer age of mathematics. The asymptotic behavior of these generalizations is calculated. In particular, we show that a previously known constant that appears in the study of exponential sums of symmetric Boolean functions is universal in the sense that it also emerges in the asymptotic behavior of all of the sequences considered in this work. Finally, in the last section, we use the power of computers and some remarkable algorithms to show that these generalizations are holonomic; i.e., they satisfy homogeneous linear recurrences with polynomial coefficients.

## 1. Introduction

Number theory and combinatorics often offer tantalizing objects that captivate the imaginations of mathematicians. Almost all of us have played with prime numbers, explored open problems like Goldbach's conjecture or drawn a lattice on a paper just to see how Catalan numbers work. Nowadays, computer technology allows us to extend the limits of our knowledge and explore these objects in a way that was almost unimaginable 40 years ago. In this work, we pay close attention to some binomial sums that come from the theory of Boolean functions. These binomial sums emerge when the problem of balancedness of these functions is considered. As it is a common practice in mathematics, the idea in this work is to study these binomial sums in a more general framework. Once the proper framework is established, we use the power of computers to expand our knowledge. We start this work with a

short survey of Boolean functions and exponential sums in an effort to make the manuscript self-contained. The expert reader may skip the majority of it.

A Boolean function is a function from the vector space $\mathbb{F}_2^n$ to $\mathbb{F}_2$, where $\mathbb{F}_2 = \{0, 1\}$ is the binary field and $n$ is some positive integer. These functions are beautiful combinatorial objects with applications to many areas of mathematics as well as outside the discipline. Some examples include combinatorics, electrical engineering, game theory, the theory of error-correcting codes, and cryptography. In the current era, efficient implementations of Boolean functions with many variables is a challenging problem due to memory restrictions of current technology. Because of this, symmetric Boolean functions are good candidates for efficient implementations.

It is known that every Boolean function can be identified with a multivariable polynomial. Let $F(X) = F(X_1, \ldots, X_n)$ be a polynomial in $n$ variables over $\mathbb{F}_2$. Assume that $F(X)$ is not a polynomial in some subset of the variables $X_1, \ldots, X_n$. The exponential sum associated to $F$ over $\mathbb{F}_2$ is

$$S(F) = \sum_{x \in \mathbb{F}_2^n} (-1)^{F(x)}. \tag{1-1}$$

A Boolean function $F(\mathbf{X})$ is called *balanced* if $S(F) = 0$, i.e., the number of zeros and the number of ones are equal in the truth table of $F$. In many applications, especially ones related to cryptography, it is important for Boolean functions to be balanced. Balancedness of Boolean functions is an active area of research with open problems even for the relatively simple symmetric case [Adolphson and Sperber 1987; Cai et al. 1996; Canteaut and Videau 2005; Castro et al. 2015; Castro and Medina 2011; 2014; Cusick and Li 2005; Cusick et al. 2008; 2009; Gao et al. 2011; 2016; Su et al. 2013].

Our interest in this work lies in symmetric Boolean functions and therefore, an important step is to try to see what exponential sums of symmetric Boolean functions look like. Let $\sigma_{n,k}$ denote the elementary symmetric polynomial in $n$ variables of degree $k$. This polynomial is formed by adding together all distinct products of $k$ distinct variables. For example,

$$\sigma_{4,3} = X_1 X_2 X_3 + X_1 X_4 X_3 + X_2 X_4 X_3 + X_1 X_2 X_4. \tag{1-2}$$

Elementary symmetric polynomials are the building blocks of symmetric Boolean functions, as every such function can be identified with an expression of the form

$$\sigma_{n,k_1} + \sigma_{n,k_2} + \cdots + \sigma_{n,k_s}, \tag{1-3}$$

where $0 \le k_1 < k_2 < \cdots < k_s$ are integers. For the sake of simplicity, we use the notation $\sigma_{n,[k_1,\ldots,k_s]}$ to denote (1-3). For example,

$$\sigma_{3,[2,1]} = \sigma_{3,2} + \sigma_{3,1} = X_1 X_2 + X_3 X_2 + X_1 X_3 + X_1 + X_2 + X_3. \tag{1-4}$$

It turns out that exponential sums of symmetric polynomials have nice representations as binomial sums. Define $A_j$ to be the set of all $(x_1, \ldots, x_n) \in \mathbb{F}_2^n$ with exactly $j$ entries equal to 1. Clearly, $|A_j| = \binom{n}{j}$ and by symmetry $\sigma_{n,k}(\boldsymbol{x}) = \binom{j}{k}$ for $\boldsymbol{x} \in A_j$. Therefore,

$$S(\sigma_{n,k}) = \sum_{j=0}^{n} \sum_{\boldsymbol{x} \in A_j} (-1)^{\sigma_{n,k}(\boldsymbol{x})} = \sum_{j=0}^{n} (-1)^{\binom{j}{k}} \binom{n}{j}. \tag{1-5}$$

In general, if $0 \le k_1 < k_2 < \cdots < k_s$ are fixed integers, then

$$S(\sigma_{n,[k_1,\ldots,k_s]}) = \sum_{j=0}^{n} (-1)^{\binom{j}{k_1} + \binom{j}{k_2} + \cdots + \binom{j}{k_s}} \binom{n}{j}. \tag{1-6}$$

Equation (1-6) is a clear computational improvement over (1-1). It also connects the problem of balancedness of symmetric Boolean functions to the intriguing problem of bisecting binomial coefficients; see [Mitchell 1990]. A solution $(\delta_0, \delta_1, \ldots, \delta_n)$ to the equation

$$\sum_{j=0}^{n} \delta_j \binom{n}{j} = 0, \quad \delta_j \in \{-1, 1\}, \tag{1-7}$$

is said to give a *bisection of the binomial coefficients* $\binom{n}{j}$, $0 \le j \le n$. Observe that a solution to (1-7) provides us with two disjoints sets $A$, $B$ such that $A \cup B = \{0, 1, 2, \ldots, n\}$ and

$$\sum_{j \in A} \binom{n}{j} = \sum_{j \in B} \binom{n}{j} = 2^{n-1}. \tag{1-8}$$

The problem of bisecting binomial coefficients is an interesting problem in its own right; however, it is out of the scope of this work.

The identity (1-6) was used by Castro and Medina [2011] to study exponential sums of symmetric Boolean functions from the point of view of integer sequences. As part of their study, they showed that the sequence $\{S(\sigma_{n,[k_1,\ldots,k_s]})\}_{n \in \mathbb{N}}$ satisfies the homogeneous linear recurrence

$$a(n) = \sum_{j=1}^{2^r-1} (-1)^{j-1} \binom{2^r}{j} a(n-j), \tag{1-9}$$

where $r = \lfloor \log_2(k_s) \rfloor + 1$; this result was first proved by Cai, Green and Thierauf [Cai et al. 1996, Theorem 3.1, p. 248]. The characteristic polynomial of (1-9) is given by

$$(t-2)\Phi_4(t-1)\Phi_8(t-1) \cdots \Phi_{2^r}(t-1), \tag{1-10}$$

where $\Phi_n(t)$ represents the $n$-th cyclotomic polynomial. This is very important, as it implies that (1-9) has an embedded nature. Before giving the formal definition

of what we mean by "embedded nature", let us explore recurrence (1-9) in order to have a better understanding of where we want to go with this term. Observe that the exponential sum of every symmetric Boolean function of degree less than 4 satisfies

$$a(n) = \sum_{j=1}^{3} (-1)^{j-1} \binom{4}{j} a(n-j), \tag{1-11}$$

the exponential sum of every symmetric Boolean function of degree less than 8 satisfies

$$a(n) = \sum_{j=1}^{7} (-1)^{j-1} \binom{8}{j} a(n-j), \tag{1-12}$$

the exponential sum of every symmetric Boolean function of degree less than 16 satisfies

$$a(n) = \sum_{j=1}^{15} (-1)^{j-1} \binom{16}{j} a(n-j), \tag{1-13}$$

and so on. This means, for example, that $\{S(\sigma_{n,[7,2]})\}_{n \in \mathbb{N}}$, for which the first few values are given by

2, 4, 6, 8, 12, 24, 58, 144, 344, 784, 1716, 3632, 7464, 14928, 29128, 55680, ...,

must satisfy (1-12) and (1-13), but not (1-11). Next is the formal definition of *embedded recurrences*.

**Definition 1.1.** Let $\{a_{f(x)}(n)\}$ be a family of integer sequences indexed by some polynomial family $\{f(x)\}$. Suppose that every sequence $\{a_{f(x)}(n)\}$ satisfies a linear recurrence. We say that these recurrences are *embedded* if there is a sequence of integers $n_1 < n_2 < n_3 < \cdots$ such that every sequence $\{a_{f(x)}(n)\}$ with the property $\deg(f) < n_l$ satisfies a global recurrence. For example, the sequences of exponential sums of symmetric Boolean functions satisfy recurrences that are embedded. In this case, $n_l = 2^l$ and the global recurrence is (1-9).

Castro and Medina [2011] also computed the asymptotic behavior of $S(\sigma_{n,[k_1,...,k_s]})$ as $n \to \infty$. To be specific, they showed that

$$\lim_{n \to \infty} \frac{1}{2^n} S(\sigma_{n,[k_1,...,k_s]}) = c_0(k_1, \ldots, k_s), \tag{1-14}$$

where

$$c_0(k_1, \ldots, k_s) = \frac{1}{2^r} \sum_{j=0}^{2^r-1} (-1)^{\binom{j}{k_1}+\cdots+\binom{j}{k_s}}. \tag{1-15}$$

They used this limit to show that a conjecture by Cusick, Li and Stănică [Cusick et al. 2008] is true asymptotically. Some of these results, especially recurrence

(1-9) and limit (1-14), were extended to some perturbations of symmetric Boolean functions [Castro and Medina 2014].

In this manuscript, we generalize the concept of exponential sums of symmetric Boolean functions by virtue of the binomial sum in (1-6) and study some of its properties. Let $d$ be a nonnegative integer. We define the *$d$-generalized exponential sum* of $\sigma_{n,[k_1,\ldots,k_s]}$ as the power sum of binomial coefficients given by

$$S_d(\sigma_{n,[k_1,\ldots,k_s]}) = \sum_{j=0}^{n}(-1)^{\binom{j}{k_1}+\cdots+\binom{j}{k_s}}\binom{n}{j}^d. \tag{1-16}$$

In a similar manner, if $Q(x) = a_0 + a_1 x + \cdots + a_t x^t$ is a polynomial, then the $Q(x)$-*generalized exponential sum* of $\sigma_{n,[k_1\ldots,k_s]}$ is defined as

$$S_{Q(x)}(\sigma_{n,[k_1,\ldots,k_s]}) = \sum_{j=0}^{n}(-1)^{\binom{j}{k_1}+\cdots+\binom{j}{k_s}}Q\left(\binom{n}{j}\right). \tag{1-17}$$

By linearity, the study of (1-17) is reduced to the study of (1-16). Thus, emphasis is made on $d$-generalized exponential sums.

It is clear that if $d = 1$, then the $d$-generalized exponential sum is just the regular exponential sum. However, we point out that $d$-generalized exponential sums generalize other combinatorial objects. For instance, when degree 0 is considered, we have $S_d(\sigma_{n,0}) = -f_{n,d}$, where

$$f_{n,d} = \sum_{j=0}^{n}\binom{n}{j}^d \tag{1-18}$$

is the $d$-th order Franel number. When $k = 1$,

$$S_d(\sigma_{n,1}) = \sum_{j=0}^{n}(-1)^j\binom{n}{j}^d \tag{1-19}$$

is the $d$-th order alternate Franel number, for which, when $d = 3$, we have the beautiful identity of Dixon

$$S_3(\sigma_{2n,1}) = \sum_{j=0}^{2n}(-1)^j\binom{2n}{j}^3 = (-1)^n\binom{2n}{n}\binom{3n}{n}. \tag{1-20}$$

Finally, the sequence $\{x_n\}$ defined by $x_0 = 0$, $x_1 = 3$ and $x_n = S_0(\sigma_{n-1,3})$ can be identified with sequence A018837 [Sloane and LeBrun 2008], which represents the minimum number of steps for a knight which starts at position $(0,0)$ to reach $(n,0)$ on an infinite chessboard.

In this article we extend some of the results that appear in [Castro and Medina 2011; 2014] to $d$-generalized exponential sums. In particular, we show that these

sequences satisfy recurrences and, as is the case for $d = 1$, there is an embedded component behind it. We also calculate the asymptotic behavior of these sequences and show that the constant $c_0(k_1, \ldots, k_s)$ is universal in the sense that it appears in the asymptotic behavior of $S_{Q(x)}(\sigma_{n,[k_1,\ldots,k_s]})$ for every polynomial $Q(x)$. The case $d = 0$ turns out to be relatively easy when compared to the case $d \neq 0$, and, as a result, we decided to discuss it now. First, it is clear that $S_0(\sigma_{n,[k_1,\ldots,k_s]}) = O(n)$. Second, if $r = \lfloor \log_2(k_s) \rfloor + 1$, then it satisfies the linear recurrence

$$a(n) = a(n-1) + a(n-2^r) - a(n-2^r-1). \tag{1-21}$$

The characteristic polynomial of (1-21) is given by

$$(t-1)^2 \Phi_2(t) \Phi_4(t) \cdots \Phi_{2^r}(t), \tag{1-22}$$

and therefore, as in the case $d = 1$, these recurrences are embedded. Finally, if $i_1, \ldots, i_p$ are all the integers between 1 and $2^r - 1$ such that $\binom{i}{k_1} + \cdots + \binom{i}{k_s} \equiv 1$ (mod 2), then it is not hard to see that

$$S_0(\sigma_{n,[k_1,\ldots,k_s]}) = n + 1 - 2\left\lceil \frac{n+1-i_1}{2^r} \right\rceil - \cdots - 2\left\lceil \frac{n+1-i_p}{2^r} \right\rceil. \tag{1-23}$$

The asymptotic behavior of $d$-generalized exponential sums is discussed in Section 2. Then, in Section 3, we use computer power to find recurrences for these sums. The reader is invited to use her favorite computer algebra system while reading this manuscript. This is not necessary, as we believe the manuscript is self-contained; however we encourage experimentation because it helps to build intuition and to cement and develop appreciation for mathematical knowledge.

## 2. Asymptotic behavior of the generalized exponential sum

The asymptotic behavior of $S(\sigma_{n,k})$ as $n \to \infty$ was used in [Castro and Medina 2011] to show a conjecture by Cusick, Li and Stănică [Cusick et al. 2008] is true for large $n$. This shows the importance of the behavior of $S(\sigma_{n,[k_1,\ldots,k_s]})$ as $n$ increases. In this section we discuss the asymptotic behavior of $\{S_d(\sigma_{n,[k_1,\ldots,k_s]})\}_{n\in\mathbb{N}}$ and show that the behavior of $\{S_d(\sigma_{n,[k_1,\ldots,k_s]})\}_{\in\mathbb{N}}$, as $n$ increases, is closely related to that of $\{S(\sigma_{n,[k_1,\ldots,k_s]})\}_{n\in\mathbb{N}}$.

We start our discussion with the case $d = 2$ and $k = 3$; that is, we consider the sequence $\{S_2(\sigma_{n,3})\}_{n\in\mathbb{N}}$. The idea for doing this is to gain insight as to what is behind the asymptotic behavior of these sequences. A proof for the general case will be provided later in this section once our intuition is solidified.

The first few values of the sequence $\{S_2(\sigma_{n,3})\}_{n\in\mathbb{N}}$ are given by

$$2, \ 6, \ 18, \ 38, \ 52, \ 124, \ 980, \ 6470, \ 31916, \ 127156, \ \ldots.$$

It is not surprising, knowing already the behavior of $S(\sigma_{n,3})$, that the value of the $n$-th term of the sequence $\{S_2(\sigma_{n,3})\}_{n\in\mathbb{N}}$ increases quite rapidly as $n \to \infty$. Now, by previous knowledge we have that

$$\lim_{n\to\infty} \frac{1}{2^n} S(\sigma_{n,3}) = \frac{1}{2},$$

where $2^n$ is the number of $n$-tuples with 0, 1 entries; thus, it is natural to consider the behavior of $S_2(n, 3)/2^n$. The reader can check via computer experimentation that $S_2(n, 3)/2^n$ seems to diverge to $\infty$, which, if true, it would imply that our sequence increases a rate that is faster than $2^n$. Taking into consideration that in this case $d = 2$, it is not a wild idea to check the behavior of $S_2(\sigma_{n,3})/2^{2n}$. In this case, the reader can convince herself that $S_2(\sigma_{n,3})/4^n \to 0$ as $n \to \infty$. Moreover, experiments on a computer suggest that

$$\lim_{n\to\infty} \frac{1}{4^n} S_2(\sigma_{n,k}) = 0 \tag{2-1}$$

for any positive integer $k$. For example, the values of $S_2(\sigma_{n,7})/4^n$ for $n = 10, 100,$ and 1000 are given by

$$0.148731, \quad 0.0426647, \quad \text{and} \quad 0.0133793,$$

respectively. Thus, it appears that $S_2(\sigma_{n,k})$ increases faster than $2^n$, but slower than $4^n$. So, what is the appropriate behavior?

To answer the question, we start by analyzing the reason behind the behavior of the regular exponential sum $S(\sigma_{n,3})$. Using the definition of $S(\sigma_{n,k})$ in terms of binomial coefficients, we see that

$$S(\sigma_{n,3}) = \sum_{j=0}^{n} (-1)^{\binom{j}{3}} \binom{n}{j} = \sum_{j=0}^{n} \binom{n}{j} - 2\sum_{j=0}^{n} \binom{n}{4j+3}$$

$$= 2^n - 2\sum_{j=0}^{n} \binom{n}{4j+3}. \tag{2-2}$$

Observe that when we divide $S(\sigma_{n,3})$ by $2^n$, we control the contribution of the negative terms. We now do the analogous thing for $S_2(\sigma_{n,3})$. Observe that

$$S_2(\sigma_{n,3}) = \sum_{j=0}^{n} (-1)^{\binom{j}{3}} \binom{n}{j}^2 = \sum_{j=0}^{n} \binom{n}{j}^2 - 2\sum_{j=0}^{n} \binom{n}{4j+3}^2$$

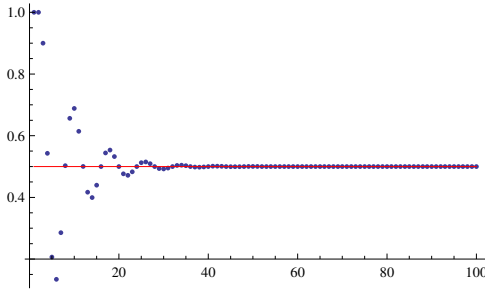$$= \binom{2n}{n} - 2\sum_{j=0}^{n} \binom{n}{4j+3}^2. \tag{2-3}$$

**Figure 1.** Graphical representation of $S_2(\sigma_{n,3})/\binom{2n}{n}$.

Therefore, it is now natural to see that dividing $S_2(\sigma_{n,3})$ by the central binomial coefficient controls the contribution of the negative terms. Figure 1 is a graphical representation of this fact. The dots correspond to $S(\sigma_{n,3})/\binom{2n}{n}$. The line corresponds to $y = \frac{1}{2}$.

It is clear now that $S_2(\sigma_{n,3})$ increases faster than $2^n$, but a bit slower than $4^n$. Its behavior is somewhat similar to that of the central binomial coefficient and by Stirling's formula we know that

$$\binom{2n}{n} \sim \frac{4^n}{\sqrt{\pi n}}. \tag{2-4}$$

Moreover, observe that

$$\lim_{n \to \infty} \binom{2n}{n}^{-1} S_2(\sigma_{n,3}) = \frac{1}{2} = c_0(3). \tag{2-5}$$

Equation (2-5) is not a coincidence, as we will show that $c_0(k_1, \ldots, k_s)$ appears in the behavior of $S_d(\sigma_{n,[k_1,\ldots,k_s]})$. We are now ready to discuss the general case.

Let $d$ be a nonnegative integer. Define $G(n, d)$ as the $d$-th order Franel number

$$G(n, d) = \sum_{j=0}^{n} \binom{n}{j}^d. \tag{2-6}$$

For $d = 0, 1, 2$, the value of $G(n, d)$ is given by

$$G(n, 0) = n + 1, \quad G(n, 1) = 2^n \quad \text{and} \quad G(n, 2) = \binom{2n}{n}. \tag{2-7}$$

Sadly, there is not a nice closed formula for $G(n, d)$ when $d > 2$. Instead, the value of $G(n, d)$ is given by the hypergeometric function

$$G(n, d) = {}_dF_{d-1}\left(-n, -n, \ldots, -n; 1, 1, \ldots, 1; (-1)^n\right). \tag{2-8}$$

The asymptotic behavior of $G(n, d)$ is already known [Pólya and Szegő 1976]:

$$G(n, d) \sim \frac{2^{dn}}{\sqrt{d}} \left(\frac{2}{\pi n}\right)^{(d-1)/2}. \tag{2-9}$$

A formal proof of (2-9) was given by Farmer and Leth [2005]. A treatment for $G(2n, d)$ using Euler's summation formula and the tail-exchange trick appears in [Graham et al. 1994]. Also, a proper adjustment to the proof of Farmer and Leth leads to the following result.

**Lemma 2.1.** *Let m and d be fixed natural numbers and i an integer such that $0 \leq i \leq m$. Then, as n increases, we have*

$$\sum_{j=0}^{n} \binom{n}{mj+i}^{d} \sim \frac{2^{dn}}{m\sqrt{d}} \left(\frac{2}{\pi n}\right)^{(d-1)/2} \sim \frac{1}{m} G(n, d). \qquad (2\text{-}10)$$

With Lemma 2.1 at hand, we are now ready to provide the asymptotic behavior of $S_d(\sigma_{n,[k_1,\dots,k_s]})$.

**Theorem 2.2.** *Let d and $k_1 < \cdots < k_s$ be fixed positive integers. Then,*

$$\lim_{n\to\infty} \frac{S_d(\sigma_{n,[k_1,\cdots,k_s]})}{G(n, d)} = c_0(k_1, \dots, k_s). \qquad (2\text{-}11)$$

*Proof.* Let $r = \lfloor \log_2(k_s) \rfloor + 1$. Let $i_1, \dots, i_p$ be all the integers between 1 and $2^r - 1$ such that $\binom{i}{k_1} + \cdots + \binom{i}{k_s} \equiv 1 \pmod 2$. It is known, see [Castro and Medina 2011], that the sequence $\{\binom{n}{k_1} + \cdots + \binom{n}{k_s} \pmod 2\}_{n\in\mathbb{N}}$ is periodic and the period is a divisor of $2^r$. Therefore, $\binom{i}{k_1} + \cdots + \binom{i}{k_s} \equiv 1 \pmod 2$ if and only if $i \equiv i_l \pmod{2^r}$ for some $i_l \in \{i_1, \dots, i_p\}$.

Using the definition of $S_d(\sigma_{n,[k_1,\dots,k_s]})$ we observe that

$$S_d(\sigma_{n,[k_1,\dots,k_s]}) = G(n, d) - 2\sum_{j=0}^{n}\left[\binom{n}{2^r \cdot j+i_1}^{d} + \cdots + \binom{n}{2^r \cdot j+i_p}^{d}\right]. \qquad (2\text{-}12)$$

Therefore, as $n \to \infty$, we have

$$S_d(\sigma_{n,[k_1,\sigma,k_s]}) \sim G(n, d) - \frac{2p}{2^r} G(n, d) = (1 - p \cdot 2^{1-r}) G(n, d). \qquad (2\text{-}13)$$

It is not hard to show that $c_0(k_1, \dots, k_s) = 1 - p \cdot 2^{1-r}$. $\qquad \square$

Using the asymptotic behavior (2-9), we obtain the following corollary.

**Corollary 2.3.** *Let d and $k_1 < \cdots < k_s$ be positive integers. Then,*

$$\lim_{n\to\infty} \frac{(\sqrt{n})^{d-1} \cdot S_d(\sigma_{n,[k_1,\dots,k_s]})}{2^{dn}} = \frac{1}{\sqrt{d}}\left(\frac{2}{\pi}\right)^{(d-1)/2} c_0(k_1, \dots, k_s). \qquad (2\text{-}14)$$

*More generally, if $Q(x) = a_0 + a_1 x + \cdots + a_t x^t$ is a polynomial and*

$$A_{Q(x)}(n) = a_0 \cdot (n+1) + a_1 \cdot 2^n + \frac{a_2}{\sqrt{2}}\left(\frac{2}{\pi \cdot n}\right)^{1/2} 2^{2n} + \cdots + \frac{a_t}{\sqrt{t}}\left(\frac{2}{\pi \cdot n}\right)^{(t-1)/2} 2^{tn}, \qquad (2\text{-}15)$$

*then,*

$$\lim_{n\to\infty} \frac{S_{Q(x)}(\sigma_{n,[k_1,\ldots,k_s]})}{A_{Q(x)}(n)} = c_0(k_1,\ldots,k_s). \tag{2-16}$$

*Proof.* This is a direct consequence of Theorem 2.2 and the asymptotic behavior of $G(n,d)$. $\qquad\square$

**Example 2.4.** Consider the case $d = 4$ and $k = 7$. We know that $c_0(7) = \frac{3}{4}$. Thus,

$$\frac{1}{\sqrt{d}}\left(\frac{2}{\pi}\right)^{(d-1)/2} c_0(k) = \frac{3}{8}\left(\frac{2}{\pi}\right)^{3/2} \approx 0.1904809078\ldots. \tag{2-17}$$

Note that

| $n$ | $(\sqrt{n})^3 S_4(\sigma_{n,7})/2^{4n}$ |
|---|---|
| 1 | 0.1250000000 |
| 10 | 0.2280899652 |
| 100 | 0.2021752897 |
| 1000 | 0.1903737868 |
| 10000 | 0.1904701935 |
| 100000 | 0.1904798364 |

**Example 2.5.** Let $k_1 = 2$, $k_2 = 3$, $k_3 = 4$, and $k_4 = 5$. Consider the polynomial $Q(x) = x^3 + 5x + 2$. The reader can check that in this case we have

$$A_{Q(x)}(n) \cdot c_0(2,3,4,5) = \frac{1}{2}\left(5 \cdot 2^n + 2(n+1) + \frac{2^{3n+1}}{\sqrt{3\pi n}}\right). \tag{2-18}$$

Corollary 2.3 states that

$$\lim_{n\to\infty} \frac{S_{Q(x)}(\sigma_{n,[2,3,4,5]})}{A_{Q(x)}(n) \cdot c_0(2,3,4,5)} = 1. \tag{2-19}$$

Figure 2 is a graphical representation of (2-19).
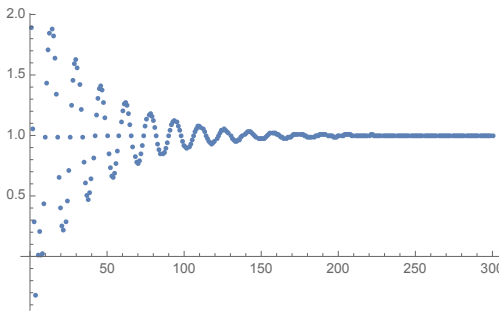


**Figure 2.** Graphical representation of $S_{Q(x)}(\sigma_{n,[2,3,4,5]})/(A_{Q(x)}(n) \cdot c_0(2,3,4,5))$ when $Q(x) = x^3 + 5x + 2$.

We conclude this section with the observation that Theorem 2.2 and Corollary 2.3 imply that the constant $c_0(k_1, \ldots, k_s)$ is universal in the sense that it appears in the asymptotic behavior of $d$-generalized exponential sums. Moreover, Theorem 2.2 is the natural generalization of limit (1-14). In the next section, we explore a generalization to recurrence (1-9).

## 3. Recurrence relations: some experiments

In this section we discuss recurrence relations for the sequences $\{S_d(\sigma_{n,[k_1,\ldots,k_s]})\}_{n\in\mathbb{N}}$. We already know that for $d = 1$, i.e., for $\{S(\sigma_{n,[k_1,\ldots,k_s]})\}_{n\in\mathbb{N}}$, we have the homogeneous linear recurrence with constant coefficients

$$a(n) = \sum_{m=1}^{2^r-1} (-1)^{m-1} \binom{2^r}{m} a(n-m), \tag{3-1}$$

where $r = \lfloor \log_2(k_s) \rfloor + 1$. See [Cai et al. 1996; Castro and Medina 2011; 2014] for more details. Experiments show that something similar happens for $\{S_d(\sigma_{n,[k_1,\ldots,k_s]})\}_{n\in\mathbb{N}}$ when $d > 1$; i.e., these sequences satisfy linear recurrences. However, as we will see, the coefficients of these recurrences are no longer constant; instead, they are polynomials in $n$. In other words, these sequences seems to be *holonomic* (this should not come as a surprise to the expert reader or to the reader aware of the work of Franel [1894; 1895] and Cusick [1989] on power sums of binomial coefficients). Therefore, for $d > 1$, the problem of finding the minimal recurrence is a hard one. Once again, the reader is encouraged to open her favorite computer algebra system while reading this section.

To show the difficulty of the problem at hand, let us consider (once again) the rather simple example $\{S_2(\sigma_{n,3})\}_{n\in\mathbb{N}}$. Note that

$$S_2(\sigma_{n,3}) = \sum_{j=0}^{n} (-1)^{\binom{j}{3}} \binom{n}{j}^2 = \binom{2n}{n} - 2\sum_{j=0}^{n} \binom{n}{4j+3}^2. \tag{3-2}$$

We already know that the central binomial coefficient satisfies a linear recurrence with nonconstant coefficients; i.e., it satisfies the recurrence

$$(n+1)a(n+1) - (4n+2)a(n) = 0. \tag{3-3}$$

Thus, it is natural to expect that if this sequence satisfies a linear recurrence, then the coefficients of the recurrence are nonconstant.

In order to find such a recurrence, we emulate what we already know about the case $d = 1$. In that case, we have

$$S(\sigma_{n,3}) = \sum_{j=0}^{n} (-1)^{\binom{j}{3}} \binom{n}{j} = 2^n - 2\sum_{j=0}^{n} \binom{n}{4j+3}. \tag{3-4}$$

The "negative" part of it, i.e., $\sum_{j=0}^{n}\binom{n}{4j+3}$, satisfies the homogeneous recurrence

$$a(n) = 4a(n-1) - 6a(n-2) + 4a(n-3). \tag{3-5}$$

It is not hard to see that 2 is a root of the characteristic polynomial of recurrence (3-5) and so $2^n$ also satisfies it. Thus, $\{S(\sigma_{n,3})\}_{n\in\mathbb{N}}$ satisfies (3-5).

In general, if $1 \le k_1 < \cdots < k_s$ are integers, $r = \lfloor \log_2(k_s) \rfloor + 1$, and $i_1, \ldots, i_p$ are all integers between 1 and $2^r - 1$ such that $\binom{i}{k_1} + \cdots + \binom{i}{k_s} \equiv 1 \pmod 2$, then

$$S(\sigma_{n,[k_1,\ldots,k_s]}) = 2^n - 2 \sum_{j=0}^{n}\left(\binom{n}{2^r j + i_1} + \cdots + \binom{n}{2^r j + i_p}\right), \tag{3-6}$$

and the negative part of (3-6) satisfies (3-1). Since 2 is a root of the characteristic polynomial of (3-1), we know $2^n$, and therefore $\{S(\sigma_{n,[k_1,\ldots,k_s]})\}_{n\in\mathbb{N}}$, satisfy (3-1).

Emulating what we did in the above paragraph, we start by looking for a recurrence for

$$\sum_{j=0}^{n}\binom{n}{4j+3}^2. \tag{3-7}$$

It is at this stage that we use the power of computers. This power, of course, is assisted by the ingenuity of a great mathematician, in this case, the great combinatorialist Doron Zeilberger [1990a]. Zeilberger's algorithm is already a built-in function in Maple and a version for Mathematica can be found at http://www.risc.jku.at/research/combinat/risc/software. Using it we obtain (with an automated proof!) that (3-7) satisfies the homogeneous linear recurrence with nonconstant coefficients

$$\sum_{j=0}^{7} p_j(n) a(n+j) = 0, \tag{3-8}$$

where the polynomials $p_j(n)$ can be found in the online supplement. Analogous to $2^n$ for $d = 1$, the central binomial coefficient satisfies (3-8). Thus, $\{S_2(\sigma_{n,3})\}_{n\in\mathbb{N}}$ satisfies (3-8).

Zeilberger's algorithm also proves that the sequences

$$\sum_{j=0}^{n}\binom{n}{4j+i}^2 \tag{3-9}$$

for $i = 0, 1, 2, 3$, satisfy (3-8) too. Since we have that

$$S_2(\sigma_{n,2}) = \binom{2n}{n} - 2\sum_{j=0}^{n}\left(\binom{n}{4j+2}^2 + \binom{n}{4j+3}^2\right),$$

$$S_2(\sigma_{n,[2,1]}) = \binom{2n}{n} - 2\sum_{j=0}^{n}\left(\binom{n}{4j+1}^2 + \binom{n}{4j+2}^2\right),$$

$$S_2(\sigma_{n,[3,2]}) = \binom{2n}{n} - 2\sum_{j=0}^{n}\binom{n}{4j+2}^2,$$

$$S_2(\sigma_{n,[3,1]}) = \binom{2n}{n} - 2\sum_{j=0}^{n}\binom{n}{4j+1}^2,$$

$$S_2(\sigma_{n,[3,2,1]}) = \binom{2n}{n} - 2\sum_{i=1}^{3}\sum_{j=0}^{n}\binom{n}{4j+i}^2,$$

all of them satisfy (3-8). In fact, for $4 \leq k_s \leq 7$, the sequence $\{S_2(\sigma_{n,[k_1,\ldots,k_s]})\}$ satisfies a recurrence of order 15 with polynomial coefficients. Moreover, every sequence $\{S_2(\sigma_{n,[k_1,\ldots,k_s]})\}$ with $1 \leq k_s \leq 7$ satisfies this recurrence of order 15. This pattern seems to hold for higher $k_s$ and any $d > 2$. If this holds true, then, as in the cases $d = 0$ and $d = 1$, these sequences satisfy recurrences that are embedded.

The expert reader may notice that it is not hard to show that $d$-generalized exponential sums, and therefore $Q(x)$-generalized exponential sums, are indeed holonomic. This follows from the fact that binomial coefficients are holonomic in both variables and from some closure properties of these sequences; a great read on this subject is [Zeilberger 1990b]. A formal proof, however, will require a proper discussion on holonomic sequences and this is out of the scope of this work.

The natural question now is: can we show that the recurrences are embedded? The answer is yes! Suppose that a sequence $\{a(n)\}$ is holonomic; that is, suppose that there exist polynomials $p_0(n), p_1(n), \ldots, p_l(n) \in \mathbb{C}[n]$ such that

$$p_l(n)a(n+l) + p_{l-1}a(n+l-1) + \cdots + p_0(n)a(n) = 0. \qquad (3\text{-}10)$$

Let $E$ be the *shift operator* that maps $a(n)$ to $a(n+1)$. Equation (3-10) can be written as $A(E)(a(n)) = 0$, where

$$A(E) = \sum_{j=0}^{l} p_j(n)E^l. \qquad (3\text{-}11)$$

The operator $A(E)$ is called an *annihilating operator* of the sequence $\{a(n)\}$. The number $l$ is called the *order* of the annihilating operator. It is not hard to see that the set of all annihilating operators of $\{a(n)\}$ forms an ideal of the ring $\mathbb{C}[n][E]$.

Consider the sequence

$$a_{d,r,i}(n) = \sum_{j=0}^{n}\binom{n}{2^r j+i}^d. \qquad (3\text{-}12)$$

Let $A_{d,r,i}(E) \in \mathbb{C}[n][E]$ be an annihilating operator for $\{a_{d,r,i}(n)\}$. Define

$$A_{d,r}(E) = \prod_{i=0}^{2^r-1} A_{d,r,i}(E). \qquad (3\text{-}13)$$

Since the set of all annihilating operators of a sequence $\{a(n)\}$ is an ideal, we know $A_{d,r}(E)(a_{d,r,i}(n)) = 0$ for every $r$, $d$ and $i$. Also, since

$$G(n, d) = \sum_{i=0}^{2^r-1} a_{d,r,i}(n), \tag{3-14}$$

we have $A_{d,r}(E)(G(n, d)) = 0$. Finally, if $r = \lfloor \log_2(k_s) \rfloor + 1$, then $S_d(\sigma_{n,[k_1,\ldots,k_s]})$ is a linear combination of $G(n, d)$ and some terms $a_{d,r,i}(n)$; therefore

$$A_{d,r}(E)(S_d(\sigma_{n,[k_1,\ldots,k_s]})) = 0 \tag{3-15}$$

and so the recurrences are embedded. To be specific, for every symmetric Boolean function of degree less than 4, the $d$-generalized exponential sum satisfies

$$A_{d,2}(E)(a(n)) = 0, \tag{3-16}$$

for every symmetric Boolean function of degree less than 8, the $d$-generalized exponential sum satisfies

$$A_{d,3}(E)(a(n)) = 0, \tag{3-17}$$

and so on.

We finish this section by noticing that the recurrences included in this work are not necessarily the minimal ones. For instance, we know that $\{S_2(\sigma_{n,2})\}_{n\in\mathbb{N}}$ satisfies (3-8). However, using the Mathematica implementation GuessMinRE, which is part of the package Guess.m written by Manuel Kauers, available at http://www.risc.jku.at/research/combinat/risc/software, we guess that $\{S_2(\sigma_{n,2})\}_{n\in\mathbb{N}}$ satisfies the recurrence

$$\sum_{j=0}^{4} q_j(n)a(n + j) = 0, \tag{3-18}$$

where

$$q_0(n) = 424 + 924n + 692n^2 + 216n^3 + 24n^4,$$
$$q_1(n) = 1280 + 2352n + 1576n^2 + 456n^3 + 48n^4,$$
$$q_2(n) = 1600 + 2780n + 1756n^2 + 480n^3 + 48n^4,$$
$$q_3(n) = -960 - 1604n - 968n^2 - 252n^3 - 24n^4,$$
$$q_4(n) = 276 + 449n + 263n^2 + 66n^3 + 6n^4.$$

This has been checked for values of $n$ up to 20000.

## Acknowledgments

# References

[Adolphson and Sperber 1987] A. Adolphson and S. Sperber, "*p*-adic estimates for exponential sums and the theorem of Chevalley–Warning", *Ann. Sci. École Norm. Sup.* (4) **20**:4 (1987), 545–556. MR Zbl

[Cai et al. 1996] J.-Y. Cai, F. Green, and T. Thierauf, "On the correlation of symmetric functions", *Math. Systems Theory* **29**:3 (1996), 245–258. MR Zbl

[Canteaut and Videau 2005] A. Canteaut and M. Videau, "Symmetric Boolean functions", *IEEE Trans. Inform. Theory* **51**:8 (2005), 2791–2811. MR Zbl

[Castro and Medina 2011] F. N. Castro and L. A. Medina, "Linear recurrences and asymptotic behavior of exponential sums of symmetric Boolean functions", *Electron. J. Combin.* **18**:2 (2011), art. id. 8, 21 pp. MR Zbl

[Castro and Medina 2014] F. N. Castro and L. A. Medina, "Asymptotic behavior of perturbations of symmetric functions", *Ann. Comb.* **18**:3 (2014), 397–417. MR Zbl

[Castro et al. 2015] F. N. Castro, O. E. González, and L. A. Medina, "A divisibility approach to the open boundary cases of Cusick–Li–Stănică's conjecture", *Cryptogr. Commun.* **7**:4 (2015), 379–402. MR Zbl

[Cusick 1989] T. W. Cusick, "Recurrences for sums of powers of binomial coefficients", *J. Combin. Theory Ser. A* **52**:1 (1989), 77–83. MR Zbl

[Cusick and Li 2005] T. W. Cusick and Y. Li, "*k*-th order symmetric SAC Boolean functions and bisecting binomial coefficients", *Discrete Appl. Math.* **149**:1-3 (2005), 73–86. MR Zbl

[Cusick et al. 2008] T. W. Cusick, Y. Li, and P. Stănică, "Balanced symmetric functions over GF(*p*)", *IEEE Trans. Inform. Theory* **54**:3 (2008), 1304–1307. MR Zbl

[Cusick et al. 2009] T. W. Cusick, Y. Li, and P. Stănică, "On a conjecture for balanced symmetric Boolean functions", *J. Math. Cryptol.* **3**:4 (2009), 273–290. MR Zbl

[Farmer and Leth 2005] J. D. Farmer and S. C. Leth, "An asymptotic formula for powers of binomial coefficients", *Math. Gazette* **89**:516 (2005), 385–391.

[Franel 1894] J. Franel, in reply to question 42 by Laisant, *L'Int. Math.* **1** (1894), 45–47. In French.

[Franel 1895] J. Franel, in reply to question 170 by Laisant, *L'Int. Math.* **2** (1895), 33–35. In French.

[Gao et al. 2011] G.-P. Gao, W.-F. Liu, and X.-Y. Zhang, "The degree of balanced elementary symmetric Boolean functions of $4k + 3$ variables", *IEEE Trans. Inform. Theory* **57**:7 (2011), 4822–4825. MR

[Gao et al. 2016] G. Gao, Y. Guo, and Y. Zhao, "Recent results on balanced symmetric Boolean functions", *IEEE Trans. Inform. Theory* **62**:9 (2016), 5199–5203. MR

[Graham et al. 1994] R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete mathematics: a foundation for computer science*, 2nd ed., Addison-Wesley, Reading, MA, 1994. MR Zbl

[Mitchell 1990] C. Mitchell, "Enumerating Boolean functions of cryptographic significance", *J. Cryptology* **2**:3 (1990), 155–170. MR Zbl

[Pólya and Szegő 1976] G. Pólya and G. Szegő, *Problems and theorems in analysis, II*, Die Grundlehren der Math. Wissenschaften **216**, Springer, New York, 1976. MR Zbl

[Sloane and LeBrun 2008] N. Sloane and M. LeBrun, "Number of steps for knight to reach $(n, 0)$ on infinite chessboard", pp. A018837 in *The online encyclopedia of integer sequences*, 2008.

[Su et al. 2013] W. Su, X. Tang, and A. Pott, "A note on a conjecture for balanced elementary symmetric Boolean functions", *IEEE Trans. Inform. Theory* **59**:1 (2013), 665–671. MR

[Zeilberger 1990a]  D. Zeilberger, "A fast algorithm for proving terminating hypergeometric identities", *Discrete Math.* **80**:2 (1990), 207–211. MR Zbl

[Zeilberger 1990b]  D. Zeilberger, "A holonomic systems approach to special functions identities", *J. Comput. Appl. Math.* **32**:3 (1990), 321–368. MR Zbl

franciscastr@gmail.com          *Department of Mathematics, University of Puerto Rico, San Juan, Puerto Rico*

oscar.gonzalez3@upr.edu          *Department of Mathematics, University of Puerto Rico, San Juan, Puerto Rico*

luis.medina17@upr.edu          *Department of Mathematics, University of Puerto Rico, San Juan, Puerto Rico*

# Coincidences among skew stable and dual stable Grothendieck polynomials

Ethan Alwaise, Shuli Chen, Alexander Clifton, Rebecca Patrias,
Rohil Prasad, Madeline Shinners and Albert Zheng

(Communicated by Jim Haglund)

The question of when two skew Young diagrams produce the same skew Schur function has been well studied. We investigate the same question in the case of stable Grothendieck polynomials, which are the $K$-theoretic analogues of the Schur functions. We prove a necessary condition for two skew shapes to give rise to the same dual stable Grothendieck polynomial. We also provide a necessary and sufficient condition in the case where the two skew shapes are ribbons.

## 1. Introduction

It is well known that the Schur functions indexed by the set of partitions $\{s_\lambda\}$ form a linear basis for the ring of symmetric functions over $\mathbb{Z}$. However, for general skew shapes $\lambda/\mu$, the corresponding Schur functions are no longer linearly independent. In fact, two different skew shapes can give rise to the same Schur function. Such skew shapes are called *Schur equivalent*. There are trivial examples of such equivalences—for instance $\langle 2 \rangle$ is clearly Schur-equivalent to $\langle 4 \rangle/\langle 2 \rangle$ as they yield the same shape positioned differently in space—and there are also many nontrivial examples (note that we use angled brackets here to denote a partition instead of parentheses to avoid ambiguity with later notation). For example, the shapes shown below are Schur equivalent [Reiner et al. 2007].
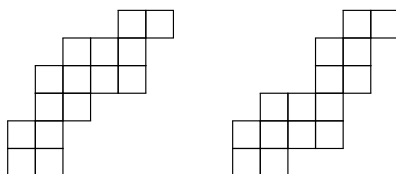


It is natural to ask when these coincidences occur. One application of this type of result involves the representation theory of $\mathrm{GL}_N(\mathbb{C})$. In this setting, equality

among skew Schur functions corresponds to equivalence of certain $GL_N(\mathbb{C})$ modules [Reiner et al. 2007]. Coincidences among skew Schur functions have been studied by Billera–Thomas–van Willigenburg [Billera et al. 2006], Reiner–Shaw–van Willigenburg [Reiner et al. 2007], and McNamara–van Willigenburg [2009], among others.

The stable and dual stable Grothendieck polynomials are natural ($K$-theoretic) analogues of Schur functions obtained as weighted generating functions over *set-valued tableaux* and *reverse plane partitions*, respectively [Buch 2002; Lam and Pylyavskyy 2007]. Roughly speaking, while the Schur functions give information about the cohomology of the Grassmannian, these analogues give information about the $K$-theory of the Grassmannian, where $K$-theory is a generalized cohomology theory. Our work concerns the combinatorics of these objects, so knowledge of cohomology theories is not necessary.

The question of coincidences among stable and dual stable Grothendieck polynomials of skew shapes was previously unstudied. After a brief background in symmetric functions, we focus on dual stable Grothendieck polynomials of ribbon shape $g_\alpha$, where a ribbon is a connected Young diagram containing no $2 \times 2$ square. For a ribbon shape $\alpha$, let $\alpha^*$ denote the shape obtained by 180-degree rotation. We prove the following theorem.

**Theorem 3.4.** *For ribbons $\alpha$ and $\beta$, we have $g_\alpha = g_\beta$ if and only if $\alpha = \beta$ or $\alpha = \beta^*$.*

We next prove two necessary conditions for dual stable Grothendieck equivalence involving *bottleneck numbers* of shape $\lambda/\mu$, $b_i^{\lambda/\mu}$.

**Theorem 3.9.** *Suppose $g_{\lambda/\mu} = g_{\gamma/\nu}$. Then*

$$b_i^{\lambda/\mu} + b_{n-i+1}^{\lambda/\mu} = b_i^{\gamma/\nu} + b_{n-i+1}^{\gamma/\nu}$$

*for $i = 1, 2, \ldots, n$, where $n$ is the number of columns in $\lambda/\mu$.*

**Corollary 3.15.** *Suppose $g_{\lambda/\mu} = g_{\gamma/\nu}$. Then*

$$\sum_{i=1}^{n} (b_i^{\lambda/\mu})^2 = \sum_{i=1}^{n} (b_i^{\gamma/\nu})^2.$$

In Section 4, we prove the following result for stable Grothendieck polynomials, where $A^t$ is the transpose or conjugate of skew shape $A$.

**Theorem 4.2.** *If $G_A = G_B$ for skew shapes $A$ and $B$, then $G_{A^t} = G_{B^t}$.*

We end by giving examples that show that stable Grothendieck equivalence does not imply dual stable Grothendieck equivalence and vice versa and by highlighting areas for future research.

## 2. Preliminaries

***Partitions and tableaux.*** A *partition* $\lambda = \langle \lambda_1, \lambda_2, \ldots, \lambda_k \rangle$ of a positive integer $n$ is a weakly decreasing sequence of positive integers $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k > 0$ whose sum is $n$. The integer $\lambda_i$ is called the $i$-th *part* of $\lambda$. We call $n$ the *size* of $\lambda$, denoted by $|\lambda| = n$. Throughout this document $\lambda$ will refer to a partition. We may visualize a partition $\lambda$ using a *Young diagram*: a collection of left-justified boxes where the $i$-th row from the top has $\lambda_i$ boxes. For example, the Young diagram of $\lambda = \langle 5, 2, 1, 1 \rangle$ is shown below.
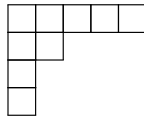


A *skew shape* $\lambda/\mu$ is a pair of partitions $\lambda = \langle \lambda_1, \ldots, \lambda_m \rangle$ and $\mu = \langle \mu_1, \ldots, \mu_k \rangle$ such that $k \leq m$ and $\mu_i \leq \lambda_i$ for all $i$. We form the Young diagram of a skew shape $\lambda/\mu$ by superimposing the Young diagrams of $\lambda$ and $\mu$ and removing the boxes that are contained in both. If $\mu$ is empty, $\lambda/\mu = \lambda$ is called a *straight shape*. Given a skew shape $\lambda/\mu$, we define its *antipodal rotation* $(\lambda/\mu)^*$ as the skew shape obtained by rotating the Young diagram of $\lambda/\mu$ by 180 degrees. For example, the Young diagrams of the skew shapes $\langle 6, 3, 1 \rangle/\langle 3, 1 \rangle$ and $(\langle 6, 3, 1 \rangle/\langle 3, 1 \rangle)^*$ are shown below.

$$\langle 6, 3, 1 \rangle/\langle 3, 1 \rangle = \quad\quad\quad\quad (\langle 6, 3, 1 \rangle/\langle 3, 1 \rangle)^* = $$
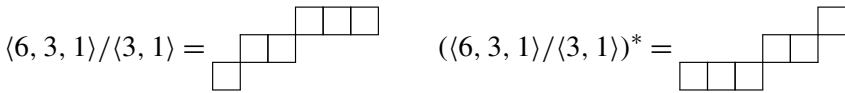


A *semistandard Young tableau* of shape $\lambda/\mu$ is a filling of the boxes of the Young diagram of $\lambda/\mu$ with positive integers such that the entries weakly increase from left to right across rows and strictly increase from top to bottom down columns. Two semistandard Young tableaux are shown below.
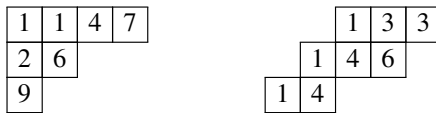


A *set-valued tableau* of shape $\lambda/\mu$ is a filling of the boxes of the Young diagram of $\lambda/\mu$ with finite, nonempty sets of positive integers such that the entries weakly increase from left to right across rows and strictly increase from top to bottom down columns. For two sets of positive integers $A$ and $B$, we say that $A \leq B$ if $\max A \leq \min B$ and $A < B$ if $\max A < \min B$. For a set-valued tableau $T$, we define $|T|$, the *size* of $T$, to be the sum of the sizes of the sets appearing as entries in $T$.

For example,

| 1, 2 | 2, 3 | 6 | 9 |
|------|------|---|---|
| 3 | 5 | | |
| 6 | 6, 7 | | |

is a set-valued tableau of shape $\lambda = \langle 4, 2, 2 \rangle$ and size 11.

A *reverse plane partition* (RPP) of shape $\lambda/\mu$ is a filling of the boxes of the Young diagram of $\lambda/\mu$ with positive integers such that the entries weakly increase both from left to right across rows and from top to bottom down columns. For example,

|   |   | 1 | 1 | 2 | 7 |
|---|---|---|---|---|---|
|   |   | 1 | 2 | 2 | 8 |
|   | 1 | 2 | 2 | 2 | |

is a reverse plane partition of shape $\langle 5, 5, 4 \rangle / \langle 1, 1 \rangle$.

***Symmetric functions.*** To each of the above fillings of a Young diagram we may associate a monomial as follows. First, let $T$ be a semistandard or set-valued tableau. We associate a monomial $x^T$ given by

$$x^T = \prod_{i \in \mathbb{N}} x_i^{m_i},$$

where $m_i$ is the number of times the integer $i$ appears as an entry in $T$. For example, the semistandard Young tableaux shown above correspond to monomials $x_1^2 x_2 x_4 x_6 x_7 x_9$ and $x_1^3 x_3^2 x_4^2 x_6$, respectively, while the set-valued tableau corresponds to monomial $x_1 x_2^2 x_3^2 x_5 x_6^3 x_7 x_9$.

Given a reverse plane partition $P$, the associated monomial $x^P$ is given by

$$x^P = \prod_{i \in \mathbb{N}} x_i^{m_i},$$

where $m_i$ is the number of columns of $P$ that contain the integer $i$ as an entry. The reverse plane partition shown above has monomial $x_1^3 x_2^3 x_7 x_8$.

We can now define the Schur functions, the stable Grothendieck polynomials, and the dual stable Grothendieck polynomials, which are all indexed by skew shapes.

We define the *Schur function* $s_{\lambda/\mu}$ by

$$s_{\lambda/\mu} = \sum_T x^T,$$

where we sum over all semistandard Young tableaux of shape $\lambda/\mu$. Note that entries may be any positive integer, so $s_{\lambda/\mu}$ will be an infinite sum where each term has

degree $|\lambda/\mu| = |\lambda| - |\mu|$. For example,

$$s_{\langle 1 \rangle} = x_1 + x_2 + x_3 + x_4 + \cdots,$$

and

$$s_{\langle 2,1 \rangle} = x_1^2 x_2 + x_1^2 x_3 + x_2^2 x_3 + \cdots + 2x_1 x_2 x_3 + 2x_1 x_2 x_4 + \cdots + 2x_4 x_8 x_{101} + \cdots.$$

Though is it not obvious from this combinatorial definition, the Schur functions are symmetric functions. In other words, each $s_{\lambda/\mu}$ is unchanged after permuting any finite subset of the infinite variable set $\{x_1, x_2, \ldots\}$. Moreover, the Schur functions indexed by straight shapes $\{s_\lambda\}$ form a basis for the ring of symmetric functions over $\mathbb{Z}$. These functions arise naturally in areas like algebraic combinatorics, representation theory, and Schubert calculus. We refer the interested reader to [Stanley 1999] for further reading on Schur functions and symmetric functions.

We next define the *stable Grothendieck polynomial*, the first of two *K-theoretic analogues* of the Schur functions. We direct the interested reader to [Buch 2002] for more on this topic and for an explanation of the connection to $K$-theory. The stable Grothendieck polynomial $G_{\lambda/\mu}$ is defined by

$$G_{\lambda/\mu} = \sum_T (-1)^{|T|-|\lambda|} x^T,$$

where we sum over all set-valued tableaux of shape $\lambda/\mu$.

Note that semistandard tableaux are set-valued tableaux where each subset has size one. It follows that each $G_{\lambda/\mu}$ will be a sum of $s_{\lambda/\mu}$ plus terms of degree greater than $|\lambda/\mu|$. While each term in a Schur function has the same degree, each stable Grothendieck polynomial is an infinite sum where terms have arbitrarily large degree. For example,

$$G_{\langle 1 \rangle} = x_1 + x_2 + \cdots - x_1 x_2 - x_2 x_3 + \cdots + x_1 x_2 x_4 x_5 x_9 + \cdots,$$

and

$$G_{\langle 2,2 \rangle / \langle 1 \rangle} = x_1^2 x_2 + 2x_1 x_2 x_3 + \cdots - 3x_1^2 x_2 x_3 - 8x_2 x_5 x_9 x_{114} - \cdots + 2x_1^2 x_2^2 x_3 + \cdots.$$

The other natural $K$-theoretic analogue of the Schur function is the *dual stable Grothendieck polynomial*. It is dual to the stable Grothendieck polynomial under the Hall inner product. We refer the reader to [Lam and Pylyavskyy 2007] for more background. We define the dual stable Grothendieck polynomial $g_{\lambda/\mu}$ by

$$g_{\lambda/\mu} = \sum_P x^P,$$

where the sum is over all reverse plane partitions of shape $\lambda/\mu$.

Again, note that semistandard Young tableaux are examples of reverse plane partitions where the columns are strictly increasing. As a result, each dual stable Grothendieck polynomial $g_{\lambda/\mu}$ is a sum of the Schur function indexed by the same shape $s_{\lambda/\mu}$ and terms of degree strictly less than $|\lambda/\mu|$. They are again infinite sums, but now each term has degree at most $|\lambda/\mu|$ and at least the number of columns in shape $\lambda/\mu$. For example,

$$g_{(2,1)} = x_1^2 x_2 + 2x_1 x_2 x_3 + \cdots + x_1 x_2 + x_1 x_3 + \cdots + x_1^2 + x_2^2 + \cdots.$$

Though it is again not obvious from the definitions, both the stable and dual stable Grothendieck polynomials are symmetric functions. We use this fact throughout this paper.

We say that two skew shapes $D_1$ and $D_2$ are *G-equivalent* or *g-equivalent* if $G_{D_1} = G_{D_2}$ or $g_{D_1} = g_{D_2}$, respectively. Since any $G_D$ contains $s_D$ as its lowest degree terms, $G_{D_1} = G_{D_2}$ implies $s_{D_1} = s_{D_2}$. Similarly, $g_{D_1} = g_{D_2}$ implies $s_{D_1} = s_{D_2}$. Furthermore, it is straightforward to check that two skew shapes that are equivalent in any of the three aforementioned senses must have the same number of rows and columns. We will implicitly use this fact throughout.

It is an easy consequence of symmetry that all three notions of skew equivalence are preserved under antipodal rotation, $^*$. We provide a proof for stable Grothendieck polynomials below.

**Proposition 2.1.** *For any skew shape $\lambda/\mu$, $G_{\lambda/\mu} = G_{(\lambda/\mu)^*}$ and $g_{\lambda/\mu} = g_{(\lambda/\mu)^*}$.*

*Proof.* We prove the result for stable Grothendieck polynomials; the argument for dual stable Grothendieck polynomials is similar. Let $x_I = x_{i_1}^{p_1} x_{i_2}^{p_2} \ldots x_{i_k}^{p_k}$ be a monomial with $i_1 < i_2 < \cdots < i_k$. It suffices to show that the $x_I$-coefficient of each of the two polynomials is equal. To do so, we construct a bijection between set-valued tableaux of shape $\lambda/\mu$ with weight monomial $x_I$ and set-valued tableaux of shape $(\lambda/\mu)^*$ with weight monomial $x_{I'} = x_{i_k+1-i_1}^{p_1} x_{i_k+1-i_2}^{p_2} \ldots x_1^{p_k}$. This bijection, which is in fact an involution, maps a tableau $T$ to the tableau $T'$ given by rotating $T$ and then replacing every entry $j$ with $i_k + 1 - j$. An example is given below, where $i_k = 5$.
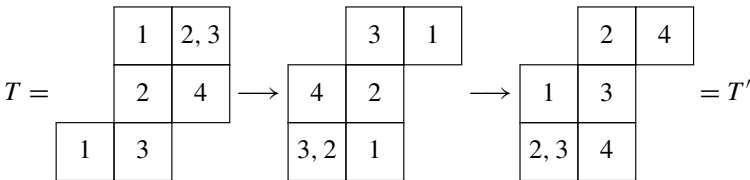


Thus, the $x_{I'}$-coefficient of $G_{(\lambda/\mu)^*}$ is equal to the $x_I$-coefficient of $G_{\lambda/\mu}$. By symmetry, the $x_{I'}$-coefficient of $G_{(\lambda/\mu)^*}$ is equal to the $x_I$-coefficient of $G_{(\lambda/\mu)^*}$, so the $x_I$-coefficients of $G_{\lambda/\mu}$ and $G_{(\lambda/\mu)^*}$ are equal, as desired. □

***Ribbon shapes.*** We will be interested in a special class of skew shapes known as *ribbons*. A skew shape $\alpha$ is called a ribbon if it is connected and contains no $2 \times 2$ rectangles. Being connected means that if there is more than one box, then each box must share an edge with another box. The shape shown below on the left is a ribbon while the shape in the middle and on the right are not. The shape in the middle contains a $2 \times 2$ rectangle and the shape on the right is not connected.
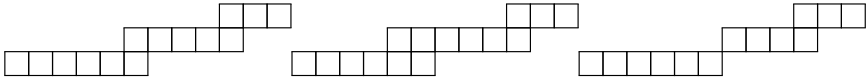


A *composition* of a positive integer $n$ is an ordered list of positive integers that sum to $n$. We will write compositions inside of parentheses. For example, $(2, 7, 4, 9)$ is a composition of 22. It is easy to see that ribbons of size $n$ are in bijection with compositions of $n$: to obtain a composition from a ribbon shape, simply read the row sizes from bottom to top. This is clearly a bijection. For this reason, we will denote a ribbon shape by the associated composition $\alpha$. For example, the we denote the ribbon shown above by $(6, 5, 3)$.

Note that one can also construct a bijection between compositions and ribbons using the sizes of the columns of $\alpha$ read from left to right. We also describe ribbon shapes this way, and we use square brackets in place of parentheses to denote this column reading. For example, the ribbon shown above may be written as $[1, 1, 1, 1, 1, 2, 1, 1, 1, 2, 1, 1]$.

Notice that the antipodal rotation $\alpha^*$ of $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k)$ is the ribbon $(\alpha_k, \alpha_{k-1}, \ldots, \alpha_1)$. We refer to $\alpha^*$ as the *reverse ribbon* of $\alpha$. For a ribbon shape $\alpha$, we denote the corresponding Schur function by $s_\alpha$ and refer to it as the *ribbon Schur function*.

We now define several binary operations on the set of ribbons as in [Reiner et al. 2007]. Here we let $\alpha = (\alpha_1, \ldots, \alpha_k)$ and $\beta = (\beta_1, \ldots, \beta_m)$ be ribbons. We define the concatenation operation

$$\alpha \cdot \beta = (\alpha_1, \cdots, \alpha_k, \beta_1, \cdots, \beta_m)$$

and the near concatenation operation

$$\alpha \odot \beta = (\alpha_1, \ldots, \alpha_{k-1}, \alpha_k + \beta_1, \beta_2, \ldots, \beta_m).$$

We let

$$\alpha^{\odot n} = \underbrace{\alpha \odot \cdots \odot \alpha}_{n}.$$

We can combine the two concatenation operations to yield a third operation $\circ$, defined by

$$\alpha \circ \beta = \beta^{\odot \alpha_1} \cdot \beta^{\odot \alpha_2} \cdots \beta^{\odot \alpha_k}.$$

**Example 2.2.** Consider ribbons $\alpha = (3, 2)$ and $\beta = (1, 2)$ shown below.

$$\alpha = \qquad \beta =$$

Then $\alpha \cdot \beta$ and $\alpha \odot \beta$ are as follows:

$$\alpha \cdot \beta = \qquad \alpha \odot \beta =$$

The operation $\alpha \circ \beta$ replaces each square of $\alpha$ with a copy of $\beta$. The copies of $\beta$ are near-concatenated if the corresponding blocks of $\alpha$ are horizontally adjacent and concatenated if the corresponding blocks of $\alpha$ are vertically adjacent.

$$\alpha \circ \beta =$$

If a ribbon $\alpha$ can be written in the form $\alpha = \beta_1 \circ \cdots \circ \beta_\ell$, we call this a *factorization* of $\alpha$. A factorization $\alpha = \beta \circ \gamma$ is called *trivial* if any of the following conditions hold:

(1) One of $\beta$ or $\gamma$ consists of a single square.

(2) Both $\beta$ and $\gamma$ consist of a single row.

(3) Both $\beta$ and $\gamma$ consist of a single column.

A factorization $\alpha = \beta_1 \circ \cdots \circ \beta_\ell$ is called *irreducible* if none of the factorizations $\beta_i \circ \beta_{i+1}$ are trivial and each $\beta_i$ has no nontrivial factorization. In [Billera et al. 2006], the authors prove that every ribbon $\alpha$ has a unique irreducible factorization. They then prove the following theorem.

**Theorem 2.3** [Billera et al. 2006]. *Two ribbons $\alpha$ and $\beta$ satisfy $s_\alpha = s_\beta$ if and only if $\alpha$ and $\beta$ have irreducible factorizations*

$$\alpha = \alpha_1 \circ \cdots \circ \alpha_k \quad and \quad \beta = \beta_1 \circ \cdots \circ \beta_k,$$

*where each $\beta_i$ is equal to either $\alpha_i$ or $\alpha_i^*$.*

In the next section, we use the above theorem to prove a necessary and sufficient condition for two ribbons to be $g$-equivalent. We also provide a necessary condition for two skew shapes to be $g$-equivalent.

## 3. Coincidences of dual stable Grothendieck polynomials

***Ribbons.*** The main result of this section is that for two ribbons $\alpha$ and $\beta$, we have $g_\alpha = g_\beta$ if and only if $\alpha = \beta$ or $\alpha = \beta^*$. We will obtain restrictions on $\alpha$ and $\beta$ by

writing the dual stable Grothendieck polynomials in terms of ribbon Schur functions and comparing the coefficients in the resulting expansions.

The next proposition requires the following ordering on ribbons. For ribbons $\alpha = [\alpha_1, \ldots, \alpha_n]$ and $\gamma = [\gamma_1, \ldots, \gamma_n]$ with the same number of columns, we write $\gamma \leq \alpha$ if $\gamma_i \leq \alpha_i$ for each $i = 1, \ldots, n$.

**Proposition 3.1.** *Let $\alpha = [\alpha_1, \ldots, \alpha_n]$ be a ribbon. The dual stable Grothendieck polynomial $g_\alpha$ can be decomposed into a sum of ribbon Schur functions as*

$$g_\alpha = \sum_{\gamma \leq \alpha} \left( \prod_{i=1}^{n} \binom{\alpha_i - 1}{\alpha_i - \gamma_i} \right) s_\gamma.$$

*Proof.* We define a map from reverse plane partitions of ribbon shape $\alpha$ to the set of semistandard Young tableaux of shape $\gamma$ where $\gamma \leq \alpha$. Given a reverse plane partition $T$ of shape $\alpha$, map $T$ to a semistandard Young tableau of shape $\gamma = [\gamma_1, \ldots, \gamma_n]$, where $\gamma_i$ is the number of distinct entries in column $i$ in $T$. Fill column $i$ of $\gamma$ with the distinct entries of column $i$ in $T$ in increasing order. This gives a semistandard Young tableau because columns are clearly strictly increasing and rows will remain weakly increasing.

This map preserves the monomial corresponding to the reverse plane partition. The map is also surjective, since any semistandard Young tableau of shape $\gamma$ where $\gamma \leq \alpha$ is mapped to by any reverse plane partition with the same entries in each column but with some entries copied.

It remains to show each semistandard Young tableau is mapped to by exactly $\prod \binom{\alpha_i - 1}{\alpha_i - \gamma_i}$ reverse plane partitions. Fix some semistandard Young tableau of shape $\gamma \leq \alpha$. We construct all possible reverse plane partitions of $\alpha$ mapping to this semistandard Young tableau column by column. Given column $i$ of $\alpha$, consider the $\alpha_i - 1$ pairs of adjacent squares in the column. Since there are $\gamma_i$ distinct entries in the column and the entries are written in weakly increasing order, $\alpha_i - \gamma_i$ of these pairs must match. A size $(\alpha_i - \gamma_i)$ subset of the $\alpha_i - 1$ pairs of adjacent squares gives a unique filling, where the given subset is the set of adjacent squares that match. Thus the number of possible fillings for each column is $\binom{\alpha_i - 1}{\alpha_i - \gamma_i}$, giving the desired formula. $\square$

**Lemma 3.2.** *Let $\alpha = [\alpha_1, \ldots, \alpha_n]$ and $\beta = [\beta_1, \ldots, \beta_n]$ be ribbons such that $g_\alpha = g_\beta$. Then for all $i = 1, \ldots, n$, we have $\alpha_i + \alpha_{n-i+1} = \beta_i + \beta_{n-i+1}$.*

*Proof.* We use Proposition 3.1 to write $g_\alpha$ and $g_\beta$ as a sum of ribbon Schur functions. Note that all terms of degree $n + 1$ in both sums are of the form $s_\gamma$, where $\gamma$ is a ribbon $(i, n - i + 1)$. Let $A$ denote the set of all compositions of $n + 1$ with weakly decreasing parts (i.e., the set of partitions of $n + 1$). It is shown in [Billera et al. 2006, Proposition 2.2] that the set $\{s_\alpha\}_{\alpha \in A}$ forms a basis for $\Lambda_{n+1}$, the degree $n + 1$ elements of the ring of symmetric functions. Then since each ribbon $(i, n - i + 1)$

is Schur equivalent to $(n-i+1, i)$, it follows that the set of Schur functions of such ribbons is linearly independent. Comparing coefficients in the respective sums gives the desired equality. $\qquad\square$

**Lemma 3.3.** *Suppose $\alpha$ and $\beta$ are ribbons such that $g_\alpha = g_\beta$, $\alpha \neq \beta$, and there exist ribbons $\sigma$, $\tau$, and $\mu$ such that $\alpha = \sigma \circ \mu$ and $\beta = \tau \circ \mu$. Then $\mu = \mu^*$.*

*Proof.* Let $\mu = [\mu_1, \ldots, \mu_t]$, $\alpha = [\alpha_1, \ldots, \alpha_n]$, and $\beta = [\beta_1, \ldots, \beta_n]$. By hypothesis, we have that $\alpha = \mu \square_1 \cdots \square_b \mu$ and $\beta = \mu \diamond_1 \cdots \diamond_s \mu$, where each $\square_i$ and $\diamond_i$ is one of the operations $\cdot$ or $\odot$. Thus each $\alpha_i$ and $\beta_i$ is equal to one of $\mu_1, \ldots, \mu_t$ or $\mu_1 + \mu_t$. Since $\alpha \neq \beta$, let $r$ be the minimal index such that $\alpha_r \neq \beta_r$. We see that $\{\alpha_r, \beta_r\} = \{\mu_t, \mu_1 + \mu_t\}$ because the first index where $\alpha$ and $\beta$ disagree corresponds to the first index $i$ where $\square_i \neq \diamond_i$. By Lemma 3.2 it follows that if $\alpha_i = \beta_i$ then $\alpha_{n-i+1} = \beta_{n-i+1}$. Hence $n - r + 1$ is the maximal index where $\alpha$ and $\beta$ disagree. Note that by the same argument we similarly have $\{\alpha_{n-r+1}, \beta_{n-r+1}\} = \{\mu_1, \mu_1 + \mu_t\}$.

We have $\alpha_r + \alpha_{n-r+1} = \beta_r + \beta_{n-r+1}$ by Lemma 3.2. Substituting the possible values of $\alpha_r \neq \beta_r$ and $\alpha_{n-r+1} \neq \beta_{n-r+1}$, we find that this equation is either

$$\mu_1 + \mu_t = 2(\mu_1 + \mu_t)$$

or

$$2\mu_1 + \mu_t = \mu_1 + 2\mu_t.$$

The first equation is a contradiction. Thus the second equation holds, implying that $\mu_1 = \mu_t$. We will show by induction that $\mu_i = \mu_{t-i+1}$, completing the proof. We have just shown the base case. For the general case, we have by Lemma 3.2,

$$\alpha_{r+i} + \alpha_{n-r-i+1} = \beta_{r+i} + \beta_{n-r-i+1}.$$

We may assume without loss of generality that $\alpha_r = \mu_t$. Then we have $\alpha_{n-r+1} = \mu_1 + \mu_t$, $\beta_r = \mu_1 + \mu_t$ and $\beta_{n-r+1} = \mu_1$. Therefore

$$\alpha_{r+i} = \mu_i, \qquad\qquad \beta_{r+i} = \mu_{i+1},$$
$$\alpha_{n-r-i+1} = \mu_{t-i}, \quad \beta_{n-r-i+1} = \mu_{t-i+1}.$$

We thus have

$$\mu_i + \mu_{t-i} = \mu_{i+1} + \mu_{t-i+1}.$$

By the inductive hypothesis $\mu_i = \mu_{t-i+1}$, so $\mu_{i+1} = \mu_{t-i}$, finishing the proof. $\quad\square$

We are now ready for the main result of this section.

**Theorem 3.4.** *For ribbons $\alpha$ and $\beta$, we have $g_\alpha = g_\beta$ if and only if $\alpha = \beta$ or $\alpha = \beta^*$.*

*Proof.* Suppose $g_\alpha = g_\beta$. Then $s_\alpha = s_\beta$. By Theorem 2.3, we have irreducible factorizations

$$\alpha = \alpha_k \circ \cdots \circ \alpha_1, \quad \beta = \beta_k \circ \cdots \circ \beta_1.$$

Here we reverse the indices for ease of induction. We prove by induction on $r$ that for $r = 1, \ldots, k$ we have

$$\alpha_r \circ \cdots \circ \alpha_1 \in \{\beta_r \circ \cdots \circ \beta_1, (\beta_r \circ \cdots \circ \beta_1)^*\}.$$

By Theorem 2.3 we have $\alpha_1 \in \{\beta_1, \beta_1^*\}$, so the base case is satisfied. Now suppose $r \geq 2$. By the inductive hypothesis we have

$$\alpha_{r-1} \circ \cdots \circ \alpha_1 \in \{\beta_{r-1} \circ \cdots \circ \beta_1, (\beta_{r-1} \circ \cdots \circ \beta_1)^*\}.$$

If $\alpha = \beta$ we are done, so we may assume otherwise. Then by letting $\mu = \alpha_{r-1} \circ \cdots \circ \alpha_1$ and applying Lemma 3.3 to $\alpha$ and either $\beta$ or $\beta^*$, we have

$$\beta_{r-1} \circ \cdots \circ \beta_1 = (\beta_{r-1} \circ \cdots \circ \beta_1)^*.$$

Since we also have that $\alpha_r \in \{\beta_r, \beta_r^*\}$ we are done. $\qquad \square$

***Necessary condition: bottlenecks.*** We now move to the case of determining equality of dual stable Grothendieck polynomials of general skew shape. We introduce the "bottleneck numbers" of a skew diagram and use these to construct closed-form expressions for certain coefficients of its dual Grothendieck polynomial. We then obtain a necessary condition for $g$-equivalence that generalizes Lemma 3.2.

For the following definition, we define an *interior horizontal edge* to be a horizontal edge of a box in a Young diagram that lies neither at the top boundary nor the bottom boundary of the Young diagram.

**Definition 3.5.** A *bottleneck edge* in a skew shape $\lambda/\mu$ is an interior horizontal edge touching both the left and right boundaries of the shape. For example, the red edges in Figure 1 are bottleneck edges. We let $b_i^{\lambda/\mu}$ denote the number of bottleneck edges in column $i$.

If the shape $\lambda/\mu$ has $n$ columns and $m$ rows, then the number of bottleneck edges in column $i$ for $i = 1, 2, \ldots, n$ is equivalently

$$b_i^{\lambda/\mu} = \big|\{1 \leq j \leq m - 1 \mid \mu_j = i - 1, \, \lambda_{j+1} = i\}\big|.$$

When the skew shape in question is clear, we will often suppress the superscript.

Bottleneck edges are related to the *row overlap compositions* defined in [Reiner et al. 2007], which we now review.

**Definition 3.6** [Reiner et al. 2007]. The *$k$-row overlap composition* $r^{(k)}$ of a skew diagram $\lambda/\mu$ with $m$ rows is $(r_1^{(k)}, \ldots, r_{m-k+1}^{(k)})$, where $r_i^{(k)}$ is the number of columns containing squares in all the rows $i, i+1, \ldots, i+k-1$.

In particular, $r^{(2)} = (\lambda_2 - \mu_1, \lambda_3 - \mu_2, \ldots, \lambda_m - \mu_{m-1})$. Thus bottleneck edges correspond to 1s in the 2-row overlap composition. When the $2, 3, \ldots, m$ row overlap compositions are written, they form a triangular array of nonnegative
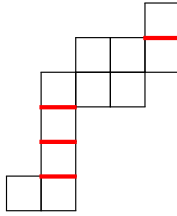
**Figure 1.** The skew shape $\langle 5, 5, 4, 2, 2, 2\rangle/\langle 4, 2, 1, 1, 1\rangle$ has three bottleneck edges in column 2 and one bottleneck edge in column 5.

integers as shown in Example 3.7. A column having $i$ bottleneck edges corresponds in the array to an equilateral triangle of 1s with side length $i$. In [Reiner et al. 2007], it is proven that the $k$-overlap compositions of two Schur equivalent shapes are permutations of each other for each $k$.

**Example 3.7.** Let $\lambda/\mu = \langle 5, 5, 4, 2, 2, 2\rangle/\langle 4, 2, 1, 1, 1\rangle$. Then the number of bottleneck edges in each column is shown below. Here, $(b_1, b_2, b_3, b_4, b_5) = (0, 3, 0, 0, 1)$. The row overlap compositions $r^{(2)}, \ldots, r^{(6)}$ are

| | | | | | |
|---|---|---|---|---|---|
| $r^{(6)}$ | | | 0 | | |
| $r^{(5)}$ | | | 0 | 0 | |
| $r^{(4)}$ | | 0 | 0 | 1 | |
| $r^{(3)}$ | 0 | 0 | 1 | 1 | |
| $r^{(2)}$ | 1 | 2 | 1 | 1 | 1 |

**Definition 3.8.** We define a *1,2-RPP* to be a reverse plane partition involving only 1s and 2s. A *mixed* column of a 1,2-RPP contains both 1s and 2s while an *$i$-pure* column contains only $i$'s.

Note the 1,2-RPPs of a given shape are in bijection with lattice paths from the upper right vertex of the shape to the lower left vertex of the shape. The corresponding 1,2-RPP can be generated from such a lattice path by filling the squares below the path with 2s and the squares above the path with 1s. Conversely, the corresponding lattice path can be recovered from a 1,2-RPP by drawing horizontal segments below the last 1 (if there are any) in a column and above the first 2 (if there are any) in a column. Vertical segments can then be drawn to connect these horizontal segments into a lattice path. Observe that mixed columns in the 1,2-RPP correspond to interior horizontal edges in the lattice path.

**Theorem 3.9.** *Let $\lambda/\mu$ be a skew shape with $n$ columns, and suppose $g_{\lambda/\mu} = g_{\gamma/\nu}$. Then*

$$b_i^{\lambda/\mu} + b_{n-i+1}^{\lambda/\mu} = b_i^{\gamma/\nu} + b_{n-i+1}^{\gamma/\nu}$$

*for $i = 1, 2, \ldots, n$.*

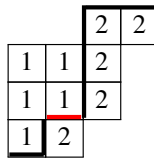|   |   | 2 | 2 |
|---|---|---|---|
| 1 | 1 | 2 |   |
| 1 | 1 | 2 |   |
| 1 | 2 |   |   |

**Figure 2.** Inside the skew shape, 1,2-RPPs correspond to lattice paths. Note the red interior horizontal edge corresponds to the boundary between the 1s and 2s in the mixed column.

Note that $\gamma/\nu$ must also have $n$ columns.

*Proof.* Fix a shape $\lambda/\mu$ with $m$ rows and $n$ columns. We will compute the coefficients for terms of the form $x_1^r x_2^{n-r+1}$ in $g_{\lambda/\mu}$. Since $g_{\lambda/\mu}$ is symmetric, we may assume without loss of generality that $r \leq n - r + 1$.

By the bijection between 1,2-RPPs and lattice paths given above, we may compute the coefficient of $x_1^r x_2^{n-r+1}$ by counting the number of lattice paths corresponding to this monomial. Note that any such lattice path must have exactly one interior horizontal edge. For each interior horizontal edge, we will count the number of lattice paths corresponding to the monomial $x_1^r x_2^{n-r+1}$ using the given edge. There are four cases: the interior horizontal edge either touches neither boundary, only the left boundary, only the right boundary, or both the left and right boundary (i.e., the edge is a bottleneck edge).

Fix an interior horizontal edge and suppose it lies in column $i$. Consider first the case where the interior horizontal edge touches neither boundary. Suppose a lattice path uses the given edge as its only interior horizontal edge. Then, as depicted in Figure 3, the lattice path must travel the top boundary until column $i$ and then drop to the horizontal edge. Then from the left endpoint of the given edge the path must drop to the bottom boundary and travel along the bottom boundary until reaching the bottom left. Hence there is a unique lattice path that uses the given edge as its only interior horizontal edge. Note that the corresponding 1,2-RPP has $i$ columns with 1s and $n - i + 1$ columns with 2s. Thus the lattice path gives the monomial $x_1^r x_2^{n-r+1}$ exactly when the edge lies in column $r$.

Next suppose the edge touches only the right boundary. Then as depicted in Figure 4, there may be multiple lattice paths using the edge: from the top right, the path may travel along the top boundary and drop down at any column before reaching column $i$. Note that the lattice path can correspond to a 1,2-RPP with between $i$ and $n$ columns containing 1s, and that the number of columns containing 1s determines the path. Similarly, if the edge touches only the left boundary, then after reaching the edge the path can drop down to the bottom boundary at any column before $i$. Hence the lattice path can correspond to a 1,2-RPP with between 1 and $i$ columns containing 1s.
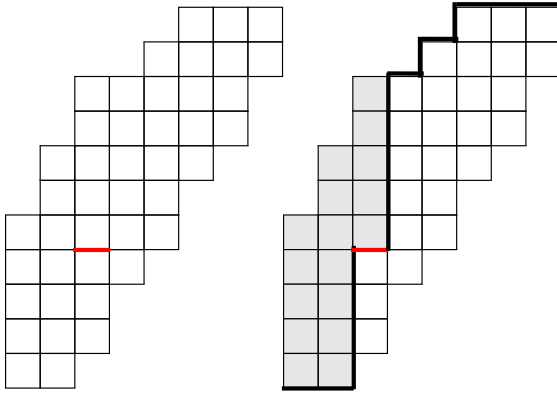
**Figure 3.** Given an interior horizontal edge touching neither bound-
ary, there is a unique lattice path with a single interior edge using
the edge. If the edge lies in column $i$, the path contains $i$ columns
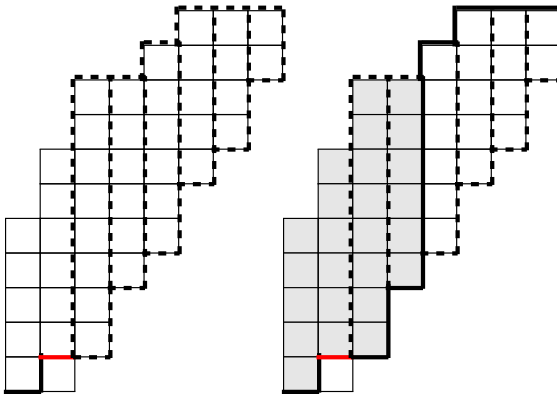with 1s and hence corresponds to the monomial $x_1^i x_2^{n-i+1}$.



**Figure 4.** When the edge touches only the right boundary, a lattice
path using this edge can now drop down from the top boundary
at any column after column $i$. However, there is a unique path
corresponding to the monomial $x_1^r x_2^{n-r+1}$ if $i \leq r$ and no possible
paths if $i > r$.

Thus we have identified three cases where there is at least one lattice path
corresponding to $x_1^r x_2^{n-r+1}$: the interior horizontal edge lies in one of column
$1, \ldots, r$ and touches the right boundary, the edge lies in column $r$ and touches
neither boundary, or the edges lies in one of column $r, \ldots, n$ and touches the left
boundary. We will consider the fourth case, where the interior edge is a bottleneck
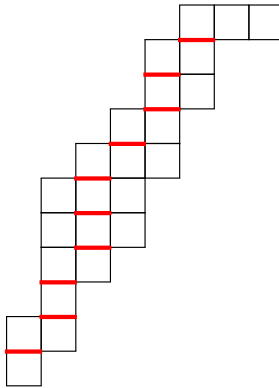
**Figure 5.** There are $m - 1$ possible edges that can be chosen as the interior horizontal edge for a lattice path. Unless the edge is a bottleneck edge, each such edge corresponds to a unique lattice path.

edge, in the next paragraph. Fix two adjacent rows, and consider the set of horizontal edges between these two rows. The columns that these edges lie in are either all to the left of column $r$, contain column $r$, or all to the right of column $r$. In any of these three cases there is exactly one valid edge, as depicted in Figure 5. That is, between any two adjacent rows there is exactly one edge that corresponds to at least one lattice path. Since there are $m$ rows, this gives $m - 1$ possible valid interior horizontal edges. Unless the edges are bottleneck edges, each possible edge corresponds to a single lattice path. It remains to count the additional lattice paths given by bottleneck edges.

Now suppose the interior horizontal edge is a bottleneck edge lying in column $i$. Then there is flexibility on both sides: there can be between 0 and $(i - 1)$ 1-pure columns to the left of column $i$ and between 0 and $(n - i)$ 2-pure columns to the right of column $i$. If there are $x$ 1-pure columns to the left of column $i$, $x$ may be between 0 and $\max(i - 1, r - 1)$. If $x$ 1-pure columns lie to the left, the remaining $(r - x - 1)$ 1-pure columns can be chosen to be to the right of column $i$ (because we assumed that $r \leq n - r + 1$). Hence there are $\max(i, r)$ possible lattice paths using a given bottleneck edge in column $i$.

We can now give a formula for the coefficient of $x_1^r x_2^{n-r+1}$. Let $k = \lceil \frac{1}{2}n \rceil$ and

$$f_i = b_i + b_{n-i+1} \quad \text{for } i = 1, 2, \ldots, k - 1.$$

If $n$ is even, let $f_k = b_k + b_{n-k+1}$ and if $n$ is odd, let $f_k = b_k$. There are always at least $m - 1$ valid lattice paths. Each bottleneck edge in column $i$ also contributes an additional $\max(i, r) - 1$ lattice paths. Hence the coefficient is

$$(m - 1) + f_2 + 2f_3 + 3f_4 + \cdots + (r - 1)f_r + (r - 1)f_{r+1} + \cdots + (r - 1)f_k.$$

Let $t_r$ denote the coefficient of $x_1^r x_2^{n-r+1}$. Note that for $2 \leq r \leq k - 1$, we have $2t_r - t_{r-1} - t_{r+1} = f_r$. Since any two $g$-equivalent shapes $\lambda/\mu$ and $\gamma/\nu$ must have the same coefficients $t_r$, it follows that for $2 \leq r \leq k - 1$ the sums $f_r = b_r + b_{n-r+1}$ are the same for the two shapes. Also, since

$$t_k = (m - 1) + f_2 + 2f_3 + 3f_4 + \cdots + (k - 1)f_k$$

is invariant for the two shapes, it then follows that $f_k$ is invariant for the two shapes.

By [Reiner et al. 2007, Corollary 8.11], we also have that $b_1 + \cdots + b_n = f_1 + \cdots + f_k$ is invariant, since the total number of bottleneck edges is the number of 1s in the 2-row overlap composition. Hence $f_1$ is invariant as well. $\square$

**Remark 3.10.** For a ribbon $[\alpha_1, \alpha_2, \ldots, \alpha_n]$, we have $b_i = \alpha_i - 1$ for $i = 1, 2, \ldots, n$. Hence Theorem 3.9 generalizes Lemma 3.2 as noted at the beginning of the section.

**Example 3.11.** It is noted in [Reiner et al. 2007] that the shapes
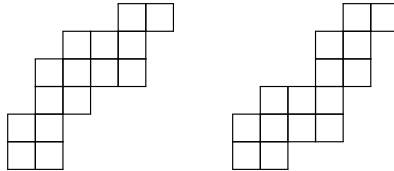


are Schur equivalent. But since $b_2 + b_5 = 2$ for the first shape and $b_2 + b_5 = 1$ for the second shape, it follows that the two shapes are not $g$-equivalent.

**Example 3.12.** Having the same bottleneck edge sequence is not sufficient for two skew shapes to be $g$-equivalent. By [Reiner et al. 2007, Theorem 7.6], the shapes $D_1$ and $D_2$ below are equivalent and have the same bottleneck edge sequence. However, upon computation it is found that they are not $g$-equivalent.
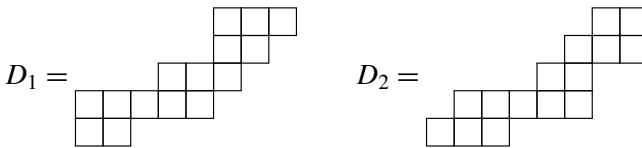


Since the bottleneck condition followed as a result of comparing terms of $g$ with degree $n + 1$ in two variables, it is natural to compute coefficients for terms of higher degree or more variables. The following result shows that terms of degree $n + 1$ and more than two variables do not impose additional constraints for two skew shapes to be $g$-equivalent.

**Proposition 3.13.** *Suppose two skew shapes $\lambda/\mu$ and $\gamma/\nu$ have the same number of rows and the polynomial $g_{\lambda/\mu}$ and $g_{\gamma/\nu}$ have same coefficient for every term of degree $n + 1$ with two variables. Then in fact these polynomials have the same coefficient for any term of degree $n + 1$.*
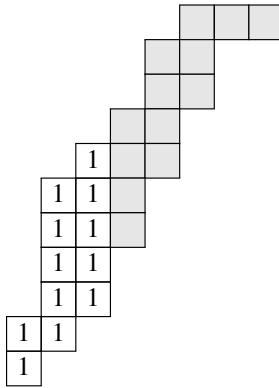
**Figure 6.** The remaining shape shaded in gray is a skew shape with $n - i_1$ columns, denoted $(\lambda/\mu)_{i_1}$.

*Proof.* Fix positive integers $i_1, i_2, \ldots, i_k$, where $k \geq 2$ is some positive integer, and let $n = \left(\sum_{j=1}^{k} i_j\right) - 1$. Given a skew diagram $\lambda/\mu$ with $n$ columns, we claim that the coefficient of $x_1^{i_1} x_2^{i_2} \ldots x_k^{i_k}$ can be expressed as a $\mathbb{Z}$-linear combination $c_0 + c_2 b_2 + \ldots c_{n-1} b_{n-1}$ of the bottleneck numbers $b_i$. Furthermore, the constant $c_0$ is known to be $(k-1)(m-1)$, where $m$ is the number of rows in $\lambda/\mu$. We proceed by induction on $k$.

The base case $k = 2$ is given in the proof of Theorem 3.9, so we may assume $k \geq 3$. We count the number of reverse plane partitions giving the monomial $x_1^{i_1} \ldots x_k^{i_k}$.

Suppose first that every column containing a 1 is in fact 1-pure. Then the first $i_1$ columns must be filled with 1s. Note the remaining squares form a skew shape with $n - i_1$ columns, as depicted in Figure 6. We henceforth use $(\lambda/\mu)_{i_1}$ to denote the skew shape given by removing the first $i_1$ columns of $\lambda/\mu$. Note $(\lambda/\mu)_{i_1}$ must be filled with a reverse plane partition giving the monomial $x_2^{i_2} \cdots x_k^{i_k}$.

Let $m'$ be the number of rows in the shape obtained by removing the first $i_1$ columns from $\lambda/\mu$. Then by induction the number of ways to fill in this shape is

$$(k-2)(m'-1) + c'_{i_1+2} b_{i_1+2} + \cdots + c'_{n-1} b_{n-1}$$

for some integers $c'_{i_1+2}, \ldots, c'_{n-1}$.

The remaining case is when the reverse plane partition has a mixed column containing a 1. Given such a reverse plane partition, consider the 1,2-RPP obtained by replacing every entry greater than or equal to 2 with 2. From this 1,2-RPP, we obtain a lattice path via our previously described bijection between 1,2-RPPs and lattice paths. Since the reverse plane partition has a mixed column containing a 1 and the total degree of the monomial $x_1^{i_1} x_2^{i_2} \ldots x_k^{i_k}$ is $n + 1$, this lattice path must have a single interior horizontal edge. As noted in Figure 5, there are $m - 1$ possibilities for the unique interior horizontal edge.
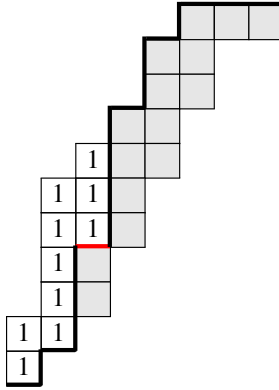
**Figure 7.** Case 1: the lattice path uses an edge in columns $1, \ldots, i_1$ touching the bottom boundary. Then the remaining shape is the union of $(\lambda/\mu)_{i_1}$ and a single column.

Consider first the interior horizontal edges in columns $1, \ldots, i_1$ touching the bottom boundary of $\lambda/\mu$. This case is depicted in Figure 7. Note that there are $m - m'$ such edges, since in total there are $m - 1$ edges touching the bottom boundary and exactly $m' - 1$ of them lie in columns $i_1 + 1, \ldots, n$. For each of these $m - m'$ edges, there is only one possible lattice path. The path starts at the top right, travels along the top boundary until it reaches the boundary between column $i_1$ and column $i_1 + 1$, drops down to the bottom boundary, and travels along the bottom boundary until the horizontal edge, traverses the edge, and then immediately drops back down to the bottom boundary and traverses it until reaching the bottom left. This lattice path determines which squares are filled with 1s. The remaining shape is a disconnected skew shape where one component is a single column and the other component is $(\lambda/\mu)_{i_1}$. There are $(k - 1)$ fillings using this lattice path, since the column below the edge may be filled with any of $2, \ldots, k$ and the remaining columns must fill $(\lambda/\mu)_{i_1}$ in increasing order. Note that unless the edge is a bottleneck edge, this is the unique lattice path using this edge.

The remaining $m' - 1$ edges are those in column $i_1$ not touching the bottom boundary and the edges in column $i_i + 1, \ldots, n$ touching the top boundary. We similarly describe a possible lattice path for each of these edges. Suppose the edge lies in column $i$. The path starts at the top right, travels along the top boundary until the boundary between column $i$ and $i + 1$, drops down to the edge and traverses it, traverses the top boundary until the boundary between column $i_1 - 1$ and column $i_1$, and drops down to the bottom boundary and traverses it until reaching the bottom left. This path determines which squares are filled with 1s. The remaining squares form a (possibly disconnected) skew shape, which must be filled with no mixed columns. Note the remaining skew shape is connected if and only if the horizontal
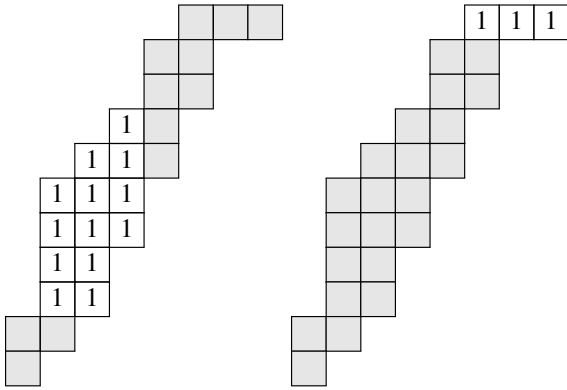
**Figure 8.** The remaining shape will have 1 or 2 components. The number of fillings is determined by $i_2, \ldots, i_k$ and the number of columns in the components.

edge was not a bottleneck edge. If the shape is connected, then filling the columns in increasing order is the only possible filling. Otherwise, this is one possible filling but there may be more.

Thus far, this gives us

$$(k-2)(m'-1) + c'_{i_1+2}b_{i_1+2} + \cdots + c'_{n-1}b_{n-1} + (k-1)(m-m') + (m'-1)$$
$$= (k-1)(m-1) + c'_{i_1+2}b_{i_1+2} + \cdots + c'_{n-1}b_{n-1}$$

fillings. It remains to show each bottleneck edge in column $i$ contributes a fixed number of additional fillings depending only on $i$.

As noted in the proof of Theorem 3.9, each bottleneck edge in column $i$ has $\min(i, n-i+1, i_1)$ possible lattice paths using that edge. Each lattice path determines which squares will be filled with 1s. Note the remaining squares will form a possibly disconnected skew shape with $n - i_1 + 1$ columns (depicted in Figure 8), which must then be filled with no mixed columns. There are a fixed number of ways to fill this shape, which depends only on $i_2, \ldots, i_k$ and the number of columns in the two components. The possible number of columns in each component is in turn determined by which column the bottleneck edge is in; see Figure 9. This finishes the proof of the claim.

Thus we have that the coefficient of $x_1^{i_1} \ldots x_k^{i_k}$ for any shape with $n = \left(\sum_{j_1}^k i_j\right) - 1$ columns is $(k-1)(m-1) + c_2 b_1 + \cdots + c_{n-1} b_{n-1}$ for some integers $c_2, \ldots, c_{n-1}$. Recall that every shape is equivalent to its 180-degree rotation, and note that a 180-degree rotation reverses the bottleneck sequence $b_1, \ldots, b_n$. Since there are shapes with arbitrary sequences of $b_1, \ldots, b_n$ (for example, the ribbon $[b_1+1, \ldots, b_n+1]$), it follows that $c_i = c_{n-i+1}$ for $i = 2, \ldots, n-1$. Recall also that the proof of Theorem 3.9 shows each sum $b_i + b_{n-i+1}$ for $i = 2, \ldots, n-1$ must be the same
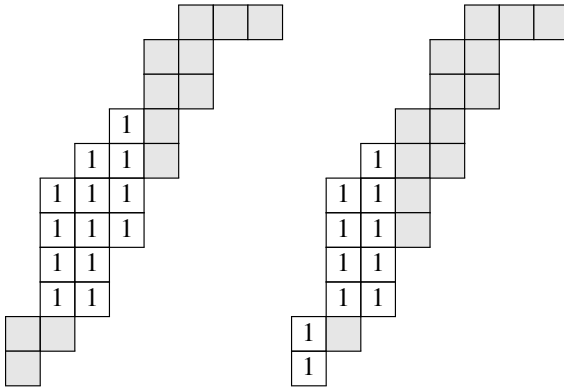
**Figure 9.** The possible numbers of columns in the components of the remaining shape are determined by which column the bottleneck edge is in. In this example, since the bottleneck edge is in column 2, the components have 1 and $n - i_1 + 1$ columns or 2 and $n - i_1 + 2$ columns.

for any two shapes such that the terms in $g$ of degree $n + 1$ with two variables are the same. Since the number of rows $m$ must be the same as well, it follows that the sum $(k - 1)(m - 1) + c_2 b_1 + \cdots + c_{n-1} b_{n-1}$ must also be the same. □

**Proposition 3.14.** *The coefficient of $x_1^2 x_2^n$ in $g_{\lambda/\mu}$ is*

$$\binom{m}{2} - \sum_{i=1}^{n} \binom{b_i + 1}{2}.$$

*Proof.* A 1,2-RPP giving the monomial $x_1^2 x_2^n$ must have no 1-pure columns, $(n - 2)$ 2-pure columns, and two mixed columns. Hence the corresponding lattice paths have two interior horizontal edges. Consider the *heights* of the interior horizontal edges. By an interior horizontal edge at height $i$ we mean the edge lies between row $i$ and row $i + 1$. Observe that given the height of the two interior horizontal edges, there is at most one lattice path using the heights; since there are no 1-pure columns, the lattice path is completely determined by the heights chosen.

There are $\binom{m}{2}$ ways to choose a pair of heights from $1, \ldots, m - 1$ (with possible repetition). Since each pair of heights contributes either 1 or 0 lattice paths, the desired coefficient is thus $\binom{m}{2}$ minus the number of pairs not giving a lattice path. These are exactly the pairs of heights where the only interior horizontal edges at those heights lie in a single column. These are precisely the pairs of bottleneck edges from the same column. For each column $i$ there are $\binom{b_i + 1}{2}$ ways to choose two of the bottlenecks in column $i$ (with possible repetition), giving the desired formula. □
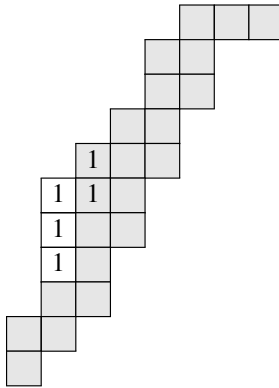
**Figure 10.** The heights determine the filling, since the column containing 1s more to the left must touch the left boundary of the shape, and the other column must touch the left boundary of the remaining shape.

By [Reiner et al. 2007, Corollary 8.11], the number of rows $m$ and the sum $b_1 + \cdots + b_n$ are invariant under $g$-equivalence. Hence we attain the following as a direct consequence of Proposition 3.14.

**Corollary 3.15.** *Suppose* $g_{\lambda/\mu} = g_{\gamma/\nu}$. *Then*

$$\sum_{i=1}^{n} (b_i^{\lambda/\mu})^2 = \sum_{i=1}^{n} (b_i^{\gamma/\nu})^2.$$

*Equivalently, the sums of the areas of the equilateral triangles of* 1*s in the row overlap compositions* $r^{(2)}, \ldots, r^{(m)}$ *are the same.*

**Remark 3.16.** One can also count various other coefficients in the dual stable Grothendieck polynomial. For terms of degree greater than $n + 1$, it is useful to define a generalization of bottleneck edges. To that end, for $i = 1, \ldots, \lambda_1 - w + 1$, the *number of width* $w$ *bottlenecks in position* $i$ is

$$b_i^{(w)} = \big|\{1 \le j \le m - 1 \mid \mu_j = i - 1, \ \lambda_{j+1} = i + w - 1\}\big|.$$

For example, let $\lambda/\mu = (5, 5, 4, 2, 2, 2)/(4, 2, 1, 1, 1, 0)$. Then the number of bottleneck edges of each width is given below. Note $b^{(1)}$ is just the previously defined bottleneck edges.

| | | | | | |
|---|---|---|---|---|---|
| $b^{(5)}$ | | | | 0 | |
| $b^{(4)}$ | | | 0 | | 0 |
| $b^{(3)}$ | | 0 | | 0 | 0 |
| $b^{(2)}$ | 0 | | 0 | 1 | 0 |
| $b^{(1)}$ | 0 | 3 | 0 | 0 | 1 |

We state the following propositions with proofs omitted for brevity.

**Proposition 3.17.** *The coefficient of $x_1^3 x_2^{n-1}$ in $g_{\lambda/\mu}$ is*

$$\left(\binom{m}{2} - \sum_{i=1}^{n}\binom{b_i^{(1)}+1}{2}\right) + \sum_{i=2}^{n-2}\binom{b_i^{(2)}+1}{2} + (m-2)\sum_{i=2}^{n-1}b_i^{(1)}$$
$$- \left(b_2^{(1)}(m-\mu_1'-1) + b_{n-1}^{(1)}(\lambda_n'-1) + \sum_{i=2}^{n-2}b_i^{(1)}b_{i+1}^{(1)}\right).$$

**Proposition 3.18.** *The coefficient of $x_1^3 x_2^n$ in $g_{\lambda/\mu}$ is*

$$\binom{m+1}{3} - \sum_{i=1}^{n}\left((m-1)\binom{b_i^{(1)}+1}{2} - 2\binom{b_i^{(1)}}{3} - b_i^{(1)}(b_i^{(1)}-1)\right)$$
$$- \sum_{i=1}^{n-1}\left(\binom{b_i^{(2)}+2}{3} + (b_i^{(1)}+b_{i+1}^{(1)})\binom{b_i^{(2)}+1}{2} + b_i^{(1)}b_i^{(2)}b_{i+1}^{(1)}\right).$$

## 4. Transposition and stable Grothendieck polynomials

Given a Young diagram $\lambda = \langle\lambda_1, \lambda_2, \ldots, \lambda_k\rangle$, we define its *transpose* Young diagram to be $\lambda^t = \langle\lambda_1', \ldots, \lambda_s'\rangle$, where $\lambda_i'$ is the number of boxes in column $i$ of $\lambda$. This operation extends to skew diagrams by setting $(\lambda/\mu)^t = \lambda^t/\mu^t$. For example, $\langle 5, 5, 2\rangle^t = \langle 3, 3, 2, 2, 2\rangle$ and $(\langle 4, 3, 1\rangle/\langle 2\rangle)^t = \langle 4, 3, 1\rangle^t/\langle 2\rangle^t = \langle 3, 2, 2, 1\rangle/\langle 1, 1\rangle$.

For skew shapes $A$ and $B$ it follows immediately from the Jacobi–Trudi identity that $s_A = s_B$ implies $s_{A^t} = s_{B^t}$. There is not yet a Jacobi–Trudi identity for stable and dual stable Grothendieck polynomials, so we must use other methods.

With the goal of proving this result for stable Grothendieck polynomials, we introduce the following definitions. We first define the symmetric function $\widetilde{K}_{\lambda/\mu}$, which was first introduced by Lam and Pylyavskyy [2007], by

$$\widetilde{K}_{\lambda/\mu} = \sum_{T} x^T,$$

where we sum over all set-valued tableaux of shape $\lambda/\mu$. It is easy to see that $K_{\lambda/\mu}$ is related to $G_{\lambda/\mu}$ by $\tilde{K}_{\lambda/\mu} = (-1)^{|\lambda/\mu|}G_{\lambda/\mu}(-x_1, -x_2, \ldots)$ and that $G_A = G_B$ if and only if $\widetilde{K}_A = \widetilde{K}_B$ for any skew shapes $A$ and $B$.

We also introduce the symmetric function $J_{\lambda/\mu}$ [Lam and Pylyavskyy 2007] using the following definition.

**Definition 4.1.** A *weak set-valued tableau* T of shape $\lambda/\nu$ is a filling of the boxes of the skew shape $\lambda/\nu$ with finite, nonempty multisets of positive integers so that

(1) the largest number in each box is strictly smaller than the smallest number in the box directly to the right of it, and

(2) the largest number in each box is less than or equal to the smallest number in the box directly below it.

In other words, we fill the boxes with multisets so that rows are strictly increasing and columns are weakly increasing. For example, the filling of shape $(3, 2, 1)$ shown below gives a weak set-valued tableau, $T$, of weight $x^T = x_1 x_2^3 x_3^3 x_4^2 x_5 x_6 x_7$.

| 1, 2 | 3, 3 | 4, 6 |
|------|------|------|
| 2, 2, 3 | 4 | |
| 5, 7 | | |

Let $J_{\lambda/\nu} = \sum_T x^T$ be the weight generating function of weak set-valued tableaux $T$ of shape $\lambda/\nu$. From [Patrias and Pylyavskyy 2016, Theorem 5.11], we know that

$$J_{\lambda/\mu}(x_1, x_2, \ldots) = (-1)^{|\lambda/\mu|} G_{(\lambda/\mu)^t}\left(-\frac{x_1}{x - x_1}, -\frac{x_2}{1 - x_2}, \ldots\right).$$

It follows from this that $G_A = G_B$ if and only if $J_{A^t} = J_{B^t}$. In addition, [Lam and Pylyavskyy 2007, Proposition 9.22] says that

$$\omega(\widetilde{K}_{\lambda/\mu}) = J_{\lambda/\mu},$$

where $\omega$ is the fundamental involution of symmetric functions that sends $s_\lambda$ to $s_{\lambda^t}$.

**Theorem 4.2.** *If $G_A = G_B$ for skew shapes $A$ and $B$, then $G_{A^t} = G_{B^t}$.*

*Proof.* If $G_A = G_B$, then $\widetilde{K}_A = \widetilde{K}_B$, and thus $J_A = \omega(\widetilde{K}_A) = \omega(\widetilde{K}_B) = J_B$. It follows that $G_{A^t} = G_{B^t}$. □

It remains open to prove this for the dual stable Grothendieck polynomials. If conjugation does preserve $g$-equivalence, then we immediately have another necessary condition on $g$-equivalence by taking a transposed version of Theorem 3.9.

**Question 4.3.** Suppose $g_A = g_B$. Does it follow that $g_{A^T} = g_{B^T}$?

## 5. Relation between $g$-equivalence and $G$-equivalence

It is natural to ask whether $g_A = g_B$ for two skew shapes $A$ and $B$ implies $G_A = G_B$, and vice versa. The following examples show that in general, neither equality implies the other.

**Example 5.1.** Based on computer computation, the shapes
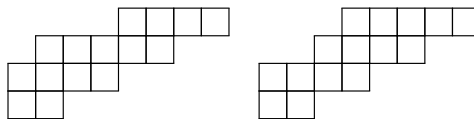
are $g$-equivalent but not $G$-equivalent. For example, the coefficients of $x_1^6 x_2^6 x_3^3 x_4$ in $G$ are $-353$ and $-354$, respectively.

**Example 5.2.** The shapes
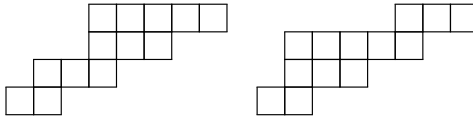


are $G$-equivalent but not $g$-equivalent. One can show $G$-equivalence through computer computation using the reverse lattice word expansion of $G_{\lambda/\mu}$ into stable Grothendieck polynomials indexed by straight shapes found in [Buch 2002]. To see the shapes are not $g$-equivalent, we notice that $b_4 + b_5 = 1$ for the shape on the left and $b_4 + b_5 = 0$ for the shape on the right.

## 6. Future explorations

*Coincidences of ribbon stable Grothendieck polynomials.* The combinatorics of ribbon stable Grothendieck polynomials seem to be more difficult than their dual stable Grothendieck and Schur counterparts. However, we still conjecture that coincidences among ribbon Grothendieck polynomials arise in precisely the same way as the dual case. While one direction of the following statement is immediate, the other direction has proven to be much more difficult.

**Conjecture 6.1.** *Let $\alpha$ and $\beta$ be ribbons. Then $G_\alpha = G_\beta$ if and only if $\beta = \alpha$ or $\beta = \alpha^*$.*

*Ribbon staircases.* A class of nontrivial skew equivalences is described in [Reiner et al. 2007, Theorem 7.30]. A *nesting* is a word consisting of the symbols left parenthesis "(", right parenthesis ")", dot ".", and vertical slash "|", where the parentheses must be properly matched. Given a skew shape that may be decomposed into a ribbon $\alpha$ in a certain manner as described in [Reiner et al. 2007], one may obtain a corresponding nesting. Theorem 7.30 in the previous work states that shapes that may be decomposed with the same ribbon $\alpha$ such that the nestings are reverses of each other are Schur equivalent.

It is interesting to consider whether these equivalences hold for $g$ and $G$ as well. For example, [Reiner et al. 2007, Corollary 7.32] states that $s_{\delta_n/\mu} = s_{(\delta_n/\mu)^T}$ for any diagram $\mu$ contained in the staircase partition $\delta_n = \langle n-1, n-2, \dots, 1 \rangle$. Computation strongly suggests the same holds true for the Grothendieck polynomials as well.

**Conjecture 6.2.** *Let $\mu$ be a diagram contained in the staircase partition $\delta_n = \langle n-1, n-2, \dots, 1 \rangle$. Then $g_{\delta_n/\mu} = g_{(\delta_n/\mu)^T}$ and $G_{\delta_n/\mu} = G_{(\delta_n/\mu)^T}$.*

However, not all equivalences described by [Reiner et al. 2007, Theorem 7.30] hold for Grothendieck polynomials.

**Question 6.3.** For which ribbons $\alpha$ and nestings $\mathcal{N}$ are the corresponding shapes $g$-equivalent or $G$-equivalent?

### Acknowledgments

### References

[Billera et al. 2006] L. J. Billera, H. Thomas, and S. van Willigenburg, "Decomposable compositions, symmetric quasisymmetric functions and equality of ribbon Schur functions", *Adv. Math.* **204**:1 (2006), 204–240. MR Zbl

[Buch 2002] A. S. Buch, "A Littlewood–Richardson rule for the $K$-theory of Grassmannians", *Acta Math.* **189**:1 (2002), 37–78. MR Zbl

[Lam and Pylyavskyy 2007] T. Lam and P. Pylyavskyy, "Combinatorial Hopf algebras and $K$-homology of Grassmannians", *Int. Math. Res. Not.* **2007**:24 (2007), art. id. rnm125, 48 pp. MR

[McNamara and van Willigenburg 2009] P. R. W. McNamara and S. van Willigenburg, "Towards a combinatorial classification of skew Schur functions", *Trans. Amer. Math. Soc.* **361**:8 (2009), 4437–4470. MR Zbl

[Patrias and Pylyavskyy 2016] R. Patrias and P. Pylyavskyy, "Combinatorics of $K$-theory via a $K$-theoretic Poirier–Reutenauer bialgebra", *Discrete Math.* **339**:3 (2016), 1095–1115. MR Zbl

[Reiner et al. 2007] V. Reiner, K. M. Shaw, and S. van Willigenburg, "Coincidences among skew Schur functions", *Adv. Math.* **216**:1 (2007), 118–152. MR Zbl

[Stanley 1999] R. P. Stanley, *Enumerative combinatorics, II*, Cambridge Studies in Advanced Mathematics **62**, Cambridge Univ. Press, 1999. MR Zbl

ealwais@emory.edu                  *Department of Mathematics and Computer Science,
                                   Emory University, Atlanta, GA 30322, United States*

sc2586@cornell.edu                 *Cornell University, Ithaca, NY 14850, United States*

aclifton@mit.edu                   *Massachusetts Institute of Technology,
                                   Cambridge, MA 02139, United States*

patri080@umn.edu                   *Université du Québec à Montréal, Montreal QC H2L 2C4,
                                   Canada*

prasad01@college.harvard.edu       *Harvard University, Cambridge, MA 02138, United States*

mfshinners@wisc.edu                *University of Wisconsin-Madison, Madison, WI 53706,
                                   United States*

iaeuy@berkeley.edu                 *University of California-Berkeley, Berkeley, CA 94720,
                                   United States*

# A probabilistic heuristic for counting components of functional graphs of polynomials over finite fields

Elisa Bellah, Derek Garton, Erin Tannenbaum and Noah Walton

(Communicated by Michael E. Zieve)

Flynn and Garton (2014) bounded the average number of components of the functional graphs of polynomials of fixed degree over a finite field. When the fixed degree was large (relative to the size of the finite field), their lower bound matched Kruskal's asymptotic for random functional graphs. However, when the fixed degree was small, they were unable to match Kruskal's bound, since they could not (Lagrange) interpolate cycles in functional graphs of length greater than the fixed degree. In our work, we introduce a heuristic for approximating the average number of such cycles of any length. This heuristic is, roughly, that for sets of edges in a functional graph, the quality of being a cycle and the quality of being interpolable are "uncorrelated enough". We prove that this heuristic implies that the average number of components of the functional graphs of polynomials of fixed degree over a finite field is within a bounded constant of Kruskal's bound. We also analyze some numerical data comparing implications of this heuristic to some component counts of functional graphs of polynomials over finite fields.

## 1. Introduction

A (*discrete*) *dynamical system* is a pair $(S, f)$ consisting of a set $S$ and a map $f : S \to S$. Given such a system, an element $s \in S$ is a *periodic point* of the system if there exists some $k \in \mathbb{Z}^{>0}$ such that $(f \circ \cdots \circ f)(s) = s$, where $f$ appears $k$ times; the smallest $k \in \mathbb{Z}^{>0}$ with this property is called the *period* of $s$. The *functional graph* of such a system, which we denote by $\Gamma(S, f)$, is the directed graph whose vertex set is $S$ and whose edges are given by the relation $s \to t$ if and only if $f(s) = t$. A *component* of such a graph is a component of the underlying undirected graph. For any $n \in \mathbb{Z}^{>0}$, let $\mathcal{K}(n)$ denote the average number of components of a random

functional graph on a set of size $n$; that is, choose any set $S$ with $|S| = n$ and let

$$\mathcal{K}(n) = n^{-n} \sum_{f:S \to S} |\{\text{components of } \Gamma(S, f)\}|.$$

Kruskal [1954] proved that

$$\mathcal{K}(n) = \tfrac{1}{2} \log n + \tfrac{1}{2}(\log 2 + C) + o(1),$$

where $C = .5772\ldots$ is Euler's constant.

Recently, researchers have begun studying the analogous situation for polynomials (and rational maps) over finite fields. More precisely, if $q$ is a prime power and $f \in \mathbb{F}_q[x]$, define $\Gamma(q, f) = \Gamma(\mathbb{F}_q, f)$ (if there is no ambiguity, we will frequently write $\Gamma_f$ for $\Gamma(q, f)$). Then we can ask the question: for $d \in \mathbb{Z}^{>0}$, what is the average number of components of $\Gamma_f$, for $f$ ranging over all polynomials over $\mathbb{F}_q$ of a fixed degree? In particular, if we define

$$\mathcal{P}(q, d) := \frac{1}{|\{f \in \mathbb{F}_q[x] \mid \deg f = d\}|} \cdot \sum_{\substack{f \in \mathbb{F}_q[x] \\ \deg f = d}} |\{\text{components of } \Gamma_f\}|,$$

then we can ask:

**Question 1.1.** For a prime power $q$ and $d \in \mathbb{Z}^{>0}$, how does $\mathcal{P}(q, d)$ compare to $\mathcal{K}(q)$?

In this paper, we recast these questions in probabilistic terms. Specifically, in Section 2, we define two families of random variables whose interaction determines the answer to Question 1.1. Briefly, both families of random variables have sample space a certain collection of subsets of $\mathbb{F}_q \times \mathbb{F}_q$ — one random variable determines if a collection is a cycle, and the other returns how many polynomials of a given degree pass though every point in a collection.

Our main result, Theorem 3.3, states that if these two families of random variables satisfy a certain "noncorrelation hypothesis", then

$$\mathcal{P}(q, d) = \mathcal{K}(q) + O(1);$$

see Heuristic 3.1 for an exact formulation of this hypothesis. In Section 2 we define and study these random variables; in particular, we compute their expected values. Next, in Section 3 we use the results from Section 2 to prove the aforementioned Theorem 3.3. Then, in Section 4, we provide numerical evidence in support of Heuristic 3.1. Finally, these results carry over easily to the analogous question for rational functions; these results make up Section 5.

Previous work of Flynn and the second author (see [Flynn and Garton 2014]) provided a partial answer to the question under discussion. In particular, they proved that if $d \geq \sqrt{q}$, then the average number of components of functional graphs of

polynomials (or rational maps) of degree $d$ over $\mathbb{F}_q$ is bounded below by [Flynn and Garton 2014, Corollary 2.3 and Theorem 3.6]

$$\tfrac{1}{2}\log q - 4.$$

To describe their method, which is the starting point for this paper, we require a definition and an observation. If a map $f$ has a periodic point $s$ of period $k$, with orbit

$$s = s_1 \xrightarrow{f} \cdots \xrightarrow{f} s_k \xrightarrow{f} s_1,$$

then we refer to its orbit as a *cycle* (cycles of length $k$ are called *$k$-cycles*); see [Vasiga and Shallit 2004] for more exposition and illustrations of the cycle structure of functional graphs. This definition is especially useful since it allows for the following observation.

**Observation 1.2.** Components of $\Gamma_f$ are in one-to-one correspondence with the cycles of $f$.

To obtain their results, Flynn and the second author used Lagrange interpolation to interpolate all the cycles of length smaller than the degree of the maps in question. Since they could not interpolate longer cycles,

- they obtained only a lower bound for $\mathcal{P}(q, d)$, and
- their result required that $d$ be at least $\sqrt{q}$.

See Remark 2.5 for a discussion on the relationship between the results of this paper and the results of [Flynn and Garton 2014]; for example, they proved that the random variables mentioned above are indeed uncorrelated in certain cases.

The cycle structure of functional graphs of polynomials over finite fields has been studied extensively in certain cases. Vasiga and Shallit [2004] studied the cycle structure of $\Gamma_f$ for the cases $f = x^2$ and $f = x^2 - 2$, as did Rogers [1996] for $f = x^2$. For any $m \in \mathbb{Z}^{>0}$, the squaring function is also defined over $\mathbb{Z}/m\mathbb{Z}$; Carlip and Mincheva [2008] addressed this situation for certain $m$. Similarly, Chou and Shparlinski [2004] studied the cycle structure of repeated exponentiation over finite fields of prime size. In the context of Pollard's rho algorithm for factoring integers (see [Pollard 1975]), researchers have provided copious data and heuristic arguments supporting the claim that quadratic polynomials produce as many "collisions" as random functions, but very little has been proven (see [Pollard 1975; Bach 1991]). For many other aspects of functional graphs besides their cycle structure, see [Flajolet and Odlyzko 1990] for a study of about twenty characteristic parameters of random mappings in various settings.

More recently, Burnette and Schmutz [2017] used the probabilistic point of view to study a similar question to the one we address here. If $f$ is a polynomial (or rational function) over $\mathbb{F}_q$, define the *ultimate period of $f$* to be the least common

multiple of the cycle lengths of $\Gamma_f$. They found a lower bound for the average ultimate period of polynomials (and rational functions) of fixed degree, whenever the degree of the maps in question, and the size of the finite field, were large enough.

## 2. Two families of random variables

In this section, we define two families of random variables and compute their expected values. The interaction of these random variables determines the answer to Question 1.1; see Remark 2.4 and the remarks that follow for details about this interaction. For the remainder of the section, fix a prime power $q$ and positive integer $d$. Now, for any set $S$ and $C \subseteq S \times S$, we say that $C$ is *consistent* if and only if it has the following property: if $(a, b), (a, c) \in C$, then $b = c$. Next, for any $k \in \mathbb{Z}^{\geq 0}$, define

$$\mathfrak{C}(q, k) = \{C \subseteq \mathbb{F}_q \times \mathbb{F}_q \mid C \text{ is consistent and } |C| = k\}.$$

Any element of $C \in \mathfrak{C}(q, k)$ defines a directed graph with vertex set $\mathbb{F}_q$ and edge set $\{s \to t \mid (s, t) \in C\}$; let $X_{q,k} : \mathfrak{C}(q, k) \to \{0, 1\}$ be the binary random variable that detects whether or not an element of $\mathfrak{C}(q, k)$ defines a graph that happens to be a $k$-cycle. If $f \in \mathbb{F}_q[x]$ and $C \in \mathfrak{C}(q, k)$, we say that $f$ *satisfies* $C$ if $f(a) = b$ for all $(a, b) \in C$. Next, we let

$$Y_{q,d,k} : \mathfrak{C}(q, k) \to \mathbb{Z}^{\geq 0}$$

be the random variable defined by

$$Y_{q,d,k}(C) = |\{f \in \mathbb{F}_q[x] \mid \deg f = d \text{ and } f \text{ satisfies } C\}|.$$

Before computing the expected values of $X_{q,k}$ and $Y_{q,d,k}$, we first mention the size of their sample space.

**Remark 2.1.** If $k \in \mathbb{Z}^{>0}$, then

$$|\mathfrak{C}(q, k)| = q^k \binom{q}{k}.$$

*Proof.* Since the elements of $\mathfrak{C}(q, k)$ are consistent, there are $\binom{q}{k}$ possible choices for the sets of abscissas for any choice of ordinates. Since the ordinates of elements of $\mathfrak{C}(q, k)$ are unrestricted, we conclude that $|\mathfrak{C}(q, k)| = \binom{q}{k} q^k$. □

**Remark 2.2.** If $k \in \{1, \ldots, q\}$, then

$$\mathbb{E}[X_{q,k}] = \frac{q(q-1) \cdots (q - (k-1))}{k |\mathfrak{C}(q, k)|} = \frac{(k-1)!}{q^k}.$$

*Proof.* Since

$$\mathbb{E}[X_{q,k}] = \frac{|\{C \in \mathfrak{C}(q, k) \mid C \text{ is a cycle}\}|}{|\mathfrak{C}(q, k)|},$$

we only need to count the number of elements in $\mathfrak{C}(q, k)$ that are cycles. Since there are

$$\frac{q(q-1)\cdots(q-(k-1))}{k}$$

cycles of length $k$, we conclude by Remark 2.1. □

**Proposition 2.3.** *If $k \in \{1, \ldots, q\}$, then $\mathbb{E}[Y_{q,d,k}] = q^{d+1-k} - q^d$.*

*Proof.* Since

$$\sum_{C \in \mathfrak{C}(q,k)} Y_{q,d,k}(C) = \sum_{C \in \mathfrak{C}(q,k)} |\{f \in \mathbb{F}_q[x] \mid \deg f = d \text{ and } f \text{ satisfies } C\}|$$

$$= \sum_{\substack{f \in \mathbb{F}_q[x] \\ \deg f = d}} |\{C \in \mathfrak{C}(q, k) \mid C \text{ is satisfied by } f\}|$$

$$= \sum_{\substack{f \in \mathbb{F}_q[x] \\ \deg f = d}} \binom{q}{k} = (q^{d+1} - q^d)\binom{q}{k},$$

we see by Remark 2.1 that

$$\mathbb{E}[Y_{q,d,k}] = |\mathfrak{C}(q, k)|^{-1} \cdot \sum_{C \in \mathfrak{C}(q,k)} Y_{q,d,k}(C)$$

$$= \frac{(q^{d+1} - q^d)\binom{q}{k}}{q^k \binom{q}{k}} = q^{d+1-k} - q^{d-k}. \qquad \Box$$

**Remark 2.4.** If we assume that $X_{q,d}$, $Y_{q,d,k}$ are uncorrelated for all $k \in \{1, \ldots, q\}$, then $\mathcal{K}(q) = \mathcal{P}(q, d)$.

*Proof.* Note that for any $k \in \{1, \ldots, q\}$,

$$\sum_{\substack{f \in \mathbb{F}_q[x] \\ \deg f = d}} |\{k\text{-cycles in } \Gamma_f\}| = \sum_{C \in \mathfrak{C}(q,k)} X_{q,k} Y_{q,d,k}(C)$$

$$= |\mathfrak{C}(q, k)| \, \mathbb{E}[X_{q,k} Y_{q,d,k}]$$

$$= |\mathfrak{C}(q, k)| \, \mathbb{E}[X_{q,k}] \, \mathbb{E}[Y_{q,d,k}] \quad \text{by assumption.}$$

Now we can apply Remarks 2.1 and 2.2, along with Proposition 2.3, to see that

$$\mathcal{P}(q, d) = \frac{|\mathfrak{C}(q, k)|}{q^{d+1} - q^d} \cdot \sum_{k=1}^{q} \mathbb{E}[X_{q,k}] \, \mathbb{E}[Y_{q,d,k}]$$

$$= \sum_{k=1}^{q} \frac{q(q-1)\cdots(q-(k-1))}{kq^k}$$

$$= \mathcal{K}(q) \quad \text{by [Kruskal 1954, Equation 16].} \qquad \Box$$

**Remark 2.5.** Unfortunately, we must face up to the fact that the random variables $X_{q,d}$, $Y_{q,d,k}$ are not uncorrelated for all $k \in \{1, \ldots, q\}$. Indeed, if they were, then the computations from Remark 2.4 would show that

$$\sum_{\substack{f \in \mathbb{F}_q[x] \\ \deg f = 2}} |\{q\text{-cycles in } \Gamma_f\}| = \frac{q!\,(q-1)}{q^{q-2}}.$$

But, if $q > 3$, then the quantity on the left is an integer, and the quantity on the right is not! In Section 3, we propose a heuristic that is more reasonable than that these two random variables are uncorrelated.

On the other hand, we should note that the variables $X_{q,d}$, $Y_{q,d,k}$ are indeed uncorrelated whenever $k \in \{1, \ldots, d\}$; this is the content of [Flynn and Garton 2014, Lemma 2.1].

## 3. The heuristic assumption and its implications

As mentioned in Remark 2.5, the variables $X_{q,d}$, $Y_{q,d,k}$ are not uncorrelated for all $k \in \{1, \ldots, q\}$. In this section, we propose a weaker heuristic for these variables, one which nevertheless implies $\mathcal{P}(q, d) = \mathcal{K}(q) + O(1)$.

**Heuristic 3.1.** For any $k \in \mathbb{Z}^{>0}$ and any $d \in \mathbb{Z}^{\geq 0}$,

$$\mathbb{E}[X_{q,k} Y_{q,d,k}] = \mathbb{E}[X_{q,k}]\,\mathbb{E}[Y_{q,d,k}] + O(q^{d-2k}).$$

Here, the implied constant depends only on $d$.

In fact, Heuristic 3.1 implies more than $\mathcal{P}(q, d) = \mathcal{K}(q) + O(1)$; we state the stronger implication here as a conjecture after one more definition. If $k \in \mathbb{Z}^{>0}$ and any $d \in \mathbb{Z}^{\geq 0}$, let

$$\mathcal{P}(q, d, k) := \frac{1}{|\{f \in \mathbb{F}_q[x] \mid \deg f = d\}|} \cdot \sum_{\substack{f \in \mathbb{F}_q[x] \\ \deg f = d}} |\{k\text{-cycles in } \Gamma_f\}|.$$

**Conjecture 3.2.** For any $k \in \mathbb{Z}^{>0}$ and any $d \in \mathbb{Z}^{\geq 0}$,

$$\mathcal{P}(q, d, k) = \frac{q(q-1)\cdots(q-(k-1))}{kq^k} + O(1/q),$$

where the implied constant depends only on $d$. In particular, $\mathcal{P}(q, d) = \mathcal{K}(q) + O(1)$.

**Theorem 3.3.** *If Heuristic 3.1 is true, then Conjecture 3.2 is true.*

*Proof.* As in the proof of Remark 2.4, Heuristic 3.1 immediately implies that

$$\sum_{\substack{f \in \mathbb{F}_q[x] \\ \deg f = d}} |\{k\text{-cycles in } \Gamma_f\}| = |\mathfrak{C}(q, k)| \big( \mathbb{E}[X_{q,k}]\,\mathbb{E}[Y_{q,d,k}] + O(q^{d-2k}) \big).$$

Next, we can apply Remarks 2.1 and 2.2, along with Proposition 2.3, to see that

$$\sum_{\substack{f\in\mathbb{F}_q[x] \\ \deg f=d}} |\{k\text{-cycles in }\Gamma_f\}|$$
$$= \frac{q(q-1)\cdots(q-(k-1))}{kq^k}(q^{d+1}-q^d)+\binom{q}{k}q^k\cdot O(q^{d-2k})$$
$$= \frac{q(q-1)\cdots(q-(k-1))}{kq^k}(q^{d+1}-q^d)+O(q^d).$$

To conclude, note that

$$\mathcal{P}(q,d,k) = \frac{1}{q^{d+1}-q^d}\cdot\sum_{\substack{f\in\mathbb{F}_q[x] \\ \deg f=d}} |\{k\text{-cycles in }\Gamma_f\}|$$
$$= \frac{q(q-1)\cdots(q-(k-1))}{kq^k}+O(1/q). \qquad\square$$

**Remark 3.4.** The available numerical data suggests that the implied constants in Heuristic 3.1 could be quite small. For example, the constant for $d=2$ seems as if it could be as small as 60 (see Section 4 for more details on the available data).

## 4. Numerical evidence

In constructing numerical evidence for Conjecture 3.2, we computed the number of cycles of every length for all polynomials in $\mathbb{F}_q[x]$

- of degree 2, up to $q=241$, and
- of degree 3 up to $q=73$.

For the remainder of the section, we will address only the quadratic case; a similar analysis works for the cubic case.

Of course, if we let $\mathfrak{Q}=\{q\in\mathbb{Z}\mid q \text{ is a prime power, and } 2\le q\le 241\}$, then for any $k\in\{1,\ldots,241\}$, there is certainly a constant — let's call it $C_k$ — for which

$$\left|\mathcal{P}(q,2,k)-\frac{q(q-1)\cdots(q-(k-1))}{kq^k}\right|\le C_k\cdot 1/q \quad\text{for all } q\in\mathfrak{Q}.$$

There are two obvious questions to ask about these constants, which we will address in turn:

- For any particular $k$, how plausible is it that

$$\left|\mathcal{P}(q,2,k)-\frac{q(q-1)\cdots(q-(k-1))}{kq^k}\right|\le C_k\cdot 1/q$$

  for *all* prime powers $q$?

- Even if

$$P(q, 2, k) = \frac{q(q-1) \cdots (q-(k-1))}{kq^k} + O(1/q)$$

for all $k \in \mathbb{Z}^{>0}$, does it seem likely that the implied constants are bounded, as asserted by Conjecture 3.2?

To answer the former question, we could plot, for various $k$,

$$P(q, 2, k) \quad \text{and} \quad \frac{q(q-1) \cdots (q-(k-1))}{kq^k} \pm C_k \cdot 1/q.$$

But, as these numbers quickly become minuscule, it is convenient to let

$$\widehat{P}(q, d, k) = \left| \{ f \in \mathbb{F}_q[x] \mid \deg f = d \} \right| \cdot P(q, d, k) = (q^{d+1} - q^d) \cdot P(q, d, k);$$

that is, $\widehat{P}(q, d, k)$ is the number of $k$-cycles appearing in functional graphs of polynomials in $\mathbb{F}_q[x]$ of degree $d$. Conjecture 3.2 predicts that this quantity is about

$$(q^{d+1} - q^d) \cdot \frac{q(q-1) \cdots (q-(k-1))}{kq^k},$$

which we will denote by $\mathcal{G}(q, d, k)$. By the definition of $C_k$, we know that for all $q \in \mathfrak{Q}$ and $k \in 1, 2, \ldots, 241$,

$$|\widehat{P}(q, 2, k) - \mathcal{G}(q, 2, k)| \le C_k(q^2 - q).$$

As two examples of the data we have compiled, we include plots of $\widehat{P}(q, 2, k)$ and $\mathcal{G}(q, 2, k) \pm C_k(q^2 - q)$ for $k = 6, 10$, where $C_6 = 59$ and $C_{10} = 14$; see Figure 1. These graphs are typical for $k \in \{1, \ldots, 241\}$.

To address the second question mentioned above, we plot the various values of $C_k$ in the hopes that they appear to be bounded. This graph is shown in Figure 2.

We should point out that the small values of $C_k$ in Figure 2 are a result of the fact that in our data, we simply found no $k$-cycles at all for all $k > 82$. So from $k = 82$ onward, the graph is simply plotting

$$\frac{241!}{(241 - k)! \cdot k \cdot 241^{k-1}}.$$

This begs two questions:

- As cycles of larger length arise for larger values of $q$, will the size of $C_k$ increase?

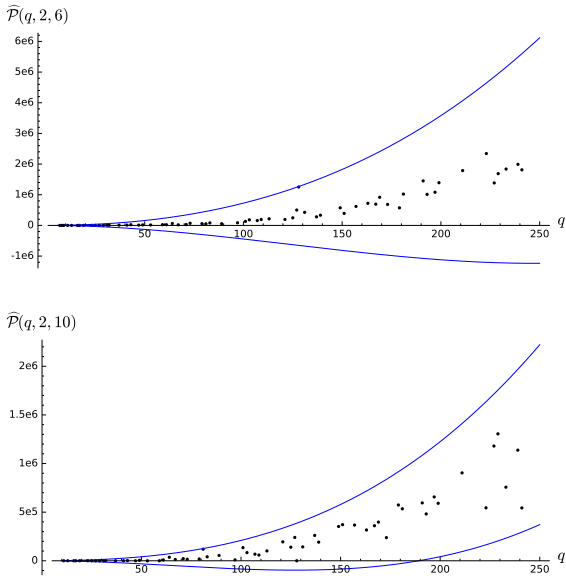- Conversely, if these cycles do not arise promptly, will this increase the size of $C_k$?

**Figure 1.** Plots of $\widehat{\mathcal{P}}(q, 2, k)$ and $\mathcal{G}(q, 2, k) \pm C_k(q^2-q)$ for $k=6, 10$. Top: $C_6 \approx 59.06$; bottom: $C_{10} \approx 14.86$.

Of course, we cannot answer these questions, but note that for the particular value of $k = 82$, the quadratic polynomials we tested yielded exactly 27722 82-cycles (all appearing when $q = 167$), whereas for $k \in \{70, \ldots, 81\}$, they yielded exactly zero. That is, this is an example of a cycle of larger length arising without affecting the maximum of the $C_k$.

As for the second question, the lack of $k$-cycles will not cause $C_k$ to rise above 60 as long as the first $k$-cycle appears in a graph for a finite field of size less than $60k$. For example, the smallest $q$ for which 62-cycles appear is $q = 128$ (which is well
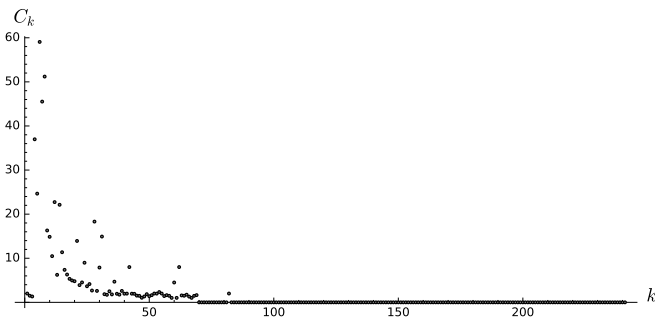


**Figure 2.** Various values of $C_k$.

under $60 \cdot 62$). The smallest cycle length that does not appear for $q \in \mathfrak{Q}$ is $k = 43$; if a 43-cycle does not appear by the time $q = 2579$, then $C_{43}$ will rise above 60. It is unfortunately beyond our abilities to determine if a 43-cycle appears by this time.

## 5. Rational functions

In this section, we briefly mention the results for rational functions, which are analogous to those for polynomials. For any prime power $q$ and $d \in \mathbb{Z}^{\geq 0}$, let

$$\mathcal{R}(q, d) := \frac{1}{|\{f \in \mathbb{P}^1(\mathbb{F}_q)[x] \mid \deg(f) = d\}|} \cdot \sum_{\substack{f \in \mathbb{P}^1(\mathbb{F}_q)[x] \\ \deg(f) = d}} |\{\text{cycles in } \Gamma(\mathbb{P}^1(\mathbb{F}_q), f)\}|.$$

If $k \in \mathbb{Z}^{>0}$, we can define $\mathcal{R}(q, d, k)$ in exactly the same way as $\mathcal{P}(q, d, k)$.

To define our new families of random variables, for any prime power $q$ and $k \in \mathbb{Z}^{>0}$, let

$$\mathfrak{T}(q, k) = \{T \subseteq \mathbb{P}^1(\mathbb{F}_q) \times \mathbb{P}^1(\mathbb{F}_q) \mid T \text{ is consistent and } |T| = k\},$$

and $V_{q,k} : \mathfrak{T}(q, k) \to \{0, 1\}$ be the binary random variable that detects whether or not an element of $\mathfrak{T}(q, k)$ is a $k$-cycle. If $d \in \mathbb{Z}^{\geq 0}$, let

$$W_{q,d,k} : \mathfrak{T}(q, k) \to \mathbb{Z}^{\geq 0}$$

be the random variable defined by

$$W_{q,d,k}(T) = |\{f \in \mathbb{F}_q(x) \mid \deg f = d \text{ and } f \text{ satisfies } T\}|.$$

The rational function analogs of Remark 2.1, Remark 2.2, Proposition 2.3 are proved as above, leading to the following conjecture, which again follows from the heuristic that the random variables $V_{q,k}$, $W_{q,d,k}$ are "uncorrelated enough".

**Conjecture 5.1.** For any $k \in \mathbb{Z}^{>0}$ and any $d \in \mathbb{Z}^{\geq 0}$,

$$\mathcal{R}(q, d, k) = \frac{(q+1)q \cdots (q - (k-2))}{k(q+1)^k} + O(1/q),$$

where the implied constant depends only on $d$. In particular, $\mathcal{R}(q, d) = \mathcal{K}(q+1) + O(1)$.

**Heuristic 5.2.** If $k \in \{1, \ldots, q\}$, and $d \in \mathbb{Z}^{\geq 0}$, then

$$\mathbb{E}[V_{q,k} W_{q,d,k}] = \mathbb{E}[V_{q,k}] \, \mathbb{E}[W_{q,d,k}] + O(q^{2d-2k}).$$

Here, the implied constant depends only on $d$.

**Theorem 5.3.** *If Heuristic 5.2 is true, then Conjecture 5.1 is true.*

*Proof.* Similar to the proof of Theorem 3.3. □

## Acknowledgments

The authors would like to thank Ian Dinwoodie, Rafe Jones, and Christopher Kramer for their help and advice.

## References

[Bach 1991]  E. Bach, "Toward a theory of Pollard's rho method", *Inform. and Comput.* **90**:2 (1991), 139–155. MR Zbl

[Burnette and Schmutz 2017]  C. Burnette and E. Schmutz, "Periods of iterated rational functions", *Int. J. Number Theory* **13**:5 (2017), 1301–1315. MR

[Carlip and Mincheva 2008]  W. Carlip and M. Mincheva, "Symmetry of iteration graphs", *Czechoslovak Math. J.* **58(133)**:1 (2008), 131–145. MR Zbl

[Chou and Shparlinski 2004]  W.-S. Chou and I. E. Shparlinski, "On the cycle structure of repeated exponentiation modulo a prime", *J. Number Theory* **107**:2 (2004), 345–356. MR Zbl

[Flajolet and Odlyzko 1990]  P. Flajolet and A. M. Odlyzko, "Random mapping statistics", pp. 329–354 in *Advances in cryptology: EUROCRYPT '89* (Houthalen, 1989), edited by J.-J. Quisquater and J. Vandewalle, Lecture Notes in Comput. Sci. **434**, Springer, Berlin, 1990. MR Zbl

[Flynn and Garton 2014]  R. Flynn and D. Garton, "Graph components and dynamics over finite fields", *Int. J. Number Theory* **10**:3 (2014), 779–792. MR Zbl

[Kruskal 1954]  M. D. Kruskal, "The expected number of components under a random mapping function", *Amer. Math. Monthly* **61** (1954), 392–397. MR Zbl

[Pollard 1975]  J. M. Pollard, "A Monte Carlo method for factorization", *Nordisk Tidskr. Inform.* **15**:3 (1975), 331–334. MR Zbl

[Rogers 1996]  T. D. Rogers, "The graph of the square mapping on the prime fields", *Discrete Math.* **148**:1-3 (1996), 317–324. MR Zbl

[Vasiga and Shallit 2004]  T. Vasiga and J. Shallit, "On the iteration of certain quadratic maps over GF($p$)", *Discrete Math.* **277**:1-3 (2004), 219–240. MR Zbl

ebellah@uoregon.edu          *Department of Mathematics, University of Oregon, Eugene, OR 97403, United States*

gartondw@pdx.edu             *Fariborz Maseeh Department of Mathematics and Statistics, Portland State University, PO Box 751, Portland, OR 97207, United States*

ejt3@pdx.edu                 *Fariborz Maseeh Department of Mathematics and Statistics, Portland State University, PO Box 751, Portland, OR 97207, United States*

nwalton@pdx.edu              *Fariborz Maseeh Department of Mathematics and Statistics, Portland State University, PO Box 751, Portland, OR 97207, United States*

msp

# Guidelines for Authors

Submissions in all mathematical areas are encouraged. All manuscripts accepted for publication in *Involve* are considered publishable in quality journals in their respective fields, and include a minimum of one-third student authorship. Submissions should include substantial faculty input; faculty co-authorship is strongly encouraged. Authors may submit manuscripts in PDF format on-line at the Submission page at the Involve website.

**Originality**. Submission of a manuscript acknowledges that the manuscript is original and 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, 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.