involve

a journal of mathematics

msp

2020 vol. 13, no. 5

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.

PRODUCTION

Silvio Levy, Scientific Editor

Cover: Alex Scorpan

See inside back cover or msp.org/involve for submission instructions. The subscription price for 2020 is US $205/year for the electronic version, and $275/year (+$35, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues and changes of subscriber address should be sent to MSP.

Involve (ISSN 1944-4184 electronic, 1944-4176 printed) at Mathematical Sciences Publishers, 798 Evans Hall #3840, c/o University of California, Berkeley, CA 94720-3840, is published continuously online. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices.

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

PUBLISHED BY

mathematical sciences publishers

nonprofit scientific publishing

http://msp.org/

INVOLVE 13:5 (2020)
https://doi.org/10.2140/involve.2020.13.721

Set-valued domino tableaux and shifted set-valued domino tableaux

Florence Maas-Gariépy and Rebecca Patrias

(Communicated by Kenneth S. Berenhaut)

We prove K-theoretic and shifted K-theoretic analogues of the bijection of Stanton and White between domino tableaux and pairs of semistandard tableaux. As a result, we obtain product formulas for certain pairs of stable Grothendieck polynomials and certain pairs of K-theoretic Q-Schur functions.

1. Introduction

Recall that a *partition* is a finite, nonincreasing sequence of positive integers $\lambda = (\lambda_1, \dots, \lambda_t)$ and that any partition can be identified with the corresponding *Young diagram* — a left-justified array of boxes with λ_i boxes in the i-th row from the top. A filling of the boxes of a Young diagram with nonnegative integers such that entries weakly increase across rows and strictly increase down columns gives a *semistandard Young tableau*. The *Schur functions* are symmetric functions that are indexed by partitions. Each element of the set of Schur functions can be defined as a weighted generating function of semistandard Young tableaux of the corresponding partition shape. The set of Schur functions forms a linear basis for the ring of symmetric functions and appears naturally in many areas of mathematics including representation theory, Schubert calculus, and gauge theory.

Of particular interest is the question of how to express a product of two Schur functions since the answer has meaning in the fields mentioned above. One way to answer this question for certain products is to consider *domino tableaux*, where a domino tableau is an array of dominoes (2×1 and 1×2 pieces) in the shape of a Young diagram, where each domino is filled with a positive integer, rows are weakly increasing and columns are strictly increasing. D. Stanton and D. White [1985] proved that for arbitrary partitions μ and ν, the product $s_\mu s_\nu$ can be written as a sum of weighted generating functions of domino tableaux. Their proof was later simplified by C. Carré and B. Leclerc [1995].

MSC2010: 05E05.
Keywords: set-valued tableaux, Schur functions, stable Grothendieck polynomials, Lagrangian Grassmannian.

A well-known analogue of the Schur functions is the set of *Q-Schur functions* $\{Q_\lambda\}$, which are indexed by partitions $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_t)$ with $\lambda_t \geq t$. Z. Chemli [2016] introduced the notion of a *shifted domino tableau* and proves the analogue of the Stanton–White result in this setting. Namely, he proves that for shifted partitions μ and ν, the product $Q_\mu Q_\nu$ can be written as a sum of weighted generating functions of shifted domino tableaux.

In addition to this shifted analogue of the Schur functions, there is also a natural K-theoretic analogue called the *stable Grothendieck polynomials*, denoted by G_λ [Fomin and Kirillov 1996; Buch 2002]. Combinatorially, we obtain the stable Grothendieck polynomials by allowing finite, nonempty subsets of positive integers to fill the boxes of a Young diagram instead of only allowing single entries. The stable Grothendieck polynomials are called K-theoretic analogues because where there is a deep connection between Schur functions and cohomology of the Grassmannian, there is the same connection between stable Grothendieck polynomials and K-theory of the Grassmannian. A reader unfamiliar with cohomology theory and K-theory need not worry; we will only address the combinatorial properties of these symmetric functions. There is also a natural K-theoretic analogue of the Q-Schur functions, i.e., a natural shifted analogue of the stable Grothendieck polynomials [Ikeda and Naruse 2013; Graham and Kreiman 2015], denoted by GQ_λ. The GQ_λ appear in the study of the K-theory of the Lagrangian Grassmannian.

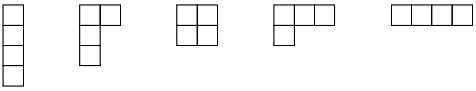
In this paper, we prove the K-theoretic analogue of both the Stanton–White result and the Chemli result, thus obtaining product formulas for pairs $G_\mu G_\nu$ and $GQ_\mu GQ_\nu$. Note that $G_\mu G_\nu$ and $GQ_\mu GQ_\nu$ expand positively in terms of the G_λ's and GQ_λ's, respectively; however, no combinatorial description of this $GQ_\mu GQ_\nu$ expansion is known. To obtain our results, we introduce the notions of *set-valued domino tableaux* and *shifted set-valued domino tableaux*.

The paper proceeds as follows. In Section 2, we review the necessary combinatorial background for the rest of the paper: tableaux, Young diagrams, symmetric functions, and Schur functions. Section 3 gives an introduction to domino tableaux, and Section 4 explains the bijection that leads to the result of Stanton and White. In Section 5, we introduce the stable Grothendieck polynomials and set-valued domino tableaux, and we prove the K-theoretic analogue of the Stanton–White result. Section 6 gives the necessary background on Q-Schur functions, shifted Young tableaux and shifted domino tableaux, and reviews the result of Chemli. We conclude in Section 7 by proving the shifted K-theoretic analogue of Chemli's result.

2. Preliminaries

We begin by reviewing basic notions related to symmetric functions, partitions, and Young tableaux. We refer the reader to [Stanley 1999] for a more in-depth study of these topics.

2A. ***Partitions and tableaux.*** A *partition* is a finite, nonincreasing sequence of positive integers $\lambda = (\lambda_1, \dots, \lambda_k)$. We say that λ is a *partition of n*, written $\lambda \vdash n$ or $|\lambda| = n$, when $\lambda_1 + \cdots + \lambda_k = n$. For example, (2, 1, 1) is a partition of 4, and there are five partitions of 4 in total. To each partition, we can associate a *Young diagram*: a left-justified array of boxes with λ_i boxes in the i-th row from the top. We often equate a partition λ with its Young diagram. The Young diagrams for the partitions of 4 are shown below:



A *semistandard Young tableau* of shape λ is a filling of the boxes of the Young diagram of shape λ with positive integers such that the entries weakly increase from left to right along rows and strictly increase down columns. A *standard Young tableau* of shape $\lambda \vdash n$ is a semistandard Young tableau of shape λ such that each positive integer $1, 2, \dots, n$ appears exactly once. For a semistandard Young tableau T, let $\mathrm{sh}(T)$ denote the shape of T and $|T|$ denote the number of boxes of T or, equivalently, the number of entries in T. For example, T_1 below is a semistandard Young tableau and T_2 is a standard Young tableau. We see that $|T_1| = |T_2| = 11$ and $\mathrm{sh}(T_1) = \mathrm{sh}(T_2) = (5, 3, 3)$:

$$T_1 = \begin{array}{|c|c|c|c|c|} \hline 1 & 1 & 1 & 3 & 4 \\ \hline 3 & 3 & 5 \\ \cline{1-3} 4 & 5 & 7 \\ \cline{1-3} \end{array} \qquad T_2 = \begin{array}{|c|c|c|c|c|} \hline 1 & 3 & 4 & 9 & 11 \\ \hline 2 & 5 & 7 \\ \cline{1-3} 6 & 8 & 10 \\ \cline{1-3} \end{array}$$

Consider a semistandard Young tableau as sitting in the southeast quadrant of the plane with top left corner at the origin and each box of side length 1. Notice that each cell of the tableau is crossed by a unique diagonal D_k, where D_k is the line $-x + k$ for some $k \in \mathbb{Z}$. For example, the boxes of T_2 with entries 1, 5, and 10 lie on D_0, while the boxes with entries 2 and 8 lie on D_{-1}. The *diagonal reading word* of a semistandard tableau T is the word obtained by reading the entries along each diagonal from northwest to southeast starting with the bottom diagonal. We insert the symbol "/" between the segments obtained from each diagonal. For example, the diagonal reading words for T_1 and T_2 are respectively $4\,/\,3, 5\,/\,1, 3, 7\,/\,1, 5\,/\,1\,/\,3\,/\,4$ and $6\,/\,2, 8\,/\,1, 5, 10\,/\,3, 7\,/\,4\,/\,9\,/\,11$. The diagonal reading word of a semistandard Young tableau defines it uniquely, a property that we will use in Theorem 4.1.

2B. ***Symmetric functions.*** A *weak composition* is a countable sequence of nonnegative integers $\alpha = (\alpha_1, \alpha_2, \dots)$ such that only finitely many α_i are nonzero. Let S_n denote the symmetric group of order $n!$, the group of all permutations of the set $\{1, 2, \dots, n\}$.

Let $x = (x_1, x_2, \dots)$ be a countable set of variables, and for a weak composition α, define x^α to be $x_1^{\alpha_1} x_2^{\alpha_2} \cdots$. A *symmetric function* is a formal power series $f(x) = \sum_\alpha c_\alpha x^\alpha$ such that $c_\alpha \in \mathbb{R}$ and such that, for any nonnegative integer n and any $\sigma \in S_n$,

$$f(x_{\sigma(1)}, x_{\sigma(2)}, \dots, x_{\sigma(n)}, x_{n+1}, \dots) = f(x_1, x_2, \dots).$$

We say that a symmetric function is *homogeneous of degree n* if each of its monomials has degree n. We denote the set of homogeneous symmetric functions of degree n by Λ^n and the set of symmetric functions by Λ.

For example,

$$f(x) = \sum_{i \le j} x_i x_j = x_1^2 + x_1 x_2 + x_1 x_3 + \cdots \in \Lambda^2$$

is a homogeneous symmetric function of degree 2,

$$g(x) = x_1 + x_2 + \cdots + x_1^2 + x_2^2 + \cdots \in \Lambda$$

is a symmetric function but is not homogeneous of any degree, and

$$h(x) = x_1 + \sum_{i \in \mathbb{N}} x_i^2$$

is not a symmetric function.

It is easy to see that Λ is an algebra with identity element $1 \in \Lambda^0$. In other words, Λ is an $\mathbb{R}$-vector space under addition and a ring under multiplication.

2C. ***Schur functions.*** The algebra of symmetric functions has many nice bases, which are well studied. We next introduce one such basis: the basis of Schur functions. This basis is of great interest because of its connections to other areas of mathematics. For example, Schur functions are closely related to the irreducible representations of both the symmetric group and the general linear group. They also appear in the area of Schubert calculus as a tool for computing the structure constants in the cohomology ring of the Grassmannian. There are many ways to define the Schur functions and we use the combinatorial definition.

Let T be a semistandard Young tableau. We can associate a monomial x^T in the variable set $(x_1, x_2, \dots)$ to T by letting the exponent of x_i be the number of times the entry i appears in T. For example, $x^{T_1} = x_1^3 x_3^3 x_4^2 x_5^2 x_7$ and $x^{T_2} = x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 x_9 x_{10} x_{11}$ for T_1 and T_2 from Section 2A.

The basis of Schur functions is indexed by partitions. We define the *Schur function* s_λ by

$$s_\lambda = s_\lambda(x) = \sum_{\mathrm{sh}(T) = \lambda} x^T,$$

where we sum over all semistandard Young tableaux T of shape λ. Note that if $\lambda \vdash n$, then $s_\lambda \in \Lambda^n$. It is easy to see that each monomial has degree n, but it is not obvious from the combinatorial definition that s_λ is indeed symmetric.

Example 2.1. We can compute that

$$s_{(2,1)} = x_1^2x_2 + x_1^2x_3 + x_1^2x_9 + x_1x_2^2 + x_1x_2x_3 + x_1x_2x_3 + x_1x_3^2 + x_2^2x_3 + x_2x_3^2 + \cdots ,$$

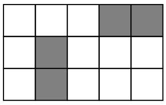
where the monomials given correspond to the semistandard tableaux below:

$$\begin{array}{|c|c|}\hline 1 & 1 \\ \hline 2 \\ \cline{1-1}\end{array}\quad \begin{array}{|c|c|}\hline 1 & 1 \\ \hline 3 \\ \cline{1-1}\end{array}\quad \begin{array}{|c|c|}\hline 1 & 1 \\ \hline 9 \\ \cline{1-1}\end{array}\quad \begin{array}{|c|c|}\hline 1 & 2 \\ \hline 2 \\ \cline{1-1}\end{array}\quad \begin{array}{|c|c|}\hline 1 & 2 \\ \hline 3 \\ \cline{1-1}\end{array}\quad \begin{array}{|c|c|}\hline 1 & 3 \\ \hline 2 \\ \cline{1-1}\end{array}\quad \begin{array}{|c|c|}\hline 1 & 3 \\ \hline 3 \\ \cline{1-1}\end{array}\quad \begin{array}{|c|c|}\hline 2 & 2 \\ \hline 3 \\ \cline{1-1}\end{array}\quad \begin{array}{|c|c|}\hline 2 & 3 \\ \hline 3 \\ \cline{1-1}\end{array}$$

Since the Schur functions form a linear basis for Λ, we know that we can express any product $s_\lambda s_\mu$ as a finite sum of Schur functions: $s_\lambda s_\mu = \sum_\nu c^\nu_{\lambda,\mu} s_\nu$. This idea has applications in representation theory and Schubert calculus, as mentioned above, and has been very well studied. The coefficients $c^\nu_{\lambda,\mu}$ are called *Littlewood–Richardson coefficients*, and there are many combinatorial rules for computing them. In Theorem 4.2, we give a rule for expressing this product as a sum over domino tableaux for certain pairs s_λ and s_μ.

3. Domino tableaux

First, define a *domino* to be a 2×1 or 1×2 rectangle inside of a Young diagram. The shaded shapes below are both dominoes.



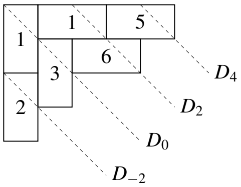
We say that a Young diagram is *pavable* if it can be written as the disjoint union of dominoes. The reader may verify that the partition (2, 2, 2) is pavable, while the partition (5, 5, 5) shown above is not (it has an odd number of boxes) and the partition (5, 3, 3, 2, 1) is not. If λ is pavable, we call any such covering a *domino paving*.

Definition 3.1. A *domino tableau of shape* λ is the filling of a domino paving of λ by positive integers such that

- entries weakly increase from left to right and
- columns strictly increase from top to bottom.

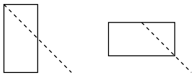
We again think of the top left corner of a domino tableau as sitting at the origin of the plane. In this setting, each domino in a domino tableau is crossed by a unique diagonal D_{2k} of equation $-x + 2k$. We define the *diagonal reading word* of a domino tableau to be the integer sequence obtained by reading northwest to southeast along each diagonal D_{2k} starting with the bottom diagonal. We again separate the entries on distinct diagonals with a slash. Unlike for the diagonal reading of semistandard Young tableaux, the diagonal reading of a domino tableau does not define it uniquely. For example, the diagonal reading "1" could refer to a single vertical domino or a single horizontal domino.

Example 3.2. The following figure represents a domino tableau of shape (5, 4, 2, 1) with its diagonals. The diagonal reading word of this tableau is 2 / 1, 3 / 1, 6 / 5:

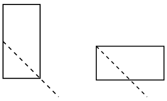


We can divide the dominoes of a domino paving into two categories depending on how they are cut by a diagonal D_{2k}:

(1) We call a domino a *type-1 domino* if the small triangle cut by the diagonal points upward:



(2) We call a domino a *type-2 domino* if the small triangle cut by the diagonal points downward:



We next define the *2-quotient* of a partition $\lambda = (\lambda_1, \dots, \lambda_k) \vdash n$, a pair of partitions (μ, ν) obtained in the following way:

(1) First define $L = (l_1, l_2, \dots, l_k)$, where $l_i = \lambda_i + k - i$ for $i \in \{1, 2, \dots, k\}$.

(2) Let M be obtained from L by successively replacing the even components of L by $0, 2, 4, \dots$ from right to left and the odd components by $1, 3, 5 \dots$ from right to left.

(3) To obtain μ, subtract the even components of L by the even components of M and divide by 2. Delete the components that are 0.

(4) To obtain ν, subtract the odd components of L by the odd components of M and divide by 2. Delete the components that are 0.

Example 3.3. Let $\lambda = (4, 2, 2, 1, 1, 1)$. Then we have

(1) $L = (4+6-1, 2+6-2, 2+6-3, 1+6-4, 1+6-5, 1+6-6) = (9, 6, 5, 3, 2, 1)$,

(2) $M = (7, 2, 5, 3, 0, 1)$,

(3) $\mu = \frac{1}{2}((6, 2) - (2, 0)) = \frac{1}{2}(4, 2) = (2, 1)$, and

(4) $\nu = \frac{1}{2}((9, 5, 3, 1) - (7, 5, 3, 1)) = \frac{1}{2}(2, 0, 0, 0) = (1, 0, 0, 0) = (1)$.

Thus the 2-quotient of $(4, 2, 2, 1, 1, 1)$ is the pair $((2, 1), (1))$.

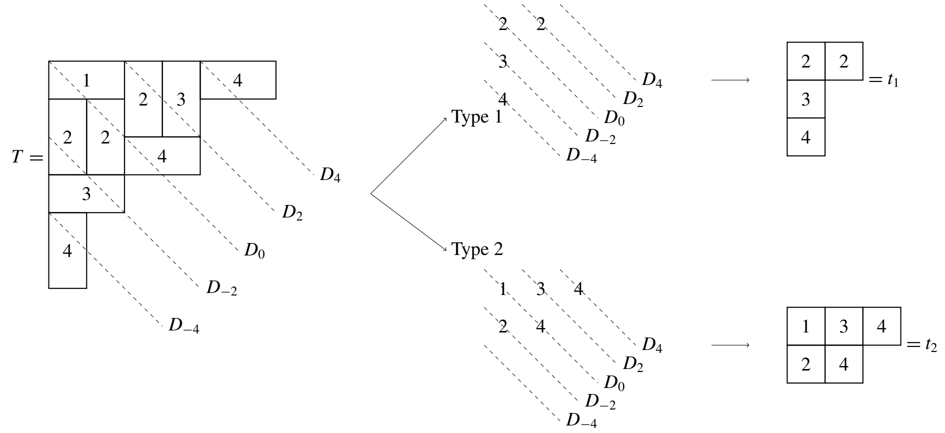
Note that this process is reversible; i.e., every pair of partitions (μ, ν) is the 2-quotient of some partition λ. We discuss the reverse procedure in the next section.

4. Bijection between domino tableaux and semistandard tableaux

We now describe the bijection used to prove the following theorem. Our main result of Section 5 is a generalization of this theorem, so it will be useful in later sections to understand this bijection.

Theorem 4.1 [Carré and Leclerc 1995; Stanton and White 1985]. *Let λ be a pavable partition with* 2*-quotient* (μ, ν)*. There is a bijection between the set of domino tableaux of shape λ and the set of pairs of Young tableaux* (t_1, t_2) *of shape* (μ, ν).

Theorem 4.1 is proven by giving an explicit bijection Γ that sends a domino tableau to the associated pair of Young tableaux. The bijection Γ consists of considering the diagonal reading of entries in type-1 dominoes and of type-2 dominoes separately. More precisely, let T be a domino tableau, and form the diagonal reading word for T. Let w_1 be the word obtained by restricting this diagonal reading word to the entries that come from type-1 dominoes and let w_2 be the word obtained by restricting the diagonal reading word for T to entries coming from type-2 dominoes. We then let semistandard tableau t_1 be the unique Young tableau with diagonal reading word w_1 and let t_2 be the unique Young tableau with diagonal reading w_2. We illustrate this bijection below using an example:



We leave it to the reader to verify that the 2-quotient of $\mathrm{sh}(T) = (6, 4, 4, 2, 1, 1)$ is $((2, 1, 1), (3, 2))$, the shape of (t_1, t_2).

The inverse algorithm, Γ^{-1}, consists of recursively constructing the domino tableau of shape λ associated to a pair of Young tableaux (t_1, t_2) of shape (μ, ν), where (μ, ν) is the 2-quotient of λ. At any step, we have a pair of Young tableaux $(t_1^{(i)}, t_2^{(i)})$ of shape $(\mu^{(i)}, \nu^{(i)})$ and the associated domino tableau $T^{(i)}$ of shape $\lambda^{(i)}$.

We start the algorithm with $\mu^{(0)}=\nu^{(0)}=\lambda^{(0)}=\varnothing$. The algorithm stops when $(t_1^{(s)},t_2^{(s)})=(t_1,t_2)$. Then we have that the domino tableau associated to (t_1,t_2) is $T^{(s)}$. We now describe the i-th step of the algorithm.

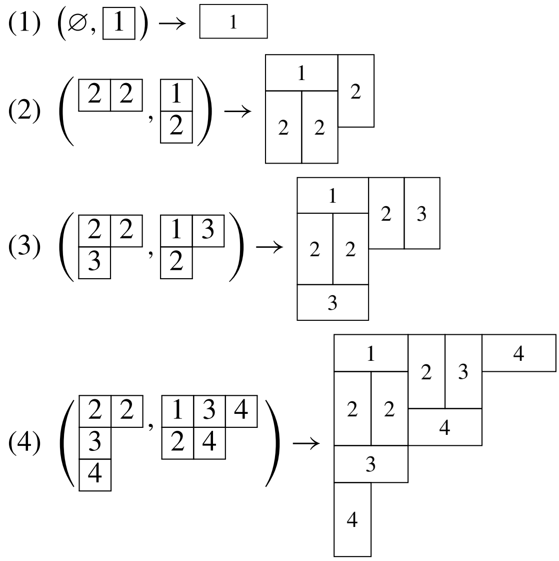
Let u_i be the smallest value appearing in (t_1,t_2) that does not appear in $(t_1^{(i-1)},t_2^{(i-1)})$. We build $(t_1^{(i)},t_2^{(i)})$ of shape $(\mu^{(i)},\nu^{(i)})$ by adding to $(t_1^{(i-1)},t_2^{(i-1)})$ all cells of (t_1,t_2) with value u_i, while preserving their original position.

To construct the domino tableau $T^{(i)}$ of shape $\lambda^{(i)}$, we use the following procedure for each diagonal, starting with the bottom diagonal: For all cells in $t_1^{(i)}$ (resp. $t_2^{(i)}$) containing the value u_i on diagonal D_k, we add to $T^{(i-1)}$ a type-1 domino (resp. a type-2 domino) with entry u_i on the corresponding diagonal D_{2k}. We then get the associated domino tableau $T^{(i)}$ of shape $\lambda^{(i)}$.

Below is an example of Γ^{-1} applied to the pair of Young tableaux

$$(t_1,t_2)=\left(\begin{array}{|c|c|}\hline 2&2\\\hline 3\\\cline{1-1}4\\\cline{1-1}\end{array},\begin{array}{|c|c|c|}\hline 1&3&4\\\hline 2&4\\\cline{1-2}\end{array}\right).$$

Notice that we recover the tableau T from the previous example:



The following is a corollary of Theorem 4.1. It is important to note that this result holds for any pair of partitions (μ,ν) because we can reverse the 2-quotient procedure, as illustrated in the example of Γ^{-1} above.

Theorem 4.2. *Let λ be a partition with 2-quotient (μ,ν). Then*

$$s_\mu s_\nu=\sum_T x^T,$$

where T runs over the set of domino tableaux of shape λ.

Proof. Each term in the product $s_\mu s_\nu$ is represented by a pair of semistandard Young tableaux of shape (μ, ν). Theorem 4.1 says that these pairs are in bijection with domino tableaux of shape λ, and the domino tableau associated to a pair (t_1, t_2) has the same multiset of entries as (t_1, t_2). □

5. *K*-theoretic generalizations

5A. ***Stable Grothendieck polynomials.*** *K*-theory is a generalized cohomology theory. As mentioned in Section 2C, the Schur functions are deeply connected to the cohomology of the Grassmannian, and play a very specific role in that cohomology. It turns out that there is a generalization of the Schur functions that plays the same role in the *K*-theory of the Grassmannian; these symmetric functions are called the stable Grothendieck polynomials and are denoted by G_λ, where λ is a partition. We will only discuss the combinatorics of *K*-theory here, and it is not necessary for the reader to have any prior knowledge of the geometry.

Stable Grothendieck polynomials were introduced in [Fomin and Kirillov 1996] as certain limits of Lascoux and Schützenberger's Grothendieck polynomials [1982]. We will give the combinatorial definition first explicitly written in [Buch 2002]. The rough idea is that G_λ is defined in the same way as s_λ except that we are allowed to fill the boxes of λ with finite subsets of integers instead of just single integers. To this end, we first define a partial ordering on subsets.

Let A and B be finite, nonempty sets of positive integers. We say that $A \leq B$ if $\max(A) \leq \min(B)$ and $A < B$ if $\max(A) < \min(B)$. For example, $\{1, 3, 4\} \leq \{4, 6, 7, 33\}$, $\{1, 3, 4\} < \{5\}$, and $\{1, 3, 4\}$ is not comparable to $\{2, 5, 7\}$.

Definition 5.1. Let λ be a partition. A *semistandard set-valued tableau of shape* λ is a filling of the Young diagram λ with finite, nonempty sets of positive integers such that

- entries are weakly increasing from left to right along the rows and
- entries are strictly increasing down columns.

Given a semistandard set-valued tableau T, we may again associate a monomial

$$x^T = x_1^{\alpha_1(T)} x_2^{\alpha_2(T)} x_3^{\alpha_3(T)} \cdots,$$

where $\alpha_i(T)$ is the number of occurrences of i in T. We let $|T|$ denote the sum of the sizes of the sets that fill T or, equivalently, the total number of integers filling T. To illustrate, the tableau T shown below has $x^T = x_1 x_3^2 x_4 x_6 x_7 x_9$ and $|T| = 8$:

$$T = \begin{array}{|c|c|c|}\hline 1,3 & 3 & 6,7 \\ \hline 4 & 5,9 \\ \cline{1-2}\end{array}$$

Note that if we pick one representative from each box of a semistandard set-valued tableau of shape λ, we obtain a semistandard Young tableau of shape λ.

We can again define the *diagonal reading word* for a set-valued tableau T in a similar way. We will read the entries northwest to southeast along each diagonal, starting with the bottom diagonal. The only difference with the diagonal reading of a semistandard Young tableau is that we use braces to identify the elements of a set that has more than one element. For example, the diagonal reading word for tableau T above is $4\,/\,\{1,3\},\{5,9\}\,/\,3\,/\,\{6,7\}$. Just like for semistandard tableaux, the diagonal reading of a set-valued tableau defines it uniquely.

Definition 5.2 [Buch 2002]. Let $\lambda \vdash n$. The *stable Grothendieck polynomial G_λ* is

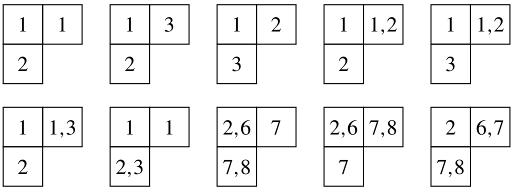
$$G_\lambda = \sum_{\mathrm{sh}(T)=\lambda} (-1)^{|T|-|\lambda|} x^T,$$

where we sum over all semistandard set-valued tableaux T of shape λ.

Example 5.3. We have

$$G_{(2,1)} = x_1^2x_2 + 2x_1x_2x_3 - x_1^2x_2^2 - 3x_1^2x_2x_3 + 3x_2x_6x_7^2x_8 \pm \cdots ,$$

where the terms shown correspond to the tableaux below:



Note that there are additional tableaux with monomial $x_2x_6x_7^2x_8$, so the coefficient of $x_2x_6x_7^2x_8$ in the full $G_{(2,1)}$ is greater than 3.

Notice that since the set of semistandard Young tableaux is contained in the set of semistandard set-valued tableaux, G_λ will contain s_λ as the set of terms of lowest degree. This observation shows us that G_λ is a generalization of s_λ, and, in particular, any formula we have for expressing a product $G_\mu G_\nu$ in terms of stable Grothendieck polynomials will restrict to a formula for $s_\mu s_\nu$ upon restriction to the lowest-degree terms. Notice also that the monomials of G_λ have arbitrarily large degree.

5B. *Set-valued domino tableaux.* In this section, we introduce the notion of a set-valued domino tableau. We will use this object to prove a K-theoretic analogue of Theorem 4.1.

If F is a domino filled with subset A, let $\max(F)$ and $\min(F)$ denote $\max(A)$ and $\min(A)$, respectively. We say a (square) box of a partition is in position (i, j) if it is in the i-th row and j-th column of that partition. Suppose the top left square of domino F_1 is in position (i, j) of pavable partition λ. We say that domino F_2 is *weakly southeast* of domino F_1 if F_2 intersects a box of λ in position (k, ℓ) with $k \geq i$ and $\ell \geq j$. In other words, at least part of F_2 is weakly southeast of the top left square of F_1.

Definition 5.4. A *set-valued domino tableau of shape* λ is the filling of a domino paving of λ with finite, nonempty sets of positive integers such that:

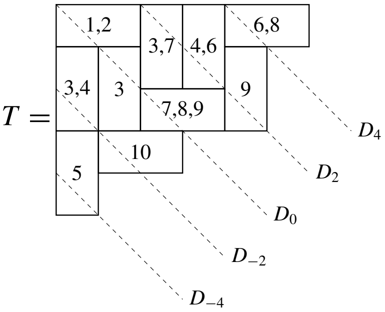
(1) Restricting to the minimum entry in each domino yields a domino tableau.

(2) If F_1 and F_2 are dominoes of the same type on neighboring diagonals and F_2 is weakly southeast of F_1, then
   - $\max(F_1) \leq \min(F_2)$ if F_1 is located on D_{2k} and F_2 is on $D_{2(k+1)}$,
   - $\max(F_1) < \min(F_2)$ if F_1 is located on $D_{2(k+1)}$ and F_2 is on D_{2k}.

Another way to state the first condition is that the minimum entries in the dominoes weakly increase from left to right along rows and strictly increase down columns.

For T a set-valued domino tableau, we again let $|T|$ denote the total number of positive integers in the filling. We can also define a diagonal reading word to be the sequence obtained by reading northwest to southeast along each diagonal D_{2k}, starting with the bottom diagonal, and separating entries on distinct diagonals with a slash. We use braces to identify the elements of a set that contains more than one element.

Example 5.5. A set-valued domino tableau of shape $\lambda = (6, 5, 5, 3, 1)$ is shown below:



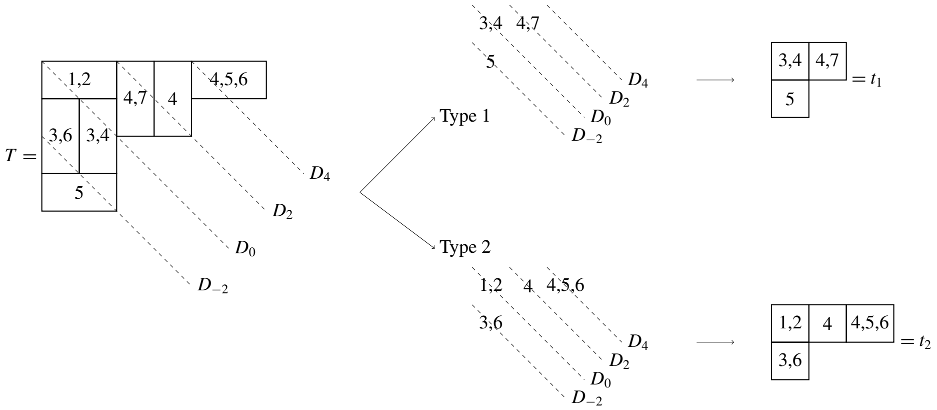
Notice, for example, that the entry $\{3, 4\}$ appears to the left of the entry $\{3\}$. This is acceptable because the corresponding dominoes have different types, and we therefore must only check that the minimum entries are weakly increasing across rows and strictly increasing down columns. Its diagonal reading word is $5 \,/\, \{3, 4\}, 10 \,/\, \{1, 2\}, 3, \{7, 8, 9\} \,/\, \{3, 7\}, \{4, 6\}, 9 \,/\, \{6, 8\}$ and $|T| = 17$.

We may now prove our main result of this section.

Theorem 5.6. *Let* λ *be a pavable partition with* 2*-quotient* (μ, ν)*. There is a bijection between the set of set-valued domino tableaux of shape* λ *and the set of pairs of semistandard set-valued tableaux of shape* (μ, ν).

Proof. To prove this theorem, we generalize the maps Γ and Γ^{-1} from the proof of Theorem 4.1. We denote our generalized bijection by Γ^* and describe how it maps a set-valued domino tableau of shape λ to a pair of semistandard set-valued tableaux (t_1, t_2) of shape (μ, ν), where (μ, ν) is the 2-quotient of λ.

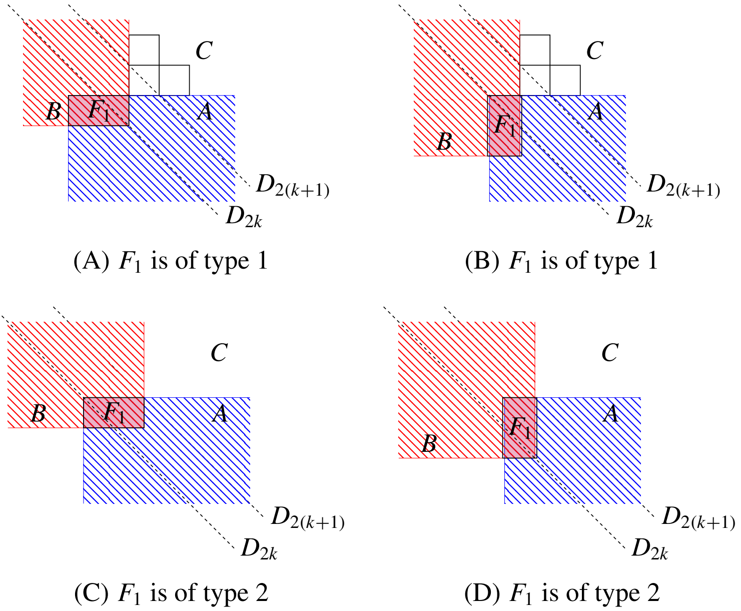
Let T be a set-valued domino tableau of shape λ and form the diagonal reading word for T. Let w_1 be the word obtained by restricting this diagonal reading word to the entries that come from type-1 dominoes and w_2 be the word obtained by restricting the diagonal reading word for T to entries coming from type-2 dominoes. We then construct t_1 to be the unique set-valued tableau with diagonal reading word w_1 and t_2 to be the unique set-valued tableau with diagonal reading word w_2. We illustrate this bijection below using an example:



Let T be a set-valued domino tableau as before, and we show that $\Gamma^*(T) = (t_1, t_2)$ is a pair of semistandard set-valued tableaux. We first show that the sets in t_1 and t_2 weakly increase along rows. Suppose box b_1 lies directly left of box b_2 in t_i. Now let F_1 and F_2 be the dominoes of T such that Γ^* sends the entries of F_1 to b_1 and those of F_2 to b_2. Note then that F_2 lies on the diagonal to the right of that of F_1. We will show that F_2 is weakly southeast of F_1.

Since taking the smallest entry in each domino of T gives a domino tableau and Γ^* restricts to Γ on domino tableaux, we know that $\min(b_1) \le \min(b_2)$, and so $\min(F_1) \le \min(F_2)$. Consider Figure 1. The first two images show the case where F_1 is type-1 and the next two show the analogous situation in the case where F_1 is type-2. Since T is a set-valued domino tableau and $\min(F_1) \le \min(F_2)$, F_2 cannot intersect region B. Also, the image shows that F_2 cannot lie completely in region C since it must be the same type as F_1. Thus F_2 must intersect region A and so is weakly southeast of F_1. Since we know F_2 lies on the diagonal to the right of the diagonal of F_1, this implies $\max(F_1) \le \min(F_2)$. Hence $\max(b_1) \le \min(b_2)$, as desired.

We next show that the entries of t_1 and t_2 strictly increase down columns. Suppose box b_1 lies directly above box b_2 in t_i and let F_1 and F_2 be as before. Note that



(A) F_1 is of type 1

(B) F_1 is of type 1

(C) F_1 is of type 2

(D) F_1 is of type 2

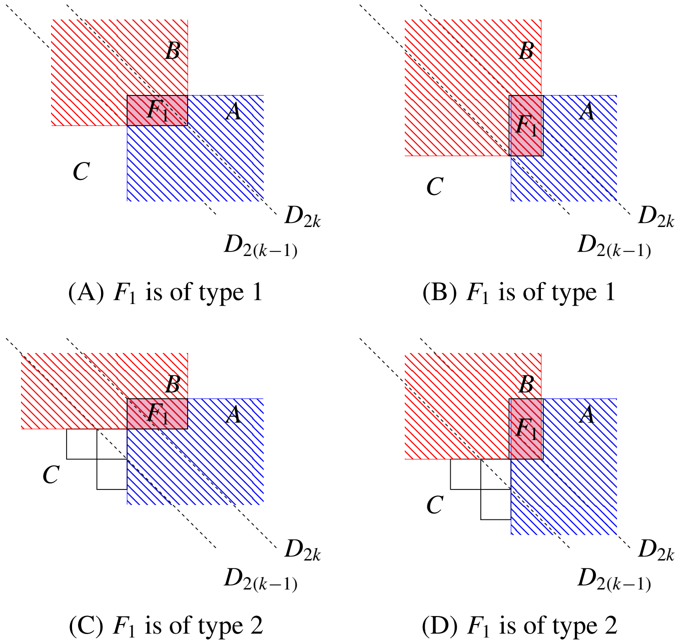
Figure 1. F_2 is southeast of F_1 when b_1 is left of b_2 in t_i.

F_2 lies on the diagonal directly left of that of F_1. Since taking the smallest entry in each domino of T gives a domino tableau and Γ^* restricts to Γ on domino tableaux, we know that $\min(b_1) < \min(b_2)$ and so $\min(F_1) < \min(F_2)$. Consider Figure 2. Since T is a set-valued domino tableau and $\min(F_1) < \min(F_2)$, F_2 cannot intersect the region B. We can also see that F_2 cannot be completely contained in region C, and hence F_2 intersects region A and is weakly southeast of F_1. Since F_2 lies on the diagonal to the left of that of F_1, then $\max(F_1) < \min(F_2)$. We conclude that $\max(b_1) < \min(b_2)$, as desired.

Note that we know that (t_1, t_2) has shape (μ, ν) by restricting Γ^* to the set of domino tableaux of shape λ. We conclude that Γ^* sends a set-valued domino tableau of shape λ to a pair of semistandard set-valued tableaux of shape (μ, ν), where (μ, ν) is the 2-quotient of λ.

We will now describe Γ^{*-1}, the inverse map of Γ^*. Let (t_1', t_2') be the semistandard tableaux obtained from (t_1, t_2) by taking only the smallest entry in each box. We may then apply Γ^{-1} to (t_1', t_2') to obtain a domino tableau T'. We then define $\Gamma^{*-1}(t_1, t_2)$ to be the set-valued domino tableau obtained from T' by reuniting each entry in T' with the rest of the subset that was with that entry in (t_1, t_2). We describe this precisely below.

Let (t_1, t_2) be a pair of semistandard set-valued tableaux of shape (μ, ν). Similarly to the description of Γ^{-1}, we recursively construct the set-valued domino tableau of shape λ associated to (t_1, t_2). At any step, we have a pair of set-valued



(A) F_1 is of type 1

(B) F_1 is of type 1

(C) F_1 is of type 2

(D) F_1 is of type 2

Figure 2. F_2 is southeast of F_1 when b_1 is over b_2 in t_i.

tableaux $(t_1^{(i)}, t_2^{(i)})$ of shape $(\mu^{(i)}, \nu^{(i)})$ and the associated set-valued domino tableau $T^{(i)}$ of shape $\lambda^{(i)}$. We start the algorithm with $\mu^{(0)} = \nu^{(0)} = \lambda^{(0)} = \varnothing$. The algorithm stops when $(t_1^{(s)}, t_2^{(s)}) = (t_1, t_2)$. Then we have that the set-valued domino tableau associated to (t_1, t_2) is $T^{(s)}$.

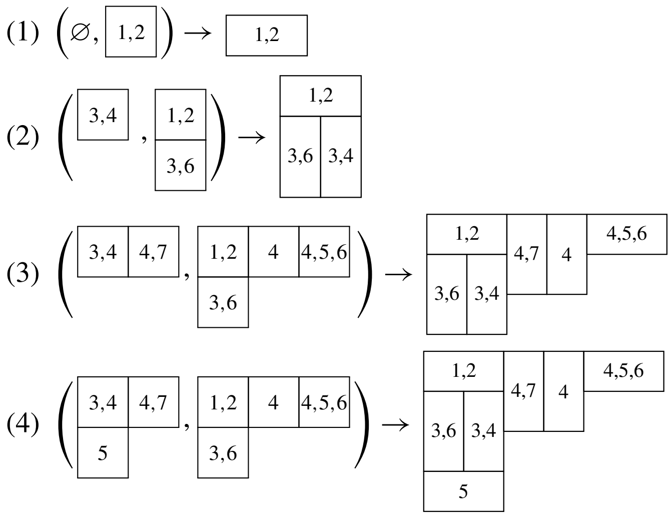
We now describe the i-th step of the algorithm. Let u_i be the smallest value appearing as the minimum entry in a box in (t_1, t_2) that does not appear as the minimum entry in a box in $(t_1^{(i-1)}, t_2^{(i-1)})$. We build $(t_1^{(i)}, t_2^{(i)})$ of shape $(\mu^{(i)}, \nu^{(i)})$ by adding to $(t_1^{(i-1)}, t_2^{(i-1)})$ all cells of (t_1, t_2) filled by a set with minimum u_i, while preserving their original position.

To construct the domino tableau $T^{(i)}$ of shape $\lambda^{(i)}$, we use the following procedure for each diagonal, starting with the bottom diagonal: For any cell in t_1 (resp. t_2) filled with a set with minimum u_i on diagonal D_k, we add to $T^{(i-1)}$ a type-1 domino (resp. a type-2 domino) filled with that set on the corresponding diagonal D_{2k}. We then get the associated domino tableau $T^{(i)}$ of shape $\lambda^{(i)}$.

Below is an example of Γ^{*-1} applied to the pair of set-valued tableaux:

$$(t_1, t_2) = \left(\begin{array}{|c|c|} \hline 3,4 & 4,7 \\ \hline 5 & \multicolumn{1}{c}{} \\ \cline{1-1} \end{array}\, , \begin{array}{|c|c|c|} \hline 1,2 & 4 & 4,5,6 \\ \hline 3,6 & \multicolumn{2}{c}{} \\ \cline{1-1} \end{array} \right).$$

Notice that we recover the tableau T from the previous example:



We need to see that Γ^{*-1} gives a set-valued domino tableau. Since Γ^{-1} and Γ^{*-1} coincide on pairs of semistandard tableaux and give a domino tableau, we see that the minimum entries of the dominoes increase weakly along rows and strictly down columns. Also, $T = \Gamma^{*-1}(t_1, t_2)$ has shape λ, since $\Gamma^{-1}(t_1', t_2')$ does.

Suppose F_1 and F_2 are dominoes of the same type of $T = \Gamma^{*-1}(t_1, t_2)$, that F_1 is on diagonal D_{2k} for some k, and that F_2 is on diagonal $D_{2(k+1)}$ and is weakly southeast of F_1. We must show that $\max(F_1) \leq \min(F_2)$.

Let b_1 and b_2 be the boxes of t_i such that Γ^{*-1} sends the entries of b_1 to F_1 and the entries of b_2 to F_2. Then b_1 lies on diagonal D_k of t_i and b_2 lies on diagonal D_{k+1}. Then b_2 is either weakly southeast of b_1 or is weakly northwest of b_1, as shown in the first image of Figure 3. However, since Γ^{*-1} restricts to Γ^{-1} on semistandard tableaux and gives a domino tableau, we know that $\min(F_1) \leq \min(F_2)$. It follows that $\min(b_1) \leq \min(b_2)$, and so b_2 must be weakly southeast of b_1. Thus $\max(b_1) \leq \min(b_2)$, which implies that $\max(F_1) \leq \min(F_2)$.

Lastly, suppose that F_1 and F_2 are dominoes of the same type of $T = \Gamma^{*-1}(t_1, t_2)$, that F_1 is on diagonal $D_{2(k+1)}$ for some k, and that F_2 is on diagonal D_{2k} and is weakly southeast of F_1. We must show that $\max(F_1) < \min(F_2)$.

Let b_1 and b_2 be as before, so b_1 lies on diagonal D_{k+1} of t_i and b_2 lies on diagonal D_k. From the second image of Figure 3, we see that either b_2 is weakly southeast of b_1 or b_2 is weakly northwest of b_1. For the same reason as in the previous argument, we know that $\min(F_1) < \min(F_2)$, so $\min(b_1) < \min(b_2)$. This means that b_2 must lie weakly southeast of b_1, and so $\max(b_1) < \min(b_2)$. We then have that $\max(F_1) < \min(F_2)$, as desired.

We conclude that Γ^{*-1} sends a pair of semistandard set-valued tableaux of shape (μ, ν) to a set-valued domino tableau of shape λ. It is clear that Γ^{*-1} and Γ^{*} are indeed inverses as they are governed by Γ^{-1} and Γ. □

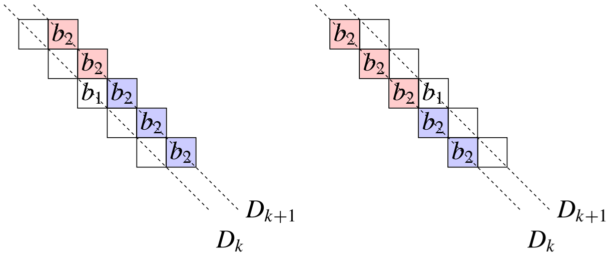


Figure 3. Possible relative positions of b_1 and b_2: case 1 is shown in red, where $\max(b_2) \leq \min(b_1)$ and case 2 in blue, where $\max(b_1) \leq \min(b_2)$.

The translation into the language of symmetric functions gives us the following theorem. Note that for pavable partition λ, the number of dominoes in a paving of λ is given by $\frac{1}{2}|\lambda|$.

Corollary 5.7. *Let λ be a partition with 2-quotient (μ, ν). Then*

$$G_\mu G_\nu = \sum_T (-1)^{|T|-\frac{1}{2}|\lambda|} x^T,$$

where we sum over all set-valued domino tableaux of shape λ.

Proof. Consider a term in the product $G_\mu G_\nu$. This monomial corresponds to a pair of semistandard set-valued tableaux: t_1 of shape μ and t_2 of shape ν. The pair (t_1, t_2) corresponds to some set-valued domino tableau T of shape λ by Theorem 5.6. It is clear from the previous bijection that $x^{t_1}x^{t_2} = x^T$.

We now examine the sign of $x^{t_1}x^{t_2}$ in $G_\mu G_\nu$. We see that it appears with sign

$$(-1)^{|t_1|-|\mu|}(-1)^{|t_2|-|\nu|} = (-1)^{|t_1|+|t_2|-(|\mu|+|\nu|)} = (-1)^{|T|-\frac{1}{2}|\lambda|}.$$

This gives the desired result. □

6. *Q*-Schur functions and shifted domino tableaux

6A. ***Q*-Schur functions.** Let λ be a partition. Define up(λ) to be the boxes of λ that lie on a diagonal weakly northeast of D_0 and down(λ) to be the boxes of λ that lie on a diagonal strictly southwest of D_0.

Let $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_k)$ be such that $\lambda_k \geqslant k$. We may form a *shifted Young tableau* of shape λ by filling the boxes of down(λ) with the symbol X and filling the boxes of up(λ) with primed and unprimed positive integers with linear order $1' < 1 < 2' < 2 < \cdots$ such that

- rows and columns are weakly increasing,
- there is at most one occurrence of i' in any row and
- there is at most one occurrence of i in any column.

For example, both tableaux below are shifted Young tableaux:

$1'$	1	$2'$	
X	2	4	

2	$3'$	3	3
X	$3'$	4	
X	X	6	

We may associate a monomial to a shifted Young tableau T by defining

$$x^T = x_1^{\beta_1(T)} x_2^{\beta_2(T)} \cdots ,$$

where $\beta_i(T)$ is the number of occurrences of i and i' in T. The tableau above on the right corresponds to monomial $x_2 x_3^4 x_4 x_6$.

The *Q-Schur function* indexed by partition $\lambda = (\lambda_1, \dots, \lambda_k)$ with $\lambda_k \geq k$, denoted by Q_λ, is then defined to be the weighted generating function over all shifted Young tableaux of shape λ:

$$Q_\lambda = \sum_{\mathrm{sh}(T)=\lambda} x^T.$$

The Q-Schur functions were introduced by I. Schur [1911] in relation to the projective representations of the symmetric and alternating groups. They have since been widely studied, for example by B. Sagan [1987] and J. Stembridge [1989].

Below are a few terms of $Q_{(3,3,3)}$ and the corresponding shifted tableaux. Note that each term shown appears with multiplicity in the full $Q_{(3,3,3)}$. For example, $x_1^3 x_2^2 x_3$ appears with coefficient 8 since each element on the diagonal of the tableau shown on the left may be primed or unprimed:

$$Q_{(3,3,3)} = x_1^3 x_2^2 x_3 + x_1^3 x_2^2 x_4 + x_1^3 x_2^2 x_5 + x_1 x_2^2 x_3^2 x_4 + x_1 x_2^3 x_3^2 + \cdots$$

$1'$	1	1
X	$2'$	2
X	X	$3'$

$1'$	1	1
X	2	2
X	X	4

$1'$	1	1
X	$2'$	2
X	X	$5'$

$1'$	$2'$	$3'$
X	$2'$	$3'$
X	X	4

$1'$	$2'$	2
X	$2'$	$3'$
X	X	3

6B. ***Shifted domino tableaux.*** We next define the notion of a shifted domino tableau, which was first introduced in [Chemli 2016].

Definition 6.1 [Chemli 2016]. Let λ be a pavable partition with 2-quotient ($\mu = (\mu_1, \dots, \mu_s)$, $\nu = (\nu_1, \dots, \nu_t)$) and fixed paving. This paving is a *shifted paving* if

- $\mu_s \geq s$ and $\nu_t \geq t$ and
- there is no vertical domino d on D_0 such that the dominoes directly left of d and adjacent to d are all strictly below D_0.

If such a paving of λ exists, we call λ a *shifted pavable partition.*

For a shifted pavable partition λ with fixed shifted paving, define up(λ) to be the dominoes of λ that lie on a diagonal weakly northeast of D_0 and down(λ) to be the dominoes of λ that lie on a diagonal strictly southwest of D_0.

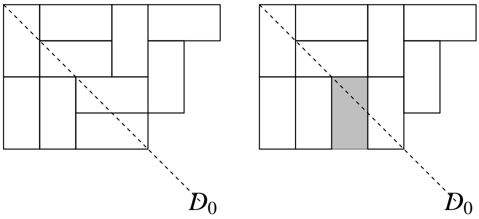
Definition 6.2 [Chemli 2016]. Given a shifted pavable partition λ with fixed shifted paving, a *shifted domino tableau* is a filling of the dominoes of $\mathrm{down}(\lambda)$ with X and the dominoes of $\mathrm{up}(\lambda)$ with primed and unprimed integers with linear order $1' < 1 < 2' < 2 < \cdots$ such that

- rows and columns are weakly increasing,
- there is at most one occurrence of i in any column, and
- there is at most one occurrence of i' in any row.

For T a shifted domino tableau, let $\mathrm{up}(T)$ be the dominoes of T that lie on a diagonal weakly northeast of D_0 along with the filling of these dominoes. We consider two shifted domino tableaux T and T' of shape λ to be *equivalent* if $\mathrm{up}(T) = \mathrm{up}(T')$. Let the *set of shifted domino tableaux* refer to the set up to equivalence.

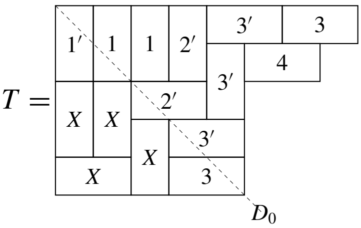
We define the *diagonal reading word* of a shifted domino tableau to be the sequence obtained from reading northwest to southeast along each diagonal D_{2k} with $k \geq 0$, starting with D_0 and separating entries on distinct diagonals with a slash. Note that equivalent domino tableaux have equal diagonal reading words.

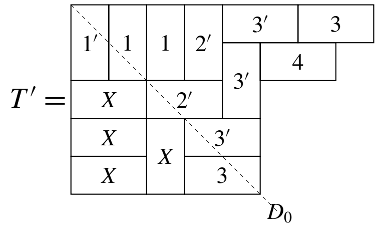
Example 6.3. Let $\lambda = (6, 5, 5, 4)$, which is a pavable partition with 2-quotient $((2, 2), (3, 3))$. Clearly, this 2-quotient respects the conditions of Definition 6.1. Below are two different pavings of λ. The paving on the left is a shifted paving, and the one on the right is not. The problematic domino is shaded in the second paving:

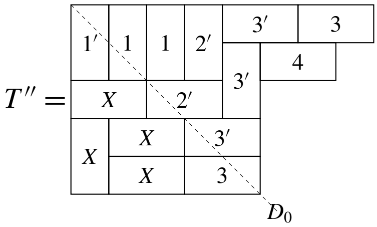


Now consider $\lambda = (5, 5, 4, 3, 3, 2)$, which is a pavable partition with 2-quotient $((3, 1, 1), (2, 2, 2))$. Since neither of the partitions of the 2-quotient respects the first condition of Definition 6.1, λ is not a shifted pavable partition.

The following tableaux T, T', and T'' are equivalent shifted domino tableaux and are thus considered equal in the set of shifted domino tableaux. The diagonal reading word for each is $1', 1, 2', 3', 3 \,/\, 1, 2', 3' \,/\, 3', 4 \,/\, 3$:





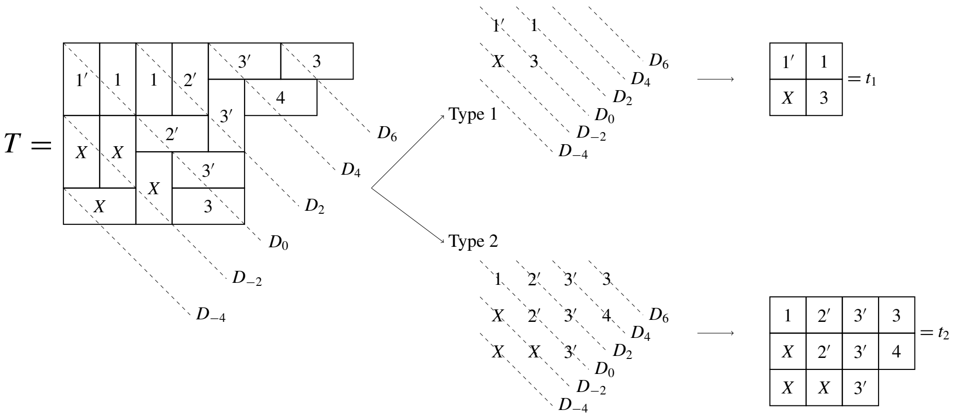


6C. *Bijection between shifted domino tableaux and shifted Young tableaux.* Chemli proves the following bijection.

Theorem 6.4 [Chemli 2016, Theorem 3.1]. *Let λ be a shifted pavable partition with 2-quotient (μ, ν). The set of shifted domino tableaux of shape λ is in bijection with the set of pairs (t_1, t_2) of shifted Young tableaux of shape (μ, ν).*

The bijection is a slight modification of the map Γ from Theorem 4.1. We call this modified map Γ_s and it goes from the set of shifted domino tableaux of shape λ to the set of pairs of shifted Young tableaux of shape (μ, ν).

The modification of Γ to obtain Γ_s is quite simple. We apply Γ to a shifted domino tableau as if it were a domino tableau, without considering if the dominoes are filled with X's or with integers. For example, let us apply Γ_s to the tableau T of Example 6.3:



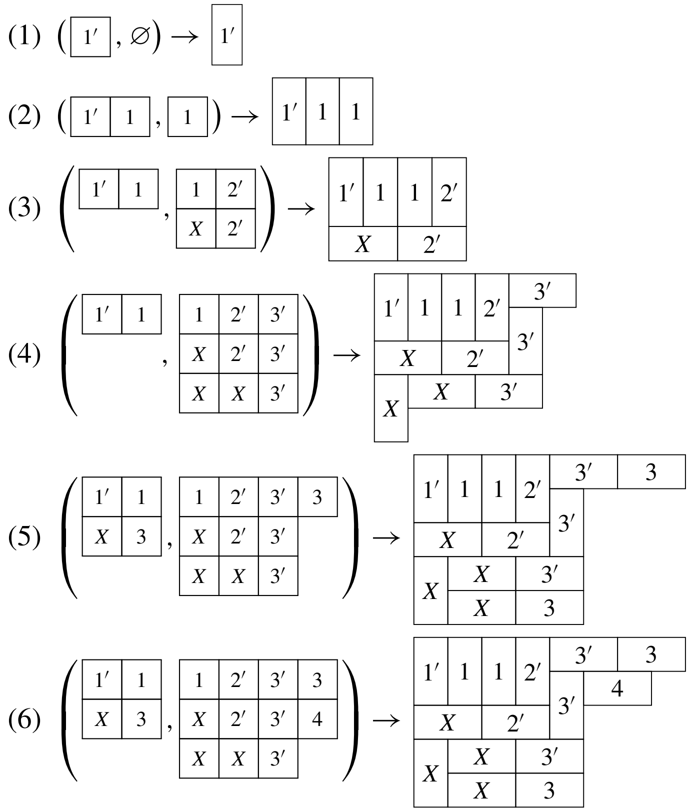
We see that Γ_s^{-1} is also very similar to Γ^{-1}. It takes as input a pair of shifted Young tableaux and outputs a shifted domino tableau such that $\Gamma_s^{-1}(\Gamma_s(T))$ is equivalent to T. However, we need to describe how the algorithm deals with the cells containing X.

Suppose we apply Γ_s^{-1} to a pair of shifted Young tableaux (t_1, t_2). At step i, we have the pair $(t_1^{(i-1)}, t_2^{(i-1)})$ of shifted Young tableaux, and u_i is the smallest integer (with respect to the relation $1' < 1 < 2' < 2 < \cdots$) appearing in (t_1, t_2) that doesn't appear in $(t_1^{(i-1)}, t_2^{(i-1)})$. If a cell of t_1 (resp. t_2) containing u_i has cells filled with X to its left, then all those cells are also added into $t_1^{(i)}$ (resp. $t_2^{(i)}$), while preserving their original positions. This ensures that at every step of the procedure, the constructed tableaux $(t_1^{(i)}, t_2^{(i)})$ are shifted Young tableaux.

For example, let's apply Γ_s^{-1} to the pair

$$(t_1, t_2) = \left(\begin{array}{|c|c|} \hline 1' & 1 \\ \hline X & 3 \\ \hline \end{array} \; , \; \begin{array}{|c|c|c|c|} \hline 1 & 2' & 3' & 3 \\ \hline X & 2' & 3' & 4 \\ \hline X & X & 3' \\ \cline{1-3} \end{array} \right)$$

of shifted Young tableaux:



Notice that we do not recover exactly the tableau T that we started with. Instead, we obtain a shifted domino tableau equivalent to T.

As a corollary, we then have the following.

Corollary 6.5 [Chemli 2016, Theorem 3.2]. *Let λ be a shifted pavable partition with 2-quotient (μ, ν). One has*

$$Q_\mu Q_\nu = \sum_{\mathrm{sh}(T)=\lambda} x^T,$$

where we sum over the set of shifted domino tableaux of shape λ.

Proof. Each term in the product $Q_\mu Q_\nu$ is represented by a pair of shifted tableaux of shape (μ, ν). Theorem 6.4 says that the set of these pairs are in bijection with the set of shifted domino tableaux of shape λ, and the shifted domino tableau associated to a pair of shifted tableaux (t_1, t_2) has the same multiset of entries as (t_1, t_2). □

7. Shifted K-theoretic generalizations

7A. ***K-theoretic Q-Schur functions.*** There is a natural K-theoretic analogue of the Q-Schur functions, introduced in [Ikeda and Naruse 2013; Graham and Kreiman 2015], called the *K-theoretic Q-Schur function* and denoted by GQ_λ. In fact, Ikeda and Naruse introduced a more general *K-theoretic factorial Q-Schur function*, but it will suffice for us to consider the restricted generality. As a natural K-theoretic analogue, the GQ_λ are related to the K-theory of the maximal isotropic Grassmannian of symplectic type.

Using the same linear order on subsets of $\{1' < 1 < 2' < 2 < \cdots\}$, where $A \le B$ if $\max(A) \le \max(B)$, we have the following definition.

Definition 7.1 [Ikeda and Naruse 2013]. Let $\lambda = (\lambda_1, \dots, \lambda_k)$ be a partition with $\lambda_k \ge k$. A *shifted set-valued Young tableau* of shape λ is a filling of the cells of $\mathrm{down}(\lambda)$ with X and the cells of $\mathrm{up}(\lambda)$ with finite, nonempty sets of positive integers such that

- entries weakly increase across rows and down columns,
- there is at most one occurrence of i in any column and
- there is at most one occurrence of i' in any row.

Let $\mathrm{up}(T)$ denote the cells of T weakly above D_0 along with their filling.

We associate to each shifted set-valued tableau T a monomial x^T,

$$x^T = x_1^{\beta_1(T)} x_2^{\beta_2(T)} \cdots,$$

where again $\beta_i(T)$ is the number of occurrences of i and i' in T. We also let $|T| = |\mathrm{up}(T)|$ denote the number of primed and unprimed integers in T. We define the diagonal reading word of shifted set-valued tableau T in the natural way. It is easy to see that the diagonal reading word uniquely defines the shifted set-valued Young tableau.

Definition 7.2 [Ikeda and Naruse 2013]. Let $\lambda = (\lambda_1, \dots, \lambda_k)$ be a partition with $\lambda_k \geq k$. The *K-theoretic Q-Schur function GQ_λ* is

$$GQ_\lambda = \sum_{\mathrm{sh}(T)=\lambda} (-1)^{|\mathrm{up}(T)|-|\mathrm{up}(\lambda)|} x^T,$$

where we sum over all shifted set-valued tableaux of shape λ and $|\mathrm{up}(\lambda)|$ denotes the number of boxes in $\mathrm{up}(\lambda)$.

Example 7.3. We may compute some monomials of

$$GQ_{(2,2)} = 4x_1^2x_2 - 2x_1^3x_2 - 4x_1x_2^2x_3 - x_1x_2^2x_3^2x_5 \pm \cdots$$

using the tableaux shown below:

$$\begin{array}{|c|c|}\hline 1' & 1 \\ \hline X & 2 \\ \hline\end{array}\quad \begin{array}{|c|c|}\hline 1 & 1 \\ \hline X & 2 \\ \hline\end{array}\quad \begin{array}{|c|c|}\hline 1' & 1 \\ \hline X & 2' \\ \hline\end{array}\quad \begin{array}{|c|c|}\hline 1 & 1 \\ \hline X & 2' \\ \hline\end{array}\quad \begin{array}{|c|c|}\hline 1',1 & 1 \\ \hline X & 2 \\ \hline\end{array}$$

$$\begin{array}{|c|c|}\hline 1',1 & 1 \\ \hline X & 2' \\ \hline\end{array}\quad \begin{array}{|c|c|}\hline 1,2 & 2 \\ \hline X & 3 \\ \hline\end{array}\quad \begin{array}{|c|c|}\hline 1',2' & 2 \\ \hline X & 3 \\ \hline\end{array}\quad \begin{array}{|c|c|}\hline 1' & 2' \\ \hline X & 2,3 \\ \hline\end{array}\quad \begin{array}{|c|c|}\hline 1' & 2' \\ \hline X & 2',3 \\ \hline\end{array}\quad \begin{array}{|c|c|}\hline 1',2' & 2,3' \\ \hline X & 3',5 \\ \hline\end{array}$$

Note that we have not listed all tableaux with monomials $x_1x_2^2x_3$ and $x_1x_2^2x_3^2x_5$. We see that the diagonal reading word for the rightmost tableau above is

$$\{1', 2'\}, \{3', 5\} \,/\, \{2, 3'\}.$$

Since shifted Young tableaux are shifted set-valued tableaux, we see that the lowest-degree terms of GQ_λ make up Q_λ.

7B. *Shifted set-valued domino tableaux.*

Definition 7.4. Let λ be a shifted pavable partition. A *shifted set-valued domino tableau* is a filling of the dominoes of $\mathrm{down}(\lambda)$ with X and the dominoes of $\mathrm{up}(\lambda)$ with finite, nonempty sets of primed and unprimed integers with linear order $\{1' < 1 < 2' < 2 < \cdots\}$ such that:

(1) Restricting to the minimum entry in each domino yields a shifted domino tableau.

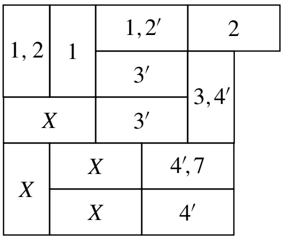
(2) If F_1 and F_2 are dominoes of the same type on neighboring diagonals and F_2 is weakly southeast of F_1, then $\max(F_1) \leq \min(F_2)$, and

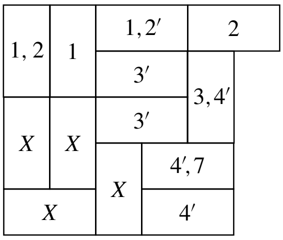
- $\max(F_1) < \min(F_2)$ if F_1 is located on D_{2k}, F_2 is on $D_{2(k+1)}$, and $\max(F_1)$ is primed, and
- $\max(F_1) < \min(F_2)$ if F_1 is located on $D_{2(k+1)}$, F_2 is on D_{2k}, and $\max(F_1)$ is unprimed.

For T a shifted set-valued domino tableau, let $\mathrm{up}(T)$ be the dominoes of T that lie on a diagonal weakly northeast of D_0 along with the filling of these dominoes. We consider two shifted set-valued tableaux T and T' to be *equivalent* if $\mathrm{up}(T) = \mathrm{up}(T')$. The *set of shifted set-valued domino tableaux* refers to the set up to equivalence.

We define the diagonal reading word in the natural way and again note that equivalent shifted set-valued domino tableaux have equal diagonal reading words.

Example 7.5. Below are two equivalent shifted set-valued domino tableaux of shape $(6, 5, 5, 5, 3)$ with diagonal reading word $\{1, 2\}, 1, 3', \{4', 7\}, 4' \,/\, \{1, 2'\}, 3', \{3, 4'\} \,/\, 2$:





We may now state the main result of this section.

Theorem 7.6. *Let λ be a shifted pavable partition with 2-quotient (μ, ν). The set of shifted set-valued domino tableaux of shape λ is in bijection with the set of pairs (t_1, t_2) of shifted set-valued tableaux of shape (μ, ν).*

Proof. We define set-valued versions of Γ_s and Γ_s^{-1} called Γ_s^* and Γ_s^{*-1} analogously to the definitions of Γ^* and Γ^{*-1} from Γ and Γ^{-1}. See Example 7.7 for an illustration.

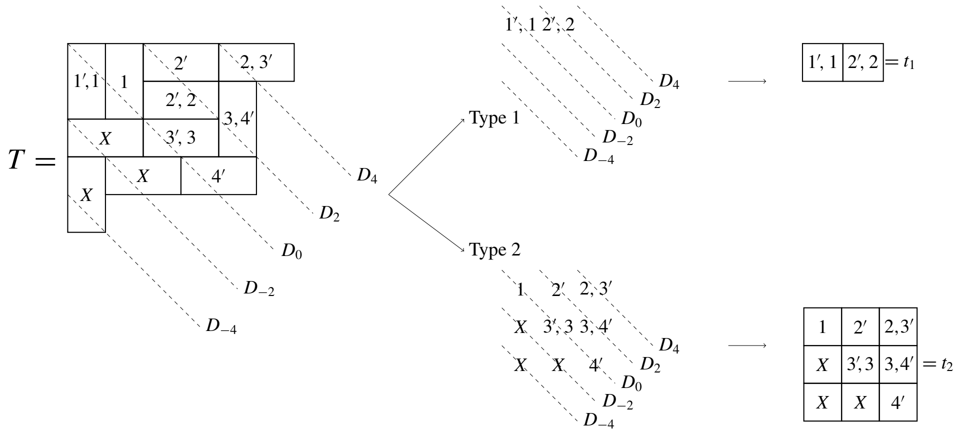
Let T be a shifted set-valued domino tableau. We show that $\Gamma_s^*(T) = (t_1, t_2)$ is a pair of shifted set-valued Young tableaux. We can use the same argument as in the proof of Theorem 5.6 to show that t_1 and t_2 are weakly increasing in rows and columns. To see there is at most one occurrence of i' in a row, suppose b_1 lies directly left of b_2 in t_i and let F_1 and F_2 be the dominoes of T such that Γ_s^* sends the entries of F_1 to b_1 and those of F_2 to b_2. Using the argument from the proof of Theorem 5.6, we know F_2 is weakly southeast of F_1. Hence if $\max(b_1)$ is primed, $\max(F_1) < \min(F_2)$, and so $\max(b_1) < \min(b_2)$. We can similarly argue that there is at most one occurrence of i in any column of t_i.

Now let (t_1, t_2) be a pair of shifted set-valued Young tableaux. We show that $T = \Gamma_s^{*-1}(t_1, t_2)$ is a shifted set-valued domino tableau. Using an argument analogous to that in the proof of Theorem 5.6, we see that restricting to the minimum entry in each domino yields a shifted domino tableau and that $\max(F_1) \leq \min(F_2)$ when F_2 is weakly southeast of F_1.

Suppose F_2 is weakly southeast of F_1, F_1 is on a diagonal D_{2k} and F_2 is on $D_{2(k+1)}$, and $\max(F_1)$ is primed. Let b_1 and b_2 be the boxes of t_i such that Γ_s^{*-1} sends the entries of b_1 to F_1 and the entries of b_2 to F_2. We have shown in the proof of Theorem 5.6 that b_2 must be weakly southeast of b_1. Since $\max(F_1)$ is primed, $\max(b_1)$ is primed. Then $\max(b_1) < \min(b_2)$ because t_i is a shifted set-valued tableau, and so $\max(F_1) < \min(F_2)$. We can similarly show that $\max(F_1) < \min(F_2)$ if F_1 is on $D_{2(k+1)}$, F_2 is on D_{2k}, and $\max(F_1)$ is unprimed.

It is clear that Γ_s^* and Γ_s^{*-1} are inverses because Γ_s and Γ_s^{-1} are. □

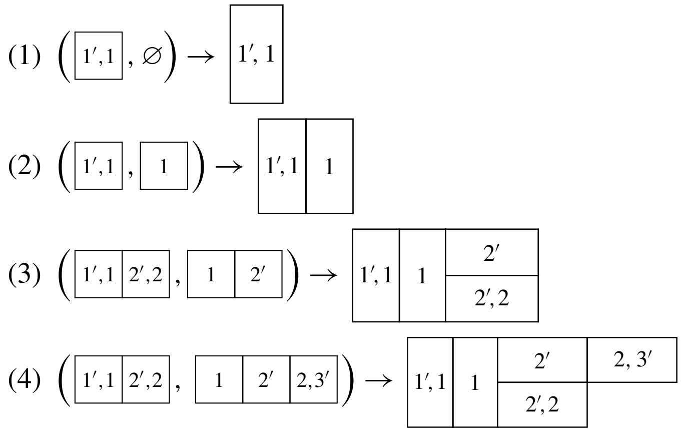
Example 7.7. Let T be the following shifted set-valued domino tableau. We apply Γ_s^* to the tableau T:

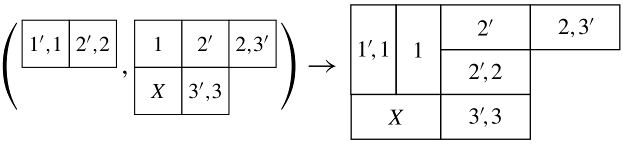


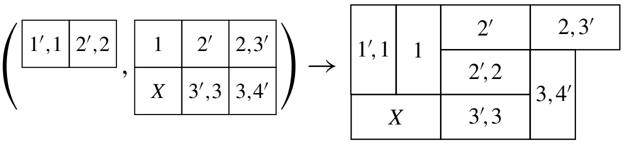
If we take the pair of shifted set-valued tableaux

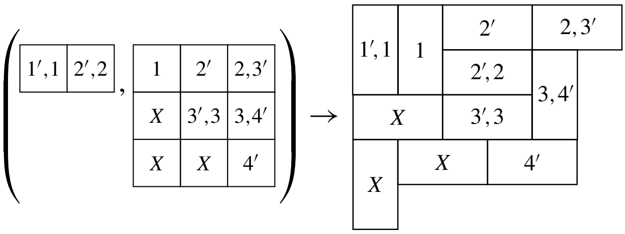
$$(t_1, t_2) = \left(\begin{array}{|c|c|}\hline 1',1 & 2',2 \\ \hline \end{array}\ ,\ \begin{array}{|c|c|c|}\hline 1 & 2' & 2,3' \\ \hline X & 3',3 & 3,4' \\ \hline X & X & 4' \\ \hline \end{array} \right),$$

we will see that we reconstruct T exactly the same way as in Section 6C, with integer set entries instead of integers.



(5)


(6)


(7)


Corollary 7.8. *Let λ be a shifted pavable partition with* 2*-quotient* (μ, ν). *Then*

$$GQ_\mu GQ_\nu = \sum_{\mathrm{sh}(T)=\lambda} (-1)^{|\mathrm{up}(T)|-|\mathrm{up}(\lambda)|} x^T,$$

where we sum over all shifted set-valued domino tableaux of shape λ, $|\mathrm{up}(T)|$ *denotes the number of positive integers in* $\mathrm{up}(T)$ *and* $|\mathrm{up}(\lambda)|$ *denotes the number of dominoes in* $\mathrm{up}(\lambda)$.

Proof. Consider a term in the product $GQ_\mu GQ_\nu$. This monomial corresponds to a pair of shifted set-valued tableaux: t_1 of shape μ and t_2 of shape ν. The pair (t_1, t_2) corresponds to some shifted set-valued domino tableau T of shape λ by Theorem 7.6. It is then clear from the previous bijection that $x^{t_1}x^{t_2} = x^T$.

We now examine the sign of $x^{t_1}x^{t_2}$ in $GQ_\mu GQ_\nu$. We see that it appears with sign

$$\begin{aligned}(-1)^{|\mathrm{up}(t_1)|-|\mathrm{up}(\mu)|}(-1)^{|\mathrm{up}(t_2)|-|\mathrm{up}(\nu)|} &= (-1)^{|\mathrm{up}(t_1)|+|\mathrm{up}(t_2)|-(|\mathrm{up}(\mu)|+|\mathrm{up}(\nu)|)} \\ &= (-1)^{|\mathrm{up}(T)|-|\mathrm{up}(\lambda)|}.\end{aligned}$$

This gives the desired result. □

Acknowledgements

This paper is the result of an undergraduate research project, where Maas-Gariépy was funded by a CRM-ISM Summer Undergraduate Scholarship. The authors are grateful to Hugh Thomas for his support and to the LaCIM community. Patrias received support from NSERC, CRM-ISM, and the Canada Research Chairs Program. We thank the anonymous reviewer for comments that improved the exposition.

References

[Buch 2002] A. S. Buch, "A Littlewood–Richardson rule for the K-theory of Grassmannians", *Acta Math.* **189**:1 (2002), 37–78. MR Zbl

[Carré and Leclerc 1995] C. Carré and B. Leclerc, "Splitting the square of a Schur function into its symmetric and antisymmetric parts", *J. Algebraic Combin.* **4**:3 (1995), 201–231. MR Zbl

[Chemli 2016] Z. Chemli, "Shifted domino tableaux", preprint, 2016. arXiv

[Fomin and Kirillov 1996] S. Fomin and A. N. Kirillov, "The Yang–Baxter equation, symmetric functions, and Schubert polynomials", *Discrete Math.* **153**:1-3 (1996), 123–143. MR Zbl

[Graham and Kreiman 2015] W. Graham and V. Kreiman, "Excited Young diagrams, equivariant K-theory, and Schubert varieties", *Trans. Amer. Math. Soc.* **367**:9 (2015), 6597–6645. MR Zbl

[Ikeda and Naruse 2013] T. Ikeda and H. Naruse, "K-theoretic analogues of factorial Schur P- and Q-functions", *Adv. Math.* **243** (2013), 22–66. MR Zbl

[Lascoux and Schützenberger 1982] A. Lascoux and M.-P. Schützenberger, "Structure de Hopf de l'anneau de cohomologie et de l'anneau de Grothendieck d'une variété de drapeaux", *C. R. Acad. Sci. Paris Sér. I Math.* **295**:11 (1982), 629–633. MR Zbl

[Sagan 1987] B. E. Sagan, "Shifted tableaux, Schur Q-functions, and a conjecture of R. Stanley", *J. Combin. Theory Ser. A* **45**:1 (1987), 62–103. MR Zbl

[Schur 1911] J. Schur, "Über die Darstellung der symmetrischen und der alternierenden Gruppe durch gebrochene lineare Substitutionen", *J. Reine Angew. Math.* **139** (1911), 155–250. MR Zbl

[Stanley 1999] R. P. Stanley, *Enumerative combinatorics, Vol. 2*, Cambridge Studies in Advanced Mathematics **62**, Cambridge University Press, 1999. MR Zbl

[Stanton and White 1985] D. W. Stanton and D. E. White, "A Schensted algorithm for rim hook tableaux", *J. Combin. Theory Ser. A* **40**:2 (1985), 211–247. MR Zbl

[Stembridge 1989] J. R. Stembridge, "Shifted tableaux and the projective representations of symmetric groups", *Adv. Math.* **74**:1 (1989), 87–134. MR Zbl

Received: 2018-07-22 Revised: 2020-02-23 Accepted: 2020-08-10

maas-gariepy.florence@courrier.uqam.ca
Laboratoire de Combinatoire et d'Informatique Mathématique,
Université du Québec à Montréal, Montréal, QC, Canada

patriasr@lacim.ca
Laboratoire de Combinatoire et d'Informatique Mathématique,
Université du Québec à Montréal, Montréal, QC, Canada

INVOLVE 13:5 (2020)
https://doi.org/10.2140/involve.2020.13.747

The first digit of the discriminant of Eisenstein polynomials as an invariant of totally ramified extensions of p-adic fields

Chad Awtrey, Alexander Gaura, Sebastian Pauli,
Sandi Rudzinski, Ariel Uy and Scott Zinzer

(Communicated by Kenneth S. Berenhaut)

Let K be an extension of the p-adic numbers with uniformizer π. Let φ and ψ be Eisenstein polynomials over K of degree n that generate isomorphic extensions. We show that if the cardinality of the residue class field of K divides $n(n-1)$, then $v(\operatorname{disc}(\varphi)-\operatorname{disc}(\psi)) > v(\operatorname{disc}(\varphi))$. This makes the first (nonzero) digit of the π-adic expansion of $\operatorname{disc}(\varphi)$ an invariant of the extension generated by φ. Furthermore we find that noncyclic extensions of degree p of the field of p-adic numbers are uniquely determined by this invariant.

1. Introduction

This paper is concerned with the classification of extensions of p-adic fields using invariants. We introduce a new invariant and show how it is related to other invariants of the extension.

For a field extension of finite degree, the discriminant of an integral basis yields invariants of the extension. A change of integral basis results in the multiplication of the discriminant by the square of a unit in the base ring, namely, by the determinant of the transformation matrix. Because the only units in the ring of integers $\mathbb{Z}$ are ± 1 and because -1 is not a square, the discriminant is an invariant of an extension of the rational numbers $\mathbb{Q}$.

When we consider extensions of the p-adic field $\mathbb{Q}_p$, the ring of p-adic numbers $\mathbb{Z}_p$ takes the place of $\mathbb{Z}$. As $\mathbb{Z}_p$ contains infinitely many units, the discriminant of an integral basis is not an invariant of an extension of $\mathbb{Q}_p$. However, since changing integral bases changes the discriminant by the square of a unit, the discriminant modulo the square of units does produce an invariant of the extension; see [Cassels 1986, Chapter 7, Section 6] for details. Nevertheless, commonly only the valuation

MSC2020: 11S05, 11S15.
Keywords: p-adic field, Eisenstein polynomial, discriminant, invariant.

of the discriminant is used as an invariant of extensions of $\mathbb{Q}_p$, since it yields information about the ramified part of the extension.

It is natural to ask whether this invariant can be refined by considering not the complete discriminant but possibly some of the digits of its p-adic expansion. To this end we restrict our investigation to discriminants of integral bases of a certain form; namely, power integral bases, which are bases of the form $1, \alpha, \alpha^2, \dots, \alpha^{n-1}$, where n is the degree of the extension. In the case of totally ramified extensions, the discriminants of such a power integral basis is the same as the discriminant of the Eisenstein polynomial with root α.

We find that the first digit of the p-adic expansion of the discriminant of generating Eisenstein polynomials is an invariant for most extensions. This allows for a finer classification of totally ramified extensions. Most of our results also hold over extensions of $\mathbb{Q}_p$; in this case we consider the π-adic expansion of the discriminant, where π is a uniformizing element of the base field. As an application we show that for extensions of $\mathbb{Q}_p$ of degree p this new invariant already yields the Galois group.

Notation. We introduce the notation that we use throughout the paper and recall some results about totally ramified extensions.

For a prime number p we denote by $\mathbb{Q}_p$ the completion of the rational numbers $\mathbb{Q}$ with respect to the p-adic exponential valuation v_p, by $\mathbb{Z}_p$ its valuation ring, and by $\overline{\mathbb{Q}}_p$ a fixed algebraic closure of $\mathbb{Q}_p$. By K we denote a finite extension of $\mathbb{Q}_p$ with valuation ring $\mathcal{O}_K$, uniformizing element π, and maximal ideal (π). The exponential valuation v_π on K is normalized such that $v_\pi(\pi) = 1$. The extensions of v_p and v_π to $\overline{\mathbb{Q}}_p$ are also denoted by v_p and v_π. We write $\underline{K}$ for the residue class field $\mathcal{O}_K/(\pi)$ of K and for $\beta \in \mathcal{O}_K$ we set $\underline{\beta} = \beta + (\pi) \in \underline{K}$.

The extensions that we consider are totally ramified extensions. In the following the ground field K will either be the p-adic field $\mathbb{Q}_p$ or a finite extension of $\mathbb{Q}_p$. Totally ramified extensions L/K can be generated by an Eisenstein polynomial $\varphi \in \mathcal{O}_K[x]$. For a root α of φ we have $\mathcal{O}_L = \mathcal{O}_K(\alpha)$. The discriminant $\operatorname{disc}(\varphi)$ of the generating polynomial φ is equal to the discriminant of the integral basis $1, \alpha, \dots, \alpha^{n-1}$ of $\mathcal{O}_L/\mathcal{O}_K$, where n is the degree of φ and thus the degree of L/K. The root α is a uniformizing element of $\mathcal{O}_L$. We write v_α for the valuation on $\mathcal{O}_L$ that is normalized such that $v_\alpha(\alpha) = 1$. For $\gamma \in K$ we have $v_\alpha(\gamma) = n \cdot v_\pi(\gamma)$. The minimal polynomial of any uniformizing element of $\mathcal{O}_L$, that is, of any $\beta \in \mathcal{O}_L$ with $v_\alpha(\beta) = 1$, is also a uniformizing element of L and its minimal polynomial is an Eisenstein polynomial.

2. Main results

In this section we present our main results. The proofs can be found in the following sections. We start by investigating how the number of common digits of the discriminants is related to the distance between the two polynomials.

For a given $n \in \mathbb{N}$ there are finitely many totally ramified extensions of K of degree n. Marc Krasner [1966] presented a formula for the number of the extensions with given degree and valuation of the discriminant. To this end he introduced a distance function d on the set of Eisenstein polynomials of a fixed degree and valuation of the discriminant; also see [Pauli and Roblot 2001, Section 4].

Definition 2.1. Let $\varphi, \psi \in \mathcal{O}_K[x]$ be Eisenstein of degree n with $v_\pi(\operatorname{disc}(\varphi)) = v_\pi(\operatorname{disc}(\psi))$. Let $\alpha \in \overline{\mathbb{Q}}_p$ be a root of φ and define the distance between φ and ψ as $d(\varphi, \psi) = v_\pi(\psi(\alpha))$.

This distance is symmetric with respect to the input polynomials and is independent of the choice of root. It satisfies the ultrametric inequality, and clearly $d(\varphi, \psi) = \infty$ if and only if $\varphi = \psi$.

We find that when we choose two polynomials φ and ψ that generate isomorphic extensions such that they are close enough, that is, when $d(\varphi, \psi)$ is large enough, we can ensure that arbitrarily many digits of the π-adic expansion of their discriminants coincide.

Theorem 2.2. *Let $\varphi, \psi \in \mathcal{O}_K[x]$ be Eisenstein of degree n with $v_\pi(\operatorname{disc}(\varphi)) = v_\pi(\operatorname{disc}(\psi)) = n + j - 1$. If $d(\varphi, \psi) > (n + 2j)/n$, then*

$$v_\pi(\operatorname{disc}(\varphi) - \operatorname{disc}(\psi)) \geq \Bigl(1 - \frac{1}{n}\Bigr) v_\pi(\operatorname{disc}(\varphi)) + d(\varphi, \psi).$$

From now on we concentrate on the first digit of the π-adic expansion of the discriminants of polynomials. We use this notation:

Definition 2.3. For $\beta \in \mathcal{O}_K$ we set $\operatorname{tc}(\beta) = \underline{(\beta/\pi^{v_\pi(\beta)})} \in \underline{K}$. This is the first nonzero digit (or the trailing coefficient) of the π-adic expansion of β as an element of $\underline{K}$.

Clearly we have:

Lemma 2.4. *For $\alpha, \beta \in K$ we have* $\operatorname{tc}(\alpha \cdot \beta) = \operatorname{tc}(\alpha) \cdot \operatorname{tc}(\beta)$.

We find that this first π-adic digit of the valuation of the generating Eisenstein polynomial is an invariant in many cases.

Theorem 2.5. *Let $\varphi, \psi \in \mathcal{O}_K[x]$ be Eisenstein and of degree n such that $K[x]/(\varphi) \cong K[x]/(\psi)$. If $(\#\underline{K} - 1) \mid (n(n-1))$ then* $\operatorname{tc}(\operatorname{disc}(\varphi)) = \operatorname{tc}(\operatorname{disc}(\psi))$.

In these cases, the trailing coefficient of the discriminant is independent of the generating Eisenstein polynomial and thus is an invariant of power integral bases of a totally ramified extension. Thus, if $(\#\underline{K} - 1) \mid (n(n-1))$, then the trailing coefficient of the discriminant is an invariant of the extension. Some classes of extensions always have the same invariant.

Proposition 2.6. *Let p be odd and $\varphi \in \mathbb{Q}_p[x]$ be Eisenstein of degree p^m such that $\mathbb{Q}_p[x]/(\varphi)$ is cyclic. Then*

$$\mathrm{tc}(\mathrm{disc}(\varphi)) = \begin{cases} 1 & \textit{if } m \textit{ is even},\\ -1 & \textit{if } m \textit{ is odd and } p \equiv 1 \bmod 4,\\ 1 & \textit{if } m \textit{ is odd and } p \equiv 3 \bmod 4. \end{cases}$$

We end with examples of information that can be obtained from this invariant. Extensions of $\mathbb{Q}_p$ of degree p have been described in detail; see [Amano 1971; Jones and Roberts 2006]. We show that they can also be classified using the discriminant and its trailing coefficient, that is, the first nonzero coefficient of its p-adic expansion.

Theorem 2.7. *Let φ be an Eisenstein polynomial of degree p in $\mathbb{Q}_p$ such that* $\mathrm{Gal}(\varphi) \not\cong C_p$ *and* $v_p(\mathrm{disc}(\varphi)) \neq 2p-1$. *Then the isomorphism class of the extension generated by φ is uniquely determined by* $v_p(\mathrm{disc}(\varphi))$ *and* $\mathrm{tc}(\mathrm{disc}(\varphi))$.

The Galois group of an Eisenstein polynomial φ of degree p over $\mathbb{Q}_p$ can be determined from the valuation of its discriminant $\mathrm{disc}(\varphi)$ and its trailing coefficient $\mathrm{tc}(\mathrm{disc}(\varphi))$.

Corollary 2.8. *Let p be an odd prime, and let φ be an Eisenstein polynomial of degree p over $\mathbb{Q}_p$. Let $v = v_p(\mathrm{disc}(\varphi))$, $j = v-p+1$, and $\underline{\gamma} = (\underline{-1})^{(p-1)/2}\,\mathrm{tc}(\mathrm{disc}(\varphi))$. Then*

$$\mathrm{Gal}(\varphi) \cong \begin{cases} C_p \rtimes C_{p-1} & \textit{if } v = 2p-1,\\ C_p & \textit{if } v = 2p-2 \textit{ and } \underline{\gamma} = \underline{p-1},\\ C_p \rtimes C_d & \textit{otherwise}, \end{cases}$$

where

$$d = \frac{p-1}{\gcd((p-1)/m,\, j)},$$

with m being the order of $\underline{aj}$ in $\mathbb{F}_p^\times$ for $\underline{a} = \underline{\gamma} \cdot (\underline{-1})^{j+1}\underline{j}^{-1}$.

3. Proof of Theorem 2.2

We recall some of the results from [Krasner 1966] (see also [Pauli and Roblot 2001]) about the distance function d. Assume that $\varphi(x) = \sum_{i=0}^{n} \varphi_i x^i \in \mathcal{O}_K[x]$ and $\psi(x) = \sum_{i=0}^{n} \psi_i x^i \in \mathcal{O}_K[x]$ are Eisenstein polynomials whose discriminants have the same valuation, say $v_\pi(\mathrm{disc}(\varphi)) = v_\pi(\mathrm{disc}(\psi)) = n + j - 1$. Denote by $\alpha = \alpha_{(1)}, \dots, \alpha_{(n)}$ the roots of φ in $\overline{\mathbb{Q}}_p$.

If $d(\varphi, \psi) > (n+2j)/n$ then there is a root $\beta \in \overline{\mathbb{Q}}_p$ of ψ such that $v_\pi(\alpha - \beta) > v_\pi(\alpha - \alpha_{(i)})$ for $2 \le i \le n$. In this case Krasner's lemma implies that $K(\alpha) = K(\beta)$. So the assumption of Theorem 2.2 implies that the extensions generated by φ and ψ are isomorphic.

Furthermore, again assuming that $v_\pi(\alpha-\beta) > v_\pi(\alpha-\alpha_{(i)})$ for some $2 \le i \le n$, we obtain another expression for the distance of two polynomials:

$$d(\varphi,\psi) = \sum_{i=1}^{n} \min\{v_\pi(\alpha-\beta), v_\pi(\alpha-\alpha_{(i)})\} = v_\pi(\alpha-\beta) + \sum_{i=2}^{n} v_\pi(\alpha-\alpha_{(i)}).$$

So we can write the valuation of $\alpha-\beta$ in terms of $d(\varphi,\psi)$ and $\operatorname{disc}(\varphi)$:

$$\begin{aligned} v_\pi(\alpha-\beta) &= d(\varphi,\psi) - \sum_{i=2}^{n} v_\pi(\alpha-\alpha_{(i)}) \\ &= d(\varphi,\psi) - v_\pi\left(\prod_{i=2}^{n}(\alpha-\alpha_{(i)})\right) = d(\varphi,\psi) - \frac{1}{n} v_\pi(\operatorname{disc}(\varphi)). \qquad (1) \end{aligned}$$

We now are ready to prove Theorem 2.2.

Proof of Theorem 2.2. Let $\alpha = \alpha_{(1)}, \dots, \alpha_{(n)}$ and β be as above. As $K(\alpha) = K(\beta)$, there is $\gamma \in K(\alpha)$ with $v_\alpha(\gamma) = 0$ such that $\beta = \alpha + \gamma\alpha^m$, where $m = v_\alpha(\alpha-\beta)$. We order the roots $\beta_{(1)}, \dots, \beta_{(n)}$ of ψ such that $\beta_{(1)} = \beta = \alpha + \gamma\alpha^m$, $\beta_{(2)} = \alpha_{(2)} + \gamma_{(2)}\alpha_{(2)}^m$ and so on. For the discriminant of ψ we get

$$\begin{aligned} \operatorname{disc}(\psi) &= \prod_{i<j}(\beta_{(i)}-\beta_{(j)})^2 = \prod_{i<j}\big((\alpha_{(i)}+\gamma_{(i)}\alpha_{(i)}^m) - (\alpha_{(j)}+\gamma_{(j)}\alpha_{(j)}^m)\big)^2 \\ &= \prod_{i<j}\big((\alpha_{(i)}-\alpha_{(j)}) + (\gamma_{(i)}\alpha_{(i)}^m - \gamma_{(j)}\alpha_{(j)}^m)\big)^2 \\ &= \prod_{i<j}(\alpha_{(i)}-\alpha_{(j)})^2 \prod_{i<j}\left(1 - \sum_{k=1}^{m}(\gamma_{(j)}\alpha_{(i)}^{m-k}\alpha_{(j)}^{k-1}) - \alpha_{(i)}^m \frac{\gamma_{(j)}-\gamma_{(i)}}{\alpha_{(j)}-\alpha_{(i)}}\right)^2 \\ &= \operatorname{disc}(\varphi) \prod_{i<j}\left(1 - \gamma_{(j)}\sum_{k=1}^{m}(\alpha_{(i)}^{m-k}\alpha_{(j)}^{k-1}) - \alpha_{(i)}^m \frac{\gamma_{(j)}-\gamma_{(i)}}{\alpha_{(j)}-\alpha_{(i)}}\right)^2. \end{aligned}$$

Let

$$C_{ij} = \left(1 - \gamma_{(j)}\sum_{k=1}^{m}(\alpha_{(i)}^{m-k}\alpha_{(j)}^{k-1}) - \alpha_{(i)}^m \frac{\gamma_{(j)}-\gamma_{(i)}}{\alpha_{(j)}-\alpha_{(i)}}\right)^2.$$

We have $v_\alpha(C_{ij}-1) \ge m-1$. With $\operatorname{disc}(\psi) = \operatorname{disc}(\varphi)\big(\prod_{i<j} C_{ij}\big)$ we get

$$\begin{aligned} v_\pi(\operatorname{disc}(\varphi) - \operatorname{disc}(\psi)) &= v_\pi\left(\operatorname{disc}(\varphi) - \operatorname{disc}(\varphi)\left(\prod_{i<j} C_{ij}\right)\right) \\ &= v_\pi(\operatorname{disc}(\varphi)) + v_\pi\left(1 - \left(\prod_{i<j} C_{ij}\right)\right) \\ &\ge v_\pi(\operatorname{disc}(\varphi)) + \frac{m-1}{n}. \end{aligned}$$

With $m = v_\alpha(\alpha-\beta) = n\cdot v_\pi(\alpha-\beta)$ and (1) we obtain

$$\begin{aligned} v_\pi(\operatorname{disc}(\varphi)-\operatorname{disc}(\psi)) &\geq v_\pi(\operatorname{disc}(\varphi)) + \frac{n\cdot v_\pi(\alpha-\beta)-1}{n} \\ &= v_\pi(\operatorname{disc}(\varphi)) + d(\varphi,\psi) - \frac{1}{n}(v_\pi(\operatorname{disc}(\varphi))+1). \qquad \square \end{aligned}$$

4. Proof of Theorem 2.5

Proof of Theorem 2.5. Let α be a root of φ. Since φ and ψ generate isomorphic extensions, there exists $\beta \in K(\alpha)$ such that $\psi(\beta) = 0$. So $\beta = \sum_{k=0}^{n-1} b_k\alpha^k$ for some $b_k \in \mathcal{O}_K$. As $v_\alpha(\beta) = v_\alpha(\alpha) = 1$, we have $v(b_1) = 0$. Let $\alpha_{(1)}, \alpha_{(2)}, \dots, \alpha_{(n)}$ be the conjugates of α and let $\sigma_1, \sigma_2, \dots, \sigma_n$ be the isomorphisms such that $\sigma_i(\alpha) = \alpha_{(i)}$. Let $\beta_{(1)}, \beta_{(2)}, \dots, \beta_{(n)}$ be the roots of ψ, defined by

$$\beta_{(i)} = \sigma_i(\beta) = \sigma_i\left(\sum_{k=0}^{n-1} b_k\alpha^k\right) = \sum_{k=0}^{n-1} \sigma_i(b_k)\sigma_i(\alpha)^k = \sum_{k=0}^{n-1} b_k\alpha_{(i)}^k.$$

We now compute the discriminant of ψ, with the goal of writing it in terms of the discriminant of φ:

$$\begin{aligned} \operatorname{disc}(\psi) &= \prod_{i<j}(\beta_{(i)}-\beta_{(j)})^2 = \prod_{i<j}\left(\sum_{k=1}^{n-1} b_k(\alpha_{(i)}^k - \alpha_{(j)}^k)\right)^2 \\ &= \prod_{i<j}(\alpha_{(i)}-\alpha_{(j)})^2 \prod_{i<j}\left[\sum_{k=1}^{n-1}\left(b_k\sum_{\ell=0}^{k-1}\alpha_{(i)}^{(k-1-\ell)}\alpha_{(j)}^{\ell}\right)\right]^2 \\ &= \operatorname{disc}(\varphi)\cdot\prod_{i<j}\left[\sum_{k=1}^{n-1}\left(b_k\sum_{\ell=0}^{k-1}\alpha_{(i)}^{(k-1-\ell)}\alpha_{(j)}^{\ell}\right)\right]^2. \end{aligned}$$

We write $\operatorname{disc}(\varphi) - \operatorname{disc}(\psi) = \operatorname{disc}(\varphi)\cdot(1-\gamma)$, where

$$\gamma = \prod_{i<j}\left[\sum_{k=1}^{n-1}\left(b_k\sum_{\ell=0}^{k-1}\alpha_{(i)}^{(k-1-\ell)}\alpha_{(j)}^{\ell}\right)\right]^2.$$

Note that γ is a symmetric polynomial in $\alpha_{(1)}, \dots, \alpha_{(n)}$. Let $e_1, \dots, e_n$ denote the elementary symmetric polynomials in $\alpha_{(1)}, \dots, \alpha_{(n)}$. By the fundamental theorem of symmetric polynomials, there is a polynomial $\gamma^* \in \mathcal{O}_K[x_1, \dots, x_n]$ such that $\gamma = \gamma^*(e_1, \dots, e_n)$. If we expand γ^*, all terms consist of sums of products of $\alpha_{(1)}, \dots, \alpha_{(n)}$ except for the constant term, $b_1^{n(n-1)}$. So

$$\gamma^* = m_{(1)}e_1 + m_{(2)}e_2 + \cdots + m_{(n)}e_n + m_{(n+1)}e_1^2 + m_{(n+2)}e_1e_2 + \cdots + b_1^{n(n-1)}$$

for some $m_{(1)}, \dots, m_{(n(n-1)-1)} \in \mathcal{O}_K$. Note that $e_1, \dots, e_n$ are exactly the coefficients of φ. Since the coefficients of φ have π-adic valuation greater than or equal to 1, π divides all of its coefficients. This implies $\underline{\gamma} = \underline{b}_1^{n(n-1)}$.

The next step is to show $\underline{b}_1^{n(n-1)} = \underline{1}$. Since both φ and ψ are Eisenstein and generate the same extension, $v_\alpha(\beta) = 1$. So

$$1 = v_\alpha(\beta) = v_\alpha\left(\sum_{k=0}^{n-1} b_k \alpha^k\right) = \min_{0 \le k \le n-1} \{v_\alpha(b_k \alpha^k)\}.$$

Equality holds because each b_k has α-adic valuation 0 or a positive multiple of n, so each term has a different valuation. For all k, we have $v_\alpha(b_k\alpha^k) \ge 1$. For $k \ge 1$, this is obviously true. For $k = 0$, we have $v_\alpha(b_0)$ is equal to 0 or a positive multiple of n. As $v_\alpha(b_0)$ must be at least 1, the lowest multiple of n it can be is n. For $k \ne 1$, we now have that $v_\alpha(b_k\alpha^k) \ge 2$. Thus $v_\alpha(b_1\alpha) = 1$, implying $v_\alpha(b_1) = 0$, i.e., $b_1 \notin (\pi)$. By the generalization of Fermat's little theorem, $b_1 \notin (\pi)$ implies $\underline{b}_1^{\#\underline{K}-1} = 1$. By assumption $(\#\underline{K} - 1) \mid n(n-1)$, so this implies $\underline{b}_1^{n(n-1)} = 1$.

We have

$$\underline{(1-\gamma)} = \underline{(1 - b_1^{n(n-1)})} = \underline{(1-1)} = \underline{0}.$$

Because $\pi \mid (1-\gamma)$ we can write $1 - \gamma = \pi \cdot c$ for some c. So

$$\operatorname{disc}(\varphi) - \operatorname{disc}(\psi) = \operatorname{disc}(\varphi) \cdot (1-\gamma) = \operatorname{disc}(\varphi) \cdot \pi \cdot c.$$

Therefore,

$$v_\pi(\operatorname{disc}(\varphi) - \operatorname{disc}(\psi)) \ge v_\pi(\operatorname{disc}(\varphi)) + 1$$

and thus $\operatorname{tc}(\operatorname{disc}(\varphi)) = \operatorname{tc}(\operatorname{disc}(\psi))$. □

5. Proof of Proposition 2.6

Because of Lemma 2.4 we only need to consider the trailing coefficients of the differences of roots in our considerations.

In the proof of the proposition we will use information obtained from the ramification polygon of the polynomial $\varphi \in \mathbb{Q}_p[x]$ under consideration. We recall some of the information that can be obtained from the ramification polygon; see [Greve and Pauli 2012] for details.

Let α be a root of φ. The ramification polynomial of an Eisenstein polynomial φ of degree n

$$\rho(x) = \frac{\varphi(\alpha x + \alpha)}{\alpha^n} = \sum_{i=0}^{n} \rho_i x^i \in \mathbb{Q}_p(\alpha)[x]$$

has the roots $(\alpha^* - \alpha)/\alpha$, where α^* is a root of φ. The Newton polygon of the ramification polynomial is called the ramification polygon of φ, it is independent of the choice of α and an invariant of the extension $\mathbb{Q}_p(\alpha) \cong \mathbb{Q}_p[x]/(\varphi)$. Its breaks can only be at powers of p and the negatives of the slopes λ of the segments are the valuations of the differences of the roots of φ. The length of the segment (in the

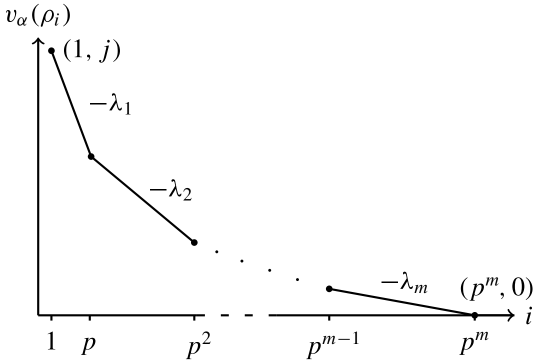


Figure 1. Ramification polygon of an Eisenstein polynomial $\varphi \in \mathbb{Q}_p[x]$ of degree p^m with $\operatorname{disc}(\varphi) = p^m + j - 1$ generating a normal cyclic extension, where $\rho(x) = \varphi(\alpha x + \alpha)/(\alpha^{p^m}) = \sum_{i=0}^{p^m} \rho_i x^i \in \mathbb{Q}_p(\alpha)[x]$ with α a root of φ is the ramification polynomial of φ.

direction of the horizontal axis) with slope λ is the number of roots α^* of φ such that $v_\alpha(\alpha - \alpha^*) = \lambda + 1$.

When $\mathbb{Q}_p[x]/(\varphi)$ is normal the differences of the roots of φ are in $\mathbb{Q}_p[x]/(\varphi)$ and thus the slopes λ_i of the segments of the ramification polygon are integral. Furthermore the roots of the residual polynomials of the segment of slope λ are of the form

$$\underline{\gamma} = \underline{\left(\frac{\alpha^* - \alpha}{\alpha^{\lambda+1}}\right)},$$

and thus $\operatorname{tc}(\alpha^* - \alpha) = \underline{\gamma}$. As $\mathbb{Q}_p[x]/(\varphi)$ is normal we have $\underline{\gamma} \in \mathbb{F}_p$. The normality also implies that the lengths of the segments of the ramification polygon are $p^i - p^{i-1}$ for $1 \le i \le m$ and that all elements of $\mathbb{F}_p^\times$ are roots of the residual polynomial of each segment.

Proof of Proposition 2.6. The polynomial φ is an Eisenstein polynomial of degree p^m. As the extension generated by φ is normal, the slopes of the segments are integers. It follows from the symmetry of the roots and the normality of the extension that the breaks in the polygon are exactly at $1, p, p^2, \ldots, p^{m-1}$; see Figure 1. So the lengths of the segments with finite slope are $p - 1, p^2 - p, \ldots, p^m - p^{m-1}$.

As $\underline{\mathbb{Q}}_p^\times = \mathbb{F}_p^\times$ only has $p - 1$ distinct elements it follows from the symmetry of the roots that we get for any root α of φ that

$$\prod_{v_p(\alpha - \alpha^*) = \lambda_1} \operatorname{tc}(\alpha - \alpha^*) = \prod_{\underline{\gamma} \in \mathbb{F}_p^\times} \underline{\gamma} = \underline{-1}, \tag{2}$$

where the α^* are roots of φ and $-\lambda_1$ is the slope of the first segment of the ramification polygon from Figure 1. Similarly, again because of normality and symmetry, taking into consideration the lengths of the segment for the second to the n-th segments with slopes λ_i of length $p^i - p^{i-1}$ using that p is odd, we get

$$\prod_{v_p(\alpha-\alpha^*)=\lambda_i} \operatorname{tc}(\alpha-\alpha^*) = \Bigl(\prod_{\underline{\gamma}\in\mathbb{F}_p^\times} \underline{\gamma}\Bigr)^{p^{i-1}} = (\underline{-1})^{p^{i-1}} = \underline{-1}. \tag{3}$$

Equations (2) and (3) yield

$$\begin{aligned}\operatorname{tc}(\operatorname{disc}(\varphi)) &= (\underline{-1})^{\frac{p^m(p^m-1)}{2}} \prod_{\alpha^*\neq\alpha} \operatorname{tc}(\alpha-\alpha^*)\\ &= (\underline{-1})^{\frac{p^m(p^m-1)}{2}} \prod_{\alpha}\prod_{i=1}^{m}\prod_{v_p(\alpha-\alpha^*)=\lambda_i} \operatorname{tc}(\alpha-\alpha^*)\\ &= (\underline{-1})^{\frac{p^m(p^m-1)}{2}} ((\underline{-1})^m)^{p^m} = (\underline{-1})^{m+\frac{p^m(p^m-1)}{2}}.\end{aligned}$$

Recall that p is odd. When m is even we have $p^m \equiv 1 \bmod 4$ so $(p^m-1)/2 \equiv 0 \bmod 4$, which implies $\operatorname{tc}(\operatorname{disc}(\varphi)) = 1$. Similarly, if m is odd and $p \equiv 1 \bmod 4$ then $(p^m-1)/2 \equiv 0 \bmod 4$, so $\operatorname{tc}(\operatorname{disc}(\varphi)) = -1$, and if m is odd and $p \equiv 3 \bmod 4$ then $(p^m-1)/2 \equiv 1 \bmod 4$, so $\operatorname{tc}(\operatorname{disc}(\varphi)) = 1$. □

6. Proof of Theorem 2.7 and Corollary 2.8

Before we get to the proof of the theorem and corollary we give some auxiliary results. In the proofs below we use that $\operatorname{tc}(\operatorname{Res}(\varphi, \varphi')) = \operatorname{tc}(\operatorname{disc}(\varphi))(\underline{-1})^{n(n-1)/2}$, where $\operatorname{Res}(\varphi, \varphi')$ denotes the resultant of φ and φ' and the degree of φ is n.

To determine the trailing coefficient of the discriminant of Eisenstein polynomials φ of degree p over $\mathbb{Q}_p$ for p odd we distinguish the three cases presented in Table 1. The case $v_p\operatorname{disc}(\varphi) = 2p-2$ in which $\operatorname{Gal}\varphi \cong C_p$ is covered by Proposition 2.6.

Lemma 6.1. *Let p be an odd prime and $\varphi\in\mathbb{Q}_p[x]$ be of degree p, with $v_p(\operatorname{disc}(\varphi)) = 2p-1$. Then $\operatorname{tc}(\operatorname{disc}(\varphi)) = (\underline{-1})^{p(p-1)/2}$.*

Proof. If φ is a degree-p polynomial with $v_p(\operatorname{disc}(\varphi)) = 2p-1$, it must be of the form $\varphi(x) = x^p + p(1+ap)$ for $a \in \{1, \dots, p-1\}$ or generate an isomorphic extension to such a polynomial [Jones and Roberts 2006, Table 2.1]. However, using Theorem 2.5, we can reduce to the case where the polynomials are exactly of this form since we are only concerned with the trailing coefficient of the discriminant. We compute $\varphi'(x) = px^{p-1}$. Then 0 is a root of φ' with multiplicity $p-1$. To show $\operatorname{tc}(\operatorname{disc}(\varphi)) = (\underline{-1})^{p(p-1)/2}$ we show $\operatorname{tc}(\operatorname{Res}(\varphi, \varphi')) = \underline{1}$. We have

$$\operatorname{Res}(\varphi, \varphi') = p^p(p(1+ap))^{p-1} = p^{2p-1}(1+ap)^{p-1}.$$

Thus

$$\operatorname{tc}(\operatorname{Res}(\varphi, \varphi')) = \underline{(1+ap)}^{p-1} = \underline{1}^{p-1} = \underline{1}. \qquad □$$

$\varphi \in \mathbb{Q}_p[x]$	parameters	$v_p(\mathrm{disc}(\varphi))$	$\mathrm{Gal}(\varphi)$
$x^p + apx^j + p$	$1 \le a \le p-1$ $1 \le j \le p-1$ $(j, a) \ne (p-1, p-1)$	$p+j-1$	$C_p \rtimes C_d$
$x^p - px^{p-1} + p(1+ap)$	$0 \le a \le p-1$	$2p-2$	C_p
$x^p + p(1+ap)$	$0 \le a \le p-1$	$2p-1$	$C_p \rtimes C_{p-1}$

Table 1. Families of generating polynomials of extensions of degree p of $\mathbb{Q}_p$ for p odd with their Galois group. We have $d = (p-1)/\gcd((p-1)/m, j)$, where m is the order of $\underline{aj}$ in $\mathbb{F}_p^\times$ for $\underline{a} = \underline{\gamma} \cdot (\underline{-1})^{j+1} \underline{j}^{-1}$. See [Jones and Roberts 2006].

Lemma 6.2. *Let φ be an Eisenstein polynomial in $\mathbb{Q}_p$ of the form $x^p + apx^j + p$ for $a, j \in \{1, \dots, p-1\}$ and j and a not both equal to $p-1$. Then* $\mathrm{tc}(\mathrm{Res}(\varphi, \varphi')) = (\underline{-1})^{j+1}\underline{aj}$.

Proof. As $\varphi(x) = x^p + apx^j + p$ we have

$$\varphi'(x) = px^{p-1} + apjx^{j-1} = px^{j-1}(x^{p-j} + aj).$$

The polynomial φ' has 0 as a root with multiplicity $j-1$ and $p-j$ roots $r_0, \dots, r_{p-j-1}$ with $r_k^{p-j} = -aj$ for $0 \le k \le p-j-1$ of multiplicity 1.

Let $\xi \in \overline{\mathbb{Q}}_p$ be a primitive $(p-j)$-th root of unity and fix a nonzero root r_0 of φ'; then the other nonzero roots are of the form $r_k = \xi^k r_{p-j}$, where $k = 1, \dots, p-j-1$. Writing $\varphi(x) = x^j(x^{p-j} + ap) + p$ and evaluating φ at the roots of φ' we obtain

$$\mathrm{Res}(\varphi,') = p^p p^{j-1} k = 0^{p-j-1}[(\xi^k r_0)^j (r_0^{p-j} + ap) + p].$$

Hence

$$\begin{aligned}
\mathrm{tc}(\mathrm{Res}(\varphi, \varphi')) &= \prod_{k=0}^{p-j-1} [(\underline{\xi}^k \underline{r}_0)^j \cdot (\underline{r}_0^{p-j} + \underline{ap}) + \underline{a}] \\
&= \prod_{k=0}^{p-j-1} [\underline{\xi}^{kj} \cdot \underline{r}_0^{j+p-j}] = \underline{r}_0^{p(p-j)} \cdot \prod_{k=0}^{p-j-1} \underline{\xi}^{kj} \\
&= (\underline{r}_0)^{p-j} \cdot \underline{\xi}^{j \sum_0^{p-j-1} k} = (\underline{-aj}) \cdot \underline{\xi}^{j \frac{(p-j)(p-j-1)}{2}} \\
&= \underline{-aj} \cdot (\underline{\xi}^{\frac{p-j}{2}})^{j(p-j-1)} = (\underline{-1})^{j+1} \underline{aj}.
\end{aligned}$$

□

Proof of Theorem 2.7. Let $\varphi \in \mathbb{Q}_p[x]$ be Eisenstein of degree p such that $\mathrm{Gal}(\varphi) \ne C_p$ and $v_p(\mathrm{disc}(\varphi)) \ne 2p-1$. For each of these extensions, there is exactly one polynomial of the form $x^p + apx^j + p$ for $a, j \in \{1, \dots, p-1\}$ where $p + j - 1$

is the valuation of the discriminant and j and a are both not equal to $p-1$ [Jones and Roberts 2006, Proposition 2.3.1].

Thus there exists some $\psi(x) = x^p + apx^j + p$ (for fixed a and j) that generates an extension isomorphic to φ. By Theorem 2.5, $v_p(\mathrm{Res}(\varphi, \varphi')) = v_p(\mathrm{Res}(\psi, \psi'))$ and $\mathrm{tc}(\mathrm{Res}(\varphi, \varphi')) = \mathrm{tc}(\mathrm{Res}(\psi, \psi'))$. With Table 1 and Lemma 6.2 we get

$$j = v_p(\mathrm{Res}(\varphi, \varphi')) - p + 1 \quad \text{and} \quad a = \mathrm{tc}(\mathrm{Res}(\varphi, \varphi')) = (\underline{-1})^{j+1}\underline{j}^{-1}.$$

No two distinct $j \in \{1, \dots, p-1\}$ have the same multiplicative inverse modulo p. Also for a fixed j, no two distinct possible values of $\mathrm{tc}(\mathrm{Res}(\varphi, \varphi'))$ give the same value of a. Thus $v_p(\mathrm{Res}(\varphi, \varphi'))$ and $\mathrm{tc}(\mathrm{Res}(\varphi, \varphi'))$ uniquely determine ψ, and therefore the extension. □

Proof of Corollary 2.8. If $v_p(\mathrm{Res}(\varphi, \varphi')) = 2p-1$, then $\mathrm{Gal}(\varphi) = C_p \rtimes C_{p-1}$.

Suppose $v_p(\mathrm{Res}(\varphi, \varphi')) = 2p-2$ and $\mathrm{tc}(\mathrm{Res}(\varphi, \varphi')) = \underline{-1}$. Then φ is either in the first or second family in Table 1, since $v_p(\mathrm{Res}(\varphi, \varphi')) = 2p-2$. By Lemma 6.2 we have $\mathrm{tc}(\mathrm{Res}(\varphi, \varphi')) = (\underline{-1})^{j+1}\underline{aj}$. Since $j = p-1$,

$$\underline{-1} = \mathrm{tc}(\mathrm{Res}(\varphi, \varphi')) = (\underline{-1})^p \underline{a(p-1)} = \underline{a}.$$

Hence φ must be in the second row of the table, that is, $\mathrm{Gal}(\varphi) = C_p$.

Otherwise, compute a and j as in Theorem 2.7. In Table 1,

$$d = \frac{p-1}{\gcd((p-1)/m, j)},$$

where m is the order of aj in $\mathbb{F}_p^\times$ [Jones and Roberts 2006]. The size of the Galois group is $p \cdot d$ and the Galois group is $C_p \rtimes C_d$. □

Acknowledgments

The results presented in this paper were obtained during the Research Experience for Undergraduates "Computational research on local fields and Galois groups" at Elon University in the summer of 2018, which was supported by Elon University, the National Security Agency (grant H98230-18-1-0012), and the University of North Carolina Greensboro. We also would like to thank the anonymous referee for the helpful comments.

References

[Amano 1971] S. Amano, "Eisenstein equations of degree p in a $\mathfrak{p}$-adic field", *J. Fac. Sci. Univ. Tokyo Sect. IA Math.* **18** (1971), 1–21. MR Zbl

[Cassels 1986] J. W. S. Cassels, *Local fields*, London Mathematical Society Student Texts **3**, Cambridge University Press, 1986. MR Zbl

[Greve and Pauli 2012] C. Greve and S. Pauli, "Ramification polygons, splitting fields, and Galois groups of Eisenstein polynomials", *Int. J. Number Theory* **8**:6 (2012), 1401–1424. MR Zbl

[Jones and Roberts 2006] J. W. Jones and D. P. Roberts, "A database of local fields", *J. Symbolic Comput.* **41**:1 (2006), 80–97. MR Zbl

[Krasner 1966] M. Krasner, "Nombre des extensions d'un degré donné d'un corps $\mathfrak{p}$-adique", pp. 143–169 in *Les tendances géom. en algèbre et théorie des nombres*, Editions du Centre National de la Recherche Scientifique, Paris, 1966. MR Zbl

[Pauli and Roblot 2001] S. Pauli and X.-F. Roblot, "On the computation of all extensions of a p-adic field of a given degree", *Math. Comp.* **70**:236 (2001), 1641–1659. MR Zbl

Received: 2019-04-22 Revised: 2020-02-17 Accepted: 2020-09-15

cawtrey@elon.edu *Department of Mathematics and Statistics, Elon University, Elon, NC, United States*

agaura@princeton.edu *Department of Mathematics, Princeton University, Princeton, NJ, United States*

s_pauli@uncg.edu *Department of Mathematics and Statistics, University of North Carolina, Greensboro, NC, United States*

sbrudzin@uncg.edu *Department of Mathematics and Statistics, University of North Carolina, Greensboro, NC, United States*

auy@andrew.cmu.edu *Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA, United States*

szinzer@aurora.edu *Department of Mathematics, Aurora University, Aurora, IL, United States*

INVOLVE 13:5 (2020)
https://doi.org/10.2140/involve.2020.13.759

Counting pseudo progressions

Jay Cummings, Quin Darcy, Natalie Hobson, Drew Horton,
Keith Rhodewalt, Morgan Throckmorton and Ry Ulmer-Strack

(Communicated by Kenneth S. Berenhaut)

An *m-pseudo progression* is an increasing list of numbers for which there are at most m distinct differences between consecutive terms. This object generalizes the notion of an arithmetic progression. We give two counts for the number of k-term m-pseudo progressions in $\{1, 2, \dots, n\}$. We also provide computer-generated tables of values which agree with both counts and graphs that display the growth rates of these functions. Finally, we present a generating function which counts k-term progressions in $\{1, 2, \dots, n\}$ whose differences are all distinct, and we discuss further directions in Ramsey theory.

1. Introduction and motivation

Arithmetic progressions have been well-studied. Their existence within partitions of $\mathbb{Z}$ is a central theme of Ramsey theory. The existence of long arithmetic progressions within the primes was famously solved by Green and Tao [2008]. Landman and Robertson [2004] recently asked how the theory changes when instead of searching for arithmetic progressions, one searches for a specific generalization of an arithmetic progression, called an m-pseudo progression; in Section 6 we provide more details on this connection. Motivated by this question, we sought to better understand m-pseudo progressions by counting them. In this paper we provide two explicit methods and formulas to count these objects, which we now define.

1A. ***Defining m-pseudo progressions.*** A k-term arithmetic progression is a list of numbers, $a_1, a_2, \dots, a_k$, for which there exists some $d \in \mathbb{Z}^+$ where $a_{i+1} - a_i = d$ for all i. We now generalize this definition to include progressions with a greater number of differences between consecutive terms.

Definition 1.1. A (k-term) *m-pseudo progression* is a list

$$a_1, a_2, \dots, a_k$$

MSC2020: 05A15.
Keywords: enumerative combinatorics, pseudo progressions.

of increasing integers from $\mathbb{Z}^+$ for which there exists a set $\{d_1, d_2, \dots, d_m\}$, where $a_{i+1} - a_i \in \{d_1, d_2, \dots, d_m\}$ for all i. If the m is not specified, it is simply called a *progression*.

For a given progression and any difference d_i, we denote by $\|d_i\|$ the number of times d_i appears as a difference in the progression.

Notice that if $m = 1$, we get the definition of an arithmetic progression. Furthermore, every m-pseudo progression is an ℓ-pseudo progression for $\ell \geq m$, because the definition of an ℓ-pseudo progression requires *at most* ℓ differences.

Example 1.2. The list $2, 5, 8, 13, 16, 21, 24$ is a 7-term 2-pseudo progression since there are at most two common differences:

$$\begin{aligned} 5-2 &= 3, & 16-13 &= 3, \\ 8-5 &= 3, & 21-16 &= 5, \\ 13-8 &= 5, & 24-21 &= 3. \end{aligned}$$

Therefore, by letting $\{d_1, d_2\} = \{3, 5\}$, it is indeed true that $a_{i+1} - a_i \in \{d_1, d_2\}$ for all i.

In this paper, we provide two methods of counting of how many k-term m-pseudo progressions there are in $\{1, 2, \dots, n\}$, and we discuss in detail some special cases.

We note that there are other generalizations of arithmetic progressions. Pseudo progressions are in fact a generalization of what are called *generalized arithmetic progressions*, sometimes also called *d-dimensional arithmetic progressions* or *quasiprogressions*, see for example [Chang 2003; Freiman 1999], which are increasing sequences $a_1, a_2, \dots, a_k$ for which there is some d such that $a_{i+1} - a_i \in \{1, \dots, d\}$ for all i. These progressions demand that all the differences are "close" to each other, while a pseudo progression simply restricts the number of such differences.

2. Counting 2-pseudo progressions

2A. ***Arithmetic progressions.*** We begin with a discussion of the solved problem of counting arithmetic progressions, which we first formally provide. It is beneficial to see this approach, as two of our main theorems generalize the ideas highlighted in this proof.

Theorem 2.1. *There are*

$$F_1 = n\left\lfloor \frac{n}{k-1} \right\rfloor - (k-1)\binom{\lfloor \frac{n}{k-1} \rfloor + 1}{2}$$

k-term arithmetic progressions in $\{1, 2, \dots, n\}$.

Proof. Fix a $d \in \mathbb{Z}^+$. Notice that if you know that a k-term arithmetic progression has common difference d, and you know the first element, then you have determined the entire progression. The possible arithmetic progressions with a common difference of d are

$$\begin{array}{ccccc} 1, & 1+d, & 1+2d, & \dots, & 1+(k-1)d, \\ 2, & 2+d, & 2+2d, & \dots, & 2+(k-1)d, \\ & & \vdots & & \\ n-(k-1)d, & n-(k-1)d+d, & n-(k-1)d+2d, & \dots, & n. \end{array}$$

That is, there are $n-(k-1)d$ such progressions in $\{1,2,\dots,n\}$.

Notice that the largest possible d is $\lfloor \frac{n}{k-1} \rfloor$, and as for any integer $m > \lfloor \frac{n}{k-1} \rfloor$, the first possible arithmetic progression is

$$1, \quad 1+m, \quad 1+2m, \quad \dots, \quad 1+(k-1)m,$$

making its last element $1+(k-1)m > n$. Adding up the number of progressions for all possible d and using the combinatorial identity $\sum_{d=1}^{L} d = \binom{L+1}{2}$, it follows that

$$\begin{aligned} \sum_{d=1}^{\lfloor \frac{n}{k-1} \rfloor} (n-(k-1)d) &= n\left\lfloor \frac{n}{k-1} \right\rfloor - (k-1) \sum_{d=1}^{\lfloor \frac{n}{k-1} \rfloor} d \\ &= n\left\lfloor \frac{n}{k-1} \right\rfloor - (k-1)\binom{\lfloor \frac{n}{k-1} \rfloor + 1}{2}. \end{aligned}$$ □

Remark 2.2. There are a few elements of this proof which will be important later. First, the maximum common difference of a k-term arithmetic progression in $\{1,2,\dots,n\}$ is $\lfloor \frac{n-1}{k-1} \rfloor$. Second, the number of k-term arithmetic progressions in $\{1,2,\dots,n\}$ with common difference d is $n-d(k-1)$. Both of these will be generalized in this paper.

2B. *2-pseudo progressions.* In this section, we provide two methods for counting the number of k-term 2-pseudo progressions in $\{1,2,\dots,n\}$. First, some notation.

Notation 2.3. Suppose

$$a_1, a_2, \dots, a_k$$

is a k-term 2-pseudo progression. We will use a and b to refer to possible differences of this progression, that is, $a_{i+1} - a_i \in \{a,b\}$ for all i. Recall, we will use $\|a\|$ to refer to the number of differences of size a and $\|b\|$ to refer to the number of differences of size b.

For instance, in Example 1.2 we have $a=3$ with $\|a\|=4$ and $b=5$ with $\|b\|=2$.

We will also wish to refer to different orderings of the differences a and b. For example, in Example 1.2 the differences are in the order 3, 3, 5, 3, 5, 3. We call this a difference pattern.

Definition 2.4. Given an m-pseudo progression

$$a_1, a_2, \ldots, a_k,$$

the list

$$a_2 - a_1, a_3 - a_2, \ldots, a_k - a_{k-1}$$

is called the progression's *difference pattern.*

Lemma 2.5. *Fix a set of differences* $\{d_1, d_2, \ldots, d_m\}$ *and numbers* $\|d_1\|$, $\|d_2\|$, $\ldots$, $\|d_m\|$. *The number of k-term m-pseudo progressions in* $\{1, 2, \ldots, n\}$ *with a fixed difference pattern is independent of the difference pattern chosen.*

Proof. Fix a set of differences $\{d_1, d_2, \ldots, d_m\}$ and their multiplicities $\|d_1\|$, $\|d_2\|$, $\ldots$, $\|d_m\|$, as given in the lemma. Next, fix two difference patterns D_1 and D_2 using the d_i from this set. Let S be the collection of all k-term m-pseudo progressions in $\{1, 2, \ldots, n\}$ with difference pattern D_1, and let T be the collection of all k-term m-pseudo progressions with difference pattern D_2. We will show a bijection between the elements in S and T.

Given an arbitrary $s \in S$, say s starts at p_0. Note that there is only one progression starting at s whose difference pattern matches D_1. Let f be the function that maps s to the progression $t \in T$ which starts at p_0, which is likewise unique since its difference pattern is again specified. Moreover, both s and t must end at $p_0 + d_1\|d_1\| + d_2\|d_2\| + \cdots + d_m\|d_m\|$, so if one is in $\{1, 2, \ldots, n\}$, then the other is too. Because a starting point uniquely determines the progression, f is invertible and hence a bijection. This shows there are the same number of elements in S and T, completing the proof. □

Once you have fixed your set of differences and their multiplicities, it will not be difficult to determine how many difference patterns match those criteria. Thus, Lemma 2.5 will be beneficial in that it allows one to focus on a special class of difference patterns, such as ones in which all instances of one of the differences occur before any instance of the second difference. The following lemma pushes this further by fixing the multiplicities but relaxing the differences themselves.

Lemma 2.6. *Fix some* $s_0, t_0 \in \{0, 1, 2, \ldots\}$, *where* $s_0 + t_0 = k - 1$, *and consider any two lists, L_1 and L_2, each consisting of s_0 copies of the variable a and t_0 copies of the variable b, in some order. Let D_1 be the collection of all k-term* 2*-pseudo progressions in* $\{1, 2, \ldots, n\}$ *whose difference pattern matches L_1* (*note that a and b can differ between progressions within D_1, provided the order of the differences matches L_1*), *and let D_2 be the collection of all k-term* 2*-pseudo progressions in* $\{1, 2, \ldots, n\}$ *whose difference pattern matches L_2* (*for appropriate substitutions of a and b*). *Then,* $|D_1| = |D_2|$.

Proof. Consider a k-term 2-pseudo progression P_1 from D_1, and suppose this progression starts at i. Then, there exist a and b for which the progression makes s_0 jumps of size a and t_0 jumps of size b (in the order matching L_1), and the last term of the progression is, therefore, $i + s_0a + t_0b$. Since all the progressions in D_2 also have s_0 copies of one difference and t_0 copies of a second, and the differences are allowed to be anything, if you simply reorder the differences in P_1 to match the ordering of L_2, you get a new 2-pseudo progression P_2 which begins at i and ends at $i + s_0a + t_0b$ and which is now in D_2.

That is, by permuting the order in which you make your jumps of size a and b, you necessarily get a new progression from the same beginning point to the same ending point. So if one of these progressions is in $\{1, 2, \dots, n\}$, then the other is too. And since this procedure is clearly invertible, we have a bijection between D_1 and D_2. □

2C. *Recursive count.* We now present our first count of 2-pseudo progressions. By Lemma 2.6, the number of m-pseudo progressions is independent on the difference pattern. Therefore, we define a simple difference pattern for which the progressions will be simpler to count, and then later scale up this count to include all m-pseudo progressions.

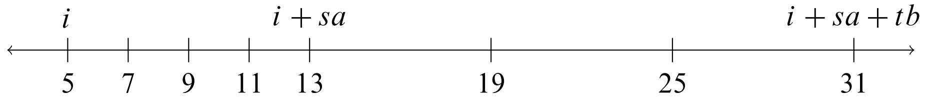
Definition 2.7. Call a 2-pseudo progression *s-t-simple* if its difference pattern is of the form

$$\underbrace{a, a, \dots, a}_{s \text{ terms}}, \underbrace{b, b, \dots, b}_{t \text{ terms}},$$

where s, t, a and b are positive integers, and let $S(n, s, t)$ be the number of s-t-simple 2-pseudo progressions in $\{1, 2, \dots, n\}$.

Note that if $a = b$, then the 2-pseudo progression is in fact an arithmetic progression and is s-t-simple whenever $s + t = k - 1$, where k is the length of the progression.

To better visualize the general case, suppose we want to create a 4-3-simple 2-pseudo progression that starts at $i = 5 \in \{1, 2, \dots, n\}$ and where $a = 2$ and $b = 6$. Since our 2-pseudo progression is of the form a, a, a, a, b, b, b, it looks like:



Here, $i + sa = 13$ and $i + sa + tb = 31$. It is also important to note that given any k-term 2-pseudo progression, $s + t = k - 1$. This is because s and t count the number of differences between terms, so the total number of those differences will always be one less than the number of terms in the progression. For instance, in the example above we have $k = 8$, $s = 4$ and $t = 3$, so we see that $4 + 3 = 8 - 1$. In general, we see that s and t are dependent on k.

Proposition 2.8. $S(n,s,t) = \sum_{a=1}^{\lfloor \frac{n-1-t}{s} \rfloor} \sum_{i=1}^{n-sa-t} \left\lfloor \frac{n-(i+sa)}{t} \right\rfloor.$

Proof. Given an s-t-simple 2-pseudo progression, let us assume that the progression starts at $i \in \{1,2,\ldots,n\}$. Since the first s common differences are of size a, the $(s+1)$-th term of the pseudo progression will be $i+sa$. After $i+sa$, the terms in the 2-pseudo progression have the characteristic that $a_{j+1} - a_j = b$ and since the pseudo progression proceeds with t common differences of size b, the 2-pseudo progression will end at $i+sa+tb$.

Since s and t are dependent on k, instead of fixing a k we choose to fix s and t, and consider the cases for which i, a, and b vary. Without loss of generality, by using the inequality $i+sa+tb \leq n$ we can see that $b \leq (n-(i+sa))/t$. We know that b must be a positive integer, and so by using the floor function we have

$$1 \leq b \leq \left\lfloor \frac{n-(i+sa)}{t} \right\rfloor.$$

Therefore, given any valid selection of a and i, the above gives the possible values of b. That is,

$$S(n,s,t) = \sum_a \sum_i \left\lfloor \frac{n-(i+sa)}{t} \right\rfloor,$$

where the sums are over all the valid values of a and of i. Thus, our focus turns to determining these valid values. We begin with the range of i. We know that $i+sa+tb \leq n$, and so we have that $i \leq n-sa-tb$. And since i is at its greatest when b is at its smallest (when $b=1$), we have

$$1 \leq i \leq n-sa-t.$$

Finally, for a fixed i, consider the valid values of a. In order to find the greatest possible value that a can be, note that a is at its greatest when both b is at its smallest (when $b=1$) and the sequence starts at the earliest index (when $i=1$). Thus, by again using the inequality $i+sa+tb \leq n$ we have that $a \leq (n-i-tb)/s$. And in the case that $b=1$ and $i=1$, we have

$$1 \leq a \leq \left\lfloor \frac{n-1-t}{s} \right\rfloor.$$

From this, we obtain our final count,

$$S(n,s,t) = \sum_{a=1}^{\lfloor \frac{n-1-t}{s} \rfloor} \sum_{i=1}^{n-sa-t} \left\lfloor \frac{n-(i+sa)}{t} \right\rfloor,$$

which completes the proof. □

Note, though, that there are many other progressions with s common differences of size a and t common differences of size b. For example,

$$\underbrace{a, a, \dots, a}_{s-2 \text{ terms}}, \underbrace{b, b, \dots, b}_{t-2 \text{ terms}}, a, b, b, a.$$

We now count these other forms.

Corollary 2.9. *Let F be a list of s copies of a and t copies of b, in some order. Let D be the collection of difference patterns which, for some substitution of a and b, match F. Then, the number of 2-pseudo progressions in $\{1, 2, \dots, n\}$ with a difference pattern in D is equal to $S(n, s, t)$.*

Proof. This follows immediately from Lemma 2.6. □

Definition 2.10. Let $F_m(n, k)$ be the number of m-pseudo progressions. When the context is clear, we will write simply F_m.

Note that $F_1(n, k)$ is the number of k-term arithmetic progressions in $\{1, 2, \dots, n\}$, which we counted in Section 2A.

Theorem 2.11. *Fix n and k. Let $a_1, a_2, \dots, a_k$ be a 2-pseudo progression where $a_{i+1} - a_i \in \{a, b\}$ for all i. Let s be defined as the number of elements such that $a_{i+1} - a_i = a$ and t be defined as the number of elements such that $a_{j+1} - a_j = b$. Then,*

$$F_2 = F_1 + \frac{1}{2}\binom{k-1}{\frac{k-1}{2}}\Big[S\Big(n, \frac{k-1}{2}, \frac{k-1}{2}\Big) - F_1\Big] + \sum_{s=1}^{\lfloor \frac{k-2}{2} \rfloor} \binom{s+t}{s}[S(n, s, t) - F_1].$$

Proof. Recall that s is the number copies of a and t is the number of copies of b, while a and b can be any integers. Since one of these integers will occur at least as many times as the other, we may assume without loss of generality that $s \le t$. First, assume $s < t$.

By Proposition 2.8, $S(n, s, t)$ counts the number of progressions of the form

$$\underbrace{a, a, \dots, a}_{s \text{ terms}}, \underbrace{b, b, \dots, b}_{t \text{ terms}}.$$

By Corollary 2.9, we know that given any form with s copies of a and t copies of b, there are $S(n, s, t)$ progressions of that form. Moreover, one can see that there are $\binom{s+t}{s}$ such forms — the $s + t$ terms in the form are each either an a or a b, and this is determined by choosing which s of these terms are an a. Now, $S(n, s, t)$ includes all the arithmetic progressions (in the case where $a = b$). Therefore,

$$S(n, s, t) - F_1$$

counts the number of progressions of any fixed form containing s copies of a and t copies of b, which contains *exactly* two distinct common differences. There are

$$\binom{s+t}{s}[S(n,s,t)-F_1]$$

2-pseudo progressions, excluding the arithmetic progressions. Note that $s+t=k-1$, and so in our current case where $s<t$ (i.e., $s \le t-1$), we know that $2s+1 \le k-1$, implying that $s \le \lfloor \frac{k-2}{2} \rfloor$. And so, in the case where $s<t$, the total number of 2-pseudo progressions which are not arithmetic progressions is

$$\sum_{s=1}^{\lfloor \frac{k-2}{2} \rfloor} \binom{s+t}{s}[S(n,s,t)-F_1].$$

Thus, among all cases in which $s<t$, the total number of 2-pseudo progressions is

$$F_1 + \sum_{s=1}^{\lfloor \frac{k-2}{2} \rfloor} \binom{s+t}{s}[S(n,s,t)-F_1].$$

The last case to consider is when $s=t$; note that this is only possible if k is odd. And since $s+t=k-1$, we have $s=t=\frac{k-1}{2}$. Just like above, given any form using $\frac{k-1}{2}$ copies of a and $\frac{k-1}{2}$ copies of b, there are $S\big(n, \frac{k-1}{2}, \frac{k-1}{2}\big)$ progressions of this form. And so there are

$$S\Big(n, \frac{k-1}{2}, \frac{k-1}{2}\Big) - F_1$$

progressions that have exactly two distinct common differences. The only difference is in the next step. Note that if we simply multiply by $\binom{s+t}{s}$, we will be over-counting by a factor of 2. Indeed, since $s=t$, any progression comes about in two ways: once when the copies of a are counted by s and the copies of b are counted by t, and once when the copies of a are counted by t and the copies of b are counted by s. Thus, the count in the $s=t$ case is

$$\frac{1}{2}\binom{k-1}{\frac{k-1}{2}}\Big[S\Big(n, \frac{k-1}{2}, \frac{k-1}{2}\Big) - F_1\Big].$$

Note that the binomial here evaluates to 0 in the event that k is even, and so including this term in the even case is consistent. This gives us our final answer

$$F_2 = F_1 + \frac{1}{2}\binom{k-1}{\frac{k-1}{2}}\Big[S\Big(n, \frac{k-1}{2}, \frac{k-1}{2}\Big) - F_1\Big] + \sum_{s=1}^{\lfloor \frac{k-2}{2} \rfloor} \binom{s+t}{s}[S(n,s,t)-F_1]. \quad \square$$

Since there are $F_\ell(n,k)$ sequences where you are allowed up to ℓ differences, and $F_{\ell-1}(n,k)$ sequences where you are allowed up to $\ell-1$ differences, there are

$$F_\ell(n,k) - F_{\ell-1}(n,k)$$

sequences with exactly ℓ differences.

If we let $\widetilde{F}_\ell(n,k)$ denote the number of ℓ-pseudo progressions with *exactly* ℓ distinct differences, and let $\widetilde{S}(n,s,t)$ likewise denote the number of s-t-simple progressions with exactly two differences, then $\widetilde{S}(n,s,t) = S(n,s,t) - F_1(n,k)$ and Theorem 2.11 has the reduced form

$$\widetilde{F}_2(n,k) = \frac{1}{2}\binom{k-1}{\frac{k-1}{2}}\widetilde{S}\Big(n,\frac{k-1}{2},\frac{k-1}{2}\Big) + \sum_{s=1}^{\lfloor\frac{k-2}{2}\rfloor}\binom{s+t}{s}\widetilde{S}(n,s,t).$$

Moreover, the above two equations imply

$$F_2(n,k) - F_1(n,k) = \frac{1}{2}\binom{k-1}{\frac{k-1}{2}}\widetilde{S}\Big(n,\frac{k-1}{2},\frac{k-1}{2}\Big) + \sum_{s=1}^{\lfloor\frac{k-2}{2}\rfloor}\binom{s+t}{s}\widetilde{S}(n,s,t),$$

which is another form of Theorem 2.11.

2D. ***Iterative count.*** As in the previous sections, we use a and b to denote the two differences in a 2-pseudo progression. In this section, we will insist that $a < b$, which in particular prohibits $a = b$. We will often refer to a k-term 2-pseudo progression in $\{1, 2, \dots, n\}$ as just a "progression" if the context is clear.

Remark 2.12. Recall, for a k-term 2-pseudo progression,

$$\|a\| + \|b\| = k - 1.$$

That is, the total number of differences of sizes a and b is equal to the total number of differences in the progression, $k - 1$. If we are considering k-term progressions, then we only need to know either $\|a\|$ or $\|b\|$ and the other will follow.

First, given a fixed number of two differences, we determine the maximum value these differences can be.

Lemma 2.13. *For a k-term 2-pseudo progression with a fixed number of differences, say $\|a\|$ and $\|b\|$ (without the sizes of the differences a and b being determined), the largest possible value for a is*

$$a_{\max} = \left\lfloor \frac{(n-1) - \|b\|}{k-1} \right\rfloor.$$

Proof. We begin by noting that similar to Remark 2.2, the maximum possible difference between the first and last terms of a 2-pseudo progression is $n-1$. Since we are assuming $a < b$, if we want to find the largest possible value for a, we can assume $a = b-1$. Thus, we must distribute the difference of $n-1$ into $k-1$ groups ($\|a\|$ groups of size a, and $\|b\|$ groups of size $a+1$). To do this, we subtract $\|b\|$ from $n-1$ in order to account for the $\|b\|$ groups of one larger value than a. The maximum possible value for the difference a is the result in the lemma. □

Similarly, with a fixed difference of size a and number of differences, say $\|a\|$ and $\|b\|$, we can determine the maximum value of the difference b in a k-term 2-pseudo progression in $\{1,2,\dots,n\}$.

Lemma 2.14. *Suppose a k-term 2-pseudo progression has difference a and some number of differences, say $\|a\|$ and $\|b\|$. Then, the maximum possible value for b is*

$$\bar{b}_a = \begin{cases} 0, & \|b\| = 0, \\ \lfloor \frac{(n-1)-a\|a\|}{\|b\|} \rfloor, & \|b\| \neq 0. \end{cases}$$

Proof. Similar to our argument in Lemma 2.13, the largest possible value between the first and final terms of a progression in $\{1,2,\dots,n\}$ is $n-1$. Thus, to determine the maximum possible b that can create a progression, we must divide $n-1$ into $k-1$ groups ($\|a\|$ groups of size a and $\|b\|$ groups of size b). In order to account for the $\|a\|$ differences of size a, we must subtract off the product $a\|a\|$ and divide the remaining $(n-1)-a\|a\|$ into $\|b\|$ groups. Thus, we have our resulting maximum above. The equality is also guaranteed, since a 2-pseudo progression with these metrics which begins at 1 will end at $1+a\|a\|+\bar{b}_a\|b\| \leq n$. □

Proposition 2.15. *Given a fixed a, b, $\|a\|$ and $\|b\|$, the number of k-term 2-pseudo progressions in a set $\{1,2,\dots,n\}$ with a fixed difference pattern is*

$$n - a\|a\| - b\|b\|.$$

Proof. Fix the integers n, k, a, b, $\|a\|$ and $\|b\|$, and a difference pattern. Similar to the argument in the proof of Theorem 2.1, we proceed by first considering a progression with initial term 1. Since our differences are of size a and b and there are $\|a\|$ a's and $\|b\|$ b's, we have that the final term in the progression will be $1+a\|a\|+b\|b\|$. In general, if a progression with these parameters has initial term p_0, then the final term will be $p_0+a\|a\|+b\|b\|$. Such a progression is valid in $\{1,2,\dots,n\}$ if this final term $p_0+a\|a\|+b\|b\| \leq n$. Such will be the case when

$$p_0 \leq n - a\|a\| - b\|b\|.$$

Thus, the total number of valid 2-pseudo progressions with these parameters is equal to the largest p_0 such that the above inequality is true. And so the result of the proposition follows. □

Proposition 2.16. *The total number of k-term 2-pseudo progressions in $\{1,2,\dots,n\}$ can be counted using the formula*

$$F_1 + \sum_{\|a\|=1}^{k-1} \sum_{a=1}^{a_{\max}} \sum_{b=a+1}^{\bar{b}_a} \Big[\binom{k-1}{\|a\|} (n - a\|a\| - b\|b\|) \Big].$$

Proof. A formula for F_1 is given in Theorem 2.1, and it counts the number of arithmetic progressions in $\{1, 2, \dots, n\}$; we now turn to count the number of 2-pseudo progressions which contain two distinct differences.

As a consequence of Proposition 2.15, the total number of k-term 2-pseudo progressions in $\{1, 2, \dots, n\}$ with a given difference pattern and fixed values of $a, b, \|a\|, \|b\| \in \mathbb{Z}^+$ is given by $n - a\|a\| - b\|b\|$. Furthermore, by Lemma 2.5, this value does not depend on the difference pattern chosen. Thus, given fixed a, b, and $\|a\|$ (which implies the value of $\|b\|$), we count the total number of progressions with these parameters by scaling the count in Proposition 2.15 by the total number of difference patterns that could occur with $\|a\|$ a's and $\|b\|$ b's. In particular, there are $\binom{k-1}{\|a\|}$ such ways, by choosing which of the total $k-1$ skips to place the $\|a\|$ skips of size a.

Now we must determine all possible values of a and b that are possible. We assume $0 < a < b$, so we have $a = 1$ being the smallest possible value for a. Thus, $b = a + 1$ is the smallest possible value for b given a value for a. This gives us the bounds for the inner two sums.

Finally, we iterate this process through all possible positive values of $\|a\|$ and $\|b\|$, by summing up all valid progressions for each value of $\|a\|$, with $1 \le \|a\| \le k - 1$. Since $\|a\| + \|b\| = k - 1$, iterating through all possible values of $\|a\|$ will indeed iterate all valid positive pairs of $\|a\|$ and $\|b\|$. □

3. Counting m-pseudo progressions

We now generalize Lemma 2.6, which will be used to generalize Corollary 2.9

Lemma 3.1. *Fix some $d_0, d_1, \dots, d_m \in \{0, 1, 2, \dots\}$, where $\sum_{i=1}^{m} d_i = k-1$, and consider any two lists, L_1 and L_2, each consisting of $\|d_i\|$ copies of the variable d_i for each i, in some order. Let D_1 be the collection of all k-term m-pseudo progressions in $\{1, 2, \dots, n\}$ whose difference pattern matches L_1 (for appropriate substitutions of $d_1, d_2, \dots, d_m$), and let D_2 be the collection of all k-term m-pseudo progressions in $\{1, 2, \dots, n\}$ whose difference pattern matches L_2 (for appropriate substitutions of $d_1, d_2, \dots, d_m$). Then, $|D_1| = |D_2|$.*

Proof. We will show a bijective correspondence between k-term m-pseudo progressions with the same number of differences $\|d_i\|$ for all i, regardless of the ordering of the differences. Let S and T be the set of m-pseudo progressions with distinct difference patterns containing differences $d_1, d_2, \dots, d_m$ such that the number of differences $\|d_i\|$ is the same in each difference pattern for all i.

Given an arbitrary $s \in S$, if s starts at, say, p_0, then it will end at $p_0 + d_1\|d_1\| + d_2\|d_2\| + \cdots + d_m\|d_m\|$. Let f be the function that maps s to the m-pseudo progression $t \in T$ starting at p_0. Note that this is well-defined, as pairing the starting term of an m-pseudo progression with that progression's difference pattern uniquely determines the progression. Any $t \in T$ is the image of the m-pseudo progression in S starting at the initial term of t. Therefore, the number of progressions with a fixed difference pattern is independent of the ordering of the differences in the difference pattern. □

3A. ***Recursive count.*** In this section we generalize the ideas of Section 2C to count k-term m-pseudo progressions in $\{1, 2, \dots, n\}$.

Definition 3.2. A progression is $(s_1, s_2, \dots, s_m)$-simple if it is of the form

$$\underbrace{a_1, a_1, \dots, a_1}_{s_1 \text{ terms}}, \underbrace{a_2, a_2, \dots, a_2}_{s_2 \text{ terms}}, \dots, \underbrace{a_m, a_m, \dots, a_m}_{s_m \text{ terms}},$$

where $s_i, a_j \in \mathbb{Z}^+$ for all i, j. Let $S(n, s_1, \dots, s_m)$ be the number of $(s_1, s_2, \dots, s_m)$-simple progressions in $\{1, 2, \dots, n\}$.

By the same reasoning as in Section 2C, there is a bijection between the set of progressions of this form and the set of progressions of any permuted form, and there are $\binom{k-1}{s_1 \cdots s_m}$ such permuted forms. And

$$S(n, s_1, \dots, s_{m+1}) = \sum_{i=1}^{n} \left\lfloor \frac{i}{s_{m+1}} \right\rfloor S(n-i, s_1, \dots, s_m),$$

since an $(s_1, s_2, \dots, s_{m+1})$-simple progression in $\{1,2,\dots,n\}$ is an $(s_1, s_2, \dots, s_m)$-simple progression in $[n-i]$, for some i, followed immediately by s_{m+1} more terms. And within the remaining i numbers, starting with the first, there are $\lfloor \frac{i}{s_{m+1}} \rfloor$ possible common differences that will keep the entire $(s_1, s_2, \dots, s_{m+1})$-simple progression in $\{1, 2, \dots, n\}$.

So in this way we have recursively found the number of $(s_1, s_2, \dots, s_m)$-simple progressions for an arbitrary m. And so

$$S(n, s_1, \dots, s_{m+1}) - F_m(n)$$

counts the number of $(s_1, s_2, \dots, s_m)$-simple progressions with *exactly* $m+1$ distinct common differences. And if $s_1 < s_2 < \cdots < s_{m+1}$, then by Lemma 3.1,

$$\binom{k-1}{s_1 \cdots s_{m+1}} [S(n, s_1, \dots, s_{m+1}) - F_m(n)]$$

counts the number of all progressions which are of a form which is a permutation of the $(s_1, s_2, \dots, s_m)$-simple form.

Otherwise, if we only assume $s_1 \le s_2 \le \cdots \le s_{m+1}$, then just like with 2-pseudo progressions, we may overcount. Indeed, if in the multiset $\{s_1, s_2, \dots, s_{m+1}\}$ we have, say, 7 appearing 3 times (say, $s_3 = s_4 = s_5 = 7$), then this alone causes the above expression to overcount by a factor of $3! = 6$ (just like in the 2-pseudo progressions, when $s = t$ implied that we overcounted by a factor of $2! = 2$). To see this, note that each such progression will have been counted once when the copies of a_3 were counted by s_3, the copies of a_4 were counted by s_4, and the copies of a_5 were counted by s_5; but will also have been counted once for each of the other 3! pairings between these a_i's and s_j's. This reasoning gives the answer of

$$\frac{1}{n_1!\,n_2!\cdots n_\ell!}\binom{k-1}{s_1\ldots s_{m+1}}[S(n,s_1,\ldots,s_{m+1})-F_m(n)],$$

where n_i is the multiplicity of the i-th distinct number in $s_1,s_2,\ldots,s_{m+1}$, and therefore ℓ is the size of $\{s_1,s_2,\ldots,s_{m+1}\}$ as a set (i.e., removing multiplicities). Note that our answer in the 2-pseudo progression case is a special case of this. Adding up all the possibilities, and adding back in the progressions with at most m distinct common differences, gives you the final count

$$F_{m+1}(n,k)= \\ F_m(n)+\sum_{s_1\le\cdots\le s_{m+1}}\frac{1}{n_1!\,n_2!\cdots n_\ell!}\binom{k-1}{s_1\ldots s_{m+1}}[S(n,s_1,\ldots,s_{m+1})-F_m(n)].$$

This also gives a particularly nice formula for $\widetilde{F}_{m+1}(n,k)$, which you recall is the number of k-term progressions in $\{1,2,\ldots,n\}$ with *exactly* $m+1$ common differences. The above reduces to

$$\widetilde{F}_{m+1}(n,k)=\sum_{s_1\le\cdots\le s_{m+1}}\frac{1}{n_1!\,n_2!\cdots n_\ell!}\binom{k-1}{s_1\cdots s_{m+1}}\widetilde{S}(n,s_1,\ldots,s_{m+1}).$$

3B. ***Iterative count.*** In this section we generalize the ideas of Section 2D to count k-term m-pseudo progressions in $\{1,2,\ldots,n\}$.

Remark 3.3. As before, the differences in an m-pseudo progression will be denoted by d_i for $1\le i\le m$. In this section, we assume

$$0<d_1<d_2<\cdots<d_m.$$

Proposition 3.4. *For any i such that $0<i\le m$, and fixed list of positive integers $\|d_1\|,\|d_2\|,\ldots,\|d_m\|$ such that $\sum_{j=1}^{m}\|d_j\|=k-1$, and fixed multiplicities $0<d_1<d_2<\cdots<d_{i-1}$, the maximum value of the difference d_i of a k-term m-pseudo progression with m distinct differences is*

$$\bar{d}_i=\left\lfloor\frac{(n-1)-\left(\sum_{j=1}^{i-1}d_j\|d_j\|+\sum_{j=i+1}^{m}((j-i)\|d_j\|)\right)}{k-1-\sum_{j=1}^{i-1}\|d_j\|}\right\rfloor.$$

Proof. Assume the setup of Proposition 3.4. We have a fixed list of numbers $\|d_1\|,\|d_2\|,\ldots,\|d_m\|$ such that $\sum_{j=1}^{m}\|d_j\|=k-1$, a difference pattern, and a fixed size of differences $d_1,d_2,\ldots,d_{i-1}$.

Similar to the proof for Lemmas 2.14 and 2.13, the largest difference between the initial and final term of a progression in $\{1, 2, \ldots, n\}$ is $n-1$. In other words, for a progression that starts at 1, the last term in the progression is, $1 + \sum_{i=1}^{m} d_i \|d_i\|$. If we want to determine a difference that is as *large* as possible, we can assume that $1 + \sum_{i=1}^{m} d_i \|d_i\|$ is as large as possible. Thus, we assume $1 + \sum_{i=1}^{m} d_i \|d_i\| = n$ (and we consider issues of whether this value is an integer later). Thus, to determine the maximum value for d_i we must remove $d_j \|d_j\|$ from $n-1$ for each known difference $d_1, d_2, \ldots, d_{i-1}$.

Similar to the proof for Lemma 2.13, since we want to determine the largest possible value for d_i, we will assume for each $j > i$ that d_j is as small as possible while still maintaining the inequality from Remark 3.3. That is, we will assume for each $j > i$ that $d_j = d_i + (j - i)$. For example, $d_{i+1} = d_i + 1$.

However, since we are determining the value of d_i, we will account for each d_j of size $d_i + (j - i)$ by removing $(j - i)\|d_j\|$ from $n-1$ and distributing the remaining value equally between the remaining $k - 1 - \sum_{j=1}^{i-1} \|d_j\|$ possible skips.

The floor of this expression gives the largest possible value for the difference of size d_i. □

Proposition 3.5. *Given fixed $d_1, d_2, \ldots, d_m$, $\|d_1\|, \|d_2\|, \ldots \|d_m\|$ such that*

$$\sum_{i=1}^{m} \|d_i\| = k - 1,$$

and difference pattern, the number of k-term m-pseudo progressions in $\{1, 2, \ldots, n\}$ with a fixed difference pattern is

$$n - \sum_{j=1}^{m} d_j \|d_j\|.$$

Proof. This result follows the same reasoning as Proposition 2.15. □

Proposition 3.6. *Given n, k and m, the total number of k-term pseudo progressions in $\{1, 2, \ldots, n\}$ with exactly m distinct differences is*

$$\sum_{\|d_1\|=1}^{k-1} \cdots \sum_{\|d_{m-1}\|=1}^{k-(m-1)} \sum_{d_1=1}^{\bar{d}_1} \cdots \sum_{d_m=d_{m-1}+1}^{\bar{d}_m} \binom{k-1}{\|d_1\|, \ldots, \|d_m\|} \left(n - \sum_{j=1}^{m} d_j \|d_j\|\right).$$

This can be written more succinctly as

$$\sum_{\mathbb{D}} \binom{k-1}{\|d_1\|, \|d_2\|, \ldots, \|d_m\|} \left(n - \sum_{j=1}^{m} d_j \|d_j\|\right),$$

where $\mathbb{D} = \{(d_1, \ldots, d_m, \|d_1\|, \ldots, \|d_m\|)$ such that $\|d_i\| \neq 0$, $\sum_{i=1}^{m} \|d_i\| = k-1$, $1 \leq d_i \leq \bar{d}_i$, and $d_i < d_j$ whenever $i < j\}$.

In order to compute the total number of k-term m-pseudo progressions (*that is*, *progressions with up to m distinct differences*), *sum the above over all m from* 1 *to m.*

Proof. We can determine the total number of m-pseudo progressions in $\{1, 2, \dots, n\}$ with a given set of fixed positive integers $d_1, d_2, \dots, d_m$, $\|d_1\|, \|d_2\|, \dots, \|d_m\|$ such that $\sum_{i=1}^{m} \|d_i\| = k-1$ and a fixed difference pattern from Proposition 3.5. By Lemma 2.5, we can scale this count by the number of possible difference patterns to determine the number of m-pseudo progressions with fixed parameters $d_1, d_2, \dots, d_m$, $\|d_1\|, \|d_2\|, \dots, \|d_m\|$. The number of such difference patterns is the number of ways to choose where the $\|d_i\|$ differences of size d_i occur for each i from 1 to m. That is, the multinomial coefficient

$$\binom{k-1}{\|d_1\|, \|d_2\|, \dots, \|d_m\|}.$$

To determine the allowable collections of numbers $d_1, d_2, \dots, d_m$ and $\|d_1\|$, $\|d_2\|, \dots, \|d_m\|$, we continue with similar reasoning as in Proposition 2.16. That is, we iterate over all possible values of $\|d_i\|$ from 1 to $k-1-(m-1)$ (in order to ensure $\|d_i\| \neq 0$ for all i) such that $\sum_{i=1}^{m} \|d_i\| = k-1$. In order to maintain the inequality from Remark 3.3 and the maximum in Proposition 3.4, we iterate over the values of d_i from $d_{i-1}+1$ to $\bar{d}_i$. All such valid lists of differences $d_1, d_2, \dots, d_m$ and amounts $\|d_1\|, \|d_2\|, \dots, \|d_m\|$ can be represented by the set $\mathbb{D}$. □

3C. *Reinterpreting combinatorial identities.* Observe that if a $(k-1)$-pseudo progression has j numbers in $[v]$ (which can occur in $\binom{v}{j}$ ways), then the other $k-j$ numbers in the pseudo progression can be anywhere in $\{v+1, v+2, \dots, n\}$, which has size $n-v$. The number of ways to complete this is $F_{k-1-j}(n-v, k-j)$. Thus,

$$F_{k-1}(n,k) = \sum_{j=0}^{k} \binom{v}{j} F_{k-1-j}(n-v, k-j).$$

Recalling that $F_{k-1}(n,k) = \binom{n}{k}$, this gives a new proof of the Chu–Vandermonde identity

$$\binom{n}{k} = \sum_{j=0}^{k} \binom{v}{j} \binom{n-v}{k-j}.$$

4. Generating functions

An m-pseudo progression places a limit of m on the number of distinct differences within such a progression. In this section, we go to the opposite extreme and ask what happens if we demand all of the differences be distinct. Indeed, below we find the generating function which counts the number of k-term pseudo progressions in $\{1, 2, \dots, n\}$ where all of the $k-1$ differences are distinct.

We will be discussing m-pseudo progressions without needing to refer to any particular m ($m = k - 1$ would suffice, except that we do not wish to allow fewer than $k - 1$ distinct differences). Therefore we will continue to refer to these as pseudo progressions without mentioning any m.

It is well known that

$$\frac{x^{\frac{k(k-1)}{2}}}{(1-x)(1-x^2)\cdots(1-x^{k-1})}$$

is the generating function for integer partitions with $k - 1$ distinct parts. That is, the coefficient of x^t in this generating function gives the number of partitions of t into $k - 1$ distinct parts: $t = p_1 + p_2 + \cdots + p_{k-1}$, where each p_i is a positive integer and $p_1 < p_2 < \cdots < p_{k-1}$.

Definition 4.1. Fix a k and n. For $t < n$, let $c(t)$ be the number of partitions of t into $k - 1$ distinct parts.

Lemma 4.2. *There are*

$$\sum_{t=\frac{k(k-1)}{2}}^{n-1} (k-1)!\,(n-t)c(t)$$

k-term pseudo progressions in $\{1, 2, \ldots, n\}$ with distinct common differences.

Proof. Given a partition of t into $k-1$ distinct parts, note that we can create a pseudo progression in $\{1, 2, \ldots, n\}$ which starts at 1, ends at $t + 1$, and whose common differences are distinct. Namely, if the partition is $t = p_1 + p_2 + \cdots + p_{k-1}$, then the pseudo progression is

$$1, \quad 1 + p_1, \quad 1 + p_1 + p_2, \quad \ldots, \quad 1 + t.$$

Also, observe that because $p_1 < p_2 < \cdots < p_{k-1}$, we in fact can find $(k-1)!$ pseudo progressions which start at 1 and end at $t + 1$ by simply considering all possible permutations of $\{p_1, \ldots, p_{k-1}\}$, and adding in the p_i in the order determined by the permutation.

Moreover, all k-term pseudo progressions with distinct common differences that start at 1 and end at $t + 1$ can be realized in this way. To see this, simply take such a pseudo progression, $1 = a_1, a_2, \ldots, a_k = t + 1$, and observe the $k - 1$ distinct common differences,

$$a_2 - a_1, \quad a_3 - a_2, \quad \ldots, \quad a_k - a_{k-1}.$$

The sum of these common differences telescopes, so their sum can be seen as $a_k - a_1 = (t + 1) - 1 = t$. And being distinct, once they are reordered in increasing order they do indeed form a partition of t with $k - 1$ distinct parts.

So there is in fact a total of $(k-1)!\,c(t)$ k-term pseudo progressions in $\{1,2,\dots,n\}$ which start at 1, end at $t+1$, and have distinct common differences. To obtain a count for all such progressions, we simply need to multiply by the number of possible starting points. The progressions could begin at 1 and end at $t+1$, begin at 2 and end at $t+2$, ..., begin at $n-t$ and end at $t+(n-t)$. In total, there are $n-t$ ways that we can "shift" these progressions which start at 1 into progression that start at higher values. Thus, by multiplying by $n-t$ we get the total number of $(n-t)(k-1)!\,c(t)$ k-term pseudo progressions in $\{1,2,\dots,n\}$ with distinct common differences.

Finally, we must sum over all possible values of t. The smallest t corresponds to the smallest value which can be partitioned into $k-1$ distinct parts, which is

$$1+2+3+\cdots+(k-1)=\frac{k(k-1)}{2}.$$

The largest possible t is $n-1$, since this corresponds to a progression which starts at 1 and ends at $t+1=n$. Thus, by summing over these possible vales of t, we get our final count

$$\sum_{t=\frac{k(k-1)}{2}}^{n-1}(k-1)!\,(n-t)c(t). \qquad \square$$

We now use this to find the generating function for the number of k-term pseudo progressions in $\{1,2,\dots,n\}$ with distinct common differences.

Theorem 4.3. *The number of k-term pseudo progressions in $\{1,2,\dots,n\}$ with distinct common differences is the coefficient on x^n in the generating function*

$$\frac{(k-1)!\,x^{1+\frac{k(k-1)}{2}}}{(1-x)^3(1-x^2)(1-x^3)\cdots(1-x^{k-1})}.$$

Proof. Recall that the generating function for the number of integer partitions with distinct parts is

$$\frac{x^{\frac{k(k-1)}{2}}}{(1-x)(1-x^2)\cdots(1-x^{k-1})}.$$

That is, the coefficient of x^t in this generating function gives $c(t)$. By scaling, the coefficient of x^n in

$$\frac{x^{n-t+\frac{k(k-1)}{2}}}{(1-x)(1-x^2)\cdots(1-x^{k-1})}$$

now gives $c(t)$. Thus, by Lemma 4.2, since there are $\sum_{t=k(k-1)/2}^{n-1}(k-1)!\,(n-t)c(t)$ k-term pseudo progressions in $\{1,2,\dots,n\}$ with distinct common differences, this value is given by the coefficient of x^n in

$$\sum_{t=\frac{k(k-1)}{2}}^{n-1} \frac{(k-1)!\,(n-t)x^{n-t+\frac{k(k-1)}{2}}}{(1-x)(1-x^2)\cdots(1-x^{k-1})}$$
$$= \frac{(k-1)!\,x^{\frac{k(k-1)}{2}}}{(1-x)(1-x^2)\cdots(1-x^{k-1})} \sum_{t=\frac{k(k-1)}{2}}^{n-1} (n-t)x^{n-t}.$$

By substituting i for $n-t$, which reverses the order of summation, the above is equivalent to

$$\frac{(k-1)!\,x^{\frac{k(k-1)}{2}}}{(1-x)(1-x^2)\cdots(1-x^{k-1})} \sum_{i=1}^{n-\frac{k(k-1)}{2}} ix^i.$$

Notice that the coefficient on x^n here is the same as in

$$\frac{(k-1)!\,x^{\frac{k(k-1)}{2}}}{(1-x)(1-x^2)\cdots(1-x^{k-1})} \sum_{i=1}^{\infty} ix^i,$$

and so this new expression also has the property that the coefficient on x^n gives the number of k-term pseudo progressions in $\{1, 2, \ldots, n\}$ with distinct common differences. Since $\sum_{i=1}^{\infty} ix^i$ has generating function $x/(x-1)^2$, this is equivalent to

$$\frac{(k-1)!\,x^{1+\frac{k(k-1)}{2}}}{(1-x)^3(1-x^2)(1-x^3)\cdots(1-x^{k-1})},$$

as desired. □

5. Symmetries

We have observed (see Section 7) that for certain small values of k, the number of k-term m-pseudo progressions in $\{1, 2, \ldots, n\}$ is equal to the number of $(n-k)$-term m-pseudo progressions in $\{1, 2, \ldots, n\}$. Indeed, the relationship seems to be related to the complement. Consider a k-term m-pseudo progression and let K be the subset of $\{1, 2, \ldots, n\}$ consisting of the elements of the progression. Then, the set $K^c = \{1, 2, \ldots, n\} \setminus K$ corresponds to an $(n-k)$-term progression.

Note that the K^c progression will include a difference of 1 whenever there are two adjacent numbers in $\{1, 2, \ldots, n\}$ which are not in K (for $k < \frac{n}{2} - 1$, this is guaranteed). The K^c progression will include a difference of 2 whenever the K progression had a term $i \in \{2, 3, \ldots, n-1\}$ for which $i-1$ and $i+1$ are not in K (for most sets K of small size, such an i will exist). For the K^c progression to have a difference of $d > 1$, the K progression would have to include $d-1$ consecutive terms.

Since terms from the K progression have to be used to create differences in the K^c progressions, $|K|$ creates a bound on how many differences the K^c can have. Indeed, by this reasoning, it is impossible for the K^c progression to have more than m differences if

$$|K| < 1 + \sum_{d=2}^{m+1} (d-1) = 1 + \binom{m+1}{2}.$$

In Section 7, these symmetries appear in our tables of values, which also show other interesting behaviors. For example, the number of 4-term 3-pseudo progressions in $\{1, 2, \dots, 14\}$ is equal to the number of 10-term 3-pseudo progressions in $\{1, 2, \dots, 14\}$.

6. Further directions

We were motivated to study this problem because of a problem in Ramsey theory, and it is in this direction that we plan to move to next.

Consider the positive integers $\mathbb{Z}^+ = \{1, 2, 3, 4, \dots\}$. An r-coloring of these integers is produced by assigning each of these integers one of r colors. The question is whether *every* r-coloring of $\mathbb{Z}^+$ contains a *k-term monochromatic arithmetic progression*. Such a progression is a collection of integers a, $a + d$, $a + 2d$, $\dots$, $a + (k-1)d$ which are all assigned the same color. Here, d is called the *common difference*. The seminal van der Waerden theorem [1927] says that given any k and r, there exists some N such that every r-coloring of $\{1, 2, 3, \dots, N\}$ contains a k-term monochromatic arithmetic progression; the smallest such N is denoted by $w(k, r)$. For example, $w(3, 2) = 9$. That is, every 2-coloring of $\{1, 2, 3, \dots, 9\}$ contains a 3-term monochromatic arithmetic progression, and furthermore it is not true that every such coloring of $\{1, 2, 3, \dots, 8\}$ does. For example, here is a 2-coloring that avoids such a progression:

$$1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8.$$

Much work has been done to try to bound $w(k, r)$. The best upper bound is

$$w(k, r) \le 2^{2^{r^{2^{2^{k+9}}}}},$$

and is due to Tim Gowers.

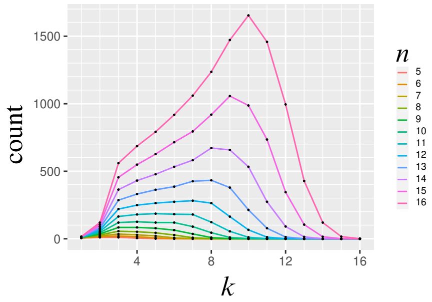
Brown, Graham and Landman [Brown et al. 1999] investigated what happens when you restrict the allowable set of arithmetic progressions. In particular, if $D \subseteq \mathbb{Z}^+$ is a set of allowable common differences, they asked whether there must still exist an N for which every r-coloring of $\{1, 2, 3, \dots, N\}$ contains a monochromatic arithmetic progression whose common difference is in D. That is, their research focused on a subset of the collection of arithmetic progressions. It seems natural then to ask what happens when you instead consider a superset of this collection.

Landman and Robertson recently asked about generalizations of van der Waerden's theorem to m-pseudo progressions. Now that m-pseudo progressions are better understood through their count, we aim to determine the smallest values of N for which every r-coloring of $\{1, 2, \dots, N\}$ contains a monochromatic m-pseudo progression.

7. Tables of values and graphs

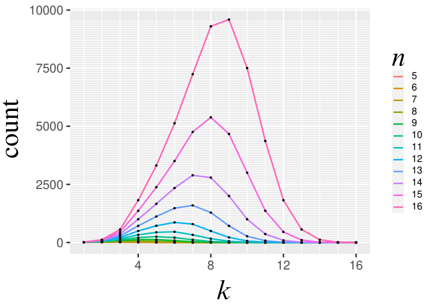
Number of k-term 2-pseudo progressions in $\{1, 2, \ldots, n\}$.

$k\downarrow n\rightarrow$	5	6	7	8	9	10	11	12	13	14	15	16
1	5	6	7	8	9	10	11	12	13	14	15	16
2	10	15	21	28	36	45	55	66	78	91	105	120
3	10	20	35	56	84	120	165	220	286	364	455	560
4	5	15	29	52	84	126	180	249	331	431	549	686
5	1	6	21	44	78	120	186	264	363	478	627	792
6	0	1	7	28	64	120	182	274	386	533	715	918
7	0	0	1	8	36	90	180	282	426	582	795	1060
8	0	0	0	1	9	45	123	264	433	672	919	1236
9	0	0	0	0	1	10	55	164	379	658	1057	1472
10	0	0	0	0	0	1	11	66	214	533	987	1654
11	0	0	0	0	0	0	1	12	78	274	735	1458
12	0	0	0	0	0	0	0	1	13	91	345	995
13	0	0	0	0	0	0	0	0	1	14	105	428
14	0	0	0	0	0	0	0	0	0	1	15	120
15	0	0	0	0	0	0	0	0	0	0	1	16
16	0	0	0	0	0	0	0	0	0	0	0	1



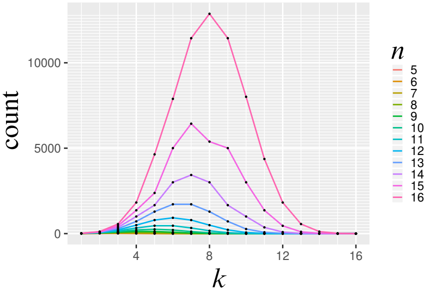
Number of k-term 3-pseudo progressions in $\{1, 2, \ldots, n\}$.

$k\downarrow n\rightarrow$	5	6	7	8	9	10	11	12	13	14	15	16
1	5	6	7	8	9	10	11	12	13	14	15	16
2	10	15	21	28	36	45	55	66	78	91	105	120
3	10	20	35	56	84	120	165	220	286	364	455	560
4	5	15	35	70	126	210	330	495	715	1001	1365	1820
5	1	6	21	56	126	252	438	720	1119	1666	2379	3312
6	0	1	7	28	84	210	462	864	1476	2343	3505	5128
7	0	0	1	8	36	120	330	792	1596	2892	4755	7240
8	0	0	0	1	9	45	165	495	1287	2793	5385	9300
9	0	0	0	0	1	10	55	220	715	2002	4669	9592
10	0	0	0	0	0	1	11	66	286	1001	3003	7504
11	0	0	0	0	0	0	1	12	78	364	1365	4368
12	0	0	0	0	0	0	0	1	13	91	455	1820
13	0	0	0	0	0	0	0	0	1	14	105	560
14	0	0	0	0	0	0	0	0	0	1	15	120
15	0	0	0	0	0	0	0	0	0	0	1	16
16	0	0	0	0	0	0	0	0	0	0	0	1



Number of k-term 4-pseudo progressions in $\{1, 2, \ldots, n\}$.

$k\downarrow n\rightarrow$	5	6	7	8	9	10	11	12	13	14	15	16
1	5	6	7	8	9	10	11	12	13	14	15	16
2	10	15	21	28	36	45	55	66	78	91	105	120
3	10	20	35	56	84	120	165	220	286	364	455	560
4	5	15	35	70	126	210	330	495	715	1001	1365	1820
5	1	6	21	56	126	252	462	792	1287	1666	2379	4368
6	0	1	7	28	84	210	462	924	1716	3003	5005	7888
7	0	0	1	8	36	120	330	792	1716	3432	6435	11440
8	0	0	0	1	9	45	165	495	1287	3003	5385	12870
9	0	0	0	0	1	10	55	220	715	1666	5005	11440
10	0	0	0	0	0	1	11	66	286	1001	3003	8008
11	0	0	0	0	0	0	1	12	78	364	1365	4368
12	0	0	0	0	0	0	0	1	13	91	455	1820
13	0	0	0	0	0	0	0	0	1	14	105	560
14	0	0	0	0	0	0	0	0	0	1	15	120
15	0	0	0	0	0	0	0	0	0	0	1	16
16	0	0	0	0	0	0	0	0	0	0	0	1



Acknowledgements

The authors would like to thank Kathryn Leonard and the Center for Undergraduate Research in Mathematics (CURM) for the amazing and helpful work they do. This project was supported by NSF grant DMS-1722563, and we are grateful to the NSF for funding CURM and supporting undergraduate research.

References

[Brown et al. 1999] T. C. Brown, R. L. Graham, and B. M. Landman, "On the set of common differences in van der Waerden's theorem on arithmetic progressions", *Canad. Math. Bull.* **42**:1 (1999), 25–36. MR Zbl

[Chang 2003] M. Chang, "Factorization in generalized arithmetic progressions and applications to the Erdős–Szemerédi sum-product problems", *Geom. Funct. Anal.* **13**:4 (2003), 720–736. MR Zbl

[Freiman 1999] G. A. Freiman, "Structure theory of set addition", pp. 1–33 in *Structure theory of set addition*, edited by J.-M. Deshouillers et al., Astérisque **258**, Société Mathématique de France, Paris, 1999. MR Zbl

[Green and Tao 2008] B. Green and T. Tao, "The primes contain arbitrarily long arithmetic progressions", *Ann. of Math.* (2) **167**:2 (2008), 481–547. MR Zbl

[Landman and Robertson 2004] B. M. Landman and A. Robertson, *Ramsey theory on the integers*, Student Mathematical Library **24**, American Mathematical Society, Providence, RI, 2004. MR Zbl

[van der Waerden 1927] B. L. van der Waerden, "Beweis einer Baudetschen Vermutung", *Nieux Archief* (2) **15** (1927), 212–216. Zbl

Received: 2019-06-07 Revised: 2020-07-24 Accepted: 2020-08-25

jay.cummings@csus.edu — *Department of Mathematics and Statistics,*
California State University, Sacramento, CA, United States

quintondarcy@csus.edu — *Department of Mathematics and Statistics,*
California State University, Sacramento, CA, United States

natalie.hobson@sonoma.edu — *Department of Mathematics and Statistics,*
Sonoma State University, Rohnert Park, CA, United States

hortond@sonoma.edu — *Department of Mathematics and Statistics,*
Sonoma State University, Rohnert Park, CA, United States

rhodewal@sonoma.edu — *Department of Mathematics and Statistics,*
Sonoma State University, Rohnert Park, CA, United States

morganthrockmorton@csus.edu — *Department of Mathematics and Statistics,*
California State University, Sacramento, CA, United States

ulmerstr@sonoma.edu — *Department of Mathematics and Statistics,*
California State University, Sacramento, CA, United States

mathematical sciences publishers msp

INVOLVE 13:5 (2020)
https://doi.org/10.2140/involve.2020.13.781

Growth series for graphs

Walter Liu and Richard Scott

(Communicated by Kenneth S. Berenhaut)

Given a graph Γ, one can associate a right-angled Coxeter group W and a cube complex Σ on which W acts. By identifying W with the vertex set of Σ, one obtains a growth series for W defined as $W(t)=\sum_{w\in W} t^{\ell(w)}$, where $\ell(w)$ denotes the minimum length of an edge path in Σ from the vertex 1 to the vertex w. The series $W(t)$ is known to be a rational function. We compute some examples and investigate the poles and zeros of this function.

1. Introduction

1.1. ***Preliminaries.*** Throughout this paper, graphs will be undirected, simple, and without loops. It is customary to think of a graph Γ as a collection of vertices with edges joining certain pairs of vertices; throughout the paper, an edge will be denoted by the *unordered* pair $\{u, v\}$ where u and v are the vertices that it joins. Given a graph Γ, we let $V(\Gamma)$ and $E(\Gamma)$ denote the vertex set and edge set, respectively.

A *simplicial complex* with vertex set V is a collection K of subsets of V such that if $\sigma \in K$ and $\tau \subseteq \sigma$, then $\tau \in K$. An element $\sigma \in K$ is called a *simplex*, and another element $\tau \in K$ is a *face of* σ if $\tau \subseteq \sigma$. Typically, the vertex set V is a collection of points in $\mathbb{R}^n$, in which case we can identify a simplex σ with its convex hull, obtaining a *geometric realization of* K as a subspace of $\mathbb{R}^n$. Because of this geometric connection, we define the *dimension of a simplex* σ to be $|\sigma|-1$, where $|\sigma|$ denotes the cardinality of σ. The *dimension* of a simplicial complex is the dimension of a largest simplex. A simplicial complex K is a *flag complex* if whenever all of the edges of a simplex σ are in K, then so is σ. Any graph Γ is a 1-dimensional simplicial complex (on its vertex set $V=V(\Gamma)$), but is a flag complex only if it has no complete subgraphs. Given any graph Γ, there is a unique flag simplicial complex K_Γ that has Γ as its 1-skeleton. It is obtained from Γ by adding subsets $\sigma \subseteq V(\Gamma)$ whenever the induced subgraph on σ is a complete

MSC2020: primary 20F55; secondary 51M20.
Keywords: growth series, right-angled Coxeter groups, graphs.
Liu was supported by the Panello Fund from the Department of Mathematics and Computer Science at Santa Clara University. Scott was partially supported by a Simon’s Foundation Collaboration Grant for Mathematicians.

subgraph (i.e., a *clique*). We shall call the complex K_Γ the *flag completion* of the graph Γ.

Another standard term we shall assume is that of the *link of simplex σ in K*. By definition, this is the subcomplex $\operatorname{Lk}(\sigma) \subseteq K$ defined by

$$\operatorname{Lk}(\sigma) = \{\tau \in K \mid \sigma \cup \tau \in K \text{ and } \sigma \cap \tau = \varnothing\}.$$

The link is useful because it records the local geometry of K in a neighborhood of σ.

1.2. ***Right-angled Coxeter groups and the Davis complex.*** Given a graph Γ, there is a corresponding *right-angled Coxeter group* W_Γ defined by a presentation with generating set $V(\Gamma)$ and relations

(1) $s^2 = 1$ for every $s \in V(\Gamma)$, and

(2) $st = ts$ for every edge $\{s, t\} \in E(\Gamma)$.

For any such group W_Γ, there is also a CAT(0)-cube complex Σ_Γ obtained from the Cayley graph for this presentation by "filling in cubes" (that is, when all of the edges of an n-dimensional cube are included, the interior of that cube is also included in the complex). The group W_Γ acts cellularly on Σ_Γ, the action is simply transitive on the vertices of Σ_Γ, and the link of every vertex of Σ_Γ is precisely the flag completion of the graph Γ. The cube complex Σ_Γ is called the *Davis complex* associated to the Coxeter group W_Γ. Further details and properties of W_Γ and Σ_Γ can be found in [Davis 2015].

1.3. ***Growth series.*** Given a graph Γ, let $K = K_\Gamma$ be its flag completion, let $W = W_\Gamma$ be the corresponding right-angled Coxeter group, and let $\Sigma = \Sigma_\Gamma$ be the corresponding Davis complex. Fix a vertex x_0 in Σ, and for any $w \in W$, let $\ell(w)$ denote the minimum number of edges in an edge-path connecting x_0 to $w \cdot x_0$ in Σ. We call $\ell(w)$ the *length* of w. Because the 1-skeleton of Σ coincides with the Cayley graph of W with respect to the generators $V(\Gamma)$, this length $\ell(w)$ is also the *word length of w* with respect to this generating set. We then define the *growth series for Γ (or for W)* to be the power series $W(t)$ defined by

$$W(t) = \sum_{w \in W} t^{\ell(w)}.$$

Given a simplicial complex K, we define the *f-polynomial of K* by

$$f(x) = \sum_{\sigma \in K} x^{|\sigma|}.$$

Note that the degree of the f-polynomial is one more than the dimension of K, and the constant term of $f(x)$ is 1 (corresponding to the empty simplex). Also note that $f(x)$ is a generating function in the sense that the coefficients of $f(x)$ record the number of simplices in K of each dimension. A related polynomial that captures the same data is the *h-polynomial*, defined by

$$h(t) = (1-t)^m f\left(\frac{t}{1-t}\right),$$

where m is one more than the dimension of K. The following proposition relates these polynomials to the growth series.

Proposition 1. *Let Γ be a graph and let $f(t)$ and $h(t)$ be the f-polynomial and h-polynomial of its flag completion K_Γ. Then*

$$W(t) = \frac{1}{f(-t/(1+t))},$$

and (*using the identity above*)

$$W(t) = \frac{(1+t)^m}{h(-t)}.$$

This formula is a special case of the known formula for the growth series of an arbitrary Coxeter group (not just right-angled). A reference for the latter is [Steinberg 1968, Theorem 1.25 and Corollary 1.29]. In [Scott 2007], the second author showed that this same formula holds not just for the Davis complex of a right-angled Coxeter group, but more generally, it holds for the growth series of *any* CAT(0) cube complex whose vertex links all have the same f-polynomial.

Remark 2. Note that Proposition 1 implies that the growth series of a graph is always (the Maclaurin series for) a rational function. Thus from the formula for $W(t)$ in terms of the h-polynomial, it follows that the radius of convergence of $W(t)$ is the minimal norm of a root of the h-polynomial.

Remark 3. Other investigation of the poles and zeroes of the h-polynomial have also appeared in the literature. A generalization of the Charney–Davis conjecture, predicting the sign of the Euler characteristic for flag triangulations of generalized homology spheres, is the real root conjecture (see [Gal 2005] for a clear statement, background, and counterexamples), which states, roughly, that the zeroes of the h-polynomial of a flag triangulation of a generalized homology sphere are all real.

1.4. ***The join of two simplicial complexes.*** Given simplicial complexes K_1 and K_2 with respective vertex sets V_1 and V_2, the *join of K_1 and K_2*, denoted by $K_1 * K_2$, is the simplicial complex with vertex set $V_1 \cup V_2$ and simplices of the form $\sigma_1 \cup \sigma_2$ for all $\sigma_1 \in K_1$ and $\sigma_2 \in K_2$.

It follows easily from the definition of the f-polynomial that if f_1 and f_2 denote the f-polynomials of K_1 and K_2, then the f-polynomial of the join $K_1 * K_2$ is given by

$$f(x) = f_1(x) f_2(x).$$

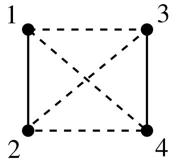


Figure 1. The join of two 1-simplices.

For the example in Figure 1, the complex K is the join of two 1-simplices, which is a 3-simplex. Since the f-polynomial of a 1-simplex is $(x+1)^2$, our formula says that the f-polynomial of K is

$$f(x) = (x+1)^2(x+1)^2 = (x+1)^4 = x^4 + 4x^3 + 6x^2 + 4x + 1.$$

2. Operations on graphs

Several standard graph theory arguments involve the contraction and deletion of edges (see, e.g., chromatic recursion in [West 1996, Theorem 5.3.6]). For flag simplicial complexes, these operations are a little more complicated.

2.1. ***Edge deletion.*** As the local geometry of an edge is fundamental to the way the edge's deletion affects the simplicial complex, it behooves us to encode that geometry in representing how the f-polynomial of the complex changes. We define the subset $L_e \subset K$ of simplices in K which have e as a face, including e itself. Then we define the *local f-polynomial* of an edge e, f_e, as

$$f_e(x) := \sum_{\sigma \in L_e} x^{\dim(\sigma)+1}$$

and the f-polynomial of K after the deletion of e, f_{K-e}, is just $f_K - f_e$. Note that the local f-polynomial of an edge can be calculated by multiplying the f-polynomial of the link of the edge by x^2, as each k-simplex in the link of an edge corresponds to a $(k+2)$-simplex containing that edge in the complex.

The example in Figure 2 shows what happens when K is the simplicial complex on the left and e is the top edge of the shaded 2-simplex. In this case,

$$f_K = x^3 + 5x^2 + 5x + 1,$$

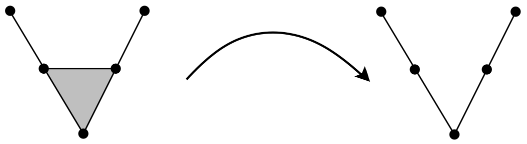


Figure 2. Edge-deletion.

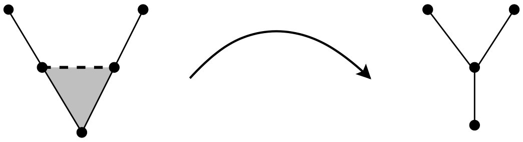


Figure 3. Edge-contraction.

and the link of e is the single vertex at the bottom of the 2-simplex (which has f-polynomial $(x+1)$), so the polynomial f_e is $x^2(x+1) = x^3 + x^2$. It follows that

$$f_{K-e} = f_K - f_e = 4x^2 + 5x + 1,$$

which is the f-polynomial of the (edge-deleted) complex on the right in Figure 2.

2.2. ***Edge contraction.*** The local f-polynomial of an edge can also be used to determine how the contraction of that edge affects the f-polynomial of the simplicial complex. To understand this it is useful to consider how contraction affects simplices with e as a face:

- contracting a 1-simplex removes it from the simplicial complex and identifies its endpoints, thereby removing one 0-simplex;
- contracting an edge of a 2-simplex removes the 2-simplex and identifies its other two edges, thereby removing one 1-simplex;
- contracting an edge of a 3-simplex removes the 3-simplex and identifies the two faces not adjacent to that edge, thereby removing one 2-simplex;

and so on. Thus, for each n-simplex in L_e, contracting e removes one n-simplex and one $(n-1)$-simplex, and therefore the f-polynomial of K after contracting e, $f_{K/e}(x)$, is

$$f_{K/e}(x) = f_K(x) - \Big(1 + \frac{1}{x}\Big) f_e(x) = f_K(x) - (x + x^2) f_{\mathrm{Lk}(e)}(x).$$

The example in Figure 3 shows what happens when the same edge from the previous example is contracted (instead of deleted). In this case, the f-polynomial for K is

$$f_K = x^3 + 5x^2 + 5x + 1,$$

and the f-polynomial for $\mathrm{Lk}(e)$ is $f_{\mathrm{Lk}(e)} = x + 1$, so our formula for the f-polynomial of K/e is

$$f_{K/e} = (x^3+5x^2+5x+1) - (x^2+x)(x+1) = 3x^2+4x+1.$$

2.3. ***Subdivision.*** The subdivision of a simplicial complex K on an edge e is the complex $\mathrm{Sd}_e(K)$ created by adding a new vertex at the midpoint of e and bisecting all simplices with e as a face. Subdividing a simplicial complex on an edge also

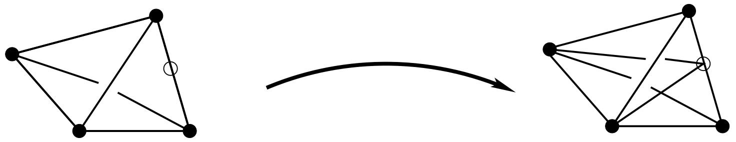


Figure 4. Edge-subdivision of a 3 simplex.

affects the f-polynomial in a way specific to the local geometry of the edge. After subdivision, each n-simplex $\sigma \subset K$ with $e \subseteq \sigma$ becomes two n-simplices bisected by an $(n-1)$-simplex.

Thus the f-polynomial after subdivision is

$$f_{\mathrm{Sd}(e)} = f_K(x) + \left(1 + \frac{1}{x}\right) f_e(x) = f_K(x) + (x + x^2) f_{\mathrm{Lk}(e)}(x).$$

The example in Figure 4 shows what happens when a 3-simplex is subdivided along one of its edges. In this case, K is the 3-simplex, whose f-polynomial is $f_K = (x+1)^4$, and e is the bisected edge, The link of e is the opposite edge, so $f_{\mathrm{Lk}(e)} = (x+1)$. Our formula gives

$$f_{\mathrm{Sd}(e)} = f_K + x(x+1) f_{\mathrm{Lk}(e)}(x) = (x+1)^4 + x(x+1)^3 = 2x^4 + 7x^3 + 9x^2 + 5x + 1.$$

Note the inverse relationship between edge subdivision and edge contraction: subdividing a complex on an edge and then contracting one of the halves of the subdivided edge returns the complex to its original state.

2.4. ***Cartesian graph product.*** Given graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, the Cartesian graph product of G_1 and G_2 is the graph $G = (V, E)$, where $V = V_1 \times V_2$ and

$$E = \bigl\{\{(u_1, u_2), (v_1, v_2)\} \bigm| (u_1 = v_1 \wedge \{u_2, v_2\} \in E_2) \vee (\{u_1, v_1\} \in E_1) \wedge (u_2 = v_2)\bigr\},$$

with $u_1, v_1 \in V_1$ and $u_2, v_2 \in V_2$. Let the f-polynomial of graph A be

$$f_A(x) = 1 + \sum_{i \in \mathbb{Z}_{\geq 0}} a_i x^{i+1}, \quad a_i \in \mathbb{Z}_{\geq 0},$$

and the f-polynomial of graph B be

$$f_B(x) = 1 + \sum_{i \in \mathbb{Z}_{\geq 0}} b_i x^{i+1}, \quad b_i \in \mathbb{Z}_{\geq 0}.$$

The Cartesian graph product $A \times B$ then has f-polynomial

$$f_{AB}(x) = b_0 f_A(x) + a_0 f_B(x) - (a_0 b_0 x + a_0 + b_0 - 1).$$

Some intuition for this formula comes from the fact that the product $A \times B$ contains a copy of A for each vertex of B and a copy of B for each vertex of A. The subtracted term corresponds to double-counted edges.

3. Examples and computations

3.1. ***Trees.*** A connected graph with n vertices and $n-1$ edges is a tree. The f-polynomial of such a tree is

$$f(x) = 1 + nx + (n-1)x^2 = (1 + (n-1)x)(1+x).$$

The h-polynomial is then

$$h(t) = 1 + (n-2)t,$$

the root of which is $1/(2-n)$. It follows (from Remark 2) that for any infinite family of trees, the radius of convergence of the growth series can be arbitrarily close to zero.

3.2. ***Cycles.*** Cycles are a basic family of graphs and a straightforward example for f- and h-polynomial computation. The k-cycle has k vertices and k edges; thus the f-polynomial for the k-cycle with $k > 3$ is $f(x) = 1 + kx + kx^2$, which has roots

$$x = \frac{-k \pm \sqrt{k^2 - 4k}}{2k}.$$

As $k \to \infty$, these roots approach 0 and -1. We can find the h-polynomial by substituting $x = t/(1-t)$ as follows:

$$\begin{aligned} f\left(\frac{t}{1-t}\right) &= 1 + \frac{kt}{1-t} + \frac{kt^2}{(1-t)^2} = \left(\frac{1}{(1-t)^2}\right)((1-t)^2 + kt(1-t) + kt^2) \\ &= \left(\frac{1}{(1-t)^2}\right)(1 - 2t + t^2 + kt - kt^2 + kt^2), \end{aligned}$$

so

$$h(t) = 1 + (k-2)t + t^2.$$

The roots now are

$$t = \frac{2 - k \pm \sqrt{(k-2)^2 - 4}}{2} = \frac{2 - k \pm \sqrt{k^2 - 4k}}{2},$$

which approach 1 and $-\infty$.

3.3. ***Paws.*** Define the *paw* as the simplicial complex constructed by attaching an edge to a vertex of a 2-simplex. More generally, a *k-paw* is a 2-simplex with a path of length k (i.e., a path with k edges) attached to a vertex of this 2-simplex (see Figure 5).

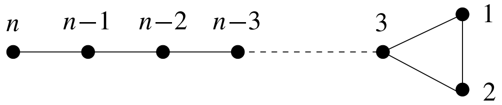


Figure 5. A k-paw with $k = n - 3$.

The f-polynomial for the k-paw is $f(x) = 1 + (k+3)x + (k+3)x^2 + x^3$. We can then find the roots: $f(x)$ factors neatly into $(1 + x)(1 + (k+2)x + x^2)$, and the quadratic roots are

$$x = \frac{-k-2 \pm \sqrt{(k+2)^2 - 4}}{2} = \frac{-k-2 \pm \sqrt{k^2+4k}}{2}$$

with limits $\lim_{k\to\infty} x = -1, -\infty$

Likewise, we can derive the h-polynomial using substitution:

$$\begin{aligned} f\left(\frac{t}{1-t}\right) &= 1 + (k+3)\left(\frac{t}{1-t}\right) + (k+3)\left(\frac{t}{1-t}\right)^2 + \left(\frac{t}{1-t}\right)^3 \\ &= \left(\frac{1}{1-t}\right)^3 \big((1-t)^3 + (k+3)t(1-t)^2 + (k+3)t^2(1-t) + t^3\big), \end{aligned}$$

so

$$\begin{aligned} h(t) &= 1-3t+3t^2-t^3+(k+3)t-2(k+3)t^2+(k+3)t^3+(k+3)t^2-(k+3)t^3+t^3 \\ &= 1+kt-kt^2, \end{aligned}$$

with roots

$$t = 1, \frac{2-k \pm \sqrt{k^2-4k}}{2}$$

and limits $\lim_{k\to\infty} x = 1, -\infty$.

3.4. ***Cycles with extra diagonals.*** Other families of graphs can be obtained by adding additional diagonals to a cycle. For example, Figure 6 shows a 12-cycle with additional edges connecting vertex i to vertex j whenever $i - j \equiv 2$ or $3 \bmod 12$. This complex has f-polynomial $f(x) = 1 + 12x + 36x^2 + 36x^3 + 12x^4$, with roots $x \approx -0.414, -0.124, -1.231 \pm 0.330i$. Substitution produces the h-polynomial $h(t) = 1 + 8t + 6t^2 - 4t^3 + t^4$, with roots $t \approx -0.706, -0.141, 2.424 \pm 2.033i$. Notably, both the f-polynomial and the h-polynomial for this graph have complex roots.

3.5. ***Possible roots of f- and h-polynomials.*** As f-polynomials have restrictions on their coefficients, it is possible to find restrictions on their possible roots. The first of these is provided by Descartes' rule of signs, which says that the maximum number of positive real roots of a polynomial is less than or equal to the number of sign changes between consecutive coefficients. Since f-polynomial coefficients

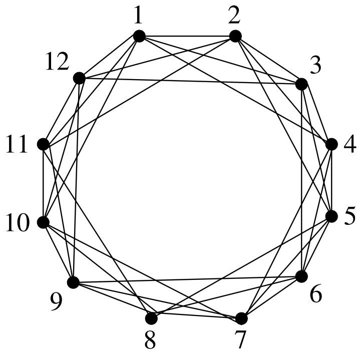


Figure 6. 12-gon with extra diagonals.

count simplices in a complex, they are always nonnegative; thus all roots of f-polynomials must be nonpositive. Due to the fact that, by definition, any simplicial complex contains exactly one -1-simplex, f-polynomials always have 1 as a constant term, and therefore 0 is not a possible root of an f-polynomial. Thus all roots of f-polynomials are negative.

Proposition 4. *The only possible integer root of the h-polynomial of a simplicial complex is* 1.

Proof. Assume the h-polynomial of a simplicial complex K has an integer root $a \in \mathbb{Z}$. Then $x = a/(1-a)$ is a root of the f-polynomial. Note that if a is an integer $\neq 1$, then $x = a/(1-a)$ is a reduced fraction. Since the f-polynomial has integer coefficients, the rational root theorem states that for any rational root of the f-polynomial, written in lowest terms p/q, p is an integer divisor of the constant term of the f-polynomial. Since $p = a$, we know a is an integer divisor of 1, so a must be 1. □

For most of the examples above, the radius of convergence of the growth series $W(t)$ (or, equivalently, the smallest norm of a root of the h-polynomial) is ≤ 1. This actually holds more generally. (The authors thank Michael Hartglass of the Department of Mathematics and Computer Science at Santa Clara University for one of the key ideas in the proof of the following theorem). The key technical tool in the proof is the use of the root test to bound the radius of convergence for a power series.

Theorem 5. *Let K be a flag simplicial complex, let W be the corresponding right-angled Coxeter group, and let Σ be the Davis complex. Assume further that K is not a single simplex. Then the h-polynomial of K has a real root in the interval* $(-1, 0)$.

Proof. We use the second formula from Proposition 1

$$W(t) = 1 + a_1 t + a_2 t^2 + \cdots + a_n t^n + \cdots = \frac{(1+t)^m}{h(-t)},$$

where a_n denotes the number of vertices in Σ that are (edge-path-) distance n from x_0. Since this series is a rational function defined in a neighborhood of 0, we know that the radius of convergence coincides with both the norm of the smallest root of $h(-t)$ and $1/\lambda$ where $\lambda = \limsup_n |a_n|^{1/n}$. Since K is not a single simplex, Σ is unbounded, so the sequence $a_1, a_2, \ldots$ has infinitely many terms ≥ 1. Hence $\lambda \geq 1$. It follows that $h(-t)$ has a root in the interval $(0, 1/\lambda) \subseteq (0, 1)$, and hence $h(t)$ has a root in the interval $(-1, 0)$. □

References

[Davis 2015] M. W. Davis, "The geometry and topology of Coxeter groups", pp. 129–142 in *Introduction to modern mathematics*, edited by S.-Y. Cheng et al., Adv. Lect. Math. (ALM) **33**, International Press, Somerville, MA, 2015. MR Zbl

[Gal 2005] S. R. Gal, "Real root conjecture fails for five- and higher-dimensional spheres", *Discrete Comput. Geom.* **34**:2 (2005), 269–284. MR Zbl

[Scott 2007] R. Scott, "Growth series for vertex-regular CAT(0) cube complexes", *Algebr. Geom. Topol.* **7** (2007), 285–300. MR Zbl

[Steinberg 1968] R. Steinberg, *Endomorphisms of linear algebraic groups*, Memoirs of the American Mathematical Society **80**, Amer. Math. Soc., Providence, RI, 1968. MR Zbl

[West 1996] D. B. West, *Introduction to graph theory*, Prentice Hall, Upper Saddle River, NJ, 1996. MR Zbl

Received: 2019-07-17 Revised: 2020-07-12 Accepted: 2020-08-11

walterbliu@gmail.com
Department of Mathematics and Computer Science,
Santa Clara University, Santa Clara, CA, United States

rscott@scu.edu
Department of Mathematics and Computer Science,
Santa Clara University, Santa Clara, CA, United States

INVOLVE 13:5 (2020)
https://doi.org/10.2140/involve.2020.13.791

Peg solitaire in three colors on graphs

Tara C. Davis, Alexxis De Lamere, Gustavo Sopena,
Roberto C. Soto, Sonali Vyas and Melissa Wong

(Communicated by Joseph A. Gallian)

Peg solitaire is a classical one-person game that has been played in various countries on different types of boards. Numerous studies have focused on the solvability of the games on these traditional boards and more recently on mathematical graphs. In this paper, we go beyond traditional peg solitaire and explore the solvability on graphs with pegs of more than one color and arrive at results that differ from previous works on the subject. This paper focuses on classifying the solvability of peg solitaire in three colors on several different types of common mathematical graphs, including the path, complete bipartite, and star. We also consider the solvability of peg solitaire on the Cartesian products of graphs.

1. Introduction

Peg solitaire has been played in cultures across the world for over 300 years. Boards can come in a variety of shapes and sizes, each with a different number of pegs. The standard rules for a game of peg solitaire state that a board starts with a certain number of pegs and one hole placed anywhere on the board, usually the center. The player must have an ending state in which only one peg is remaining on the board by removing all other pegs achieved by jumping adjacent pegs over one another. There is a rich history of problems posed and solved on different types of boards [Beasley 1985] and recently there has been renewed interest in the subject [Bell 2007; 2008].

In the last few years Robert A. Beeler and his students have studied peg solitaire on graphs by modifying the rules as follows [Beeler et al. 2017; Beeler and Paul Hoilman 2011; 2012; Beeler and Walvoort 2015]: suppose there is an edge connecting vertices v_1 and v_2 and a second edge connecting vertices v_2 and v_3 with pegs in vertices v_2 and v_3 (see Figure 1). Then a player may jump the peg in v_3 over the peg in v_2 to obtain the result in Figure 1. Following [Beeler and Walvoort 2015], we denote such a jump by $v_3 \cdot \vec{v}_2 \cdot v_1$. Once the player jumps one peg over the other, as seen in Figure 1, the peg that was jumped over is then removed. Although

MSC2010: primary 05C57; secondary 91A43.
Keywords: peg solitaire, combinatorial games, games on graphs.

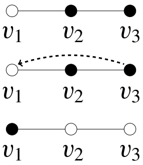


Figure 1. Example of a move in peg solitaire.

others have created and analyzed different variations of peg solitaire on graphs [Beeler and Rodriguez 2012; Engbers and Weber 2018; Engbers and Stocker 2015; Loeb and Wise 2015], the goal of this paper is to generalize the results in [Beeler and Paul Hoilman 2011] by adding an additional, "third," color to the game. We recognize that although the pegs in our version of the game come in two colors, it might be best to think of the "third" color as the white holes, since, as explained in Section 2, our version of the game is closely aligned with arithmetic in $\mathbb{Z}_3$. In Section 2 we present the rules of peg solitaire on graphs with pegs with different colors, in Section 3 we give results on different graphs with these new rules, and in Section 4 we discuss some open questions.

2. Three-color peg solitaire

As described in [Beeler and Paul Hoilman 2011], each game consists of a graph, $G = (V, E)$, with $|V| \geq 2$, that serves as the board, and a starting state S depicting the initial placement of the pegs. As in that paper, we also assume that all graphs are finite and undirected with no loops or multiple edges and that graphs are always connected. In this version of peg solitaire we include two different types of pegs and adjust the criteria for what moves can be executed. The starting state of each game includes the following components:

- a singleton set S_0 consisting of the vertex that has a peg of color 0, also called the initial hole in the board;
- a set S_1 which consists of vertices that have pegs of color 1; and
- a set S_2 which is made up of vertices that have pegs of color 2.

Thus a starting state S will be denoted as $S = (S_0, S_1, S_2)$

As in [Beeler and Paul Hoilman 2011] we note that a player can only execute a jump $x \cdot \vec{y} \cdot z$ if $xy, yz \in E$ and if z has color 0, while x and y do not. A difference in the rules is that when a player jumps a peg with the same color, the peg that has been jumped over then switches to the second color (see Figure 2). Moreover, when adjacent pegs have different colors, then the peg that jumps over the different colored peg creates a hole (see Figure 3).

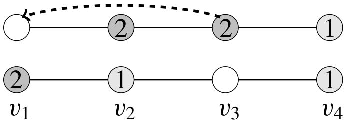


Figure 2. Example of moving a peg over a peg of the same color.

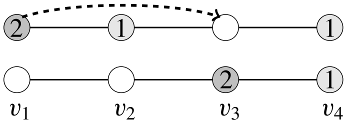


Figure 3. Example of moving a peg over a peg of a different color.

Moreover we note that we can translate the color rules in terms of modular arithmetic as follows: If we allow the darker gray color in Figure 2 to correspond to the number 2 and the lighter gray color to correspond to the number 1, then when a peg jumps over another peg into a hole, the middle peg is replaced with the result of an addition modulo 3. This generalizes the rules in [Beeler and Paul Hoilman 2011], since in that paper all pegs have color 1, and the result of any jump is $1+1=0 \bmod 2$, a hole. Using modular arithmetic we obtain the following notation for Figure 2:

$$\begin{matrix} 0 & 2 & 2 & 1 \\ 2 & 1 & 0 & 1. \end{matrix}$$

In other words, when the peg in vertex v_3 jumped over the peg in vertex v_2 we add the values of the pegs in v_3 and v_2 modulo 3 and obtain the number 1 in vertex v_2. This corresponds to the move in Figure 2. The new notation would express the move in Figure 3 as

$$\begin{matrix} 2 & 1 & 0 & 1 \\ 0 & 0 & 2 & 1. \end{matrix}$$

Note that there is a game that is closely related to the one in the previous paragraph and which can be played in a similar manner, namely the game

$$0 \quad 1 \quad 1 \quad 2,$$

where each color in the original game is replaced with the opposite[1] color. Following the same two moves highlighted above we obtain the following configurations of the game:

$$\begin{matrix} 0 & 1 & 1 & 2 \\ 1 & 2 & 0 & 2 \\ 0 & 0 & 1 & 2. \end{matrix}$$

[1] The opposite color refers to the fact that 1 and 2 are additive inverses in $\mathbb{Z}_3$.

Thus if Γ is a game on a graph G with starting state $S_1 = \{u_1, u_2, \dots, u_m\}$, $S_2 = \{v_1, v_2, \dots, v_n\}$, and $S_0 = \{v_o\}$, then the game Γ' given by $S_1' = \{v_1, v_2, \dots, v_n\}$, $S_2' = \{u_1, u_2, \dots, u_m\}$, and $S_0 = \{v_o\}$ is called the *opposite* of Γ. As in the example above, the properties of the addition table for $\mathbb{Z}_3$ imply that any move made on a game in three colors will be mirrored in the opposite game. This observation allows us to reduce later arguments in the sense that the actual colors being used in the game do not matter, only whether the peg that jumps is of the same or a different color than the one that is being jumped over.

Similar to the traditional game of peg solitaire, the goal of this version is to have one remaining peg, regardless of the color, on the board at the end of the game. The game is in its terminal state when one of two things happen. Firstly, the game is in its terminal state if the goal has been reached and there is only one peg remaining on the board. In the second option, the game is in its terminal state if the board has reached a point where there are no remaining moves to solve the game. For example, there may be no allowable moves remaining (see Proposition 3.4). Otherwise, there may be allowable moves creating a repeating loop in which no additional pegs are removed (see Theorem 3.7).

We define the *terminal state* $T = (T_0, T_1, T_2)$ of each game as follows:

- a set T_0 containing the vertices with remaining pegs of color 0, or holes;
- a set T_1 that consists of the vertices with remaining pegs of color 1; and
- a set T_2 that consists of the vertices with remaining pegs of color 2.

Thus the goal of playing each game is to reach a terminal state T, where $|T_1 \cup T_2| = 1$. We say that a game on a graph, i.e., a graph G together with a starting state S, has been *won* if there is a sequence of moves leading to a terminal state with $|T_1 \cup T_2| = 1$. We say that a game on a graph is a *losing game* if it is not a winning game: no sequence of moves leads to a terminal state with $|T_1 \cup T_2| = 1$. This leads us to the following definitions, where we intuitively think that a game is solvable if we can win at least one game; and freely solvable if we can win every game.

Definition 2.1 (solvable). A graph $G = (V, E)$ is solvable if there exists a starting state $S = (S_0, S_1, S_2)$ that has an associated terminal state $T = (T_0, T_1, T_2)$ such that $|T_1 \cup T_2| = 1$.

Definition 2.2 (freely solvable). A graph $G = (V, E)$ is freely solvable if for every starting state $S = (S_0, S_1, S_2)$, there exists an associated terminal state $T = (T_0, T_1, T_2)$ such that $|T_1 \cup T_2| = 1$.

Definition 2.3 (k-solvable). A graph $G = (V, E)$ is k-solvable for $k \in \mathbb{N}$ when k is the minimal value such that there exists a starting state $S = (S_0, S_1, S_2)$, with associated terminal state $T = (T_0, T_1, T_2)$ such that $|T_1 \cup T_2| = k$, and all vertices in $T_1 \cup T_2$ are nonadjacent.

We finish this section by noting our first result following from the discussion thus far.

Lemma 2.4. *Let G be a graph and let Γ be a game on G. Then Γ is a winning game if and only if its opposite is a wining game.*

In the next section we determine the solvability of many of the same graphs considered in [Beeler and Paul Hoilman 2011].

3. Results

3A. ***Games on path and cyclic graphs.*** We begin by showing that path graphs with n vertices, $n \geq 3$, are solvable but not freely solvable. We remark that this result provides a distinction between the results occurring in two [Beeler and Paul Hoilman 2011] versus three colors. We further note that while, at this point, we do not have an example of a graph that is solvable in two colors yet not solvable in three colors, there are individual games where this is the case, such as in P_4. See Theorem 2.3 in [Beeler and Paul Hoilman 2011] for these examples.

A path graph P_n has n vertices and $n-1$ edges connecting each of the vertices along a line. The cycle graph C_n is the same as the path graph with one additional edge connecting the first and last vertex. For more on path graphs and other basic graph theory terminology, refer to [West 2001]. In the proofs below we label the vertices as in Figure 4.

Theorem 3.1. *The path graph on n vertices, P_n, is solvable for $n \geq 2$. Moreover, P_n is not freely solvable for $n \geq 3$.*

Proof. We begin by noting that we can win every game on the path graph P_2; thus P_2 is freely solvable. We now show that P_n is solvable for all $n \geq 3$, proceeding by induction on the number of vertices. Suppose that we can win the game, Γ, with starting state $S_0 = \{v_1\}$, $S_1 = \{v_2, \dots, v_n\}$, and $S_2 = \varnothing$; i.e., we can win the game

$$0\ 1\ 1\ 1\ 1\ 1\ \cdots\ 1.$$

Consider a similar game in P_{n+1}, denoted by Γ', with starting state $S'_0 = \{v_1\}$, $S'_1 = \{v_2, \dots, v_n, v_{n+1}\}$, and $S'_2 = \varnothing$,

$$0\ 1\ 1\ 1\ 1\ 1\ \cdots\ 1\ 1.$$

Then by moving $v_3 \cdot \vec{v}_2 \cdot v_1$ and then back $v_1 \cdot \vec{v}_2 \cdot v_3$, we see that we now have holes in vertices v_1 and v_2 and pegs of the color 1 in vertices $v_3, \dots, v_{n+1}$. Now if

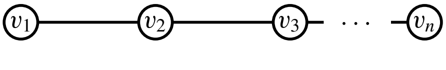


Figure 4. A path graph with n vertices.

we consider the vertices $v_2, \dots, v_{n+1}$ on their own, we see that their configuration matches the starting state of the game Γ. We can play the rest of the game Γ' as we would the game Γ and win the game. Thus P_n is solvable for all n.

We can show that P_n is not freely solvable for $n \geq 3$ in a similar manner. We begin by noting that we cannot win the game 1 0 2 in P_3. Now suppose $n = 2k + 1$ for some integer k, and consider the game Γ in P_n with starting state given by $S_0 = \{v_1\}$, $S_1 = \{v_2, v_4, \dots, v_{2k}\}$, and $S_2 = \{v_3, v_5, \dots, v_{2k+1}\}$, i.e.,

$$0\ 1\ 2\ 1\ 2\ 1\ \cdots\ 1\ 2.$$

Note that we only have one move, $v_3 \cdot \vec{v}_2 \cdot v_1$, at our disposal and we are left with

$$2\ 0\ 0\ 1\ 2\ 1\ \cdots\ 1\ 2.$$

We are again left with one possible move, $v_5 \cdot \vec{v}_4 \cdot v_3$, and now obtain

$$2\ 0\ 2\ 0\ 0\ 1\ \cdots\ 1\ 2.$$

We continue in this manner, always forced to make one move, $v_{2i+1} \cdot \vec{v}_{2i} \cdot v_{2i-1}$, as i goes from 3 to k so that our terminal state becomes $T_0 = \{v_2, v_4, \dots, v_{2k}, v_{2k+1}\}$, $T_1 = \varnothing$, and $T_2 = \{v_1, v_3, \dots, v_{2k-1}\}$, i.e.,

$$2\ 0\ 2\ 0\ 2\ 0\ 2\ \cdots\ 2\ 0\ 0.$$

Thus $|T_1 \cup T_2| \neq 1$ and since this is the only way to play this particular game we conclude that P_n is not freely solvable for n odd.

Now consider a game in P_{2k}, $k \geq 2$, with the starting state $S_0 = \{v_1\}$, $S_1 = \{v_2, v_4, \dots, v_{2k}\}$, $S_2 = \{v_3, v_5, \dots, v_{2k-1}\}$, given by

$$0\ 1\ 2\ 1\ 2\ 1\ \cdots\ 2\ 1.$$

Then just as before we note that we are forced to make one move each time we try to play, namely, $v_{2i+1} \cdot \vec{v}_{2i} \cdot v_{2i-1}$, as i goes from 1 to $k - 1$. Our terminal state is similar to the one above: $T_0 = \{v_2, v_4, \dots, v_{2k-2}, v_{2k-1}\}$, $T_1 = \{v_{2k}\}$, and $T_2 = \{v_1, v_3, \dots, v_{2k-3}\}$, illustrated as

$$2\ 0\ 2\ 0\ 2\ 0\ 2\ \cdots\ 2\ 0\ 1.$$

Thus, $|T_1 \cup T_2| \neq 1$ and so P_n is not freely solvable when n is even. □

The following result follows immediately from Theorem 3.1 and the fact that P_n is a spanning subgraph of C_n.

Corollary 3.2. *The cyclic graph on n vertices, C_n, is solvable for $n \geq 2$.*

The following lemma will be useful in a couple of proofs, both in this paper and a subsequent paper, so we state it here. Moreover, this lemma introduces an interesting phenomenon with cycle graphs alluded to below.

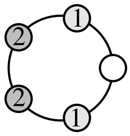


Figure 5. A game that cannot be won in C_5.

Lemma 3.3. *The cycle graph on three vertices, C_3, is freely solvable.*

Proof. Let v_1, v_2, and v_3 be the three vertices of the graph C_3 such that there are edges between v_1 and v_2, v_2 and v_3, and v_3 and v_1. Without loss of generality, suppose that $S_0 = \{v_1\}$. Then we have two cases to consider.

- Case 1: The two pegs on the graph, in vertices v_2 and v_3, are different colors. Then by $v_3 \cdot \vec{v}_2 \cdot v_1$ we win this game.
- Case 2: The two pegs on the graph, in vertices v_2 and v_3, are the same color. Then by $v_3 \cdot \vec{v}_2 \cdot v_1$ we obtain a game equivalent to the game in Case 1. □

At this time we cannot determine if C_n is freely solvable for all $n \geq 2$, but playing each game in $C_2, C_3, C_4, \ldots, C_{11}$, with the aid of a computer program [Sopena 2019] we have determined that each of these graphs are freely solvable with the exception of C_5. The game in Figure 5 is a game that cannot be won in C_5, which we prove below.

Proposition 3.4. *The game Γ with starting state $S_0 = \{v_1\}$, $S_1 = \{v_2, v_5\}$, and $S_2 = \{v_3, v_4\}$, or any equivalent game, cannot be won.*

Proof. Note that our first move is either $v_3 \cdot \vec{v}_2 \cdot v_1$ or $v_4 \cdot \vec{v}_5 \cdot v_1$. Regardless, we create a situation where there is a vertex (either v_2 or v_5) that has a peg of one color, while the vertices adjacent to it have pegs of a different color. Moreover, the other two vertices have holes. Again, there is a choice of two moves but either one leads to a terminal state with two pegs. □

3B. ***Games on complete bipartite graphs.*** In this section we will prove a result that complements Theorem 2.7 of [Beeler and Paul Hoilman 2011], namely that the complete bipartite graph is freely solvable. The complete bipartite graph, $K_{m,n}$, has as its vertex set the union of disjoint sets of vertices $X = \{x_1, \ldots, x_m\}$ and $Y = \{y_1, \ldots, y_n\}$. The set of edges E connects all vertices from X to all vertices in Y, but none from X to other vertices in X and none from Y to other vertices in Y.

Theorem 3.5. *The complete bipartite graph $K_{m,n}$ is freely solvable for all $m, n > 1$.*

Proof. Without loss of generality, let $S_0 = \{x_1\}$. The first step is to remove the peg in y_1 by $x_2 \cdot \vec{y}_1 \cdot x_1$. If the pegs in x_2 and y_1 are of the same color, we add a second move of jumping back $x_1 \cdot \vec{y}_1 \cdot x_2$. After relabeling, we now have holes in x_1 and y_1.

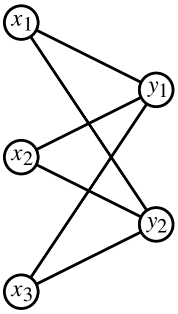


Figure 6. An example of a complete bipartite graph, $K_{3,2}$.

We repeat this action $n-2$ more times to remove all but one peg in Y until we have holes in exactly the vertices $y_1, \dots, y_{n-1}$ and x_1.

Now we will remove the peg in x_2 by jumping $y_n \cdot \vec{x}_2 \cdot y_{n-1}$. If the pegs in x_2 and y_n are of the same color, we add a second move of jumping $y_{n-1} \cdot \vec{x}_2 \cdot y_n$. After relabeling, we now have holes in x_1, x_2, and $y_1, \dots, y_{n-1}$. We repeat this action $m-2$ more times to remove all pegs in X until we have holes in all vertices except y_n. □

Corollary 3.6. *The complete graph, K_n, is freely solvable for $n > 1$.*

Proof. When $n = 2$, the graph is P_2, which is clearly freely solvable. When $n = 3$, the graph is C_3, which is freely solvable by Lemma 3.3. For $n > 3$, the complete bipartite graph is a spanning subgraph of the complete graph, so the result follows from Theorem 3.5. □

Theorem 3.7. *The star graph $K_{1,n}$ is $(n-1)$-solvable for $n > 2$.*

Proof. Note that for $n = 1$, the graph is P_2 and so is freely solvable, and for $n = 2$ the graph is P_3, which is also solvable. We remind the reader that the graph $K_{1,n}$ has vertex set $\{x_1, y_1, y_2, \dots, y_n\}$ with an edge between x_1 and y_i for all $1 \leq i \leq n$, and no other edges.

If $S_0 = \{x_1\}$, then there are no possible moves. Otherwise, note that a peg in x_1 is the only one that can possibly be removed from the graph. Suppose without loss of generality that $S_0 = \{y_n\}$ and that the first jump is $y_1 \cdot \vec{x}_1 \cdot y_n$. If the color of y_1 is the same as x_1, then the peg in x_1 has not been removed. Therefore, we should assume that the color of y_1 is different than x_1, in which case the move creates a hole in x_1 and a hole in y_1. There are no possible moves after this. The terminal state has $|T_1 \cup T_2| = n-1$. □

3C. *Cartesian products of graphs.* In this section we show that if we have a graph G that is solvable, and H is any graph, then the Cartesian product of the two graphs, G and H, is also solvable. If $V(G)$ is the set of vertices a graph G and $V(H)$ is the set of vertices of graph H, then the set of vertices of the Cartesian

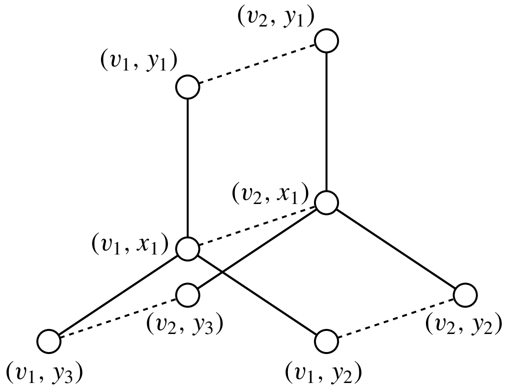


Figure 7. An example of the Cartesian product of P_2 and $K_{1,3}$.

product $G \square H$ is given by $V(G \square H) = \{(u, v) \mid u \in V(G),\ v \in V(H)\}$. Moreover, as in [Beeler and Paul Hoilman 2011], we define G_u to be the copy of G induced by $u \in V(H)$. Note that Figure 7 shows us the Cartesian product $P_2 \square K_{1,3}$ and that there are new edges introduced between both copies of $K_{1,3}$.

We begin by making a remark about a strategy that we have used before (see proof of Theorem 3.1) and will be used heavily hereafter.

Remark 3.8 (there and back strategy). Let x, y and z be three vertices with edges xy and yz. Suppose that a hole exists in either vertex x or z and that the other two vertices contain pegs of the same color. Without loss of generality suppose the hole is in the vertex z. Then we remove the peg in vertex y by $x \cdot \vec{y} \cdot z$ followed by $z \cdot \vec{y} \cdot x$. We shall denote this "there and back move" by $z \cdot \overset{\leftrightarrows}{y} \cdot x$.

Theorem 3.9. *Let G be a solvable graph and let T be a tree. Then $G \square T$ is solvable.*

Proof. Let $V(G)$ be the vertices of G and $V(T)$ be the vertices of T. Furthermore, let u_r be the root of T, suppose that the height of T is n, and let U_i be the set of all descendants of u_r at height i for $i \in \{1, \dots, n\}$. Consider a winning game, Γ, in G with starting state (S_0, S_1, S_2) and associated winning terminal state (T_0, T_1, T_2). Define the following starting state (S'_0, S'_1, S'_2) for $G \square T$:

- $S'_0 = \{(v_0, u_r)\}$ for $S_0 = \{v_0\}$.
- S'_1 contains all of the vertices (v, u) such that $v \in S_1$ and $u = u_r$ or $u \in U_i$ for $i \in \{1, 2, \dots, n\}$.
- S'_2 contains all of the vertices (v, u) such that $v \in S_2$ and $u = u_r$ or $u \in U_i$ for $i \in \{1, 2, \dots, n\}$.

Note that the description above has placed a copy of the winning game Γ in each copy of G. We now fill in the vertices that are copies of the hole in vertex (v_0, u_r).

We choose and fix a vertex in $V(G)$ that is adjacent to v_0, call it v', and note the color of the peg in v'. Without loss of generality suppose that v' is in S_1. Let $(v_0, u) \in S_1'$ for all $u \in U_i$ with $i \in \{1, \dots, n\}$. In other words, all of the vertices in $\{v_0\} \square T$ that are not (v_0, u_r) have a peg of the same color as the peg in v'.

The proof proceeds in three steps, the first of which removes pegs that are in vertices (v_0, u) with $u \in U_i$ with $i \in \{1, \dots, n\}$ so that we have holes in all of the vertices in $\{v_0\} \square T$. The second step wins the game Γ now found in each copy of T. The final step eliminates the remaining pegs except for one.

We begin the first step by inductively using Remark 3.8. We take the pegs in vertices $(v', u) \in G \square U_1$ to remove all pegs in vertices $(v_0, u) \in G \square U_1$. In other words, we are removing the pegs in vertices $(v_0, u) \in G \square U_1$ by $(v', u) \cdot \overleftrightarrow{(v_0, u)} \cdot (v_0, u_r)$ for every $u \in U_1$. We now inductively use the vertices $(v', u) \in G \square U_2$ to remove all pegs in vertices $(v_0, u) \in G \square U_2$ via $(v', u) \cdot \overleftrightarrow{(v_0, u)} \cdot (v_0, u_r)$ and proceed until all of the vertices $(v_0, u) \in G \square T$ have holes. Thus there is a copy of Γ in each copy of G given by $G \square \{u\}$ for $u \in T$.

The second step follows easily as it involves winning the game Γ on each copy of G on T. This leads to a peg on one vertex for each copy of G on T, say (t, u) for some $t \in G$ and for every $u \in T$. Again we proceed inductively but this time from the "bottom up". We begin by using the pegs in vertices (t, u) for all $u \in U_{n-1}$ to remove the pegs in vertices (t, w) for all $w \in U_n$ by $(t, u) \cdot \overleftrightarrow{(t, w)} \cdot (v, w)$ for some $v \in G$ adjacent to $t \in G$. We then repeat this step using the pegs in vertices (t, u) for all $u \in U_{n-2}$ to remove the pegs in (t, w) for all $w \in U_{n-1}$ in a similar manner until we reach the top of the tree. That is, after this process the last peg is in vertex (t, u_r). □

Corollary 3.10. *Let G be a solvable graph and let H be any graph. Then $G \square H$ is solvable.*

Proof. Every graph has a spanning tree T. Now we use the spanning tree of H and Theorem 3.9. □

We leave the proof of the following corollary to the reader.

Corollary 3.11. *Let G be a solvable graph and T a tree. Then $G \square T^n$ is solvable for all $n \geq 1$.*

4. Discussion

We conclude this paper with some open questions related to this project. Following [Beeler et al. 2017; Beeler and Walvoort 2015] it is natural to ask about solvability of peg solitaire in three colors on caterpillars and trees. It would be interesting to explore peg solitaire in three colors on more traditional boards as well. As noted in Section 3A, it remains a natural open question, building on the characterization of free solvability of cycle graphs in two colors given by [Beeler and Paul Hoilman 2011], as to whether the cycle graph C_n is freely solvable in three colors. We would also propose creating a three-color variant of fool's solitaire to compare with the results in [Beeler and Rodriguez 2012].

Acknowledgement

The authors are grateful to Lauren Grimley, Kenan Ince, Gizem Karaali, and Boyan Kostadinov and the American Institute for Mathematics Research Experience for Undergraduates for Faculty Program for their support and guidance on this project. This work is part of a project that was initiated at a workshop that was a part of the Research Experiences for Undergraduate Faculty (REUF) program. REUF is a program of the American Institute of Mathematics (AIM) and the Institute for Computational and Experimental Mathematics (ICERM), made possible by the support from the National Science Foundation (NSF) through DMS 1620073 to AIM and 1620080 to ICERM. We would also like to thank the two anonymous referees for their helpful comments which contributed to the improvement of this paper.

References

[Beasley 1985] J. D. Beasley, *The ins and outs of peg solitaire*, Recreations in Mathematics **2**, Oxford University Press, 1985. MR

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

[Beeler and Walvoort 2015] R. A. Beeler and C. A. Walvoort, "Peg solitaire on trees with diameter four", *Australas. J. Combin.* **63** (2015), 321–332. MR Zbl

[Beeler et al. 2017] R. A. Beeler, H. Green, and R. T. Harper, "Peg solitaire on caterpillars", *Integers* **17** (2017), art. id. G1. MR Zbl

[Bell 2007] G. I. Bell, "A fresh look at peg solitaire", *Math. Mag.* **80**:1 (2007), 16–28. MR Zbl

[Bell 2008] G. I. Bell, "Solving triangular peg solitaire", *J. Integer Seq.* **11**:4 (2008), art. id. 08.4.8. 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

[Engbers and Weber 2018] J. Engbers and R. Weber, "Merging peg solitaire on graphs", *Involve* **11**:1 (2018), 53–66. 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

[Sopena 2019] G. Sopena, "Python implementation of peg solitaire", GitHub repository, 08 2019, available at https://github.com/gustavo-sopena/Peg-Solitaire.

[West 2001] D. B. West, *Introduction to graph theory*, 2nd ed., Prentice Hall, Upper Saddle River, NJ, 2001. Zbl

Received: 2019-09-13 Revised: 2020-03-26 Accepted: 2020-08-06

tdavis@hpu.edu *Department of Mathematics, Hawaii Pacific University,*
Honolulu, HI, United States

adelame1@my.hpu.edu *Department of Mathematics, Hawaii Pacific University,*
Honolulu, HI, United States

gsopena@csu.fullerton.edu *Department of Mathematics, California State University,*
Fullerton, CA, United States

rcsoto@fullerton.edu *Department of Mathematics, California State University,*
Fullerton, CA, United States

sonalivyas@csu.fullerton.edu *Department of Mathematics, California State University,*
Fullerton, CA, United States

melissawong929@gmail.com *Department of Mathematics, California State University,*
Fullerton, CA, United States

mathematical sciences publishers msp

INVOLVE 13:5 (2020)
https://doi.org/10.2140/involve.2020.13.803

Disagreement networks

Florin Catrina and Brian Zilli

(Communicated by Jonathon Peterson)

We first describe an observation based on an analysis of data regarding the outcomes of decisions in cases considered by the United States Supreme Court. Based on this observation, we propose a simple model aiming toward producing an objective notion of an *ideology index*. As an initial step in justifying this concept we produce explicit formulas for the highest-energy eigenvectors of reversible Markov chains with rank-2 transition matrices.

1. Introduction

The idea for this article came about when Zilli noticed a strong correlation between the entries of an eigenvector built intrinsically only from the number of disagreements between justices on the United States Supreme Court and an extrinsic index defined roughly as the percent conservative decisions throughout the justices' tenure, where each decision's direction is determined by a scheme developed by political scientists.

The instances of disagreements between justices on the United States Supreme Court (or any voting body wherein pairwise agreement or disagreement between members can be discerned from vote records) can be expressed as weights on edges of an undirected graph. That is, each justice is represented by a node, and the edge connecting two nodes has weight equal to the number of times the corresponding justices disagreed on the outcome of a case they both participated in. It was hypothesized that techniques of spectral decomposition applied to such a graph could allow for the quantification of each justice's ideology. The significance of this approach is that it does not require legal analysis of the ideological implications of voting in a particular way on each case, which one would intuit as being essential in studying judicial ideology. We apply this method to data from the Washington University Law School's Supreme Court Database [SCDB 2018] and compare the results to the percent conservative measure of ideology resulting from the curated, legal analysis of each case included in the database. The results correlate well (see

MSC2020: primary 05C50, 60J20; secondary 91F10.
Keywords: reversible Markov chain, highest-energy eigenvector, ideal point.

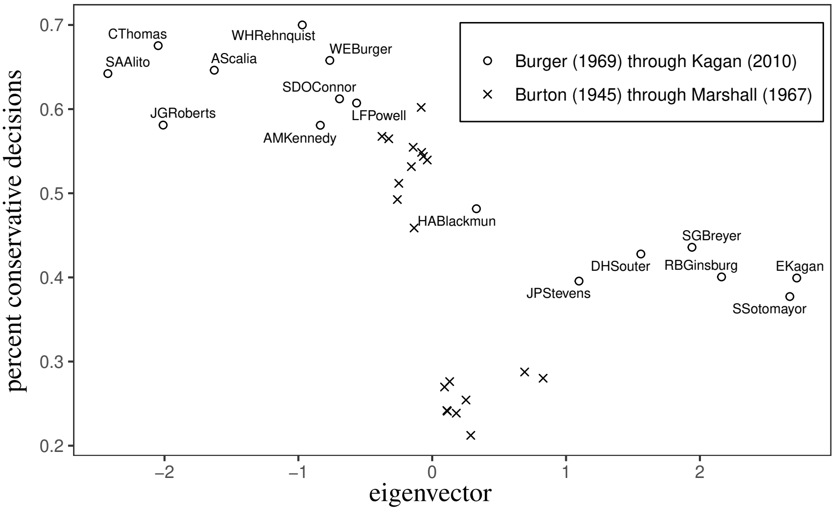


Figure 1. We construct a disagreement graph for decisions rendered from 1946 to 2018 (sourced from [SCDB 2018]) and derive the highest-energy eigenvector for the random walk on this graph. The plot shows the correlation between the eigenvector value and the percent conservative decisions (calculated from the expert classifications in the database) of each justice from Burton (appointed in 1945) through Kagan (2010): $R^2 = 0.349$. When considering only those justices from Burger (1969) through Kagan (represented by $\circ$ on the plot), the correlation is much stronger: $R^2 = 0.831$.

Figure 1), and we believe that this method may be applied to yield insight into bodies for which such extensive expert analysis does not exist.

Specifically, we consider a weighted graph where the nodes represent justices, and the weight on the edge between nodes x and y is a positive integer that represents the number of cases on which justices x and y disagreed. Such a weighted graph induces a reversible Markov chain. We propose the use of the (normalized) eigenvector $\boldsymbol{v}$ that corresponds to the highest energy as an objective measure of the ideological inclination of each of the justices, and we term it *ideology index*. The ideology index for the justice represented by node x will be the value $v(x)$ in the eigenvector $\boldsymbol{v}$.

There are two reasons that led us to consider weights given by disagreements rather than agreements. Firstly, instances of agreement between two judges include, along with cases which have ideological implications, the cases which were relatively uncontroversial (and would not be as indicative of judicial ideology). Conversely, the very existence of a disagreement between judges indicates that the case was contentious. That is, if one were to count the agreements between judges, there would be no objective way to discern which of those agreements were a result of the judges hearing a relatively noncontentious case and which of the agreements were along ideological grounds, whereas all disagreements are attributable to contention between the judges, which is more likely caused by differences in their ideology. Secondly, when we used the agreements data, the strongest correlation to ideology was with the eigenvector corresponding to the eighth-largest eigenvalue (for a 37×37 probability transition matrix). However, we were not able to find why this particular eigenvector would have the strongest correlation with ideology. On the other hand, when using disagreements, the strongest correlation was with the highest-energy eigenvector. In hindsight, this is something that could have been expected from considering the extreme case of bipartite graphs.

While developed independently, we believe that our model has similarities to existing *ideal point* models, such as that of [Martin and Quinn 2002]. In an ideal point model, one considers a justice's ideological position as a point in multidimensional Euclidean space, and the ideological implication of each decision direction on each case is also represented as a point in that space. In such models, the justice will vote in the direction whose point is closest to his or her own in Euclidean distance. Our model is simpler in many of its assumptions compared to Martin and Quinn's, specifically in that we do not consider a justice's ideal point as dynamic throughout his or her career. On the other hand, if one has to describe a justice's voting record with just one number, we consider the measure given by this model to be suitable and objective.

The structure of the paper is as follows. In Section 2 we collect a number of background results on the random walk associated to a reversible Markov chain and discuss the spectrum of the associated probability transition matrix. Section 3 contains the analysis of the highest-energy eigenvector for reversible Markov chains with rank-2 transition matrices. The explicit formulas obtained in the rank-2 case are then used to obtain estimates of the spectrum of a perturbed (higher-rank) "disagreement matrix". In Section 4 we conclude with a review of our results and an invitation to the continuation of this study.

2. Random walk preliminaries

In this section we introduce briefly the notion of random walk associated to a reversible Markov chain and some spectral properties; for details we refer to Section 1.5 in [Chung 1997]. An alternative point of view is that of random walks on electrical networks. For a thorough presentation of this approach we refer to [Doyle and Snell 1984] or Chapter 9 in [Levin et al. 2009].

A *network* is an undirected connected graph $G = (V, E)$, together with nonnegative weights $w(x, y)$ defined for all $x, y \in V$. The numbers $w(x, y)$ are *conductances*, which are assumed positive if (x, y) is an edge in E and zero otherwise. Note that we may allow *loops* (i.e., an edge from a vertex to itself) in which case $(x, x) \in E$ and $w(x, x)$ is positive. We assume the collection of weights forms a symmetric $N \times N$ matrix W, where $N = |V|$ is the number of vertices in the graph; i.e.,

$$\text{for any two vertices } x, y \in V, \quad w(x, y) = w(y, x).$$

Define the *degree* of a vertex $x \in V$ by

$$d(x) = \sum_{y \in V} w(x, y). \tag{2-1}$$

We denote by $\mathbf{1}$ the column vector with N entries all equal to 1.

The matrix P with entries

$$P(x, y) := \frac{w(x, y)}{d(x)} \text{ satisfies } P\mathbf{1} = \mathbf{1}, \text{ i.e.,} \sum_{y \in V} P(x, y) = 1 \text{ for every fixed } x. \tag{2-2}$$

Therefore, the vertices V are the states of a Markov chain with probability transition matrix P. An associated *random walk* is a sequence of random variables $\{X_i\}_{i \geq 0}$ satisfying for all $i \geq 0$ and all $x, y \in V$ that

$$\operatorname{Prob}(X_{i+1} = y \mid X_i = x) = P(x, y).$$

Note that even though the matrix of weights W is symmetric, in general P is not a symmetric matrix.

On the set of nodes V we define a probability distribution $\boldsymbol{\pi}$ (think of it as a row vector) with entries given by

$$\pi(x) = \frac{d(x)}{\sum_y d(y)} = \frac{d(x)}{\operatorname{vol} G},$$

where $\operatorname{vol} G = \sum_y d(y)$ is the total degree (the sum of the degrees of all vertices in V). It is easy to check that the Markov chain thus defined is *reversible* with respect to the probability distribution $\boldsymbol{\pi}$. By definition, this means that for any $x, y \in V$ it holds that

$$\pi(x) P(x, y) = \pi(y) P(y, x). \tag{2-3}$$

A probability distribution $\boldsymbol{\sigma}$ on the set of states V of a Markov chain is called *stationary* if and only if

$$\boldsymbol{\sigma} P = \boldsymbol{\sigma},$$

i.e., $\boldsymbol{\sigma}$ is a left eigenvector of P with eigenvalue 1. A direct check using (2-3) shows that a reversible distribution $\boldsymbol{\pi}$ is also stationary.

A connected graph is *bipartite* if and only if V is the union of two nonempty disjoint sets V_1 and V_2 such that $w(x, y) = 0$ whenever $x, y \in V_1$ or $x, y \in V_2$. The case of bipartite graphs will be examined in what follows only as a limiting case.

Throughout most of this work the graph G will be assumed nonbipartite. On a nonbipartite connected graph any associated random walk is *ergodic*; i.e., there exists a unique stationary distribution $\boldsymbol{\sigma}$ on V such that for any initial distribution $\boldsymbol{\mu}$ it holds that

$$\lim_{s\to\infty} \boldsymbol{\mu} P^s = \boldsymbol{\sigma}.$$

One of the major questions in the theory of reversible Markov chains is characterizing the speed in the convergence

$$\lim_{s\to\infty} \boldsymbol{\mu} P^s = \boldsymbol{\pi}. \tag{2-4}$$

In answering this question, the more powerful techniques use the spectrum (the eigenvalues) and the eigenvectors of P.

The fact that $P\mathbf{1} = \mathbf{1}$ in (2-2) expresses the fact that $\mathbf{1}$ is an eigenvector of P corresponding to eigenvalue 1. Actually, when coupled with the fact that the entries of P are nonnegative, this relation means that every one of the rows of P is a probability distribution on V. By applying P to a column vector $\boldsymbol{u}$ each entry in the resulting vector $P\boldsymbol{u}$ is an average of the entries of $\boldsymbol{u}$, weighted with respect to the corresponding row of P. This averaging characteristic makes P into a contraction with respect to every convex norm on $\mathbb{R}^N$ (this follows from Jensen's inequality). For example, on $\mathbb{R}^N$ (which is identified with the space $L^2(V)$ of functions from V to $\mathbb{R}$) consider the norm

$$\|\boldsymbol{u}\|_2 = \left(\sum_x u^2(x)\right)^{\frac{1}{2}}.$$

Then for every $\boldsymbol{u} \in L^2(V)$ it holds that

$$\|P\boldsymbol{u}\|_2 \le \|\boldsymbol{u}\|_2.$$

In particular, if $\boldsymbol{u}$ is an eigenvector of P with eigenvalue α we have

$$\|P\boldsymbol{u}\|_2 = \|\alpha\boldsymbol{u}\|_2 = |\alpha|\,\|\boldsymbol{u}\|_2 \le \|\boldsymbol{u}\|_2.$$

Therefore, every eigenvalue of P has to satisfy $|\alpha| \le 1$.

Another consequence of reversibility is the fact that all eigenvalues of P are real. This is due to the fact that relations (2-3) imply that P is similar to a symmetric matrix S. Indeed, let $\boldsymbol{d}$ denote the vector of degrees, with entries given by (2-1), and let D denote the diagonal matrix with $\boldsymbol{d}$ on the diagonal (and zero everywhere else). Then, it is clear that

$$S := D^{\frac{1}{2}} P D^{-\frac{1}{2}} = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$$

is a symmetric matrix. Since S and P have the same (real) eigenvalues, and because of the contractive property of P, $|\alpha| \le 1$, it follows that all eigenvalues of P are contained in the interval $[-1, 1]$. Let

$$\Pi = \frac{1}{\operatorname{vol} G} D = \operatorname{diag}(\boldsymbol{\pi}).$$

From the spectral decomposition

$$S = \sum_{i=1}^{N} \alpha_i \phi_i \phi_i^{\mathsf{T}}, \tag{2-5}$$

with $\{\phi_i\}_i$ a set of orthonormal eigenvectors of the matrix S, we get that the eigenvectors of P are given by $\Pi^{-1/2}\phi_i$ and are orthonormal in L^2_π (see below for the definition of L^2_π).

We have already seen that $\alpha_1 = 1$ is an eigenvalue of P with eigenvector $\mathbf{1}$. It is a standard argument (again based on the averaging property of P) to show that when the graph G is connected, $\alpha_1 = 1$ is a simple eigenvalue. Therefore the remaining $N - 1$ eigenvalues of P are strictly less than 1.

The second-largest eigenvalue α_2 of P is of great importance in estimating the convergence speed in (2-4). By analogy with the spectral theory on manifolds, one defines the matrix

$$L = \operatorname{Id} - P,$$

which is called the (normalized) *Laplacian* of the weighted graph G. The spectrum of L consists of eigenvalues $\lambda = 1 - \alpha \in [0, 2]$. We define the inner product space $L^2_\pi(V)$, weighted with respect to the stationary distribution $\boldsymbol{\pi}$, where the inner product is given by

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle_\pi = \sum_x \pi(x) u(x) v(x).$$

Proposition 1. *The equality*

$$\langle \boldsymbol{u}, L\boldsymbol{v} \rangle_\pi = \langle L\boldsymbol{u}, \boldsymbol{v} \rangle_\pi = \frac{1}{\operatorname{vol} G} \sum_{(x,y)\in E} w(x, y)(u(x) - u(y))(v(x) - v(y)) \tag{2-6}$$

holds.

Proof. If we denote by $\delta_{x,y}$ the Kronecker delta function, we have

$$\langle \boldsymbol{u}, L\boldsymbol{v} \rangle_\pi = \sum_{x\in V} \pi(x) u(x)(L\boldsymbol{v})(x) = \sum_{x,y\in V} \frac{d(x)}{\operatorname{vol} G} u(x)(\delta_{x,y} - P(x, y))v(y),$$

therefore

$$\begin{aligned} \langle \boldsymbol{u}, L\boldsymbol{v} \rangle_\pi &= \frac{1}{\operatorname{vol} G} \left\{ \sum_{x\in V} d(x)u(x)v(x) - \sum_{x,y\in V} w(x, y)u(x)v(y) \right\} \\ &= \frac{1}{\operatorname{vol} G} \left\{ \sum_{x,y\in V} w(x, y)u(x)(v(x) - v(y)) \right\}. \end{aligned}$$

For every edge $(x, y) \in E$ with $x \neq y$ (recall that loops are allowed, but note that the terms in the sum for which $x = y$ are zero), terms involving x and y appear twice in the sum above, with x and y reversed. That is, for each edge (x, y) with $x \neq y$, there are two corresponding terms in the sum:

$$w(x, y)u(x)(v(x) - v(y)) \quad \text{and} \quad w(y, x)u(y)(v(y) - v(x)).$$

Since W is symmetric, $w(x, y) = w(y, x)$, and the sum of these terms is

$$w(x, y)(u(x) - u(y))(v(x) - v(y)).$$

Therefore, indexing the sum over edges, we have (2-6). □

For $\boldsymbol{u} \in \mathbb{R}^N$ we define its *energy* by

$$\mathcal{E}(\boldsymbol{u}) := \langle \boldsymbol{u}, L\boldsymbol{u} \rangle_\pi = \frac{1}{\operatorname{vol} G} \sum_{(x,y)\in E} w(x, y)(u(x) - u(y))^2. \tag{2-7}$$

The method of Lagrange multipliers gives that the eigenvalues of L are precisely the critical levels of $\mathcal{E}$ subject to the constraint $\langle \boldsymbol{u}, \boldsymbol{u} \rangle_\pi = 1$.

The minimum level of the energy $\mathcal{E}$ gives $\lambda_1 = 0$, and it is achieved by the constant eigenvector $\boldsymbol{u} = \mathbf{1}$. This corresponds to the largest eigenvalue of P, $\alpha_1 = 1 - \lambda_1 = 1$. The second-largest eigenvalue of P, $\alpha_2 = 1 - \lambda_2$, corresponds to the second smallest eigenvalue of L, given by

$$\lambda_2 = \inf_{\langle \boldsymbol{u}, \mathbf{1} \rangle_\pi = 0} \frac{\langle \boldsymbol{u}, L\boldsymbol{u} \rangle_\pi}{\langle \boldsymbol{u}, \boldsymbol{u} \rangle_\pi}.$$

The difference

$$1 - \alpha_2 = \lambda_2$$

is called the *spectral gap* of the graph and it gives a measure of the speed of convergence in (2-4).

In this paper we will be concerned on the contrary with the smallest eigenvalue of P, which we denote by $\beta = \alpha_N$, and with its corresponding eigenvector. In terms of the eigenvalues of L, this corresponds to the largest eigenvalue, that is, the maximum of $\mathcal{E}$ restricted to the ellipsoid $\langle \boldsymbol{u}, \boldsymbol{u} \rangle_\pi = 1$. For this reason, we will call the eigenvector corresponding to β the *highest-energy eigenvector.* As we have seen above, β has to be a number in the interval $[-1, 1)$. For a connected graph, -1 is an eigenvalue if and only if G is bipartite. In our general setting G is not bipartite; therefore except for the one eigenvalue 1, the remaining eigenvalues (including β) will be in $(-1, 1)$.

3. Disagreement matrices

Assume we have a graph with a fixed number N of nodes (the justices), and each node x has a probability $p_x \in [0, 1]$ associated to it. In our proposed model, we assume there exists an abstract notion of, say, ideal conservator, which indicates how a pure ideological conservative would vote on every case. Then, p_x will represent the probability that justice x will vote with the ideal conservator. We will assume that for each justice, this probability is constant throughout their career. The probabilities p_x will be collected as entries of a vector $\boldsymbol{p}$ which will be fixed throughout. Our assumption that each p_x is constant throughout a justice's career is a major simplification in our model over earlier approaches, as for example in [Martin and Quinn 2002].

For each x, let $q_x := 1 - p_x$ also in $[0, 1]$. The *disagreement matrix* of this system is defined as the matrix B with entries

$$B_{xy} = \begin{cases} 0 & \text{if } x = y, \\ p_x q_y + p_y q_x & \text{if } x \neq y. \end{cases}$$

It has zero on the diagonal since a justice always votes in agreement with her/himself and the off-diagonal entries represent the probabilities of disagreement (justices x and y vote contrary to each other).

3.1. ***Rank-2 weight matrices.*** In this subsection we study the highest-energy eigenvector in a simpler (and less realistic) case when the diagonal entries in the disagreement matrix are not equal to zero. In $\mathbb{R}^N$, let $\boldsymbol{p}$ and $\boldsymbol{p}^\intercal$ denote the column and the row vector, respectively, with entries $p_x \in [0, 1]$, and similarly define the vectors $\boldsymbol{q}$ and $\boldsymbol{q}^\intercal$ with entries $q_x = 1 - p_x \in [0, 1]$. We discuss the spectrum of the probability transition matrix P_T associated with the rank-2 weight matrix $T := \boldsymbol{p}\boldsymbol{q}^\intercal + \boldsymbol{q}\boldsymbol{p}^\intercal$. We note that B is equal to T except for the diagonal entries, which are replaced by zero.

We begin with the following:

Proposition 2. *In the generic case*, *i.e.*, $\boldsymbol{p}$ *is not a constant vector*, *we have that* $\boldsymbol{p}$ *and* $\boldsymbol{q}$ *are linearly independent vectors. In particular*, T *is a matrix of rank* 2.

Proof. Assume that for real numbers a and b we have $a\boldsymbol{p} + b\boldsymbol{q} = \boldsymbol{0}$. Then

$$\boldsymbol{0} = a\boldsymbol{p} + b(\boldsymbol{1} - \boldsymbol{p}) = (a - b)\boldsymbol{p} + b\boldsymbol{1}, \quad \text{i.e.,} \quad (b - a)\boldsymbol{p} = b\boldsymbol{1}.$$

Since $\boldsymbol{p}$ is nonconstant, from the last equality we must have $b = a = 0$; therefore the vectors $\boldsymbol{p}$ and $\boldsymbol{q}$ are linearly independent.

Next, we show that T has rank 2. Let $\boldsymbol{p}^\perp$ denote the $(N-1)$-dimensional space of vectors in $\mathbb{R}^N$ which are orthogonal to $\boldsymbol{p}$, and similarly define $\boldsymbol{q}^\perp$. Since $\boldsymbol{p}$ and

$\boldsymbol{q}$ are linearly independent vectors we have

$$\dim(\boldsymbol{p}^\perp \cap \boldsymbol{q}^\perp) = N - 2.$$

Since

$$\boldsymbol{p}^\perp \cap \boldsymbol{q}^\perp \subset \ker T,$$

we have that

$$\dim(\ker T) \geq N - 2. \tag{3-1}$$

On the other hand, the vectors $T\boldsymbol{p}$ and $T\boldsymbol{q}$ are linearly independent. Indeed, $aT\boldsymbol{p} + bT\boldsymbol{q} = \mathbf{0}$ implies

$$\mathbf{0} = a(\boldsymbol{p}^\mathsf{T}\boldsymbol{q})\boldsymbol{p} + a|\boldsymbol{p}|^2\boldsymbol{q} + b|\boldsymbol{q}|^2\boldsymbol{p} + b(\boldsymbol{p}^\mathsf{T}\boldsymbol{q})\boldsymbol{q},$$

and since $\boldsymbol{q}$ and $\boldsymbol{p}$ are linearly independent we must have

$$a|\boldsymbol{p}|^2 + b(\boldsymbol{p}^\mathsf{T}\boldsymbol{q}) = 0 \quad \text{and} \quad a(\boldsymbol{p}^\mathsf{T}\boldsymbol{q}) + b|\boldsymbol{q}|^2 = 0.$$

The determinant of this system is

$$\begin{vmatrix} |\boldsymbol{p}|^2 & \boldsymbol{p}^\mathsf{T}\boldsymbol{q} \\ \boldsymbol{p}^\mathsf{T}\boldsymbol{q} & |\boldsymbol{q}|^2 \end{vmatrix} = |\boldsymbol{p}|^2|\boldsymbol{q}|^2 - (\boldsymbol{p}^\mathsf{T}\boldsymbol{q})^2 > 0$$

because the value of the determinant equals the square of the area of the parallelogram spanned by $\boldsymbol{p}$ and $\boldsymbol{q}$. Therefore, again we must have $b = a = 0$. Since $T\boldsymbol{p}$ and $T\boldsymbol{q}$ are linearly independent, it means

$$\dim(\operatorname{range} T) \geq 2. \tag{3-2}$$

From the rank-nullity theorem we have

$$\dim(\ker T) + \dim(\operatorname{range} T) = N.$$

From this, it follows that both inequalities (3-1) and (3-2) are in fact equalities; i.e., T has rank 2. □

Define

$$s_p = \sum_x p_x \quad \text{and} \quad s_q = \sum_x q_x, \quad \text{both in } [0, N], \text{ with } s_p + s_q = N.$$

Since we work with the weight matrix $T = \boldsymbol{p}\boldsymbol{q}^\mathsf{T} + \boldsymbol{q}\boldsymbol{p}^\mathsf{T}$, the degree of node x is the sum of entries on row x of T,

$$d(x) = \sum_y T_{xy} = p_x\Big(\sum_y q_y\Big) + q_x\Big(\sum_y p_y\Big) = p_x s_q + q_x s_p.$$

Note that we can write

$$d(x) = (N - s_p)p_x + s_p(1 - p_x),$$

and since both numbers $N - s_p$ and s_p are strictly positive, we have that $d(x) > 0$ for every x. We collect the degrees of all nodes in the vector $\boldsymbol{d}$, therefore

$$\boldsymbol{d} = s_q \boldsymbol{p} + s_p \boldsymbol{q},$$

and denote by $D := \operatorname{diag}(\boldsymbol{d})$ the diagonal matrix with entries $d(x)$ on the diagonal. Define the probability transition matrix

$$P_T := D^{-1} T.$$

We have the following:

Theorem 3. *In the generic case*, *i.e.*, $\boldsymbol{p}$ *is not a constant vector*, *the matrix* P_T *has*

(1) *the largest eigenvalue is equal to* 1, *it is simple*, *and has corresponding eigenvector* $\mathbf{1}$;

(2) *eigenvalue* 0 *with multiplicity* $N - 2$ *and corresponding eigenspace consisting of all vectors orthogonal to both* $\boldsymbol{p}$ *and* $\boldsymbol{q}$, *i.e.*,

$$\ker P_T = \ker T = \boldsymbol{p}^{\perp} \cap \boldsymbol{q}^{\perp};$$

(3) *simple eigenvalue* $\beta \in [-1, 0)$, *with eigenvector* $\boldsymbol{v}$ *given explicitly in the formulas* (3-5) *and* (3-11) *below.*

Remark 4. As a consequence of the formulas (3-5) and (3-11) we will note that the eigenvector $\boldsymbol{v}$ is monotonic in $\boldsymbol{p}$. This means that as we order the vertices x of the graph in increasing order of $\boldsymbol{p}$, the values in the entries of $\boldsymbol{v}$ will also be either in increasing or decreasing order.

Another consequence of these formulas are estimates (explicitly in terms of $\boldsymbol{p}$ or $\boldsymbol{d}$) for the maximum values $\|\boldsymbol{v}\|_\infty$ given in (3-6) and (3-12).

Proof of Theorem 3. Part (1) follows from the general theory presented in Section 2.

Part (2) is immediate from the fact that the matrix D^{-1} is diagonal with no zeros on the diagonal, together with the fact from Proposition 2 that $\operatorname{rank} T = 2$.

For part (3) we will distinguish the following two cases:

Case 1: $s_p = N/2$. Since $\boldsymbol{q} = \mathbf{1} - \boldsymbol{p}$ we have

$$s_q = N - s_p = \frac{N}{2}, \quad \boldsymbol{d} = s_q \boldsymbol{p} + s_p \boldsymbol{q} = \frac{N}{2}\mathbf{1} \quad \text{and} \quad D = \frac{N}{2}\,\mathrm{Id}\,.$$

Also,

$$|\boldsymbol{q}|^2 = \sum_x q_x^2 = \sum_x (1 - p_x)^2 = \sum_x (1 - 2p_x + p_x^2) = N - 2s_p + \sum_x p_x^2 = |\boldsymbol{p}|^2,$$

and

$$\boldsymbol{p}^{\mathsf{T}} \boldsymbol{q} = \boldsymbol{p}^{\mathsf{T}} (\mathbf{1} - \boldsymbol{p}) = s_p - |\boldsymbol{p}|^2 = \frac{N}{2} - |\boldsymbol{p}|^2.$$

Consider the vector

$$\boldsymbol{u} = \boldsymbol{p} - \boldsymbol{q} = 2\boldsymbol{p} - \mathbf{1}.$$

Then

$$T\boldsymbol{u} = (\boldsymbol{p}\boldsymbol{q}^{\mathsf{T}} + \boldsymbol{q}\boldsymbol{p}^{\mathsf{T}})(\boldsymbol{p}-\boldsymbol{q}) = (\boldsymbol{p}^{\mathsf{T}}\boldsymbol{q})(\boldsymbol{p}-\boldsymbol{q}) + |\boldsymbol{p}|^2\boldsymbol{q} - |\boldsymbol{q}|^2\boldsymbol{p} = \Big(\frac{N}{2} - 2|\boldsymbol{p}|^2\Big)(\boldsymbol{p}-\boldsymbol{q}).$$

Therefore

$$T\boldsymbol{u} = \left(1 - \frac{4|\boldsymbol{p}|^2}{N}\right)\frac{N}{2}\boldsymbol{u} = \left(1 - \frac{4|\boldsymbol{p}|^2}{N}\right)D\boldsymbol{u}. \tag{3-3}$$

Note that because $s_p = N/2$, by the Cauchy–Schwarz inequality we have

$$\frac{N^2}{4} = \Big(\sum_x p_x\Big)^2 \le N \sum_x p_x^2 = N|\boldsymbol{p}|^2,$$

i.e.,

$$1 \le \frac{4|\boldsymbol{p}|^2}{N}.$$

In fact, the inequality above is strict as the only possibility for equality is when $\boldsymbol{p}$ is constant.

Since $P_T = D^{-1}T$, by multiplying (3-3) on the left by D^{-1}, we get that $P_T\boldsymbol{u} = \beta\boldsymbol{u}$ with

$$\beta = 1 - \frac{4|\boldsymbol{p}|^2}{N} < 0. \tag{3-4}$$

The stationary distribution in this case is the uniform distribution, i.e., $\pi(x) = 1/N$. It is common to normalize the eigenvectors in the L^2_π norm. That is, if $\boldsymbol{v} = c\boldsymbol{u}$, we require

$$1 = \|\boldsymbol{v}\|_\pi^2 = \frac{c^2}{N}|\boldsymbol{u}|^2 = \frac{c^2}{N}(|\boldsymbol{p}|^2 + |\boldsymbol{q}|^2 - 2\boldsymbol{p}^{\mathsf{T}}\boldsymbol{q}) = c^2\left(\frac{4|\boldsymbol{p}|^2}{N} - 1\right).$$

Therefore

$$\boldsymbol{v} = c(\boldsymbol{p} - \boldsymbol{q}), \quad \text{with } c = \left(\frac{4|\boldsymbol{p}|^2}{N} - 1\right)^{-\frac{1}{2}}. \tag{3-5}$$

We note that, because $\boldsymbol{v} = c(\boldsymbol{p} - \boldsymbol{q}) = c(2\boldsymbol{p} - 1)$ with $c > 0$, we have that $\boldsymbol{v}$ is monotonically increasing in $\boldsymbol{p}$; that is, nodes x with larger p_x also have larger v_x.

Because $-\mathbf{1} \le 2\boldsymbol{p} - 1 \le \mathbf{1}$, from (3-5) we get the pointwise estimate for the entries of $\boldsymbol{v}$,

$$\|\boldsymbol{v}\|_\infty = c\max\{(2p_{\max} - 1), (1 - 2p_{\min})\} \le \frac{1}{\sqrt{4|\boldsymbol{p}|^2/N - 1}}. \tag{3-6}$$

Case 2: $s_p \ne N/2$. In this case,

$$\boldsymbol{d} = s_q\boldsymbol{p} + s_p\boldsymbol{q} = s_p\mathbf{1} + (N - 2s_p)\boldsymbol{p}$$

is not a constant vector anymore, and together $\mathbf{1}$ and $\boldsymbol{d}$ form a basis for the range of T. Indeed, from Proposition 2 we know that $\operatorname{rank} T = 2$. In fact, the range of T is equal to the span of the vectors $\boldsymbol{p}$ and $\boldsymbol{q}$. That $\boldsymbol{p}$ is in the range of T can be seen by applying T to the component of $\boldsymbol{q}$ orthogonal to $\boldsymbol{p}$

$$\operatorname{orth}_{\boldsymbol{p}} \boldsymbol{q} := \boldsymbol{q} - \frac{\boldsymbol{p}^{\mathsf{T}}\boldsymbol{q}}{|\boldsymbol{p}|^2}\boldsymbol{p}.$$

Then

$$T \operatorname{orth}_{\boldsymbol{p}} \boldsymbol{q} = \left(|\boldsymbol{q}|^2 - \frac{(\boldsymbol{p}^{\mathsf{T}}\boldsymbol{q})^2}{|\boldsymbol{p}|^2}\right)\boldsymbol{p}.$$

Since the vectors $\boldsymbol{p}$ and $\boldsymbol{q}$ are not proportional, the Cauchy–Schwarz inequality

$$(\boldsymbol{p}^{\mathsf{T}}\boldsymbol{q})^2 \le |\boldsymbol{p}|^2|\boldsymbol{q}|^2$$

is strict and therefore the coefficient of $\boldsymbol{p}$ above is strictly positive. By a similar argument, $\boldsymbol{q}$ is in the range of T. Since the vectors

$$\mathbf{1} = \boldsymbol{p} + \boldsymbol{q} \quad \text{and} \quad \boldsymbol{d} = T\mathbf{1}$$

are in the range of T and are linearly independent, they form a basis for the range of T.

We introduce the notation

$$A(\boldsymbol{d}) = \frac{\sum_x d(x)}{N} \quad \text{and} \quad H(\boldsymbol{d}) = \frac{N}{\sum_x 1/d(x)}$$

for the arithmetic and harmonic means, respectively, of the entries of the vector of degrees $\boldsymbol{d}$. Consider the vector

$$\boldsymbol{u} = \mathbf{1} - A(\boldsymbol{d})D^{-1}\mathbf{1}. \tag{3-7}$$

We show that for a $\beta \in [-1, 0)$ specified below in (3-8), we have $P_T\boldsymbol{u} = \beta\boldsymbol{u}$.

First we note that because $\boldsymbol{d}$ is not a constant vector, we have

$$D\boldsymbol{u} = \boldsymbol{d} - A(\boldsymbol{d})\mathbf{1} \neq \mathbf{0}.$$

From this, together with

$$\mathbf{1}^{\mathsf{T}} D\boldsymbol{u} = 0,$$

we deduce that $\mathbf{1}$ and $D\boldsymbol{u}$ are nonzero, perpendicular, vectors in the range of T.

Since the range of T is 2-dimensional, if we show that $T\boldsymbol{u}$ is also perpendicular to $\mathbf{1}$, it follows that for some real β,

$$T\boldsymbol{u} = \beta D\boldsymbol{u}, \quad \text{i.e.,} \quad P_T\boldsymbol{u} = \beta\boldsymbol{u}.$$

We show that $\mathbf{1}^{\mathsf{T}}T\boldsymbol{u} = 0$. Indeed, we have

$$\mathbf{1}^{\mathsf{T}}T\boldsymbol{u} = (T\mathbf{1})^{\mathsf{T}}(\mathbf{1} - A(\boldsymbol{d})D^{-1}\mathbf{1}) = \boldsymbol{d}^{\mathsf{T}}(\mathbf{1} - A(\boldsymbol{d})D^{-1}\mathbf{1}) = NA(\boldsymbol{d}) - A(\boldsymbol{d})\boldsymbol{d}^{\mathsf{T}}D^{-1}\mathbf{1} = 0.$$

The only thing left to check is that $-1 \leq \beta < 0$. We do this by calculating the trace of P_T. On one hand, since the trace of a square matrix equals the sum of its eigenvalues, and $N-2$ of the eigenvalues of P_T are zero, we have $\operatorname{Tr} P_T = 1+\beta$, and therefore

$$\beta = \operatorname{Tr} P_T - 1. \tag{3-8}$$

On the other hand,

$$\begin{aligned}\operatorname{Tr} P_T &= \sum_x \frac{2p_x q_x}{d(x)} = \sum_x \frac{2p_x q_x}{s_q p_x + s_p q_x}\\ &= \sum_x \frac{2}{s_q/q_x + s_p/p_x} \leq \sum_x \frac{q_x/s_q + p_x/s_p}{2} = 1.\end{aligned}$$

The inequality above follows from the term-by-term inequality between the harmonic and arithmetic means of two positive numbers

$$\frac{2}{s_q/q_x + s_p/p_x} \leq \frac{q_x/s_q + p_x/s_p}{2},$$

with equality if and only if

$$\frac{q_x}{s_q} = \frac{p_x}{s_p}.$$

Since the equality cannot hold for all x (otherwise $\boldsymbol{p}$ would be constant), we obtain that $0 < \operatorname{Tr} P_T < 1$, that is, $-1 < \beta < 0$. In the calculation above, we implicitly assumed that for all x we have $p_x \in (0, 1)$. If p_x is equal to either 0 or 1, then the term $2p_x q_x/d(x)$ corresponding to x in $\operatorname{Tr} P_T$ is zero, and thus strictly smaller than

$$\frac{q_x/s_q + p_x/s_p}{2}.$$

As we shall see in the next subsection, if for every x we have that p_x is equal to either 0 or 1, then $\operatorname{Tr} P_T = 0$ and therefore $\beta = -1$. This is the case of bipartite graphs.

Alternatively, one may estimate β from $T\boldsymbol{u} = \beta D\boldsymbol{u}$ by multiplying by $\boldsymbol{u}^\intercal$ on the left. We get

$$\beta = \frac{\boldsymbol{u}^\intercal T\boldsymbol{u}}{\boldsymbol{u}^\intercal D\boldsymbol{u}}. \tag{3-9}$$

The denominator in (3-9) is readily calculated as

$$\boldsymbol{u}^\intercal D\boldsymbol{u} = (\mathbf{1}^\intercal - A(\boldsymbol{d})\mathbf{1}^\intercal D^{-1})D\boldsymbol{u} = \mathbf{1}^\intercal D\boldsymbol{u} - A(\boldsymbol{d})\mathbf{1}^\intercal\boldsymbol{u}.$$

Since $\mathbf{1}^\intercal D\boldsymbol{u} = 0$, we get

$$\boldsymbol{u}^\intercal D\boldsymbol{u} = -A(\boldsymbol{d})\Bigl(N - A(\boldsymbol{d})\sum_x \frac{1}{d(x)}\Bigr) = -A(\boldsymbol{d})\Bigl(N - N\frac{A(\boldsymbol{d})}{H(\boldsymbol{d})}\Bigr),$$

which means

$$\boldsymbol{u}^{\mathsf{T}} D\boldsymbol{u} = NA(\boldsymbol{d})\left(\frac{A(\boldsymbol{d})}{H(\boldsymbol{d})} - 1\right) > 0. \tag{3-10}$$

The fact that the quantity in parentheses above is positive follows from the inequality between harmonic and arithmetic means. The calculation that shows that $\boldsymbol{u}^{\mathsf{T}} T\boldsymbol{u} < 0$, however, can be quite tedious.

The equilibrium measure in this case is not uniform anymore. It will be given by

$$\pi(x) = \frac{d(x)}{\sum_y d(y)} = \frac{d(x)}{NA(\boldsymbol{d})}.$$

Again, we would like the eigenvector corresponding to eigenvalue β to be normalized in the L^2 norm weighted by $\boldsymbol{\pi}$. That is, if $\boldsymbol{v} = c\boldsymbol{u}$, we require

$$1 = \|\boldsymbol{v}\|_\pi^2 = \frac{c^2}{NA(\boldsymbol{d})}\boldsymbol{u}^{\mathsf{T}} D\boldsymbol{u} = \frac{c^2}{NA(\boldsymbol{d})} NA(\boldsymbol{d})\left(\frac{A(\boldsymbol{d})}{H(\boldsymbol{d})} - 1\right) = c^2\left(\frac{A(\boldsymbol{d})}{H(\boldsymbol{d})} - 1\right).$$

Therefore the normalized eigenvector corresponding to eigenvalue β is

$$\boldsymbol{v} = c(\mathbf{1} - A(\boldsymbol{d})D^{-1}\mathbf{1}), \quad \text{with } c = \left(\frac{A(\boldsymbol{d})}{H(\boldsymbol{d})} - 1\right)^{-\frac{1}{2}}. \tag{3-11}$$

Since we are in the case $s_p \neq N/2$, we have that

$$\boldsymbol{d} = s_q\boldsymbol{p} + s_p\boldsymbol{q} = (s_q - s_p)\boldsymbol{p} + s_p\mathbf{1}$$

is monotonic in $\boldsymbol{p}$ and therefore so is the vector $D^{-1}\mathbf{1}$. From the formula (3-11) we conclude that $\boldsymbol{v}$ is monotonic in $\boldsymbol{p}$. Also, observe that $\boldsymbol{v}$ changes sign precisely between the values of $\boldsymbol{d}$ which are above and below $A(\boldsymbol{d})$.

From the monotonicity of $\boldsymbol{v}$ we conclude that

$$\max_x |v_x| = \|\boldsymbol{v}\|_\infty \quad \text{for } x \text{ such that either } p_x = p_{\min} \text{ or } p_x = p_{\max}.$$

Precisely, we have

$$\|\boldsymbol{v}\|_\infty = \frac{\max\{(1 - A(\boldsymbol{d})/d_{\max}), (A(\boldsymbol{d})/d_{\min} - 1)\}}{\sqrt{A(\boldsymbol{d})/H(\boldsymbol{d}) - 1}}, \tag{3-12}$$

completing the proof. □

In the next subsection we discuss briefly the case when every p_x is either 0 or 1.

3.2. ***The case $p_x \in \{0, 1\}$.*** In our setting with $p_x \in [0, 1]$, because every term in $\operatorname{Tr} P_T$ is nonnegative, we only get $\operatorname{Tr} P_T = 0$ when all the p_x are categorical, i.e.,

$p_x = 0$ or $p_x = 1$. Assume m of p_x are equal to 1 and $n = N - m$ are equal to 0; then the matrix T has the form

$$T = \left[\begin{array}{c|c} \mathbf{0}_{m\times m} & \mathbf{1}_{m\times n} \\ \hline \mathbf{1}_{n\times m} & \mathbf{0}_{n\times n} \end{array}\right],$$

where the diagonal zero blocks are of size $m \times m$ and $n \times n$. The degrees are

$$d_x = n \quad \text{when } p_x = 1 \qquad \text{and} \qquad d_x = m \quad \text{when } p_x = 0.$$

The graph associated to the Markov chain with probability transition matrix $P_T = D^{-1}T$ is bipartite and P_T has the smallest eigenvalue $\beta = -1$. The corresponding eigenvector $\boldsymbol{v}$ has

$$v_x = 1 \quad \text{when } p_x = 1 \qquad \text{and} \qquad v_x = -1 \quad \text{when } p_x = 0.$$

Indeed, both formulas given in (3-5) and in (3-11) reduce to $\beta = -1$ and to the eigenvector $\boldsymbol{v}$ above:

If $m = n = N/2$ we have $|\boldsymbol{p}|^2 = N/2$ so that (3-5) yields $c = 1$.

If $m < n$, then

$$A(\boldsymbol{d}) = \frac{2mn}{N}, \quad H(\boldsymbol{d}) = \frac{Nmn}{m^2 + n^2},$$

and from (3-11) we get

$$\boldsymbol{u} = \frac{n-m}{N}\begin{bmatrix} 1 \\ \vdots \\ 1 \\ -1 \\ \vdots \\ -1 \end{bmatrix} \quad \text{and} \quad c = \frac{N}{n-m},$$

and thus $\boldsymbol{v} = c\boldsymbol{u}$ as above.

3.3. ***The eigenvalues of the transition matrix P_B.*** A more realistic model would consider a disagreement matrix B equal to T everywhere except on the diagonal, where the entries are made equal to 0 (a justice never votes contrary to their own vote). As before, we define the degrees vector $\boldsymbol{d}_B$ with entries

$$d_B(x) = \sum_y B_{xy} = \sum_{y \neq x} T_{xy} = p_x s_q + q_x s_p - 2p_x q_x.$$

Let $P_B = D_B^{-1}B$ be the probability transition matrix associated to B, where D_B is the diagonal matrix with $\boldsymbol{d}$ on the diagonal. The study of the spectrum and the eigenvectors of P_B is itself an interesting problem.

Observe that for $N = 2$ we have

$$P_B = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

and its eigenvalues are 1 and -1 with eigenvectors given respectively by

$$\mathbf{1} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

In general, the matrices B, and consequently P_B, no longer have rank 2. However, numerical simulations lead (at least in the case when N is relatively large and the values of $\boldsymbol{p}$ are well distributed in [0, 1]) to the following

Conjecture 5. *Besides the eigenvalue $\lambda = 1$, the remaining $N - 1$ eigenvalues are negative with $N-2$ of them close to* 0, *while the smallest is less than β* (*the smallest eigenvalue of P_T*).

Conjecture 6. *The highest-energy eigenvector of P_B seems to remain monotonic in $\boldsymbol{p}$ and it is close to the corresponding eigenvector of P_T.*

We should warn the reader that our use of the word "conjecture" in this paper is not in the sense it is used traditionally in mathematics, but more with the meaning of "guess".

While the conjecture about the sign of the eigenvalues of P_B is true, as we prove in the Theorem 8 next, the conjecture about the monotonicity in $\boldsymbol{p}$ of the highest-energy eigenvector of P_B fails. This was again observed numerically for small values of N ($N = 3$ and $N = 4$) and it naturally leads to the following:

Question 7. *What conditions on $\boldsymbol{p}$ ensure the monotonicity* (*in $\boldsymbol{p}$*) *of the highest-energy eigenvector of P_B?*

We now prove the following:

Theorem 8. *The probability transition matrix P_B has the largest eigenvalue* 1, *while the remaining $N - 1$ eigenvalues are negative*, *with the smallest satisfying $\beta_B \le \beta_T$*, *with equality if and only if $p_x \in \{0, 1\}$ for all x* (*hence the bipartite case from Section 3.2*).

Proof. Our method of proof is to relate the eigenvalues of P_B to those of P_T in Section 3.1. For this we denote by $\boldsymbol{e}$ the vector with entries $e(x) = 2p_xq_x$, and by $E := \operatorname{diag}(\boldsymbol{e})$, the diagonal matrix with $\boldsymbol{e}$ on the diagonal. Note that

$$T = B + E \quad \text{and} \quad \boldsymbol{d}_T = \boldsymbol{d}_B + \boldsymbol{e}.$$

Let $\boldsymbol{u}$ be a nonconstant eigenvector of P_B corresponding to an eigenvalue $\alpha < 1$. We show that the eigenvalue of $\operatorname{Id} - P_B$ given by $\lambda = 1 - \alpha$ is greater than 1 and therefore $\alpha < 0$. For this we estimate the energy $\mathcal{E}_B(\boldsymbol{u})$ given by (2-7) with $w(x, y) = p_x q_y + p_y q_x$.

From the spectral decomposition (2-5) we have that $\boldsymbol{u}$ is perpendicular to $\mathbf{1}$ with respect to $\boldsymbol{\pi}$, and therefore with respect to the weight $\boldsymbol{d}_B$; i.e.,

$$\langle \boldsymbol{u}, \mathbf{1}\rangle_{d_B} = \langle \boldsymbol{u}, \boldsymbol{d}_B\rangle = \sum_{x\in V} d_B(x)u(x) = 0.$$

Consider the (unique) decomposition

$$\boldsymbol{u} = a\mathbf{1} + \boldsymbol{v},$$

where a is a real number, and $\boldsymbol{v}$ is perpendicular to $\mathbf{1}$ with respect to $\boldsymbol{d}_T$. Since $\boldsymbol{u}$ is nonconstant, the vector $\boldsymbol{v}$ is nonzero. The coefficient of $\mathbf{1}$ in the decomposition of $\boldsymbol{u}$ above is the scalar projection of $\boldsymbol{u}$ on $\mathbf{1}$ with respect to the inner product weighted by $\boldsymbol{d}_T$, and it can be calculated as

$$a = \frac{\langle \boldsymbol{u}, \boldsymbol{e}\rangle}{\operatorname{vol}(G_T)} = \frac{\langle \boldsymbol{u}, \boldsymbol{e}\rangle}{2s_p s_q}. \tag{3-13}$$

Indeed, from the fact that $\boldsymbol{d}_T = \boldsymbol{d}_B + \boldsymbol{e}$, we obtain

$$\langle \boldsymbol{u}, \boldsymbol{d}_T\rangle = \underbrace{\langle \boldsymbol{u}, \boldsymbol{d}_B\rangle}_{=0} + \langle \boldsymbol{u}, \boldsymbol{e}\rangle, \qquad \text{therefore} \quad \langle a\mathbf{1}, \boldsymbol{d}_T\rangle + \underbrace{\langle \boldsymbol{v}, \boldsymbol{d}_T\rangle}_{=0} = \langle \boldsymbol{u}, \boldsymbol{e}\rangle.$$

Since $\langle \mathbf{1}, \boldsymbol{d}_T\rangle = \operatorname{vol}(G_T) = 2s_p s_q$, formula (3-13) follows.

Observe that because $\boldsymbol{v}$ is perpendicular to $\mathbf{1}$ with respect to the weight $\boldsymbol{d}_T$ then $\boldsymbol{v}$ is in the sum of the eigenspaces of T with nonpositive eigenvalues. Therefore the energy $\mathcal{E}_T(\boldsymbol{v})$ is bounded from below by $\|\boldsymbol{v}\|^2_{\pi_T} = \langle \boldsymbol{v}, \boldsymbol{v}\rangle_{\pi_T}$. This means,

$$\operatorname{vol}(G_T)\mathcal{E}_T(\boldsymbol{v}) \geq \langle \boldsymbol{v}, \boldsymbol{v}\rangle_{d_T} = \sum_{x\in V} d_T(x)v^2(x) = \sum_{x\in V} d_B(x)v^2(x) + \sum_{x\in V} e(x)v^2(x).$$

As $v(x) = (u(x) - a)$ for every x, we have

$$\begin{aligned}\sum_{x\in V} d_B(x)v^2(x) &= \sum_{x\in V} d_B(x)(u(x)-a)^2\\ &= \sum_{x\in V} d_B(x)u^2(x) + a^2\operatorname{vol}(G_B) - 2a\underbrace{\sum_{x\in V} d_B(x)u(x)}_{=0}.\end{aligned}$$

Since $\mathcal{E}_T(\boldsymbol{u}) = \mathcal{E}_T(\boldsymbol{v})$, we obtain

$$\operatorname{vol}(G_B)\mathcal{E}_B(\boldsymbol{u}) = \operatorname{vol}(G_T)\mathcal{E}_T(\boldsymbol{v}) \geq \sum_{x\in V} d_B(x)u^2(x) + a^2\operatorname{vol}(G_B) + \sum_{x\in V} e(x)v^2(x).$$

Therefore

$$\mathcal{E}_B(\boldsymbol{u}) \geq \|\boldsymbol{u}\|^2_{\pi_B} + a^2 + \frac{1}{\operatorname{vol}(G_B)}\sum_{x\in V} e(x)v^2(x),$$

and so

$$\mathcal{E}_B(\boldsymbol{u}) = (1-\alpha)\|\boldsymbol{u}\|^2_{\pi_B} \geq \|\boldsymbol{u}\|^2_{\pi_B}, \quad \text{i.e.,} \quad \alpha \leq 0.$$

Observe that from (3-13) we have $a^2 > 0$ unless $\boldsymbol{u} \perp \boldsymbol{e}$. Also, under the assumption that $p_x \in (0, 1)$ for all x, we have $\sum_{x\in V} e(x)v^2(x) > 0$. Thus, the only cases of equality $\mathcal{E}_B(\boldsymbol{u}) = \|\boldsymbol{u}\|^2_{\pi_B}$ happen when $a = 0$, i.e., $\boldsymbol{u} = \boldsymbol{v}$, and $u(x) = 0$ for any x with $p_x \in (0, 1)$.

To estimate the highest-energy eigenvalue β_B of P_N, we use the variational characterization

$$(1-\beta_B) = \max_{\boldsymbol{u}\neq\boldsymbol{0}} \frac{\mathcal{E}_B(\boldsymbol{u})}{\|\boldsymbol{u}\|^2_{\pi_B}}.$$

Let $\boldsymbol{v}$ be the eigenvector corresponding to the highest-energy eigenvalue β_T of P_T and consider a test vector $\boldsymbol{u} = a\mathbf{1} + \boldsymbol{v}$, where a is picked such that $\boldsymbol{u}$ is perpendicular to $\mathbf{1}$ with respect to the weight $\boldsymbol{d}_B$. Then,

$$(1-\beta_B) \geq \frac{\mathcal{E}_B(\boldsymbol{u})}{\|\boldsymbol{u}\|^2_{\pi_B}} = \operatorname{vol}(G_T)\frac{\mathcal{E}_T(\boldsymbol{u})}{\sum_{x\in V} d_B(x)u^2(x)} = (1-\beta_T)\frac{\sum_{x\in V} d_T(x)v^2(x)}{\sum_{x\in V} d_B(x)u^2(x)}.$$

As before,

$$\begin{aligned}\sum_{x\in V} d_T(x)v^2(x) &= \sum_{x\in V} d_B(x)v^2(x) + \sum_{x\in V} e(x)v^2(x)\\ &= \operatorname{vol}(G_B)a^2 + \sum_{x\in V} d_B(x)u^2(x) + \sum_{x\in V} e(x)v^2(x) \geq \sum_{x\in V} d_B(x)u^2(x).\end{aligned}$$

Therefore $(1-\beta_B) \geq (1-\beta_T)$ with strict inequality if $p_x \in (0, 1)$ for all x. □

4. Conclusions and future directions

There are two aspects associated with the subject of this paper. One is the modeling aspect together with the proposal of a new measure, the ideology index, in the concrete case of the United States Supreme Court, and the other aspect consists of the theoretical investigations of further simplifications of the model.

Regarding the model, we begin with a data set that records only the number of cases on which any given pair of justices voted opposite to each other. From this matrix we create a network to which we associate a reversible Markov chain. The states of the Markov chain are the nodes of the network, and in our model each node corresponds to a Supreme Court justice. We singled out the eigenvector of the transition probability matrix that corresponds to the smallest eigenvalue, normalized so that it has length 1 in L^2_π (weighted by the reversible measure $\boldsymbol{\pi}$). Then, we analyzed the function, which we termed the ideology index, which assigns to each node the corresponding entry in this eigenvector. We should remark that this function gives only a *relative* measure, as opposed to an *absolute* measure, in the sense that, in order to assign a meaning to the value at x, one has to relate it to the values at the other nodes in the graph. On the other hand, it is an objective function, free from any subjective interpretation, being defined entirely from the recorded data.

The [SCDB 2018] contains data compiled by legal scholars indicating whether a case carried ideological implications and the direction (conservative or liberal) the justices voted in every such case. For each justice, we used these data to calculate the frequency with which their votes coincided with conservative ideology, and observed a strong correlation between these frequencies and the entries in the eigenvector corresponding to the smallest eigenvalue. This observation was one of the main motivations for undergoing this study.

For the theoretical investigation, in order to study the eigenvalues of a transition matrix of a reversible Markov chain P_B obtained from a disagreement matrix $B = B(\boldsymbol{p})$ with $\boldsymbol{p} \in [0, 1]^N$, our approach was to study the spectrum of P_T for the matrix $T = \boldsymbol{p}\boldsymbol{q}^\intercal + \boldsymbol{q}\boldsymbol{p}^\intercal$ first. We found explicit formulas for the highest-energy eigenvector of P_T. It would be interesting to investigate whether similar formulas for the highest-energy eigenvector of P_B exist. Even without such formulas, numerical simulations suggest that the two eigenvectors are close to each other, and therefore have similar properties as for example monotonicity in $\boldsymbol{p}$ or boundedness estimates. One should then decide whether such properties grant the entries of the highest-energy eigenvector of P_B the quality of an objective measure of an ideology index.

Confirmation for the validity and robustness of our model can be achieved by selecting appropriately the vector $\boldsymbol{p}$ and comparing the output of the numerical implementation to the results given by formula (3-5) or (3-11). We propose two different ways to obtain an acceptable vector $\boldsymbol{p}$ adapted to the model. The first method is the one we used in order to obtain the y-axis values in the plot in Figure 1.

To obtain the vector $\boldsymbol{p}$, we used the frequencies calculated from the [SCDB 2018], as described above. The ability to calculate such frequencies was dependent upon the case-by-case legal analysis of ideology included in the database. Such comprehensive analysis may not exist for other data sets.

An alternative method that can be used to obtain the vector $\boldsymbol{p}$ is the following procedure. Consider data collected over a continuous time period, spanning the careers of N justices. In the following, by *case* we mean only those cases in the Supreme Court on which there was not an unanimous decision. Let C denote the $N \times N$ symmetric matrix with entries $C(x, y)$ representing the number of cases justices x and y disagreed on. Denote by $n(x, y)$ the number of cases justices x and y served together. For a vector $\boldsymbol{p} \in [0, 1]^N$ define the disagreement matrix $B = B(\boldsymbol{p})$ to be equal to $(\boldsymbol{p}\boldsymbol{q}^\intercal + \boldsymbol{q}\boldsymbol{p}^\intercal)$ off the diagonal, and zero on the diagonal. A natural candidate for $\boldsymbol{p}$ is obtained by minimizing over $\boldsymbol{p} \in [0, 1]^N$ the square distance between the matrix C and the matrix with entries $n(x, y)B(x, y)$. By this we mean minimizing the function

$$f(\boldsymbol{p}) := \sum_{x\neq y}(n(x,y)(p_x+p_y-2p_xp_y)-C(x,y))^2.$$

The general direction would be to investigate whether one can obtain useful estimates on the minimum value of the function f, and whether these estimates are transferable further to the distance between the highest-energy eigenvector of P_B and that of P_C. Should this program be successful, the highest-energy eigenvector of P_C would provide an objective measure of the proposed *ideology index*.

Acknowledgements

Most of this work was done while Zilli was enrolled in the Applied and Computational Mathematics Master of Arts program at St. John's University. Both authors acknowledge the support received from this institution.

The authors thank the referee for their careful reading of the paper and for useful suggestions that lead to a better exposition.

References

[Chung 1997] F. R. K. Chung, *Spectral graph theory*, CBMS Regional Conference Series in Mathematics **92**, Amer. Math. Soc., Providence, RI, 1997. MR Zbl

[Doyle and Snell 1984] P. G. Doyle and J. L. Snell, *Random walks and electric networks*, Carus Mathematical Monographs **22**, Math. Assoc. Amer., Washington, DC, 1984. MR Zbl

[Levin et al. 2009] D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov chains and mixing times*, Amer. Math. Soc., Providence, RI, 2009. MR Zbl

[Martin and Quinn 2002] A. D. Martin and K. M. Quinn, "Dynamic ideal point estimation via Markov chain Monte Carlo for the U.S. Supreme Court, 1953–1999", *Political Anal.* **10**:2 (2002), 134–153.

[SCDB 2018] H. J. Spaeth, L. Epstein, A. D. Martin, J. A. Segal, T. J. Ruger, and S. C. Benesh, "The Supreme Court database", website, 2018, available at http://supremecourtdatabase.org. version 2018, release 02.

Received: 2019-09-29 Revised: 2020-06-18 Accepted: 2020-09-08

catrinaf@stjohns.edu *Department of Mathematics and Computer Science,*
St John's University, Queens, NY, United States

bzilli@iastate.edu *Department of Mathematics, Iowa State University,*
Ames, IA, United States

INVOLVE 13:5 (2020)
https://doi.org/10.2140/involve.2020.13.823

Rings whose subrings have an identity

Greg Oman and John Stroud

(Communicated by Kenneth S. Berenhaut)

Let R be a ring. A nonempty subset S of R is a *subring* of R if S is closed under negatives, addition, and multiplication. We determine the rings R for which every subring S of R has a multiplicative identity (which need not be the identity of R).

1. Introduction

Let R be a ring (assumed only to be associative, not necessarily commutative or with 1). Recall that a nonempty subset S of R is a *subring* of R if S is closed under negatives, addition, and multiplication.[1] Suppose now that R has a 1. It is easy to see that a subring S of R need not have an identity. For instance, the subring $2\mathbb{Z}$ of $\mathbb{Z}$ consisting of the even integers has no multiplicative identity. In fact, it is easy to see that the only subrings of $\mathbb{Z}$ which have an identity are $\{0\}$ and $\mathbb{Z}$.

On the other extreme, consider the seemingly uninteresting ring $\mathbb{Z}/6\mathbb{Z}$. The subrings of $\mathbb{Z}/6\mathbb{Z}$ are

$$S_1 := \{\bar{0}\}, \quad S_2 := \{\bar{0}, \bar{3}\}, \quad S_3 := \{\bar{0}, \bar{2}, \bar{4}\} \quad \text{and} \quad S_4 := \mathbb{Z}/6\mathbb{Z}.$$

One easily verifies that $\bar{0}$ is the identity of S_1, $\bar{3}$ is the identity of S_2, $\bar{4}$ is the identity of S_3, and $\bar{1}$ is the identity of S_4. Hence every subring of $\mathbb{Z}/6\mathbb{Z}$ has an identity.

The purpose of this note is to classify the rings with the above property enjoyed by $\mathbb{Z}/6\mathbb{Z}$. That is, we shall find all rings R up to isomorphism with the property that every subring of R has an identity. This work is related to results in the literature. For example, in [Gilmer and Heinzer 1992], the authors studied commutative rings with identity with the property that every proper unital subring is Artinian; they show that such rings are precisely the Artinian rings for which every unital subring is Artinian. Now suppose that X is an infinite set and R is a binary relation on X. For $2 \le \kappa \le |X|$, we say that (X, R) is κ-homogeneous if any two subsets of size κ are isomorphic (with the order induced by R). Such structures were classified in [Droste 1989]. The first author has conducted related research in universal and commutative algebra; see [Oman 2009a; 2009b; 2011; 2017].

MSC2010: primary 16B99; secondary 13A99, 12E99.
Keywords: absolutely algebraic field, Jacobson's theorem, reduced ring.

[1] Some authors define a subring S of a ring R with identity 1_R to be *unital* if $1_R \in S$. In fact, it is commonplace for many authors to consider only rings with identity and only subrings which are unital.

2. Results

To streamline terminology, let us agree to call a nonzero ring R with the property that every subring of R has an identity *strongly unital*. As the example in the Introduction shows, the identities of the subrings need not all be the same, in stark contrast to the fact that the additive identity of a subring coincides with the additive identity of the ambient ring.

We begin by recalling that an element α of a ring R is *nilpotent* if there is a positive integer n such that $\alpha^n = 0$. If R has no nonzero nilpotent elements, then R is said to be *reduced*.

Proposition 1. *Every strongly unital ring is reduced.*

Proof. Suppose that R is a strongly unital ring and let $\alpha \in R$. As is well known, it suffices to prove that if $\alpha^2 = 0$, then $\alpha = 0$. Thus suppose $\alpha^2 = 0$ and set $S := \{n\alpha : n \in \mathbb{Z}\}$. One checks at once that S is a subring of R with trivial multiplication. But S has an identity, and therefore $S = \{0\}$. We deduce that $\alpha = 0$, as desired. □

Recall next that if R is a ring with identity, then the so-called *prime subring* $P(R)$ of R is the subring of R generated by 1_R. Since $P(R)$ is a homomorphic image of $\mathbb{Z}$, it is clear that $P(R) \cong \mathbb{Z}$ or $P(R) \cong \mathbb{Z}/n\mathbb{Z}$ for some positive integer n. We now record a trivial but useful observation and then prove several lemmas.

Observation 2. Suppose that a ring R is strongly unital. Then so is every nontrivial subring of R.

Lemma 3. *Let R be a strongly unital ring. Then $P(R) \cong \mathbb{Z}/m\mathbb{Z}$ for some square-free integer $m > 1$.*

Proof. Let R be strongly unital. Now, $P(R) \cong \mathbb{Z}/m\mathbb{Z}$ for some integer $m \geq 0$. Since $2\mathbb{Z}$ is a nonunital subring of $\mathbb{Z}$, we see that $m \neq 0$. As R is a nontrivial ring, $m \neq 1$. This shows that $m > 1$. Invoking Proposition 1, we deduce that $P(R)$ is reduced, and hence m is square-free. □

The following lemma is a well-known result in elementary field theory, but since its proof is short, we include it.

Lemma 4. *Let F be a finite field and let $f(X) \in F[X]$ be a nonzero polynomial. Then $F[X]/\langle f(X)\rangle$ is finite.*

Proof. Suppose that F is a finite field, and fix some nonzero polynomial $f(X) \in F[X]$ of degree $n \geq 0$. As is well known, the polynomial ring $F[X]$ is a Euclidean domain. Thus via the division algorithm, every member of the quotient ring $F[X]/\langle f(X)\rangle$ can be expressed in the form as $\langle f(X)\rangle + r(X)$, where $r(X) \in F[X]$ is zero or of degree less than n. It follows that $|F[X]/\langle f(X)\rangle| \leq |F|^n$, and therefore $F[X]/\langle f(X)\rangle$ is finite. □

Lemma 5. *Suppose that R is a ring with identity. The polynomial ring $R[X]$ is not strongly unital.*

Proof. If R is the trivial ring, then $R[X]$ is also trivial, and thus by definition is not strongly unital. Now suppose that R is nontrivial. Then it is easy to see that the subring $XR[X]$ (the subring of polynomials with constant term 0) does not have an identity: if $f(X) \in XR[X]$, then $X \cdot f(X) \neq X$. □

We are almost equipped to prove our next proposition; first we comment on notation. Let R be a ring with identity and let S be a subring of R contained in $Z(R)$, the center of R. Further, let $a \in R$. Then we define

$$S[a] = \{s_0 + s_1 a + \cdots + s_n a^n : n \in \mathbb{N}, s_i \in S\} = \{f(a) : f(X) \in S[X]\}. \quad (2\text{-}1)$$

Observe that $S[a]$ is a subring of R containing S, but it need *not* contain a. However, if R is unital with identity 1_R and $1_R \in S$, then $S[a]$ contains a and, moreover, $S[a]$ is the smallest subring of R containing S and a.

Proposition 6. *Suppose R is a strongly unital ring. Then for every $\alpha \in R$, there exists a positive integer n (depending on α) such that $\alpha^n = \alpha$. Therefore, R is commutative.*

Proof. Let R be a strongly unital ring and let $\alpha \in R$ be arbitrary. Recall from Lemma 3 that $P(R) \cong \mathbb{Z}/m\mathbb{Z}$ for some integer $m > 1$ which is square-free; say $m = p_1 \cdots p_k$, where the p_i are distinct primes. It follows that $P(R)$ is the internal direct sum of rings $S_1, \ldots, S_k$, where $S_i \cong \mathbb{Z}/p_i\mathbb{Z}$ for $i = 1, \ldots, k$. Clearly $P(R) \subseteq Z(R)$, and hence $S_i \subseteq Z(R)$ for $i = 1, \ldots, k$. It is straightforward to check that

$$P(R)[\alpha] = (S_1 + \cdots + S_k)[\alpha] = S_1[\alpha] + \cdots + S_k[\alpha]. \quad (2\text{-}2)$$

Fix i with $1 \leq i \leq k$. Recall that $S_i \cong \mathbb{Z}/p_i\mathbb{Z}$. Thus there are ring surjections $f : \mathbb{Z}/p_i\mathbb{Z}[X] \to S_i[X]$ and (by (2-1)) $g : S_i[X] \to S_i[\alpha]$. Letting K be the kernel of the composition, we have $S_i[\alpha] \cong \mathbb{Z}/p_i\mathbb{Z}[X]/K$. If K is trivial, then $S_i[\alpha] \cong \mathbb{Z}/p_i\mathbb{Z}[X]$. But then by Observation 2, $\mathbb{Z}/p_i\mathbb{Z}[X]$ is strongly unital, contradicting Lemma 5. We conclude that K is nontrivial. Invoking Lemma 4 (and using the fact that $\mathbb{Z}/p\mathbb{Z}[X]$ is a PID), $S_i[\alpha]$ is finite. As $1 \leq i \leq k$ was arbitrary, it follows from (2-2) above that

$$P(R)[\alpha] \text{ is a finite ring.} \quad (2\text{-}3)$$

Note that Proposition 1 implies that $P(R)[\alpha]$ is reduced. Thus as is well known (and an easy consequence of the Chinese remainder theorem),

$$P(R)[\alpha] \cong F_1 \times \cdots \times F_j \quad \text{for some finite fields } F_1, \ldots, F_j. \quad (2\text{-}4)$$

Now, for any $a \in F_i^\times$, $1 \leq i \leq j$, we have $a^{|F_i|-1} = 1$. We deduce that for any $\beta \in F_1 \times \cdots \times F_j$, we have $\beta^{(|F_1|-1)\cdots(|F_j|-1)+1} = \beta$. But then there is a positive integer n such that $\beta^n = \beta$. Applying (2-4), we see that there is a positive integer m

such that $\alpha^m = \alpha$. That R is commutative is now immediate from Jacobson's theorem; see [Herstein 1964, p. 367]. □

Remark 7. The fact that the integer m in the above proof is square-free is essential to our proof of (2-3). Indeed, suppose that $n > 1$ is an integer which is not square-free, and let N be the nilradical of $\mathbb{Z}/n\mathbb{Z}$. Then N is nontrivial and proper. Therefore, $\mathbb{Z}/n\mathbb{Z}[X]/N[X] \cong ((\mathbb{Z}/n\mathbb{Z})/N)[X]$ is infinite. Hence it is not the case that $\mathbb{Z}/n\mathbb{Z}[X]/K$ is finite for every nonzero ideal K of $\mathbb{Z}/n\mathbb{Z}[X]$.

We pause now to recall more terminology. If R is a ring and I is a (two-sided) ideal of R, then I is *indecomposable* if there do not exist nonzero ideals I_1 and I_2 of R such that $I = I_1 \oplus I_2$. A ring R is indecomposable if it is indecomposable as an ideal of itself. Our next lemma may be in the literature, but we could not locate a source. Therefore, we present a self-contained proof.

Lemma 8. *Let R be a ring, and suppose that R does not contain an ideal which is an infinite internal direct sum of nonzero ideals of R. Then $R = I_1 \oplus \cdots \oplus I_n$ for some indecomposable ideals $I_1, \dots, I_n$ of R.*

Proof. We proceed by contraposition. Thus let R be a ring, and suppose that R is not a finite direct sum of indecomposable ideals. Then R is not indecomposable as an ideal, and hence $R = I_1 \oplus J_1$ for some nonzero ideals I_1 and J_1. Since R is not a finite direct sum of indecomposable ideals, we may assume without loss of generality that J_1 is not indecomposable. Hence $J_1 = I_2 \oplus J_2$ for some nonzero ideals I_2 and J_2. Now, $R = I_1 \oplus I_2 \oplus J_2$. Again, R is not a finite direct sum of indecomposable ideals, and so we may assume without loss of generality that J_2 is not indecomposable. Thus $J_2 = I_3 \oplus J_3$ for some nonzero ideals I_3 and J_3. Thus $R = I_1 \oplus I_2 \oplus I_3 \oplus J_3$. Proceeding recursively, we see that R contains an ideal which is an infinite internal direct sum of nonzero ideals of R, and the proof is complete. □

We are almost ready to classify the strongly unital rings. First, we establish a final lemma and recall a couple of definitions. The lemma is a special case of a more general result in the literature; in [Jans 1964, p. 22], the author establishes that a left Artinian ring with no nonzero nilpotent left ideals is a semisimple ring with identity. Our next lemma is an immediate consequence of this fact.

Lemma 9. *Every finite reduced commutative ring has an identity.*

Before stating our main theorem, we remind the reader that if F is a field, then the *prime subfield* of F is the subfield of F generated by 1. It is easy to see that if F has characteristic p, then the prime subfield of F is isomorphic to $\mathbb{Z}/p\mathbb{Z}$; if F has characteristic 0, then the prime subfield of F is isomorphic to $\mathbb{Q}$. Finally, F is called *absolutely algebraic* if F is algebraic over its prime subfield. Our main result and its proof conclude this note.

Theorem 10. *Let R be a ring. Then R is strongly unital if and only if there is a positive integer n such that $R \cong F_1 \times \cdots \times F_n$, where each F_i is an absolutely algebraic field of prime characteristic.*

Proof. Assume first that R is a ring which is strongly unital. We claim that

$$\text{there is no ideal of } R \text{ which is an infinite direct sum of nonzero ideals of } R. \tag{2-5}$$

Suppose not, and let X be an infinite index set and $\{I_x : x \in X\}$ an enumeration of nonzero ideals of R which generate a direct sum. Because X is infinite, it is clear that $\bigoplus_{x \in X} I_x$ does not have a multiplicative identity. However, $\bigoplus_{x \in X} I_x$ is an ideal of R, and hence also a subring of R. This contradicts the assumption that R is strongly unital, and (2-5) is verified. It now follows from Lemma 8 (and the fact that by definition R is nontrivial) that there exist nonzero indecomposable ideals $I_1, \ldots, I_n$ of R such that $R = I_1 \oplus \cdots \oplus I_n$. Observe that the map $(i_1, \ldots, i_n) \mapsto i_1 + \cdots + i_n$ is a ring isomorphism between the external direct product $I_1 \times \cdots \times I_n$ of the rings $I_1, \ldots, I_n$ and R. We record this below:

$$R \cong I_1 \times \cdots \times I_n \text{ (as rings).} \tag{2-6}$$

To finish proving the first implication of the theorem, it remains only to show that

$$I_k \text{ is an absolutely algebraic field of prime characteristic for } 1 \le k \le n. \tag{2-7}$$

Clearly it suffices to prove the assertion for $I := I_1$. Toward this end, since I is a subring of R and R is strongly unital, there is some $1_I \in I$ which is a multiplicative identity for I. We claim that

$$\text{the only idempotents of } I \text{ are } 0 \text{ and } 1_I. \tag{2-8}$$

Indeed, if $e \ne 0, 1_I$ is an idempotent of I, then I decomposes as $I = Ie \oplus I(1_I - e)$. Now observe that both Ie and $I(1_I - e)$ are nonzero ideals of R, and we have contradicted the fact that I is indecomposable. We now easily show that I is a field. Recall from Proposition 6 that R is commutative. Next, let $r \in I$ be nonzero. Then clearly $Ir \subseteq I$ is a nonzero ideal of R, and hence has an identity element e^*. Because e^* is idempotent, we deduce from (2-8) that $e^* = 0$ or $e^* = 1_I$. Also $e^* \ne 0$, lest $Ir = \{0\}$. We conclude that $e^* = 1_I$, and thus $1_I \in Ir$. But this means that r is invertible, and I is a field, as claimed. Finally, Proposition 6 implies that I is absolutely algebraic of prime characteristic.

Conversely, suppose $R \cong F_1 \times \cdots \times F_n$, where each F_k is an absolutely algebraic field of prime characteristic, and let S be a subring of R. We shall prove that S has an identity. We may of course assume that S is nontrivial. For $1 \le i \le n$, let $\pi_i : R \to F_i$ be the projection map onto the i-th coordinate. Further, set $\pi(S) := \{1 \le i \le n : \pi_i(S) \text{ is nontrivial}\}$. Without loss of generality, we may suppose that

$\pi(S)=\{1,2,\dots,r\}$ for some r with $1\le r\le n$. For $1\le i\le r$, let $x_i\in S$ be such that

$$\pi_i(x_i)\neq 0. \tag{2-9}$$

Now let S' be the subring of S generated by $x_1,\dots,x_r$. Further, for each i with $1\le i\le n$, let K_i be the prime subfield of F_i. It is clear that, up to isomorphism, S' is a subring of

$$K_1(\pi_1(x_1),\pi_1(x_2),\dots,\pi_1(x_r))\times\cdots\times K_n(\pi_n(x_1),\pi_n(x_2),\dots,\pi_n(x_r)). \tag{2-10}$$

Recall that each K_i is finite and each $\pi_i(x_j)$ is algebraic over K_i. But then it follows that each $K_i(\pi_i(x_1),\pi_i(x_2),\dots,\pi_i(x_r))$ is a finite field, and we conclude from (2-10) that S' is finite. Applying Lemma 9, we see that S' has a multiplicative identity $1_{S'}:=(e_1,\dots,e_r,0,\dots,0)$. We claim that $1_{S'}$ is also an identity for S. Toward this end, it clearly suffices to prove that $e_i=1$ for $1\le i\le r$ (here, 1 is the multiplicative identity of F_i). To see this, simply note that $1_{S'}\cdot x_i=x_i$. Therefore, $\pi_i(1_{S'})\cdot\pi_i(x_i)=\pi_i(x_i)$. Applying (2-9) and the fact that F_i is a field, we deduce that $e_i=\pi_i(1_{S'})=1$, and the proof is complete. □

Acknowledgment

The authors thank the anonymous referees, whose comments improved the exposition of this note.

References

[Droste 1989] M. Droste, "k-homogeneous relations and tournaments", *Quart. J. Math. Oxford Ser.* (2) **40**:157 (1989), 1–11. MR Zbl

[Gilmer and Heinzer 1992] R. Gilmer and W. Heinzer, "An application of Jónsson modules to some questions concerning proper subrings", *Math. Scand.* **70**:1 (1992), 34–42. MR Zbl

[Herstein 1964] I. N. Herstein, *Topics in algebra*, Blaisdell, New York, 1964. MR Zbl

[Jans 1964] J. P. Jans, *Rings and homology*, Holt, Rinehart and Winston, New York, 1964. MR Zbl

[Oman 2009a] G. Oman, "More results on congruent modules", *J. Pure Appl. Algebra* **213**:11 (2009), 2147–2155. MR Zbl

[Oman 2009b] G. Oman, "On modules M for which $N\cong M$ for every submodule N of size $|M|$", *J. Commut. Algebra* **1**:4 (2009), 679–699. MR Zbl

[Oman 2011] G. Oman, "On elementarily κ-homogeneous unary structures", *Forum Math.* **23**:4 (2011), 791–802. MR Zbl

[Oman 2017] G. Oman, "Elementarily λ-homogeneous binary functions", *Algebra Universalis* **78**:2 (2017), 147–157. MR Zbl

Received: 2020-01-06 Revised: 2020-06-19 Accepted: 2020-08-06

goman@uccs.edu *Department of Mathematics, University of Colorado,*
Colorado Springs, CO, United States

jstroud@uccs.edu *Department of Physics, University of Colorado,*
Colorado Springs, CO, United States

INVOLVE 13:5 (2020)
https://doi.org/10.2140/involve.2020.13.829

Simple graphs of order 12 and minimum degree 6 contain K_6 minors

Ryan Odeneal and Andrei Pavelescu

(Communicated by Ronald Gould)

We prove that every simple graph of order 12 which has minimum degree 6 contains a K_6 minor, thus proving Jørgensen's conjecture for graphs of order 12. In the process, we establish several lemmata linking the existence of K_6 minors for graphs to their size or degree sequence, by means of their clique sum structure. We also establish an upper bound for the order of graphs where the 6-connected condition is necessary for Jørgensen's conjecture.

1. Introduction

All the graphs considered in this article are simple (nonoriented, without loops or multiple edges). For a graph G, *a minor of G* is any graph that can be obtained from G by a sequence of vertex deletions, edge deletions, and simple edge contractions. A simple edge contraction means identifying its endpoints, deleting that edge, and deleting any double edges thus created. A graph G is called *apex* if it has a vertex v such that $G - v$ is planar, where $G - v$ is the subgraph of G obtained by deleting vertex v and all edges of G incident to v. Jørgensen [1994] stated the following conjecture:

Conjecture 1. Let G be 6-connected graph which does not have a K_6 minor. Then G is apex.

This result relates to Hadwiger's conjecture [1943], which states:

Conjecture 2. For every integer $t \geq 1$, if a loopless graph G has no K_t minor, then it is $(t-1)$-colorable.

Conjecture 2 is known to be true for $t \leq 6$. For $t = 5$, the conjecture is equivalent to Appel and Haken's 4-Color theorem [1989]. For $t = 6$, Robertson, Seymour, and

MSC2010: 05C10, 05C83.
Keywords: Jorgensen's conjecture, K_6 minors, minimum degree 6.

Thomas [1993] proved it using a result of Mader. Mader [1968b] proved that a minimal counterexample to Conjecture 2 for $t = 6$ has to be 6-connected. Together with Jørgensen's conjecture, it would provide another proof that Conjecture 2 holds for $t = 6$, along with more information about the structure of graphs with no K_6 minors.

Jørgensen himself took steps towards proving Conjecture 1. In [Jørgensen 1994], he proved that every graph G with at most 11 vertices and minimal degree $\delta(G)$ at least 6 is contractible to a K_6. In his proof, he used the following result of [Mader 1968a]:

Theorem 3. *Every simple graph with minimal degree at least* 5 *either has a minor isomorphic to* K_6^- *or it has a minor isomorphic to the icosahedral graph.*

The icosahedral graph is the only 5-regular planar graph on 12 vertices. Mader [1968a] also proved the following theorem.

Theorem 4. *For every integer* $2 \le t \le 7$ *and every simple graph* G *of order* $n \ge t-1$ *which has no minor isomorphic to* K_t, G *has at most* $(t-2)n - \binom{t-1}{2}$ *edges.*

Note that for $t = 6$, the theorem implies that every graph G of order n and size $4n - 9$ or more has a K_6 minor.

Jørgensen [1988] classified the graphs of order n and size $4n - 10$.

Theorem 5. *Let* p *be a natural number*, $5 \le p \le 7$. *Let* G *be a graph with* n *vertices and* $(p-2)n - \binom{p}{2}$ *edges that is not contractible to* K_p. *Then either* G *is an* MP_{p-5}*-cockade or* $p = 7$ *and* G *is the complete* 4*-partite graph* $K_{2,2,2,3}$.

For $p = 6$, this theorem shows that any graph G of order n and size $4n - 10$ either contains a K_6 minor, or it is an MP_1-cockade. The following is Jørgensen's definition of an MP_1-cockade.

Definition 6. MP_1-cockades are defined recursively as follows:

(1) K_5 is an MP_1-cockade and if H is a 4-connected maximal planar graph then $H * K_1$ is an MP_1-cockade.

(2) Let G_1 and G_2 be disjoint MP_1-cockades, and let x_1, x_2, x_3, and x_4 be the vertices of a K_4 subgraph of G_1 and let y_1, y_2, y_3, and y_4 be the vertices of a K_4 subgraph of G_2. Then the graph obtained from $G_1 \cup G_2$ by identifying x_j and y_j, for $j = 1, 2, 3, 4$, is an MP_1-cockade.

For two graphs G_1 and G_2, we denote by $G_1 * G_2$ the graph with vertex set $V(G_1) \sqcup V(G_2)$ and edge set $E(G_1) \sqcup E(G_2) \sqcup E'$, where E' is the set of edges with one endpoint in $V(G_1)$ and the other endpoint in $V(G_2)$. In $G_1 * v$, we call v *a cone* over G_1. A graph G is the *clique sum* of G_1 and G_2 over K_p if $V(G) = V(G_1) \cup V(G_2)$, $E(G) = E(G_1) \cup E(G_2)$ and the subgraphs induced by $V(G_1) \cap V(G_2)$ in both G_1 and G_2 are complete of order p. In this context, an MP_1-cockade is either a cone over a 4-connected maximal planar graph or the clique sum over K_4 of two smaller MP_1-cockades.

Kawarabayashi, Norine, Thomas, and Wollan [2018] proved that Conjecture 1 holds for sufficiently large graphs. Little is known about the validity of Conjecture 1 for small order graphs. In this paper, we prove that Jørgensen's conjecture holds for graphs of order 12 in a more general setting.

Theorem 7. *Let G be a simple graph of order* 12 *and assume that* $\delta(G) \geq 6$, *where* $\delta(G)$ *denotes the minimal degree of G. Then G contains a K_6 minor.*

Note that the theorem implies Jørgensen's conjecture is vacuously true for graphs of order 12.

2. Main theorem

For a graph G, we denote by $V(G)$ its vertex set and by $E(G)$ its edge set. The size of $V(G)$ is called *the order* of G, and the cardinality of $E(G)$ is called *the size* of G. For $n \geq 1$, K_n denotes the complete graph of order n and K_n^- denotes the complete graph of order n with one edge removed. If $v_1, v_2, \dots, v_k$ are vertices of G, then $\langle v_1, v_2, \dots, v_k \rangle_G$ denotes the subgraph of G induced by these vertices. If v is a vertex of G, then $N_G[v]$ is the subgraph of G induced by v and the vertices adjacent to v in G (the closed neighborhood of v). Let $N_G(v)$ denote the subgraph of G induced by all the vertices adjacent to v (the open neighborhood of v). If S is a subset of $V(G)$, then $G - S$ is the subgraph of G obtained by deleting all of the vertices in S and all the edges of G to which S is incident.

The following lemma is a corollary of Theorem 4.

Lemma 8. *Let G be a simple graph of order n and size* $4n - 10$. *If $G - v$ is planar, then v cones over $G - v$.*

Proof. Since $G - v$ is planar of order $n - 1$, it has at most $3(n-1) - 6 = 3n - 9$ edges. This implies that v has at least $4n - 10 - (3n - 9) = n - 1$ neighbors, and the conclusion follows. □

Proof of Theorem 7. Let G denote a simple graph of order 12 and minimal degree $\delta(G)$ at least six. It follows that G has at least size 36. By Theorem 4, if the size of G is at least 39, then G contains a K_6 minor. We shall prove Theorem 7 by considering the size of G, $36 \leq |E(G)| \leq 38$.

Case 1: Assume $|E(G)| = 38$. By Theorem 5, either G contains a K_6 minor, or G is apex, or G is the clique sum over K_4 of two MP_1 cockades.

If G is isomorphic to $H * K_1$, where H is a maximal planar graph on 11 vertices, then $\delta(H) \geq 5$ and, by Theorem 3, it follows that H has a K_6^- minor and thus G has a K_6 minor.

Assume that G is the clique sum over $S \simeq K_4$ of two MP_1-cockades. If $G - S$ has more than two connected components, then at least one of them has at most two vertices. But this contradicts the fact that $\delta(G) \geq 6$. So $G - S = Q_1 \sqcup Q_2$. Furthermore, unless $|Q_1| = |Q_2| = 4$, the graph either contains a K_7 subgraph (if $|Q_1|$=3), or $\delta(G) < 6$ (if $1 \leq |Q_1| \leq 2$). Since $\delta(G) \geq 6$, it follows that each vertex of Q_i connects to at least two other vertices of Q_i, for $i = 1, 2$ respectively.

Without loss of generality, let $Q_1 = \langle v_1, v_2, v_3, v_4 \rangle_G$. If Q_i is not isomorphic to K_4 for any of $i = 1, 2$, say $v_1 v_2 \notin E(Q_1)$, since v_1 and v_2 must both connect to both v_3 and v_4, contracting the edges $v_1 v_3$ and $v_2 v_4$ produces a minor of G which contains a K_6 subgraph induced by v_1, v_2, and the four vertices of S. If, on the other hand, $Q_1 \simeq Q_2 \simeq K_4$, as $\delta(G) \geq 6$, it follows that there are at least 12 edges between each of the Q_i and S. That would imply that $|E(G)| \geq 6 + 12 + 6 + 12 + 6 = 42$, a contradiction. It follows that for $|E(G)| = 38$, G has a K_6 minor.

Case 2: Assume $|E(G)| = 37$. Since $\delta(G) \geq 6$, it follows that the degree sequence of G is either $(6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 8)$ or $(6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 7, 7)$. In either of the situations, we shall need the following lemma:

Lemma 9. *Let M denote a graph of order* 11 *and size* 34 *such that* $\delta(M) \geq 5$. *Assume that M is not apex and has at most four vertices of degree* 5. *Then M contains a K_6 minor.*

Proof. By Theorem 5, either M contains a K_6 minor or is an MP_1-cockade. Since M is not apex, it follows that M is the clique sum over $S \simeq K_4$ of two MP_1-cockades. If $M - S$ has more than two connected components, $Q_1, Q_2, \ldots$, then at least one of them, say Q_1, has at most two vertices. As $|Q_1| = 1$ would violate the condition $\delta(M) \geq 5$, it follows that $|Q_1| = 2$ and the subgraph of M induced by Q_1 and S forms a K_6. So $M - S = Q_1 \sqcup Q_2$ and, without loss of generality, $Q_1 = \langle v_1, v_2, v_3 \rangle_M$. Unless $Q_1 \simeq K_3$, since $\delta(M) \geq 5$, it follows that at least two vertices of Q_1 connect to all the vertices of S and thus, via an edge contraction, they induce a K_6 minor of M.

If $Q_1 \simeq K_3$, there have to be exactly nine edges connecting the vertices of Q_1 to those of S. If there are more than nine, the subgraph induced by the vertices of Q_1 and S has seven vertices and more than $3 + 9 + 6 = 18$ edges; thus it contains a K_6 minor by Theorem 4. If there are less than nine, then at least one of the vertices of Q_1 has degree less than 5. So all the vertices of Q_1 have degree 5, and the subgraph induced by the vertices of Q_1 and S has exactly 18 edges. If L denotes the set of edges connecting the vertices of Q_2 to the vertices of S, then $|L| + |E(Q_2)| = 34 - 18 = 16$. On the other hand, since Q_2 can have at most one vertex of degree 5 in M, it follows that $|L| + 2|E(Q_2)| \geq 6 + 6 + 6 + 5 = 23$. Subtracting the last two equalities we get $|E(Q_2)| \geq 7$, a contradiction as Q_2 has four vertices. □

Assume the vertex degree sequence for G is $(6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 8)$. Furthermore, without loss of generality, we may assume $\deg_G(v_1) = 6$, $\deg_G(v_8) = 8$, and that $N_G[v_1] = \langle v_1, \dots, v_7\rangle_G$. Let $N = \langle v_2, v_3, \dots, v_7\rangle_G$, $H = \langle v_8, \dots, v_{12}\rangle_G$, and let L denote the set of edges of G with one endpoint in N and the other in H. The handshaking lemma provides the following relations between the sizes of $E(N)$, L, and $E(H)$:

$$2|E(N)| + |L| = 30, \quad 2|E(H)| + |L| = 32.$$

If $|E(H)| = 10$, that is, $H \simeq K_5$, as every vertex of H must have at least two neighbors in N ($\delta(G) \geq 6$), contracting the edges of $N_G[v_1]$, produces a K_6 minor of G.

If $|E(H)| \leq 9$, then $|L| \geq 14$ and thus $|E(N)| \leq 8$. It follows that there is a vertex of N, say v_2, such that $\deg_N(v_2) \leq 2$. If $\deg_N(v_2) < 2$, contracting the edge v_1v_2 would produce a minor of G of order 11 and size at least 35, which would contain a K_6 minor by Theorem 4.

If $\deg_N(v_2) = 2$, then contracting the edge v_1v_2 would produce a minor M of G of order 11 and size precisely 34. Furthermore, since v_2 neighbors exactly three vertices of H, the maximum degree of M is 8, so it cannot be apex, according to Lemma 8. Lastly, M has at most two vertices of degree 5, since $\deg_N(v_2) = 2$. By Lemma 9, M has a K_6 minor, and therefore so does G.

Assume the vertex degree sequence for G is $(6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 7, 7)$. If the degree-7 vertices are connected in G, deleting the edge connecting them would produce a 6-regular subgraph of order 36, to be dealt with in the last case of the proof. So, without loss of generality, assume that $\deg_G(v_1) = 6$, $\deg_G(v_8) = 7$, and $N_G[v_1] = \langle v_1, \dots, v_7\rangle_G$. Let $N = \langle v_2, v_3, \dots, v_7\rangle_G$, $H = \langle v_8, \dots, v_{12}\rangle_G$, and let L denote the set of edges of G with one endpoint in N and the other in H. If $\deg(v_i) = 7$, for some $9 \leq i \leq 12$, the same argument as before shows that $|E(N)| \leq 8$ and thus G contains a K_6 minor. So we may assume $\deg_G(v_7) = 7$ and $v_7v_8 \notin E(G)$. Using the handshaking lemma, we get

$$2|E(N)| + |L| = 31, \quad 2|E(H)| + |L| = 31.$$

If $|E(H)| = 10$, contracting the edges of $N_G[v_1]$, produces a K_6 minor of G.

If $|E(H)| \leq 9$, then $|L| \geq 13$ and thus $|E(N)| \leq 9$. If $|E(N)| \leq 8$, it follows that there is a vertex of N, v_i, such that $\deg_N(v_i) \leq 2$. If $\deg_N(v_i) < 2$, contracting the edge v_1v_i would produce a minor of G of order 11 and size at least 35, which would contain a K_6 minor by Theorem 4.

Assume $\deg_N(v_i) = 2$. Contracting the edge v_1v_i produces a minor M of G of order 11 and size 34. Moreover, since for $2 \leq j \leq 7$, v_j neighbors at most four of the vertices of H, M cannot be apex. Lastly, M has at most two vertices of degree 5, since $\deg_N(v_i) = 2$. By Lemma 9, M has a K_6 minor, and therefore so does G.

It follows that N is 3-regular, $L = 13$ and $|E(H)| = 9$; that is, $H \simeq K_5^-$. If the missing edge of H has v_8 as its endpoint, and since $\deg_G(v_8) = 7$, it follows that v_8 neighbors four vertices of N. As the other endpoint, say v_9, neighbors three vertices

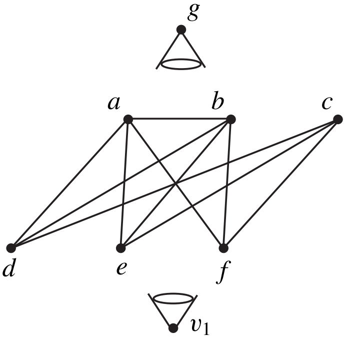


Figure 1. Contracting the edges cd and eg produces a K_6-minor.

of N, it follows that there exists $2 \leq i \leq 6$ such that v_i is a common neighbor of v_8 and v_9. Contracting the edges $v_i v_8$ and $v_1 v_j$, for $2 \leq j \leq 7$, $j \neq i$, one obtains a K_6 minor of G, as every vertex of H neighbors at least one of the v_j.

Assume the missing edge of H is $v_9 v_{10}$. Since N is 3-regular of order 6, it is isomorphic to either $K_{3,3}$ or the prism graph. If $N \simeq K_{3,3}$, as v_8 neighbors three vertices of N, contracting the edge connecting v_8 to one of its neighbors in N and all the edges in the subgraph induced by v_9, v_{10}, v_{11}, and v_{12}, produces a minor of G isomorphic to the graph in Figure 1. This minor has a K_6-minor.

If N is isomorphic to the prism graph in Figure 2, with the same labeling, for any vertex of v of H, the subgraph of G induced by its neighbors among the vertices of N must be complete (clique), since otherwise contracting v to one of its neighbors and the edges of $H - v$ produces a minor of G isomorphic to the graph in Figure 3. This graph has a K_6-minor.

If v_9 and v_{10} share a common neighbor among the vertices of N, say v_i, then contracting the edge $v_i v_9$ and $v_1 v_j$, for $2 \leq j \leq 7$, $j \neq i$, produces a K_6-minor. If v_9 and v_{10} have no common neighbor among the vertices of N, since v_9, v_{10} and v_8 each have exactly three neighbors among the vertices of N, and v_7 is not adjacent to v_8, up to a relabeling of v_9 and v_{10}, it must be that v_8 and v_9 together with v_2, v_4, and v_6 induce a K_5 subgraph of G. Contracting all the edges of $\langle v_1, v_7, v_{10}, v_{11}, v_{12} \rangle_G$ produces a K_6-minor of G.

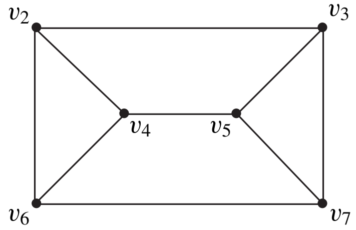


Figure 2. The graph N, the open neighborhood of v_1.

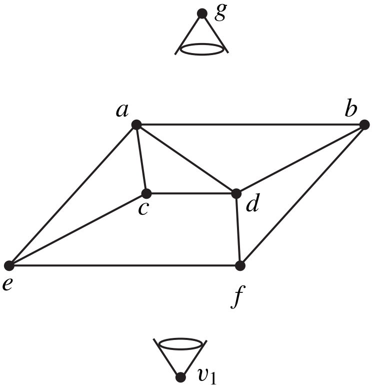


Figure 3. Contracting the edges bg and ef produces a K_6-minor.

Case 3: Assume $|G| = 36$; that is, G is a 6-regular graph. Let $v_2, v_3, \ldots, v_7$ be the neighbors of some vertex v_1 in G and let $N = \langle v_2, v_3, \ldots, v_7 \rangle_G = N_G(v_1)$ be the open neighborhood of v_1. Let $H = \langle v_8, \ldots, v_{12} \rangle_G$ and let L denote the subset of $E(G)$ of edges having one endpoint in N and the other in H. Then, as before, since the degree of every vertex in G is 6, we have

$$2|E(H)| + |L| = 30, \quad 2|E(N)| + |L| = 30;$$

thus $|E(N)| = |E(H)|$. If $|E(H)| = 10$, then $H \simeq K_5$. Since $\delta(G) = 6$ by hypothesis, each vertex in H must be adjacent to a vertex in N. Contracting the edges of $N_G[v_1]$ produces a K_6 minor of G. It follows that $|E(N)| \leq 9$ and, unless N is 3-regular, there exists at least a vertex of N which has at most two neighbors in N. There are two possible remaining cases: either the open neighborhood of every vertex of G is 3-regular, or there is a vertex of G whose open neighborhood contains a vertex of degree at most 2. Jørgensen [1994] proved that in a 6-regular graph, if the open neighborhood of every vertex of G is 3-regular, then any connected component of the graph is isomorphic to either $K_{3,3,3}$ or the complement of the Petersen graph. Since both contain K_6 minors, it suffices to consider the case $\deg_N(v_i) \leq 2$, for some $2 \leq i \leq 7$.

If, for some $2 \leq i \leq 7$, $\deg_N(v_i) = 0$, then contracting the edge $v_1 v_i$ produces a minor of G of order 11 and size 35. By Theorem 4, this minor has a K_6 minor.

If, for some $2 \leq i \leq 7$, $\deg_N(v_i) = 1$, then contracting the edge $v_1 v_i$ produces a minor M of G of order 11 and size 34. Furthermore, the degree sequence of this minor would be (5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 9); hence, by Lemma 8, M cannot be apex. By Lemma 9, M would have a K_6 minor.

We may then assume that, for $2 \leq i \leq 7$, $\deg_N(v_i) \geq 2$ and, without loss of generality, the neighbors of v_2 in N are v_3 and v_4.

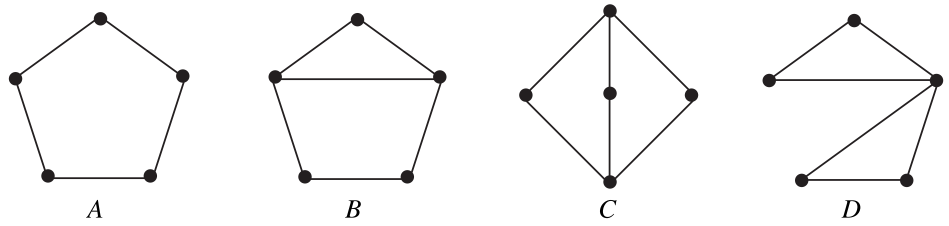


Figure 4. Graphs of order 5 and size at most 6, with minimum degree 2.

Subcase 3.1: Assume that $v_3v_4 \in E(G)$. Contracting the edge v_1v_2 produces a minor M of G of order 11 and size 33. Furthermore, the degree sequence of this minor is $(5, 5, 6, 6, 6, 6, 6, 6, 6, 6, 8)$. Via a relabeling, for the rest of this subcase, we may assume that $\deg_M(v_1) = \deg_M(v_2) = 5$, and that $\deg_M(v_6) = 8$. As before, let $N = \langle v_2, v_3, v_4, v_5, v_6\rangle_M$ be the subgraph of M induced by all the neighbors of v_1, $H = \langle v_7, \dots, v_{11}\rangle_M$, and L the subset of edges of M with one endpoint in N and the other in H. Adding degrees we get

$$2|E(H)| + |L| = 30, \quad 2|E(N)| + |L| = 26;$$

thus $|E(H)| = |E(N)| + 2$. Note that $|E(H)| \le 8$, since if $|E(H)| = 10$, contracting the edges of $N_M[v_1]$ produces a K_6 minor of M; if $|E(H)| = 9$, that is, $H \simeq K_5^-$, then, without loss of generality, assume $v_7v_8 \notin E(H)$. Since $\deg_M(v_7) = \deg_M(v_8) = 6$, it follows that both v_7 and v_8 have each three neighbors among the five vertices of N; thus they share a common neighbor in N, say v_i. Contracting the edges v_iv_7 and v_1v_j, for $2 \le j \ne i \le 6$, we obtain a K_6 minor of M.

We may then assume that $|E(H)| \le 8$, and thus $|E(N)| \le 6$. If any vertex v_i of N has at most one neighbor in N, contracting the edge v_1v_i would produce a minor of M of order 10 and size at least 31. By Theorem 4, this would contain a K_6 minor. So every vertex of N has at least two neighbors in N. If follows that $5 \le |E(N)| \le 6$ and, by [Read and Wilson 1998], N is isomorphic to one of the four graphs in Figure 4.

If $\deg_N(v_2) = 2$, then contracting the edge v_1v_2 produces a minor of M of order 10 and size 30, with degree sequence $(5, 6, 6, 6, 6, 6, 6, 6, 6, 7)$. The following lemma shows that M contains a K_6 minor.

Lemma 10. *Let M denote a graph of order* 10 *and size* 30 *such that* $\delta(M) \ge 5$. *Assume that M is not apex and has at most five vertices of degree* 5. *Then M contains a K_6 minor.*

Proof. By Theorem 5, either M contains a K_6 minor or is an MP_1-cockade. Since M is not apex, it follows that M is the clique sum over $S \simeq K_4$ of two MP_1-cockades. Since $\delta(M) \ge 5$, any connected component of $M - S$ is of size at least 2. If any of the connected components of $M - S$ has exactly two vertices, then that component together with S induces a K_6 subgraph of M. The only situation left to discuss is when $M - S$ has exactly two size-3 connected components, Q_1 and Q_2. At least one of them, say Q_1, contains at most two vertices of degree 5. If we denote by L' the set of edges connecting the vertices of Q_1 to the vertices of S, then

$$2|E(Q_1)| + |L'| \geq 6 + 5 + 5 = 16.$$

Hence

$$|E(Q_1)| + |L'| \geq 13 \quad \Longrightarrow \quad |E(Q_1)| + |L'| + |E(S)| \geq 19;$$

thus Q_1 and S induce a subgraph of M of order 7 and size 19. By Theorem 4, this subgraph contains a K_6 minor. □

If $\deg_N(v_2) = 3$, then N is isomorphic to either graph B or graph C in Figure 4. Furthermore, v_2 neighbors in N a vertex (say v_3) of total degree 6 which has degree 2 in N and does not neighbor the vertex of degree 8. Contracting the edge v_1v_3 produces a minor P of M of order 10 and size 30. Furthermore, the degree sequence of this minor is (4, 5, 6, 6, 6, 6, 6, 6, 7, 8) and thus it is not apex. By Theorem 5, P either contains a K_6 minor or it is the clique sum over $S \simeq K_4$ of two MP_1-cockades. Every vertex of S has degree at least 5 since it must connect to every connected component of $P - S$. Let Q_1 denote the connected component of $P - S$ which contains the vertex of degree 4. Let H denote the graph induced by the vertices of $P - (S \cup Q_1)$ and let L'' denote the set of edges of P with one endpoint in S and the other in H.

If $|V(Q_1)| = 1$, then

$$\begin{aligned} |L''| + |E(H)| &= 20, \\ |L''| + 2|E(H)| &\geq 6 + 6 + 6 + 6 + 5 = 29. \end{aligned}$$

It follows that $|E(H)| \geq 9$; that is, $H \simeq K_5^-$ or $H \simeq K_5$. For $H \simeq K_5$, as every vertex of H is adjacent to at least a vertex of S, contracting S produces a K_6 minor of P. If $H \simeq K_5^-$, assume $a, b \in V(H)$ and $ab \notin E(H)$. If $\deg_P(a) = 5$ or $\deg_P(b) = 5$, then a and b share at least one common neighbor s in S. Contracting the edge sa and then contracting the edges of the graph induced by the vertices of $Q_1 \cup S - \{s\}$ produces a K_6 minor of P, as every vertex of H other than a or b has degree at least 6 and must therefore be adjacent to at least two vertices of S. Finally, if $\deg_P(a) \geq 6$ and $\deg_P(b) \geq 6$, then a and b share at least two neighbors among the vertices of S, say s_1 and s_2. Since $V(H) \backslash \{a, b\}$ contains at most one vertex of degree 5, that vertex is adjacent to at most one of $\{s_1, s_2\}$, say s_1. Contracting the edge s_2a and then contracting the edges of the graph induced by the vertices of $Q_1 \cup S - \{s_2\}$ produces a K_6 minor of P.

If $|V(Q_1)| = 2$, then $\langle Q_1 \cup S\rangle_P \simeq K_6^-$ and

$$\begin{aligned}|L''| + |E(H)| &= 16,\\ |L''| + 2|E(H)| &\geq 6+6+6+6 = 24.\end{aligned}$$

It follows that $|E(H)| \geq 8$, which is a contradiction (H has only four vertices). It follows that $|V(Q_1)| = 3$ and that $\langle S \cup H\rangle_P$ contains a K_7^- minor.

If $\deg_N(v_2) = 4$, then $N \simeq D$ of Figure 4 and contracting the edge connecting v_1 to the common neighbor of v_2 and v_6 in N produces a minor P of M of order 10 and size 30, with degree sequence $(4, 6, 6, 6, 6, 6, 6, 6, 7, 7)$, where the two degree-7 vertices and the degree-4 vertex form a triangle. Theorem 5 shows that P is a clique sum over $S \simeq K_4$ of two MP_1-cockades. Furthermore $P - S$ has exactly two connected components, Q_1 and Q_2. As any vertex that's part of the clique has at least degree 5 in P, we may assume that the vertex of degree 4 is a vertex of Q_1. Unless $|Q_1| = 1$, both Q_1 and Q_2 will contain vertices of degree at least 6 in P; hence $|Q_1| = |Q_2| = 3$. But this implies that contracting any edge incident to the vertex of degree 4 in Q_1 produces a K_6 minor of the graph induced by $Q_1 \sqcup S$.

If $|V(Q_1)| = 1$, then let L'' denote the set of edges in P with one endpoint in K_4 and the other in Q_2. It follows that

$$\begin{aligned}|L''| &= 7+7+6+6-12-4 = 10,\\ |E(Q_2)| &= 10;\end{aligned}$$

hence $Q_2 \simeq K_5$ and thus contracting the edges of the subgraph induced by $Q_1 \sqcup K_4$ produces a K_6 minor.

Subcase 3.2: Assume that $v_3v_4 \notin E(G)$. Contracting the edge v_1v_2 produces a minor M of G of order 11 and size 33. Furthermore, the degree sequence of this minor is $(5, 5, 6, 6, 6, 6, 6, 6, 6, 6, 8)$. Via a relabeling, for the rest of this subcase, we may assume that $\deg_M(v_1) = \deg_M(v_7) = 5$, and that $\deg_M(v_6) = 8$. As before, let $N = \langle v_2, v_3, v_4, v_5, v_6\rangle_M$ be the subgraph of M induced by all the neighbors of v_1, $H = \langle v_7, \ldots, v_{11}\rangle_M$, and L the subset of edges of M with one endpoint in N and the other in H. Adding degrees we get

$$2|E(H)| + |L| = 29, \quad 2|E(N)| + |L| = 27;$$

thus $|E(H)| = |E(N)|+1$. Furthermore, if $|E(H)| = 10$, that is, $H \simeq K_5$, contracting v_1v_i for $2 \leq i \leq 6$ produces a K_6 minor.

Assume $|E(H)| = 9$ and $|E(N)| = 8$. Then by [Read and Wilson 1998], N is isomorphic to one of the two graphs in Figure 5.

If $N \simeq A$ in Figure 5, as all vertices of N have minimum degree 6, contracting the vertices of H (which is connected) to a single point and then further contracting

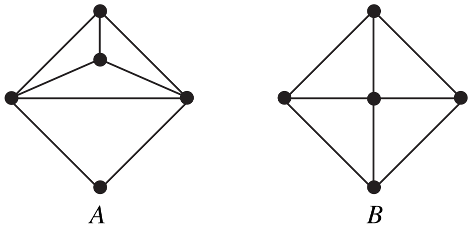


Figure 5. Graphs of order 5 and size 8, with minimum degree 2.

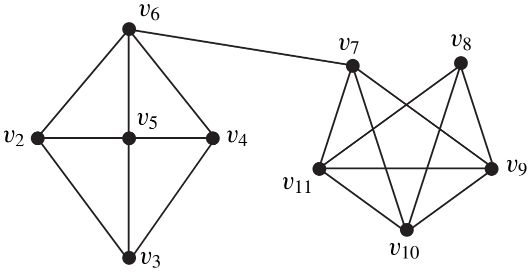


Figure 6. Subcase 3.2.1: $\deg_N(v_6) = 3$ and $\deg_H(v_7) = 3$.

the edge joining the newly obtained point and the only vertex of degree 2 in N produces a K_6 minor.

If $N \simeq B$ in Figure 5, then we distinguish four cases based on the position of v_6 and v_7 inside N and H, respectively.

Subcase 3.2.1: Assume $\deg_N(v_6) = 3$ and $\deg_H(v_7) = 3$; see Figure 6. Without loss of generality, assume $\deg_H(v_8) = 3$; that is, v_7v_8 is the only edge missing in the complete graph on the vertices of H. If v_8 neighbors v_6, contracting the edge v_6v_8 and all the edges of $\langle v_1, v_2, v_3, v_4, v_5\rangle_M$ produces a K_6 minor of M. If v_6 does not neighbor v_8, then contracting the edges of $\langle v_1, v_2, v_3, v_4, v_5, v_8\rangle_M$ produces a K_6 minor of M.

Subcase 3.2.2: Assume $\deg_N(v_6) = 4$ and $\deg_H(v_7) = 3$; see Figure 7. Without loss of generality, we may assume $\deg_H(v_8) = 3$. If v_7 and v_8 share a neighbor in N, say v_j, then contracting v_jv_7 and then all the edges of $\langle v_1, N - v_j\rangle_M$, we obtain

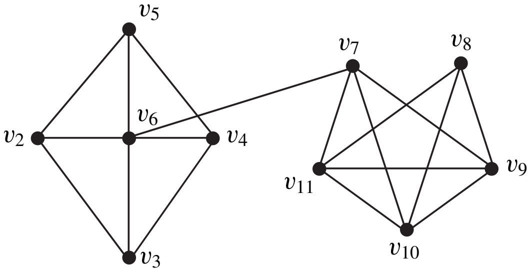


Figure 7. Subcase 3.2.2: $\deg_N(v_6) = 4$ and $\deg_H(v_7) = 3$.

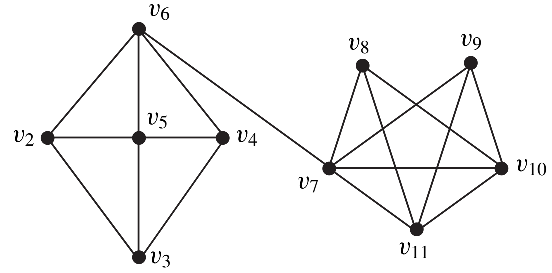


Figure 8. Subcase 3.2.3: $\deg_N(v_6) = 3$ and $\deg_H(v_7) = 4$.

a K_6 minor of G. So, as a set, $\{v_7, v_8\}$ neighbors all the vertices of N. Without loss of generality, $v_5v_7, v_6v_7, v_5v_9 \in E(M)$, $v_3v_5 \notin E(N)$. If v_3 neighbors either of v_{10} or v_{11}, say v_{10}, then contracting the edges v_3v_{10}, v_5v_9 will connect v_3 and v_5, contracting all the edges of $\langle v_2, v_7, v_8, v_{11}\rangle_M$ will connect v_2 and v_4, and thus we obtain a K_6 minor. But then, v_3 must neighbor v_9, and contracting the edge v_3v_9 and all the edges of $\langle v_7, v_8, v_{10}, v_{11}, v_2\rangle_M$ produces a K_6 minor.

Subcase 3.2.3: Assume $\deg_N(v_6) = 3$ and $\deg_H(v_7) = 4$; see Figure 8. Without loss of generality, we may assume $\deg_H(v_8) = \deg_H(v_9) = 3$ and $\deg_N(v_5) = 4$. Since v_8 and v_9 each connect to three vertices of N, they have a common vertex in N. If this vertex is not v_6, say v_i, then contracting the edges v_iv_8 and v_1v_j for $2 \leq j \neq i \leq 6$, we obtain a K_6 minor of M. So $\{v_7, v_8, v_9\}$, as a set, neighbors all the vertices of N. Since v_3 does not neighbor v_7 and cannot neighbor both v_8 and v_9, it must neighbor one of v_{10} or v_{11}. But then, contracting the edges of $\langle v_6, v_{10}, v_{11}\rangle_M$ and then contracting the edges of $\langle v_7, v_8, v_9, v_2\rangle_M$, we obtain a K_6 minor of M.

Subcase 3.2.4: Assume $\deg_N(v_6) = 4$ and $\deg_H(v_7) = 4$; see Figure 9. Without loss of generality, we may assume that $\deg_H(v_8) = \deg_H(v_9) = 3$. Since both v_8 and v_9 connect to three vertices of N, they must share at least one common neighbor in N. If that common neighbor is not v_6, say v_2, contracting the edges v_2v_8, and v_1v_i for $3 \leq i \leq 6$, we obtain a K_6 minor of M. It follows that v_6 neighbors v_7, v_8 and v_9 and that, as a set, $\{v_8, v_9\}$ neighbors all the vertices of N. If $\{v_{10}, v_{11}\}$ neighbors,

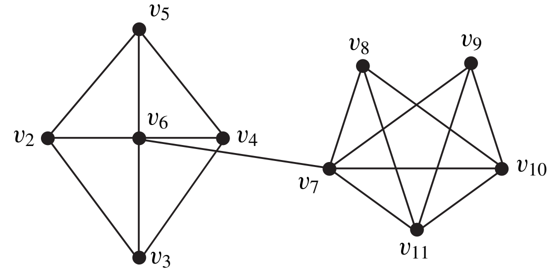


Figure 9. Subcase 3.2.4: $\deg_N(v_6) = 4$ and $\deg_H(v_7) = 4$.

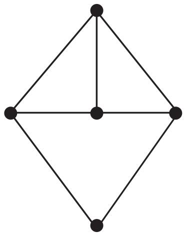


Figure 10. The unique graph of order 5, size 7, and minimum degree 2, with exactly one vertex of degree 2.

as a set, any nonneighbors in N, say v_2 and v_4, then contracting the edges $v_{10}v_{11}$, v_2v_{10}, and all the edges of $\langle v_5, v_7, v_8, v_9\rangle_M$, we obtain a K_6 minor of M. If not, as v_6 neighbors neither v_{10} nor v_{11}, we know that v_{10} and v_{11} must both neighbor an edge of N not incident to v_6, say v_2v_3. But since v_2 neighbors $v_1, v_3, v_5, v_6, v_{10}$ and v_{11} and one of v_8 or v_9, it follows that v_2 has degree 7 in M, a contradiction.

It follows that $|E(N)| \leq 7$. If any of $v_2, \dots, v_6$ have degree at most 1 in N, contracting the edge connecting that vertex to v_1 would produce minor of M of order 10 and size at least 31, which would have a K_6 minor by Theorem 4. This shows that $\delta(N) \geq 2$ and that $|E(N)| \geq 5$. Furthermore, if any of the degree-6 neighbors of v_1 have degree 2 in N, contracting the edge connecting that neighbor to v_1 would produce a nonapex graph of order 10 and size 30, with minimal degree at least 5, and at most three vertices of degree 5. By Lemma 10, this graph would contain a K_6 minor. This observation handles the cases $|E(N)| = 5$ and $|E(N)| = 6$, since any graph on five vertices with minimum degree 2 and size at most 6 has at least two vertices of degree exactly 2.

Assume that $|E(N)| = 7$, $\delta(N) = 2$, and N has only one vertex of degree 2. Then N is isomorphic to the graph in Figure 10. Furthermore, $\deg_N(v_6)$ is 2; thus v_6 neighbors all the vertices of H. If v_7 neighbors two or more vertices of N, then contracting v_1v_i for $2 \leq i \leq 5$ and H to one of its K_4 minors (H has 8 edges and 5 vertices, by Theorem 4 it has a K_4 minor) we obtain a K_6 minor. It follows that v_7 connects to v_8, v_9, v_{10} and v_{11}. Furthermore, the open neighborhood of v_7 contains exactly eight edges. By symmetry between v_1 and v_7 and Subcase 3.2, $|E(N)| = 8$, it follows that M has a K_6 minor. □

3. Future explorations

(1) Is it true that any simple graph of order at most 14 and minimum degree at least 6, which is not apex, contains a K_6-minor? Note that in the proof of Theorem 7, we used weaker versions of Lemmas 9 and 10. Similar lemmas hold for graphs of orders 13 and 12, respectively. They provide a first step in generalizing Theorem 7 for graphs of order at most 14.

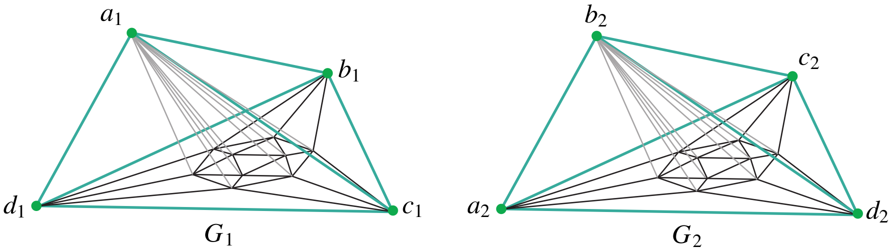


Figure 11. Graphs G_1 and G_2 are identified along the highlighted tetrahedra to obtain the graph G.

(2) The result of this paper shows that, for graphs of order 12, weaker assumptions are needed for the conclusion of Jørgensen's conjecture to be true. What is the minimum $n > 12$ for which the condition of minimum degree 6 is no longer sufficient and the 6-connected condition is needed? Such n would have to be at most 22, as the following example demonstrates.

Let $G_1 \simeq G_2 \simeq K_1 * \mathrm{Ic}$, where Ic denotes the icosahedral graph (5-regular, maximal planar, order 12). Let G denote a clique sum over K_4 of G_1 and G_2, done in such a way that the cones are not identified to each other (so that the maximum degree of G is 15). In Figure 11, a_1 and a_2, b_1 and b_2, c_1 and c_2, and d_1 and d_2 are respectively identified. Then $\delta(G) = 6$, G is not apex, and it has no K_6 minor.

Acknowledgements

The authors would like to thank Elena Pavelescu for the useful comments and the help with the illustrations.

References

[Appel and Haken 1989] K. Appel and W. Haken, *Every planar map is four colorable*, Contemporary Mathematics **98**, American Mathematical Society, Providence, RI, 1989. MR Zbl

[Hadwiger 1943] H. Hadwiger, "Über eine Klassifikation der Streckenkomplexe", *Vierteljschr. Naturforsch. Ges. Zürich* **88** (1943), 133–142. MR Zbl

[Jørgensen 1988] L. K. Jørgensen, "Extremal graphs for contractions to K_7", *Ars Combin.* **25**:C (1988), 133–148. MR Zbl

[Jørgensen 1994] L. K. Jørgensen, "Contractions to K_8", *J. Graph Theory* **18**:5 (1994), 431–448. MR Zbl

[Kawarabayashi, Norine, Thomas, and Wollan 2018] K.-i. Kawarabayashi, S. Norine, R. Thomas, and P. Wollan, "K_6 minors in large 6-connected graphs", *J. Combin. Theory Ser. B* **129** (2018), 158–203. MR Zbl

[Mader 1968a] W. Mader, "Homomorphiesätze für Graphen", *Math. Ann.* **178** (1968), 154–168. MR Zbl

[Mader 1968b] W. Mader, "Über trennende Eckenmengen in homomorphiekritischen Graphen", *Math. Ann.* **175** (1968), 243–252. MR Zbl

[Read and Wilson 1998] R. C. Read and R. J. Wilson, *An atlas of graphs*, The Clarendon Press, New York, 1998. MR Zbl

[Robertson, Seymour, and Thomas 1993] N. Robertson, P. Seymour, and R. Thomas, "Hadwiger's conjecture for K_6-free graphs", *Combinatorica* **13**:3 (1993), 279–361. MR Zbl

Received: 2020-01-28 Revised: 2020-06-27 Accepted: 2020-06-28

rao1521@jagmail.southalabama.edu
Department of Mathematics and Statistics,
University of South Alabama, Mobile, AL, United States

andreipavelescu@southalabama.edu
Department of Mathematics and Statistics,
University of South Alabama, Mobile, AL, United States

INVOLVE 13:5 (2020)
https://doi.org/10.2140/involve.2020.13.845

Mixed volume of small reaction networks

Nida Obatake, Anne Shiu and Dilruba Sofia

(Communicated by Suzanne Lenhart)

An important invariant of a chemical reaction network is its maximum number of positive steady states. This number, however, is in general difficult to compute. Nonetheless, there is an upper bound on this number — namely, a network's mixed volume — that is easy to compute. Moreover, recent work has shown that, for certain biological signaling networks, the mixed volume does not greatly exceed the maximum number of positive steady states. Continuing this line of research, we further investigate this overcount and also compute the mixed volumes of small networks, those with only a few species or reactions.

1. Introduction

For chemical reaction networks, information about steady states — both their number and their nature (stability, etc.) — yields insight into a network's capacity for processing information. Therefore, there have been numerous investigations into the capacity for multiple steady states, especially for networks arising from biology; see, e.g., [Banaji and Pantea 2016; Conradi et al. 2017; Craciun et al. 2006; Dickenstein et al. 2019; Joshi and Shiu 2015; Torres and Feliu 2019].

The next step, determining the maximum number of steady states of a given network, is more difficult. Indeed, this question, mathematically, asks us to compute the maximum number of positive roots of a family of parametrized polynomial systems. Therefore, we are interested in upper bounds on this maximum number that are easy to compute.

One such bound, introduced in [Obatake et al. 2019], is the mixed volume of a network; see also the closely related definitions in [Gross et al. 2016; Gross and Hill 2019]. This bound is surprisingly good for certain biological signaling networks, with the "mixed-volume overcount" — the difference between the mixed volume and the maximum number of steady states — no more than 2 or 4 [Obatake et al. 2019]. Related results for three infinite families of networks are obtained in [Gross and Hill 2019] and for a Wnt shuttle model in [Gross et al. 2016].

MSC2020: primary 52A39, 12D10, 37C10; secondary 65H04, 80A30.
Keywords: chemical reaction network, steady state, Newton polytope, mixed volume.

Here we further investigate the mixed volume and the mixed-volume overcount, with a focus on small networks, those with just a few species or reactions. Our results are as follows. First, for networks with only one species, we show how to read off the mixed volume (and mixed-volume overcount) directly from the network (Theorems 3.2 and 3.4), and conclude that the mixed-volume overcount can be arbitrarily large (Corollary 3.3). Next, we investigate networks with two species and two reactions and show that among those that are at-most-bimolecular, nearly all have mixed-volume overcount 0 (Theorem 3.13). Thus, the mixed volume is an excellent bound for such networks.

The outline of our work is as follows. We provide background in Section 2. Our main results are presented in Section 3, and we end with a discussion in Section 4.

2. Background

Below, we give background on chemical reaction systems (Section 2A), their steady states (Section 2B), mixed volume (Section 2C), and networks having only one species (Section 2D).

2A. *Chemical reaction systems.*

Definition 2.1. A *reaction network* $G := (\mathcal{S}, \mathcal{C}, \mathcal{R})$ consists of three finite sets: a set of *species* $\mathcal{S} := \{A_1, A_2, \dots, A_s\}$, a set $\mathcal{C} := \{y_1, y_2, \dots, y_p\}$ of *complexes* (finite nonnegative-integer combinations of the species), and a set of *reactions*, which are ordered pairs of complexes, excluding diagonal pairs: $\mathcal{R} \subseteq (\mathcal{C} \times \mathcal{C}) \smallsetminus \{(y, y) \mid y \in \mathcal{C}\}$.

Throughout our work, s and r denote the numbers of species and reactions, respectively. A reaction network is *genuine* if every species takes part in at least one reaction.

Writing the i-th complex as $y_{i1}A_1 + y_{i2}A_2 + \cdots + y_{is}A_s$ (here, $y_{ij} \in \mathbb{Z}_{\geq 0}$ is the *stoichiometric coefficient* of A_j for $j = 1, 2, \dots, s$), this complex is *at-most-bimolecular* if $y_{i1} + y_{i2} + \cdots + y_{is} \leq 2$. A reaction network is *at-most-bimolecular* if every complex in the network is at-most-bimolecular.

It is customary to write a reaction (y_i, y_j) as $y_i \to y_j$, and y_i is called the *reactant* and y_j is the *product*. Also, a reaction $y_i \to y_j$ is *reversible* if its reverse reaction $y_j \to y_i$ is also in $\mathcal{R}$, and we denote such a pair by $y_i \rightleftharpoons y_j$. A reaction $y_i \to y_j$ defines the *reaction vector* $y_j - y_i$, which encodes the net change in each species resulting from the reaction. The *stoichiometric matrix* Γ is the $s \times r$ matrix whose k-th column is the reaction vector of the k-th reaction. Each reaction comes with a *rate constant* κ_{ij}, which is a positive parameter.

Next, we let $x_1, x_2, \dots, x_s$ represent the concentrations of the s species, which we view as functions $x_i(t)$ of time t. Also, we define the monomial $\boldsymbol{x}^{y_i} := x_1^{y_{i1}} x_2^{y_{i2}} \cdots x_s^{y_{is}}$.

A *chemical reaction system* is the dynamical system that arises, via mass-action kinetics, from a chemical reaction network $(\mathcal{S}, \mathcal{C}, \mathcal{R})$ and a choice of rate constants $(\kappa_{ij}^*) \in \mathbb{R}_{>0}^r$ (recall that r is the number of reactions) as follows:

$$\frac{d\boldsymbol{x}}{dt} = \sum_{y_i \to y_j \text{ is in } \mathcal{R}} \kappa_{ij} \boldsymbol{x}^{y_i} (y_j - y_i) =: f_\kappa(\boldsymbol{x}). \tag{1}$$

Viewing the rate constants as a vector of parameters $\kappa = (\kappa_1, \kappa_2, \ldots, \kappa_m)$, we have polynomials $f_{\kappa,i} \in \mathbb{Q}[\kappa, x]$ for $i = 1, 2, \ldots, s$. For simplicity, we will write f_i rather than $f_{\kappa,i}$.

The *stoichiometric subspace*, $S := \operatorname{span}(\{y_j - y_i \mid y_i \to y_j \text{ is in } \mathcal{R}\})$, is the vector subspace of $\mathbb{R}^s$ spanned by all reaction vectors $y_j - y_i$. Thus, $S = \operatorname{im}(\Gamma)$, where Γ is the stoichiometric matrix. Let $d = s - \operatorname{rank}(\Gamma)$. A *conservation-law matrix* of G, denoted by W, is a row-reduced $(d \times s)$-matrix whose rows form a basis of the orthogonal complement of S.

A trajectory $x(t)$ that starts at a positive vector $x(0) = x^0 \in \mathbb{R}_{>0}^s$ remains, for all positive time, in the following *stoichiometric compatibility class* with respect to the *total-constant vector* $c := Wx^0 \in \mathbb{R}^d$:

$$\mathcal{S}_c := \{x \in \mathbb{R}_{\geq 0}^s \mid Wx = c\}. \tag{2}$$

Example 2.2. Consider the network $G = \{2A \xrightarrow{k_1} 2B,\ B \xrightarrow{k_2} A\}$. This network has $r = 2$ nonreversible reactions involving $p = 4$ distinct complexes — represented as vectors $(2, 0)$, $(0, 2)$, $(0, 1)$, $(1, 0)$ — on $s = 2$ species, A and B. Also, the network is genuine and at-most-bimolecular. The stoichiometric matrix of G is

$$\Gamma = \begin{bmatrix} -2 & 1 \\ 2 & -1 \end{bmatrix}.$$

The stoichiometric subspace S, which has dimension $d = 1$, is spanned by $(1, -1)^{\mathrm{T}}$, and a conservation-law matrix of G is $W = [1 \ \ 1]$. Let $\boldsymbol{x}(t) = (x_1(t), x_2(t)) \in \mathbb{R}_{\geq 0}^2$ denote the vector of concentrations of species A and B. A conservation law for G is $x_1 + x_2 = c_1$ for $c_1 \in \mathbb{R}_{\geq 0}$. The chemical reaction system of G arising from mass-action kinetics is

$$\frac{d\boldsymbol{x}}{dt} = \begin{pmatrix} -2k_1x_1^2 + k_2x_2 \\ 2k_1x_1^2 - k_2x_2 \end{pmatrix}.$$

2B. *Steady states.* For a chemical reaction system, a *steady state* is a nonnegative concentration vector $\boldsymbol{x}^* \in \mathbb{R}_{\geq 0}^s$ at which the right-hand side of the ODEs (1) vanish: $f_\kappa(\boldsymbol{x}^*) = 0$. We will focus on *positive steady states* $\boldsymbol{x}^* \in \mathbb{R}_{>0}^s$.

To analyze steady states in a stoichiometric compatibility class, we use conservation laws in place of linearly dependent steady-state equations as follows. Let $I = \{i_1 < i_2 < \cdots < i_d\}$ denote the indices of the first nonzero coordinate of the rows of the conservation-law matrix W. For a total-constant vector c, define the function $f_{c,\kappa} : \mathbb{R}^s_{\geq 0} \to \mathbb{R}^s$ as

$$f_{c,\kappa,i} = f_{c,\kappa}(x)_i := \begin{cases} f_i(x) & \text{if } i \notin I, \\ (Wx - c)_k & \text{if } i = i_k \in I. \end{cases} \tag{3}$$

The system (3) is called the system *augmented by conservation laws*.

Remark 2.3. For networks without conservation laws, the augmented system is just the original system $f_{c,\kappa}$ in (1).

Definition 2.4. (1) A network is *multistationary* if there exist positive rate constants κ_{ij} such that, for the corresponding chemical system (1), there is some stoichiometric compatibility class (2) having two or more positive steady states.

(2) A network *admits k positive steady states* (for some $k \in \mathbb{Z}_{\geq 0}$) if there exists a choice of positive rate constants so that the resulting mass-action system has exactly k positive steady states in some stoichiometric compatibility class.

The *maximum number of positive steady states* of a network G is the maximum value of k (with $k \in \mathbb{Z}_{\geq 0}$) for which G admits k positive steady states.

2C. *Mixed volume.* Here we recall from [Obatake et al. 2019] the mixed volume of a network, which is in general an upper bound on the maximum number of positive steady states. For background on convex and polyhedral geometry, see [Ewald 1996; Ziegler 1995]. In particular, for a polynomial $f = b_1 x^{\sigma_1} + b_2 x^{\sigma_2} + \cdots + b_\ell x^{\sigma_\ell} \in \mathbb{R}[x_1, x_2, \ldots, x_s]$, where the exponent vectors $\sigma_i \in \mathbb{Z}^s$ are distinct and $b_i \neq 0$ for all i, the *Newton polytope* of f is the convex hull of its exponent vectors: $\mathrm{Newt}(f) := \mathrm{conv}\{\sigma_1, \sigma_2, \ldots, \sigma_\ell\} \subseteq \mathbb{R}^s$.

Definition 2.5. Let $P_1, P_2, \ldots, P_s \subseteq \mathbb{R}^s$ be polytopes. The volume of the Minkowski sum $\lambda_1 P_1 + \lambda_2 P_2 + \cdots + \lambda_s P_s$ is a degree-s homogeneous polynomial in nonnegative variables $\lambda_1, \lambda_2, \ldots, \lambda_s$. In this polynomial, the coefficient of $\lambda_1 \lambda_2 \cdots \lambda_s$, denoted by $\mathrm{Vol}(P_1, P_2, \ldots, P_s)$, is the *mixed volume* of $P_1, P_2, \ldots, P_s$.

Definition 2.6. Let G be a network with s species, r reactions, and a $d \times s$ conservation-law matrix W. Let $f_{c,\kappa}$, as in (3), denote the resulting system augmented by conservation laws. Let $c^* \in \mathbb{R}^d_{\neq 0}$, and let $\kappa^* \in \mathbb{R}^r_{>0}$ be generic. Let $P_1, P_2, \ldots, P_s \subset \mathbb{R}^s$ be the Newton polytopes of $f_{c^*,\kappa^*,1}, f_{c^*,\kappa^*,2}, \ldots, f_{c^*,\kappa^*,s}$, respectively. The *mixed volume of G* (with respect to W) is the mixed volume of $P_1, P_2, \ldots, P_s$.

The mixed volume (Definition 2.6) is well-defined [Obatake et al. 2019, Remark 8]. The next result follows from Bernstein's theorem [1975]; see [Obatake et al. 2019, Proposition 8]:

Proposition 2.7. *For every network, the following inequality relates the maximum number of positive steady states and the mixed volume* (*with respect to any conservation-law matrix*):

$$\textit{maximum number of positive steady states} \le \textit{mixed volume}. \tag{4}$$

The *mixed-volume overcount* measures how tight the bound (4) is. Of particular interest are networks with 0 mixed-volume overcount, because for these networks, the mixed volume precisely and efficiently calculates the maximum number of positive steady states.

Definition 2.8. The *mixed-volume overcount* of a reaction network G is

$$(\text{mixed volume of } G) - (\text{maximum number of positive steady states of } G).$$

An example considered in earlier work is the extracellular signal-regulated kinase (ERK) network. This is an important biological signaling network known to be multistationary (and also bistable) [Obatake et al. 2019; Rubinstein et al. 2016]. For the ERK network and several simplified versions of the network, the mixed-volume overcount is 2 — for the fully irreversible and reduced subnetworks — or (conjectured to be) 4 — for the full network and the subnetwork obtained by removing one reaction (specifically, the reaction k_{on}) [Obatake et al. 2019, Proposition 9 and Conjecture 1].

Here, our focus is not on directly investigating multistationarity. Instead, we analyze the mixed-volume overcount for small networks as a tool for future work studying multistationarity in larger networks (e.g., the ERK network and the network in Example 3.15).

2D. *One-species networks.*

Here we recall some definitions from [Joshi and Shiu 2017].

Definition 2.9. Let G be a reaction network containing only one species A. Each reaction of G therefore has the form $aA \to bA$, where $a, b \ge 0$ and $a \ne b$. Let m be the number of (distinct) reactant complexes, and let $a_1 < a_2 < \cdots < a_m$ be the stoichiometric coefficients. The *arrow diagram* of G, denoted by $\rho = (\rho_1, \ldots, \rho_m)$, is the element of $\{\to, \leftarrow, \leftarrow\!\bullet\!\to\}^m$ with

$$\rho_i := \begin{cases} \to & \text{if for all reactions } a_i A \to bA \text{ in } G\text{, we have } b > a_i, \\ \leftarrow & \text{if for all reactions } a_i A \to bA \text{ in } G\text{, we have } b < a_i, \\ \leftarrow\!\bullet\!\to & \text{otherwise.} \end{cases}$$

Definition 2.10. For nonnegative integers $T \ge 0$, a *T-alternating network* is a one-species network with exactly $T + 1$ reactions and with arrow diagram $\rho \in \{\to, \leftarrow\}^{T+1}$ such that, if $T \ge 1$, we have $\rho_i = \to$ if and only if $\rho_{i+1} = \leftarrow$ for all $i \in \{1, 2, \ldots, T\}$.

Example 2.11. Consider the network

$$G = \{0 \leftarrow A \rightarrow 2A \rightleftharpoons 3A\}.$$

Two 1-alternating subnetworks of G, $\{A \rightarrow 2A, 2A \leftarrow 3A\}$ and $\{2A \rightarrow 3A, 2A \leftarrow 3A\}$, have arrow diagram $(\rightarrow, \leftarrow)$. On the other hand, $\{0 \leftarrow A, A \rightarrow 2A\}$ is *not* a 1-alternating subnetwork of G: its arrow diagram is $(\leftarrow\!\bullet\!\rightarrow)$. Finally, $\{0 \leftarrow A, 2A \rightarrow 3A, 2A \leftarrow 3A\}$ is a 2-alternating subnetwork of G with arrow diagram $(\leftarrow, \rightarrow, \leftarrow)$.

The following result follows directly from [Joshi and Shiu 2017, Theorem 3.6] and its proof:

Proposition 2.12 (number of steady states for one-species networks). *Let G be a reaction network with only one species* (*and at least one reaction*). *Then*, *the maximum number of positive steady states of G equals the maximum value of $T \in \mathbb{Z}_{\geq 0}$ for which G has a T-alternating subnetwork.*

3. Results

In Section 3A, we characterize the mixed volume and mixed-volume overcount of networks with only one reaction or one species. As a consequence, we show that the mixed-volume overcount can be arbitrarily large (Corollary 3.3). Subsequently, in Section 3B, we show that nearly all (genuine) networks with two species and two reactions have mixed-volume overcount 0 (Theorem 3.13).

3A. *Networks with only one reaction or one species.*

Proposition 3.1 (mixed volume of one-reaction networks). *For a network with only a single reaction*, *the mixed volume is* 0 *and the mixed-volume overcount is* 0.

Proof. Let G be a network with only one reaction. The right-hand side of the ODE consists of a single monomial, so the Newton polytope is just a point (the exponent vector of the monomial). Hence, the mixed volume of G is 0, and so the mixed-volume overcount is 0, by Proposition 2.7. □

Theorem 3.2 (mixed volume of one-species networks). *Let G be a reaction network that contains only one species A. Let m be the number of* (*distinct*) *reactant complexes*, *and let $a_1 < a_2 < \cdots < a_m$ be their stoichiometric coefficients. Then*

$$\textit{mixed volume of } G = a_m - a_1.$$

Proof. As G has only one species, there are no conservation laws and only one differential equation. In this equation, the leading monomial is $x_1^{a_m}$, and the lowest-degree monomial is $x_1^{a_1}$. The Newton polytope of this single polynomial is therefore the line segment between a_1 and a_m. Thus, by definition, the mixed volume of G is $a_m - a_1$. □

Corollary 3.3. *The mixed-volume overcount can be arbitrarily large.*

Proof. Consider the network $0 \underset{k_2}{\overset{k_1}{\rightleftharpoons}} nA$, where $n \in \mathbb{N}$. The right-hand side of the mass-action ODEs (1) is the polynomial $-k_2a^n + k_1$, which has precisely one positive real root (namely, $a = \sqrt[n]{k_1/k_2}$). However, by Theorem 3.2, the mixed volume is n. So, the mixed-volume overcount is $n - 1$. □

Theorem 3.4 (one-species networks with mixed-volume overcount 0). *Let G be a reaction network that contains only one species A. Let m be the number of* (*distinct*) *reactant complexes*, *and let $a_1 < a_2 < \cdots < a_m$ be their stoichiometric coefficients. Then G has mixed-volume overcount* 0 *if and only if G has an* $(m-1)$-*alternating subnetwork and $a_i = a_1 + i - 1$ for all $i \in \{2, 3, \dots, m\}$.*

Proof. This result follows directly from Proposition 2.12 and Theorem 3.2. □

Example 3.5 (Example 2.11 continued). By Theorem 3.4, the network from Example 2.11 has mixed-volume overcount 0. Indeed, it is a one-species network with three distinct reactant complexes (note that 0 is not a reactant complex in this network) satisfying $a_i = a_1 + i - 1$ for $i \in \{2, 3\}$ (here the notation is as in Theorem 3.4 with $a_1 = 1$), and it has a 2-alternating subnetwork.

3B. ***Networks with two species and two reactions.*** Up to relabeling species, there are 210 genuine, at-most-bimolecular networks with two species and two reactions [Banaji 2017]. These networks, which were enumerated by Banaji, were originally available at https://reaction-networks.net/networks/. A file containing the list of these networks also can be found in the repository https://github.com/neeedz/mixedvolume. Here we determine that 92% of these networks have mixed-volume overcount 0 (Theorem 3.13); the 16 exceptional networks are listed in Table 1.

The following result, which follows directly from [Joshi and Shiu 2017, Lemmas 2.7 and 4.1, and Theorem 4.8] (cf. Corollary 4.12 and the preceding paragraph in that work), implies that the 210 networks we consider in this subsection are *not* multistationary.

Proposition 3.6. *If G is an at-most-bimolecular reaction network with exactly two species and two reactions*, *then the maximum number of positive steady states of G is at most* 1. *Moreover*, *this maximum number is* 1 *if the two reaction vectors of G are negative scalar multiples of each other*, *and* 0 *otherwise.*

Proposition 3.6 and the definition of mixed-volume overcount directly yield the following:

Corollary 3.7. *Let G be an at-most-bimolecular reaction network with exactly two species and two reactions. If the mixed volume of G is at least* 2, *then the mixed-volume overcount is at least* 1.

We use the following procedure to compute (by using PHCpack [Gross et al. 2013], as in [Obatake et al. 2019]) the mixed-volume overcount of a two-species, two-reaction network:

Procedure 3.8. *Input*: A two-species, two-reaction network G.
Output: The mixed-volume overcount of G.

(0) Compute the system augmented by conservation laws (3), denoted by $f_{c,\kappa}$, for some choice of conservation-law matrix W.

(1) Compute the mixed volume of G as follows. Viewing the two polynomials in $f_{c,\kappa}$ as polynomials in x_1 and x_2, substitute 1 for all coefficients; let poly1 and poly2 be the resulting polynomials. Next, run the following Macaulay2 code:

```
loadPackage "PHCpack"
S = CC[x1,x2];
F = {poly1 , poly2};
mixedVolume(F)
```

(2) Compute the maximum number of positive steady states:

(a) If G has no linear conservation laws, the maximum number of positive steady states is 0.

(b) If G has a linear conservation law, determine the maximum number of positive steady states of G by analyzing the possible numbers of positive roots of $f_{c,\kappa} = 0$ (or by other means, e.g., if applicable, Proposition 3.6).

(3) Output the difference between the mixed volume (from step (1)) and the maximum number of positive steady states (from step (2)).

Proof of correctness of Procedure 3.8. The correctness of step (1) is due to the fact that mixed volume considers only the supports of polynomials. The correctness of step (2a) follows from [Joshi and Shiu 2017, Lemma 4.1]. Step (2b) is correct by construction of $f_{c,\kappa}$. Finally, the correctness of step (3) follows directly from the definition of mixed-volume overcount (Definition 2.8). □

Example 3.9. Consider $G = \{A + B \xrightarrow{k_1} 2B \xleftarrow{k_2} 2A\}$.

(0) The system augmented by conservation laws is

$$\begin{cases} f_1(x_1, x_2) = x_1 + x_2 - c_1, \\ f_2(x_1, x_2) = 2k_2 x_1^2 + k_1 x_1 x_2. \end{cases} \tag{5}$$

(1) Take $k_1 = 2k_2 = -c_1 = 1$ in (5), and compute the mixed volume of the resulting polynomial system. The mixed volume of the network is 1.

(2) We compute the maximum number of steady states:

(a) There is a linear conservation law (namely, f_1), so continue to step (2b).

(b) The reaction vectors, $(-1, 1)$ and $(-2, 2)$, are *not* negative scalar multiples of each other. So, by Proposition 3.6, the maximum number of positive steady states is 0. Alternatively, notice that $f_2(x_1^*, x_2^*) > 0$ when $x_1^*, x_2^* > 0$, and so $f_{c,\kappa} = 0$ never has positive roots.

(3) The mixed-volume overcount is $1 - 0 = 1$.

Next we provide two more examples of genuine two-species, two-reaction networks. These examples show that determining the maximum number of positive steady states by analyzing the roots of $f_{c,\kappa} = 0$ (step (2b) of Procedure 3.8) is not straightforward in general.

Example 3.10 (Example 2.2 continued). Recall the genuine two-species, two-reaction network $\{2A \xrightarrow{k_1} 2B,\ B \xrightarrow{k_2} A\}$. Using Procedure 3.8, we show below that the mixed-volume overcount of the network is 1.

(0) The system augmented by conservation laws is

$$\begin{cases} f_1(x_1, x_2) = x_1 + x_2 - c_1, \\ f_2(x_1, x_2) = 2k_1 x_1^2 - k_2 x_2. \end{cases} \tag{6}$$

(1) Take $2k_1 = -k_2 = -c_1 = 1$ in (6), and compute the mixed volume of the resulting polynomial system. The mixed volume of the network is 2.

(2) We compute the maximum number of steady states:

(a) There is a linear conservation law (namely, f_1), so continue to step (2b).

(b) The reaction vectors are $(-2, 2)$ and $(1, -1)$, which are negative scalar multiples of each other. So, by Proposition 3.6, the maximum number of positive steady states is 1. Alternatively, we analyze the roots of $f_{c,\kappa} = 0$, as follows. First, $f_1 = 0$ yields $x_2 = c_1 - x_1$, which we substitute into $f_2 = 0$ to get

$$g(x_1) = 2k_1 x_1^2 - k_2(c_1 - x_1) = 2k_1 x_1^2 + k_2 x_1 - k_2 c_1.$$

This is a quadratic in x_1 with positive leading coefficient and negative vertical intercept (since $k_1, k_2, c_1 > 0$). Thus, for every choice of $k_1, k_2, c_1 > 0$, the quadratic has a unique positive real root in x_1, namely,

$$x_1^* = \frac{-k_2 + \sqrt{k_2^2 + 8c_1 k_1 k_2}}{4k_1}.$$

Therefore, the maximum number of steady states is at most 1. In fact, this number is 1: when $k_1 = \frac{1}{2}$, $k_2 = 1$ and $c_1 = 2$, there is a unique positive steady state, namely, $(x_1^*, x_2^*) = (1, 1)$.

(3) The mixed-volume overcount is $2 - 1 = 1$.

Example 3.11. Let $G = \{2A \xrightarrow{k_1} 2B \xrightarrow{k_2} A + B\}$.

(0) The system augmented by conservation laws is

$$\begin{cases} f_1(x_1, x_2) = x_1 + x_2 - c_1, \\ f_2(x_1, x_2) = 2k_1x_1^2 - k_2x_2^2. \end{cases} \tag{7}$$

(1) Take $2k_1 = -k_2 = -c_1 = 1$ in (7), and compute the mixed volume of the resulting polynomial system. The mixed volume of the network is 2.

(2) We compute the maximum number of steady states:

(a) There is a linear conservation law (namely, f_1), so continue to step (2b).

(b) The reaction vectors, $(-2, 2)$ and $(1, -1)$, are negative scalar multiples of each other. So, Proposition 3.6 implies that the maximum number of positive steady states is 1. An alternate approach is as follows. We solve $f_2 = 0$ for x_2 (and use the fact that we are interested in only positive x_1, x_2), which yields $x_2^* = (\sqrt{2k_1/k_2})x_1^*$. Next, we substitute this expression into $f_1 = 0$ and then solve to obtain $x_1^* = c_1/(1 + \sqrt{2k_1/k_2})$. Thus, the network always admits a unique positive steady state (x_1^*, x_2^*).

(3) The mixed-volume overcount is $2 - 1 = 1$.

Remark 3.12. The approaches that we present in this section for computing the maximum number of steady states of a network (steps (2a) and (2b) of Procedure 3.8) rely on the fact that the networks are at-most-bimolecular and have only two reactions and two species. In general, however, completing step (2) is not straightforward: as mentioned in the Introduction, it requires counting the number of positive real roots of a parametrized polynomial system. This complication further motivates the need for graphical, algebraic, and geometric tools for counting positive steady states, in order to bypass a direct analysis of the polynomial system $f_{c,\kappa} = 0$.

By applying Procedure 3.8, we obtain a classification of genuine, at-most-bimolecular networks with two species and two reactions (Theorem 3.13).

Theorem 3.13 (mixed volume of two-species, two-reaction networks). *Let G be a genuine, at-most-bimolecular network with two species and two reactions. Then G has mixed-volume overcount* 0 *if and only if G is* (*up to relabeling species*) ***not*** *one of the* 16 *networks listed in Table 1. Moreover, each network in Table 1 has mixed-volume overcount* 1.

Proof. Using Procedure 3.8 we computed the mixed-volume overcount for all genuine two-species, two-reaction networks; see `MV-overcount-2s-2r-networks.csv` in the repository https://github.com/neeedz/mixedvolume. More details are as follows. Among the 210 networks, 185 of them have mixed volume 0 and thus have mixed-volume overcount 0. For the remaining 25 networks (see the Appendix), it

	network	mixed vol.
(1)	$2A \to 2B \to A+B$	2
(2)	$2A \to 2B, \quad B \to A$	2
(3)	$2A \to A, \quad B \to A+B$	2
(4)	$B \to A, \quad 2A \to A+B$	2
(5)	$B \to A, \quad 2B \to A+B$	1
(6)	$2A \rightleftharpoons 2B$	2
(7)	$2A \to A+B \leftarrow 2B$	2
(8)	$B \to A, \quad 2B \to 2A$	1
(9)	$B \to 2B, \quad A \to A+B$	1
(10)	$2B \to 0, \quad A \to A+B$	2
(11)	$A \rightleftharpoons 2B$	2
(12)	$A+B \to 2B \leftarrow 2A$	1
(13)	$2A \to A+B \to 2B$	1
(14)	$2A \to A, \quad A+B \to B$	1
(15)	$A+B \rightleftharpoons 0$	2
(16)	$B \to A, \quad A+B \to 2A$	1

Table 1. Genuine, at-most-bimolecular networks with two species and two reactions for which the mixed-volume overcount is nonzero. Each network has mixed-volume overcount 1.

is straightforward to compute the maximum number of positive steady states using Proposition 3.6 or by directly analyzing the system $f_{c,\kappa} = 0$ as in Examples 3.9–3.11. □

We end this section by investigating why the networks in Table 1 have nonzero mixed-volume overcount. These 16 networks fall into four classes:

(1) Networks (3), (9), (10), and (14) are essentially one-species networks (for each network, one of the two ODEs is 0), and so can be analyzed using the results in Section 3A.

(2) Networks (6), (11), and (15) consist of a single pair of reversible reactions, so (e.g., by Proposition 3.6) the maximum number of positive steady states is 1.

(3) Networks (5), (8), (12), (13), and (16) have one species that is consumed in every reaction (while the other species is produced). Thus, the maximum number of positive steady states is 0.

(4) Networks (1), (2), (4), and (7) (and also networks (3), (6), (10), (11), and (15)) have mixed volume 2, so, by Corollary 3.7, the mixed-volume overcount is at least 1.

Remark 3.14. In Examples 3.10 and 3.11, we computed the maximum number of positive steady states (step (2) of Procedure 3.8) by reducing the system $f_{c,\kappa} = 0$ to a single univariate polynomial and then checking that the positive roots (which can be viewed as "partial solutions") can be extended to positive roots of the original system. Doing this for general networks, however, is difficult. Indeed, for readers with knowledge of algebraic geometry, we note that the extension theorem [Cox et al. 2007, pp. 118–120] requires an algebraically closed field and polynomials with a certain shape.

Example 3.15. Consider the following network with three species and ten reactions:

$$\begin{gathered} 0 \rightleftharpoons A, \quad 0 \rightleftharpoons B, \quad 0 \rightleftharpoons C, \\ 2A \rightleftharpoons A+B \rightleftharpoons B+C. \end{gathered}$$

This network has no conservation laws, and its augmented system is

$$\begin{cases} f_1 = k_1 - k_2x_1 - k_7x_1^2 + (k_8 - k_9)x_1x_2 + k_{10}x_2x_3, \\ f_2 = k_3 - k_4x_2 + k_7x_1^2 - k_8x_1x_2, \\ f_3 = k_5 - k_6x_3 + k_9x_1x_2 - k_{10}x_2x_3. \end{cases}$$

Analyzing this augmented system is challenging, and determining the maximum number of steady states of the network is not straightforward. This number is at least 2 [Joshi and Shiu 2015], and we compute that its mixed volume is 6. What is the mixed-volume overcount? Our wish is to answer this question in the future through a generalized version of Procedure 3.8. Such a procedure would then give us a way to study multistationarity in this and other large networks.

4. Discussion

Recall that our interest in the mixed volume of a reaction network comes from the fact that it bounds the maximum number of positive steady states. We saw in previous work that this bound is surprisingly good for certain signaling networks, and here we again found that this bound performs well for small networks that are at-most-bimolecular. As networks arising in biological applications are typically at-most-bimolecular, we might expect the mixed-volume overcount to be low for biological networks of small to medium size.

Another future research direction pertains to one aim of this work, which is to read off the mixed volume directly from a network. We now can do this for networks with just one reaction or one species (Section 3A). As for at-most-bimolecular networks with two reactions and two species, the mixed volume is (with the exception of the 16 networks in Table 1) exactly the maximum number of positive steady states, which can be ascertained using results in [Joshi and Shiu 2017]. We would like similar results for networks with more reactions or more species.

Continuing this line of investigation, we ask, *How do operations on networks affect the mixed volume* (*and thus the mixed-volume overcount*)? For instance, in Table 1, networks (1) and (7) can be obtained from each other by “stretching” one reaction (without changing the reactant or reaction vector); and similarly for networks (2) and (4). Moreover, this operation does not affect the mixed volume or the overcount. (This line of investigation therefore would be somewhat similar in spirit to [Rojas 1994; Bihan and Soprunov 2019].) Indeed, having a list of operations and their effect on the mixed volume would greatly aid our classification of networks.

Appendix: Networks with nonzero mixed volume

Below, we list the 25 genuine two-species, two-reaction networks with nonzero mixed volume, together with their maximum number of positive steady states and their augmented systems. The first 16 networks here coincide with those listed in Table 1.

network	mixed vol.	max #	system
$2A \to 2B \to A+B$	2	1	$\begin{cases} a+b-c_1, \\ 2k_1a^2-k_2b^2 \end{cases}$
$2A \to 2B, \quad B \to A$	2	1	$\begin{cases} a+b-c_1, \\ 2k_1a^2-k_2b \end{cases}$
$2A \to A, \quad B \to A+B$	2	1	$\begin{cases} -k_1a^2+k_2b, \\ 0 \end{cases}$
$B \to A, \quad 2A \to A+B$	2	1	$\begin{cases} a+b-c_1, \\ k_2a^2-k_1b \end{cases}$
$B \to A, \quad 2B \to A+B$	1	0	$\begin{cases} a+b-c_1, \\ -k_2b^2-k_1b \end{cases}$
$2A \rightleftharpoons 2B$	2	1	$\begin{cases} a+b-c_1, \\ 2k_1a^2-2k_2b^2 \end{cases}$
$2A \to A+B \leftarrow 2B$	2	1	$\begin{cases} a+b-c_1, \\ k_1a^2-k_2b^2 \end{cases}$
$B \to A, \quad 2B \to 2A$	1	0	$\begin{cases} a+b-c_1, \\ -k_1b-2k_2b^2 \end{cases}$
$B \to 2B, \quad A \to A+B$	1	0	$\begin{cases} 0, \\ k_1b+k_2a \end{cases}$

network	mixed vol.	max #	system
$2B \to 0, \quad A \to A+B$	2	1	$\begin{cases} 0, \\ -2k_1b^2+k_2a \end{cases}$
$A \rightleftharpoons 2B$	2	1	$\begin{cases} a+b-c_1, \\ -k_2b^2+k_1a \end{cases}$
$A+B \to 2B \leftarrow 2A$	1	0	$\begin{cases} a+b-c_1, \\ 2k_2a^2+k_1ab \end{cases}$
$2A \to A+B \to 2B$	1	0	$\begin{cases} a+b-c_1, \\ k_1a^2+k_2ab \end{cases}$
$2A \to A, \quad A+B \to B$	1	0	$\begin{cases} -k_2a^2-k_1ab, \\ 0 \end{cases}$
$A+B \rightleftharpoons 0$	2	1	$\begin{cases} -k_1ab+k_2, \\ a-b \end{cases}$
$B \to A, \quad A+B \to 2A$	1	0	$\begin{cases} a+b-c_1, \\ -k_1b-k_2ab \end{cases}$
$0 \to 2B, \quad A+B \to A$	1	1	$\begin{cases} 0, \\ -k_2ab+2k_1 \end{cases}$
$2B \to 0, \quad A+B \to A$	1	1	$\begin{cases} 0, \\ -2k_1b^2-k_2ab \end{cases}$
$A+B \to 2A \to 2B$	1	1	$\begin{cases} a+b-c_1, \\ 2k_2a^2-k_1ab \end{cases}$
$A+B \to 2B \to A+B$	1	1	$\begin{cases} a+b-c_1, \\ k_1ab-k_2b^2 \end{cases}$
$A+B \to 2B, \quad B \to A$	1	1	$\begin{cases} a+b-c_1, \\ k_1ab-k_2b \end{cases}$
$A \to 0, \quad B \to A+B$	1	1	$\begin{cases} -k_1a+k_2b, \\ 0 \end{cases}$
$A \rightleftharpoons B$	1	1	$\begin{cases} a+b-c_1, \\ k_1a-k_2b \end{cases}$
$A+B \to A, \quad 0 \to B$	1	1	$\begin{cases} 0, \\ -k_1ab+k_2 \end{cases}$
$A+B \rightleftharpoons A$	1	1	$\begin{cases} 0, \\ -k_1ab+k_2a \end{cases}$

Acknowledgements

This research was initiated by Sofia in the 2019 REU in the Department of Mathematics at Texas A&M University, supported by the NSF (DMS-1757872). Obatake and Shiu were partially supported by the NSF (DMS-1752672). Obatake holds an American Fellowship from AAUW. We thank Taylor Brysiewicz for helpful discussions, and we are grateful for the comments of two diligent referees.

References

[Banaji 2017] M. Banaji, "Counting chemical reaction networks with NAUTY", preprint, 2017. arXiv

[Banaji and Pantea 2016] M. Banaji and C. Pantea, "Some results on injectivity and multistationarity in chemical reaction networks", *SIAM J. Appl. Dyn. Syst.* **15**:2 (2016), 807–869. MR Zbl

[Bernstein 1975] D. N. Bernstein, "The number of roots of a system of equations", *Funkcional. Anal. i Priložen.* **9**:3 (1975), 1–4. In Russian; translated in *Funct. Anal. Appl.***9**:3 (1975), 183–185. MR Zbl

[Bihan and Soprunov 2019] F. Bihan and I. Soprunov, "Criteria for strict monotonicity of the mixed volume of convex polytopes", *Adv. Geom.* **19**:4 (2019), 527–540. MR Zbl

[Conradi et al. 2017] C. Conradi, E. Feliu, M. Mincheva, and C. Wiuf, "Identifying parameter regions for multistationarity", *PLoS Comput. Biol.* **13**:10 (2017), art. id. 1005751.

[Cox et al. 2007] D. Cox, J. Little, and D. O'Shea, *Ideals, varieties, and algorithms*: *an introduction to computational algebraic geometry and commutative algebra*, 3rd ed., Springer, 2007. MR Zbl

[Craciun et al. 2006] G. Craciun, Y. Tang, and M. Feinberg, "Understanding bistability in complex enzyme-driven reaction networks", *Proc. Natl. Acad. Sci. USA* **103**:23 (2006), 8697–8702. Zbl

[Dickenstein et al. 2019] A. Dickenstein, M. Pérez Millán, A. Shiu, and X. Tang, "Multistationarity in structured reaction networks", *Bull. Math. Biol.* **81**:5 (2019), 1527–1581. MR Zbl

[Ewald 1996] G. Ewald, *Combinatorial convexity and algebraic geometry*, Graduate Texts in Mathematics **168**, Springer, 1996. MR Zbl

[Gross and Hill 2019] E. Gross and C. Hill, "The steady-state degree and mixed volume of a chemical reaction network", preprint, 2019. arXiv

[Gross et al. 2013] E. Gross, S. Petrović, and J. Verschelde, "Interfacing with PHCpack", *J. Softw. Algebra Geom.* **5** (2013), 20–25. MR Zbl

[Gross et al. 2016] E. Gross, H. A. Harrington, Z. Rosen, and B. Sturmfels, "Algebraic systems biology: a case study for the Wnt pathway", *Bull. Math. Biol.* **78**:1 (2016), 21–51. MR Zbl

[Joshi and Shiu 2015] B. Joshi and A. Shiu, "A survey of methods for deciding whether a reaction network is multistationary", *Math. Model. Nat. Phenom.* **10**:5 (2015), 47–67. MR Zbl

[Joshi and Shiu 2017] B. Joshi and A. Shiu, "Which small reaction networks are multistationary?", *SIAM J. Appl. Dyn. Syst.* **16**:2 (2017), 802–833. MR Zbl

[Obatake et al. 2019] N. Obatake, A. Shiu, X. Tang, and A. Torres, "Oscillations and bistability in a model of ERK regulation", *J. Math. Biol.* **79**:4 (2019), 1515–1549. MR Zbl

[Rojas 1994] J. M. Rojas, "A convex geometric approach to counting the roots of a polynomial system", *Theoret. Comput. Sci.* **133**:1 (1994), 105–140. MR Zbl

[Rubinstein et al. 2016] B. Y. Rubinstein, H. H. Mattingly, A. M. Berezhkovskii, and S. Y. Shvartsman, "Long-term dynamics of multisite phosphorylation", *Mol. Biol. Cell.* **27**:4 (2016), 2331–2340.

[Torres and Feliu 2019] A. Torres and E. Feliu, “Symbolic proof of bistability in reaction networks”, preprint, 2019. arXiv

[Ziegler 1995] G. M. Ziegler, *Lectures on polytopes*, Graduate Texts in Mathematics **152**, Springer, 1995. MR Zbl

Received: 2020-05-01 Revised: 2020-09-10 Accepted: 2020-09-20

nida@math.tamu.edu *Department of Mathematics, Texas A&M University,*
College Station, TX, United States

annejls@math.tamu.edu *Department of Mathematics, Texas A&M University,*
College Station, TX, United States

sofia@math.umass.edu *Department of Mathematics and Statistics,*
University of Massachusetts, Amherst, MA, United States

INVOLVE 13:5 (2020)
https://doi.org/10.2140/involve.2020.13.861

Counting profile strings from rectangular tilings

Anthony Petrosino, Alissa Schembor and Kathryn Haymaker

(Communicated by Arthur T. Benjamin)

We study the enumeration of profiles, which are outlines that occur when tiling a rectangular board with squares, dominoes, and trominoes. Profiles of length m correspond to a special subset of the set $\{0, 1, 2, 3\}^m$, called profile strings. Profiles and their corresponding strings first appeared in the enumeration of the tilings of rectangular $2\times n$ and $3\times n$ boards with squares, dominoes, and trominoes. Profiles also play a role in enumerating the tilings of an $m\times n$ board for any fixed $m \geq 2$. We describe how profiles arise when enumerating tilings, and we prove that the number of profile strings of length m equals $m \cdot 3^{m-1}$.

1. Background

A *tiling* of an $m\times n$ board using square tiles (1×1), domino tiles (1×2 or 2×1), and I-tromino tiles (1×3 or 3×1) is a configuration of tiles that covers every position in the board. In this paper we will use the word "tromino" to mean exclusively an I-tromino. Figure 1 shows a tiling of a 3×4 board using two squares, two dominoes, and two trominoes. Throughout this paper, the variable x denotes the placement of a square tile, y denotes the placement of a domino tile, and z denotes the placement of a tromino tile.

Read [1982] and Haymaker and Robertson [2017] determined recursions for the number of tilings of a $2\times n$ board with squares and dominoes, and squares, dominoes, and trominoes, respectively, using the technique of building a tiling from the leftmost, topmost untiled position. The outline of the right-hand side of a partially tiled board is called a *profile*, and counting the number of profiles of length m is the main goal of this paper.

We first define a building process that begins on the leftmost, topmost open square.

Definition 1. Consider a blank or partially tiled $m\times n$ board. A *building move* is when a tile is placed in the leftmost column that has an uncovered square, in the topmost square that is uncovered.

MSC2020: 05A19, 05B45.
Keywords: rectangular tilings, enumeration, induction.

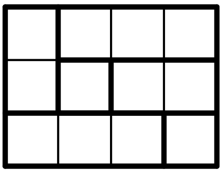


Figure 1. An example tiling of a 3×4 board with squares, dominoes, and trominoes.

There is a one-to-one correspondence between tilings of an $m \times n$ rectangular board and sequences of building moves that result in a complete $m \times n$ tiling [Read 1982]. That is, every tiling can be obtained from a unique sequence of building moves.

Definition 2. The *target square* of a partial tiling being built using building moves is the topmost square in the leftmost column with an uncovered square.

For an example of a tiling obtained using building moves, see Figure 2. In each partial tiling in Figure 2, the "×" denotes the position of the target square.

Definition 3. The m-row *profile* of a partial tiling of an $m \times n$ board is the outline of the rightmost edge of the partial tiling created by building moves. A *profile string* of an m-row profile is a string of length m where positions are assigned values of 0, 1, 2, or 3 depending on how many squares to the right of the rightmost fully tiled column are covered with tiles. We define the *length* of the profile to be m, the number of rows in the profile.

For example, the profile of the second partial tiling shown in Figure 2 has corresponding profile string 110.

Definition 4. The length-m *A profile* is the profile whose outline is a vertical line. The *A profile string* is $\underbrace{0 \cdots 0}_{m}$, the all-zero string of length m.

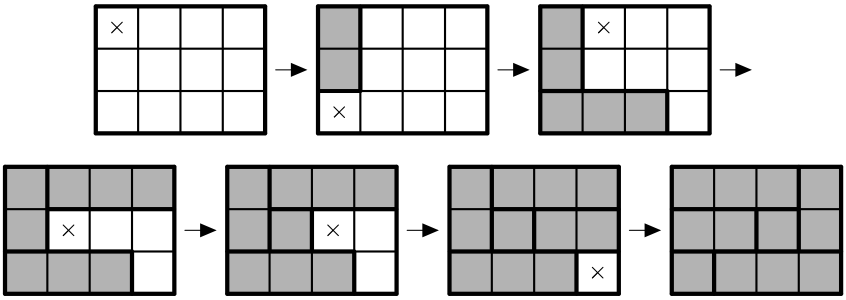


Figure 2. Sequence of profile strings from the building process for this 3×4 tiling: 000 → 110 → 002 → 302 → 201 → 110 → 000.

Next, we give an example of how building moves result in a sequence of profiles and profile strings.

Example 5 (illustrating profile strings with the 3×4 tiling). Figure 2 begins with an untiled 3×4 board, placing the target square in the leftmost-topmost position, indicated by the “×”. This profile is represented by the profile string 000 (the A profile). This is followed by the placement of a vertical domino, causing the target square to move down to the bottom row, and resulting in a profile string of 110. Next, a horizontal tromino is placed on top of the target square position, causing the target square to now move to the top row. Although this profile string appears to be 113, it should be noted that strings are denoted by the number of squares beyond the rightmost fully tiled column. Because there is one fully tiled column, the profile string reduces to 002. The process of constructing a tiling arrangement proceeds with the placement of squares, dominoes, or trominoes on top of the target square. Once the construction is complete, it will result in a 000 profile string, which is exactly how the example began.

Remark 6. The profile of a partially tiled board is the *shape* of the right-hand side of the board; the same shape/profile can result from different tile configurations. In fact, for any profile, it is possible to construct the profile shape using only “horizontal” tiles. To see this, consider a partial tiling that contains a vertical domino, for example. A specific example is the partial tiling on the far left of the middle row in Figure 2. This partial tiling can be changed into a different partial tiling with only horizontal tiles by replacing the vertical domino with two squares stacked vertically. The outline of the right side of the board remains the same; i.e., the profile remains the same. Similarly, if a partial tiling contains a vertical tromino, that piece can be replaced by three vertically stacked squares, yielding the same profile. This property of profiles becomes important in Section 3 when we count profiles using a construction method with only horizontal tiles.

2. Profile-type strings

To describe the shape of profiles in a numerical sense, we define profile-type strings. We use profile-type strings in order to further analyze the behavior of profiles resulting from the tiling of an $m\times n$ board. In Section 3, we will show that profile-type strings give a combinatorial description of all possible profiles that could result from the tiling of an $m\times n$ board using square tiles, domino tiles, and I-tromino tiles.

Definition 7. A string $x_1, x_2, \dots, x_m$ in $\{0, 1, 2, 3\}^m$ is called a *profile-type* string if it satisfies the following conditions:

(1) There exists an $i \in \{1, 2, \dots, m\}$ such that $x_i = 0$.

(2) If $x_i = 0$, then $x_j \neq 3$ for all $j > i$.

In words, a profile-type string is a string with at least one 0, where no 3 occurs to the right of a 0 in the string.

For example, the following are profile-type strings with $m = 5$: 31202, 11302, 00122. On the other hand, the following strings of length 5 are not profile-type strings: 12321, 00130, 00123.

The $m = 2$ case contains six profile-type strings

$$\begin{gathered} 00,\ 20,\ 10, \\ 30,\ 01,\ 02, \end{gathered} \tag{2-1}$$

while the $m = 3$ case contains 27 profile-type strings

$$\begin{gathered} 000,\ 300,\ 200,\ 110,\ 100,\ 330,\ 320,\ 310,\ 230, \\ 220,\ 210,\ 002,\ 001,\ 120,\ 130,\ 101,\ 202,\ 201, \\ 011,\ 010,\ 102,\ 302,\ 301,\ 012,\ 022,\ 021,\ 020. \end{gathered} \tag{2-2}$$

Proposition 8. *There are* 108 *profile-type strings for the $m = 4$ case.*

Proof. We enumerate the profile-type strings in $\{0, 1, 2, 3\}^4$ as follows. Let a, b, c, d represent a profile-type string. Then we count the following cases:

(1) If $a = 0$, then $b, c, d \in \{0, 1, 2\}$ by Definition 7, so there are $3^3 = 27$ such strings.

(2) If $a \in \{1, 2, 3\}$ and $b = 0$, then $c, d \in \{0, 1, 2\}$, so there are 27 such strings.

(3) If $a, b \in \{1, 2, 3\}$ and $c = 0$, then $d \in \{0, 1, 2\}$, so there are 27 such strings.

(4) If $a, b, c \in \{1, 2, 3\}$ and $d = 0$, there are 27 strings of this type, given the options for a, b, and c.

The total number of such strings is $4(27) = 108$. □

Theorem 9. *The number of profile-type strings of length m is equal to $m \cdot 3^{m-1}$.*

Proof. We will go through each position of a length-m profile-type string, anchor a 0 onto each position, and see how many different values are possible for all of the positions other than the fixed position. If there is a 0 anchored on a fixed position, we will say that there cannot be any 0 on any position before the fixed position, in order to avoid repeats in the enumeration.

Let x_i represent the i-th entry in the profile-type string.

The cases begin as follows:

(1) Use x_1 as the anchor, with $x_1 = 0$; then $x_2, \dots, x_m \in \{0, 1, 2\}$ as there cannot be a 3 following a 0 in a profile-type string, by definition.

(2) Next use x_2 as the anchor position, so $x_2 = 0$; then $x_1 \in \{1, 2, 3\}$ as there are no restrictions on the first position, besides not having 0 to avoid double

counting. Further, $x_3, \dots, x_m \in \{0, 1, 2\}$ as there cannot be a 3 following a 0 in a profile-type string, by definition.

In general, for $k = 1, \dots, m$, the k-th case is to anchor $x_k = 0$, and assume $x_i \in \{1, 2, 3\}$ for each i satisfying $1 \leq i < k$. Therefore by the definition of profile-type strings, $x_j \in \{0, 1, 2\}$ for each j satisfying $k < j \leq m$.

Each of these m cases has 3^{m-1} strings: there are three options for every position, other than in the fixed position in which we anchored a 0; as there are $m-1$ positions with three options (every position but the fixed position in a length-m profile-type string), there are 3^{m-1} options. There are m cases, so the total number of profile-type strings is $m \cdot 3^{m-1}$. □

3. Counting profiles

In this section, we prove that the profile-type strings of length m are in one-to-one correspondence with the profile strings of length m. That is, we can drop the word "type" in the name profile-type strings!

As a motivating example, we first look at the case of $m = 2$.

Example 10 (profiles and profile-type strings for $m = 2$). Notice that the length-2 strings that satisfy the conditions in Definition 7 are given in (2-1).

On the other hand, the profiles with two rows are in Figure 3, and their corresponding profile strings are

$$A : 00, \quad B : 20, \quad C : 10, \quad D : 30, \quad E : 01, \quad F : 02.$$

By inspection, we see that the profile strings for profiles A–F in Figure 3 are precisely the profile-type strings listed in (2-1).

Definition 11. A *profile construction* is the process of sequentially adding squares, dominoes, and trominoes to the A-profile until the desired profile results. The order in which squares, dominoes, and trominoes are added must correspond with proper

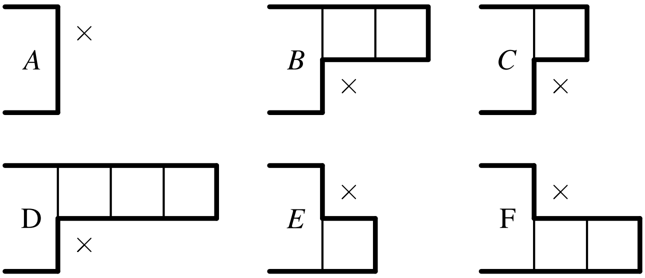


Figure 3. This figure shows all six profiles for $m = 2$. We see that there is a one-to-one correspondence between these profiles and the profile-type strings for $m = 2$.

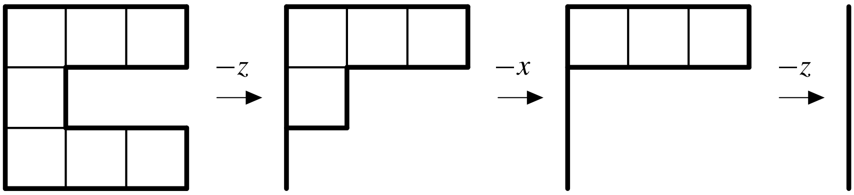


Figure 4. Deconstruction of the length-3 profile 202 to the A-profile in the -1 column position.

building moves; that is, at each step a square, domino, or tromino must be added onto the target square.

Definition 12. A *profile deconstruction* is the process of sequentially removing squares, dominoes, and trominoes from a profile until all that is left is the A-profile. The order in which squares, dominoes, and trominoes are removed must correspond with the construction of a profile; that is, squares, dominoes, and trominoes must be removed based on where the previous target square was located.

Remark 13. Profile deconstructions are not unique; there are different ways to deconstruct the same profile.

Example 14 (deconstruction and reconstruction with an additional row). We will shortly see an induction proof that involves adding a row to a profile. In this example we show a deconstruction of the profile with string 202 to the A-profile in column -1. Then we add a row and construct the profile with string 2020, as a motivating example for the inductive step.

In Figure 4, the initial profile will be deconstructed to the -1 column position. In order to do so, the deconstruction first removes a tromino from the bottom row, followed by the removal of a square from the middle row, and finally the removal of a tromino from the top row. This allows for the figure to be fully deconstructed to the A-profile in column -1.

In Figure 5, an additional row is then added to the profile. This begins the process of reconstructing the profile 2020 by adding the following (in this particular order, from top to bottom): tromino, square, tromino, square.

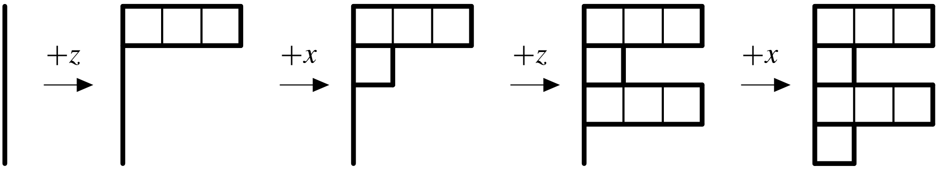


Figure 5. Construction of the $m = 4$ profile 2020 by undoing the deconstruction in Figure 4, with an additional row.

Lemma 15. *Every profile of length m can be deconstructed to the A-profile of length m in column -1.*

Proof. The following algorithm deconstructs a profile of length m to the A-profile in the -1 column. Let $\mathcal{P}_m \subseteq \{0, 1, 2, 3\}^m$ be the set of profile strings of length m.

Input: $x_1, x_2, \dots, x_m \in \mathcal{P}_m$. The algorithm runs through $k = m, m-1, \dots, 1$. If $x_k \in \{0, 1, 2\}$, remove a length-$(x_k + 1)$-tile from the k-th row and proceed to row $k-1$. If $x_k = 3$, remove a combined square and tromino from the k-th row and proceed to row $k-1$. The algorithm terminates when $k = 1$. The resulting profile is an A-profile at the -1 column. □

To reconstruct the profile from A at the -1 column, place the removed pieces from top to bottom; in the rows consisting of a square followed by a tromino, place the squares first. Then in the 0 column, place the trominoes from top to bottom. These steps comprise two phases of reconstruction.

Two-phase reconstruction, starting with the A-profile in the -1 column:

- Phase 1: If $x_k \in \{0, 1, 2\}$, then place a length-(x_k+1) tile in row k. If $x_k = 3$, phase 1 is to place a square tile in row k.
- Phase 2: Place trominoes in all rows where $x_k = 3$.

Notice that this two-phase process is designed to follow the rules of building moves and target squares, while resulting in the desired profile.

Example 16 (two-phase reconstruction of 3120).

- Phase 1: In the first portion of Figure 6, we see the result of phase 1 of construction to obtain the profile with string 3120. Note that for the desired profile, $x_1 = 3$, $x_2 = 1$, $x_3 = 2$, and $x_4 = 0$. Begin phase 1 by placing a square (length-1 tile) in row 1, as $x_1 = 3$. Next, place a domino (length-2 tile) in row 2, as $x_2 = 1$. Next, place a tromino (length-3 tile) in row 3, as $x_3 = 2$. Next, place a square (length-1 tile) in row 4, as $x_4 = 0$.
- Phase 2: Place a tromino (length-3 tile) in row 1, as $x_1 = 3$. Phase 2 is also shown in Figure 6.

Next we give a lemma that contributes to establishing the one-to-one correspondence between profiles and profile-type strings.

Lemma 17. *If $x_1, \dots, x_m, x_{m+1}$ is a profile-type string of length $m + 1$ where at least one of $x_1, \dots, x_m$ equals 0, then the string $\sigma = x_1, x_2, \dots, x_m$ is a profile-type string of length m.*

Proof. Assume that $x_1, \dots, x_m, x_{m+1}$ is a profile-type string, and at least one of $x_1, \dots, x_m$ equals 0. Let σ denote the string $x_1, x_2, \dots, x_m$. Condition (1) in Definition 7 holds for the string σ, since it contains at least one zero.

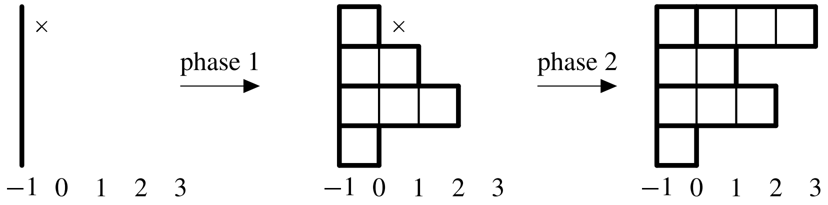


Figure 6. Two phase construction of the profile with string 3120, starting at the -1 column position.

To show that condition (2) in Definition 7 holds, we proceed by contradiction. Suppose that the string σ is not a profile-type string of length m because it violates condition (2). If σ violates condition (2), then there exist i, j such that $1 \le i < j \le m$, and $x_i = 0$ and $x_j = 3$ in σ. The entries x_i and x_j are identical in the original string $x_1, x_2, \ldots, x_m, x_{m+1}$, which implies that it also violates condition (2). This is a contradiction as we assumed that the original string was a profile-type string. □

The main result of this paper is next.

Theorem 18. *The m-row profiles are in one-to-one correspondence with the profile-type strings of length m.*

Proof. The theorem contains two claims:

(1) An m-row profile string is a profile-type string of length m (i.e., a string that satisfies Definition 7).

(2) A profile-type string of length m corresponds to an m-row profile.

To prove the first claim, let $x_1, \ldots, x_m$ be a profile string of length m corresponding to a profile P. By the definition of profile strings, at least one entry x_j must be 0, so the string satisfies condition (1) in Definition 7. Suppose that $x_j = 0$ and $x_k = 3$ for some $k > j$. Then in the construction of the profile P, a square, domino, or tromino must have been placed in the k-th row, while the j-th row contained an empty square in a position farther up and to the left. This contradicts the building move process, so no such j and k exist.

To prove the second claim, that every profile-type string corresponds to a profile string, we consider two cases and use induction.

We use induction on m. The base case is $m = 2$. See Example 10 for details of the one-to-one correspondence in the base case.

The inductive hypothesis is that there is a one-to-one correspondence between m-row profiles and profile-type strings of length m.

We will show the correspondence holds for $m + 1$.

Case 1: Suppose that the profile-type string $S = x_1, x_2, \ldots, x_{m+1}$ has $x_{m+1} = 0$, and $x_j \neq 0$ for all $j = 1, \ldots, m$. Then we construct a profile with the profile string S as follows:

Algorithm for constructing a length-$(m+1)$ profile, Case 1:

(1) Start with the A-profile of length $m+1$.

(2) From top to bottom, place tiles of length x_i horizontally in row i for $i = 1, 2, \ldots, m$.

(3) Leave row $m+1$ unchanged.

The resulting profile has corresponding string $x_1, x_2, \ldots, x_m, x_{m+1}$, where x_{m+1} is the only entry that equals 0. (Notice that this case does not involve the inductive hypothesis.)

Case 2: For this case, consider the profile-type string $S = x_1, x_2, \ldots, x_m, x_{m+1}$, where at least one of $x_1, \ldots, x_m$ equals 0.

Thus $x_1, \ldots, x_m$ is a profile-type string of length m by Lemma 17, so by the inductive hypothesis there is a profile P of length m that corresponds to it.

Algorithm for constructing a length-$(m+1)$ profile, Case 2:

(1) Deconstruct the profile P to the A-profile.

(2) Add a row to the A-profile. We are now working with an $(m+1) \times n$ array.

(3) Declare the column that the A-profile is on as column -1. The outline of the resulting profile will be at most four columns to the right of this column.

(4) Implement phase 1 of the profile reconstruction process for $x_1, \ldots, x_m$ in the first m rows. As $x_{m+1} \in \{0, 1, 2\}$, place a length-$(x_{m+1}+1)$ tile in row $m+1$.

(5) Implement phase 2 of the profile reconstruction process for $x_1, \ldots, x_m$ in the first m rows. (Since $x_{m+1} \neq 3$, there is no phase 2 for row $m+1$).

This algorithm shows that there is a profile P with profile string S, proving our claim that every profile-type string S corresponds to a profile. □

Corollary 19. *The number of m-row profiles is $m \cdot 3^{m-1}$.*

Proof. By Theorem 18, the profiles are in one-to-one correspondence with profile-like strings, and by Theorem 9 there are $m \cdot 3^{m-1}$ profile-type strings of length m. □

Remark 20. Notice that the case where the only zero in the string S is $x_{m+1} = 0$ does not rely on the deconstruction and reconstruction of $x_1, \ldots, x_m$. This is because constructing a profile where only the final entry is 0 requires only one phase, working from top to bottom and placing exactly the length-x_i tile in row i.

4. Conclusions and open questions

In this paper we proved the one-to-one correspondence of profile-type strings of length m with profile strings of length m, partly using induction. This understanding allowed for the creation of a basic formula for the number of profiles in terms of m for any given $m \times n$ board.

An open question that stems from this work is: how can we generalize the profile-type string definition for tilings with more types of tiles, for example, L-trominoes or quadrominoes?

Acknowledgments

The authors would like to thank Richard Filingeri for help with programming for counting profiles for small m.

References

[Haymaker and Robertson 2017] K. Haymaker and S. Robertson, "Counting colorful tilings of rectangular arrays", *J. Integer Seq.* **20**:5 (2017), art. id. 17.5.8. MR Zbl

[Read 1982] R. C. Read, "The dimer problem for narrow rectangular arrays: a unified method of solution, and some extensions", *Aequationes Math.* **24**:1 (1982), 47–65. MR Zbl

Received: 2020-06-04 Revised: 2020-08-16 Accepted: 2020-08-17

apetros1@villanova.edu *Villanova University, Villanova, PA, United States*

aschembo@villanova.edu *Villanova University, Villanova, PA, United States*

kathryn.haymaker@villanova.edu *Villanova University, Villanova, PA, United States*

INVOLVE 13:5 (2020)
https://doi.org/10.2140/involve.2020.13.871

Isomorphisms of graded skew Clifford algebras

Richard G. Chandler and Nicholas Engel

(Communicated by Vadim Ponomarenko)

A regular algebra of global dimension $n+1$ is often called a quantum $\mathbb{P}^n$. In 2011, Nafari, Vancliff and Zhang showed that graded skew Clifford algebras (GSCAs) could be used to classify most quadratic quantum $\mathbb{P}^2$s. Some time later, Chandler, Tomlin and Vancliff used their work with certain families of GSCAs to develop a conjecture on the quantum space of a generic quadratic quantum $\mathbb{P}^3$. These results suggest that (the isomorphism classes of) GSCAs are likely to play a fundamental role in the classification of quadratic quantum $\mathbb{P}^3$s. In this article, we will discuss some of these results on GSCAs and discuss new results on isomorphisms between GSCAs.

Introduction

The notion of a graded skew Clifford algebra was defined in [Cassidy and Vancliff 2010] as a generalization of graded Clifford algebras (see [Le Bruyn 1995]). Therein, it was shown that GSCAs had many properties analogous to those of graded Clifford algebras. In addition, several families of quadratic regular graded skew Clifford algebras of global dimension 4 that were not twists of graded Clifford algebras were produced. It was shown that many of the algebras in these families had a one-dimensional line scheme [Shelton and Vancliff 2002a; 2002b] and a point scheme [Artin et al. 1990] consisting of twenty distinct points; hence, these algebras became candidates for generic quantum $\mathbb{P}^3$s (in the language of [Chandler and Vancliff 2015]).

The line scheme and point scheme of some of these families were explicitly computed in [Chandler and Vancliff 2015; Tomlin and Vancliff 2018]. The line scheme of the generic member of the family studied in [Chandler and Vancliff 2015] was found to consist of a spatial elliptic curve, four planar elliptic curves and two nonsingular conics; the line scheme of the generic member of the family studied in [Tomlin and Vancliff 2018] was found to consist of four planar elliptic curves and four nonsingular conics. In both cases, the computations suggested that

MSC2020: 16S37, 16S38.
Keywords: graded skew Clifford algebras, μ-symmetry, isomorphisms.

the nonsingular conics were coming from a spatial elliptic curve in which one of the defining polynomials factored. This suggests the following conjecture.

Conjecture 0.1 [Chandler and Vancliff 2015]. There exists a class of generic quadratic quantum $\mathbb{P}^3$s in which any representative member has a line scheme that is isomorphic to the union of two spatial elliptic curves and four planar elliptic curves.

At present, no regular algebra with such a line scheme is known. However, given the results in [Chandler and Vancliff 2015; Tomlin and Vancliff 2018], it is possible that some GSCA may satisfy this property and thus be candidates for generic quantum $\mathbb{P}^3$s.

In [Nafari et al. 2011] it was shown that most quadratic quantum $\mathbb{P}^2$s may be constructed as a twist by an automorphism (in the sense of [Artin et al. 1991]) of either a graded skew Clifford algebra or of an Ore extension (see [Goodearl and Warfield 2004]) of a regular graded skew Clifford algebra of global dimension 2. The only regular algebras of global dimension 3 that were not classified in this way are some that have a point scheme of an elliptic curve (classified as type E and an open subset of type A in [Artin and Schelter 1987]). It is currently unclear as to whether an analogous result holds for quantum $\mathbb{P}^3$s.

These results suggest that GSCAs are likely to play some role in the classification of quantum $\mathbb{P}^3$s (and hence the isomorphism classes of such algebras will be of interest).

The article is outlined as follows. In Section 1, we will introduce preliminary notions that will be relevant to the discussions in the article. In Section 2, we will give our main results. In particular, we will discuss two types of isomorphisms: one in which the degree-1 generators of the algebras are identified and the degree-2 generators are related via matrix multiplication and another in which the degree-2 generators of the algebras are identified and the degree-1 generators are related via a diagonal mapping.

1. Preliminaries

Throughout this article, we adopt the following notation:

- $\Bbbk$ denotes an algebraically closed field with $\operatorname{char}(\Bbbk) \neq 2$.
- $\Bbbk^\times$ denotes the nonzero elements of $\Bbbk$.
- $\mathbb{M}_n(\Bbbk)$ denotes the ring of $n \times n$ matrices with entries in $\Bbbk$.
- M_{ij} denotes the ij-th entry of the matrix M and M^T denotes the transpose of M.
- If S is a positively graded, connected $\Bbbk$-algebra, then S_n denotes the homogeneous degree-n elements of S.
- If V is a $\Bbbk$-vector space, then V^n denotes the vector space $\underbrace{V \times V \times \cdots \times V}_{n \text{ times}}$.

Definition 1.1 [Cassidy and Vancliff 2010]. Let $\mu_{ij} \in \mathbb{k}^\times$, where $1 \le i, j \le n$, such that $\mu_{ij}\mu_{ji} = 1$ for all $i \ne j$ and $\mu_{ii} = 1$ for all i. We write $\mu = (\mu_{ij}) \in \mathbb{M}_n(\mathbb{k})$. A matrix $M \in \mathbb{M}_n(\mathbb{k})$ is called μ-symmetric if $M_{ij} = \mu_{ij} M_{ji}$ for all i, j.

The set of all $n \times n$ μ-symmetric matrices forms a $\mathbb{k}$-vector space, denoted by $\mathrm{Sym}_n^\mu(\mathbb{k})$. The standard basis of $\mathrm{Sym}_n^\mu(\mathbb{k})$ is

$$\mathfrak{B} = \{E_{ii} : 1 \le i \le n\} \cup \{E_{ij} + \mu_{ji} E_{ji} : 1 \le i < j \le n\},$$

where $E_{ij} \in \mathbb{M}_n(\mathbb{k})$ denotes the matrix with a 1 in the ij-th position and zeroes elsewhere.

In [Cassidy and Vancliff 2010], it was shown that the notion of identifying a quadratic form and a symmetric matrix could be generalized to μ-symmetric matrices as follows. Let $\mu \in \mathbb{M}_n(\mathbb{k})$ be as in Definition 1.1 and $M \in \mathrm{Sym}_n^\mu(\mathbb{k})$. Let S denote the $\mathbb{k}$-algebra on $z_1, z_2, \dots, z_n$ with defining relations $z_j z_i = \mu_{ij} z_i z_j$ for all $i, j = 1, \dots, n$ so that S is a skew polynomial ring. Then the image of $q = z^T M z$, where $z = [z_1\, z_2\, \cdots\, z_n]^T$, in S is a quadratic form. Furthermore, if $q = \sum_{i \le j} a_{ij} z_i z_j \in S_2$, one may associate to q an $M \in \mathrm{Sym}_n^\mu(\mathbb{k})$ defined by $M_{ii} = a_{ii}$ for all i and, for $i < j$, $M_{ij} = a_{ij}/2$.

Definition 1.2 [Cassidy and Vancliff 2010]. Let $\mu \in \mathbb{M}_n(\mathbb{k})$ be as in Definition 1.1 and $M_1, \dots, M_n \in \mathrm{Sym}_n^\mu(\mathbb{k})$. A graded skew Clifford algebra $A = A(\mu, M_1, \dots, M_n)$ associated to μ and $M_1, \dots, M_n$ is a graded $\mathbb{k}$-algebra on degree-1 generators $x_1, \dots, x_n$ and on degree-2 generators $y_1, \dots, y_n$ with defining relations given by

- $x_i x_j + \mu_{ij} x_j x_i = \sum_{k=1}^n (M_k)_{ij} y_k$ for all $i, j = 1, \dots, n$; and,
- any additional (degree-3 or degree-4) relations necessary in order to guarantee the existence of a normalizing sequence that spans $\bigoplus_{k=1}^n \mathbb{k} y_k$.

Note that if a graded skew Clifford algebra is regular, then it will only have relations of the form $x_i x_j + \mu_{ij} x_j x_i = \sum_{k=1}^n (M_k)_{ij} y_k$ for all $i, j = 1, \dots, n$.

2. Main results

We first examine the case where the degree-1 generators of each algebra are identified and the degree-2 generators are related via a mapping by matrix multiplication. Our algebras will be $\mathbb{k}$-algebras whose defining relations are of the form $x_i x_j + \mu_{ij} x_j x_i = \sum_{k=1}^n (M_k)_{ij} y_k$. Each such algebra will map onto a graded skew Clifford algebra since the defining relations of a graded skew Clifford algebra will include these relations and (possibly) additional relations necessary to guarantee the existence of a normalizing sequence, as required by Definition 1.2.

Theorem 2.1. *Let $\mu \in \mathbb{M}_n(\mathbb{k})$ be as in Definition 1.1 and let*

$$M_1, M_2, \dots, M_n, N_1, N_2, \dots, N_n \in \mathrm{Sym}_n^\mu(\mathbb{k}).$$

Let A be a $\Bbbk$-algebra on degree-1 generators $x_1, x_2, \dots, x_n$ and degree-2 generators $y_1, y_2, \dots, y_n$ with defining relations

$$x_i x_j + \mu_{ij} x_j x_i = \sum_{k=1}^{n} (M_k)_{ij} y_k \quad \textit{for all } i, j.$$

Similarly, let B be a $\Bbbk$-algebra on degree-1 generators $X_1, X_2, \dots, X_n$ and degree-2 generators $Y_1, Y_2, \dots, Y_n$ with defining relations

$$X_i X_j + \mu_{ij} X_j X_i = \sum_{k=1}^{n} (N_k)_{ij} Y_k \quad \textit{for all } i, j.$$

Given a nonsingular matrix $P \in \mathbb{M}_n(\Bbbk)$, A is isomorphic to B under a map $\varphi : A \to B$ defined by

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \overset{\varphi}{\longmapsto} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \quad \textit{and} \quad \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \overset{\varphi}{\longmapsto} P \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

if and only if $N = P^T M$, where $N = [N_1\, N_2\, \cdots\, N_n]^T$, $M = [M_1\, M_2\, \cdots\, M_n]^T \in (\mathrm{Sym}_n^{\mu}(\Bbbk))^n$.

Proof. Let $\mathfrak{B}$ denote the standard $\Bbbk$-basis of $\mathrm{Sym}_n^{\mu}(\Bbbk)$. We may write the defining relations of A as the matrix equation

$$\begin{bmatrix} 2x_1^2 \\ x_1x_2+\mu_{12}x_2x_1 \\ \vdots \\ 2x_n^2 \end{bmatrix} = \begin{bmatrix} (M_1)_{11} & (M_2)_{11} & \cdots & (M_n)_{11} \\ (M_1)_{12} & (M_2)_{12} & \cdots & (M_n)_{12} \\ \vdots & \vdots & \ddots & \vdots \\ (M_1)_{nn} & (M_2)_{nn} & \cdots & (M_n)_{nn} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

Applying the map φ to the above equation we obtain the defining relations of B:

$$\begin{bmatrix} 2X_1^2 \\ X_1X_2+\mu_{12}X_2X_1 \\ \vdots \\ 2X_n^2 \end{bmatrix} = \begin{bmatrix} (M_1)_{11} & (M_2)_{11} & \cdots & (M_n)_{11} \\ (M_1)_{12} & (M_2)_{12} & \cdots & (M_n)_{12} \\ \vdots & \vdots & \ddots & \vdots \\ (M_1)_{nn} & (M_2)_{nn} & \cdots & (M_n)_{nn} \end{bmatrix} P \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}.$$

Hence

$$\begin{bmatrix} (M_1)_{11} & (M_2)_{11} & \cdots & (M_n)_{11} \\ (M_1)_{12} & (M_2)_{12} & \cdots & (M_n)_{12} \\ \vdots & \vdots & \ddots & \vdots \\ (M_1)_{nn} & (M_2)_{nn} & \cdots & (M_n)_{nn} \end{bmatrix} P = \begin{bmatrix} (N_1)_{11} & (N_2)_{11} & \cdots & (N_n)_{11} \\ (N_1)_{12} & (N_2)_{12} & \cdots & (N_n)_{12} \\ \vdots & \vdots & \ddots & \vdots \\ (N_1)_{nn} & (N_2)_{nn} & \cdots & (N_n)_{nn} \end{bmatrix}.$$

The k-th columns of the matrices in the previous equation are the $\mathfrak{B}$-coordinate vector of M_k and N_k, respectively. Thus,

$$\begin{aligned}\begin{bmatrix}N_1 & \cdots & N_n\end{bmatrix} = \begin{bmatrix}M_1 & \cdots & M_n\end{bmatrix} P \quad &\Longrightarrow \quad \begin{bmatrix}N_1\\ \vdots\\ N_n\end{bmatrix} = P^T \begin{bmatrix}M_1\\ \vdots\\ M_n\end{bmatrix}\\ &\Longrightarrow \quad N = P^T M. \qquad \square\end{aligned}$$

Example 2.2. Let

$$\mu = \begin{bmatrix}1 & i\\ -i & 1\end{bmatrix}, \quad M_1 = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}, \quad M_2 = \begin{bmatrix}0 & i\\ 1 & 0\end{bmatrix}, \quad N_1 = \begin{bmatrix}2 & -i\\ -1 & 2\end{bmatrix}, \quad N_2 = \begin{bmatrix}i & i\\ 1 & i\end{bmatrix}.$$

Then $A = A(\mu, M_1, M_2)$ is given by

$$A = \frac{\Bbbk\langle x_1, x_2, y_1, y_2\rangle}{\langle 2x_1^2 - y_1, 2x_2^2 - y_1, x_1x_2 + ix_2x_1 - iy_2\rangle}$$

and $B = B(\mu, N_1, N_2)$ is given by

$$B = \frac{\Bbbk\langle X_1, X_2, Y_1, Y_2\rangle}{\langle 2X_1^2 - 2Y_1 - iY_2, 2X_2^2 - 2Y_1 - iY_2, X_1X_2 + iX_2X_1 + iY_1 - iY_2\rangle}.$$

One can compute that

$$\begin{cases}N_1 = 2M_1 - M_2,\\ N_2 = iM_1 + M_2\end{cases} \quad \Longrightarrow \quad \begin{bmatrix}N_1\\ N_2\end{bmatrix} = \begin{bmatrix}2 & -1\\ i & 1\end{bmatrix} \begin{bmatrix}M_1\\ M_2\end{bmatrix}.$$

Hence,

$$P = \begin{bmatrix}2 & i\\ -1 & 1\end{bmatrix}.$$

The above theorem implies that $A \cong B$ under the map $x_i \mapsto X_i$ for all i, $y_1 \mapsto 2Y_1 + iY_2$ and $y_2 \mapsto -Y_1 + Y_2$, which is easily verified.

Corollary 2.3. *Let μ, A, B, M, N, φ and P be as in Theorem 2.1. If q_k, $p_k \in S$ denote the quadratic forms associated to M_k and N_k, respectively, then A is isomorphic to B under φ if and only if*

$$\begin{bmatrix}p_1\\ p_2\\ \vdots\\ p_n\end{bmatrix} = P^T \begin{bmatrix}q_1\\ q_2\\ \vdots\\ q_n\end{bmatrix}.$$

Proof. The result follows immediately from multiplying each entry of the vectors N and M in the equation $N = P^T M$ by z^T and z on the left and right, respectively. $\square$

Example 2.4. Continuing Example 2.2, the quadratic forms associated to M_1, M_2, N_1 and N_2 are $q_1 = z_1^2 + z_2^2$, $q_2 = 2iz_1z_2$, $p_1 = 2z_1^2 - 2iz_1z_2 + 2z_2^2$ and $p_2 = iz_1^2 + 2iz_1z_2 + iz_2^2$, respectively. It is easily verified that

$$\begin{cases} p_1 = 2q_1 - q_2, \\ p_2 = iq_1 + q_2 \end{cases} \quad \Longrightarrow \quad \begin{bmatrix} p_1 \\ p_2 \end{bmatrix} = \begin{bmatrix} 2 & -i \\ i & 1 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \end{bmatrix}.$$

We now examine the case where the degree-2 generators of each algebra are identified and the degree-1 generators of the algebra are related via a diagonal mapping.

Theorem 2.5. *Let $\mu \in \mathbb{M}_n(\Bbbk)$ be as in Definition 1.1 and let*

$$M_1, M_2, \dots, M_n, N_1, N_2, \dots, N_n \in \mathrm{Sym}_n^{\mu}(\Bbbk).$$

Let A be a $\Bbbk$-algebra on degree-1 generators $x_1, x_2, \dots, x_n$ and degree-2 generators $y_1, y_2, \dots, y_n$ with defining relations

$$x_ix_j + \mu_{ij}x_jx_i = \sum_{k=1}^{n} (M_k)_{ij}y_k \quad \textit{for all } i, j.$$

Similarly, let B be a $\Bbbk$-algebra on degree-1 generators $X_1, X_2, \dots, X_n$ and degree-2 generators $Y_1, Y_2, \dots, Y_n$ with defining relations

$$X_iX_j + \mu_{ij}X_jX_i = \sum_{k=1}^{n} (N_k)_{ij}Y_k \quad \textit{for all } i, j.$$

Then $\varphi : A \to B$ is an isomorphism defined $x_i \overset{\varphi}{\longmapsto} \alpha_i X_i$ and $y_i \overset{\varphi}{\longmapsto} Y_i$ for $\alpha_i \in \Bbbk^{\times}$ and for all i if and only if $(N_k)_{ij} = (M_k)_{ij}/(\alpha_i\alpha_j)$ for all i, j, k.

Proof. Applying the map φ to the defining relations of A yields the defining relations of B:

$$\alpha_i\alpha_jX_iX_j + \mu_{ij}\alpha_i\alpha_jX_jX_i = \sum_{k=1}^{n}(M_k)_{ij}Y_k \quad \Longrightarrow \quad X_iX_j + \mu_{ij}X_jX_i = \sum_{k=1}^{n} \frac{(M_k)_{ij}}{\alpha_i\alpha_j}Y_k.$$

Hence, $(N_k)_{ij} = (M_k)_{ij}/(\alpha_i\alpha_j)$. □

Example 2.6. Let

$$\mu = \begin{bmatrix} 1 & i \\ -i & 1 \end{bmatrix}, \quad M_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad M_2 = \begin{bmatrix} 0 & i \\ 1 & 0 \end{bmatrix}, \quad N_1 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad N_2 = \begin{bmatrix} 0 & -1 \\ i & 0 \end{bmatrix}.$$

Then $A = A(\mu, M_1, M_2)$ is given by

$$A = \frac{\Bbbk\langle x_1, x_2, y_1, y_2 \rangle}{\langle 2x_1^2 - y_1, 2x_2^2 - y_1, x_1x_2 + ix_2x_1 - iy_2 \rangle}$$

and $B = B(\mu, N_1, N_2)$ is given by

$$B = \frac{\Bbbk\langle X_1, X_2, Y_1, Y_2\rangle}{\langle 2X_1^2 - Y_1, 2X_2^2 + Y_1, X_1X_2 + iX_2X_1 + Y_2\rangle}.$$

In checking to see if $(N_k)_{ij} = (M_k)_{ij}/(\alpha_i\alpha_j)$, we arrive at the equations

$$\begin{cases} 1 = 1/\alpha_1^2, \\ -1 = 1/\alpha_2^2, \\ -1 = i/(\alpha_1\alpha_2), \\ i = 1/(\alpha_1\alpha_2) \end{cases} \implies \begin{cases} \alpha_1 = \pm 1, \\ \alpha_2 = \mp i. \end{cases}$$

The above theorem implies that $A \cong B$ under both the maps φ_+ and φ_- defined by $\varphi_\pm : x_1 \mapsto \pm X_1,\ x_2 \mapsto \mp i X_2$ and $y_k \mapsto Y_k$, for all k.

Corollary 2.7. *Let $\mu, A, B, M_1, M_2, \dots, M_n, N_1, N_2, \dots, N_n$, and φ be as in Theorem 2.5. If $q_k = \sum_{i\le j} a_{k_{ij}} z_i z_j \in S$ and $p_k = \sum_{i\le j} b_{k_{ij}} z_i z_j \in S$ denote the quadratic forms associated to M_k and N_k, respectively, then A is isomorphic to B under φ if and only if $b_{k_{ij}} = a_{k_{ij}}/(\alpha_i\alpha_j)$.*

Proof. The result follows immediately from $(N_k)_{ij} = (M_k)_{ij}/(\alpha_i\alpha_j)$ and the definition of the quadratic form in [Cassidy and Vancliff 2010]. □

Example 2.8. Continuing 2.6, the quadratic forms associated to M_1, M_2, N_1 and N_2 are $q_1 = z_1^2 + z_2^2$, $q_2 = 2iz_1z_2$, $p_1 = z_1^2 - z_2^2$ and $p_2 = -2z_1z_2$, respectively. It is easily verified that $b_{k_{ij}} = a_{k_{ij}}/(\alpha_i\alpha_j)$ holds for these forms.

The reader should note that the more general case in which the degree-1 generators of an algebra are related via a mapping by matrix multiplication is much more complicated. Such maps can transform graded skew Clifford algebras into isomorphic algebras whose defining relations cannot be directly written in the form $x_ix_j + \mu_{ij}x_jx_i = \sum_{k=1}^n (M_k)_{ij} y_k$ without changing the presentation of the algebra, as illustrated in the following example.

Example 2.9. Let

$$\mu = \begin{bmatrix} 1 & i \\ -i & 1 \end{bmatrix}, \quad M_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad M_2 = \begin{bmatrix} 0 & i \\ 1 & 0 \end{bmatrix}.$$

Then $A = A(\mu, M_1, M_2)$ is given by

$$A = \frac{\Bbbk\langle x_1, x_2, y_1, y_2\rangle}{\langle 2x_1^2 - y_1, 2x_2^2 - y_1, x_1x_2 + ix_2x_1 - iy_2\rangle}.$$

Consider the isomorphism $\varphi : A \to \varphi(A)$ defined by $x_1 \mapsto X_1 - X_2,\ x_2 \mapsto X_2$, and $y_k \mapsto Y_k$ for all k. Applying φ to the defining relations of A yields the relations

$$\begin{aligned} 2X_1^2 + 2X_2^2 - 2X_1X_2 - 2X_2X_1 - Y_1 &= 0, \\ 2X_2^2 - Y_1 &= 0, \\ X_1X_2 + iX_2X_1 - (1+i)X_2^2 - iY_2 &= 0. \end{aligned}$$

These relations cannot all be written in the form $X_iX_j + \mu_{ij}X_jX_i = \sum_{k=1}^n (N_k)_{ij}Y_k$ without a change of presentation; hence, the image of the algebra does not present as a graded skew Clifford algebra.

References

[Artin and Schelter 1987] M. Artin and W. F. Schelter, "Graded algebras of global dimension 3", *Adv. in Math.* **66**:2 (1987), 171–216. MR Zbl

[Artin et al. 1990] M. Artin, J. Tate, and M. Van den Bergh, "Some algebras associated to automorphisms of elliptic curves", pp. 33–85 in *The Grothendieck Festschrift, Vol. I*, edited by P. Cartier et al., Progr. Math. **86**, Birkhäuser, Boston, MA, 1990. MR Zbl

[Artin et al. 1991] M. Artin, J. Tate, and M. Van den Bergh, "Modules over regular algebras of dimension 3", *Invent. Math.* **106**:2 (1991), 335–388. MR Zbl

[Cassidy and Vancliff 2010] T. Cassidy and M. Vancliff, "Generalizations of graded Clifford algebras and of complete intersections", *J. Lond. Math. Soc.* (2) **81**:1 (2010), 91–112. Correction in **90**:2 (2014), 631–636. MR Zbl

[Chandler and Vancliff 2015] R. G. Chandler and M. Vancliff, "The one-dimensional line scheme of a certain family of quantum $\mathbb{P}^3$s", *J. Algebra* **439** (2015), 316–333. MR Zbl

[Goodearl and Warfield 2004] K. R. Goodearl and R. B. Warfield, Jr., *An introduction to noncommutative Noetherian rings*, 2nd ed., London Mathematical Society Student Texts **61**, Cambridge University Press, 2004. MR Zbl

[Le Bruyn 1995] L. Le Bruyn, "Central singularities of quantum spaces", *J. Algebra* **177**:1 (1995), 142–153. MR Zbl

[Nafari et al. 2011] M. Nafari, M. Vancliff, and J. Zhang, "Classifying quadratic quantum $\mathbb{P}^2$s by using graded skew Clifford algebras", *J. Algebra* **346** (2011), 152–164. MR Zbl

[Shelton and Vancliff 2002a] B. Shelton and M. Vancliff, "Schemes of line modules, I", *J. London Math. Soc.* (2) **65**:3 (2002), 575–590. MR Zbl

[Shelton and Vancliff 2002b] B. Shelton and M. Vancliff, "Schemes of line modules, II", *Comm. Algebra* **30**:5 (2002), 2535–2552. MR Zbl

[Tomlin and Vancliff 2018] D. Tomlin and M. Vancliff, "The one-dimensional line scheme of a family of quadratic quantum $\mathbb{P}^3$s", *J. Algebra* **502** (2018), 588–609. MR Zbl

Received: 2020-06-22 Revised: 2020-08-07 Accepted: 2020-08-10

richard.chandler@untdallas.edu *Department of Mathematics and Information Sciences,*
University of North Texas at Dallas, Dallas, TX, United States

nicholas.engel@mavs.uta.edu *Department of Mathematics, University of Texas at Arlington,*
Arlington, TX, United States

INVOLVE 13:5 (2020)
https://doi.org/10.2140/involve.2020.13.879

Eta-quotients of prime or semiprime level and elliptic curves

Michael Allen, Nicholas Anderson,
Asimina Hamakiotes, Ben Oltsik and Holly Swisher

(Communicated by Ken Ono)

From the modularity theorem proven by Wiles, Taylor, Conrad, Diamond, and Breuil, we know that all elliptic curves are modular. It has been shown by Martin and Ono exactly which are represented by eta-quotients, and some examples of elliptic curves represented by modular forms that are linear combinations of eta-quotients have been given by Pathakjee, RosnBrick, and Yoong.

In this paper, we first show that eta-quotients which are modular for any congruence subgroup of level N coprime to 6 can be viewed as modular for $\Gamma_0(N)$. We then categorize when even-weight eta-quotients can exist in $M_k(\Gamma_1(p))$ and $M_k(\Gamma_1(pq))$ for distinct primes p, q. We conclude by providing some new examples of elliptic curves whose corresponding modular forms can be written as a linear combination of eta-quotients, and describe an algorithmic method for finding additional examples.

1. Introduction and statement of results

Dedekind's eta-function $\eta(\tau)$, defined for $\tau \in \mathbb{H} := \{\tau \in \mathbb{C} : \mathrm{Im}(\tau) > 0\}$ by

$$\eta(\tau) := q^{1/24} \prod_{n=1}^{\infty} (1 - q^n),$$

where $q := e^{2\pi i \tau}$, is arguably the most well-known half-integral weight modular form. Its modular transformation properties for a matrix $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z})$ are given by

$$\eta(A\tau) = v(A)(c\tau + d)^{1/2}\eta(\tau), \tag{1}$$

where

$$v(A) := \begin{cases} (d/|c|)\, e^{(i\pi/12)((a+d)c - bd(c^2-1) - 3c)} & \text{if } c \equiv 1 \pmod 2, \\ (c/|d|)\, e^{(i\pi/12)((a+d)c - bd(c^2-1) + 3d - 3 - 3cd)} & \text{if } c \equiv 0 \pmod 2. \end{cases} \tag{2}$$

MSC2020: primary 11F20, 11F37; secondary 11G05.
Keywords: eta-quotients, modular forms, elliptic curves.
This work was supported by NSF grant DMS-1359173.

Dedekind's eta-function has been featured prominently in work motivated by Ramanujan's study of the partition counting function $p(n)$ (see [Ono 2000; Ahlgren and Ono 2001] for example), as well as in the representation theory of the monster group, studied by Conway and Norton [1979], Borcherds [1992], and others. Due to its expression as a simple infinite product, it is easy to compute expansions numerically. Thus it is useful when a modular form can be expressed in terms of products or quotients of $\eta(\tau)$. By *eta-quotient of level N and integer weight k_f* we mean a function of the form

$$f(\tau)=\prod_{\delta\,|\,N}\eta(\delta\tau)^{r_\delta},$$

where N is a positive integer, $r_\delta\in\mathbb{Z}$, and $k_f=\frac{1}{2}\sum r_\delta\in\mathbb{Z}$.

Work on various classifications of eta-quotients has been of interest; see in particular [Dummit, Kisilevsky and McKay 1985; Martin 1996; Lemke Oliver 2013]. Moreover, the famous works [Wiles 1995; Taylor and Wiles 1995] proving Fermat's last theorem, and in particular the Shimura–Taniyama conjecture, showed that elliptic curves were attached to modular forms, and thus the question of when these modular forms can be expressed in terms of eta-quotients is a natural one.

Theorem 1.1 (modularity theorem [Diamond and Shurman 2005]). *Every elliptic curve E over $\mathbb{Q}$ with conductor N has an L-function*

$$L(E,s)=\sum_{n=1}^{\infty}\frac{a_E(n)}{n^s}$$

such that the Fourier series

$$f(\tau)=\sum_{n=1}^{\infty}a_E(n)q^n,\qquad q=e^{2\pi i\tau},\ \tau\in\mathbb{H},$$

represents a level-N cusp form of weight 2.

In light of Theorem 1.1, Martin and Ono [1997] classified all eta-quotients which are weight-2 newforms, as well as their associated elliptic curves. It is natural to ask when elliptic curves are associated to modular forms that can be written as linear combinations of eta-quotients. Recently, Pathakjee, RosnBrick, and Yoong [2012] demonstrated four such examples, utilizing spaces of cusp forms which are spanned by eta-quotients. Further work on when spaces of modular forms are spanned by eta-quotients has been done in [Rouse and Webb 2015; Arnold-Roksandich, James and Keaton 2018a; Kilford 2007], for example.

Given a congruence subgroup group $\Gamma\subseteq\mathrm{SL}_2(\mathbb{Z})$, we use the notation $S_k(\Gamma)$, $M_k(\Gamma)$, and $M_k^!(\Gamma)$ to denote the complex vector spaces of weight-k cusp forms, holomorphic modular forms, and weakly holomorphic modular forms, respectively. When $\Gamma=\Gamma_0(N)$ for a positive integer N, we write $S_k(\Gamma_0(N),\chi)$, $M_k(\Gamma_0(N),\chi)$, and $M_k^!(\Gamma_0(N), \chi)$ to denote the spaces of weight-k cusp forms, holomorphic modular forms, and weakly holomorphic modular forms, respectively, with Nebentypus χ.

In this paper, we first show that eta-quotients which are modular for any congruence subgroup of level N can be viewed as modular for $\Gamma_0(N)$. We then categorize when eta-quotients of even weight can exist in $M_k(\Gamma_1(p))$ and $M_k(\Gamma_1(pq))$ for distinct primes p, q. We conclude by providing some new examples of elliptic curves whose corresponding modular forms can be written as a linear combination of eta-quotients, and describe an algorithmic method for finding additional examples.

1A. *Viewing eta-quotients over* $\Gamma_0(N)$.

We first review a few key theorems about the modularity of eta-quotients and their orders of vanishing at cusps.

The following well-known theorem originating in [Newman 1957; 1959; Gordon and Hughes 1993], provides explicit modularity properties with respect to $\Gamma_0(N)$ for eta-quotients of specific shapes.

Theorem 1.2 [Ono 2004, Theorem 1.64]. *Let f be the eta-quotient of level N given by*

$$f(\tau) = \prod_{\delta \mid N} \eta(\delta\tau)^{r_\delta}.$$

If f satisfies both

$$\sum_{\delta \mid N} \delta r_\delta \equiv 0 \pmod{24} \quad \textit{and} \quad \sum_{\delta \mid N} \frac{N}{\delta} r_\delta \equiv 0 \pmod{24},$$

then for $k = \frac{1}{2}\sum_{\delta \mid N} r_\delta \in \mathbb{N}$ and $\chi(d) = ((-1)^k s/d)$, where $s = \prod_{\delta \mid N} \delta^{r_\delta}$, we have $f \in M_k^!(\Gamma_0(N), \chi)$, i.e., for all $M = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma_0(N)$,

$$f(M\tau) = \chi(d)(c\tau + d)^k f(\tau).$$

Remark 1.3. In the case where $\gcd(N, 6) = 1$, the two conditions above are equivalent. This is because any $\delta \mid N$ must satisfy $\gcd(\delta, 24) = 1$, and hence δ is its own inverse modulo 24. Thus,

$$\sum_{\delta \mid N} \frac{N}{\delta} r_\delta \equiv N \sum_{\delta \mid N} \delta r_\delta \pmod{24}.$$

As $\gcd(N, 24) = 1$, N is not a zero divisor in $\mathbb{Z}/24\mathbb{Z}$, so

$$N \sum_{\delta \mid N} \delta r_\delta \equiv 0 \pmod{24} \quad \text{if and only if} \quad \sum_{\delta \mid N} \delta r_\delta \equiv 0 \pmod{24}.$$

The next theorem originating in [Ligozat 1974], and further appearing in [Biagioli 1990; Martin 1996], provides a mechanism for calculating the relative orders of vanishing at cusps for eta-quotients satisfying Theorem 1.2.

Recall that if h is the width of the cusp r with respect to the group Γ, then the *order of vanishing relative to the group* Γ of a modular form f at the cusp r is given by $v_\Gamma(f, r) = h \cdot \mathrm{inv}_r(f)$, where $\mathrm{inv}_r(f)$ is the invariant order of vanishing of f at r and is always integral. We note that for the cusp at infinity, we always have $v_\Gamma(f, r) = \mathrm{inv}_r(f)$.

Theorem 1.4 [Ono 2004, Theorem 1.65]. *Let c, d, and N be positive integers with $d \mid N$ and* $\gcd(c, d) = 1$. *Then if $f(\tau)$ is an eta-quotient satisfying the conditions given in Theorem 1.2 for N, then the order of vanishing for $f(\tau)$ at the cusp c/d relative to* $\Gamma_0(N)$ *is*

$$v_{\frac{c}{d}} := v_{\Gamma_0(N)}\left(f, \frac{c}{d}\right) = \frac{N}{24}\sum_{\delta \mid N} \frac{\gcd(d, \delta)^2 r_\delta}{\gcd(d, N/d) d\delta}.$$

Calculating orders of vanishing of modular forms at cusps can be extremely useful, due to the following result known as Sturm's bound.

Theorem 1.5 (Sturm's bound [Murty, Dewar and Graves 2015]). *Let Γ be a congruence subgroup and $f \in M_k(\Gamma)$. Let $r_1, \dots, r_t$ be the Γ-inequivalent cusps of Γ. If*

$$\sum_{i=1}^{t} \mathrm{inv}_{r_i}(f) > \frac{k[\mathrm{SL}_2(\mathbb{Z}) : \{\pm I\}\Gamma]}{12},$$

then $f = 0$.

Remark 1.6. We note that Theorem 1.5 provides a direct way to check if two modular forms are equal by considering their q-expansions. Namely, if $f, g \in M_k(\Gamma)$ and their Fourier expansions at $i\infty$ agree to a power of q past the bound in Theorem 1.5, then they must be equal.

Our first theorem shows that when $\gcd(N, 6) = 1$, any eta-quotient which is modular for a congruence subgroup of level N is in fact modular for $\Gamma_0(N)$ for some Nebentypus character χ. We note that the hypothesis that $\gcd(N, 6) = 1$ is necessary; for example, one can show that $\eta(\tau)^2$ is a weight-1 modular form for $\Gamma(12)$.[1]

Remark 1.7. Let N be a positive integer with $\gcd(N, 6) = 1$ and suppose

$$f(\tau) = \prod_{\delta \mid N} \eta(\delta\tau)^{r_\delta} \in M^!_{k_f}(\Gamma(N)).$$

Then since $\left(\begin{smallmatrix} 1 & N \\ 0 & 1 \end{smallmatrix}\right) \in \Gamma(N)$, it follows that f has a Fourier expansion that is a power series in $q^{1/N}$. Of course, as an eta-quotient f can also be expressed as a power series in $q^{1/24}$, in particular with first term $q^{1/24 \sum \delta r_\delta}$. Since $\gcd(N, 24) = 1$,

[1] See https://mathoverflow.net/questions/297049/is-eta-tau2-a-modular-form-of-weight-1-on-gamma12/297053, where Jeremy Rouse describes why this is true.

it follows that $\sum_{\delta \mid N} \delta r_\delta \equiv 0 \pmod{24}$. Thus by Theorem 1.2 and Remark 1.3 we can conclude that $f(\tau) \in M^!_{k_f}(\Gamma_0(N), \chi)$, where χ is defined as in Theorem 1.2.

Remark 1.8. Let N be a positive integer with $\gcd(N, 6) = 1$. Using that fact that $M^!_k(\Gamma_0(N), \chi) \subset M^!_k(\Gamma_1(N))$ as well as Remarks 1.7 and 1.3, it follows that $f(\tau) = \prod_{\delta \mid N} \eta(\delta\tau)^{r_\delta} \in M^!_{k_f}(\Gamma_1(N))$ if and only if $f(\tau)$ satisfies the conditions in Theorem 1.2.

From Remark 1.8, we see that despite there being many more cusps of $\Gamma_1(N)$ than $\Gamma_0(N)$, we only need to consider the cusps of $\Gamma_0(N)$ when calculating orders of vanishing of eta-quotients in $M^!_k(\Gamma_1(N))$. In [Martin 1996], a complete set of representatives for the cusps of $\Gamma_0(N)$ is given by

$$C_{\Gamma_0(N)} = \left\{ \frac{a_c}{c} \in \mathbb{Q} : c \mid N,\ 1 \le a_c \le N,\ \gcd(a_c, N) = 1 \right. \\ \left. \text{and } a_c \equiv a'_c \pmod{\gcd(c, N/c)} \text{ if and only if } a_c = a'_c \right\}.$$

In the case where N is square-free, $\gcd(c, N/c) = 1$ for all c which divide N. This gives the following complete set of representatives:

$$C_{\Gamma_0(N)} = \left\{ \frac{1}{d} \in \mathbb{Q} : d \mid N \right\}. \tag{3}$$

1B. *Classifications.* Our first results deal with classifying when eta-quotients can exist in spaces of holomorphic modular forms of even weight and prime or semiprime level. The following theorem uses techniques from [Arnold-Roksandich, James and Keaton 2018a] and also addresses an error in Corollary 3.2 of that paper.

Theorem 1.9. *Let $p \ge 5$ be prime, set $h = \frac{1}{2}\gcd(p-1, 24)$, and let k be an even integer. Then there exists $f = \eta(\tau)^{r_1}\eta(p\tau)^{r_p} \in M_k(\Gamma_1(p))$ if and only if both of the following conditions hold*:

(1) $h \mid k$.

(2) *It is not the case that $p \ne 5$, $p \equiv 5 \pmod{24}$, and $k = 2$.*

We also prove a theorem of similar flavor for eta-quotients of semiprime level, which extends [Arnold-Roksandich, James and Keaton 2018b, Theorem 5.8].

Theorem 1.10. *Let $p, q \ge 5$ be distinct primes, $N = pq$, and k an even integer. Let $h = \frac{1}{2}\gcd(24, p-1, q-1)$. Then there exists $f(\tau) = \prod_{\delta \mid N} \eta(\delta\tau)^{r_\delta} \in M_k(\Gamma_1(pq))$ if and only if both of the following conditions hold*:

(1) $h \mid k$.

(2) *It is not the case that $(p, q) \pmod{24} \in \{(1, 5), (5, 1), (5, 5)\}$, $p > 5$ and $q > 5$, and $k = 2$.*

1C. ***Writing modular forms attached to elliptic curves in terms of eta-quotients.*** Utilizing our work in Section 3, we explore when $S_2(\Gamma_0(pq))$ has a basis consisting of eta-quotients. This allows us to provide examples of elliptic curves attached via the modularity theorem (Theorem 1.1) to linear combinations of eta-quotients.

Theorem 1.11. *Let E be an elliptic curve with conductor* 35. *Then E is associated via the modularity theorem to the modular form $f(\tau) \in S_2(\Gamma_0(35))$ given by*

$$f(\tau) = \eta(\tau)^2\eta(35\tau)^2 + \eta(5\tau)^2\eta(7\tau)^2.$$

We are also able to provide an example even when $S_2(\Gamma_0(pq))$ does not have a basis consisting of eta-quotients.

Theorem 1.12. *Let E be an elliptic curve with conductor* 55. *Then E is associated via the modularity theorem to the modular form $f(\tau) \in S_2(\Gamma_0(55))$ given by*

$$f(\tau) = \sum_{i=1}^{40} c_i \frac{g_i(\tau)}{a(\tau)},$$

where in Section 4 the coefficients c_i are given in Table 2, and the eta-quotients $g_i(\tau)/a(\tau)$ are given in Table 3. The first three $g_i(\tau)/a(\tau)$, $i = 1, 2, 3$, are holomorphic modular forms.

1D. ***Outline of the rest of the paper.*** The rest of the paper is devoted to proving our results, and surrounding discussions. In Section 2, we prove Theorem 1.9, and in Section 3, we prove Theorem 1.10, discussing also square-free level cases. In Section 4, we prove Theorems 1.11 and 1.12.

2. Eta-quotients of prime level

In this section our goal is to prove Theorem 1.9, which classifies when eta-quotients can exist in $M_k^!(\Gamma_1(p))$ for p prime. And of course by Remark 1.7 we know that eta-quotients in $M_k^!(\Gamma_1(p))$ are actually in $M_k^!(\Gamma_0(p), \chi)$ for some character χ. By (3), there are two cusps of $\Gamma_0(p)$: 1 and $1/p$. We first state the following useful result of [Arnold-Roksandich, James and Keaton 2018a], but offer an alternative proof which we will see in Section 3 generalizes to square-free levels.

Lemma 2.1 [Arnold-Roksandich, James and Keaton 2018a, Theorem 1.2]. *Fix a prime $p \geq 5$ and let $h = \frac{1}{2}\gcd(p-1, 24)$. There exists $f(\tau) = \eta(\tau)^{r_1}\eta(p\tau)^{r_p}$ in $M_k^!(\Gamma_1(p))$ if and only if $h \mid k$.*

Proof. Suppose there exists $f(\tau) = \eta(\tau)^{r_1}\eta(p\tau)^{r_p}$ in $M_k^!(\Gamma_1(p))$. By Remark 1.8 we see that f satisfies the conditions in Theorem 1.4, and thus

$$v_1 = \tfrac{1}{24}(pr_1 + r_p). \tag{4}$$

Since $k = \frac{1}{2}(r_1 + r_p)$, (4) is equivalent to

$$24v_1 - (p-1)r_1 = 2k. \tag{5}$$

Equation (5) can be viewed as a linear Diophantine equation in variables v_1 and r_1. In this light, the existence of a solution implies $\gcd(p-1, 24) \mid 2k$. As $p \neq 2$, we know $p-1$ is even and so $2 \mid \gcd(24, p-1)$. Therefore, $h = \frac{1}{2}\gcd(p-1, 24)$ is an integer and $h \mid k$.

Conversely, if $h \mid k$, then there exist integers v_1, r_1 which give a solution to (5). Plugging these into (4) gives an integer r_p. From (4) we see the first condition of Theorem 1.2 is satisfied, and so, since $(p, 6) = 1$, Remark 1.3 implies that $f(\tau) \in M_k^!(\Gamma_1(p))$. □

Our proof of Theorem 1.9 will also rely on the following result of [Rouse and Webb 2015].

Theorem 2.2 [Rouse and Webb 2015, Theorem 2]. *Suppose*

$$f(\tau) = \prod_{\delta \mid N} \eta(\delta\tau)^{r_\delta} \in M_{k_f}(\Gamma_0(N)).$$

Then we have

$$\sum_{\delta \mid N} |r_\delta| \leq 2k \prod_{p \mid N} \left(\frac{p+1}{p-1}\right)^{\min\{2, \operatorname{ord}_p(N)\}},$$

where $\operatorname{ord}_p(N)$ *denotes the p-adic valuation of N.*

Remark 2.3. The proof of this theorem carries through for spaces $M_k(\Gamma_0(N), \chi)$ with character. Thus in light of Remark 1.8, Theorem 2.2 holds for any eta-quotient in $M_k(\Gamma_1(N))$.

We also require two additional lemmas which will allow us to deal with certain cases in the proof of Theorem 1.9. These lemmas both deal with the value $k(p+1)/12$, where $p \geq 5$ is prime, and k is an even integer such that $h = \frac{1}{2}\gcd(p-1, 24) \mid k$. Note that in this case $k(p+1)/12$ must be an integer, since if 3 doesn't divide $p+1$, then 3 must divide $p-1$, and so 3 divides h and thus k.

Lemma 2.4. *Fix a prime $p \geq 5$, and let k be a positive even integer. If $k(p+1)/12$ is even, then there exists an eta-quotient in $M_k(\Gamma_1(p))$.*

Proof. Consider the eta-quotient

$$f(\tau) = \eta(\tau)^k \eta(p\tau)^k.$$

Since $k(p+1)/12$ is even, and by Remark 1.3, we see that $f(\tau)$ satisfies the conditions of Theorem 1.2, and so $f(\tau) \in M_k^!(\Gamma_1(p))$. Moreover, by Theorem 1.4,

$$v_1 = v_{1/p} = \tfrac{1}{24}k(p+1), \tag{6}$$

which is a positive integer. Thus $f(\tau) \in M_k(\Gamma_1(p))$. □

Remark 2.5. Note that we did not need the assumption that $h \mid k$ in the previous lemma. This is because when $k(p+1)/12$ is even, we must have $h \mid k$. To see this, observe that if $k(p+1)/12 = 2n$, then $2k = 24n - k(p-1)$, and so $\gcd(24,\, p-1) \mid 2k$.

The last lemma is a simple divisibility argument.

Lemma 2.6. *Let p be prime, $h = \frac{1}{2}\gcd(p-1, 24)$, and k be an even integer. If $h \mid k$ and $k(p+1)/12$ is odd, then $(p-1)/(2h)$ is odd.*

Proof. We first note that if $p = 2$, then $(p-1)/(2h) = 1$ and so is odd regardless of k. We now let $p \geq 3$. Suppose $(p-1)/(2h)$ is even. By the definition of h, $(p-1)/(2h)$ and $12/h$ are relatively prime, so $12/h$ must be odd. But then 12 is an odd integer times h, so we have that $4 \mid h$, and so $4 \mid k$. Thus $8 \mid k(p+1)$, since p is odd. Therefore $k(p+1)/12$ is even, which contradicts our assumptions. □

We are now able to prove Theorem 1.9.

Proof of Theorem 1.9. First, we assume that there exists

$$f(\tau) = \eta(\tau)^{r_1}\eta(p\tau)^{r_p} \in M_k(\Gamma_1(p)).$$

By Lemma 2.1, we have that $h \mid k$. To complete this direction of the proof, we show that there are no eta-quotients in $M_2(\Gamma_1(p))$ when $p \equiv 5 \pmod{24}$ and $p > 5$. To do this, we recall Remark 2.3, and employ Theorem 2.2. Fix $p \equiv 5 \pmod{24}$ with $p > 5$, and suppose

$$f(\tau) = \eta(\tau)^{r_1}\eta(p\tau)^{r_p} \in M_2(\Gamma_1(p)).$$

Theorem 2.2 gives us that

$$|r_1| + |r_p| \leq 4\left(\frac{p+1}{p-1}\right).$$

This upper bound decreases as p increases, and so will be largest when $p = 29$, which gives a bound less than 5. Since r_1 and r_p are integers, we have

$$|r_1| + |r_p| \leq 4. \tag{7}$$

Moreover, by Theorem 1.4 we have

$$24v_1 = pr_1 + r_p,$$
$$24v_{1/p} = r_1 + pr_p.$$

Since $f(\tau) \in M_k(\Gamma_1(p))$, it must be that $v_1, v_{1/p} \geq 0$. This guarantees that $r_1, r_p \geq 0$. Namely, if $r_1 < 0$, then using (7), we see that

$$24v_{1/p} \leq -29|r_1| + |r_p| < 0,$$

which contradicts that $v_{1/p} \geq 0$. Similarly, $r_p < 0$ implies that $v_1 < 0$.

Since $p \equiv 5 \pmod{24}$, Remark 1.8 implies that

$$r_1 + 5r_p \equiv 0 \pmod{24}. \tag{8}$$

However, there is no pair of nonnegative integers r_1, r_p, not both zero, that satisfy both (7) and (8). Thus we have a contradiction and so no such eta-quotient $f(\tau)$ can exist.

We now prove the converse by construction. Suppose that $p \geq 5$ is a prime and k is an even integer such that $h = \frac{1}{2}\gcd(p-1, 24) \mid k$. We construct an eta-quotient in $M_k(\Gamma_1(p))$ for every such case of p and k except when $k = 2$ and $p \equiv 5 \pmod{24}$, with $p > 5$.

We first consider the case where $p \not\equiv 5 \pmod 8$. By Lemma 2.4, it suffices to prove that $k(p+1)/12$ is even, which, since $k(p+1)/12$ is an integer, means showing that $8 \mid k(p+1)$. We already know that both k and $p+1$ are even, so it suffices to show that either k or $p+1$ is divisible by 4. If $p \equiv 3 \pmod 4$, then $4 \mid p+1$ so we are done. If not, then $p \equiv 1 \pmod 8$ and we have that $4 \mid h$ and so $4 \mid k$.

We now consider the case when $p \equiv 5 \pmod 8$. In this case either $p \equiv 13 \pmod{24}$ or $p \equiv 5 \pmod{24}$. We will show the existence of an eta-quotient in $M_h(\Gamma_1(p))$, since if $f(\tau) \in M_h(\Gamma_1(p))$, then $f(\tau)^{k/h} \in M_k(\Gamma_1(p))$. Suppose that $p \equiv 13 \pmod{24}$. Then $h = 6$, and using Theorems 1.2 and 1.4 one can quickly check that

$$\eta(\tau)^9 \eta(p\tau)^3 \in M_6(\Gamma_1(p)).$$

Next, when $p = 5$ we have $h = 2$, and so

$$\frac{\eta(p\tau)^5}{\eta(\tau)} \in M_2(\Gamma_1(p)).$$

Finally, suppose $p \equiv 5 \pmod{24}$, with $p \neq 5$ and $k \neq 2$. It is sufficient to show that there exist eta-quotients $f_4(\tau) \in M_4(\Gamma_1(p))$ and $f_6(\tau) \in M_6(\Gamma_1(p))$ since either $4 \mid k$ or $4 \mid (k-6)$ and thus either $f_4^{k/4} \in M_k(\Gamma_1(p))$ or $f_4^{(k-6)/4} f_6 \in M_k(\Gamma_1(p))$. We check using Theorems 1.2 and 1.4 that

$$f_4(\tau) := \eta(\tau)^4 \eta(p\tau)^4 \in M_4(\Gamma_1(p)),$$
$$f_6(\tau) := \eta(\tau)^9 \eta(p\tau)^3 \in M_6(\Gamma_1(p)),$$

which resolves the final case. □

3. Eta-quotients of semiprime and square-free level

In this section, we generalize the results of Section 2 to semiprime level, and further to square-free levels where possible. We begin with a square-free generalization of Lemma 2.1 which extends [Arnold-Roksandich, James and Keaton 2018a, Theorem 5.8].

Theorem 3.1. *Let $k \in \mathbb{Z}$ and let $N = p_1 \cdots p_t$, a product of distinct primes with $p_i \geq 5$. Define $h = \frac{1}{2}\gcd(p_1 - 1, \ldots, p_t - 1, 24)$. There exists*

$$f(\tau) = \prod_{\delta \mid N} \eta(\delta\tau)^{r_\delta} \in M_k^!(\Gamma_1(N))$$

if and only if $h \mid k$.

Proof. Suppose there exists

$$f(\tau) = \prod_{\delta \mid N} \eta(\delta\tau)^{r_\delta} \in M_k^!(\Gamma_1(N)).$$

By Remark 1.8 we see that f satisfies the conditions in Theorem 1.4, and thus we can compute

$$v_{1/N} = \frac{1}{24}\sum_{\delta \mid N} \delta r_\delta. \tag{9}$$

Since $k = k_f = \frac{1}{2}\sum_{\delta \mid N} r_\delta$, (9) is equivalent to

$$2k = 24v_{1/N} - \sum_{\delta \mid N} (\delta - 1) r_\delta. \tag{10}$$

Equation (10) can be viewed as a linear Diophantine equation in the variables $v_{1/N}$ and r_δ for each $\delta \mid N$ with $\delta \neq 1$. Thus, the existence of an integer solution to (10) implies that

$$\gcd(\delta_1 - 1, \ldots, \delta_s - 1, 24) \mid 2k, \tag{11}$$

where $\delta_1, \ldots, \delta_s$ are the divisors of N. However, we note that

$$\gcd(\delta_1 - 1, \ldots, \delta_s - 1, 24) = \gcd(p_1 - 1, \ldots, p_t - 1, 24) = 2h, \tag{12}$$

which follows from the fact that $ab - 1 = (a-1)(b-1) + (a-1) + (b-1)$, so if $d \mid (a-1)$ and $d \mid (b-1)$, then $d \mid (ab-1)$. As each p_i is odd, we have that $2 \mid \gcd(p_1 - 1, \ldots, p_t - 1, 24)$. Therefore, h is an integer and we see from (11) and (12) that $h \mid k$ as desired.

Conversely, if $h \mid k$, then there exist integers $v_{1/N}$ and r_δ for each $\delta \mid N$ with $\delta \neq 1$, which give a solution to (10). Additionally defining

$$r_1 = 2k - \sum_{\delta \mid N,\, \delta \neq 1} r_\delta$$

gives that $k = \frac{1}{2}\sum_{\delta \mid N} r_\delta$. Thus from (10) we see that

$$\sum_{\delta \mid N} \delta r_\delta = 24 v_{1/N} \equiv 0 \pmod{24},$$

and so by Remark 1.3,

$$f(\tau) := \prod_{\delta \mid N} \eta(\delta\tau)^{r_\delta} \in M_k^!(\Gamma_1(N)). \qquad \square$$

The next theorem computes the sum of the orders of vanishing of an eta-quotient in $M_k^!(\Gamma_1(N))$ in terms of the divisor function

$$\sigma_a(n) := \sum_{d \mid n} d^a.$$

Theorem 3.2. *Let $N = p_1 \cdots p_t$, a product of distinct primes with $p_i \geq 5$, and let $f(\tau) = \prod_{\delta \mid N} \eta(\delta\tau)^{r_\delta} \in M_{k_f}^!(\Gamma_1(N))$. Then*

$$\sum_{d \mid N} v_{1/d} = \frac{k\sigma_1(N)}{12}.$$

Proof. Since N is square-free, Theorem 1.4 gives that

$$\sum_{d \mid N} v_{1/d} = \sum_{d \mid N} \frac{N}{24} \sum_{\delta \mid N} \frac{\gcd(d,\delta)^2 r_\delta}{d\delta} = \frac{1}{24} \sum_{\delta \mid N} r_\delta \sum_{d \mid N} \frac{N \gcd(d,\delta)^2}{d\delta}. \tag{13}$$

Since $\sum_{\delta \mid N} r_\delta = 2k$, it suffices to show that the inner sum in (13) is $\sigma_1(N)$. We thus fix $\delta \mid N$, and aim to show that, for each divisor d' of N, there is a unique divisor $d \mid N$ such that

$$\frac{N \gcd(d,\delta)^2}{d\delta} = d'. \tag{14}$$

We first show existence. Given $d' \mid N$, set $d = \gcd(d',\delta)\gcd(N/d', N/\delta)$, which clearly divides N. Since δ, d', and d are all square-free, we observe that $\gcd(d,\delta) = \gcd(d',\delta)$, since any prime divisor of d and δ must divide $\gcd(d',\delta)$.

Our goal is to show (14), which we rewrite as

$$\frac{N}{\gcd(N/d', N/\delta)} = \frac{\delta d'}{\gcd(d',\delta)}. \tag{15}$$

The left-hand side of (15) is simply the product of the prime divisors of N which are also prime divisors of δ or d'. But this is also the right-hand side of (15) so we are done and have shown existence.

To determine uniqueness, we suppose $d_1, d_2 \mid N$ such that

$$\frac{N \gcd(d_1,\delta)^2}{d_1\delta} = \frac{N \gcd(d_2,\delta)^2}{d_2\delta},$$

namely that $d_2 \gcd(d_1,\delta)^2 = d_1 \gcd(d_2,\delta)^2$. Since N is square-free, comparing the prime factorizations of d_1, d_2, δ and considering cases yields that $d_1 = d_2$ as desired.

Since we have shown that

$$\sum_{d \mid N} \frac{N \gcd(d,\delta)^2}{d\delta} = \sigma_1(N), \tag{16}$$

we are now finished. □

Remark 3.3. When N is square-free, $\sigma_1(N) = \prod_{p \mid N}(p+1)$. Thus when $N = p_1 \cdots p_t$, a product of distinct primes with $p_i \geq 5$, and

$$f(\tau) = \prod_{\delta \mid N} \eta(\delta\tau)^{r_\delta} \in M^!_{k_f}(\Gamma_1(N)),$$

we have by Theorem 3.2 that

$$\sum_{d \mid N} v_{1/d} = \frac{k\sigma_1(N)}{12} = \frac{k\prod_{p \mid N}(p+1)}{12}.$$

From this we can observe that $k\sigma_1(N)/12$ must always be an integer whenever k is an even integer such that $h \mid k$ for $h = \frac{1}{2}\gcd(p_1 - 1, \dots, p_t - 1, 24)$. This is because if 3 doesn't divide $p+1$ for any $p \mid N$, then 3 must divide $p-1$ for all $p \mid N$, and so 3 divides h and thus k.

We next generalize Lemma 2.4 to square-free levels.

Lemma 3.4. *Let $N = p_1 \cdots p_t$, a product of distinct primes with $p_i \geq 5$, and define $h = \frac{1}{2}\gcd(p_1 - 1, \dots, p_t - 1, 24)$. Suppose k is a positive integer such that $h \mid k$, $2^{t-1} \mid k$, and $4 \mid k$ if $t = 2$. Then there exists*

$$f(\tau) = \prod_{\delta \mid N} \eta(\delta\tau)^{r_\delta} \in M_k(\Gamma_1(N)).$$

Proof. Define $f(\tau)$ by

$$f(\tau) = \prod_{\delta \mid N} \eta(\delta\tau)^{k/2^{t-1}}.$$

As N is square-free, there are 2^t divisors of N, and so

$$\sum_{\delta \mid N} \frac{k}{2^{t-1}} = 2k.$$

Additionally, by our hypothesis $2^t \mid (k\sigma_1(N)/12)$, so we have

$$\sum_{\delta \mid N} \delta\left(\frac{k}{2^{t-1}}\right) = \left(\frac{k}{2^{t-1}}\right)\sigma_1(N) = \frac{k\sigma_1(N)}{2^{t-1}} \equiv 0 \pmod{24}.$$

Thus, $f(\tau)$ satisfies Theorem 1.2, and so by Remark 1.8, $f(\tau) \in M^!_k(\Gamma_1(N))$. Recalling (16), we have by Theorem 1.4 that

$$v_{1/d} = \frac{k}{2^{t-1}} \cdot \frac{\sigma_1(N)}{24} = \frac{k\sigma_1(N)}{2^t \cdot 12},$$

which is a positive integer by our hypotheses since $2^t \mid \sigma_1(N)$. Thus $f(\tau) \in M_k(\Gamma_1(N))$. □

We are now ready to prove Theorem 1.10.

Proof of Theorem 1.10. First, we assume that

$$f(\tau)=\eta(\tau)^{r_1}\eta(p\tau)^{r_p}\eta(q\tau)^{r_q}\eta(N\tau)^{r_N}\in M_k(\Gamma_1(pq)),$$

where $N=pq$ is a product of distinct primes $p,q\geq 5$, and k is an even integer. By Theorem 3.1, we have that $h\mid k$. To complete this forward direction of the proof, we show that there are no eta-quotients in $M_2(\Gamma_1(pq))$ when $(p,q)\equiv(1,5),(5,1),(5,5) \pmod{24}$ and $p,q>5$. Without loss of generality, we may assume p and q are chosen so that p has a lesser or equal residue modulo 24.

Recalling Remark 2.3, we use Theorem 2.2. Fix $(p,q)\equiv(1,5),(5,5) \pmod{24}$ with $p,q>5$, and suppose

$$f(\tau)=\eta(\tau)^{r_1}\eta(p\tau)^{r_p}\eta(q\tau)^{r_q}\eta(N\tau)^{r_N}\in M_2(\Gamma_1(pq)).$$

Then by Theorem 2.2,

$$|r_1|+|r_p|+|r_q|+|r_N|\leq 4\left(\frac{p+1}{p-1}\right)\left(\frac{q+1}{q-1}\right).$$

This upper bound decreases as p and q increases, so the bound will be largest when $p=29$ and $q=53$. This gives a bound less than 5, so, since the r_δ are integers,

$$|r_1|+|r_p|+|r_q|+|r_N|\leq 4. \tag{17}$$

Moreover, by Theorem 1.4,

$$\begin{aligned}24v_1&=Nr_1+qr_p+pr_q+r_N,\\24v_{1/p}&=qr_1+Nr_p+r_q+pr_N,\\24v_{1/q}&=pr_1+r_p+Nr_q+qr_N,\\24v_{1/N}&=r_1+pr_p+qr_q+Nr_N.\end{aligned}$$

Since $f(\tau)\in M_k(\Gamma_1(N))$, we have that $v_1,v_{1/p},v_{1/q},v_{1/N}\geq 0$. It follows that $r_1,r_p,r_q,r_N\geq 0$. Namely, if $r_1<0$, then using (17), and, assuming without loss of generality that $p<q$, we see that

$$24v_1\leq -pq|r_1|+q|r_p|+q|r_q|+q|r_N|\leq -pq|r_1|+4q<-pq|r_1|+pq\leq 0,$$

which contradicts that $v_1\geq 0$. Similarly, if r_p, r_q, or r_N is negative we get contradictions for the nonnegativity of $v_{1/p}$, $v_{1/q}$, and $v_{1/N}$ respectively.

By Remark 1.8, we also know that

$$r_1+r_p+5r_q+5r_N\equiv 0 \pmod{24} \tag{18}$$

if $(p,q)\equiv(1,5) \pmod{24}$, and

$$r_1+5r_p+5r_q+r_N\equiv 0 \pmod{24} \tag{19}$$

if $(p,q)\equiv(5,5) \pmod{24}$.

However, neither (18) nor (19) have nonnegative integer solutions which satisfy (17), which is a contradiction. Thus we have shown that there are no eta-quotients in $M_2(\Gamma_1(pq))$ for $(p, q) \equiv (1, 5), (5, 1), (5, 5) \pmod{24}$ and $p, q > 5$.

We now prove the converse by construction. Namely, we need to show that if $h \mid k$ and it is not the case that $(p, q) \pmod{24} \in \{(1, 5), (5, 1), (5, 5)\}$, $p, q \neq 5$, and $k = 2$, then there does exist an eta-quotient in $M_k(\Gamma_1(pq))$.

We first note that setting $t = 2$ in Lemma 3.4 guarantees the existence of an eta-quotient in $M_k(\Gamma_1(pq))$ for distinct primes $p, q \geq 5$, when k is an even integer divisible by h such that $k(p+1)(q+1)/12$ is divisible by 4. Since 4 divides $k(p+1)(q+1)/12$ whenever 4 divides any one of $p+1$, $q+1$, or k, it suffices to consider only the cases of p, q, and k when $p+1 \equiv q+1 \equiv k \equiv 2 \pmod 4$. Consider the possible residues for (p, q) modulo 24, ordering so that the residue of p is no larger than that of q. We may immediately disregard the cases $(1, 1)$, $(1, 17)$, and $(17, 17)$, since in each $4 \mid h$, and thus since $h \mid k$ they are covered by Lemma 3.4. This leaves the cases $(1, 5)$, $(1, 13)$, $(5, 5)$, $(5, 13)$, $(5, 17)$, $(13, 13)$, and $(13, 17)$. It suffices to show there exists an eta-quotient in $M_h(\Gamma_1(N))$, since if $f(\tau) \in M_h(\Gamma_1(p))$, then $f(\tau)^{k/h} \in M_k(\Gamma_1(p))$. The following table gives such eta-quotients for the cases $(1, 13)$, $(5, 13)$, $(5, 17)$, $(13, 13)$, and $(13, 17)$:

case	h	eta-quotient
$(1, 13)$	6	$\eta(\tau)^{11}\eta(q\tau)$
$(5, 13)$	2	$\eta(p\tau)\eta(q\tau)^2\eta(N\tau)$
$(5, 17)$	2	$\eta(p\tau)\eta(q\tau)\eta(N\tau)^2$
$(13, 13)$	6	$\eta(q\tau)\eta(N\tau)^{11}$
$(13, 17)$	2	$\eta(p\tau)^2\eta(q\tau)\eta(N\tau)$

This leaves the cases $(1, 5)$ and $(5, 5)$, both of which have $h = 2$. If either p or q is 5, then $M_2(\Gamma(5)) \subset M_2(\Gamma(N))$, and so by Theorem 1.9 there will exist an eta-quotient in $M_2(\Gamma(N))$. We thus assume that neither p nor q is equal to 5. As every even integer $k \geq 4$ can be written as a linear combination of 4 and 6, it suffices to show the existence of an eta-quotient in both $M_4(\Gamma_1(N))$ and $M_6(\Gamma_1(N))$. By Lemma 3.4, we have the existence of an eta-quotient in $M_4(\Gamma_1(N))$. Moreover, one can check using Theorem 1.2 that $\eta(\tau)^3\eta(N\tau)^9 \in M_6(\Gamma_1(N))$ when $(p, q) \equiv (1, 5) \pmod{24}$, and $\eta(q\tau)^3\eta(N\tau)^9 \in M_6(\Gamma_1(N))$ when $(p, q) \equiv (5, 5) \pmod{24}$. □

4. Elliptic curves and eta-Quotients

In this section, we prove Theorems 1.11 and 1.12, and conclude by describing the method we used to find these examples.

Proof of Theorem 1.11. First using dimension formulas (Theorems 3.5.1 and 3.1.1 in [Diamond and Shurman 2005] for example), we calculate that $\dim_{\mathbb{C}} S_2(\Gamma_0(35)) = 3$.

By Theorems 1.2 and 1.4, we see that the following three eta-quotients are members of $S_2(\Gamma_0(35))$:

$$\begin{aligned} g_1(\tau) &:= \eta(\tau)\eta(5\tau)\eta(7\tau)\eta(35\tau), \\ g_2(\tau) &:= \eta(\tau)^2\eta(35\tau)^2, \\ g_3(\tau) &:= \eta(5\tau)^2\eta(7\tau)^2. \end{aligned}$$

The q-expansions of g_1, g_2, g_3 begin with

$$\begin{aligned} g_1(\tau) &= q^2 - q^3 - q^4 + q^8 + q^9 + O(q^{10}), \\ g_2(\tau) &= q^3 - 2q^4 - q^5 + 2q^6 + q^7 + 2q^8 - 2q^9 + O(q^{10}), \\ g_3(\tau) &= q - 2q^6 - 2q^8 + O(q^{10}). \end{aligned}$$

Since each q-expansion starts with a different power of q, we can quickly determine that g_1, g_2, g_3 are linearly independent. Thus, they form a basis of $S_2(\Gamma_0(35))$, and by Theorem 1.1, any elliptic curve of conductor 35 must be a linear combination of g_1, g_2, and g_3. However, one can see for example from the L-functions and modular forms database [LMFDB 2019] that there is only one isogeny class of elliptic curves of conductor 35 [LMFDB 2019, elliptic curve isogeny class 35.a], and thus only one attached modular form, $f(\tau) \in S_2(\Gamma_0(35))$ [LMFDB 2019, modular form 35.2.1.a]. The q-expansion of f begins with

$$f(\tau) = q + q^3 - 2q^4 - q^5 + q^7 - 2q^9 + O(q^{10}),$$

and since the bound in Theorem 1.5 is 8 in this case, we see by Remark 1.6 that $f(\tau) = g_2(\tau) + g_3(\tau)$, as desired. □

We now turn to the proof of Theorem 1.12, which requires more finesse.

Proof of Theorem 1.12. As with conductor 35, there is only one isogeny class of elliptic curves of conductor 55 [LMFDB 2019, elliptic curve isogeny class 55.a], and thus only one attached modular form, $f(\tau) \in S_2(\Gamma_0(55))$ [LMFDB 2019, modular form 55.2.1.a]. The q-expansion of f begins with

$$f(\tau) = q + q^2 - q^4 + q^5 - 3q^8 - 3q^9 + q^{10} - q^{11} + 2q^{13} + O(q^{15}).$$

The bound in Theorem 1.5 is 12 in this case, so we see by Remark 1.6 that modular forms in $S_2(\Gamma_0(55))$ are determined by their Fourier coefficients up to q^{13}.

Using dimension formulas, we calculate that $\dim_{\mathbb{C}} S_2(\Gamma_0(55)) = 5$. However, there are only three linearly independent eta-quotients in $S_2(\Gamma_0(55))$ given by

$$\begin{aligned} g_1(\tau) &:= \eta(\tau)\eta(5\tau)\eta(11\tau)\eta(55\tau), \\ g_2(\tau) &:= \eta(\tau)^2\eta(11\tau)^2, \\ g_3(\tau) &:= \eta(5\tau)^2\eta(55\tau)^2, \end{aligned}$$

with q-expansions beginning with

$$\begin{aligned} g_1(\tau) &= q^3 - q^4 - q^5 + q^9 + 2q^{10} - 2q^{13} + O(q^{15}), \\ g_2(\tau) &= q - 2q^2 - q^3 + 2q^4 + q^5 + 2q^6 - 2q^7 - 2q^9 - 2q^{10} \\ &\qquad\qquad + q^{11} - 2q^{12} + 4q^{13} + 4q^{14} + O(q^{15}), \\ g_3(\tau) &= q^5 - 2q^{10} + O(q^{15}). \end{aligned}$$

We thus use a method originating in [Rouse and Webb 2015], which is to multiply $f(\tau)$ by an eta-quotient $a(\tau)$ in order to push the product $f(\tau)a(\tau)$ into a higher weight space that is generated by eta-quotients. Then f can be written as a linear combination of weakly holomorphic eta-quotients in $M_2^!(\Gamma_0(55))$.

Consider the eta-quotient

$$a(\tau) = \eta(\tau)^3\eta(5\tau)^3\eta(11\tau)^3\eta(55\tau)^3.$$

We see that $a(\tau) \in S_6(\Gamma_0(55))$ by Theorems 1.2 and 1.4, and so $a(\tau)f(\tau) \in S_8(\Gamma_0(55))$. Since the bound from Theorem 1.5 is 48 in this case, modular forms in $S_8(\Gamma_0(55))$ are determined by their Fourier coefficients up to q^{49}. We see that the q-expansion for $a(\tau)$ begins with

$$\begin{aligned} a(\tau) = q^9 &- 3q^{10} + 5q^{12} - 3q^{14} + 2q^{15} - 15q^{17} + 9q^{19} + 18q^{20} + 9q^{21} - 15q^{23} \\ &- 33q^{24} - 6q^{25} - 6q^{26} + 25q^{27} + 45q^{28} + 33q^{29} - 49q^{30} - 63q^{31} + 45q^{34} \\ &+ 60q^{35} + 45q^{36} - 15q^{37} - 75q^{38} - 62q^{39} - 78q^{40} + 66q^{41} + 15q^{42} \\ &- 15q^{43} + 21q^{45} + 117q^{46} - 15q^{47} + 55q^{48} - 63q^{49} + O(q^{50}), \end{aligned}$$

whereas the expansion for $f(\tau)$ begins with

$$\begin{aligned} f(\tau) = q &+ q^2 - q^4 + q^5 - 3q^8 - 3q^9 + q^{10} - q^{11} + 2q^{13} - q^{16} + 6q^{17} \\ &- 3q^{18} - 4q^{19} - q^{20} - q^{22} + 4q^{23} + q^{25} + 2q^{26} + 6q^{29} - 8q^{31} \\ &+ 5q^{32} + 6q^{34} + 3q^{36} - 2q^{37} - 4q^{38} - 3q^{40} + 2q^{41} + 4q^{43} \\ &+ q^{44} - 3q^{45} + 4q^{46} - 12q^{47} - 7q^{49} + O(q^{50}). \end{aligned}$$

Thus we see that the q-expansion for the product $f(\tau)a(\tau)$ starts with

$$\begin{aligned} f(\tau)a(\tau) = q^{10} &- 2q^{11} - 3q^{12} + 4q^{13} + 9q^{14} - 6q^{15} - 6q^{16} + 4q^{17} - 6q^{18} \\ &- 10q^{19} - 8q^{20} + 30q^{21} + 28q^{22} - 8q^{23} - 33q^{24} + 20q^{25} \\ &+ 22q^{26} - 66q^{27} - 25q^{28} + 12q^{29} + 21q^{30} - 52q^{31} + 44q^{32} \\ &+ 96q^{33} + 19q^{34} + 82q^{35} - 105q^{36} - 86q^{37} + 29q^{38} + 108q^{39} \\ &- 148q^{40} - 106q^{41} + 159q^{42} - 260q^{43} - 136q^{44} + 36q^{45} \\ &+ 261q^{46} - 22q^{47} + 309q^{48} + 430q^{49} + O(q^{50}). \end{aligned}$$

i	$g_i(\tau)$	i	$g_i(\tau)$
1	$\eta(\tau)^4\eta(5\tau)^4\eta(11\tau)^4\eta(55\tau)^4$	21	$\eta(\tau)^5\eta(5\tau)^{-1}\eta(11\tau)^3\eta(55\tau)^9$
2	$\eta(\tau)^5\eta(5\tau)^3\eta(11\tau)^5\eta(55\tau)^3$	22	$\eta(\tau)^9\eta(5\tau)^3\eta(11\tau)^5\eta(55\tau)^{-1}$
3	$\eta(\tau)^3\eta(5\tau)^5\eta(11\tau)^3\eta(55\tau)^5$	23	$\eta(\tau)^{-1}\eta(5\tau)^5\eta(11\tau)^3\eta(55\tau)^9$
4	$\eta(\tau)^2\eta(5\tau)^6\eta(11\tau)^2\eta(55\tau)^6$	24	$\eta(\tau)^4\eta(11\tau)^2\eta(55\tau)^{10}$
5	$\eta(\tau)^6\eta(5\tau)^2\eta(11\tau)^6\eta(55\tau)^2$	25	$\eta(\tau)^3\eta(5\tau)\eta(11\tau)\eta(55\tau)^{11}$
6	$\eta(\tau)^7\eta(5\tau)\eta(11\tau)\eta(55\tau)^7$	26	$\eta(\tau)^2\eta(5\tau)^2\eta(55\tau)^{12}$
7	$\eta(\tau)^6\eta(5\tau)^6\eta(11\tau)^2\eta(55\tau)^2$	27	$\eta(\tau)\eta(5\tau)^3\eta(11\tau)^{-1}\eta(55\tau)^{13}$
8	$\eta(\tau)^7\eta(5\tau)\eta(11\tau)^7\eta(55\tau)$	28	$\eta(\tau)^6\eta(5\tau)^{-2}\eta(11\tau)^{-2}\eta(55\tau)^{14}$
9	$\eta(\tau)\eta(5\tau)^7\eta(11\tau)\eta(55\tau)^7$	29	$\eta(5\tau)^4\eta(11\tau)^{-2}\eta(55\tau)^{14}$
10	$\eta(\tau)^8\eta(11\tau)^2\eta(55\tau)^6$	30	$\eta(11\tau)^2\eta(55\tau)^{14}$
11	$\eta(\tau)^7\eta(5\tau)^5\eta(11\tau)^3\eta(55\tau)$	31	$\eta(\tau)\eta(5\tau)^{-1}\eta(11\tau)^{-3}\eta(55\tau)^{19}$
12	$\eta(\tau)\eta(5\tau)^3\eta(11\tau)^5\eta(55\tau)^7$	32	$\eta(\tau)\eta(5\tau)^7\eta(11\tau)^7\eta(55\tau)$
13	$\eta(\tau)^5\eta(5\tau)^7\eta(11\tau)\eta(55\tau)^3$	33	$\eta(\tau)^2\eta(5\tau)^2\eta(11\tau)^6\eta(55\tau)^6$
14	$\eta(5\tau)^8\eta(55\tau)^8$	34	$\eta(\tau)^3\eta(5\tau)\eta(11\tau)^7\eta(55\tau)^5$
15	$\eta(\tau)^8\eta(11\tau)^8$	35	$\eta(\tau)^2\eta(5\tau)^6\eta(11\tau)^8$
16	$\eta(\tau)^9\eta(5\tau)^{-1}\eta(11\tau)^3\eta(55\tau)^5$	36	$\eta(\tau)^6\eta(5\tau)^2\eta(55\tau)^8$
17	$\eta(\tau)^5\eta(5\tau)^3\eta(11\tau)^{-1}\eta(55\tau)^9$	37	$\eta(5\tau)^8\eta(11\tau)^6\eta(55\tau)^2$
18	$\eta(\tau)^{-1}\eta(5\tau)^9\eta(11\tau)^5\eta(55\tau)^3$	38	$\eta(\tau)^4\eta(11\tau)^8\eta(55\tau)^4$
19	$\eta(\tau)^3\eta(5\tau)^5\eta(11\tau)^9\eta(55\tau)^{-1}$	39	$\eta(\tau)^8\eta(5\tau)^4\eta(11\tau)^4$
20	$\eta(\tau)^3\eta(5\tau)^9\eta(11\tau)^5\eta(55\tau)^{-1}$	40	$\eta(5\tau)^4\eta(11\tau)^4\eta(55\tau)^8$

Table 1. Eta-quotient basis of $S_8(\Gamma_0(55))$.

Moreover, we calculate that $\dim_{\mathbb{C}} S_8(\Gamma_0(55)) = 40$, and compute a basis of $S_8(\Gamma_0(55))$ consisting only of eta-quotients $\{g_1, \dots, g_{40}\}$, which are given in Table 1.

From their q-expansions, we calculate that

$$f(\tau)a(\tau) = \sum_{i=1}^{40} c_i g_i(\tau),$$

where the coefficients c_i are given in Table 2.

Thus, we can write

$$f(\tau) = \sum_{i=1}^{40} c_i \frac{g_i(\tau)}{a(\tau)}, \tag{20}$$

where the simplified eta-quotients $g_i(\tau)/a(\tau)$ are given in Table 3. □

We conclude by describing in an algorithmic fashion the method we used to obtain Theorems 1.11 and 1.12. We have seen that both of these theorems, once

i	c_i
1	−60086495559293093894 97/81950623845145992456 2
2	−5025208905035516963 66/409753119225729962281
3	−18079466846617647763574/12292593576771898868 43
4	−9707713817545985330315/16390124769029198491 24
5	−2560946835925829940 17/81950623845145992456 2
6	−801755327323567495694/10782976821729735849 5
7	−5830988018825221370539/12292593576771898868430
8	−779344877836568799883/2458518715354379773686
9	822221079673196383713 5/12292593576771898868 43
10	−11036620216580334273019/24585187153543797736860
11	−12290417121726893676 04/61462967883859494342 15
12	−11156155051409968414091 2/20487655961286498114 05
13	10711441338208896203 6/20487655961286498114 05
14	−3413588205914099736 60/409753119225729962281
15	−1526125951371645395 75/2458518715354379773686
16	0
17	−7213279306787331549849/81950623845145992456 2
18	329520533895881663 94/409753119225729962281
19	305225190274329079 15/2458518715354379773686
20	0
21	−1552252178781785948225/81950623845145992456 2
22	0
23	157425889819966975 24/110744086277224314 13
24	−11628435331878803001 1856/12292593576771898868 43
25	−53256089410452110948 2105/16390124769029198491 24
26	−151470056826256099503 3025/2458518715354379773686
27	−56158633830253942950 0115/81950623845145992456 2
28	−1498106192582562197 80/409753119225729962281
29	−17076657283616798676 2205/81950623845145992456 2
30	−81633970222789942752773/409753119225729962281
31	15156834377209809454 8465/81950623845145992456 2
32	84501754869059292050 2/61462967883859494342 15
33	−34304558248987380189 7249/12292593576771898868430
34	−75023415460898748324 0971/24585187153543797736860
35	−2970842391989560281 46/61462967883859494342 15
36	−30834650074592581387973/20487655961286498114 05
37	−9228870808743396499/10782976821729735849 5
38	−4131906244671465777242/409753119225729962281
39	−305225190274329079 15/2458518715354379773686
40	−13146446749054646251719/81950623845145992456 2

Table 2. Coefficients c_i.

i	$g_i(\tau)/a(\tau)$	i	$g_i(\tau)/a(\tau)$
1	$\eta(\tau)\eta(5\tau)\eta(11\tau)\eta(55\tau)$	21	$\eta(\tau)^2\eta(5\tau)^{-4}\eta(55\tau)^6$
2	$\eta(\tau)^2\eta(11\tau)^2$	22	$\eta(\tau)^6\eta(11\tau)^2\eta(55\tau)^{-4}$
3	$\eta(5\tau)^2\eta(55\tau)^2$	23	$\eta(\tau)^{-4}\eta(5\tau)^2\eta(55\tau)^6$
4	$\eta(\tau)^{-1}\eta(5\tau)^3\eta(11\tau)^{-1}\eta(55\tau)^3$	24	$\eta(\tau)\eta(5\tau)^{-3}\eta(11\tau)^{-1}\eta(55\tau)^7$
5	$\eta(\tau)^3\eta(5\tau)^{-1}\eta(11\tau)^3\eta(55\tau)^{-1}$	25	$\eta(5\tau)^{-2}\eta(11\tau)^{-2}\eta(55\tau)^8$
6	$\eta(\tau)^4\eta(5\tau)^{-2}\eta(11\tau)^{-2}\eta(55\tau)^4$	26	$\eta(\tau)^{-1}\eta(5\tau)^{-1}\eta(11\tau)^{-3}\eta(55\tau)^9$
7	$\eta(\tau)^3\eta(5\tau)^3\eta(11\tau)^{-1}\eta(55\tau)^{-1}$	27	$\eta(\tau)^{-2}\eta(11\tau)^{-4}\eta(55\tau)^{10}$
8	$\eta(\tau)^4\eta(5\tau)^{-2}\eta(11\tau)^4\eta(55\tau)^{-2}$	28	$\eta(\tau)^3\eta(5\tau)^{-5}\eta(11\tau)^{-5}\eta(55\tau)^{11}$
9	$\eta(\tau)^{-2}\eta(5\tau)^4\eta(11\tau)^{-2}\eta(55\tau)^4$	29	$\eta(\tau)^{-3}\eta(5\tau)\eta(11\tau)^{-5}\eta(55\tau)^{11}$
10	$\eta(\tau)^5\eta(5\tau)^{-3}\eta(11\tau)^{-1}\eta(55\tau)^3$	30	$\eta(\tau)^{-3}\eta(5\tau)^{-3}\eta(11\tau)^{-1}\eta(55\tau)^{11}$
11	$\eta(\tau)^4\eta(5\tau)^2\eta(55\tau)^{-2}$	31	$\eta(\tau)^{-2}\eta(5\tau)^{-4}\eta(11\tau)^{-6}\eta(55\tau)^{16}$
12	$\eta(\tau)^{-2}\eta(11\tau)^2\eta(55\tau)^4$	32	$\eta(\tau)^{-2}\eta(5\tau)^4\eta(11\tau)^4\eta(55\tau)^{-2}$
13	$\eta(\tau)^2\eta(5\tau)^4\eta(11\tau)^{-2}$	33	$\eta(\tau)^{-1}\eta(5\tau)^{-1}\eta(11\tau)^3\eta(55\tau)^3$
14	$\eta(\tau)^{-3}\eta(5\tau)^5\eta(11\tau)^{-3}\eta(55\tau)^5$	34	$\eta(5\tau)^{-2}\eta(11\tau)^4\eta(55\tau)^2$
15	$\eta(\tau)^5\eta(5\tau)^{-3}\eta(11\tau)^5\eta(55\tau)^{-3}$	35	$\eta(\tau)^{-1}\eta(5\tau)^3\eta(11\tau)^5\eta(55\tau)^{-3}$
16	$\eta(\tau)^6\eta(5\tau)^{-4}\eta(55\tau)^2$	36	$\eta(\tau)^3\eta(5\tau)^{-1}\eta(11\tau)^{-3}\eta(55\tau)^5$
17	$\eta(\tau)^2\eta(11\tau)^{-4}\eta(55\tau)^6$	37	$\eta(\tau)^{-3}\eta(5\tau)^5\eta(11\tau)^3\eta(55\tau)^{-1}$
18	$\eta(\tau)^{-4}\eta(5\tau)^6\eta(11\tau)^2$	38	$\eta(\tau)\eta(5\tau)^{-3}\eta(11\tau)^5\eta(55\tau)$
19	$\eta(5\tau)^2\eta(11\tau)^6\eta(55\tau)^{-4}$	39	$\eta(\tau)^5\eta(5\tau)\eta(11\tau)\eta(55\tau)^{-3}$
20	$\eta(5\tau)^6\eta(11\tau)^2\eta(55\tau)^{-4}$	40	$\eta(\tau)^{-3}\eta(5\tau)\eta(11\tau)\eta(55\tau)^5$

Table 3. Eta-quotients $g_i(\tau)/a(\tau)$.

discovered, can be proved in a straightforward manner. However, how to find results like these may not be immediately apparent. Our approach, which is inspired by [Pathakjee, RosnBrick and Yoong 2012], utilizes results from Section 3 and has the potential to generate many new examples. We note that many of the steps require the aid of mathematical software to be practical; we used SageMath [2018]. We do not know whether the process we outline will necessarily terminate.

Step 1: Fix a semiprime $N = pq$ satisfying the conditions in Theorem 1.10 with $k = 2$ that is the conductor of an elliptic curve E. By the modularity theorem (Theorem 1.1) we know that E has an associated modular form $f(\tau) \in S_2(\Gamma_0(N))$.

Step 2: Compute $d_N = \dim_{\mathbb{C}} S_2(\Gamma_0(N))$.

Step 3: Compute all partitions of $(p+1)(q+1)/6$ into exactly four parts, and construct distinct rearrangements in order to get a complete list of all possible tuples $(v_1, v_{1/p}, v_{1/q}, v_N) \in \mathbb{N}^4$ satisfying

$$v_1 + v_{1/p} + v_{1/q} + v_{1/N} = \tfrac{1}{6}(p+1)(q+1). \tag{21}$$

By Theorem 3.2, we know that any eta-quotient in $S_2(\Gamma_0(N))$ must have orders of vanishing $v_1, v_{1/p}, v_{1/q}, v_{1/N} \geq 1$ that satisfy (21).

Step 4: For each tuple $(v_1, v_{1/p}, v_{1/q}, v_N)$ obtained in Step 3, use Theorem 1.4 to construct the system of four equations in the four unknowns (r_1, r_p, r_q, r_N)

$$\begin{aligned}24v_1 &= Nr_1 + qr_p + pr_q + r_N,\\ 24v_{1/p} &= qr_1 + Nr_p + r_q + pr_N,\\ 24v_{1/q} &= pr_1 + r_p + Nr_q + qr_N,\\ 24v_{1/N} &= r_1 + pr_p + qr_q + Nr_N,\end{aligned}$$

and solve for the unique solution $(r_1, r_p, r_q, r_N) \in \mathbb{Q}^4$.

Step 5: For each tuple (r_1, r_p, r_q, r_N) from Step 4 that has integer entries, let

$$g(\tau) = \eta(\tau)^{r_1}\eta(p\tau)^{r_p}\eta(q\tau)^{r_q}\eta(N\tau)^{r_N},$$

and use Theorems 1.2 and 1.4 to check whether $g(\tau) \in S_2(\Gamma_0(N))$. List all such $g \in S_2(\Gamma_0(N))$.

Step 6: Construct a maximally sized linearly independent set of eta-quotients from the list in Step 5, using linear algebra.

Step 7: If the set from Step 6 has size d_N, then it forms a basis of $S_2(\Gamma_0(N))$. In this case, compute the Sturm bound from Theorem 1.5 and write f as a linear combination of the basis from Step 6. If not, go to Step 8.

Step 8: Repeat Steps 2–6 for weights $2, 4, 6, \ldots$ until a weight k is found such that $S_k(\Gamma_0(N))$ has a basis of eta-quotients, and $S_{k-2}(\Gamma_0(N))$ contains an eta-quotient $a(\tau)$. Compute the Sturm bound from Theorem 1.5 and write $f(\tau)a(\tau)$ as a linear combination of the basis. Divide through by $a(\tau)$ to write $f(\tau)$ as a linear combination of eta-quotients in $M_2^!(\Gamma_0(N))$.

References

[Ahlgren and Ono 2001] S. Ahlgren and K. Ono, "Congruence properties for the partition function", *Proc. Natl. Acad. Sci. USA* **98**:23 (2001), 12882–12884. MR Zbl

[Arnold-Roksandich, James and Keaton 2018a] A. Arnold-Roksandich, K. James, and R. Keaton, "Counting eta-quotients of prime level", *Involve* **11**:5 (2018), 827–844. MR Zbl

[Arnold-Roksandich, James and Keaton 2018b] A. Arnold-Roksandich, K. James, and R. Keaton, "Proceedings paper for REU project involving counting eta-quotients", preprint, 2018. arXiv

[Biagioli 1990] A. J. F. Biagioli, "The construction of modular forms as products of transforms of the Dedekind eta function", *Acta Arith.* **54**:4 (1990), 273–300. MR Zbl

[Borcherds 1992] R. E. Borcherds, "Monstrous moonshine and monstrous Lie superalgebras", *Invent. Math.* **109**:2 (1992), 405–444. MR Zbl

[Conway and Norton 1979] J. H. Conway and S. P. Norton, "Monstrous moonshine", *Bull. London Math. Soc.* **11**:3 (1979), 308–339. MR Zbl

[Diamond and Shurman 2005] F. Diamond and J. Shurman, *A first course in modular forms*, Graduate Texts in Mathematics **228**, Springer, 2005. MR Zbl

[Dummit, Kisilevsky and McKay 1985] D. Dummit, H. Kisilevsky, and J. McKay, "Multiplicative products of η-functions", pp. 89–98 in *Finite groups: coming of age* (Montreal, 1982), edited by J. McKay, Contemp. Math. **45**, Amer. Math. Soc., Providence, RI, 1985. MR Zbl

[Gordon and Hughes 1993] B. Gordon and K. Hughes, "Multiplicative properties of η-products, II", pp. 415–430 in *A tribute to Emil Grosswald: number theory and related analysis*, edited by M. Knopp and M. Sheingorn, Contemp. Math. **143**, Amer. Math. Soc., Providence, RI, 1993. MR Zbl

[Kilford 2007] L. J. P. Kilford, "Generating spaces of modular forms with η-quotients", *JP J. Algebra Number Theory Appl.* **8**:2 (2007), 213–226. MR Zbl

[Lemke Oliver 2013] R. J. Lemke Oliver, "Eta-quotients and theta functions", *Adv. Math.* **241** (2013), 1–17. MR Zbl

[Ligozat 1974] G. Ligozat, *Courbes modulaires de genre* 1, Ph.D. thesis, Université Paris XI, 1974, available at http://sites.mathdoc.fr/PMO/PDF/L_LIGOZAT-78.pdf. MR

[LMFDB 2019] The LMFDB Collaboration, "The L-functions and modular forms database", electronic reference, 2019, available at http://www.lmfdb.org. Online; accessed 3 August 2018.

[Martin 1996] Y. Martin, "Multiplicative η-quotients", *Trans. Amer. Math. Soc.* **348**:12 (1996), 4825–4856. MR Zbl

[Martin and Ono 1997] Y. Martin and K. Ono, "Eta-quotients and elliptic curves", *Proc. Amer. Math. Soc.* **125**:11 (1997), 3169–3176. MR Zbl

[Murty, Dewar and Graves 2015] M. R. Murty, M. Dewar, and H. Graves, *Problems in the theory of modular forms*, Institute of Mathematical Sciences Lecture Notes **1**, Hindustan Book Agency, New Delhi, 2015. MR Zbl

[Newman 1957] M. Newman, "Construction and application of a class of modular functions", *Proc. London Math. Soc.* (3) **7** (1957), 334–350. MR Zbl

[Newman 1959] M. Newman, "Construction and application of a class of modular functions, II", *Proc. London Math. Soc.* (3) **9** (1959), 373–387. MR Zbl

[Ono 2000] K. Ono, "Distribution of the partition function modulo m", *Ann. of Math.* (2) **151**:1 (2000), 293–307. MR Zbl

[Ono 2004] K. Ono, *The web of modularity: arithmetic of the coefficients of modular forms and q-series*, CBMS Regional Conference Series in Mathematics **102**, Amer. Math. Soc., Providence, RI, 2004. MR Zbl

[Pathakjee, RosnBrick and Yoong 2012] D. Pathakjee, Z. RosnBrick, and E. Yoong, "Elliptic curves, eta-quotients and hypergeometric functions", *Involve* **5**:1 (2012), 1–8. MR Zbl

[Rouse and Webb 2015] J. Rouse and J. J. Webb, "On spaces of modular forms spanned by eta-quotients", *Adv. Math.* **272** (2015), 200–224. MR Zbl

[SageMath 2018] The Sage Developers, *SageMath, the Sage Mathematics Software System*, 2018, available at http://www.sagemath.org. Version 8.3.

[Taylor and Wiles 1995] R. Taylor and A. Wiles, "Ring-theoretic properties of certain Hecke algebras", *Ann. of Math.* (2) **141**:3 (1995), 553–572. MR Zbl

[Wiles 1995] A. Wiles, "Modular elliptic curves and Fermat's last theorem", *Ann. of Math.* (2) **141**:3 (1995), 443–551. MR Zbl

Received: 2020-08-04 Accepted: 2020-08-11

allenm3@oregonstate.edu *Oregon State University, Corvallis, OR, United States*

n.anderson@se19.qmul.ac.uk *Queen Mary University of London, London, United Kingdom*

asimina.hamakiotes@uconn.edu *University of Connecticut, Storrs, CT, United States*

benjamin.oltsik@uconn.edu *University of Connecticut, Storrs, CT, United States*

swisherh@math.oregonstate.edu *Department of Mathematics, Oregon State University, Corvallis, OR, United States*

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.

involve

2020 vol. 13 no. 5

1944-4176(2020)13:5;1-P