

involve

a journal of mathematics

Editorial Board

Kenneth S. Berenhaut, *Managing Editor*

| | |
|----------------------|------------------------|
| Colin Adams | Suzanne Lenhart |
| John V. Baxley | Chi-Kwong Li |
| Arthur T. Benjamin | Robert B. Lund |
| Martin Bohner | Gaven J. Martin |
| Nigel Boston | Mary Meyer |
| Amarjit S. Budhiraja | Emil Minchev |
| Pietro Cerone | Frank Morgan |
| Scott Chapman | Mohammad Sal Moslehian |
| Jem N. Corcoran | Zuhair Nashed |
| Toka Diagana | Ken Ono |
| Michael Dorff | Timothy E. O'Brien |
| Sever S. Dragomir | Joseph O'Rourke |
| Behrouz Emamizadeh | Yuval Peres |
| Joel Foisy | Y.-F. S. Pétermann |
| Errin W. Fulp | Robert J. Plemmons |
| Joseph Gallian | Carl B. Pomerance |
| Stephan R. Garcia | Bjorn Poonen |
| Anant Godbole | James Propp |
| Ron Gould | József H. Przytycki |
| Andrew Granville | Richard Rebarber |
| Jerrold Griggs | Robert W. Robinson |
| Sat Gupta | Filip Saidak |
| Jim Haglund | James A. Sellers |
| Johnny Henderson | Andrew J. Sterge |
| Jim Hoste | Ann Trenk |
| Natalia Hritonenko | Ravi Vakil |
| Glenn H. Hurlbert | Antonia Vecchio |
| Charles R. Johnson | Ram U. Verma |
| K. B. Kulasekera | John C. Wierman |
| Gerry Ladas | Michael E. Zieve |
| David Larson | |



involve

msp.org/involve

EDITORS

MANAGING EDITOR

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

BOARD OF EDITORS

| | | | |
|----------------------|---|-------------------------|---|
| Colin Adams | Williams College, USA colin.c.adams@williams.edu | David Larson | Texas A&M University, USA larson@math.tamu.edu |
| John V. Baxley | Wake Forest University, NC, USA baxley@wfu.edu | Suzanne Lenhart | University of Tennessee, USA lenhart@math.utk.edu |
| Arthur T. Benjamin | Harvey Mudd College, USA benjamin@hmc.edu | Chi-Kwong Li | College of William and Mary, USA ckli@math.wm.edu |
| Martin Bohner | Missouri U of Science and Technology, USA bohner@mst.edu | Robert B. Lund | Clemson University, USA lund@clemson.edu |
| Nigel Boston | University of Wisconsin, USA boston@math.wisc.edu | Gaven J. Martin | Massey University, New Zealand g.j.martin@massey.ac.nz |
| Amarjit S. Budhiraja | U of North Carolina, Chapel Hill, USA budhiraj@email.unc.edu | Mary Meyer | Colorado State University, USA meyer@stat.colostate.edu |
| Pietro Cerone | La Trobe University, Australia P.Cerone@latrobe.edu.au | Emil Minchev | Ruse, Bulgaria eminchev@hotmail.com |
| Scott Chapman | Sam Houston State University, USA scott.chapman@shsu.edu | Frank Morgan | Williams College, USA frank.morgan@williams.edu |
| Joshua N. Cooper | University of South Carolina, USA cooper@math.sc.edu | Mohammad Sal Moselehian | Ferdowsi University of Mashhad, Iran moslehian@ferdowsi.um.ac.ir |
| Jem N. Corcoran | University of Colorado, USA corcoran@colorado.edu | Zuhair Nashed | University of Central Florida, USA znashed@mail.ucf.edu |
| Toka Diagana | Howard University, USA tdiagana@howard.edu | Ken Ono | Emory University, USA ono@mathcs.emory.edu |
| Michael Dorff | Brigham Young University, USA mdorff@math.byu.edu | Timothy E. O'Brien | Loyola University Chicago, USA tobrie1@luc.edu |
| Sever S. Dragomir | Victoria University, Australia sever@matilda.vu.edu.au | Joseph O'Rourke | Smith College, USA orourke@cs.smith.edu |
| Behrouz Emamizadeh | The Petroleum Institute, UAE bemamizadeh@pi.ac.ae | Yuval Peres | Microsoft Research, USA peres@microsoft.com |
| Joel Foisy | SUNY Potsdam foisyjs@potsteam.edu | Y.-F. S. Pétermann | Université de Genève, Switzerland petermann@math.unige.ch |
| Errin W. Fulp | Wake Forest University, USA fulp@wfu.edu | Robert J. Plemmons | Wake Forest University, USA rplemmons@wfu.edu |
| Joseph Gallian | University of Minnesota Duluth, USA jgallian@d.umn.edu | Carl B. Pomerance | Dartmouth College, USA carl.pomerance@dartmouth.edu |
| Stephan R. Garcia | Pomona College, USA stephan.garcia@pomona.edu | Vadim Ponomarenko | San Diego State University, USA vadim@sciences.sdsu.edu |
| Anant Godbole | East Tennessee State University, USA godbole@etsu.edu | Bjorn Poonen | UC Berkeley, USA poonen@math.berkeley.edu |
| Ron Gould | Emory University, USA rg@mathcs.emory.edu | James Propp | U Mass Lowell, USA jpropp@cs.uml.edu |
| Andrew Granville | Université Montréal, Canada andrew@dms.umontreal.ca | József H. Przytycki | George Washington University, USA przytyck@gwu.edu |
| Jerrold Griggs | University of South Carolina, USA griggs@math.sc.edu | Richard Rebarber | University of Nebraska, USA rrebarbe@math.unl.edu |
| Sat Gupta | U of North Carolina, Greensboro, USA sngupta@uncg.edu | Robert W. Robinson | University of Georgia, USA rwr@cs.uga.edu |
| Jim Haglund | University of Pennsylvania, USA jhaglund@math.upenn.edu | Filip Saidak | U of North Carolina, Greensboro, USA f_saidak@uncg.edu |
| Johnny Henderson | Baylor University, USA johnny_henderson@baylor.edu | James A. Sellers | Penn State University, USA sellersj@math.psu.edu |
| Jim Hoste | Pitzer College jhoste@pitzer.edu | Andrew J. Sterge | Honorary Editor andy@ajsterge.com |
| Natalia Hritonenko | Prairie View A&M University, USA nahritonenko@pvamu.edu | Ann Trenk | Wellesley College, USA atrenk@wellesley.edu |
| Glenn H. Hurlbert | Arizona State University, USA hurlbert@asu.edu | Ravi Vakil | Stanford University, USA vakil@math.stanford.edu |
| Charles R. Johnson | College of William and Mary, USA crjohnso@math.wm.edu | Antonia Vecchio | Consiglio Nazionale delle Ricerche, Italy antonia.vecchio@cnr.it |
| K. B. Kulasekera | Clemson University, USA kk@ces.clemson.edu | Ram U. Verma | University of Toledo, USA verma99@msn.com |
| Gerry Ladas | University of Rhode Island, USA gladas@math.uri.edu | John C. Wierman | Johns Hopkins University, USA wierman@jhu.edu |
| | | Michael E. Zieve | University of Michigan, USA zieve@umich.edu |

PRODUCTION


Silvio Levy, Scientific Editor

See inside back cover or msp.org/involve for submission instructions. The subscription price for 2015 is US \$140/year for the electronic version, and \$190/year (+\$35, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues from the last three years and changes of subscribers address should be sent to MSP.

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

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

PUBLISHED BY

 **mathematical sciences publishers**
nonprofit scientific publishing

<http://msp.org/>

© 2015 Mathematical Sciences Publishers

Efficient realization of nonzero spectra by polynomial matrices

Nathan McNew and Nicholas Ormes

(Communicated by Kenneth S. Berenhaut)

A theorem of Boyle and Handelmann gives necessary and sufficient conditions for an n -tuple of nonzero complex numbers to be the nonzero spectrum of some matrix with nonnegative entries, but is not constructive and puts no bound on the necessary dimension of the matrix. Working with polynomial matrices, we constructively reprove this theorem in a special case, with a bound on the size of the polynomial matrix required to realize a given polynomial.

1. Introduction

H. R. Suleĭmanova [1949] posed a question: Given an n -tuple of complex numbers $\sigma := (\lambda_1, \lambda_2, \dots, \lambda_n)$, when is there an $n \times n$ matrix A with nonnegative entries such that $\det(I - At) = \prod_{i=1}^n (t - \lambda_i)$? This problem has come to be known as the nonnegative inverse eigenvalue problem, or NIEP. (See [Eggleston et al. 2004] for a survey article on the problem.) Although there have been some significant advances, the general NIEP as stated remains open. One major advance was proven by Boyle and Handelmann [1991]. They characterized the n -tuples that could be appended with zeros and subsequently realized as the eigenvalues of a nonnegative matrix in the above sense. Their proof relied heavily on results from symbolic dynamics and was not constructive (see [Lind and Marcus 1995] for more on symbolic dynamics and the NIEP). Very recently, Laffey [2012] proved a version of their result by constructive means, although his result is not in quite as general a setting as Boyle and Handelmann's. In this paper we provide a construction different from that of Laffey's. The result of our construction is a matrix with polynomial entries, as opposed to real entries, and we describe a simple way to construct a matrix over the reals based on the polynomial matrix. This construction makes use of weighted directed graphs and is described further in [Boyle 1993].

MSC2010: 15A18, 15B48, 05C50.

Keywords: nonnegative matrices, eigenvalues, power series, nonnegative inverse eigenvalue problem.

2. Preliminaries

A nonnegative matrix is *primitive* if it is a square matrix and some power of it is a matrix with strictly positive entries. The nonnegative inverse eigenvalue problem is generally studied in terms of primitive matrices, since given conditions for an n -tuple to be realized by a primitive matrix, one can easily extend to the general nonnegative case; for example, see [Boyle and Handelman 1991; Friedland 2012].

There are several known necessary conditions for an n -tuple of complex numbers σ to be realizable by a primitive matrix:

- (1) $\sigma = \bar{\sigma}$. (For every complex number in σ , its complex conjugate is also in σ .)
- (2) There exists $\lambda_i \in \sigma$ such that $\lambda_i \in \mathbb{R}_+$ and $\lambda_i > |\lambda_j|$ for $j \neq i$.
- (3) For all $k \in \mathbb{N}$, the k -th moment of σ , $s_k = \sum_{i=1}^n \lambda_i^k$, is nonnegative. Moreover, for all $k \in \mathbb{N}$, if $s_k > 0$, then for all $n \in \mathbb{N}$, $s_{nk} > 0$.

The first condition simply reflects the fact that for the polynomial $\prod_{i=1}^n (t - \lambda_i)$ to have real coefficients, any complex roots must come in conjugate pairs. As a result of the first condition, the NIEP can be reformulated as follows: Given a polynomial $p(t) \in \mathbb{R}[t]$, is there a nonnegative matrix A such that $p(t)$ is the characteristic polynomial of A (i.e., $p(t) = \det(It - A) = \prod_{i=1}^n (t - \lambda_i)$)? In this case, σ is the list of the roots of the polynomial with multiplicity.

The second comes as a result of the Perron–Frobenius theorem (e.g., see [Berman and Plemmons 1979; Minc 1988]). One of the consequences of this theorem is that a primitive matrix A must have a positive real eigenvalue that exceeds the modulus of all other eigenvalues. This positive real eigenvalue is often referred to as the *Perron eigenvalue* of the primitive matrix A .

The third condition is found by observing that if $\det(It - A) = \prod_{i=1}^n (t - \lambda_i)$ then the trace of A^k is $s_k = \sum_{i=1}^n \lambda_i^k$ for all $k \in \mathbb{N}$. Thus if A is nonnegative, so too must be s_k , and if A^k has a positive trace, then A^{nk} does as well for all $n \in \mathbb{N}$.

Boyle and Handelman [1991] proved that the above necessary conditions are sufficient to find a natural number N such that σ can be augmented by N zeros and then realized by a nonnegative primitive matrix. Restating more precisely, they proved the following, which we'll hereafter refer to as the Boyle–Handelman theorem:

Theorem 2.1 (spectral theorem). *Let $\sigma = (\lambda_1, \lambda_2, \dots, \lambda_n) \in \mathbb{C}^n$. There is an $N \geq 0$ and a nonnegative primitive matrix A such that*

$$\det(It - A) = t^N \prod_{i=1}^n (t - \lambda_i)$$

if and only if:

- (1) $\sigma = \bar{\sigma}$.
- (2) There exists $\lambda_i \in \sigma$ such that $\lambda_i \in \mathbb{R}_+$ and $\lambda_i > |\lambda_j|$ for $j \neq i$.
- (3) For all $k \in \mathbb{N}$, the k -th moment of σ , $s_k = \sum_{i=1}^n \lambda_i^k$, is nonnegative, and if $s_k > 0$ then $s_{nk} > 0$ for all $n \in \mathbb{N}$.

Observe that there is an $N \geq 0$ such that $\det(It - A) = t^N \prod_{i=1}^n (t - \lambda_i)$ if and only if $\det(I - At) = \prod_{i=1}^n (1 - \lambda_i t)$, and this will provide us a convenient way to reformulate the theorem. With this, the Boyle–Handelman theorem characterizes which polynomials $q(t) \in \mathbb{R}[t]$ can be the *reverse characteristic polynomial* $\det(I - At)$ for a nonnegative primitive matrix A over the reals.

3. Graphs and polynomial matrices

Let G be a weighted directed graph on N vertices with weights in \mathbb{R}_+ . Then the *adjacency matrix* A of G is the $N \times N$ matrix in which the (i, j) element is the weight of the edge running from vertex i to vertex j . The *characteristic polynomial* of this matrix (and of the associated graph G) is the polynomial $\chi_A(t) = \det(It - A)$. The *reverse characteristic polynomial* of the graph is the polynomial $\chi_A^{-1}(t) = \det(I - At)$. Of course, the process is reversible. Given a matrix A over \mathbb{R}_+ , one can easily construct a weighted directed graph G which has A as its adjacency matrix. One simply includes an edge with weight $A(i, j)$ between each pair of vertices i and j .

As we will show, a directed graph G can also be represented by a polynomial matrix $M(t)$ over $t\mathbb{R}_+[t]$, i.e., a matrix $M(t)$ whose entries are polynomials with nonnegative coefficients without constant terms besides 0. This generally allows for presentations of adjacency matrices of smaller size. This process of constructing a polynomial matrix from a graph is also reversible. We begin by describing the reverse process.

Given an $N \times N$ polynomial matrix $M(t)$ over $t\mathbb{R}_+[t]$, the construction of the corresponding weighted digraph G can be carried out as follows. Assign N “primary” vertices with labels $1, 2, \dots, N$. Then for each term $c_n t^p$ in the polynomial in the (i, j) position of $A(t)$, construct a path of length p from vertex i to vertex j in which the first edge is weighted c_n and each additional edge (if $p > 1$) is weighted 1. If $p > 1$ then the $p - 1$ additional “secondary” vertices in the new path are disjoint from the original N primary vertices and from secondary vertices used in any other path.

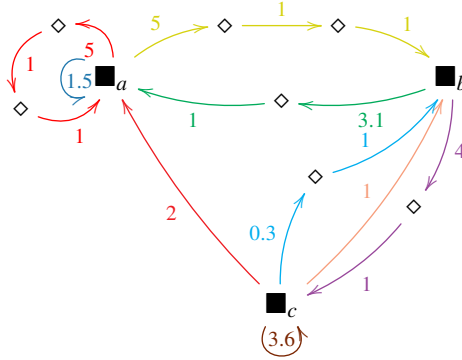
Example 3.1. Take, for example, the matrix over $t\mathbb{R}_+[t]$ given by

$$M(t) = \begin{bmatrix} 5t^3 + 1.5t & 9t^3 & 0 \\ 3.1t^2 & 0 & 4t^2 \\ 2t & 0.3t^2 + t & 3.6t \end{bmatrix}.$$

Note that

$$\det(I - M(t)) = 6t^7 + 48.44t^6 - 29.7t^5 + 22.8t^4 - 9t^3 + 5.4t^2 - 5.1t + 1.$$

From this matrix, by the method described above, we can construct the graph G below (large squares denote the primary vertices, diamonds denote secondary connecting vertices.)



Having constructed this graph, we can now construct the adjacency matrix for the graph A_G , ignoring color and the difference between primary and secondary vertices.

Suppose a directed graph G has m vertices numbered $1, 2, \dots, m$. Then we can define the $m \times m$ matrix A_G where $A_G(i, j)$ is the weight of the edges from vertex i to j .

Example 3.2. Continuing with the above example, we obtain the adjacency matrix A_G for the graph G :

$$A_G = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 5 & 1.5 & 9 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 & 3.6 & 0 & 0.3 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3.1 & 0 & 0 & 4 & 0 \end{bmatrix}.$$

We compute both the characteristic and reverse characteristic polynomials of this adjacency matrix and graph, and obtain

$$\det(It - A_G) = t^{10} - 5.1t^9 + 5.4t^8 - 9t^7 + 22.8t^6 - 29.7t^5 + 48.44t^4 + 6t^3,$$

$$\det(I - A_G t) = 6t^7 + 48.44t^6 - 29.7t^5 + 22.8t^4 - 9t^3 + 5.4t^2 - 5.1t + 1.$$

We note that for the above example,

$$\det(I - M(t)) = \det(I - A_G t),$$

and this is no coincidence.

Theorem 3.3. *Let $M(t)$ be a matrix over $t\mathbb{R}_+[t]$, and suppose G is the directed graph constructed from $M(t)$ by the aforementioned construction. Suppose A_G is the adjacency matrix for G . Then*

$$\det(I - A_G t) = \det(I - M(t)).$$

Proof. Fix i, j such that $M(t)_{i,j}$ is a polynomial of degree greater than 1. Then for each term $c_{i,j,n}t^n$, where $n > 1$ and $c_{i,j,n} > 0$, there is a path in the graph from vertex i to vertex j of length n , and thus $n + 1$ rows (indexed k_1, k_2, \dots, k_{n+1}) in the matrix A_G corresponding to each of the $n + 1$ vertices along this path. (Note that k_1 corresponds to primary vertex i and k_{n+1} corresponds to primary vertex j) Each of these rows and columns (except k_1 and k_{n+1}) will have only one nonzero term, in the (k_h, k_{h+1}) position, and $c_{i,j,n} = \prod_{h=1}^n (A_G)_{k_h, k_{h+1}}$.

Each of these additional $n - 1$ rows can be removed from the matrix $I - A_G t$ without changing the determinant by the following row operations, working backwards from $h = n$ to 2:

- (1) From row k_{h-1} , subtract row k_h scaled by the entry in position (k_{h-1}, k_h) .
- (2) From column k_{n+1} , subtract column k_h multiplied by the entry in position (k_h, k_{n+1}) .

This sequence results in the product of the terms in positions (k_{h-1}, k_h) and (k_h, k_{n+1}) appearing in position (k_{h-1}, k_{n+1}) and only a 1 remaining in both row and column k_h . Thus, after repeating this process for all the intermediate vertices, there will be a term equivalent to the product of their weights times t raised to the length of the chain added to the (k_1, k_{n+1}) position and a 1 in the primary diagonal for each row/column associated with each intermediate vertex. The determinant can be expanded by minors at each of these 1s, thus reducing the size of the matrix.

Repeating this process for each such $c_{i,j,n}$ term in $I - M(t)$ (and switching rows as necessary at the end) will produce the matrix $I - M(t)$ from $I - A_G t$ without changing the determinant. \square

The process of constructing a polynomial matrix $M(t)$ from a weighted directed graph G can be done by simply letting the (i, j) entry of $M(t)$ be $w(i, j)t$, where $w(i, j)$ is the sum of the weights of the edges from i to j . An alternative approach, which could be more efficient in terms of the size of $M(t)$, would be to identify secondary vertices as those which have at most one edge coming in and one out. Then the coefficient of t^k in the (i, j) entry of $M(t)$ is the sum of the weights of the paths of length k from primary vertex i to primary vertex j .

4. Our approach

Our approach is to study the nonnegative inverse eigenvalue problem, and specifically the Boyle–Handelman theorem, in terms of polynomial matrices rather than matrices over \mathbb{R}_+ . We attempt to reprove the Boyle–Handelman theorem in certain cases by constructing an “efficient” polynomial matrix (in terms of the size of the matrix, without any bound on the degree of polynomials used in that matrix) that realizes a given polynomial. If we were able to bound both the size of the matrix and the degree of the polynomials used then we would be able to bound the size of the corresponding matrix over \mathbb{R}_+ . In this vein, we will make use of polynomials which are truncations of the power series for $p(t)^{1/N}$.

We proceed forward assuming that $p(t)$ is a polynomial over \mathbb{R} of the form $p(t) = \prod_{i=1}^n (1 - \lambda_i t)$, where $\sigma = (\lambda_1, \lambda_2, \dots, \lambda_n)$ satisfies the conditions of the Boyle–Handelman theorem with a strengthened version of the third condition: for all $k \in \mathbb{N}$, $s_k > 0$. We will say that $p(t)$, or perhaps σ , satisfying these conditions satisfies BH+. Below we prove that in such a case, the power series expansion for $p(t)^{1/N}$ has nonpositive coefficients after the constant term. More recently, this result was proven using different means by Laffey, Loewy and Šmigoc [Laffey et al. 2013].

Theorem 4.1. *Assume that $p(t) = \prod_{i=1}^d (1 - \lambda_i t)$ satisfies BH+. Then there is an $N \geq 1$ such that the power series expansion for $p(t)^{1/N}$ is of the form*

$$p(t)^{1/N} = 1 - \sum_{k=1}^{\infty} r_k t^k,$$

where $r_k \geq 0$ for all $k \geq 1$.

Proof. Recall that the power series expansion for $(1 - t)^{1/N}$ is given by

$$(1 - t)^{1/N} = \sum_{k=0}^{\infty} \binom{1/N}{k} t^k,$$

where $\binom{1/N}{k}$ is a generalized binomial coefficient, given by

$$\binom{1/N}{k} = \frac{1/N(1/N - 1)(1/N - 2) \cdots (1/N - k + 1)}{k(k - 1)(k - 2) \cdots 1}.$$

Then

$$\begin{aligned} p(t)^{1/N} &= \prod_{i=1}^d (1 - \lambda_i t)^{1/N} \\ &= \prod_{i=1}^d \left(\sum_{k=0}^{\infty} \binom{1/N}{k} (-\lambda_i)^k t^k \right) = \prod_{i=1}^d \left(1 - \sum_{k=1}^{\infty} \binom{1/N}{k} |\lambda_i^k t^k| \right). \end{aligned}$$

The k -th coefficient of this series is given by

$$r_k = \left| \binom{1/N}{k} (\lambda_1^k + \lambda_2^k + \dots + \lambda_d^k) + \sum (-1)^l \left| \binom{1/N}{k_1} \dots \binom{1/N}{k_d} \right| \lambda_1^{k_1} \lambda_2^{k_2} \dots \lambda_d^{k_d} \right|,$$

where the second sum ranges over all combinations of nonnegative k_i such that $k_1 + k_2 + \dots + k_d = k$, where $l \geq 2$ is the number of nonzero k_i , and where $k_{i_1}, k_{i_2}, \dots, k_{i_l}$ are these nonzero values.

Factoring $\left| \binom{1/N}{k} \right| \lambda_1^k$ out of this expression (and assuming that λ_1 is the Perron eigenvalue), the first term above becomes

$$1 + \left(\frac{\lambda_2}{\lambda_1} \right)^k + \dots + \left(\frac{\lambda_d}{\lambda_1} \right)^k,$$

which approaches 1 as $k \rightarrow \infty$ and is always positive (by BH+). Therefore, this term has a uniform lower bound $\delta > 0$ (which does not depend on k or N).

The absolute value of the second term is at most

$$\sum \left| \frac{\binom{1/N}{k_1} \binom{1/N}{k_2} \dots \binom{1/N}{k_d}}{\binom{1/N}{k}} \right| \left| \frac{\lambda_1}{\lambda_1} \right|^{k_1} \left| \frac{\lambda_2}{\lambda_1} \right|^{k_2} \dots \left| \frac{\lambda_d}{\lambda_1} \right|^{k_d}.$$

Now observe that for $l \geq 2$ and $N \geq 2$,

$$\begin{aligned} & \left| \frac{\binom{1/N}{k_1} \binom{1/N}{k_2} \dots \binom{1/N}{k_d}}{\binom{1/N}{k}} \right| \\ &= \left| \frac{\frac{1}{N}(\frac{1}{N}-1)\dots(\frac{1}{N}-k_1+1)}{k_1!} \frac{1}{N}(\frac{1}{N}-1)\dots(\frac{1}{N}-k_2+1)}{k_2!} \dots \frac{1}{N}(\frac{1}{N}-1)\dots(\frac{1}{N}-k_d+1)}{k_d!}}{\frac{1}{N}(\frac{1}{N}-1)\dots(\frac{1}{N}-k+1)}{k!}} \right| \\ &= \left| \left(\frac{1}{N} \right)^{l-1} \frac{k!}{k_1! k_2! \dots k_d!} \frac{\left((\frac{1}{N}-1) \dots (\frac{1}{N}-k_1+1) \right) \dots \left((\frac{1}{N}-1) \dots (\frac{1}{N}-k_d+1) \right)}{\left(\frac{1}{N}-1 \right) \dots \left(\frac{1}{N}-k+1 \right)} \right| \\ &< \left| \left(\frac{1}{N} \right)^{l-1} \frac{k!}{k_1! k_2! \dots k_d!} \frac{(k_{i_1}-1)! (k_{i_2}-1)! \dots (k_{i_l}-1)!}{(k-1)!} \right| \\ &= \left| \left(\frac{1}{N} \right)^{l-1} \frac{k}{k_{i_1} k_{i_2} \dots k_{i_l}} \right| \\ &= \left| \frac{1}{N} \frac{k}{k_{i_1} k_{i_2} \dots k_{i_l} N^{l-2}} \right|. \end{aligned}$$

Since $k_{i_1} k_{i_2} \dots k_{i_l} N^{l-2}$ is minimized when $l = 2$ and $k_{i_1} = k - 1$, we have

$$\left| \frac{1}{N} \frac{k}{k_{i_1} k_{i_2} \dots k_{i_l} N^{l-2}} \right| < \left| \frac{1}{N} \frac{k}{k-1} \right| < \frac{2}{N}.$$

Also note that

$$\sum \left| \frac{\lambda_1}{\lambda_1} \right|^{k_1} \left| \frac{\lambda_2}{\lambda_1} \right|^{k_2} \cdots \left| \frac{\lambda_d}{\lambda_1} \right|^{k_d} < \frac{1}{1 - \left| \frac{\lambda_2}{\lambda_1} \right|} \frac{1}{1 - \left| \frac{\lambda_3}{\lambda_1} \right|} \cdots \frac{1}{1 - \left| \frac{\lambda_d}{\lambda_1} \right|} = M,$$

by expanding the right-hand side into a product of geometric series. Therefore, there is a uniform upper bound of the form $(2/N)M$, where M does not depend on k or N .

Then all we need to do is choose N such that $\delta > (2/N)M$. \square

Using this result, we pose the following question:

Question 4.2. Let $p(t)$ be a polynomial which satisfies the condition that there exists $N \geq 1$ such that $p(t)^{1/N} = 1 - \sum_{k=1}^{\infty} r_k t^k$, where $r_k \geq 0$ for all $k \geq 1$. Then does there exist an $N \times N$ polynomial matrix $M(t)$ with nonnegative coefficients such that $\det(I - M(t)) = p(t)$?

As a result of Theorems 3.3 and 4.1, answering in the affirmative would be (nearly) equivalent to proving the Boyle–Handelman theorem (with the exception of the strengthening of the third condition in Theorem 4.1.) Such an answer would further give a constructive proof and would have a bound on the size of the polynomial matrix required to realize a given polynomial. Without putting a bound on the degree of the polynomial matrix, however, this conjecture does not establish any bounds on the size of the regular matrix over \mathbb{R}_+ . If, however, the size of the polynomial matrix and the degrees of polynomials used in the matrix could both be bounded, then a bound on the size of the realizing regular matrix could be achieved.

At the moment we are able to prove the above conjecture for the cases $N = 1, 2, 3$.

5. Cases $N = 1, 2$

Case $N = 1$. The case where $N = 1$ is trivial. If $p(t)^1 = 1 - r(t)$, where $r(t)$ has no negative coefficients, then the matrix $A(t) = [r(t)]$ suffices, and

$$\det(I - A(t)) = \det([1 - r(t)]) = 1 - r(t) = p(t).$$

Case $N = 2$. Suppose $p(t)^{1/2} = 1 - r(t)$, where $r(t)$ has no negative coefficients. Then let $q(t)$ be the polynomial that results when the power series $r(t)$ is truncated to a degree- n polynomial, where n is greater than or equal to the degree of $p(t)$. Consider the polynomial $(1 - q(t))^2$.

The first n terms of this polynomial will sum to $p(t)$. Let $R(t) = (1 - q(t))^2 - p(t)$. Then $R(t)$ will be a polynomial with lowest-order term of degree $n + 1$ and highest degree of $2n$, and is described by

$$R(t) = \sum_{i=n+1}^{2n} \sum_{j+k=i} q_j q_k t^i,$$

where q_i is the coefficient of the t^i term in $q(t)$. Since all the q_i are nonnegative, $R(t)$ will contain only nonnegative terms.

Then construct the matrix

$$A(t) = \begin{bmatrix} q(t) & R(t)/t \\ t & q(t) \end{bmatrix};$$

we have

$$\det(I - A(t)) = (1 - q(t))^2 - R(t) = p(t).$$

Example 5.1 ($N = 2$). Consider the polynomial $p(t) = 1 - 3t - 2t^2 + 4t^3$. The power series of $p(t)^{1/2}$ is

$$p(t)^{1/2} = 1 - \frac{3t}{2} - \frac{17t^2}{8} - \frac{19t^3}{16} - \frac{517t^4}{128} - \frac{2197t^5}{256} + \dots$$

Let $q(t) = \frac{3t}{2} + \frac{17t^2}{8} + \frac{19t^3}{16}$. Then

$$(1 - q(t))^2 = 1 - 3t - 2t^2 + 4t^3 + \frac{517t^4}{64} + \frac{323t^5}{64} + \frac{361t^6}{256},$$

and

$$R(t) = (1 - q(t))^2 - p(t) = \frac{517t^4}{64} + \frac{323t^5}{64} + \frac{361t^6}{256}.$$

We can then construct the matrix $A(t)$ as described above:

$$A(t) = \begin{bmatrix} \frac{3t}{2} + \frac{17t^2}{8} + \frac{19t^3}{16} & \frac{517t^3}{64} + \frac{323t^4}{64} + \frac{361t^5}{256} \\ t & \frac{3t}{2} + \frac{17t^2}{8} + \frac{19t^3}{16} \end{bmatrix},$$

and $A(t)$ realizes the original polynomial $p(t) = 1 - 3t - 2t^2 + 4t^3$.

6. The case $N = 3$

The $N = 3$ case extends the ideas used in the $N = 2$ case, but is much more complicated since the “left over” terms of the $(1 - q(t))^3$ term cannot be assumed to be all positive. In this case, we work with the matrix

$$A(t) = \begin{bmatrix} q(t) & \alpha(t) & \beta(t) \\ 0 & q(t) & t \\ t & 0 & q(t) \end{bmatrix},$$

where $q(t)$ is a truncation of the power series $r(t) = 1 - p(t)^{1/3}$ of some degree n at least as large as the degree of $p(t)$. In this case,

$$\det(I - A(t)) = (1 - q(t))^3 - t^2\alpha(t) - t\beta(t)(1 - q(t)).$$

In what follows, we will denote by b_m , q_m and r_m the coefficients of the term t^m in the polynomials $\beta(t)$, $q(t)$ and power series $r(t)$ respectively, and by $[f(t)]_m$ the coefficient of t^m in a more complicated polynomial expression, $f(t)$.

Were $R(t) = (1 - q(t))^{1/3} - p(t)$ strictly positive, then this remainder could be accommodated by the $\alpha(t)$ term, as in the $N = 2$ case, and the $\beta(t)$ term would not be needed. However, this is in fact never the case. Consider the highest-order term of $R(t)$. This term (of degree $3n$) will have coefficient $(-q_n)^3$. Thus $R(t)$ will necessarily contain at least one negative coefficient, and in practice usually has many more.

On the other hand, the lowest-order term of $R(t)$ will always be positive. Since this term has degree $n + 1$, greater than the degree of $p(t)$, the coefficient of the term of order $n + 1$ in the polynomial $(1 - r(t))^3 = p(t)$ must be 0. The only “missing” term of degree $n + 1$ when expanding $(1 - q(t))^3$ is $3(-r_{n+1})$. Thus the coefficient of the lowest-order term in $R(t)$, $[R(t)]_{n+1} = 3r_{n+1}$, is positive.

Since negative terms exist in $R(t)$, the $\beta(t)$ polynomial term must be used. Any term $b_m t^m$ in $\beta(t)$ is multiplied by $t(1 - q(t))$ in the determinant of $I - A(t)$ and thus has the effect of decreasing the $(m + 1)$ -th coefficient of $(1 - q(t))^3 - t\beta(t)(1 - q(t))$ and increasing the $(m + 2)$ -th through $(m + n + 1)$ -th coefficients. The end goal is to construct the polynomial $\beta(t)$ in such a way that the remainder polynomial $d(t) = (1 - q(t))^3 - t\beta(t)(1 - q(t)) - p(t)$ has all positive coefficients.

Note that before we include any terms in $\beta(t)$, $\beta(t)$ is zero, so we have $d(t) = R(t)$. We can take the lowest-order term of $d(t)$, which we know to be positive, and include it in $\beta(t)$. This is, in a sense, the largest that this coefficient of $\beta(t)$ can be. If it were any larger then the lowest-order term in the resulting polynomial $d(t)$ would be negative. But it also provides the maximum benefit in terms of increasing the coefficients of terms with higher powers in $d(t)$.

If the next lowest-order term of the resulting $d(t)$ is also positive then we can repeat the process, including this term in $\beta(t)$ as well. This process can be continued either until a negative coefficient is reached or until the entire remaining $d(t)$ is positive. (Success!) In the case that a negative coefficient is reached, one can try again with a larger value of n , meaning that we include more terms in $q(t)$, truncating the power series $r(t)$ at a later point.

Example 6.1 ($N = 3$). Let $p(t) = 1 - 5t + 7t^2 - 3t^3$. Then,

$$p(t)^{1/2} = 1 - \frac{5t}{2} + \frac{3t^2}{8} - \frac{9t^3}{16} - \frac{189t^4}{128} - \frac{891t^5}{256} \dots$$

We cannot use a 2×2 matrix since the power series of $p(t)^{1/2}$ is not of the correct form. The power series of $p(t)^{1/3}$ is of the correct form, however, and

$$p(t)^{1/3} = 1 - \frac{5t}{3} - \frac{4t^2}{9} - \frac{76t^3}{81} - \frac{508t^4}{243} - \frac{3548t^5}{729} \dots$$

We let $q(t)$ be this power series truncated to 3 terms:

$$q(t) = \frac{5t}{3} + \frac{4t^2}{9} + \frac{76t^3}{81}.$$

Then

$$(1 - q(t))^3 = 1 - 5t + 7t^2 - 3t^3 + \frac{508t^4}{81} - \frac{1532t^5}{243} - \frac{3536t^6}{2187} - \frac{32528t^7}{6561} - \frac{23104t^8}{19683} - \frac{438976t^9}{531441}.$$

Only the first term of $R(t)$ is positive, and

$$R(t) = \frac{508t^4}{81} - \frac{1532t^5}{243} - \frac{3536t^6}{2187} - \frac{32528t^7}{6561} - \frac{23104t^8}{19683} - \frac{438976t^9}{531441}.$$

Including this term as the first term in $\beta(t)$, we have

$$(1 - q(t))^3 - \frac{508t^4}{81}(1 - q(t)) = 1 - 5t + 7t^2 - 3t^3 + \frac{112t^5}{27} + \frac{2560t^6}{2187} + \frac{6080t^7}{6561} - \frac{23104t^8}{19683} - \frac{438976t^9}{531441}.$$

Thus we now have an additional positive term which can be included in $\beta(t)$. Repeating this process twice more, we eventually get

$$(1 - q(t))^3 - \left(\frac{508t^4}{81} + \frac{112t^5}{27} + \frac{17680t^6}{2187} \right) (1 - q(t)) = 1 - 5t + 7t^2 - 3t^3 + \frac{106576t^7}{6561} + \frac{41408t^8}{6561} + \frac{3592064t^9}{531441},$$

which is $p(t)$ plus a polynomial with only positive coefficients, which can then be chosen to be $\alpha(t)$ (after dividing out a factor of t^2) in the matrix. Bringing all of these polynomials together, we can construct the matrix

$$A(t) = \begin{bmatrix} \frac{5t}{3} + \frac{4t^2}{9} + \frac{76t^3}{81} & \frac{106576t^5}{6561} + \frac{41408t^6}{6561} + \frac{3592064t^7}{531441} & \frac{508t^3}{81} + \frac{112t^4}{27} + \frac{17680t^5}{2187} \\ 0 & \frac{5t}{3} + \frac{4t^2}{9} + \frac{76t^3}{81} & t \\ t & 0 & \frac{5t}{3} + \frac{4t^2}{9} + \frac{76t^3}{81} \end{bmatrix}$$

such that $A(t)$ realizes the original polynomial $p(t)$.

At this point in our research a computer program was written which ran through the steps of this “greedy algorithm” to determine whether such a matrix could be constructed for trial polynomials $p(t)$ which satisfied the condition that the power series of $p(t)^{1/3} - 1$ had all negative coefficients. All cubic polynomials with integer coefficients less than 100 were tested and no counterexamples were found.

The goal of this algorithm can be reformulated as constructing a polynomial

$$b(t) = \sum_{i=M+1}^{3n} b_i t^i$$

such that $p(t) - (1 - q(t))^3 + b(t)(1 - q(t))$ has coefficient 0 for all terms with degree $3n$ or less. Then if $b(t)$ has only positive terms, the realizing matrix can be

easily constructed. In the following propositions, we demonstrate that it is always possible to construct such a $b(t)$ with all positive coefficients.

First, we note the following:

Proposition 6.2. *Let $p(t)^{1/3} = 1 - r(t) = 1 - q(t) - s(t)$, where $q(t)$ is polynomial of degree n equal to the power series $r(t)$ truncated to degree n and $s(t)$ is a power series consisting of the remaining terms in $r(t)$. Then*

$$b_m = 3[s(t)(1 - q(t) - s(t))]_m.$$

Proof. By the construction of $b(t)$, for all $m < 3n$,

$$[p(t) - (1 - q(t))^3 + b(t)(1 - q(t))]_m = 0,$$

$$[b(t)(1 - q(t))]_m = [(1 - q(t))^3 - p(t)]_m,$$

$$p(t) = ((1 - q(t) - s(t))^3 = (1 - q(t))^3 - 3s(t)(1 - q(t))^2 + 3s(t)^2(1 - q(t)) - s(t)^3.$$

Plugging this expression in for $p(t)$ above, we have

$$[b(t)(1 - q(t))]_m = [3s(t)(1 - q(t))^2 - 3s(t)^2(1 - q(t)) + s(t)^3]_m.$$

The lowest-order term of $s(t)^3$ will have degree $3n + 3$, so it can be dropped, giving

$$\begin{aligned} [b(t)(1 - q(t))]_m &= [3s(t)(1 - q(t))^2 - 3s(t)^2(1 - q(t))]_m \\ &= [(1 - q(t))3s(t)(1 - q(t) - s(t))]_m \\ &= [(1 - q(t))3s(t)(1 - q(t) - s(t))]_m. \end{aligned}$$

Thus,

$$[b(t)]_m = b_m = 3[s(t)(1 - q(t) - s(t))]_m. \quad \square$$

Alternatively, we can write this result in terms of $r(t)$ as

$$b_m = 3[s(t)(1 - q(t) - s(t))]_m = 3\left[r_m + \sum_{i=1}^{m-n} r_i r_{m-i}\right].$$

Proposition 6.3. *Assume $p(t)$ satisfies the conditions of the Boyle–Handelman theorem as well as our strengthened third condition. Let λ_1 be the Perron root of $p(t)$ and suppose $p(t)^{1/3} = 1 - r(t)$. A good estimate of the coefficients r_n of $r(t)$ is*

$$\left| \binom{1/3}{n} \right| \lambda_1^n (a(1/\lambda_1))^{1/3},$$

where $a(t)$ is the polynomial

$$a(t) = \frac{p(t)}{1 - \lambda_1 t}$$

and λ_1 is the Perron root of $p(t)$. By a “good estimate” we mean that

$$\lim_{n \rightarrow \infty} \frac{r_n}{\left| \binom{1/3}{n} \right| \lambda_1^n (a(1/\lambda_1))^{1/3}} = 1.$$

Proof. We begin with two subclaims.

Subclaim 1. *Let $\epsilon > 0$ be given. Then there exists an $N > 0$ such that for any $n > N$ and for any j with $0 < j < n$,*

$$\left| \frac{\binom{1/3}{n-j}}{\binom{1/3}{n}} \right| < \frac{n}{n-j} (1 + \epsilon)^j.$$

Proof. First, note that

$$\begin{aligned} \left| \frac{\binom{1/3}{n-j}}{\binom{1/3}{n}} \right| &= \left| \frac{\frac{\frac{1}{3}(\frac{1}{3}-1)\cdots(\frac{1}{3}-(n-j-1))}{(n-j)!}}{\frac{\frac{1}{3}(\frac{1}{3}-1)\cdots(\frac{1}{3}-(n-1))}{n!}} \right| \\ &= \left| \frac{n!}{(n-j)!} \frac{1}{\left(\frac{1}{3} - (n-j)\right)\left(\frac{1}{3} - (n-j+1)\right)\cdots\left(\frac{1}{3} - (n-1)\right)} \right| \\ &= \frac{n!}{(n-j)!} \frac{1}{\left|\frac{1-3(n-j)}{3}\right| \left|\frac{1-3(n-j+1)}{3}\right| \cdots \left|\frac{1-3(n-1)}{3}\right|} \\ &= \frac{n!}{(n-j)!} \frac{1}{(n-j) \left|1 - \frac{1}{3(n-j)}\right| (n-j+1) \left|1 - \frac{1}{3(n-j+1)}\right| \cdots (n-1) \left|\frac{1}{3(n-1)}\right|} \\ &= \frac{n}{n-j} \prod_{k=1}^j \frac{1}{1 - \frac{1}{3(n-k)}}, \end{aligned}$$

and

$$\log \left(\prod_{k=1}^j \frac{1}{1 - \frac{1}{3(n-k)}} \right)^{1/j} = \frac{1}{j} \sum_{k=1}^j -\log \left(1 - \frac{1}{3(n-k)} \right).$$

Since the denominator $1 - 1/(3(n-k))$ decreases with k ,

$$\begin{aligned} 0 &< \frac{1}{j} \sum_{k=1}^j -\log \left(1 - \frac{1}{3(n-k)} \right) \\ &\leq \frac{1}{n-1} \sum_{k=1}^{n-1} -\log \left(1 - \frac{1}{3(n-k)} \right) = \frac{1}{n-1} \sum_{k=1}^{n-1} -\log \left(1 - \frac{1}{3k} \right). \end{aligned}$$

The last expression above is the average of the first $n-1$ terms of the form $-\log(1 - 1/3k)$. Since these terms tend to 0 as $k \rightarrow \infty$, the average of them does as well. Thus there exists an N such that for all $n \geq N$,

$$\frac{1}{n-1} \sum_{k=1}^{n-1} -\log \left(1 - \frac{1}{3k} \right) < \log(1 + \epsilon),$$

and for $n \geq N$ and for any j with $1 < j < n$,

$$\log \left(\prod_{k=1}^j \frac{1}{1 - \frac{1}{3(n-k)}} \right)^{1/j} < \log(1 + \epsilon).$$

Therefore,

$$\prod_{k=1}^j \frac{1}{1 - \frac{1}{3(n-k)}} < (1 + \epsilon)^j,$$

and

$$\left| \frac{\binom{1/3}{n-j}}{\binom{1/3}{n}} \right| < \frac{n}{n-j} (1 + \epsilon)^j. \quad \square$$

Subclaim 2. Let $\epsilon > 0$ be given and fix $K > 0$. There exists an $N > K$ such that for any $n > N$ and for any j with $0 < j < K$,

$$1 < \left| \frac{\binom{1/3}{n-j}}{\binom{1/3}{n}} \right| < 1 + \epsilon.$$

Proof. Let $\epsilon_1 = (1 + \epsilon)^{1/(K+1)} - 1$. Then by Subclaim 1, there exists an $N_1 > K$ such that for all $n \geq N_1$ and for every j with $0 < j < K < N_1$,

$$\left| \frac{\binom{1/3}{n-j}}{\binom{1/3}{n}} \right| < \frac{n}{n-j} (1 + \epsilon_1)^j \leq \frac{n}{n-K} (1 + \epsilon_1)^K.$$

Since $\lim_{n \rightarrow \infty} n/(n-K) = 1$, there exists N_2 such that for all $n \geq N_2$,

$$\frac{n}{n-K} \leq (1 + \epsilon_1).$$

Let $N = \max(N_1, N_2)$. Then for all $n \geq N$ and j with $0 < j < K$,

$$\left| \frac{\binom{1/3}{n-j}}{\binom{1/3}{n}} \right| < \frac{n}{n-K} (1 + \epsilon_1)^K < (1 + \epsilon_1)(1 + \epsilon_1)^K = (1 + \epsilon_1)^{K+1} = (1 + \epsilon). \quad \square$$

We use these two subclaims to show that given $\epsilon > 0$, there exists an N such that for all $n > N$,

$$\left| \frac{r_n}{\left| \binom{1/3}{n} \right| \lambda_1^n a(1/\lambda_1)^{1/3}} - 1 \right| < \epsilon.$$

Let $\alpha(t) = 1 + \sum_{i=1}^{\infty} \alpha_i t^i$ denote the power series expansion for $a(t)^{1/3} = (p(t)/(1 - \lambda_1 t))^{1/3}$ at $t = 0$. Then

$$p(t)^{1/3} = 1 - r(t) = (1 - \lambda_1 t)^{1/3} \alpha(t) = \left(1 - \sum_{i=1}^{\infty} \left| \binom{1/3}{i} \right| \lambda_1^i t^i \right) \left(1 + \sum_{i=1}^{\infty} \alpha_i t^i \right).$$

We can then write r_n as

$$\begin{aligned} r_n &= \left| \binom{1/3}{n} \right| \lambda_1^n - \alpha_n + \sum_{k=1}^{n-1} \left| \binom{1/3}{n-k} \right| \alpha_k \lambda_1^{n-k} \\ &= \left| \binom{1/3}{n} \right| \lambda_1^n \left(1 + \sum_{k=1}^{n-1} \left| \frac{\binom{1/3}{n-k}}{\binom{1/3}{n}} \right| \alpha_k \lambda_1^{-k} - \frac{\alpha_n}{\left| \binom{1/3}{n} \right|} \lambda_1^{-n} \right). \end{aligned}$$

Let $\delta = \frac{1}{5} \epsilon a(1/\lambda_1)^{1/3}$. If λ_2 is the root of $a(t) = p(t)/(1 - \lambda_1 t)$ with the greatest modulus (i.e., for all λ_i roots of $a(t)$, $|\lambda_2| \geq |\lambda_i|$) then the power series $\alpha(t) = a(t)^{1/3}$ has radius of convergence $1/|\lambda_2|$, which is greater than $1/\lambda_1$. Now, for some $K > 0$ and $n > K_1$, we can write

$$1 + \sum_{k=1}^{n-1} \left| \frac{\binom{1/3}{n-k}}{\binom{1/3}{n}} \right| \alpha_k \lambda_1^{-k} - \frac{\alpha_n}{\left| \binom{1/3}{n} \right|} \lambda_1^{-n} = a(1/\lambda_1)^{1/3} \quad (6-1)$$

$$+ \left(1 + \sum_{k=1}^K \alpha_k \lambda_1^{-k} - a(1/\lambda_1)^{1/3} \right) \quad (6-2)$$

$$+ \sum_{k=1}^K \left(\left| \frac{\binom{1/3}{n-k}}{\binom{1/3}{n}} \right| - 1 \right) \alpha_k \lambda_1^{-k} \quad (6-3)$$

$$+ \sum_{k=K+1}^{n-1} \left| \frac{\binom{1/3}{n-k}}{\binom{1/3}{n}} \right| \alpha_k \lambda_1^{-k} \quad (6-4)$$

$$- \frac{\alpha_n}{\left| \binom{1/3}{n} \right|} \lambda_1^{-n}. \quad (6-5)$$

We can now make each of the terms (6-2) through (6-5) small as follows:

(6-2): Since $1/\lambda_1$ lies in the radius of convergence of $\alpha(t)$, $1 + \sum_{k=1}^{K_1} \alpha_k \lambda_1^{-k}$ converges to $a(1/\lambda_1)^{1/3}$. So for some $K_1 > 0$,

$$\left| 1 + \sum_{k=1}^{K_1} \alpha_k \lambda_1^{-k} - a(1/\lambda_1)^{1/3} \right| < \delta.$$

(6-4): This term is less than

$$\sum_{k=K+1}^{n-1} \left| \frac{\binom{1/3}{n-k}}{\binom{1/3}{n}} \right| |\alpha_k| \lambda_1^{-k}.$$

Fix ϵ_2 such that $(1 + \epsilon_2)/\lambda_1 < 1/|\lambda_2|$. Then by Subclaim 1, there exists a K_2 such that for all $n > K_2$ and $j < n$,

$$\left| \frac{\binom{1/3}{n-j}}{\binom{1/3}{n}} \right| < \frac{n}{n-j} (1 + \epsilon)^j.$$

Then for all $n > K_2$,

$$\begin{aligned} \sum_{k=K_2+1}^{n-1} \left| \frac{\binom{1/3}{n-k}}{\binom{1/3}{n}} \right| |\alpha_k| \lambda_1^{-k} &< \sum_{k=K_2+1}^{n-1} \frac{n}{n-k} (1+\epsilon_2)^k |\alpha_k| \lambda_1^{-k} \\ &= \sum_{k=K_2+1}^{n-1} \left(1 - \frac{k}{n-k}\right) |\alpha_k| \left(\frac{1+\epsilon_2}{\lambda_1}\right)^k \\ &< \sum_{k=K_2+1}^{n-1} |\alpha_k| \left(\frac{1+\epsilon_2}{\lambda_1}\right)^k + \sum_{k=K_2+1}^{n-1} k |\alpha_k| \left(\frac{1+\epsilon_2}{\lambda_1}\right)^k. \end{aligned}$$

Since $\alpha(t)$ converges absolutely, $(1+\epsilon_2)/\lambda_1$ lies within the radius of convergence of both of these series. Thus there exists a $K_3 \geq K_2$ such that for all $n > k_3$, both

$$\sum_{k=K_3+1}^{n-1} |\alpha_k| \left(\frac{1+\epsilon_2}{\lambda_1}\right)^k < \delta$$

and

$$\sum_{k=K_3+1}^{n-1} k |\alpha_k| \left(\frac{1+\epsilon_2}{\lambda_1}\right)^k < \delta.$$

Thus,

$$\sum_{k=K_3+1}^{n-1} \left| \frac{\binom{1/3}{n-k}}{\binom{1/3}{n}} \right| |\alpha_k| \lambda_1^{-k} < 2\delta.$$

(6-5): For sufficiently large n ,

$$\begin{aligned} \frac{\alpha_n}{\binom{1/3}{n}} \lambda_1^{-n} &\leq \left| \frac{n!}{\frac{1}{3}(\frac{1}{3}-1) \cdots (\frac{1}{3}-(n-1))} \right| |\alpha_n| \lambda_1^{-n} \\ &= \left| \frac{1}{\frac{1}{3}(\frac{1}{3}-1)(\frac{1}{3}-2)} \right| \left| \frac{2}{\frac{1}{3}-3} \right| \cdots \left| \frac{n-2}{\frac{1}{3}-(n-1)} \right| (n-1)(n) |\alpha_n| \lambda_1^{-n} \\ &< \frac{27}{10} (n-1)(n) |\alpha_n| \lambda_1^{-n}. \end{aligned}$$

The series $\sum (n-1)(n) \alpha_n t^n$ has radius of convergence greater than $1/\lambda_1$ and converges absolutely, so the sequence $(n-1)(n) \alpha_n t^n$ is Cauchy. Thus there exists K_5 such that for all $n > K_5$,

$$\frac{\alpha_n}{\binom{1/3}{n}} \lambda_1^{-n} < \frac{27}{10} (n-1)(n) |\alpha_n| \lambda_1^{-n} < \delta.$$

At this point we fix K in the equation above so that $K = \max(K_1, K_2, K_3, K_4, K_5)$ and look at the remaining term.

(6-3): This term is less than

$$\sum_{k=1}^{K_1} \left(\left| \frac{\binom{1/3}{n-k}}{\binom{1/3}{n}} \right| - 1 \right) |\alpha_k| \lambda_1^{-k}.$$

Let

$$\epsilon_2 = \frac{\delta}{\sum_{k=1}^{K_1} |\alpha_k| \lambda_1^{-k}}.$$

Then by Subclaim 2, since K is fixed, there exists an $N > K$ such that for all $n > N$ and j with $0 < j \leq K$,

$$1 < \left| \frac{\binom{1/3}{n-j}}{\binom{1/3}{n}} \right| < 1 + \epsilon_2.$$

Thus, for all $n > N$,

$$\sum_{k=1}^{K_1} \left(\left| \frac{\binom{1/3}{n-k}}{\binom{1/3}{n}} \right| - 1 \right) |\alpha_k| \lambda_1^{-k} < \sum_{k=1}^{K_1} ((1 + \epsilon_2) - 1) |\alpha_k| \lambda_1^{-k} = \epsilon_2 \sum_{k=1}^{K_1} |\alpha_k| \lambda_1^{-k} = \delta.$$

Combining the above, for $K = \max(K_1, K_2, K_3, K_4, K_5)$ and $n > N$,

$$1 + \sum_{k=1}^{n-1} \left| \frac{\binom{1/3}{n-k}}{\binom{1/3}{n}} \right| |\alpha_k| \lambda_1^{-k} - \frac{\alpha_n}{\left| \binom{1/3}{n} \right|} \lambda_1^{-n} < a(1/\lambda_1)^{1/3} + 5\delta = a(1/\lambda_1)^{1/3}(1 + \epsilon).$$

So,

$$\begin{aligned} & \left| \frac{r_n}{\left| \binom{1/3}{n} \right| \lambda_1^n a(1/\lambda_1)^{1/3}} - 1 \right| \\ &= \left| \frac{\left| \binom{1/3}{n} \right| \lambda_1^n \left(1 + \sum_{k=1}^{n-1} \left| \binom{1/3}{n-k} \right| / \left| \binom{1/3}{n} \right| \right) |\alpha_k| \lambda_1^{-k} - \alpha_n / \left| \binom{1/3}{n} \right| \lambda_1^{-n}}{\left| \binom{1/3}{n} \right| \lambda_1^n a(1/\lambda_1)^{1/3}} - 1 \right| \\ &< \left| \frac{\left| \binom{1/3}{n} \right| \lambda_1^n (a(t)^{1/3}(1 + \epsilon))}{\left| \binom{1/3}{n} \right| \lambda_1^n a(1/\lambda_1)^{1/3}} - 1 \right| = \epsilon. \quad \square \end{aligned}$$

Proposition 6.4. *Let $1 - c(t) = (p(t))^{2/3}$. Then there exists an N such that for $k > N$, we have $c_k \geq 0$.*

Proof. By the same method as above, a good approximation for c_n is

$$\left| \binom{2/3}{n} \right| \lambda_1^n (q(1/\lambda_1))^{2/3}.$$

Note that $q(1/\lambda_1)$ must be positive since $q(0) = 1$ and $q(t)$ has no root between 0 and $1/\lambda_1$. \square

We can now return to our polynomial $b(t)$, which was constructed such that $p(t) - (1 - q(t))^3 + b(t)(1 - q(t))$ has coefficient 0 for all terms with degree $3n$ or less (n is the degree of $q(t)$).

From Proposition 6.2,

$$[b(t)]_m = b_m = 3[s(t)(1 - q(t) - s(t))]_m,$$

where $p(t)^{1/3} = 1 - r(t) = 1 - q(t) - s(t)$. We can write

$$\begin{aligned} p(t)^{2/3} &= 1 - c(t) = (1 - q(t) - s(t))^2 \\ &= 1 - 2q(t) - 2s(t) + 2q(t)s(t) + q(t)^2 + s(t)^2. \end{aligned}$$

Thus for $n < m \leq 2n$,

$$b_m = 3[s(t)(1 - q(t) - s(t))]_m = \frac{3}{2}(c_n + [q(t)^2]_n),$$

and for $2n < m \leq 3n$,

$$b_m = 3[s(t)(1 - q(t) - s(t))]_m = \frac{3}{2}(c_n - [s(t)^2]_n).$$

So if n is large enough that $c_m \geq 0$ for $m \geq n$, we have

$$b_m = 3[s(t)(1 - q(t) - s(t))]_m = \frac{3}{2}(c_m + [q(t)^2]_m) \geq 0.$$

Then it remains to show that

$$b_m = 3[s(t)(1 - q(t) - s(t))]_m = \frac{3}{2}(c_m - [s(t)^2]_m) \geq 0$$

for $2n < m \leq 3n$.

From Propositions 6.3 and 6.4 above, we can use the approximations

$$s_n \approx \left| \binom{1/3}{n} \right| \lambda_1^n (q(1/\lambda_1))^{1/3} \quad \text{and} \quad c_n \approx \left| \binom{2/3}{n} \right| \lambda_1^n (q(1/\lambda_1))^{2/3}.$$

Note that for $2n < m \leq 3n$,

$$\begin{aligned} \sum_{i, m-i > n} \left| \binom{1/3}{i} \binom{1/3}{m-i} \right| &\leq \sum_{i, m-i > n} \left| \binom{1/3}{n+1} \binom{1/3}{m-(n+1)} \right| \\ &= (m - 2n - 1) \left| \binom{1/3}{n+1} \binom{1/3}{m-(n+1)} \right|. \end{aligned}$$

Proposition 6.5. *For $2n < m \leq 3n$, there exists d with $0 < d < 1$ such that*

$$\frac{(m - 2n - 1) \left| \binom{1/3}{n+1} \binom{1/3}{m-(n+1)} \right|}{\left| \binom{2/3}{m} \right|} \leq 1 - d.$$

Proof. We first prove a couple of subclaims.

Subclaim 3. For a fixed value of n , the expression

$$\frac{(m - 2n - 1) \left| \binom{1/3}{n+1} \binom{1/3}{m-(n+1)} \right|}{\left| \binom{2/3}{m} \right|}$$

is strictly increasing in the range $2n < m < 3n$.

Proof. First note that the denominator of this term, $\left| \binom{2/3}{m} \right|$, is strictly decreasing for increasing m . We can now show that the numerator of this term is strictly increasing by looking at the ratio of consecutive terms. We have

$$\begin{aligned} \frac{(m - 2n - 1) \left| \binom{1/3}{n+1} \binom{1/3}{m-(n+1)} \right|}{(m - 2n) \left| \binom{1/3}{n+1} \binom{1/3}{m-n} \right|} &= \left| \frac{(m - 2n - 1)(m - n)}{(m - 2n)(1/3 - (m - n - 1))} \right| \\ &= \left| \left(1 - \frac{1}{m - 2n}\right) \frac{m - n}{4/3 - (m - n)} \right| \\ &= \frac{1 - \frac{1}{m - 2n}}{1 - \frac{4/3}{m - n}}. \end{aligned} \quad (6-6)$$

Then we can compare $1/(m - 2n)$ to $(4/3)/(m - n)$ by looking at their ratio,

$$\frac{\frac{4/3}{m-n}}{\frac{1}{m-2n}} = \frac{4(m - 2n)}{3(m - n)}.$$

This term is strictly increasing in the range $2n < m < 3n$. It is equal to 0 when $m = 2n$ and equal to $2/3$ when $m = 3n$. Thus for $2n < m < 3n$, we have $1/(m - 2n) > (4/3)/(m - n)$ and $1 - 1/(m - 2n) < 1 - (4/3)/(m - n)$. So the ratio in (6-6) is less than 1, demonstrating that for $2n < m < 3n$,

$$(m - 2n - 1) \left| \binom{1/3}{n+1} \binom{1/3}{m-(n+1)} \right| < (m - 2n) \left| \binom{1/3}{n+1} \binom{1/3}{m-n} \right|. \quad \square$$

Thus it suffices to consider the largest possible value of m , $3n$, which gives us

$$\frac{(n - 1) \left| \binom{1/3}{n+1} \binom{1/3}{2n-1} \right|}{\left| \binom{2/3}{3n} \right|}.$$

Subclaim 4. For all $n \geq 1$,

$$\frac{(n - 1) \left| \binom{1/3}{n+1} \binom{1/3}{2n-1} \right|}{\left| \binom{2/3}{3n} \right|} < \frac{n \left| \binom{1/3}{n} \binom{1/3}{2n} \right|}{\left| \binom{2/3}{3n} \right|}.$$

Proof. Again, by looking at their ratio, the denominators cancel, leaving

$$\begin{aligned}
\frac{(n-1) \binom{1/3}{n+1} \binom{1/3}{2n-1}}{n \binom{1/3}{n} \binom{1/3}{2n}} &= \frac{(n-1) \binom{1/3}{n+1} \binom{1/3}{2n-1}}{n \binom{1/3}{n} \binom{1/3}{2n}} \\
&= \frac{n-1}{n} \frac{1/3-n}{n+1} \frac{2n}{1/3-2n-1} \\
&= \frac{n-1}{n} \frac{n(1-\frac{1}{3n})}{n+1} \frac{2n}{(2n-1)(1-\frac{1}{3(2n-1)})} \\
&= \frac{(n-1)(2n)}{(n+1)(2n-1)} \frac{(1-\frac{1}{3n})}{(1-\frac{1}{3(2n-1)})}.
\end{aligned}$$

Now, we can observe that

$$\frac{1-\frac{1}{3n}}{1-\frac{1}{3(2n-1)}} < 1$$

and

$$\frac{(n-1)(2n)}{(n+1)(2n-1)} = \frac{2n^2-2n}{2n^2+n-1} < 1, \quad (n \geq 1),$$

and thus their product is less than 1. □

Subclaim 5. *The terms*

$$\frac{n \binom{1/3}{n} \binom{1/3}{2n}}{\binom{2/3}{3n}}$$

are strictly decreasing for increasing values of n .

Proof. We again compute the ratio of consecutive terms, and find

$$\begin{aligned}
\left(\frac{(n+1) \binom{1/3}{n+1} \binom{1/3}{2n-1}}{\binom{2/3}{3n}} \right) / \left(\frac{n \binom{1/3}{n} \binom{1/3}{2n-1}}{\binom{2/3}{3n}} \right) &= \frac{n+1}{n} \frac{1/3-n}{n+1} \frac{(1/3-2n)(1/3-(2n+1))}{(2n+1)(2n+2)} \\
&\quad \times \frac{(3n+1)(3n+2)(3n+3)}{(2/3-3n)(2/3-(3n+1))(2/3-(3n+2))} \\
&= \frac{(1-\frac{1}{3n}) \frac{2n}{2n+2} (1-\frac{1}{6n}) (1-\frac{1}{6n+3})}{\frac{3n}{3n+3} (1-\frac{2}{9n}) (1-\frac{2}{9n+3}) (1-\frac{2}{9n+6})}.
\end{aligned}$$

We now define

$$\begin{aligned} f(x) &= \frac{\left(1 - \frac{1}{3x}\right)\left(1 - \frac{1}{6x}\right)\left(1 - \frac{1}{6x+3}\right)}{\left(1 - \frac{2}{9x}\right)\left(1 - \frac{2}{9x+3}\right)\left(1 - \frac{2}{9x+6}\right)} \\ &= \frac{3(3x-1)(6x-1)(3x+1)(3x+1)(3+2)}{x(2x+1)(9x-1)(9x-2)(9x+4)}. \end{aligned}$$

We can compute the derivative of this function,

$$\begin{aligned} f'(x) &= \frac{d}{dx} \left(\frac{3(3x-1)(6x-1)(3x+1)(3x+1)(3+2)}{x(2x+1)(9x-1)(9x-2)(9x+4)} \right) \\ &= \frac{6(104976x^7 + 130491x^6 + 49167x^5 - 1485x^4 - 4239x^3 - 258x^2 + 140x + 8)}{x^2(1458x^4 + 1215x^3 + 135x^2 - 70x - 8)^2}. \end{aligned}$$

The numerator of this function factors as

$$6(3x+1)(8+116x-606x^2-2421x^3+5778x^4+31833x^5+34992x^6).$$

Thus, the only roots of $f'(x)$ can be at $x = -1/3$ or where

$$8+116x-606x^2-2421x^3+5778x^4+31833x^5+34992x^6=0.$$

This has no solutions in $[1, \infty)$, since

$$\begin{aligned} 8+116x-606x^2-2421x^3+5778x^4+31833x^5+34992x^6 \\ = (34992x^6-2421x^3)+(31833x^5-606x^2)+5778x^4+116x+8 \end{aligned}$$

and each term above is strictly positive for $n \geq 1$. By a similar argument, the denominator of $f'(x)$ has no roots in $[1, \infty)$. We can calculate $f'(1) = \frac{1672800}{7452900} \approx 0.2244 > 0$, and thus $f'(x) > 0$ for all $x \in [1, \infty)$. Thus $f(x)$ is strictly increasing on $[1, \infty)$ and

$$\lim_{x \rightarrow \infty} f(x) = \lim_{x \rightarrow \infty} \frac{\left(1 - \frac{1}{3x}\right)\left(1 - \frac{1}{6x}\right)\left(1 - \frac{1}{6x+3}\right)}{\left(1 - \frac{2}{9x}\right)\left(1 - \frac{2}{9x+3}\right)\left(1 - \frac{2}{9x+6}\right)} = 1.$$

So $f(x) < 1$ for all $x \in [1, \infty)$. The terms

$$\frac{\binom{n}{\lfloor n^{1/3} \rfloor} \binom{n}{\lfloor 2n^{1/3} \rfloor}}{\binom{n}{\lfloor 3n^{1/3} \rfloor}}$$

are strictly decreasing for increasing values of n . □

We can then evaluate this expression at $n = 1$ and find

$$\frac{\binom{1}{1} \binom{1}{2}}{\binom{1}{3}} = \frac{3}{4}.$$

Thus for all n, m with $2n < m \leq 3n$,

$$\frac{(m-2n-1) \left| \binom{1/3}{n+1} \binom{1/3}{m-(n+1)} \right|}{\left| \binom{2/3}{m} \right|} \leq \frac{3}{4} = 1 - \frac{1}{4}.$$

So, we can choose $d = 1/4$ and the proposition is valid. \square

Now, write

$$c_m - [s(t)^2]_m = c_m - \sum_{i, m-i > n} s_i s_{m-i}.$$

For convenience we define $A = \left| \binom{2/3}{m} \right|$ and $B = (m-2n-1) \left| \binom{1/3}{n+1} \binom{1/3}{m-(n+1)} \right|$. Choose $\delta > 0$ such that $Ad > \delta(A - 2B - \delta B)$.

By the propositions above, we can choose n such that for all $m > n$,

$$c_m > (1-\delta) \left| \binom{2/3}{m} \right| \lambda_1^m (q(1/\lambda_1))^{2/3}$$

and

$$s_m < (1+\delta) \left| \binom{1/3}{m} \right| \lambda_1^m (q(1/\lambda_1))^{1/3}.$$

So,

$$\begin{aligned} & c_m - [s(t)^2]_m \\ & > (1-\delta) \left| \binom{2/3}{m} \right| \lambda_1^m (q(1/\lambda_1))^{2/3} \\ & \quad - \sum_{i, m-i > n} \left((1+\delta) \left| \binom{1/3}{i} \right| \lambda_1^i (q(1/\lambda_1))^{1/3} \right) \left((1+\delta) \left| \binom{1/3}{m-i} \right| \lambda_1^{m-i} (q(1/\lambda_1))^{1/3} \right) \\ & = (1-\delta) \left| \binom{2/3}{m} \right| \lambda_1^m (q(1/\lambda_1))^{2/3} - \sum_{i, m-i > n} (1+\delta)^2 \left| \binom{1/3}{i} \right| \left| \binom{1/3}{m-i} \right| \lambda_1^m (q(1/\lambda_1))^{2/3} \\ & = (\lambda_1^m (q(1/\lambda_1))^{2/3}) \left((1-\delta) \left| \binom{2/3}{m} \right| - (1+\delta)^2 \sum_{i, m-i > n} \left| \binom{1/3}{i} \right| \left| \binom{1/3}{m-i} \right| \right) \\ & \geq (\lambda_1^m (q(1/\lambda_1))^{2/3}) \left((1-\delta) \left| \binom{2/3}{m} \right| - (1+\delta)^2 (m-2n-1) \left| \binom{1/3}{n+1} \binom{1/3}{m-(n+1)} \right| \right). \end{aligned}$$

In terms of A and B defined above, the term in parentheses in the last line can be expanded to

$$A - \delta A - B - 2\delta B - \delta^2 B.$$

Then since $B/A \leq 1-d$, we have $A - B \geq Ad$, and the expression above is greater than or equal to

$$Ad - \delta A - 2\delta B - \delta^2 B = Bd - \delta(A - 2B - \delta B).$$

By our choice of δ above, this is strictly greater than or equal to 0, so we are done.

Namely, this demonstrates that we can construct a polynomial $b(t)$ of degree at most $3n$ with positive coefficients such that $p(t) - (1 - q(t))^3 + b(t)(1 - q(t))$ has coefficient 0 for all terms with degree $3n$ or less.

Let $d(t) = (1 - q(t))^3 - b(t)(1 - q(t)) - p(t)$. Since n is the degree of $q(t)$, $q(t)^3$ will have degree $3n$ as well, so any remaining terms in $d(t)$ will be the result of trailing terms in the product of $b(t)$ and $q(t)$. Since both of these polynomials contain only positive coefficients, $d(t)$ will as well. As a result, $p(t) = (1 - q(t))^3 - b(t)(1 - q(t)) - d(t)$ and we can construct the matrix

$$A(t) = \begin{bmatrix} q(t) & d(t)/t^2 & b(t)/t \\ 0 & q(t) & t \\ t & 0 & q(t) \end{bmatrix}$$

such that $I - A(t)$ has determinant $p(t)$.

7. Further work

The obvious next step in this research would be to continue to study this problem for larger values of N and to develop constructions for correspondingly larger polynomial matrices. Already for the case $N = 4$ at least a slightly new method will be required. The logical progression to a 4×4 matrix would be to construct

$$M(t) = \begin{bmatrix} q(t) & \alpha(t) & \beta(t) & \gamma(t) \\ 0 & q(t) & t & 0 \\ 0 & 0 & q(t) & t \\ t & 0 & 0 & q(t) \end{bmatrix}.$$

In this case $I - M(t)$ has determinant

$$(1 - q(t))^4 - \alpha(t)t^3 - \beta(t)(1 - q(t))t^2 - \gamma(t)(1 - q(t))^2t.$$

Ignoring the $\gamma(t)$ term (i.e., letting $\gamma(t) = 0$) results in a problem identical to the $N = 3$ case; however, it does not appear that this method will suffice for all polynomials which satisfy the condition that $p(t)^{1/4}$ has all negative coefficients. Thus it is likely that a solution will require use of the $\gamma(t)$ polynomial; however, the same “greedy” algorithm cannot be used. Whereas $1 - q(t)$ had all negative coefficients except for the leading 1, meaning that each coefficient of $\beta(t)$ “helped” all of the coefficients of higher order by making them more positive, $(1 - q(t))^2$ will not in general have that property. So coefficients of $\lambda(t)$ would correct some terms while “hindering” others by making them more negative. It is also possible that a different matrix configuration, utilizing more of the positions occupied by t or 0 is required.

Clearly, the ideal result would be a general proof that demonstrated this result for all N .

Another interesting possibility for research would be to look at the degrees of polynomials required in this construction and to attempt to constrain them. As mentioned before, if a given construction could control the size of both the polynomial matrix and the degrees of the polynomials used in the matrix, then it would put a constraint on the required size of the matrix over \mathbb{R}_+ described in the original problem.

Interestingly, the results given here for $N = 1$ and 2 already constrain the degree of the polynomials used. (For a polynomial of degree d , the $N = 1$ requires only a polynomial of degree d and the $N = 2$ case requires a matrix with polynomials of degree at most $2d$.) However, the polynomials required in the $N = 3$ case may currently have arbitrarily high degree.

References

- [Berman and Plemmons 1979] A. Berman and R. J. Plemmons, *Nonnegative matrices in the mathematical sciences*, Academic Press, New York, 1979. MR 82b:15013 Zbl 0484.15016
- [Boyle 1993] M. Boyle, “Symbolic dynamics and matrices”, pp. 1–38 in *Combinatorial and graph-theoretical problems in linear algebra* (Minneapolis, MN, 1991), edited by R. A. Brualdi et al., IMA Vol. Math. Appl. **50**, Springer, New York, 1993. MR 94g:58062 Zbl 0844.58023
- [Boyle and Handelman 1991] M. Boyle and D. Handelman, “The spectra of nonnegative matrices via symbolic dynamics”, *Ann. of Math. (2)* **133**:2 (1991), 249–316. MR 92d:58057 Zbl 0735.15005
- [Eggleston et al. 2004] P. D. Eggleston, T. D. Lenker, and S. K. Narayan, “The nonnegative inverse eigenvalue problem”, *Linear Algebra Appl.* **379** (2004), 475–490. MR 2005b:15040 Zbl 1040.15009
- [Friedland 2012] S. Friedland, “A note on the nonzero spectra of irreducible matrices”, *Linear Multilinear Algebra* **60**:11–12 (2012), 1235–1238. MR 2989758 Zbl 1257.15006
- [Laffey 2012] T. J. Laffey, “A constructive version of the Boyle–Handelman theorem on the spectra of nonnegative matrices”, *Linear Algebra Appl.* **436**:6 (2012), 1701–1709. MR 2890950 Zbl 1241.15007
- [Laffey et al. 2013] T. J. Laffey, R. Loewy, and H. Šmigoc, “Power series with positive coefficients arising from the characteristic polynomials of positive matrices”, preprint, 2013. arXiv 1205.1933
- [Lind and Marcus 1995] D. Lind and B. Marcus, *An introduction to symbolic dynamics and coding*, Cambridge University Press, 1995. MR 97a:58050 Zbl 1106.37301
- [Minc 1988] H. Minc, *Nonnegative matrices*, Wiley, New York, 1988. MR 89i:15001 Zbl 0638.15008
- [Suleĭmanova 1949] H. R. Suleĭmanova, “Stochastic matrices with real characteristic numbers”, *Doklady Akad. Nauk SSSR (N.S.)* **66** (1949), 343–345. In Russian. MR 11,4d Zbl 0035.20903

Received: 2011-09-05 Revised: 2014-02-01 Accepted: 2014-03-05

nathan.g.mcnew@dartmouth.edu

*Department of Mathematics, Dartmouth College,
6188 Kemeny Hall, Hanover, NH 03755, United States*

normes@du.edu

*Department of Mathematics, University of Denver,
2280 South Vine Street, Denver, CO 80208, United States*

The number of convex topologies on a finite totally ordered set

Tyler Clark and Tom Richmond

(Communicated by Kenneth S. Berenhaut)

We give an algorithm to find the number $T_{\text{cvx}}(n)$ of convex topologies on a totally ordered set X with n elements, and present these numbers for $n \leq 10$.

1. Introduction

A subset B of poset (X, \leq) is *increasing* if $x \in B$ and $y \geq x$ imply $y \in B$, and is *convex* if $x, z \in B$ and $x \leq y \leq z$ imply $y \in B$. An n -point totally ordered set X may be labeled $X = \{1, 2, \dots, n\}$, where $1 < 2 < \dots < n$. This set will be denoted $[1, n]$, and in general, $[a, b]$ will denote $\{a, a+1, \dots, b\} \subset \mathbb{N}$ with the natural order from \mathbb{N} . A topology on (X, \leq) is *convex* if it has a base of convex sets, or equivalently, if each point has a neighborhood base of convex sets. Because of these equivalent characterizations, convex topologies are often called *locally convex* topologies. (See [Nachbin 1965]). For finite sets, every point j has a minimal neighborhood $\text{MN}(j)$, which is the intersection of all neighborhoods of j . It is convenient to identify a topology on $[1, n]$ with its base $\{\text{MN}(j) : j \in [1, n]\}$ of minimal neighborhoods of each point. Finite topological spaces are used in computer graphics, where the Euclidean plane is modeled by a topology on a finite set of pixels. If $a < b < c$ in a finite poset with a topology, if c is “near” a and there is any compatibility between the topology and order, we would expect b to also be near a . This is the convexity condition, which is a natural, weak compatibility condition between a topology and order assumed in most applications. We will consider the number of convex topologies on a finite totally ordered set $[1, n]$.

An excellent reference on finding the number $T(n)$ of topologies on an n -element set is [Erné and Stege 1991]. Currently, $T(n)$ is known for $n \leq 18$. A standard approach to counting topologies on a finite set X is to employ the one-to-one correspondence between a topology τ on X and the associated specialization quasiorder defined by $x \leq y$ if and only if x is in the closure of y . This correspondence

MSC2010: 05A15, 06F30, 54A10, 54F05.

Keywords: convex topology, totally ordered set, number of topologies.

dates back to [Alexandroff 1937]. (See [Richmond 1998] for a survey of this connection.) One approach to counting the convex topologies would be to find a (bordered) characterization of convex topologies using some compatibility between the specialization order and the given total order. Fruitful results in this direction have not been found.

For $j \in [1, n]$, a convex subset $N(j)$ of $[1, n]$ containing j has the form $[a, b]$, where $1 \leq a \leq j \leq b \leq n$. There are j choices for a and $n - j + 1$ choices for b , and thus $j(n + j - 1)$ choices for $N(j)$. Since a base of minimal neighborhoods for a locally convex topology on $[1, n]$ consists of one convex subset $N(j)$ for each $j \in [1, n]$, we see that

$$\prod_{j=1}^n (j)(n + j - 1) = (n!)^2$$

gives an upper bound on $T_{\text{cvx}}(n)$. Of course, arbitrarily selecting a convex set $N(j)$ containing j for each $j \in [1, n]$ is unlikely to give a base for a topology, so this upper bound is not sharp.

2. Nested convex topologies

Stephen [1968] gave a recursive formula for the number of nested topologies (or equivalently, ordered partitions) on an n -point set X , generating the sequence 1, 3, 13, 75, 541, 4683, 47293, \dots , which is A000670 in *The On-Line Encyclopedia of Integer Sequences* (OEIS); see [Sloane 2014]. If $X = [1, n]$ is a totally ordered set with n elements, let $T_{\text{Nest}}(n)$ be the number of nested convex topologies on X , and let $T_{\text{Nest}}(n, k)$ be the number of those convex topologies consisting of k nested nonempty open sets U_1, U_2, \dots, U_k , where $X = U_1 \supset U_2 \supset \dots \supset U_k \neq \emptyset$. Since the indiscrete topology is the only nested topology with one nonempty open set, $T_{\text{Nest}}(n, 1) = 1$. Suppose we have found $T_{\text{Nest}}(m, j)$ for all $m \leq n$ and $j \leq k$. To find $T_{\text{Nest}}(n, k + 1)$, note that $X = U_1 \supset U_2 \supset U_3 \supset \dots \supset U_{k+1} \neq \emptyset$ implies that U_2 must contain at least k elements and at most $n - 1$ elements. If $|U_2| = j$, there are $n - j + 1$ ways to choose U_2 as a convex subset of X , and $T_{\text{Nest}}(j, k)$ ways to complete the nested convex topology $\{U_2, \dots, U_{k+1}\}$ on the j -point totally ordered set U_2 . Thus, we have

$$T_{\text{Nest}}(n, k + 1) = \sum_{j=k}^{n-1} (n - j + 1) T_{\text{Nest}}(j, k) = \sum_{m=2}^{n-k+1} m \cdot T_{\text{Nest}}(n - m + 1, k),$$

where the second equality follows from the substitution $m = n - j + 1$. In Table 1, we tabulate the values of $T_{\text{Nest}}(n, k)$ for $n, k \leq 10$.

| $n \backslash k$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|------------------|---|----|-----|-------|-------|--------|--------|-------|-------|-----|
| 1 | 1 | | | | | | | | | |
| 2 | 1 | 2 | | | | | | | | |
| 3 | 1 | 5 | 4 | | | | | | | |
| 4 | 1 | 9 | 16 | 8 | | | | | | |
| 5 | 1 | 14 | 41 | 44 | 16 | | | | | |
| 6 | 1 | 20 | 85 | 146 | 112 | 32 | | | | |
| 7 | 1 | 27 | 155 | 377 | 456 | 272 | 64 | | | |
| 8 | 1 | 35 | 259 | 833 | 1,408 | 1,312 | 640 | 128 | | |
| 9 | 1 | 44 | 406 | 1,652 | 3,649 | 4,712 | 3,568 | 1,472 | 256 | |
| 10 | 1 | 54 | 606 | 3,024 | 8,361 | 14,002 | 14,608 | 9,312 | 3,328 | 512 |

Table 1. $T_{\text{Nest}}(n, k)$, the number of topologies on a totally ordered n -point set consisting of k nested convex sets.

This table (sequence A056242 in the OEIS [Mallows 2014]) is also used by Hwang and Mallow [1995] to count the number of *order-consecutive partitions* of $X = \{1, 2, \dots, n\}$, which they define as follows: An ordered list S_1, S_2, \dots, S_m of subsets of X is an order-consecutive partition of X if $\{S_1, \dots, S_m\}$ is a partition of X and each of the sets $\bigcup_{j=1}^k S_j$ ($1 \leq k \leq m$) is a consecutive set of integers. If $\{S_1, \dots, S_m\}$ is an order-consecutive partition, clearly $\{S_1, S_1 \cup S_2, S_1 \cup S_2 \cup S_3, \dots, X\}$ is a nested convex topology on X . Conversely, any nested convex topology $\tau = \{U_1, U_2, \dots, U_k\}$ on $X = \{1, 2, \dots, n\}$ generates the order-consecutive partition $U_1, U_2 \setminus U_1, U_3 \setminus U_2, \dots, U_k \setminus U_{k-1}$.

It is easy to confirm from our formula for $T_{\text{Nest}}(n, k)$ that $T_{\text{Nest}}(n, n) = 2^{n-1}$ and $T_{\text{Nest}}(n, 2) = \Delta_n - 1$, where Δ_n is the n -th triangular number.

Now, we note that

$$T_{\text{Nest}}(n) = \sum_{k=1}^n T_{\text{Nest}}(n, k).$$

This sequence, whose first few elements are

$$(T_{\text{Nest}}(n))_{n=1}^{10} = (1, 3, 10, 34, 116, 396, 1352, 4616, 15760, 53808),$$

appears as A007052 in the OEIS [Mallows et al. 2014], where it is noted that

$$T_{\text{Nest}}(n) = 4 T_{\text{Nest}}(n - 1) - 2 T_{\text{Nest}}(n - 2) \quad \text{for } n > 2.$$

Solving this recurrence relation by standard techniques gives

$$T_{\text{Nest}}(n) = \frac{(2 + \sqrt{2})^n + (2 - \sqrt{2})^n}{4}.$$

Nested convex topologies have as much inclusion as possible. Not only are they totally ordered by inclusion, but they maximize “overlap”. The other extreme would be to have as little inclusion and overlap as possible. This suggests considering mutually disjoint collections. A collection \mathcal{D} of mutually disjoint convex subsets of X is not a basis for a topology if $\bigcup \mathcal{D} \neq X$, but $\mathcal{D} \cup \{X\}$ is always a basis for a convex topology on X . The authors have shown that the number of topologies on an n -element totally ordered set having a base consisting of a mutually disjoint collection \mathcal{D} of convex sets, or such a collection \mathcal{D} together with X , is $F_{2n+1} - 1$, where F_k is the k -th Fibonacci number [Clark and Richmond 2010].

3. An algorithm for $T_{\text{cvx}}(n)$

We now present a recursive algorithm to find the number $T_{\text{cvx}}(n)$ of convex topologies on a totally ordered set $[1, n]$. It is easy to check that $T_{\text{cvx}}(1) = 1 = T(1)$ and $T_{\text{cvx}}(2) = 4 = T(2)$. That is, the only topology on a 1-point set is convex, as are all four topologies on a 2-point set.

Suppose $T_{\text{cvx}}(n)$ is known. To find $T_{\text{cvx}}(n+1)$, note that each convex topology on $[1, n+1]$, when restricted to $[1, n]$, gives a unique convex topology on $[1, n]$. Thus, we may count $T_{\text{cvx}}(n)$ by looping through each topology τ counted in $T_{\text{cvx}}(n)$, adding $n+1$ as the greatest point, adjusting the minimal neighborhoods of $j \in [1, n]$, and defining the minimal neighborhood of $n+1$ so that the subspace topology on $[1, n]$ is still τ . That is, considering how each topology on $[1, n]$ may be appropriately expanded to $[1, n+1]$ gives a complete, unduplicated count of the convex topologies on $[1, n+1]$.

Step 1: Redefining minimal neighborhoods of $j \in [1, n]$. We loop through all convex topologies τ on $[1, n]$. The simplest way to extend τ to $[1, n+1]$ so that the restriction of the extension is still τ would be to keep the minimum neighborhoods of each $j \in [1, n]$ unchanged. However, we may also expand some of the minimal neighborhoods of points $j \in [1, n]$ to include $n+1$. To maintain convexity and to guarantee a topology on $[1, n+1]$ whose restriction to $[1, n]$ agrees with τ , the minimal neighborhood $\text{MN}(j)$ of j can be expanded to include $n+1$ if and only if $\text{MN}(j)$ already includes n . If $n \in \text{MN}(j) \subseteq \text{MN}(k)$ and $\text{MN}(j)$ is expanded to include $n+1$, then $\text{MN}(k)$ must also be expanded to include $n+1$, for otherwise $\text{MN}(k)$ would be a neighborhood of j not including $n+1$, contrary to the hypothesis that the minimal neighborhood of j was to include $n+1$.

As an immediate consequence, if $n \in \text{MN}(j) = \text{MN}(k)$, then $\text{MN}(j)$ is expanded to include $n+1$ if and only if $\text{MN}(k)$ is. That is, a single basis element which happens to be the minimal neighborhood of distinct points j and k is still treated as a single entity in the expansion process.

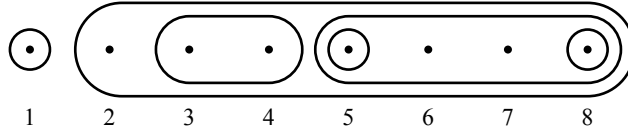


Figure 1. A sample topology on $[1, 8]$.

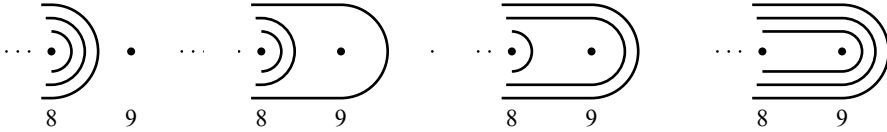


Figure 2. Possible expansions of minimal neighborhoods containing previous right endpoint: none, outermost one, outermost two, outermost three.

Thus, if $\mathcal{B} = \{MN(1), MN(2), \dots, MN(n)\}$ has m distinct sets containing n , we expand the outermost k of these to include $n + 1$, looping as k goes from 1 to m .

For example, consider the convex topology τ on $[1, 8]$ having a base of minimal neighborhoods $\mathcal{B} = \{\{1\}, [2, 8], [3, 4], \{5\}, [5, 8], \{8\}\}$, as shown in Figure 1.

We may add 9 to this topology without changing any of the minimal neighborhoods of j for $j \in [1, 8]$, or since $MN(2), MN(6) = MN(7)$, and $MN(8)$ include the right endpoint 8, they may be extended to include the added point 9. Since $8 \in MN(8) \subset MN(7) = MN(6) \subset MN(2)$, we note that $MN(6)$ is expanded if and only if $MN(7)$ is expanded, so we do not need to treat $MN(6)$ and $MN(7)$ as distinct basis elements and we may effectively ignore the duplicate $MN(7)$. Also, if $MN(6)$ is expanded, then $MN(6) \subset MN(2)$ implies that $MN(2)$ would also have to be expanded. Repeating this idea, we may expand nothing except the outermost (i.e., longest) minimal neighborhood containing 8, namely $MN(2)$, the outermost two minimal neighborhoods containing 8, namely $MN(2)$ and $MN(6)$, or the outermost three, $MN(2), MN(6)$, and $MN(8)$. See Figure 2.

Step 2: Defining the minimal neighborhood of the added point. Having determined the expansion of minimal neighborhoods of $j \in [1, n]$, it remains to define the minimal neighborhood $MN(n+1)$ of $n+1$. Clearly we must have $n+1 \in MN(n+1)$. The convexity condition and our need to retain the original topology τ on $[1, n]$ as a subspace imply that $MN(n+1)$ must be of form $\{n+1\} \cup I$, where I is increasing and open in τ . The final condition is the minimality of the neighborhood $MN(n+1)$. In Step 1, we may have expanded some neighborhoods of n to contain $n+1$ and, if so, the minimal neighborhood of $n+1$ must be contained in each of these previously defined neighborhoods of $n+1$. Thus, $MN(n+1)$ must be of the form $\{n+1\} \cup I$,

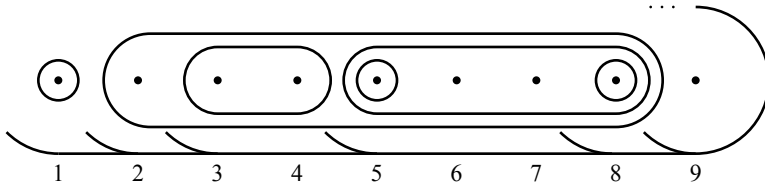


Figure 3. Possible choices for $MN(9)$ if no minimal neighborhoods $MN(j)$ are expanded for $j \in [1, 8]$.

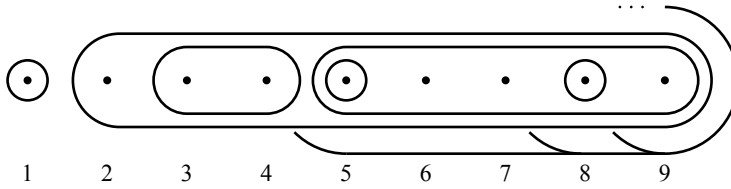


Figure 4. Possible choices for $MN(9)$ if $MN(2)$ and $MN(6)$ are expanded to include 9.

where I is increasing and τ -open, and I is contained in the innermost (shortest) neighborhood $MN(j)$ which was expanded in Step 1.

Continuing the example presented above, we may expand none of the original minimal neighborhoods of $j \in [1, 8]$ to include 9, and then the minimal neighborhood $MN(9)$ of 9 may be defined as $\{9\} \cup I$, where I is an increasing τ -open set in any of the six ways suggested in Figure 3.

Figure 4 shows the three possible choices for the minimal neighborhood $MN(9)$ if the outermost two minimal neighborhoods containing 8, namely $MN(2)$ and $MN(6)$, have been expanded to include 9.

A computer implementation of this algorithm yields the values for $T_{\text{cvx}}(n)$ shown in Table 2 below. With the $T_{\text{cvx}}(2) = 4$ convex topologies on $[1, 2]$ as input, the computer implementation loops through all the topologies τ on $[1, n]$, adds $n + 1$, determines the number m of distinct minimal neighborhoods of $j \in [1, n]$ containing n , expands the outermost k of these to contain $n + 1$ (as k goes from 0 to m), determines the increasing τ -open sets, defines the minimal neighborhood $MN(n + 1)$ of $n + 1$ as $\{n + 1\} \cup I$, where I is one of the increasing τ -open sets contained in the smallest $MN(j)$ previously expanded to include $n + 1$, and, at each selection of an option above, increments the $T_{\text{cvx}}(n + 1)$ counter and records the data for this new topology on $[1, n + 1]$ required for the next iteration.

The efficiency of this algorithm can be improved by eliminating duplication of computations. For example, if p is the largest integer with $MN(p) = X$ for two topologies s and t which agree to the right of p , then the computation for s duplicates that for t , as noted by a helpful referee.

| n | $T_{\text{Nest}}(n)$ | $T_{\text{cvx}}(n)$ | $T(n)$ |
|-----|----------------------|---------------------|-------------------|
| 1 | 1 | 1 | 1 |
| 2 | 3 | 4 | 4 |
| 3 | 10 | 21 | 29 |
| 4 | 34 | 129 | 355 |
| 5 | 116 | 876 | 6,942 |
| 6 | 396 | 6,376 | 209,527 |
| 7 | 1,352 | 48,829 | 9,535,241 |
| 8 | 4,616 | 388,771 | 642,779,354 |
| 9 | 15,760 | 3,191,849 | 63,260,289,423 |
| 10 | 53,808 | 26,864,936 | 8,977,053,873,043 |

Table 2. The numbers $T_{\text{Nest}}(n)$ and $T_{\text{cvx}}(n)$ of nested convex topologies and convex topologies on an n -point totally ordered set, and the number $T(n)$ of topologies on an n -point set.

The numbers $T_{\text{cvx}}(n)$ in Table 2 were also verified for $n \leq 8$ without the algorithm using an exhaustive generation scheme. For comparison, we also include the number $T_{\text{Nest}}(n)$ of nested convex topologies and the number $T(n)$ of topologies on n points in the table.

References

- [Alexandroff 1937] P. Alexandroff, “Diskrete Räume”, *Mat. Sb. (N.S.)* **2:3** (1937), 501–518.
- [Clark and Richmond 2010] T. Clark and T. Richmond, “Collections of mutually disjoint convex subsets of a totally ordered set”, *Fibonacci Quart.* **48:1** (2010), 77–79. Currently also at <http://people.wku.edu/tom.richmond/Papers/FibQuarterly.pdf>. MR 2011e:11029 Zbl 1211.05017
- [Erné and Stege 1991] M. Erné and K. Stege, “Counting finite posets and topologies”, *Order* **8:3** (1991), 247–265. MR 93b:06004 Zbl 0752.05002
- [Hwang and Mallows 1995] F. K. Hwang and C. L. Mallows, “Enumerating nested and consecutive partitions”, *J. Combin. Theory Ser. A* **70:2** (1995), 323–333. MR 96e:05014 Zbl 0819.05005
- [Mallows 2014] C. Mallows, “A056242: Triangle read by rows”, entry A056242 in *The on-line encyclopedia of integer sequences* (<http://oeis.org>), 2014.
- [Mallows et al. 2014] C. Mallows, N. J. A. Sloane, and S. Plouffe, “A007052: Number of order-consecutive partitions of n ”, entry A007052 in *The on-line encyclopedia of integer sequences* (<http://oeis.org>), 2014.
- [Nachbin 1965] L. Nachbin, *Topology and order*, Van Nostrand Mathematical Studies **4**, D. Van Nostrand, Princeton, N.J., 1965. MR 36 #2125 Zbl 0131.37903
- [Richmond 1998] T. A. Richmond, “Quasiorders, principal topologies, and partially ordered partitions”, *Internat. J. Math. Math. Sci.* **21:2** (1998), 221–234. MR 99h:06003 Zbl 0898.54005
- [Sloane 2014] N. J. A. Sloane, “A000670: Fubini numbers”, entry A000670 in *The on-line encyclopedia of integer sequences* (<http://oeis.org>), 2014.

[Stephen 1968] D. Stephen, "Topology on finite sets", *Amer. Math. Monthly* **75** (1968), 739–741.
MR 38 #2725 Zbl 0191.20701

Received: 2011-10-19 Revised: 2013-06-14 Accepted: 2013-08-07

thomas.clark973@topper.wku.edu *Department of Mathematics, Western Kentucky University,
1906 College Heights Boulevard,
Bowling Green, KY 42101, United States*

tom.richmond@wku.edu *Department of Mathematics, Western Kentucky University,
1906 College Heights Boulevard,
Bowling Green, KY 42101, United States*

Nonultrametric triangles in diametral additive metric spaces

Timothy Faver, Katelynn Kochalski, Mathav Kishore Murugan,
Heidi Verheggen, Elizabeth Wesson and Anthony Weston

(Communicated by Toka Diagana)

We prove that a diametral additive metric space is not ultrametric if and only if it contains a diameter attaining nonultrametric triangle.

1. Introduction

Diameter and diametral pairs of points in ultrametric spaces have been the subject of recent extensive studies, including [Dordovskyi et al. 2011]. In this paper we show that if a diametral additive metric space of diameter Δ is not ultrametric, then it must contain a nonultrametric triangle of diameter Δ .

We begin by recalling some preliminary definitions and background information.

Definition 1.1. A metric space (X, d) is said to be *ultrametric* if for all $x, y, z \in X$, we have

$$d(x, y) \leq \max\{d(x, z), d(y, z)\}.$$

Equivalently, a metric space (X, d) is ultrametric if and only if any given three points in X can be relabeled as x, y, z so that $d(x, y) \leq d(x, z) = d(y, z)$.

Interesting examples of ultrametric spaces include the rings Z_p of p -adic integers, the Baire space B_{\aleph_0} , non-Archimedean normed fields and rings of meromorphic functions on open regions of the complex plane. There is an immense literature surrounding ultrametrics, as they have been intensively studied by topologists, analysts, number theorists and theoretical biologists. For example, [de Groot 1956] characterized ultrametric spaces up to homeomorphism as the strongly zero-dimensional metric spaces. In numerical taxonomy, on the other hand, every finite

MSC2010: primary 54E35; secondary 51F99.

Keywords: ultrametric spaces, additive metric spaces, tree metrics.

The research presented in this paper was undertaken at the 2011 Cornell University *Summer Mathematics Institute* (SMI). The authors would like to thank the Department of Mathematics and the Center for Applied Mathematics at Cornell University for supporting this project, and the National Science Foundation for its financial support of the SMI through NSF grant DMS-0739338.

ultrametric space is known to admit a natural hierarchical description called a *dendogram*. This has significant ramifications in theoretical biology. See, for instance, [Gordon 1987].

In fact, ultrametrics are special instances of a more general class of metrics which are termed additive. As we have noted in Definition 1.1, ultrametrics are defined by a stringent three point criterion. The class of additive metrics satisfy a more relaxed four point criterion. The formal definition is as follows.

Definition 1.2. A metric space (X, d) is said to be *additive* if for all x, y, z, w in X , we have

$$d(x, y) + d(z, w) \leq \max\{d(x, z) + d(y, w), d(x, w) + d(y, z)\}.$$

Equivalently, a metric space (X, d) is additive if and only if any given four points in X can be relabeled as x, y, z, w so that $d(x, y) + d(z, w) \leq d(x, z) + d(y, w) = d(x, w) + d(y, z)$.

Recall that a *metric tree* is a connected graph (T, E) without cycles or loops in which each edge $e \in E$ is assigned a positive length $|e|$. The distance $d_T(x, y)$ between any two vertices $x, y \in T$ is then defined to be the sum of the lengths of the edges that make up the unique minimal geodesic from x to y . A brief but important paper, [Buneman 1974, Theorem 2], showed that a finite metric space is additive if and only if it is a tree metric in the sense of the following definition.

Definition 1.3. A metric d on a set X is said to be a *tree metric* if there exists a finite metric tree (T, E, d_T) such that

- (1) X is contained in the vertex set T of the tree, and
- (2) $d(x, y) = d_T(x, y)$ for all $x, y \in X$.

In other words, d is a tree metric if (X, d) is isometric to a metric subspace of some metric tree.

Ultrametrics form a very special subclass of the collection of all additive metrics. Indeed, there is a close relationship between ultrametric spaces and the leaf sets or end spaces of certain trees. This type of identification is discussed more formally in [Holly 2001; Fiedler 1998].

The notion of a diametral metric space is recalled in the following definition. It is a well-known result of mathematical analysis that all compact metric spaces are diametral [Kaplansky 1977, Theorem 68].

Definition 1.4. Let (X, d) be a metric space.

- (1) The *diameter* of a metric space (X, d) is defined to be the quantity $\Delta = \sup\{d(x, y) : x, y \in X\}$. If we need to be more explicit about the underlying metric space, we will write $\text{diam } X$ or $\text{diam}(X, d)$ instead of Δ .
- (2) (X, d) is *diametral* if there exist points $x, y \in X$ such that $d(x, y) = \Delta$.

A metric space (X, d) is not ultrametric if it contains a “bad” triangle $\{x, y, z\} \subseteq X$; i.e., x, y, z such that

$$d(x, y) > \max\{d(x, z), d(y, z)\}.$$

In the case of a nonultrametric diametral additive metric space (X, d) , we will see that there is always a bad triangle whose base length equals $\text{diam } X$. Such triangles are the subject of the following definition.

Definition 1.5. Let (X, d) be a metric space of diameter $\Delta < \infty$. We say that a subset $T = \{x, y, z\}$ of three distinct points from X forms a *diameter nonultrametric triangle* if (T, d) is not ultrametric and $\text{diam } T = \text{diam } X$.

2. Nonultrametric triangles in diametral additive metric spaces

In this section we show that every nonultrametric diametral additive metric space (X, d) contains a diameter nonultrametric triangle. We further note that this result is not true in the more general class of diametral metric spaces. Thus the assumption of additivity is necessary.

Henceforth we will assume that $|X| \geq 3$. The following lemma treats the cases $|X| = 3$ or 4.

Lemma 2.1. *Let (X, d) be a three or four point additive metric space. If X is not ultrametric, then X contains a diameter nonultrametric triangle.*

Proof. The lemma is true by inspection if $|X| = 3$, so we will assume that $|X| = 4$. Let $X = \{x, y, z, a\}$ and suppose that $d(a, z) = \Delta$, where Δ is the diameter of X . If X is not ultrametric, then there exist three distinct points in X that do not satisfy the ultrametric inequality. That is, there exists a three point subset of X that is not ultrametric. Consider the three point subsets of X : $\{x, y, z\}$, $\{x, y, a\}$, $\{y, z, a\}$, $\{x, z, a\}$.

Case 1: $\{y, z, a\}$ is not ultrametric. Since $a, z \in \{y, z, a\}$ and $d(a, z) = \Delta$, we see that $\text{diam}\{y, z, a\} = \Delta$. Then $\{y, z, a\}$ forms a diameter nonultrametric triangle by definition.

Case 2: $\{x, z, a\}$ is not ultrametric. The argument proceeds analogously to Case 1 and is omitted.

Case 3: $\{x, y, z\}$ is not ultrametric. If $\max\{d(a, x), d(x, z)\} < \Delta$, then $\{x, z, a\}$ is not ultrametric, and if $\max\{d(a, y), d(y, z)\} < \Delta$, then $\{y, z, a\}$ is not ultrametric. Then we are reduced to Cases 1 and 2. Suppose that $\max\{d(a, x), d(x, z)\} = \max\{d(a, y), d(y, z)\} = \Delta$. If $d(x, z) = \Delta$, then $\text{diam}\{x, y, z\} = \Delta$ and so $\{x, y, z\}$ forms a diameter nonultrametric triangle. The same occurs if $d(y, z) = \Delta$. Now let $d(a, x) = d(a, y) = \Delta$. As we are assuming that the metric space (X, d) is additive and that $d(a, z) = \Delta$, it follows from [Buneman 1974, Theorem 2] that x, y and z are equidistant from a in some finite metric tree. In particular, no three point subset

of $\{x, y, z, a\}$ that includes a can lie on a common geodesic in this tree. Thus x, y and z must be leaves in the minimal subtree generated by the vertices $\{x, y, z, a\}$. The vertex a may or may not be a leaf in this subtree. However, if a is a leaf in this subtree, we may replace it with the vertex a' in the subtree that minimizes $d(x, a')$ subject to the constraint $d(x, a') = d(y, a') = d(z, a')$. So, by proceeding in this way (if necessary) and by ignoring all irrelevant internal vertices in the subtree, it follows that $\{x, y, z\}$ forms the leaf set of a centered metric tree that has at most five vertices. Thus $\{x, y, z\}$ is ultrametric by [Fiedler 1998, Theorem 2.2].

Case 4: $\{x, y, a\}$ is not ultrametric. The argument proceeds analogously to Case 3 and is omitted. \square

Theorem 2.2. *A diametral additive metric space (X, d) is not ultrametric if and only if X contains a diameter nonultrametric triangle.*

Proof. (\Rightarrow) We prove the contrapositive of the forward implication. Let (X, d) be a diametral metric space with diameter Δ . Suppose X contains no diameter nonultrametric triangles. We may choose $a, b \in X$ with $d(a, b) = \Delta$. Let $x, y, z \in X$ be given. We show that the ultrametric inequality holds for x, y, z . Without loss of generality, we may assume that $x \neq a, b$. Consider the set $X' = \{a, b, x\}$. Clearly $\text{diam}(X', d) = \Delta$. If X' is not ultrametric, then X' forms a diameter nonultrametric triangle in X . So X' must be ultrametric. Thus $d(a, x) = \Delta$ or $d(b, x) = \Delta$. Without loss of generality, we may assume that $d(a, x) = \Delta$. Now consider $X'' = \{a, x, y, z\}$. By construction, $\text{diam}(X'', d) = \Delta$. It follows that any diameter nonultrametric triangle of X'' is also a diameter nonultrametric triangle of X . However, X contains no diameter nonultrametric triangles. So X'' contains no diameter nonultrametric triangles. By Lemma 2.1, X'' is ultrametric. Hence $d(x, y) \leq \max\{d(x, z), d(y, z)\}$, and so X is ultrametric.

(\Leftarrow) Any metric space that contains a diameter nonultrametric triangle is not ultrametric. \square

The following example shows that the forward implication of Theorem 2.2 may fail if the metric space is not assumed to be additive. Consider any nonultrametric metric triangle $(\{x, y, z\}, d)$. Let Δ denote the diameter of this triangle. We may assume that $\Delta = d(x, y) > \max\{d(x, z), d(y, z)\}$. Now adjoin a fourth point a at distance $\Delta + \varepsilon$ from x, y and z where $\varepsilon > 0$. The resulting four point diametral metric space is not additive and contains no diameter nonultrametric triangles.

Acknowledgments

The example following Theorem 2.2 is due to the referee of a previous paper. Comments by that referee motivated this paper in no small measure.

References

- [Buneman 1974] P. Buneman, “A note on the metric properties of trees”, *J. Combinatorial Theory Ser. B* **17** (1974), 48–50. MR 51 #218 Zbl 0286.05102
- [Dordovskiy et al. 2011] D. Dordovskiy, O. Dovgoshey, and E. Petrov, “Diameter and diametrical pairs of points in ultrametric spaces”, *p-Adic Numbers Ultrametric Anal. Appl.* **3**:4 (2011), 253–262. MR 2012k:54043 Zbl 06105084
- [Fiedler 1998] M. Fiedler, “Ultrametric sets in Euclidean point spaces”, *Electron. J. Linear Algebra* **3** (1998), 23–30. MR 99e:51015 Zbl 0897.54020
- [Gordon 1987] A. D. Gordon, “A review of hierarchical classification”, *J. Roy. Statist. Soc. Ser. A* **150**:2 (1987), 119–137. MR 88d:62104 Zbl 0616.62086
- [de Groot 1956] J. de Groot, “Non-archimedean metrics in topology”, *Proc. Amer. Math. Soc.* **7** (1956), 948–953. MR 18,325a Zbl 0072.40201
- [Holly 2001] J. E. Holly, “Pictures of ultrametric spaces, the p -adic numbers, and valued fields”, *Amer. Math. Monthly* **108**:8 (2001), 721–728. MR 1865659 Zbl 1039.12003
- [Kaplansky 1977] I. Kaplansky, *Set theory and metric spaces*, 2nd ed., Chelsea Publishing Co., New York, 1977. MR 56 #5297 Zbl 0397.54002

Received: 2012-06-20 Accepted: 2013-01-10

| | |
|----------------------|--|
| tef36@drexel.edu | <i>Department of Mathematics, Drexel University, Philadelphia, PA 19104, United States</i> |
| kdk7rn@virginia.edu | <i>Department of Mathematics, University of Virginia, Charlottesville, VA 22904, United States</i> |
| mkm233@cornell.edu | <i>Center for Applied Mathematics, Cornell University, Ithaca, NY 14853, United States</i> |
| heidiv@sas.upenn.edu | <i>Department of Economics, University of Pennsylvania, Philadelphia, PA 19104, United State</i> |
| enw27@cornell.edu | <i>Center for Applied Mathematics, Cornell University, Ithaca, NY 14853, United States</i> |
| westona@canisius.edu | <i>Faculty of Arts and Sciences, Australian Catholic University, North Sydney, NSW 2060, Australia</i> |
| and | |
| | <i>Department of Mathematics and Statistics, Canisius College, Buffalo, NY 14208, United States</i> |

An elementary approach to characterizing Sheffer A-type 0 orthogonal polynomial sequences

Daniel J. Galiffa and Tanya N. Riston

(Communicated by Zuhair Nashed)

In 1939, Sheffer published “Some properties of polynomial sets of type zero”, which has been regarded as an indispensable paper in the theory of orthogonal polynomials. Therein, Sheffer basically proved that every polynomial sequence can be classified as belonging to exactly one type. In addition to various interesting and important relations, Sheffer’s most influential results pertained to completely characterizing all of the polynomial sequences of the most basic type, called A-type 0, and subsequently establishing which of these sets were also orthogonal. However, Sheffer’s elegant analysis relied heavily on several characterization theorems. In this work, we show all of the Sheffer A-type 0 orthogonal polynomial sequences can be characterized by using only the generating function that defines this class and a monic three-term recurrence relation.

1. Introduction

In his seminal work, I. M. Sheffer [1939] basically showed that every polynomial sequence can be classified as belonging to exactly one *type*. The majority of his paper was dedicated to developing a wealth of aesthetic results regarding the most basic type, entitled *A-type 0*. This included various interesting characterization theorems. Moreover, one of Sheffer’s most important results was his classification of the A-type 0 orthogonal sets, which are often simply called the Sheffer sequences. Sheffer attributed these orthogonal sets to J. Meixner, who originally discovered them in [Meixner 1934]. The Sheffer sequences (also called Meixner polynomials) are now known to be the very well-studied and applicable Laguerre, Hermite, Charlier, Meixner, Meixner–Pollaczek and Krawtchouk polynomials—refer to [Koekoek and Swarttouw 1996] for details regarding these polynomials and the references therein for additional theory and applications.

MSC2010: 33C45.

Keywords: A-type 0, generating functions, orthogonal polynomials, recurrence relations, Sheffer sequences.

In this paper, we develop and employ an elementary method for characterizing all of the aforementioned Sheffer A-type 0 orthogonal polynomials (Meixner polynomials) that is entirely different than Sheffer's approach. Furthermore, the analysis herein comprises the *most basic* complete characterization of the Sheffer sequences. We also mention that although a very terse overview of the essence of our methodology (for obtaining necessary conditions) is essentially addressed in [Ismail 2009, pp. 524, 525], the rigorous details of applying our approach do not appear anywhere in the current literature.

Since the publication of [Sheffer 1939], a wealth of papers have been written related to the Sheffer sequences, many of which are quite recent. One such work that also develops a basic-type of characterization is [Di Bucchianico and Loeb 1994]. Other papers include [Al-Salam and Verma 1970; Di Bucchianico 1994; Di Nardo et al. 2011; Dominici 2007; Hofbauer 1981; Popa 1997; 1998; Shukla and Rapeli 2011]. In addition, a very large amount of work has been completed pertaining to the theory and applications of specific A-type 0 orthogonal sets, e.g., [Akleyek et al. 2010; Chen et al. 2011; Coffey 2011; Coulembier et al. 2011; Dueñas and Marcellán 2011; Ferreira et al. 2008; Hutník 2011; Khan et al. 2011; Kuznetsov 2008; Miki et al. 2011; Mouayn 2010; Sheffer 1941; Vignat 2011; Wang et al. 2011; Wang and Wong 2011; Yalçınbaşı et al. 2011]. Indeed, research on the Sheffer sequences is an active area and important in its own right. Therefore, our current characterization of such a class is certainly of interest.

In order to sufficiently lay the foreground for our analysis, we first discuss all of the preliminary definitions and terminologies that are utilized throughout this paper. Then, we give a concise overview of Sheffer's method for determining the A-type 0 orthogonal polynomial sequences. We conclude this section by briefly summarizing the sections that follow.

1A. Preliminaries. Throughout this work, we make use of each of the following definitions and terminology.

Definition 1.1. We always assume that a *set* or *sequence* of polynomials $\{P_n(x)\}_{n=0}^{\infty}$ is such that each $P_n(x)$ has degree exactly n .

Definition 1.2. A set of polynomials $\{p_n(x)\}_{n=0}^{\infty}$ is *monic* if $p_n(x) - x^n$ is of degree at most $n - 1$, or equivalently if the leading coefficient of each $p_n(x)$ is unitary.

Definition 1.3. The set of polynomials $\{P_n(x)\}_{n=0}^{\infty}$ is *orthogonal* if it satisfies one of the two following weighted inner product conditions:

$$\text{Continuous : } \langle P_m(x), P_n(x) \rangle = \int_{\Omega_1} P_m(x) P_n(x) w(x) dx = \alpha_n \delta_{m,n}, \quad (1-1)$$

$$\text{Discrete : } \langle P_m(x), P_n(x) \rangle = \sum_{\Omega_2} P_m(x) P_n(x) w(x) = \beta_n \delta_{m,n}, \quad (1-2)$$

where $\delta_{m,n}$ denotes the Kronecker delta

$$\delta_{m,n} := \begin{cases} 1 & \text{if } m = n, \\ 0 & \text{if } m \neq n, \end{cases}$$

with $\Omega_1 \subseteq \mathbb{R}$, $\Omega_2 \subseteq \{0, 1, 2, \dots\}$ and $w(x) > 0$, called the *weight function*.

The Laguerre, Hermite and Meixner–Pollaczek polynomials satisfy a continuous orthogonality relation of the form (1-1). On the other hand, the Charlier, Meixner and Krawtchouk polynomials satisfy a discrete orthogonality relation of the form (1-2); see [Koekoek and Swarttouw 1996].

Definition 1.4. We write each of our orthogonal polynomials in the *hypergeometric form* (${}_rF_s$) as

$${}_rF_s \left(\begin{matrix} a_1, \dots, a_r \\ b_1, \dots, b_s \end{matrix} \middle| z \right) = \sum_{k=0}^{\infty} \frac{(a_1, \dots, a_r)_k}{(b_1, \dots, b_s)_k} \frac{z^k}{k!}, \quad (1-3)$$

where the *Pochhammer symbol* $(a)_k$ is defined as

$$(a)_k := a(a+1)(a+2) \cdots (a+k-1), \quad (a)_0 := 1,$$

with

$$(a_1, \dots, a_j)_k := (a_1)_k \cdots (a_j)_k.$$

The sum (1-3) terminates if one of the numerator parameters is a negative integer; e.g., if one such parameter is $-n$, then (1-3) is a finite sum over $0 \leq k \leq n$.

Definition 1.5. We define a *linear generating function* for a polynomial sequence $\{P_n(x)\}_{n=0}^{\infty}$ by

$$\sum_{\Lambda} \xi_n P_n(x) t^n = F(x, t),$$

where $\{\xi_n\}_{n=0}^{\infty}$ is a sequence in n , independent of x and t , with $\Lambda \subseteq \{0, 1, 2, \dots\}$. Moreover, we say that the function $F(x, t)$ *generates* the set $\{P_n(x)\}_{n=0}^{\infty}$.

It is important to mention that a linear generating function need not converge, as several relationships can be derived when $F(x, t)$ is divergent. For example, by expanding each of the generating functions used in this paper as a *formal* power series in t , the respective polynomial $P_k(x)$ can be determined by evaluating the coefficient of t^k .

It is well-known that a necessary and sufficient condition for a set of polynomials $\{P_n(x)\}_{n=0}^{\infty}$ to be orthogonal is that it satisfies a three-term recurrence relation (see [Rainville 1960]), which can be written in different forms. In particular, we utilize the following two forms in this work, and adhere to the nomenclature used in [Al-Salam 1990].

Definition 1.6 (the three-term recurrence relations). It is a necessary and sufficient condition that an orthogonal set $\{P_n(x)\}_{n=0}^{\infty}$ satisfies an *unrestricted three-term recurrence relation* of the form

$$P_{n+1}(x) = (A_n x + B_n)P_n(x) - C_n P_{n-1}(x), \quad A_n A_{n-1} C_n > 0, \\ \text{where } P_{-1}(x) = 0 \text{ and } P_0(x) = 1. \quad (1-4)$$

If $p_n(x)$ represents the monic form of $P_n(x)$, then it is a necessary and sufficient condition that $\{p_n(x)\}_{n=0}^{\infty}$ satisfies the following *monic three-term recurrence relation*

$$p_{n+1}(x) = (x + b_n)p_n(x) - c_n p_{n-1}(x), \quad c_n > 0, \\ \text{where } P_{-1}(x) = 0 \text{ and } P_0(x) = 1. \quad (1-5)$$

1B. A summary of Sheffer's A type-0 analysis. In order to determine each of the A-type 0 orthogonal sets previously discussed, Sheffer first developed a characterization theorem, which gave necessary and sufficient conditions for a polynomial sequence to be A-type 0 via a linear generating function. Meixner [1934] essentially determined which orthogonal sets satisfy the A-type 0 generating function using a different approach than Sheffer. Meixner used the A-type 0 generating function as the *definition* of the A-type 0 class. In our present work, we follow Meixner's convention. The reader can also refer to [Al-Salam 1990] for a concise overview of Meixner's analysis. In addition, for rigorous developments of the methods of Sheffer and Meixner, as well as related results, extensions and applications, see [Galiffa 2013].

Definition 1.7. A polynomial set $\{P_n(x)\}_{n=0}^{\infty}$ is classified as A-type 0 if there exist $\{a_j\}_{j=0}^{\infty}$ and $\{h_j\}_{j=1}^{\infty}$ such that

$$A(t)e^{xH(t)} = \sum_{n=0}^{\infty} P_n(x)t^n, \quad (1-6)$$

with

$$A(t) := \sum_{n=0}^{\infty} a_n t^n, \quad a_0 = 1 \quad \text{and} \quad H(t) := \sum_{n=1}^{\infty} h_n t^n, \quad h_1 = 1. \quad (1-7)$$

To determine which orthogonal sets satisfy (1-6), Sheffer utilized a monic three-term recurrence relation of the form

$$P_n(x) = (x + \lambda_n)P_{n-1}(x) - \mu_n P_{n-2}(x), \quad n = 1, 2, \dots \quad (1-8)$$

Along with several additional results, Sheffer essentially established the following:

Theorem 1.8. *A necessary and sufficient condition for an A-type 0 set $\{P_n(x)\}_{n=0}^{\infty}$ to satisfy (1-8) is that*

$$\lambda_n = \alpha + bn \quad \text{and} \quad \mu_n = (n-1)(c + dn),$$

with $c + dn \neq 0$ for $n > 1$.

In other words, Sheffer proved that in order for an A-type 0 set $\{P_n(x)\}_{n=0}^{\infty}$ defined by (1-6) to be orthogonal, it must be that λ_n is at most linear in n and μ_n is at most quadratic in n .

Since in our present work we make use of a monic three-term recurrence relation of the form (1-5), i.e., the contemporary form, we scale (1-8) via $n \mapsto n + 1$, giving

$$P_{n+1}(x) = (x + \lambda_{n+1})P_n(x) - \mu_{n+1}P_{n-1}(x), \quad n = 0, 1, 2, \dots, \quad (1-9)$$

and the recursion coefficients in Theorem 1.8 therefore take on the form

$$\lambda_{n+1} = (\alpha + b) + bn, \quad \mu_{n+1} = (c + d)n + dn^2. \quad (1-10)$$

Theorem 1.8, again along with additional results, eventually led Sheffer to the following characterizing theorem, which yields all of the general A-type 0 orthogonal sets in terms of their linear generating functions, and which is written below using the same notation as in [Sheffer 1939].

Theorem 1.9. *A polynomial set $\{P_n(x)\}_{n=0}^{\infty}$ is A-type 0 and orthogonal if and only if $A(t)e^{xH(t)}$ in (1-6) is of one of the following forms:*

$$A(t)e^{xH(t)} = \mu(1 - bt)^c \exp\left\{\frac{d + atx}{1 - bt}\right\}, \quad abc\mu \neq 0, \quad (1-11)$$

$$A(t)e^{xH(t)} = \mu \exp[t(b + ax) + ct^2], \quad ac\mu \neq 0, \quad (1-12)$$

$$A(t)e^{xH(t)} = \mu e^{ct}(1 - bt)^{d+ax}, \quad abc\mu \neq 0, \quad (1-13)$$

$$A(t)e^{xH(t)} = \mu(1 - t/c)^{d_1+x/a}(1 - t/b)^{d_2-x/a}, \quad abc\mu \neq 0, \quad b \neq c. \quad (1-14)$$

By judiciously choosing each of the parameters in (1-11)–(1-14) we can achieve all of the Sheffer A-type 0 orthogonal sets. For emphasis, we write each of these parameter selections below and then display the corresponding generating function as it appears in [Koekoek and Swarttouw 1996]. We also call upon each of these generating relations in Section 4.

The Laguerre polynomials. In (1-11), we select the parameters as $\mu = 1$, $a = -1$, $b = 1$, $c = -(\alpha + 1)$ and $d = 0$ to obtain

$$\sum_{n=0}^{\infty} L_n^{(\alpha)}(x)t^n = (1 - t)^{-(\alpha+1)} \exp\left(\frac{xt}{t - 1}\right). \quad (1-15)$$

The Hermite polynomials. With the assignments $\mu = 1$, $a = 2$, $b = 0$ and $c = -1$ in (1-12), we have

$$\sum_{n=0}^{\infty} \frac{1}{n!} H_n(x)t^n = \exp(2xt - t^2). \quad (1-16)$$

The Charlier polynomials. If in (1-13) we choose $\mu = 1$, $a = 1$, $b = 1/\alpha$, $c = 1$, and $d = 0$, then we obtain

$$\sum_{n=0}^{\infty} \frac{1}{n!} C_n(x; \alpha) t^n = e^t \left(1 - \frac{t}{\alpha}\right)^x. \quad (1-17)$$

The Meixner polynomials. In (1-14), we select $\mu = 1$, $a = 1$, $b = 1$, c arbitrary, $d_1 = 0$ and $d_2 = -\beta$ leading to

$$\sum_{n=0}^{\infty} \frac{(\beta)_n}{n!} M(x; \beta, c) t^n = \left(1 - \frac{t}{c}\right)^x (1-t)^{-(x+\beta)}. \quad (1-18)$$

The Meixner–Pollaczek polynomials. Taking $\mu = 1$, $a = -i$, $b = e^{i\phi}$, $c = e^{-i\phi}$ and $d_1 = d_2 = -\lambda$ in (1-14) leads to

$$\sum_{n=0}^{\infty} P_n^{(\lambda)}(x; \phi) t^n = (1 - e^{i\phi} t)^{-\lambda+ix} (1 - e^{-i\phi} t)^{-\lambda-ix}. \quad (1-19)$$

The Krawtchouk polynomials. Lastly, selecting $\mu = 1$, $a = 1$, $b = -1$, $c = p/(1-p)$, $d_1 = 0$ and $d_2 = N$ in (1-14) yields

$$\sum_{n=0}^N C(N, n) K_n(x; p, N) t^n = \left(1 - \frac{1-p}{p} t\right)^x (1+t)^{N-x}, \quad (1-20)$$

for $x = 0, 1, 2, \dots, N$, where $C(N, n)$ denotes the binomial coefficient.

Interestingly enough, Sheffer only stated (1-15) and (1-16) by their names, i.e., the Laguerre and Hermite polynomials respectively. Moreover, at the time when [Sheffer 1939] was published, the remaining orthogonal polynomials were not yet commonly referred to by the names above; the exception to this being the Charlier polynomials, which were called the Poisson–Charlier polynomials by Meixner [1934] and others.

1C. An overview of our present A-type 0 analysis. Our current work amounts to determining which A-type 0 polynomial sequences are also orthogonal by utilizing only the generating function (1-6) and the monic three-term recurrence relation (1-9), without calling upon any additional relationships. It is in this regard that our approach is elementary. The remainder of this paper is organized as follows.

In Section 2, we derive necessary conditions for the Sheffer A-type 0 recursion coefficients λ_{n+1} and μ_{n+1} as in (1-10), which in fact comprise only the terms a_1 , a_2 , h_2 and h_3 in (1-7). In Section 3, we prove that the A-type 0 orthogonal sets are necessarily the monic forms of the Laguerre, Hermite, Charlier, Meixner, Meixner–Pollaczek and Krawtchouk polynomials by appropriately selecting the parameters a_1 , a_2 , h_2 and h_3 . As a supplement to this analysis, in Section 4 we first derive

linear generating functions for each of the monic forms of the A-type 0 orthogonal sets using (1-15)–(1-20). From these relations, we obtain the same parameter values as those in Section 3. We conclude this paper in Section 5 by showing that the conditions on the recursion coefficients λ_{n+1} and μ_{n+1} from Section 2 are also sufficient. This provides six additional basic characterizations of our orthogonal sets.

2. Deriving the Sheffer A-type 0 recursion coefficients

In this section, we derive necessary conditions for the recursion coefficients λ_{n+1} and μ_{n+1} to be as in (1-10). In order to do this, we first determine the coefficients of x^n , x^{n-1} and x^{n-2} of the arbitrary Sheffer A-type 0 polynomial $P_n(x)$ in (1-6), which we label as $c_{n,0}$, $c_{n,1}$ and $c_{n,2}$, respectively. To obtain these values, we compare the coefficients of $x^k t^n$ for $k = n, n-1, n-2$ on both sides of (1-6). After these leading coefficients are discovered, we substitute our polynomial $P_n(x) = c_{n,0}x^n + c_{n,1}x^{n-1} + c_{n,2}x^{n-2} + \mathcal{O}(x^{n-3})$ into the three-term recurrence relation (1-4) and derive a system of simultaneous linear equations, the solution of which yields the recursion coefficients A_n , B_n and C_n as in (1-4). We then transform the resulting unrestricted recurrence relation into monic form, which gives λ_{n+1} and μ_{n+1} .

We begin by expanding the left side of (1-6) and accounting for $h_1 = 1$ via (1-7):

$$\begin{aligned} \sum_{n=0}^{\infty} a_n t^n \cdot \exp(x(t + h_2 t^2 + h_3 t^3 + \dots)) \\ &= \sum_{n=0}^{\infty} a_n t^n \cdot \exp(xt) \cdot \exp(h_2 x t^2) \cdot \exp(h_3 x t^3) \dots \\ &= \sum_{k_0=0}^{\infty} a_{k_0} t^{k_0} \cdot \sum_{k_1=0}^{\infty} \frac{(xt)^{k_1}}{k_1!} \cdot \sum_{k_2=0}^{\infty} \frac{(h_2 x t^2)^{k_2}}{k_2!} \cdot \sum_{k_3=0}^{\infty} \frac{(h_3 x t^3)^{k_3}}{k_3!} \dots \end{aligned}$$

We next express the general term in each of the products above as

$$a_{k_0} t^{k_0} \cdot \frac{x^{k_1} t^{k_1}}{k_1!} \cdot \frac{h_2^{k_2} x^{k_2} t^{2k_2}}{k_2!} \cdot \frac{h_3^{k_3} x^{k_3} t^{3k_3}}{k_3!} \dots \quad (2-1)$$

Thus, discovering the coefficient of $x^r t^s$ is equivalent to determining all of the nonnegative integer solutions $\{k_0, k_1, k_2, \dots\}$ to the linear Diophantine equations

$$k_1 + k_2 + k_3 + \dots = r, \quad (2-2)$$

$$k_0 + k_1 + 2k_2 + 3k_3 + \dots = s, \quad (2-3)$$

where (2-2) represents the x -exponents and (2-3) the t -exponents. We can now discover the coefficients $x^n t^n$, $x^{n-1} t^n$ and $x^{n-2} t^n$, which we partition into the three parts below.

The coefficient of $x^n t^n$. For this case, we subtract (2-2) from (2-3) with $r = n$ and $s = n$, yielding

$$k_0 + k_2 + 2k_3 + \cdots = 0.$$

It is then readily seen that k_1 is a free variable and $k_0 = k_2 = k_3 = \cdots = 0$. Thus, from substituting these values into (2-3) with $s = n$, we see that $k_1 = n$, and after comparing with (2-1) we observe that the coefficient of $x^n t^n$ is $1/n!$.

The coefficient of $x^{n-1} t^n$. Here, we subtract (2-2) from (2-3) with $r = n - 1$ and $s = n$, which gives

$$k_0 + k_2 + 2k_3 + \cdots = 1,$$

yielding two cases:

Case 1. $k_0 = 1$ and $k_2 = k_3 = \cdots = 0$. Substituting these values into (2-3) gives $k_1 = n - 1$, and via (2-1) we achieve

$$\frac{a_1}{(n-1)!}.$$

Case 2. $k_2 = 1$ and $k_0 = k_3 = \cdots = 0$. Now, (2-3) becomes $k_1 = n - 2$, and from (2-1) we have

$$\frac{h_2}{(n-2)!}.$$

Therefore, we know that the coefficient of $x^{n-1} t^n$ is

$$\frac{a_1}{(n-1)!} + \frac{h_2}{(n-2)!}.$$

The coefficient of $x^{n-2} t^n$. Lastly, we subtract (2-2) from (2-3) with $r = n - 2$ and $s = n$, and obtain

$$k_0 + k_2 + 2k_3 = 2,$$

which has four solutions, yielding four cases. In the same way as in the previous cases, we see that the coefficient of $x^{n-2} t^n$ is

$$\frac{a_2}{(n-2)!} + \frac{a_1 h_2 + h_3}{(n-3)!} + \frac{h_2^2}{2!(n-4)!};$$

the details have been omitted for brevity. Thus, we have established the following:

Lemma 2.1. *For the Sheffer A-type 0 polynomial $P_n(x) = c_{n,0}x^n + c_{n,1}x^{n-1} + c_{n,2}x^{n-2} + \mathcal{O}(x^{n-3})$ as in (1-6), we have*

$$\begin{aligned} c_{n,0} &= \frac{1}{n!}, & c_{n,1} &= \frac{a_1}{(n-1)!} + \frac{h_2}{(n-2)!}, \\ c_{n,2} &= \frac{a_2}{(n-2)!} + \frac{a_1 h_2 + h_3}{(n-3)!} + \frac{h_2^2}{2!(n-4)!}. \end{aligned} \tag{2-4}$$

Proof. See the above analysis. \square

We now have the following result:

Theorem 2.2. *The Sheffer A-type 0 recursion coefficients A_n , B_n and C_n satisfying (1-4) are given by*

$$\begin{aligned} A_n &= \frac{1}{n+1}, & B_n &= \frac{a_1 + 2h_2n}{n+1}, \\ C_n &= \frac{a_1^2 - 2a_2 + 2a_1h_2 - 4h_2^2 + 3h_3 + (4h_2^2 - 3h_3)n}{n+1}. \end{aligned} \quad (2-5)$$

Proof. We see that upon substituting $P_n(x) = c_{n,0}x^n + c_{n,1}x^{n-1} + c_{n,2}x^{n-2} + \mathcal{O}(x^{n-3})$ into the three-term recurrence relation (1-4), we obtain

$$\begin{aligned} c_{n+1,0}x^{n+1} + c_{n+1,1}x^n + c_{n+1,2}x^{n-1} + \mathcal{O}(x^{n-2}) \\ &= A_n c_{n,0}x^{n+1} + A_n c_{n,1}x^n + A_n c_{n,2}x^{n-1} + \mathcal{O}(x^{n-2}) \\ &\quad + B_n c_{n,0}x^n + B_n c_{n,1}x^{n-1} + B_n c_{n,2}x^{n-2} + \mathcal{O}(x^{n-3}) \\ &\quad - C_n c_{n-1,0}x^{n-1} - C_n c_{n-1,1}x^{n-2} - C_n c_{n-1,2}x^{n-3} + \mathcal{O}(x^{n-4}). \end{aligned}$$

Thus, comparing the coefficients of x^{n+1} , x^n and x^{n-1} above results in the lower-triangular simultaneous system of linear equations

$$\begin{bmatrix} c_{n,0} & 0 & 0 \\ c_{n,1} & c_{n,0} & 0 \\ c_{n,2} & c_{n,1} & -c_{n-1,0} \end{bmatrix} \begin{bmatrix} A_n \\ B_n \\ C_n \end{bmatrix} = \begin{bmatrix} c_{n+1,0} \\ c_{n+1,1} \\ c_{n+1,2} \end{bmatrix}.$$

Since the diagonal terms $c_{n,0}$ and $c_{n-1,0}$ are nonzero by Definition 1.1, the solution to the above system is unique and determined via Gauss–Jordan Elimination to be

$$\begin{aligned} A_n &= \frac{c_{n+1,0}}{c_{n,0}}, & B_n &= \frac{c_{n+1,1}c_{n,0} - c_{n+1,0}c_{n,1}}{c_{n,0}^2}, \\ C_n &= \frac{c_{n+1,0}(c_{n,0}c_{n,2} - c_{n,1}^2) + c_{n,0}(c_{n+1,1}c_{n,1} - c_{n+1,2}c_{n,0})}{c_{n-1,0}c_{n,0}^2}. \end{aligned}$$

Substituting (2-4) accordingly yields our desired result. \square

We now determine λ_{n+1} and μ_{n+1} . To accomplish this, we must derive a monic three-term recurrence relation of the form (1-9) from the recursion coefficients (2-5). Thus, we replace $P_n(x)$ with $d_n Q_n(x)$ in (1-4), resulting in

$$Q_{n+1}(x) = \frac{d_n}{d_{n+1}} A_n x Q_n(x) + \frac{d_n}{d_{n+1}} B_n Q_n(x) - \frac{d_{n-1}}{d_{n+1}} C_n Q_{n-1}(x).$$

Therefore, we require

$$\frac{d_n}{d_{n+1}} A_n = 1,$$

which is a first-order linear difference equation readily solved via iterations to be

$$d_n = \frac{1}{n!}.$$

Then, we have

$$\lambda_{n+1} = \frac{d_n}{d_{n+1}} B_n = a_1 + 2h_2 n, \quad (2-6)$$

$$\mu_{n+1} = \frac{d_{n-1}}{d_{n+1}} C_n = (a_1^2 - 2a_2 + 2a_1 h_2 - 4h_2^2 + 3h_3)n + (4h_2^2 - 3h_3)n^2. \quad (2-7)$$

Thus, we have shown that λ_{n+1} is at most linear in n and that μ_{n+1} is at most quadratic in n . Hence, we have the following statement:

Theorem 2.3. *For a polynomial sequence $\{P_n(x)\}_{n=0}^\infty$ to be A-type 0 and orthogonal, the recursion coefficients λ_{n+1} and μ_{n+1} in*

$$P_{n+1}(x) = (x + \lambda_{n+1})P_n(x) - \mu_{n+1}P_{n-1}(x), \quad n = 0, 1, 2, \dots$$

must necessarily be of the form

$$\lambda_{n+1} = c_1 + c_2 n \quad \text{and} \quad \mu_{n+1} = c_3 n + c_4 n^2, \quad c_1, \dots, c_4 \in \mathbb{R},$$

with $\mu_{n+1} > 0$.

Interestingly enough, the parameters c_1, \dots, c_4 above are only in terms of the first two nonunitary coefficients of t in $A(t)$ and $H(t)$ of (1-7), i.e., a_1, a_2, h_2 and h_3 . Furthermore, in regard to Sheffer's analysis, we can readily write λ_{n+1} and μ_{n+1} in Theorem 2.3 as in (1-10) and uniquely determine the parameters α, b, c and d .

Corollary 2.4. *The parameters α, b, c and d in the Sheffer A-type 0 monic recursion coefficients λ_{n+1} and μ_{n+1} of (1-10) are*

$$\alpha = a_1 - 2h_2, \quad b = 2h_2, \quad c = a_1^2 - 2a_2 + 2a_1 h_2 - 8h_2^2 + 6h_3 \quad \text{and} \quad d = 4h_2^2 - 3h_3.$$

3. The Sheffer A-type 0 orthogonal polynomials

In this section, we prove the following theorem, which relies on the analysis conducted in Section 2:

Theorem 3.1. *The following orthogonal polynomial sequences all necessarily belong to the Sheffer A-type 0 class:*

$$\left\{ (-1)^n n! L_n^{(\alpha)}(x), \quad \{2^{-n} H_n(x)\}, \quad \{(-a)^n C_n(x; a)\}, \quad \left\{ \frac{c^n(\beta)_n}{(c-1)^n} M_n(x; \beta, c) \right\}, \right. \\ \left. \{ (2 \sin \phi)^{-n} n! P_n^{(\lambda)}(x; \phi) \}, \quad \{ (-N)_n p^n K_n(x; p, N) \}; \right.$$

these are respectively the monic forms of the Laguerre, Hermite, Charlier, Meixner, Meixner–Pollaczek and Krawtchouk polynomials, as defined in (1-15)–(1-20).

Proof. We first substitute λ_{n+1} and μ_{n+1} as in (2-6) and (2-7), respectively, into (1-9). We therefore see that every A-type 0 orthogonal set must necessarily satisfy a monic three-term recurrence of the form

$$P_{n+1}(x) = [x + a_1 + 2h_2n]P_n(x) - [(a_1^2 - 2a_2 + 2a_1h_2 - 4h_2^2 + 3h_3)n + (4h_2^2 - 3h_3)n^2]P_{n-1}(x). \quad (3-1)$$

We now separately consider each of the monic three-term recurrence relations for the Laguerre, Hermite, Charlier, Meixner, Meixner–Pollaczek and Krawtchouk polynomials, and then uniquely determine the values that the parameters a_1 , a_2 , h_2 and h_3 must take in each case.

The Laguerre polynomials. The Laguerre polynomials satisfy a monic three-term recurrence relation of the form

$$\mathcal{L}_{n+1}^{(\alpha)}(x) = (x - (\alpha + 1) - 2n)\mathcal{L}_n^{(\alpha)}(x) - (\alpha n + n^2)\mathcal{L}_{n-1}^{(\alpha)}(x), \quad (3-2)$$

where

$$\mathcal{L}_n^{(\alpha)}(x) := (-1)^n n! L_n^{(\alpha)}(x) \quad (3-3)$$

with

$$L_n^{(\alpha)}(x) := \frac{(\alpha + 1)_n}{n!} {}_1F_1\left(\begin{matrix} -n \\ \alpha + 1 \end{matrix} \middle| x\right).$$

Therefore, comparing (3-2) with (3-1), we see that

$$a_1 = -(\alpha + 1), \quad a_2 = \frac{1}{2}(\alpha + 1)(\alpha + 2), \quad h_2 = -1, \quad h_3 = 1. \quad (3-4)$$

Thus, $\{(-1)^n n! L_n^{(\alpha)}(x)\}_{n=0}^{\infty}$ is a Sheffer A-type 0 orthogonal set.

The Hermite polynomials. The monic recurrence relation for the Hermite polynomials is

$$\mathcal{H}_{n+1}(x) = x\mathcal{H}_n(x) - \frac{1}{2}n\mathcal{H}_{n-1}(x), \quad (3-5)$$

where

$$\mathcal{H}_n(x) := 2^{-n} H_n(x) \quad (3-6)$$

with

$$H_n(x) := 2^n x^n {}_2F_0\left(\begin{matrix} -n/2, (1-n)/2 \\ - \end{matrix} \middle| -\frac{1}{x^2}\right).$$

From comparing (3-5) with (3-1), we obtain

$$a_1 = 0, \quad a_2 = -1/4, \quad h_2 = 0, \quad h_3 = 0. \quad (3-7)$$

Thus, $\{2^{-n} H_n(x)\}_{n=0}^{\infty}$ is a Sheffer A-type 0 orthogonal set.

The Charlier polynomials. The Charlier polynomials satisfy a monic three-term recurrence relation of the form

$$\mathcal{C}_{n+1}(x) = (x - a - n)\mathcal{C}_n(x) - an\mathcal{C}_{n-1}(x) \quad (3-8)$$

where

$$\mathcal{C}_n(x) := (-1)^n a^n C_n(x; a) \quad (3-9)$$

with

$$C_n(x; a) := {}_2F_0\left(\begin{matrix} -n, -x \\ - \end{matrix} \middle| -\frac{1}{a}\right).$$

Therefore, weighing (3-8) against (3-1), we see that

$$a_1 = -a, \quad a_2 = \frac{a^2}{2!}, \quad h_2 = -1/2, \quad h_3 = 1/3, \quad (3-10)$$

and we conclude that $\{(-1)^n a^n C_n(x; a)\}_{n=0}^{\infty}$ is a Sheffer A-type 0 orthogonal set.

The Meixner polynomials. The monic three-term recurrence relation for the Meixner polynomials is

$$\begin{aligned} &\mathcal{M}_{n+1}(x) \\ &= \left(x + \frac{c\beta}{c-1} + \frac{c+1}{c-1}n\right)\mathcal{M}_n(x) - \left(\frac{\beta-1}{(c-1)^2}cn + \frac{c}{(c-1)^2}n^2\right)\mathcal{M}_{n-1}(x), \end{aligned} \quad (3-11)$$

where

$$\mathcal{M}_n(x) := (\beta)_n \left(\frac{c}{c-1}\right)_n^{nM}(x; \beta, c), \quad (3-12)$$

with

$$M_n(x; \beta, c) := {}_2F_1\left(\begin{matrix} -n, -x \\ \beta \end{matrix} \middle| 1 - \frac{1}{c}\right).$$

Then, from comparing (3-11) with (3-1) we arrive at

$$a_1 = \frac{c\beta}{c-1}, \quad a_2 = \frac{c^2\beta(\beta+1)}{2(c-1)^2}, \quad h_2 = \frac{c+1}{2(c-1)}, \quad h_3 = \frac{1+c+c^2}{3(c-1)^2}. \quad (3-13)$$

Hence, we have shown that $\{c^n(\beta)_n/(c-1)^n M_n(x; \beta, c)\}_{n=0}^{\infty}$ is a Sheffer A-type 0 orthogonal set.

The Meixner–Pollaczek polynomials. The Meixner–Pollaczek polynomials have the monic three-term recurrence relation

$$\mathcal{P}_{n+1}(x) = \left(x + \frac{\lambda}{\tan \phi} + \frac{n}{\tan \phi}\right)\mathcal{P}_n(x) - \left(\frac{2\lambda-1}{4\sin^2 \phi}n + \frac{n^2}{4\sin^2 \phi}\right)\mathcal{P}_{n-1}(x), \quad (3-14)$$

where

$$\mathcal{P}_n(x) := \frac{n!}{(2\sin \phi)^n} P_n^{(\lambda)}(x; \phi), \quad (3-15)$$

with

$$P_n^{(\lambda)}(x; \phi) := \frac{(2\lambda)_n}{n!} e^{in\phi} {}_2F_1\left(\begin{matrix} -n, \lambda + ix \\ 2\lambda \end{matrix} \middle| 1 - e^{-2i\phi}\right), \quad \lambda > 0 \quad \text{and} \quad \phi \in (0, \pi).$$

After comparing (3-14) with (3-1), we obtain

$$\begin{aligned} a_1 &= \lambda \cot \phi, & a_2 &= \frac{4 \cos^2 \phi \lambda (\lambda + 1) - 2\lambda}{8 \sin^2 \phi}, \\ h_2 &= \frac{1}{2} \cot \phi, & h_3 &= \frac{1}{4} \cot^2 \phi - \frac{1}{12}. \end{aligned} \quad (3-16)$$

Thus, $\{(2 \sin \phi)^{-n} n! P_n^{(\lambda)}(x; \phi)\}_{n=0}^{\infty}$ is a Sheffer A-type 0 orthogonal set.

The Krawtchouk polynomials. The Krawtchouk polynomials have

$$\begin{aligned} \mathcal{K}_{n+1}(x) &= [x - pN + (2p - 1)n] \mathcal{K}_n(x) \\ &\quad - [p(1 - p)(N + 1)n - p(1 - p)n^2] \mathcal{K}_{n-1}(x) \end{aligned} \quad (3-17)$$

as a monic recurrence relation, where

$$\mathcal{K}_n(x) := (-N)_n p^n K_n(x; p, N) \quad (3-18)$$

with

$$K_n(x; p, N) := {}_2F_1\left(\begin{matrix} -n, -x \\ -N \end{matrix} \middle| \frac{1}{p}\right), \quad n = 0, 1, 2, \dots, N.$$

After equating the recursion coefficients in (3-17) with those in (3-1) it follows that

$$a_1 = -Np, \quad a_2 = \frac{1}{2}(N - 1)Np^2, \quad h_2 = p - \frac{1}{2}, \quad h_3 = p^2 - p + \frac{1}{3} \quad (3-19)$$

and therefore $\{(-N)_n p^n K_n(x; p, N)\}_{n=0}^{\infty}$ is a Sheffer A-type 0 orthogonal set.

Hence, we have now established the theorem. \square

4. Verification of parameters via generating function expansion

Here, we supplement the analysis of the previous two sections by implementing a procedure for discovering the a_1 , a_2 , h_2 and h_3 parameters for each of the Sheffer A-type 0 orthogonal polynomials obtained in Section 3 by using their corresponding generating functions. This analysis yields explicit power series expansions for $A(t)$ and $H(t)$ in (1-7) for each of the A-type 0 orthogonal sets.

The method used throughout this section is as follows. Momentarily, let us assume that $P_n(x)$ is a Sheffer A-type 0 orthogonal polynomial and $p_n(x)$ is its corresponding monic form. Then notice via the proof of Theorem 3.1 that these polynomials must be related in the following way

$$p_n(x) = a_n b^n P_n(x), \quad (4-1)$$

where b is a polynomial parameter, a function of a polynomial parameter, or a constant and a_n is a sequence in n . Furthermore, let us assume that $\{P_n(x)\}_{n=0}^{\infty}$ has a linear generating function of the form

$$\sum_{\Lambda} c_n P_n(x) t^n = F(x, t).$$

Then, we uniquely determine d_n such that $a_n d_n = c_n$, multiply (4-1) by $d_n t^n$ and sum over Λ to obtain

$$\sum_{\Lambda} d_n p_n(x) t^n = \sum_{\Lambda} c_n P_n(x) (bt)^n = F(x, bt). \quad (4-2)$$

The relation (4-2) is a generating function for $\{p_n(x)\}_{n=0}^{\infty}$. Simply stated, we see that it was achieved via the transformation $t \mapsto bt$ of the generating function for $\{P_n(x)\}_{n=0}^{\infty}$. After deriving a relation of the form (4-2), we then can determine $A(t)$ and $H(t)$ and construct their Maclaurin series expansions, from which we can deduce a_1, a_2, h_2 and h_3 and compare them accordingly with those of Section 3.

The Laguerre polynomials. Multiplying relation (3-3) by $t^n/n!$ and summing for $n = 0, 1, 2, \dots$ gives

$$\sum_{n=0}^{\infty} \frac{1}{n!} \mathcal{L}_n^{(\alpha)}(x) t^n = \sum_{n=0}^{\infty} L_n^{(\alpha)}(x) (-t)^n = (1+t)^{-(\alpha+1)} \exp\left(\frac{xt}{1+t}\right)$$

via (1-15). This yields the following relations for $A(t)$ and $H(t)$:

$$\begin{aligned} A(t) &= (1+t)^{-(\alpha+1)} = \sum_{n=0}^{\infty} \frac{(-1)^n (\alpha+1)_n}{n!} t^n \\ &= 1 - (\alpha+1)t + \frac{1}{2}(\alpha+1)(\alpha+2)t^2 + \dots, \end{aligned}$$

and

$$H(t) = \frac{t}{1+t} = \sum_{n=1}^{\infty} (-1)^{n+1} t^n = t - t^2 + t^3 + \dots.$$

Thus, we see that a_1, a_2, h_2 and h_3 above correspond exactly with those in (3-4).

The Hermite polynomials. We multiply the relation (3-6) by $t^n/n!$, sum for $n = 0, 1, 2, \dots$ and then utilize (1-16) to obtain

$$\sum_{n=0}^{\infty} \frac{1}{n!} \mathcal{H}_n(x) t^n = \sum_{n=0}^{\infty} \frac{1}{n!} H_n(x) \left(\frac{1}{2}t\right)^n = \exp\left(xt - \frac{1}{4}t^2\right).$$

Upon writing this relation in the form $A(t)e^{xH(t)}$, we see that

$$\sum_{n=0}^{\infty} \frac{1}{n!} \mathcal{H}_n(x) t^n = \exp(-\frac{1}{4}t^2) \exp(xt),$$

which gives the following expressions for $A(t)$ and $H(t)$:

$$A(t) = \exp(-\frac{1}{4}t^2) = \sum_{n=0}^{\infty} \frac{(-1)^n t^{2n}}{2^{2n} n!} = 1 - \frac{1}{4}t^2 + \dots, \quad H(t) = t.$$

Hence, we realize that a_1, a_2, h_2 and h_3 are exactly the same as those in (3-7).

The Charlier polynomials. By multiplying the relation (3-9) by $t^n/n!$, summing for $n = 0, 1, 2, \dots$ and then using (1-17), we have

$$\sum_{n=0}^{\infty} \frac{1}{n!} \mathcal{C}_n(x; a) t^n = \sum_{n=0}^{\infty} \frac{1}{n!} C_n(x; a) (-at)^n = e^{-at} (1+t)^x.$$

We then put this result in the form $A(t)e^{xH(t)}$:

$$\sum_{n=0}^{\infty} \frac{1}{n!} \mathcal{C}_n(x; a) t^n = e^{-at} e^{x \ln(1+t)},$$

which leads to the relations for $A(t)$ and $H(t)$

$$A(t) = e^{-at} = \sum_{n=0}^{\infty} \frac{(-a)^n t^n}{n!} = 1 - at + \frac{a^2}{2!} t^2 + \dots,$$

$$H(t) = \ln(1+t) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1} t^n}{n} = t - \frac{1}{2}t^2 + \frac{1}{3}t^3 + \dots,$$

and we observe that a_1, a_2, h_2 and h_3 above are indiscernible from those in (3-10).

The Meixner polynomials. We multiply (3-12) by $t^n/n!$, sum for $n = 0, 1, 2, \dots$ and then use (1-18), which gives

$$\begin{aligned} \sum_{n=0}^{\infty} \frac{1}{n!} \mathcal{M}(x; \beta, c) t^n &= \sum_{n=0}^{\infty} \frac{(\beta)_n}{n!} M(x; \beta, c) \left(\frac{ct}{c-1} \right)^n \\ &= \left(1 - \frac{t}{c-1} \right)^x \left(1 - \frac{ct}{c-1} \right)^{-(x+\beta)}. \end{aligned}$$

Rewriting this result in the form $A(t)e^{xH(t)}$, we see that

$$\sum_{n=0}^{\infty} \frac{1}{n!} \mathcal{M}(x; \beta, c) t^n = (1 - ct/(c-1))^{-\beta} \exp\left(x \ln\left(\frac{c-1-t}{c-1-ct}\right)\right),$$

which in turn gives the following relations for $A(t)$ and $H(t)$:

$$\begin{aligned} A(t) &= (1 - ct/(c-1))^{-\beta} \\ &= \sum_{n=0}^{\infty} \frac{(\beta)_n c^n}{n!(c-1)^n} t^n = 1 + \frac{c\beta}{c-1} t + \frac{c^2\beta(\beta+1)}{2(c-1)^2} t^2 + \dots, \\ H(t) &= \ln(c-1-t) - \ln(c-1-ct) \\ &= \sum_{n=1}^{\infty} \frac{c^n - 1}{n(c-1)^n} t^n = t + \frac{c+1}{2(c-1)} t^2 + \frac{1+c+c^2}{3(c-1)^2} t^3 + \dots, \end{aligned}$$

and hence a_1, a_2, h_2 and h_3 are identical to those in (3-13).

The Meixner–Pollaczek polynomials. We multiply the relation (3-15) by $t^n/n!$ and sum for $n = 0, 1, 2, \dots$. We then use (1-19), and obtain

$$\begin{aligned} \sum_{n=0}^{\infty} \frac{1}{n!} \mathcal{P}_n(x; \phi) t^n \\ = \sum_{n=0}^{\infty} P_n(x; \phi) \left(\frac{t}{2 \sin \phi} \right)^n = \left(1 - \frac{e^{i\phi t}}{2 \sin \phi} \right)^{-\lambda + ix} \left(1 - \frac{e^{-i\phi t}}{2 \sin \phi} \right)^{-\lambda - ix}. \end{aligned}$$

Rewriting this result in the form $A(t)e^{xH(t)}$, we have

$$\begin{aligned} \sum_{n=0}^{\infty} \frac{1}{n!} \mathcal{P}_n(x; a) t^n \\ = \left[\left(1 - \frac{e^{i\phi t}}{2 \sin \phi} \right) \left(1 - \frac{e^{-i\phi t}}{2 \sin \phi} \right) \right]^{-\lambda} \exp \left[x \ln \left(\left(\frac{1 - e^{i\phi t}/(2 \sin \phi)}{1 - e^{-i\phi t}/(2 \sin \phi)} \right)^i \right) \right], \end{aligned}$$

which leads to $A(t)$ and $H(t)$ below:

$$\begin{aligned} A(t) &= \left[\left(1 - \frac{e^{i\phi t}}{2 \sin \phi} \right) \left(1 - \frac{e^{-i\phi t}}{2 \sin \phi} \right) \right]^{-\lambda} = \sum_{n=0}^{\infty} \left[\sum_{k=0}^n \frac{(\lambda)_k (\lambda)_{n-k} e^{i(n-2k)\phi}}{2^n k! (n-k)! \sin^n \phi} \right] t^n \\ &= 1 + \lambda \cot \phi t + \left(\frac{4 \cos^2 \phi \lambda (\lambda + 1) - 2\lambda}{8 \sin^2 \phi} \right) t^2 + \dots, \\ H(t) &= i \left[\ln(1 - e^{i\phi t}/(2 \sin \phi)) - \ln(1 - e^{-i\phi t}/(2 \sin \phi)) \right] \\ &= \sum_{n=1}^{\infty} \frac{\sin(n\phi)}{2^{n-1} n \sin^n \phi} t^n = t + \frac{1}{2} \cot \phi t^2 + \left(\frac{1}{4} \cot^2 \phi - \frac{1}{12} \right) t^3 + \dots, \end{aligned}$$

and the a_1, a_2, h_2 and h_3 above are the same as those in (3-16).

The Krawtchouk polynomials. Finally, we multiply (3-18) by $t^n/n!$, sum for $n = 0, 1, 2, \dots$, use the fact that

$$\frac{(-N)_n}{n!} = \frac{(-1)^n N!}{n!(N-n)!} = (-1)^n C(N, n),$$

and (1-20) in order to obtain

$$\begin{aligned} \sum_{n=0}^{\infty} \frac{1}{n!} \mathcal{K}_n(x; p, N) t^n &= \sum_{n=0}^{\infty} \frac{(-N)_n p^n}{n!} K_n(x; p, N) t^n \\ &= \sum_{n=0}^{\infty} C(N, n) K_n(x; p, N) (-pt)^n \\ &= (1 + (1-p)t)^x (1-pt)^{N-x}. \end{aligned}$$

We write this result in the form $A(t)e^{xH(t)}$:

$$\sum_{n=0}^{\infty} \frac{1}{n!} \mathcal{K}_n(x; p, N) t^n = (1-pt)^N \exp\left(x \ln\left(\frac{1+(1-p)t}{1-pt}\right)\right).$$

Then, $A(t)$ and $H(t)$ are

$$A(t) = (1-pt)^N = \sum_{n=0}^{\infty} \frac{(-N)_n p^n}{n!} t^n = 1 - Npt + \frac{1}{2}(N-1)Np^2t^2 + \dots,$$

$$\begin{aligned} H(t) &= \ln(1 + (1-p)t) - \ln(1-pt) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}(1-p)^n + p^n}{n} t^n \\ &= t + \left(p - \frac{1}{2}\right)t^2 + \left(p^2 - p + \frac{1}{3}\right)t^3 + \dots \end{aligned}$$

and our a_1, a_2, h_2 and h_3 above correspond exactly to those in (3-19).

5. A proof for sufficiency: the “inverse method”

We have thus far established necessary conditions for the A-type 0 recursion coefficients (2-6) and (2-7). We next show that these conditions are also sufficient and thus achieve a complete characterization of all of the A-type 0 orthogonal sets. Namely, we prove that given (3-1), (1-6) must follow. We call our approach the “inverse method”, which is a procedure for obtaining a linear generating function from a three-term recurrence relation and therefore reverses the analysis conducted in Sections 2 and 3. The method is as follows.

Assume $\{P_n(x)\}_{n=0}^{\infty}$ is a polynomial set that satisfies a three-term recurrence relation of the form (1-4). We first multiply this relation by $c_n t^n$, where c_n is a certain function in n that is independent of x and t , and sum for $n = 0, 1, 2, \dots$. Then, from the assignment $F(t; x) := \sum_{n=0}^{\infty} c_n P_n(x) t^n$, we obtain a first-order

differential equation in t , with x regarded as a parameter. The initial condition for this equation is $F(0; x) = 1$, via the initial condition $P_0(x) = 1$ in (1-4). The existence and uniqueness of the solution to this differential equation are guaranteed, and the solution will be a generating function for the set $\{P_n(x)\}_{n=0}^{\infty}$.

We now apply the inverse method to each of the unrestricted three-term recurrence relations of our A-type 0 orthogonal sets — as a byproduct, additional fundamental characterizations (differential equations) are obtained for our generating functions (1-15)–(1-20). To derive each of these relations, we first substitute (1-10) into (1-9), which leads to

$$P_{n+1}(x) = xP_n(x) + (\alpha + b + bn)P_n(x) - ((c + d)n + dn^2)P_{n-1}(x).$$

Here, we use (1-10) as opposed to (2-6) and (2-7) for ease of notation. Now define $P_n(x) := e_n Q_n(x)$, and note that our relation directly above becomes

$$e_{n+1}Q_{n+1}(x) = xe_nQ_n(x) + (\alpha + b + bn)e_nQ_n(x) - ((c + d)n + dn^2)e_{n-1}Q_{n-1}.$$

Taking $e_n := n!$ and dividing both sides by $n!$, we have

$$(n + 1)Q_{n+1}(x) = xQ_n(x) + (\alpha + b + bn)Q_n(x) - (c + d + dn)Q_{n-1}(x), \quad (5-1)$$

which is the unrestricted three-term recurrence relation for the Sheffer A-type 0 orthogonal polynomials. We apply Corollary 2.4 accordingly to determine the recurrence coefficients for each case.

We begin by writing out the rigorous details for the Laguerre case. For the subsequent cases, we outline only the salient details. In these cases, we first display the unrestricted three-term recurrence relation, which we henceforth call UTTRR. Then, we display the c_n and the corresponding definition of F . Finally, we write the resulting differential equation (labeled DE) and its unique solution, which will be the corresponding Sheffer A-type 0 generating function.

The Laguerre polynomials. Using (3-4), Corollary 2.4 and (5-1), we obtain

$$(n + 1)L_{n+1}^{(\alpha)}(x) - (2n + \alpha + 1 - x)L_n^{(\alpha)}(x) + (n + \alpha)L_{n-1}^{(\alpha)}(x) = 0.$$

We next multiply both sides of this relation by t^n ($c_n \equiv 1$) and sum for $n = 0, 1, 2, \dots$, which yields

$$\begin{aligned} \sum_{n=0}^{\infty} (n + 1)L_{n+1}^{(\alpha)}(x)t^n - 2 \sum_{n=1}^{\infty} nL_n^{(\alpha)}(x)t^n - (\alpha + 1 - x) \sum_{n=0}^{\infty} L_n^{(\alpha)}(x)t^n \\ + \sum_{n=1}^{\infty} nL_{n-1}^{(\alpha)}(x)t^n + \alpha \sum_{n=0}^{\infty} L_{n-1}^{(\alpha)}(x)t^n = 0. \quad (5-2) \end{aligned}$$

We next assign $F := F(t; x) := \sum_{n=0}^{\infty} L_n^{(\alpha)}(x)t^n$, accounting for the fact that $\dot{F}(t; x) = \sum_{n=1}^{\infty} nL_n^{(\alpha)}(x)t^{n-1}$ by $\dot{F} := (\partial/\partial t)F(t; x)$. Recalling that $L_{-1}^{(\alpha)}(x) = 0$ from (1-4), we see that (5-2) becomes

$$\dot{F} - 2t\dot{F} - (\alpha + 1 - x)F + \sum_{n=2}^{\infty} nL_{n-1}^{(\alpha)}(x)t^n + \alpha tF = 0 \quad (5-3)$$

and also observe that

$$\sum_{n=1}^{\infty} nL_{n-1}^{(\alpha)}(x)t^n = \sum_{n=2}^{\infty} (n-1)L_{n-1}^{(\alpha)}(x)t^n + \sum_{n=1}^{\infty} L_{n-1}^{(\alpha)}(x)t^n = t^2\dot{F} + tF.$$

Then, we can put (5-3) in standard form:

$$\dot{F} + \left[\frac{x + (\alpha + 1)(t - 1)}{1 - 2t + t^2} \right] F = 0; \quad F(0; x) = 1. \quad (5-4)$$

The integrating factor in (5-4) turns out to be

$$\mu = \exp \left[\int \frac{x + (\alpha + 1)(t - 1)}{1 - 2t + t^2} dt \right]$$

and, through partial fraction decomposition, we attain the general solution

$$F(t; x) = c(x, \alpha)(t - 1)^{-(\alpha+1)} \exp\left(\frac{x}{t-1}\right).$$

Therefore, using our initial condition in (5-4) to determine $c(x, \alpha)$, we establish the solution

$$F(t; x) = \sum_{n=0}^{\infty} L_n^{(\alpha)}(x)t^n = (t - 1)^{-(\alpha+1)} \exp\left(\frac{xt}{t-1}\right), \quad (5-5)$$

which is the Sheffer A-type 0 generating function for the Laguerre polynomials.

The Hermite polynomials.

$$\text{UTTR: } H_{n+1}(x) = 2xH_n(x) - 2nH_{n-1}(x)$$

$$c_n: 1/n!$$

$$F: F(t; x) := \sum_{n=0}^{\infty} (1/n!)H_n(x)t^n$$

$$\text{DE: } \dot{F} - 2(x-t)F = 0; \quad F(0; x) = 1$$

$$\text{Solution: } F(t; x) = \sum_{n=0}^{\infty} (1/n!)H_n(x)t^n = \exp(2xt - t^2)$$

The Charlier polynomials.

$$\begin{aligned} \text{UTTRR: } & -xC_n(x; a) = aC_{n+1}(x; a) - (n+a)C_n(x; a) + nC_{n-1}(x; a) \\ c_n: & 1/n! \\ F: & F(t; x, a) := \sum_{n=0}^{\infty} (1/n!)C_n(x; a)t^n \\ \text{DE: } & \dot{F} - (1+x/(t-a))F = 0; \quad F(0; x, a) = 1 \\ \text{Solution: } & F(t; x, a) = \sum_{n=0}^{\infty} (1/n!)C_n(x; a)t^n = e^t(1-t/a)^x \end{aligned}$$

The Meixner polynomials.

$$\begin{aligned} \text{UTTRR: } & (c-1)xM_n(x; \beta, c) = \\ & c(\beta+n)M_{n+1}(x; \beta, c) - [n+c(\beta+n)]M_n(x; \beta, c) + nM_{n-1}(x; \beta, c) \\ c_n: & (\beta)_n 1/n! \\ F: & F(t; x, \beta, c) := \sum_{n=0}^{\infty} (\beta)_n / (n!) M_n(x, \beta, c) t^n \\ \text{DE: } & \dot{F} + \left(\frac{(c-1)x + (c-t)\beta}{(1+c-t)t-c} \right) F = 0; \quad F(0; x, \beta, c) = 1 \\ \text{Solution: } & F(t; x, \beta, c) = (1-t/c)^x (1-t)^{-(x+\beta)} \end{aligned}$$

Remark 5.1. For establishing this differential equation, we made use of the identity $(\beta)_n = (\beta)_{n-1}(\beta+n-1)$.

The Meixner–Pollaczek polynomials.

$$\begin{aligned} \text{UTTRR: } & (n+1)P_{n+1}^{(\lambda)}(x; \phi) - 2[x \sin \phi + (n+\lambda) \cos \phi]P_n^{(\lambda)}(x; \phi) \\ & + (n+2\lambda-1)P_{n-1}^{(\lambda)}(x; \phi) = 0 \\ c_n: & 1 \\ F: & F(t; x, \lambda, \phi) := \sum_{n=0}^{\infty} P_n^{(\lambda)}(x; \phi) t^n \\ \text{DE: } & \dot{F} + 2 \left(\frac{\lambda(t-\cos \phi) - x \sin \phi}{1-2 \cos \phi t + t^2} \right) F = 0; \quad F(0; x, \lambda, \phi) = 1 \\ \text{Solution: } & F(t; x, \lambda, \phi) = \sum_{n=0}^{\infty} P_n^{(\lambda)}(x; \phi) t^n = (1-e^{i\phi}t)^{-\lambda+ix} (1-e^{-i\phi}t)^{-\lambda-ix} \end{aligned}$$

The Krawtchouk polynomials.

$$\begin{aligned} \text{UTTRR: } & -xK_n(x; P, N) = p(N-n)K_{n+1}(x; P, N) \\ & - [p(N-n) + n(1-p)]K_n(x; P, N) + n(1-p)K_{n-1}(x; P, N) \\ c_n: & \binom{N}{n} \\ F: & F(t; x, p, N) := \sum_{n=0}^N \binom{N}{n} K_n(x, p, N) t^n \\ \text{DE: } & \dot{F} + \left(\frac{x/(p+tp-t) - N}{1+t} \right) F = 0; \quad F(0; x, p, N) = 1 \\ \text{Solution: } & F(t; x, p, N) = \sum_{n=0}^N \binom{N}{n} K_n(x, p, N) t^n \\ & = (1 - ((1-p)/p)t)^x (1+t)^{N-x} \end{aligned}$$

We now have the following statement:

Theorem 5.2. *Given the monic recursion coefficients corresponding to each of the A-type 0 orthogonal sets of Laguerre, Hermite, Charlier, Meixner, Meixner–Pollaczek and Krawtchouk, there exists a generating function of the form (1-6).*

Hence, Theorem 2.3 in conjunction with Theorem 5.2 establishes the following culminating statement:

Theorem 5.3. *A necessary and sufficient condition for $\{P_n(x)\}_{n=0}^{\infty}$ to be a Sheffer A-type 0 orthogonal set is that the monic recursion coefficients λ_{n+1} and μ_{n+1} , as respectively in (2-6) and (2-7), have the form*

$$\lambda_{n+1} = c_1 + c_2n \quad \text{and} \quad \mu_{n+1} = c_3n + c_4n^2, \quad c_1, \dots, c_4 \in \mathbb{R},$$

with $\mu_{n+1} > 0$.

Finally, we mention that this paper solves Problem 1 in Section 3.9 of [Galiffa 2013].

References

- [Akleyek et al. 2010] S. Akleyek, M. Cenk, and F. Özbudak, “Polynomial multiplication over binary fields using Charlier polynomial representation with low space complexity”, pp. 227–237 in *Progress in cryptology—INDOCRYPT 2010*, Lecture Notes in Comput. Sci. **6498**, Springer, Berlin, 2010. MR 2012h:94138 Zbl 1253.94040
- [Al-Salam 1990] W. A. Al-Salam, “Characterization theorems for orthogonal polynomials”, pp. 1–24 in *Orthogonal polynomials: Theory and Practice* (Columbus, OH, 1989), NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci. **294**, Kluwer Acad. Publ., Dordrecht, 1990. MR 92g:42011 Zbl 0704.42021
- [Al-Salam and Verma 1970] W. A. Al-Salam and A. Verma, “Generalized Sheffer polynomials”, *Duke Math. J.* **37** (1970), 361–365. MR 41 #7175 Zbl 0205.07603
- [Chen et al. 2011] X. Chen, Z. Chen, Y. Yu, and D. Su, “An unconditionally stable radial point interpolation meshless method with Laguerre polynomials”, *IEEE Trans. Antennas and Propagation* **59**:10 (2011), 3756–3763. MR 2893572
- [Coffey 2011] M. W. Coffey, “On finite sums of Laguerre polynomials”, *Rocky Mountain J. Math.* **41**:1 (2011), 79–93. MR 2012k:33020 Zbl 1214.33004
- [Coulembier et al. 2011] K. Coulembier, H. De Bie, and F. Sommen, “Orthogonality of Hermite polynomials in superspace and Mehler type formulae”, *Proc. Lond. Math. Soc.* (3) **103**:5 (2011), 786–825. MR 2852289 Zbl 1233.33004
- [Di Bucchianico 1994] A. Di Bucchianico, “Representations of Sheffer polynomials”, *Stud. Appl. Math.* **93**:1 (1994), 1–14. MR 95h:60131 Zbl 0818.33010
- [Di Bucchianico and Loeb 1994] A. Di Bucchianico and D. Loeb, “A simpler characterization of Sheffer polynomials”, *Stud. Appl. Math.* **92**:1 (1994), 1–15. MR 95h:05016 Zbl 0795.05018
- [Di Nardo et al. 2011] E. Di Nardo, H. Niederhausen, and D. Senato, “A symbolic handling of Sheffer polynomials”, *Ann. Mat. Pura Appl.* (4) **190**:3 (2011), 489–506. MR 2012h:05041 Zbl 1292.05052
- [Dominici 2007] D. Dominici, “Some remarks on a paper by L. Carlitz”, *J. Comput. Appl. Math.* **198**:1 (2007), 129–142. MR 2007g:33012 Zbl 1115.33007

- [Dueñas and Marcellán 2011] H. Dueñas and F. Marcellán, “The holonomic equation of the Laguerre–Sobolev-type orthogonal polynomials: a non-diagonal case”, *J. Difference Equ. Appl.* **17**:6 (2011), 877–887. MR 2012f:42046 Zbl 1219.42012
- [Ferreira et al. 2008] C. Ferreira, J. L. López, and P. J. Pagola, “Asymptotic approximations between the Hahn-type polynomials and Hermite, Laguerre and Charlier polynomials”, *Acta Appl. Math.* **103**:3 (2008), 235–252. MR 2010c:33024 Zbl 1168.33309
- [Galiffa 2013] D. J. Galiffa, *On the higher-order Sheffer orthogonal polynomial sequences*, Springer, New York, 2013. MR 3013644 Zbl 1262.33011
- [Hofbauer 1981] J. Hofbauer, “A representation of Sheffer polynomials in terms of a differential equation for their generating functions”, *Aequationes Math.* **23**:2-3 (1981), 156–168. MR 84k:33014 Zbl 0505.33011
- [Hutník 2011] O. Hutník, “Wavelets from Laguerre polynomials and Toeplitz-type operators”, *Integral Equations Operator Theory* **71**:3 (2011), 357–388. MR 2852192 Zbl 06048404
- [Ismail 2009] M. E. H. Ismail, *Classical and quantum orthogonal polynomials in one variable*, Encyclopedia of Mathematics and its Applications **98**, Cambridge University Press, Cambridge, 2009. MR 2010i:33001 Zbl 1172.42008
- [Khan et al. 2011] M. A. Khan, A. H. Khan, and M. Singh, “Integral representations for the product of Krawtchouk, Meixner, Charlier and Gottlieb polynomials”, *Int. J. Math. Anal. (Ruse)* **5**:1-4 (2011), 199–206. MR 2776557 Zbl 1235.42022
- [Koekoek and Swarttouw 1996] R. Koekoek and R. F. Swarttouw, “The Askey-scheme of hypergeometric orthogonal polynomials and its q -analogue”, Reports of the Faculty of Technical Mathematics and Information, No. 98–17, Delft University of Technology, 1996, Available at <http://aw.twi.tudelft.nl/~koekoek/askey/>.
- [Kuznetsov 2008] A. Kuznetsov, “Expansion of the Riemann Ξ function in Meixner–Pollaczek polynomials”, *Canad. Math. Bull.* **51**:4 (2008), 561–569. MR 2010a:11165 Zbl 1173.41003
- [Meixner 1934] J. Meixner, “Orthogonale Polynomsysteme Mit Einer Besonderen Gestalt Der Erzeugenden Funktion”, *J. London Math. Soc.* **S1-9**:1 (1934), 6–13. MR 1574715 Zbl 0008.16205
- [Miki et al. 2011] H. Miki, L. Vinet, and A. Zhedanov, “Non-Hermitian oscillator Hamiltonians and multiple Charlier polynomials”, *Phys. Lett. A* **376**:2 (2011), 65–69. MR 2859302 Zbl 1255.81143
- [Mouayn 2010] Z. Mouayn, “A new class of coherent states with Meixner–Pollaczek polynomials for the Gol’dman–Krivchenkov Hamiltonian”, *J. Phys. A* **43**:29 (2010), 295201. MR 2011m:81163 Zbl 1193.81051
- [Popa 1997] E. C. Popa, “Note on Sheffer polynomials”, *Octagon Math. Mag.* **5**:2 (1997), 56–57. MR 1619516
- [Popa 1998] E. C. Popa, “On the Sheffer polynomials”, *Bul. Ştiinţ. Univ. Politeh. Timiş. Ser. Mat. Fiz.* **43(57)**:1 (1998), 21–23. MR 99k:33032 Zbl 0974.44006
- [Rainville 1960] E. D. Rainville, *Special functions*, Macmillan, New York, 1960. MR 21 #6447 Zbl 0092.06503
- [Sheffer 1939] I. M. Sheffer, “Some properties of polynomial sets of type zero”, *Duke Math. J.* **5** (1939), 590–622. MR 1,15c Zbl 0022.01502
- [Sheffer 1941] I. M. Sheffer, “Some applications of certain polynomial classes”, *Bull. Amer. Math. Soc.* **47** (1941), 885–898. MR 3,111f Zbl 0027.39503
- [Shukla and Rapeli 2011] A. K. Shukla and S. J. Rapeli, “An extension of Sheffer polynomials”, *Proyecciones* **30**:2 (2011), 265–275. MR 2852353 Zbl 1247.33025

- [Vignat 2011] C. Vignat, “Old and new results about relativistic Hermite polynomials”, *J. Math. Phys.* **52**:9 (2011), 093503. MR 2012i:33019 Zbl 1272.33012
- [Wang and Wong 2011] X.-S. Wang and R. Wong, “Global asymptotics of the Meixner polynomials”, *Asymptot. Anal.* **75**:3-4 (2011), 211–231. MR 2012k:33026 Zbl 1252.33009
- [Wang et al. 2011] J. Wang, W. Qiu, and R. Wong, “Uniform asymptotics for Meixner–Pollaczek polynomials with varying parameters”, *C. R. Math. Acad. Sci. Paris* **349**:19-20 (2011), 1031–1035. MR 2012h:33017 Zbl 1231.33015
- [Yalçınbaş et al. 2011] S. Yalçınbaş, M. Aynigül, and M. Sezer, “A collocation method using Hermite polynomials for approximate solution of pantograph equations”, *J. Franklin Inst.* **348**:6 (2011), 1128–1139. MR 2012f:65115 Zbl 1221.65187

Received: 2012-08-20

Revised: 2013-05-07

Accepted: 2013-12-27

djg34@psu.edu

*Department of Mathematics, Penn State Erie, The Behrend
College, Erie, PA 16563, United States*

tnr5033@psu.edu

*Department of Mathematics, Penn State Erie, The Behrend
College, Erie, PA 16563, United States*

Average reductions between random tree pairs

Sean Cleary, John Passaro and Yasser Toruno

(Communicated by Robert W. Robinson)

There are a number of measures of degrees of similarity between rooted binary trees. Many of these ignore sections of the trees which are in complete agreement. We use computational experiments to investigate the statistical characteristics of such a measure of tree similarity for ordered, rooted, binary trees. We generate the trees used in the experiments iteratively, using the Yule process modeled upon speciation.

1. Introduction

Rooted binary trees arise in a wide range of settings, from biological evolutionary trees to efficient structures for searching datasets. There are a number of measures of tree similarity which arise in these settings. Here we investigate a measure which is relevant for ordered, rooted, binary trees of the same size. Examples of trees satisfying such conditions include some binary search trees. Our approach is to consider pairs of such trees of increasing size n , selected via a random process, and investigate the degree of commonality given by a natural measure of the degree to which they agree completely on peripheral subtrees. Using experimental evidence, we find that the degree of commonality appears to grow linearly with tree size, and we estimate the average behavior.

There are a number of processes for selecting trees randomly. One method that is commonly studied is the uniform distribution on trees, where each tree is equally likely to be selected. Some properties of the reduction behavior of trees selected uniformly at random have been investigated by Cleary, Elder, Rechnitzer and Taback [Cleary et al. 2010] while studying statistical properties of Thompson's group F , showing that a tree pair selected from the uniform distribution on tree pairs is almost surely unreduced in the sense described below. The common subtrees investigated here via reduction are a particular case of common edges, where in the

MSC2010: 05C05, 68P05.

Keywords: random binary tree pairs.

Partial funding provided by NSF grants 0811002 and 1417820. Sean Cleary was partially supported by grant 234548 from the Simons Foundation.

common edge case the collections of common edges need not be peripheral. That is, in the more general case they need not include the complete subtree, extending to the leaves. For common edges of all types, the average number of common edges with respect to the uniform selection of trees at random case has been examined experimentally by Chu and Cleary [2013] and asymptotically by Cleary, Reznitzer and Wong [Cleary et al. 2013]. Asymptotically, the expected number of reductions of a tree pair selected uniformly at random is

$$\frac{16 - 5\pi}{\pi}n + \frac{7\pi - 20}{\pi} + O\left(\frac{\log n}{n}\right),$$

for reductions of a more general type, which is about

$$0.092958n + 0.633802 + O\left(\frac{\log n}{n}\right).$$

The experimental results in [Chu and Cleary 2013] show quick convergence to the dominant linear term of $0.092958n$. For the particular subtree peripheral reductions (that is, subtree reductions) considered here, a similar generating function analysis gives the asymptotic number of trees as $(7 - 4\sqrt{3})n$, which is about $0.0717968n$ when tree pairs are selected uniformly at random. So on average more than three quarters of the expected common edges lie in expected common peripheral subtrees.

Here, instead of considering trees selected uniformly at random, we study a process for generating trees at random motivated by biological questions, called the Yule process [Yule 1925; Harding 1971], also known as uniform speciation. A tree is grown iteratively from the root. At each step, a leaf is selected uniformly at random from the leaves present at that stage, and a new sibling pair is attached at that leaf, and then the process is iterated until we have a tree with the appropriate number of leaves. Such a distribution of trees also can arise from a variety of insertion scenarios in tree-structured data.

The distribution of the number of sibling pairs (“cherries”) of unordered trees was investigated by McKenzie and Steel [2000] for both the uniform and Yule tree distributions — asymptotically, there are $n/3$ expected sibling pairs for the Yule distribution and $n/4$ for the uniform distribution. Here we find experimentally that the expected number of subtree reductions is also larger for the Yule distribution than the uniform distribution, with almost 13% expected subtree reduction compared to the expected reduction of about 7% in the uniform case.

2. Background and definitions

We consider rooted binary trees on n leaves with a natural left-to-right order on leaves, numbered from 1 to n . The internal nodes of the trees we refer to as nodes and the external nodes we refer to as leaves. Two children of the same node which

are leaves form a sibling pair and their leaf numbers are necessarily of the form i and $i + 1$ for some i .

A tree pair (S, T) is *reduced* if there are no sibling pairs with leaves numbered i and $i + 1$ in S which have a corresponding sibling pair i and $i + 1$ as leaves in T . An *elementary reduction* for a tree pair (S, T) with n leaves with a common sibling pair $(i, i + 1)$ is a tree pair (S', T') with $n - 1$ leaves, where the common sibling pair has been removed in both S and T and the leaves have been appropriately renumbered. A *reduction* of a tree pair diagram is a sequence of elementary reductions. There may be many possible elementary reductions for an unreduced tree pair and thus many possible reductions, but for a given tree pair (S, T) , there is a unique reduced tree pair (S'', T'') which is itself a reduction of (S, T) and which has the property that any possible sequence of reductions from (S, T) will terminate in that reduced tree pair. An example of tree pair reduction is given in Figure 1.

The subtrees that are eliminated during the reduction process for a tree pair (S, T) are portions of the tree in which S and T agree completely. There are a number of metrics on spaces of trees of interest, coming from biological questions, database efficiency questions and more abstract approaches. For all of the standard metrics on spaces of trees with an order on the leaves, the parts of the trees which are in complete agreement do not contribute to the distance. That is, if a tree pair (S, T) reduces to a tree pair (S', T') , the distance of interest between S and T is the same as the distance between S' and T' . The fact that the trees S' and T' may be considerably smaller is of good use, particularly for distances which are

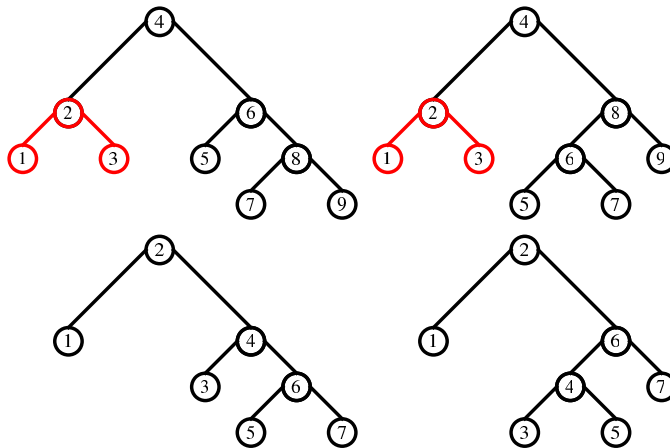


Figure 1. An unreduced tree pair and its reduction to a reduced tree pair. The top unreduced tree pair has a common subtree containing the sibling pair of nodes 1 and 3 in both trees, shown in red, which is then removed and the nodes renumbered, resulting in the lower tree pair which is reduced.

difficult to compute. Given that the best known algorithm for rotation distance is of exponential running time, and that many tree metrics of biological interest are proven to be of class NP, even a marginal reduction in the sizes of trees under consideration is worthwhile. This analysis is an effort to understand the degree to which such reductions typically reduce the size of tree pairs.

We generate trees using the Yule or *speciation* method as follows. We begin with a single node with two leaves, and then randomly select from the leaves and replace that leaf with a node with its own two leaves, renumbering the leaves as needed. We then choose randomly from the three current leaves, replacing that chosen leaf with a node and two leaves, and continue enlarging the tree in this process until it is the desired size.

As shown in Figure 2, there may be more than one way to generate a given tree using the Yule process. The process is generally more likely to generate balanced trees than stringy ones, so the distribution on trees is different than that for the uniform random selection of trees, as described in [Harding 1971]. This is also related to the difference in expected number of sibling pairs described in [McKenzie and Steel 2000].

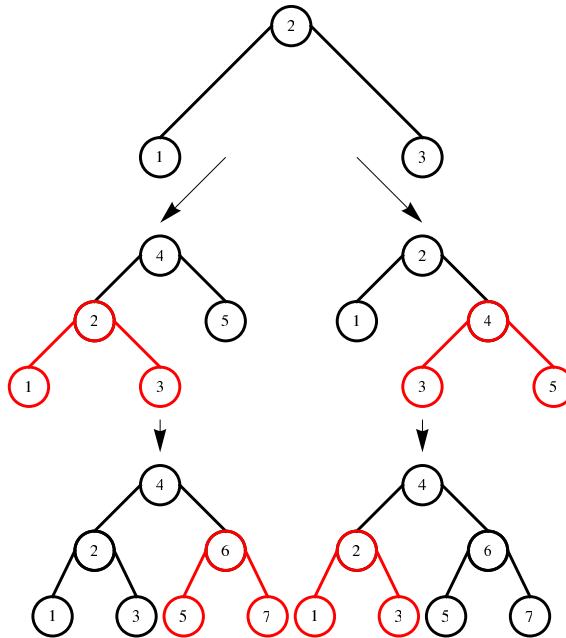


Figure 2. Some trees can be generated in several ways via the Yule process, such as this balanced tree with four leaves which can be generated in two ways. Every other tree with four leaves can be generated in just one way, resulting in a nonuniform distribution of random tree selection.

3. Experiments and conclusions

We constructed programs in C to create tree pairs of a specified size and count the reductions, iterating to obtain average values. Tree pairs with trees ranging from size 100 to 29,000 were generated and the total size of common subtrees was calculated and recorded for each pair generated, with the results summarized in Table 1. Generally, there were around 1000 tree pairs of each size generated and analyzed, sufficient to give small error bars in the analysis. The average reductions grew linearly, with about 12.8% average reduction in size, significantly more than the corresponding value of about 7.1% in the corresponding case for trees generated uniformly at random. As indicated in Figures 3 and 4, the relationship appears to be

| Tree size range | Average total subtree reduction | σ subtree reduction |
|-----------------|---------------------------------|----------------------------|
| 100– 2000 | 0.12846 | 0.013829 |
| 2001– 8000 | 0.12781 | 0.006034 |
| 8001–15000 | 0.12775 | 0.003462 |
| 15001–29000 | 0.12773 | 0.002402 |

Table 1. Average total size of common subtrees and corresponding sample standard deviations.

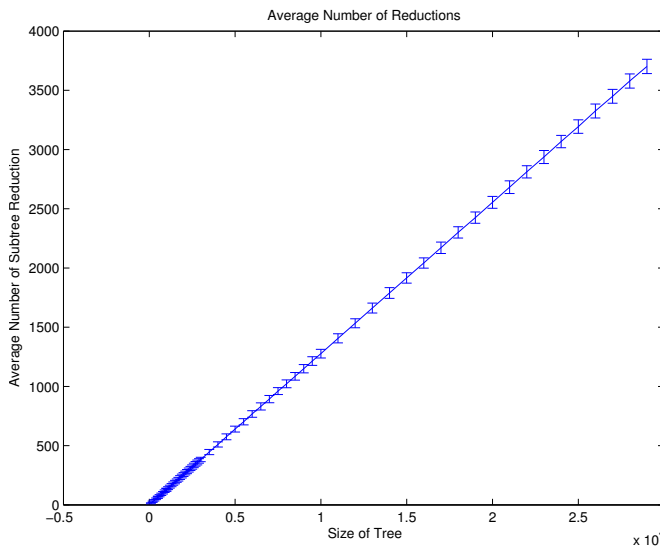


Figure 3. The average number of reductions grows linearly with tree size, with tight error bars from the sample sizes used over this range. The slope of the line of best fit is about 0.127. Error bars indicate 3 standard deviations from the sample averages.

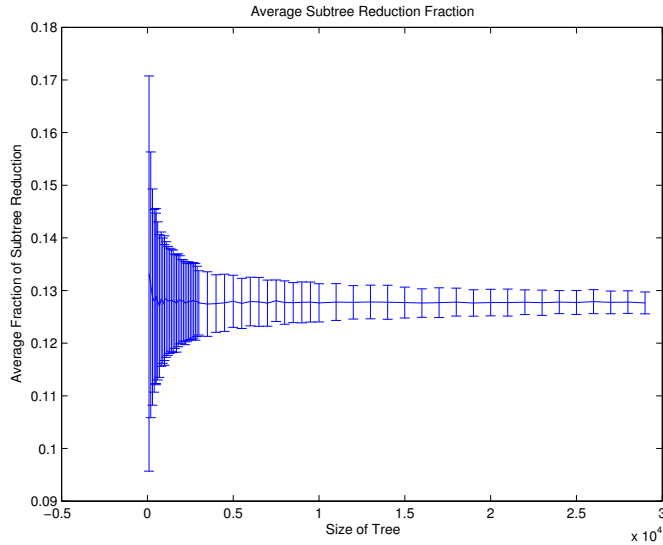


Figure 4. The average fraction of the tree pairs which are eliminated in the reduction process is close to 0.127, over the range shown. Error bars indicate 3 standard deviations from the sample averages.

linear, and a linear regression to the data gives an excellent fit with r^2 value of within one-millionth of 1. The line of best fit for the experimental data is $0.1277n + 0.268$.

What we find is that the fraction of the trees which reduce appears larger for the Yule distribution than for the uniform distribution.

References

- [Chu and Cleary 2013] T. Chu and S. Cleary, “Expected conflicts in pairs of rooted binary trees”, *Involve* **6**:3 (2013), 323–332. MR 3101764 Zbl 1274.05066
- [Cleary et al. 2010] S. Cleary, M. Elder, A. Rechnitzer, and J. Taback, “Random subgroups of Thompson’s group F ”, *Groups Geom. Dyn.* **4**:1 (2010), 91–126. MR 2011e:20062 Zbl 1226.20034
- [Cleary et al. 2013] S. Cleary, A. Rechnitzer, and T. Wong, “Common edges in rooted trees and polygonal triangulations”, *Electron. J. Combin.* **20**:1 (2013), Paper 39, 22. MR 3035049 Zbl 1267.05249
- [Harding 1971] E. F. Harding, “The probabilities of rooted tree-shapes generated by random bifurcation”, *Advances in Appl. Probability* **3** (1971), 44–77. MR 43 #8162 Zbl 0241.92012
- [McKenzie and Steel 2000] A. McKenzie and M. Steel, “Distributions of cherries for two models of trees”, *Math. Biosci.* **164**:1 (2000), 81–92. MR 2001e:92010 Zbl 0947.92021
- [Yule 1925] G. Yule, “A mathematical theory of evolution, based upon the conclusions of Dr. J. C. Willis, F.R.S.”, *Royal Society of London Philosophical Transactions, Series B* **213** (1925), 21–87.

cleary@sci.ccny.cuny.edu *Department of Mathematics,
The City College of New York and the CUNY Graduate Center,
City University of New York, NAC R8133,
160 Convent Avenue, New York, NY 10031, United States*

john.a.passaro@gmail.com *Department of Mathematics, The City College of New York,
City University of New York, New York, NY 10031,
United States*

yltoruno@gmail.com *Department of Computer Science,
The City College of New York, City University of New York,
New York, NY 10031, United States*

Growth functions of finitely generated algebras

Eric Fredette, Dan Kubala, Eric Nelson,
Kelsey Wells and Harold W. Ellingsen, Jr.

(Communicated by Joseph A. Gallian)

We study the growth of finitely presented two-generator monomial algebras. In particular, we seek to improve an upper bound found by the last author. Our search lead us to a connection to de Bruijn graphs and a drastically improved bound.

The growth of algebras has been long studied by algebraists; it goes hand-in-hand with the Gelfand–Kirillov dimension of algebras. An excellent source is [Krause and Lenagan 2000]. Throughout this paper F denotes a field and $0 \in \mathbb{N}$. We focus our work on the growth of algebras of the form $F\langle x, y \rangle / I$, where I is an ideal of the free algebra $F\langle x, y \rangle$ generated by finitely many monomials. Such an algebra is called a *finitely presented two-generator monomial algebra*. It is customary to refer to monomials as words. Let A be one of these algebras. We consider the set \mathcal{B} of all words in x and y that do not have any of the words in the generators for I as factors or subwords. It is standard to show that the image of \mathcal{B} is a basis for A . Instead of referring to images of words, we will view the multiplication on A as follows. For any words u and v in \mathcal{B} , uv is simply uv if uv has no generator of I as a subword, and $uv = 0$ otherwise. We define the *length* of a word to be the number of letters in it, counting repetitions. Now we can define a function $g : \mathbb{N} \rightarrow \mathbb{N}$ by setting $g(n)$ to be the number of words in \mathcal{B} of length at most n . This function g is called a *growth function* for A and the *growth* of A is essentially the type of function g is, such as a polynomial of some degree or an exponential. Let's consider a couple of examples.

Example 1 (Determine a growth function for $A = F\langle x, y \rangle$, the free algebra in two variables). Then the set \mathcal{B} consists of all of the words in x and y , such as 1 (the word of length zero), x , y , x^2 , xy , yx , and y^2 . Now, given an $n \in \mathbb{N}$, we see that there are two choices for each of the n letters in a word of length n , and so there are 2^n words of length n in \mathcal{B} . Thus $g(n) = \sum_{i=0}^n 2^i = 2^{n+1} - 1$. In this case the growth of A is exponential.

MSC2010: 16P90, 68R15.

Keywords: growth of algebras, de Bruijn graphs.

This work was done during the Potsdam/Clarkson REU during Summer 2012, funded by NSF DMS 1004531.

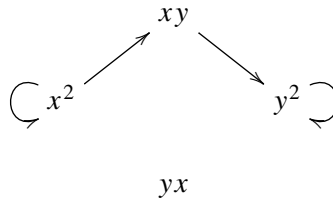
Example 2 (Determine a growth function for $A = F\langle x, y \rangle / (xy)$). Now \mathcal{B} consists of all of the words in x and y that do not have xy as a subword. A few of them are $1, x, y, x^2, yx,$ and y^2 . Since xy is a subword of $x^2y, x^2y \notin \mathcal{B}$. Let $n \in \mathbb{N}$. Since no word having xy as a subword is in \mathcal{B} , the words of length n in \mathcal{B} are of the form $y^k x^{n-k}$ for $k = 0, 1, \dots, n$. We see that there are $n + 1$ of these and thus

$$g(n) = \sum_{i=0}^n (i + 1) = \frac{n^2 + 3n + 2}{2}.$$

The growth function is a quadratic polynomial, so we say that A has quadratic growth.

These two examples were fairly straightforward as there were very few generators for the ideals. We can only imagine how complicated the counting could get when there are several generators. It could easily become a combinatorial nightmare. However we were fortunate that Ufnarovskiĭ [1982] came up a very nice way to overcome this. He considered the cycle structure of a particular directed graph, which is constructed as follows. Consider one of our algebras, with $d + 1$ being the maximum length of the words that generate the ideal, where $d \geq 2$. The set of vertices of the directed graph is the set of all words in x and y of length d in \mathcal{B} . We draw an arrow from a vertex u to a vertex v provided $ua = bv \in \mathcal{B}$, where $a, b \in \{x, y\}$. This graph is called the *overlap graph* for A and will be denoted Γ_A .

Example 3 (Construct Γ_A for $A = F\langle x, y \rangle / I$ where $I = (yx^2, y^2x, xyx, yxy)$). Since the maximum length of generators for I is 3, $d = 2$. Since all of the generators for I have length 3, the vertices for Γ_A are the words in \mathcal{B} of length 2: x^2, y^2, xy, yx . Notice that the words in \mathcal{B} of length 3 are $x^3, y^3, x^2y,$ and xy^2 . Here is Γ_A :



We have an arrow from x^2 to x^2 because $x^3 \in \mathcal{B}$ and $x^3 = (x^2)x = x(x^2)$. Also we have an arrow from x^2 to xy as $x^2y \in \mathcal{B}$ and $x^2y = (x^2)y = x(xy)$. Even though $(y^2)x = y(yx)$, there is no arrow from y^2 to yx , as $y^2x \notin \mathcal{B}$.

The following theorem yields the connection between the overlap graph and the growth of the algebra.

Theorem 4 [Ufnarovskiĭ 1982]. *Let $A = F\langle x, y \rangle / I$, where I is generated by finitely many monomials of maximum length $d + 1$ for some $d \geq 2$, and let Γ_A be the overlap graph for A . Then:*

- (1) *There is a one-to-one correspondence between words in \mathcal{B} of length $d + j$ and paths in Γ_A of length j for each $j \in \mathbb{N}$. (We define the length of a path to be the number of arrows in it, counting repetitions).*
- (2) *If Γ_A has two intersecting cycles, then the growth of A is exponential.*
- (3) *If Γ_A has no intersecting cycles, then the growth of A is polynomial of degree s , where s is the maximal number of distinct cycles on a path in Γ_A .*

Referring to Example 3 above, we see that Γ_A has no intersecting cycles, but does have two distinct cycles on a path. So its growth is degree two, or quadratic, as we have already seen. Given $d \geq 2$ in Theorem 4, we wish to determine the highest-possible-degree polynomial that bounds the growth for A . In [Ellingsen Jr. 1993] it was shown that $2^d - d + 1$ is an upper bound for this degree.

Now we come to the connection to de Bruijn graphs. We are very grateful to Dr. Jo Ellis-Monaghan of St. Michael's College in Vermont for making us aware of them. It turns out that the overlap graphs for our algebras can be considered as subgraphs of de Bruijn graphs, with the only difference being that de Bruijn used 0 and 1 instead of x and y . For a given $d \geq 2$, the vertices of the *de Bruijn graph* B_d are all of the binary d -tuples, and there is an arrow from the binary d -tuple $u = u_1u_2 \cdots u_d$ to the binary d -tuple $v = v_1v_2 \cdots v_d$ if and only if $u_2u_3 \cdots u_d = v_1v_2 \cdots v_{d-1}$, that is, $u_1u_2 \cdots u_dv_d = u_1v_1v_2 \cdots v_d$. Replacing 0 and 1 with x and y yields the overlap graph using all the words in x and y of length d with all possible arrows. After some online searching, the student authors found that much work has been done on de Bruijn graphs, the most remarkable of which is the following theorem proven by Mykkeltveit [1972], but originally conjectured by Golomb.

Theorem 5. *For any $d \geq 2$, the maximum number of simultaneous disjoint cycles in B_d is $Z(d) = (1/d) \sum_{k|d} \phi(k)2^{d/k}$, where ϕ is Euler's phi function.*

Our main theorem follows.

Theorem 6. *Let $d \geq 2$ and let I be an ideal of $F\langle x, y \rangle$ generated by finitely many words of maximum length $d + 1$. If the growth function for $A = F\langle x, y \rangle / I$ is not exponential, then the maximum possible polynomial degree for the growth of A is $Z(d)$.*

Proof. Let $d \geq 2$, let I be an ideal of $F\langle x, y \rangle$ generated by finitely many words of maximum length $d + 1$ and let $A = F\langle x, y \rangle / I$. Assume that the growth of A is not exponential. Let Γ be the overlap graph for the words of length d with all possible arrows and Γ_A the overlap graph for A . By the previous theorem we know that there are at most $Z(d)$ disjoint cycles in B_d , which is identical to Γ . Thus there can be at most $Z(d)$ distinct cycles on any path in Γ . Since Γ_A is a subgraph of Γ , $Z(d)$ is also the maximum possible number of distinct cycles in Γ_A . Hence by Ufnarovskii's theorem the maximum possible polynomial degree for the growth of A is $Z(d)$. \square

The following table illustrates the drastic improvement of the new upper bound:

| d | $2^d - d + 1$ | $Z(d)$ |
|-----|---------------|--------|
| 2 | 3 | 3 |
| 3 | 6 | 4 |
| 4 | 13 | 6 |
| 5 | 28 | 8 |
| 6 | 59 | 14 |
| 7 | 122 | 20 |
| 8 | 249 | 36 |
| 9 | 504 | 60 |

We have found explicitly that this bound is sharp for $d \in \{2, 3, 4, 5, 6, 7\}$ [Flores et al. 2009; Hunt 2002], and are working on the conjecture that it is sharp for all $d \geq 2$.

References

- [Ellingsen Jr. 1993] H. W. Ellingsen Jr., *Growth of algebras, words, and graphs*, Ph.D. thesis, Virginia Polytechnic Institute and State University, 1993, Available at http://scholar.lib.vt.edu/theses/available/etd-10242005-124052/restricted/LD5655.V856_1993.E455.pdf.
- [Flores et al. 2009] L. A. H. Flores, B. George, and B. Schlomer, “Growth functions of finitely generated algebras”, REU paper, SUNY Potsdam/Clarkson, 2009, Available at http://www.uaeh.edu.mx/docencia/P_Lectura/icbi/assignatura/GrowthFunctionsFinitelyGeneratedAlgebras.pdf.
- [Hunt 2002] D. J. Hunt, “Constructing higher-order de Bruijn graphs”, Master’s thesis, Naval Postgraduate School, 2002, Available at <http://www.dtic.mil/dtic/tr/fulltext/u2/a404934.pdf>.
- [Krause and Lenagan 2000] G. R. Krause and T. H. Lenagan, *Growth of algebras and Gelfand–Kirillov dimension*, Revised ed., Graduate Studies in Mathematics **22**, American Mathematical Society, Providence, RI, 2000. MR 2000j:16035
- [Mykkeltveit 1972] J. Mykkeltveit, “A proof of Golomb’s conjecture for the de Bruijn graph”, *J. Combinatorial Theory Ser. B* **13** (1972), 40–45. MR 48 #1985
- [Ufnarovskii 1982] V. A. Ufnarovskii, “Criterion for the growth of graphs and algebras given by words”, *Mat. Zametki* **31**:3 (1982), 465–472, 476. In Russian; translated in *Math. Notes*, **31**(3), March 1982, 238–241. MR 83f:05026

Received: 2012-11-19 Revised: 2013-08-30 Accepted: 2013-08-31

| | |
|-------------------------|---|
| fredetee@clarkson.edu | <i>Clarkson University, 10 Clarkson Avenue, P.O. Box 7728, Potsdam, NY 13699, United States</i> |
| dj kubala@gmail.com | <i>Providence College, 1 Cunningham Square, Providence, RI 02918, United States</i> |
| nelsoner193@potsdam.edu | <i>SUNY Potsdam, 44 Pierrepont Avenue, Potsdam, NY 13676, United States</i> |
| kelseywells@gmail.com | <i>University of Nebraska-Lincoln, 1400 R St, Lincoln, NE 68588, United States</i> |
| ellinghw@potsdam.edu | <i>SUNY Potsdam, 44 Pierrepont Avenue, Potsdam, NY 13676, United States</i> |

A note on triangulations of sumsets

Károly J. Böröczky and Benjamin Hoffman

(Communicated by Andrew Granville)

For finite subsets A and B of \mathbb{R}^2 , we write $A + B = \{a + b : a \in A, b \in B\}$. We write $\text{tr}(A)$ to denote the common number of triangles in any triangulation of the convex hull of A using the points of A as vertices. We consider the conjecture that $\text{tr}(A + B)^{\frac{1}{2}} \geq \text{tr}(A)^{\frac{1}{2}} + \text{tr}(B)^{\frac{1}{2}}$. If true, this conjecture would be a discrete two-dimensional analogue to the Brunn–Minkowski inequality. We prove the conjecture in three special cases.

1. Introduction

We write A, B to denote finite subsets of \mathbb{R}^d , and $|\cdot|$ to stand for their cardinality. For objects X_1, \dots, X_k in \mathbb{R}^d , $[X_1, \dots, X_k]$ denotes their convex hull. Our starting point is two classical results. One is due to Freiman from the 1960s; namely,

$$|A + B| \geq |A| + |B| - 1, \quad (1)$$

with equality if and only if A and B are arithmetic progressions of the same difference. The other result, the Brunn–Minkowski inequality, dates back to the 19th century. It says that if $X, Y \subset \mathbb{R}^d$ are compact sets, then

$$\lambda(X + Y)^{\frac{1}{d}} \geq \lambda(X)^{\frac{1}{d}} + \lambda(Y)^{\frac{1}{d}},$$

where λ stand for the Lebesgue measure, and equality holds if X and Y are convex homothetic sets. This theorem has been successfully applied to estimating the size of a sumset, for example by Ruzsa, Green, and Tao. In turn, various discrete analogues of the Brunn–Minkowski inequality have been established in papers by Bollobás and Leader, Gardner and Gronchi, Green and Tao and, most recently, by Gryniewicz and Serra in the planar case. All these papers use the method of compression, which changes a finite set into a set better suited for sumset estimates, but which cannot control the convex hull. See [Freiman 1973; 2002] for the earlier history, and [Ruzsa 2009] and [Tao and Vu 2006] for thorough surveys.

MSC2010: 11B75, 52C05.

Keywords: additive combinatorics, sumsets, Brunn–Minkowski inequality, triangulations.

Böröczky is supported by OTKA 109789.

Unfortunately the known analogues are not as simple in their form as the original Brunn–Minkowski inequality. A formula due to Gardner and Gronchi says that if A is not contained in any affine subspace of \mathbb{R}^d , then

$$|A + B| \geq (d!)^{-\frac{1}{d}} (|A| - d)^{\frac{1}{d}} + |B|^{\frac{1}{d}}.$$

In this paper, we discuss a more direct version of the Brunn–Minkowski inequality in the plane, which would improve Freiman’s inequality if both A and B are two-dimensional.

In the planar case ($d = 2$), a recent conjecture by Matolcsi and Ruzsa (personal communication, 2009) might point to the right version of the Brunn–Minkowski inequality. Let A be a finite noncollinear point set in \mathbb{R}^2 . We write $\text{tr } A$ to denote the common number of triangles in any triangulation of $[A]$ using the points of A as vertices. If b_A and i_A denote the number of points of A in $\partial[A]$ and $\text{int}[A]$, then the Euler formula yields

$$\text{tr } A = b_A + 2i_A - 2. \quad (2)$$

If Π is a polygon with vertices in \mathbb{Z}^2 , and $A = \mathbb{Z}^2 \cap \Pi$, then Pick’s theorem says that

$$\text{tr } A = 2\lambda(\Pi).$$

Now the Ruzsa–Matolcsi conjecture proposes that if A and B in the plane are not collinear, then

$$\text{tr}(A + B)^{\frac{1}{2}} \geq \text{tr}(A)^{\frac{1}{2}} + \text{tr}(B)^{\frac{1}{2}}. \quad (3)$$

We note that equality holds if for a polygon Π whose vertices are in \mathbb{Z}^2 and integers $k, m \geq 1$, we have $A = \mathbb{Z}^2 \cap k\Pi$ and $B = \mathbb{Z}^2 \cap m\Pi$.

In this paper, we verify (3) in some special cases. To present our main idea we note that if $\alpha, \beta > 0$, then

$$(\alpha + \beta)^2 \leq 2(\alpha^2 + \beta^2), \quad (4)$$

with equality if and only if $\alpha = \beta$. Thus conjecture (3) follows from

$$\text{tr}(A + B) \geq 2[\text{tr } A + \text{tr } B]. \quad (5)$$

This inequality does not hold in general. For example, let Π be a polygon with vertices in \mathbb{Z}^2 , and let $A = \mathbb{Z}^2 \cap k\Pi$ and $B = \mathbb{Z}^2 \cap m\Pi$ for integers $k, m \geq 1$. If $k \neq m$, then we have equality in the Brunn–Minkowski theorem for $X = [A]$ and $Y = [B]$. Still, as we verify, (5) holds in several interesting cases.

The triangulation conjecture (3) can be written in the following form.

Conjecture 1 (main conjecture). *If A and B are finite noncollinear sets \mathbb{R}^2 , then*

$$\sqrt{2i_{A+B} + b_{A+B} - 2} \geq \sqrt{2i_A + b_A - 2} + \sqrt{2i_B + b_B - 2}.$$

In turn, (5) is equivalent with

$$2i_{A+B} + b_{A+B} \geq 4i_A + 4i_B + 2b_A + 2b_B - 6. \quad (6)$$

2. Remarks on the boundary

In the following, we need the notion of exterior normal. A vector u is an exterior normal at x_0 to $[A]$, where $x_0 \in A$, if

$$u \cdot x_0 = \max\{u \cdot x : x \in A\}.$$

It immediately follows that only points in the boundary of $[A]$ will have nonzero exterior normals. It also follows that if $a + b$ is a boundary point of $[A + B]$ for $a \in A$ and $b \in B$, then an exterior unit normal u at $a + b$ to $[A + B]$ is an exterior unit normal at a to $[A]$, and at b to $[B]$. We conclude the following:

Lemma 2. *If A and B are finite noncollinear sets in \mathbb{R}^2 , and $a \in A$ and $b \in B$, then $a + b$ lies on the boundary of $[A + B]$ with nonzero exterior unit normal vector u if and only if u is an exterior normal to $[A]$ at a and to $[B]$ at b .*

For a unit vector u , and finite set A , define the collinear set of points

$$A_u = \{x \in A : u \cdot x = \max_{y \in A}(u \cdot y)\}.$$

Lemma 3. *For any finite noncollinear sets A and B in \mathbb{R}^2 , we have*

$$b_{A+B} \geq b_A + b_B,$$

with equality if and only if the inequalities $|A_u| \geq 2$ and $|B_u| \geq 2$ for a unit vector u imply that A_u and B_u are arithmetic progressions of the same difference.

Proof. For a finite collinear set C , let $S(C) = |C| - 1$, namely, the number of segments the points of C divide the line into. Therefore if C and D are contained in parallel lines, then $S(C + D) \geq S(C) + S(D)$, with equality if and only if $|C| = 1$, $|D| = 1$, or C and D are arithmetic progressions of the same difference. Applying this observation to $C = A_u$ and $D = B_u$ for each unit vector which is an exterior normal to a side of $[A + B]$ yields the lemma. \square

3. Sums with unique representation for each point

In this section we consider the case where representation of points in $A + B$ is unique. We say that the representation is unique when for all $x \in A + B$, if $x = a_1 + b_1$ and $x = a_2 + b_2$, then $a_1 = a_2$ and $b_1 = b_2$.

Theorem 4. *If the representation of points in $A + B$ is unique, then Conjecture 1 holds.*

Proof. From the previous section, we see that whether $x = a + b \in A + B$ lies on the boundary of $[A + B]$ depends only on the exterior normals of $a \in A$ and $b \in B$. So applying any transformation to A or B that preserves $|A + B|$, $\text{tr } A$, $\text{tr } B$, and the exterior normals of A and B will also preserve $\text{tr}(A + B)$. Note that scalar multiplication by ϵ , where $\epsilon A = \{\epsilon a : a \in A\}$, satisfies the latter three conditions immediately. Since the representation of points in $A + B$ is unique, picking ϵ so that the representation of points in $\epsilon A + B$ is also unique will satisfy the first condition.

We pick ϵ small enough so that, for fixed $b \in B$, letting $\epsilon A + b = \{a + b : a \in \epsilon A\}$, for any $x \in \epsilon A + B$, if $x \in [\epsilon A + b]$, then $x \in \epsilon A + b$. Geometrically, this amounts to shrinking A enough that $\epsilon A + B$ looks like a little copy of A placed at each point in B . It follows that the representation of points in $\epsilon A + B$ is unique, and hence $\text{tr}(\epsilon A + B) = \text{tr}(A + B)$.

Assume without loss of generality that $\text{tr } A = \text{tr}(\epsilon A) \geq \text{tr } B$. We begin to draw lines between points in $\epsilon A + B$ to form a partial triangulation, which can be extended to a triangulation of $\epsilon A + B$. For each $b \in B$, draw lines on $\epsilon A + b$ that form a triangulation of that set. Then, consider a triangulation T of B . For $b_1, b_2 \in B$ that are connected by a line in T , consider $\epsilon A + b_1$ and $\epsilon A + b_2$. Pick a point $b_1^* \in \epsilon A + b_1$ that has exterior normal $b_2 - b_1$ in $[\epsilon A + b_1]$. Pick a point $b_2^* \in \epsilon A + b_2$ that has exterior normal $b_1 - b_2$ in $[\epsilon A + b_2]$. Now, in $\epsilon A + B$, draw a line between b_1^* and b_2^* . Geometrically, we have mimicked a triangulation of A at each little copy of A , and a triangulation of B on a large scale, treating each little copy of A as a point in B . Letting $\text{ptr}(\epsilon A + B)$ denote the number of polygons enclosed in this partial triangulation, it follows that

$$\text{tr}(A + B) = \text{tr}(\epsilon A + B) \geq \text{ptr}(\epsilon A + B) = |B| \text{tr } A + \text{tr } B. \quad (7)$$

Conjecture 1 then follows from the inequality

$$\sqrt{|B| \text{tr } A + \text{tr } B} \geq \sqrt{\text{tr } A} + \sqrt{\text{tr } B}. \quad (8)$$

Since $|B| \geq 3$ and $\text{tr } A \geq \text{tr } B$, $(|B| - 2) \text{tr } A \geq \text{tr } B$ holds, which then implies (8). \square

4. The case $i_A = i_B = 1$

We see that Lemma 3 yields that (6), and in turn Conjecture 1, would follow from

$$2i_{A+B} \geq 4i_A + 4i_B + b_A + b_B - 6, \quad (9)$$

which we have already noted does not always hold. However, in the remainder of this paper we show it holds for two special cases. The proof of the first case is simple:

Theorem 5. *When $i_A = i_B = 1$, Conjecture 1 holds.*

Proof. From Lemma 2, it follows that if $a \in A_{\text{int}} = \{a \in A : a \in \text{int}[A]\}$, then $a + B \subset (A + B)_{\text{int}}$. So by (1), since i_A and i_B are nonempty, $i_{A+B} \geq i_A + |B| - 1$,

and similarly $i_{A+B} \geq i_B + |A| - 1$. Thus, since $|A| = i_A + b_A$ and $|B| = i_B + b_B$, we have

$$2i_{A+B} \geq 2i_A + 2i_B + b_A + b_B - 2. \quad (10)$$

In the case that $i_A = i_B = 1$, (9) follows. □

5. The case $|A| = b_A$ and $|B| = b_B$

We now turn to the case $|A| = b_A$ and $|B| = b_B$, or in other words, both A and B lie on the boundary of their convex hulls. In this case, (9) becomes

$$2i_{A+B} \geq b_A + b_B - 6. \quad (11)$$

The bad news is that (11) does not always hold. Let

$$\tilde{A} = \{(0, 0), (1, 0), (0, 1)\} = \{(x, y) \in \mathbb{N}^2 : x + y \leq 1\},$$

$$\tilde{B} = \{(0, 0), (1, 0), (0, 1), (2, 0), (1, 1), (0, 2)\} = \{(x, y) \in \mathbb{N}^2 : x + y \leq 2\}.$$

Therefore $|\tilde{A}| = b_{\tilde{A}} = 3$, $|\tilde{B}| = b_{\tilde{B}} = 6$, and $\tilde{A} + \tilde{B} = \{(x, y) \in \mathbb{N}^2 : x + y \leq 3\}$ yields $i_{\tilde{A}+\tilde{B}} = 1$. In particular, (11) fails to hold for \tilde{A} and \tilde{B} , but the good news is that Conjecture 1 does hold for them.

We note that $\tilde{B} = \tilde{A} + \tilde{A}$. Actually, if A is any set of three noncollinear points, and $B = A + A$, then there exists a linear transformation φ such that A is a translate of $\varphi\tilde{A}$, and B is a translate of $\varphi\tilde{B}$. Therefore (11) does not hold for that A and B , as well. However, in the remainder of the paper, we prove the following theorem. From this result Conjecture 1 holds for the case when $|A| = b_A$ and $|B| = b_B$.

Theorem 6. *If A and B are finite noncollinear sets in \mathbb{R}^2 such that $|A| = b_A$, $|B| = b_B$ and (11) fails to hold, then either $|A| = 3$, and B is a translate of $A + A$, or $|B| = 3$, and A is a translate of $B + B$.*

To prove Theorem 6, we consider a unit vector v not parallel to any side of $[A]$ or $[B]$. We think of v as pointing vertically upwards. Let $l_{v,A}$ and $r_{v,A}$ be the leftmost and rightmost vertices of $[A]$, respectively. We note that $l_{v,A}$ and $r_{v,A}$ are unique, because v is not parallel to any side of $[A]$. Similarly, let $l_{v,B}$ and $r_{v,B}$ be the (unique) leftmost and rightmost vertices of $[B]$, respectively.

Remember that v points upwards. We observe that $l_{v,A}$ and $r_{v,A}$ divide the boundary of $[A]$ into one “upper” and one “lower” polygonal arc. Let $A_{v,\text{upp}}$ and $A_{v,\text{low}}$ denote the set of points of A in the upper and lower polygonal arcs, respectively, excluding $l_{v,A}$ and $r_{v,A}$. For $a \in A$, we have

$$a \in A_{v,\text{upp}} \text{ if and only if } u \cdot v > 0 \text{ for any unit exterior normal } u \text{ to } [A] \text{ at } a, \quad (12)$$

$$a \in A_{v,\text{low}} \text{ if and only if } u \cdot v < 0 \text{ for any unit exterior normal } u \text{ to } [A] \text{ at } a. \quad (13)$$

In addition, as $l_{v,A}$ and $r_{v,A}$ are excluded, we have

$$|A_{v,\text{upp}}| + |A_{v,\text{low}}| = b_A - 2. \quad (14)$$

Similarly, $l_{v,B}$ and $r_{v,B}$ divide the boundary of $[B]$ into an “upper” and a “lower” polygonal arc; let $B_{v,\text{upp}}$ and $B_{v,\text{low}}$ denote the set of points of B in the upper and lower polygonal arcs, respectively, excluding $l_{v,B}$ and $r_{v,B}$. For $b \in B$, we have

$$b \in B_{v,\text{upp}} \text{ if and only if } u \cdot v > 0 \text{ for any unit exterior normal } u \text{ to } [B] \text{ at } b, \quad (15)$$

$$b \in B_{v,\text{low}} \text{ if and only if } u \cdot v < 0 \text{ for any unit exterior normal } u \text{ to } [B] \text{ at } b, \quad (16)$$

$$|B_{v,\text{upp}}| + |B_{v,\text{low}}| = b_B - 2. \quad (17)$$

Lemma 7. *Let A and B be finite noncollinear sets in \mathbb{R}^2 , and let v be a unit vector not parallel to any side of $[A]$ or $[B]$. If $A_{v,\text{upp}}$, $A_{v,\text{low}}$, $B_{v,\text{upp}}$ and $B_{v,\text{low}}$ are all nonempty, then (11) holds.*

Proof. Lemma 2, (12) and (16) yield that $A_{v,\text{upp}} + B_{v,\text{low}} \subset \text{int}[A + B]$; therefore

$$i_{A+B} \geq |A_{v,\text{upp}} + B_{v,\text{low}}| \geq |A_{v,\text{upp}}| + |B_{v,\text{low}}| - 1.$$

In addition, Lemma 2, (13) and (15) yield that $A_{v,\text{low}} + B_{v,\text{upp}} \subset \text{int}[A + B]$; therefore

$$i_{A+B} \geq |A_{v,\text{low}} + B_{v,\text{upp}}| \geq |A_{v,\text{low}}| + |B_{v,\text{upp}}| - 1.$$

We deduce from (14) and (17) that

$$2i_{A+B} \geq |A_{v,\text{upp}}| + |B_{v,\text{low}}| + |A_{v,\text{low}}| + |B_{v,\text{upp}}| - 2 = b_A + b_B - 6. \quad \square$$

In other words, Lemma 7 says that if (11) does not hold, then at least one of the sets $A_{v,\text{upp}}$, $A_{v,\text{low}}$, $B_{v,\text{upp}}$ and $B_{v,\text{low}}$ is empty. We observe that replacing v by $-v$ simply exchanges $A_{v,\text{upp}}$ and $A_{v,\text{low}}$ on the one hand, and $B_{v,\text{upp}}$ and $B_{v,\text{low}}$ on the other hand. Therefore Proposition 9 will refine Lemma 7. Before that, we verify another auxiliary statement. Let $[p, q]$ denote the closed line segment with end points $p, q \in \mathbb{R}^2$.

Lemma 8. *Let A and B be finite noncollinear sets in \mathbb{R}^2 , and let v be a unit vector not parallel to any side of $[A]$ or $[B]$. If $A_{v,\text{low}} = \emptyset$, then $i_{A+B} \geq |B_{v,\text{upp}}| - 2$, where equality would imply that $B_{v,\text{low}} \subset [l_{v,B}, r_{v,B}]$, and the segments $[l_{v,A}, r_{v,A}]$ and $[l_{v,B}, r_{v,B}]$ are parallel.*

Proof. We drop the reference to v in the notation. After applying a linear transformation fixing v , we may assume that

$$w \cdot v = 0 \text{ for } w = r_A - l_A. \quad (18)$$

We may also assume that

$$l_A \cdot v = r_A \cdot v = 0. \quad (19)$$

If $r_B \cdot v > l_B \cdot v$, then we reflect both A and B through the line $\mathbb{R}v$. This keeps v , but interchanges the roles of l_A and r_A on the one hand, and the roles of l_B and r_B on the other hand. Therefore we may assume that

$$r_B \cdot v \leq l_B \cdot v. \quad (20)$$

Understanding exterior normals helps bound interior points in $[A + B]$. As A has some point above $[l_A, r_A]$ by $A_{\text{low}} = \emptyset$, (18) yields that

$$\text{either } u \cdot w > 0 \text{ or } u = -v \text{ for any exterior unit normal } u \text{ at } r_A \text{ to } [A], \quad (21)$$

$$\text{either } u \cdot w < 0 \text{ or } u = -v \text{ for any exterior unit normal } u \text{ at } l_A \text{ to } [A]. \quad (22)$$

We may assume that $B_{\text{upp}} \neq \emptyset$ (otherwise Lemma 8 trivially holds). We subdivide B_{upp} into the sets

$$B_{\text{upp}}^- = \{b \in B_{\text{upp}} : u \cdot w < 0 \text{ for any exterior unit normal } u \text{ at } b \text{ to } [B]\}, \quad (23)$$

$$B_{\text{upp}}^+ = \{b \in B_{\text{upp}} : u \cdot w > 0 \text{ for any exterior unit normal } u \text{ at } b \text{ to } [B]\}, \quad (24)$$

$$B_{\text{upp}}^0 = \{b \in B_{\text{upp}} : v \text{ is an exterior unit normal } u \text{ at } b \text{ to } [B]\}. \quad (25)$$

Since for any $b \in B$, the set of all exterior unit normals u at b to $[B]$ is an arc of the unit circle, the sets B_{upp}^- , B_{upp}^+ and B_{upp}^0 are pairwise disjoint, and their union is B_{upp} . In addition, we define

$$\tilde{B}_{\text{upp}}^- = \begin{cases} \{l_B\} \cup B_{\text{upp}}^- & \text{if there exists } b \in B \text{ with } b \cdot v < l_B \cdot v, \\ B_{\text{upp}}^- & \text{if } b \cdot v \geq l_B \cdot v \text{ for all } b \in B. \end{cases} \quad (26)$$

It follows that if $b \in \tilde{B}_{\text{upp}}^-$, then

$$\text{either } u \cdot w < 0 \text{ or } u = v \text{ for an exterior unit normal } u \text{ to } [B] \text{ at } b. \quad (27)$$

Turning to B_{upp}^0 , if $B_{\text{upp}}^0 \neq \emptyset$, then there exist $l_B^0, r_B^0 \in B_{\text{upp}}^0$ such that $r_B^0 - l_B^0 = sv$ for $s \geq 0$, and

$$B_{\text{upp}}^0 = B \cap [l_B^0, r_B^0], \quad (28)$$

$$v \cdot b_0 = \max\{v \cdot b : b \in B\} = H \text{ for all } b_0 \in B_{\text{upp}}^0. \quad (29)$$

To estimate i_{A+B} , we deduce from Lemma 2, and from (21) and (27) on the one hand, from (22) and (24) on the other hand, that

$$\begin{aligned} r_A + \tilde{B}_{\text{upp}}^- &\subset \text{int}[A + B] \text{ if } B_{\text{upp}}^- \neq \emptyset, \\ l_A + B_{\text{upp}}^+ &\subset \text{int}[A + B] \text{ if } B_{\text{upp}}^+ \neq \emptyset. \end{aligned} \quad (30)$$

We claim that if $\tilde{B}_{\text{upp}}^- \neq \emptyset$ and $B_{\text{upp}}^+ \neq \emptyset$, then

$$|(r_A + \tilde{B}_{\text{upp}}^-) \cap (l_A + B_{\text{upp}}^+)| \leq 1. \quad (31)$$

We observe that $r_A + x = l_A + y$ if and only if $y - x = w$, and hence $x \cdot v = y \cdot v$. However, if $x_1, x_2 \in \tilde{B}_{\text{upp}}^-$ and $y_1, y_2 \in B_{\text{upp}}^+$ with $x_1 \cdot v = y_1 \cdot v < x_2 \cdot v = y_2 \cdot v$, then $(y_2 - x_2) \cdot v < (y_1 - x_1) \cdot v$, which in turn yields (31). We conclude by (19), (29), (30) and (31) that

$$|\{z \in (A + B) \cap \text{int}[A + B] : z \cdot v < H\}| \geq |\tilde{B}_{\text{upp}}^-| + |B_{\text{upp}}^+| - 1. \quad (32)$$

We recall that there exists some $p \in A_{\text{upp}}$, and hence $p \cdot v > 0$ by $l_A \cdot v = 0$. Thus if $B_{\text{upp}}^0 \neq \emptyset$, and $z \in \{l_A, r_A\} + B_{\text{upp}}^0$ is different from $l_A + l_B^0$ and $r_A + r_B^0$, then these two points of $A + B$ lie left and right from z . Since $(l_A + l_B) \cdot v < z \cdot v$ and $(p + l_B^0) \cdot v > z \cdot v$, we have $z \in \text{int}[A + B]$. In particular, $|\{l_A, r_A\} + B_{\text{upp}}^0| \geq |B_{\text{upp}}^0| + 1$ yields that

$$|\{z \in (A + B) \cap \text{int}[A + B] : z \cdot v = H\}| \geq |B_{\text{upp}}^0| - 1. \quad (33)$$

Adding (32) and (33) implies $i_{A+B} \geq |B_{\text{upp}}| - 2$. If $i_{A+B} = |B_{\text{upp}}| - 2$, then $\tilde{B}_{\text{upp}}^- = B_{\text{upp}}^-$, and hence $r_B \cdot v = l_B \cdot v$ by (20) and (26), and $B_{\text{low}} \subset [l_B, r_B]$. In particular, (18) implies that $[l_{v,A}, r_{v,A}]$ and $[l_{v,B}, r_{v,B}]$ are parallel. \square

Proposition 9. *Let A and B be finite noncollinear sets in \mathbb{R}^2 , and let v be a unit vector not parallel to any side of $[A]$ or $[B]$. If (11) does not hold, then possibly after exchanging A and B , or v and $-v$, we have the following:*

- (i) $A_{v,\text{low}} = \emptyset$.
- (ii) $B_{v,\text{low}} \subset [l_{v,B}, r_{v,B}]$.
- (iii) $[l_{v,A}, r_{v,A}]$ and $[l_{v,B}, r_{v,B}]$ are parallel.
- (iv) Either $B_{v,\text{low}} = \emptyset$ and $b_B = b_A$, or $|B_{v,\text{upp}}| = |A_{v,\text{upp}}| + |B_{v,\text{low}}| + 1$ and $b_B > b_A$.

Proof. We drop the reference to v in the notation. To present the argument, we make some preparations. Again using that (11) does not hold, Lemma 7 yields that possibly after exchanging A and B , or v and $-v$, we may assume that

$$A_{\text{low}} = \emptyset.$$

Possibly after exchanging A and B again, we may assume that

$$\text{if } B_{\text{low}} = \emptyset, \text{ then } b_B \geq b_A. \quad (34)$$

Since (11) does not hold, we have

$$i_{A+B} < \frac{1}{2}(b_A + b_B) - 3. \quad (35)$$

First we show that

$$\begin{aligned} \text{either } |B_{\text{upp}}| &= \frac{b_A + b_B}{2} - 2, \quad B_{\text{low}} = \emptyset \text{ and } b_A = b_B, \\ \text{or } |B_{\text{upp}}| &> \frac{b_A + b_B}{2} - 2. \end{aligned} \quad (36)$$

If $B_{\text{low}} = \emptyset$, then $b_B \geq b_A$ by (34), and hence

$$|B_{\text{upp}}| = b_B - 2 \geq \frac{b_A + b_B}{2} - 2,$$

with equality only if $b_A = b_B$.

If $B_{\text{low}} \neq \emptyset$, then we use that $A_{\text{upp}} \neq \emptyset$ by $A_{\text{low}} = \emptyset$. Thus Lemma 2, (12) and (16) yield that $A_{\text{upp}} + B_{\text{low}}$ lies in the interior of $[A + B]$. Combining this fact with (17) leads to

$$\begin{aligned} i_{A+B} &\geq |A_{\text{upp}} + B_{\text{low}}| \geq |A_{\text{upp}}| + |B_{\text{low}}| - 1 \\ &= b_A - 2 + b_B - 2 - |B_{\text{upp}}| - 1 = b_A + b_B - |B_{\text{upp}}| - 5. \end{aligned} \quad (37)$$

Therefore

$$|B_{\text{upp}}| > \frac{b_A + b_B}{2} - 2$$

by (35), proving (36).

It follows from (35) and (36) that $i_{A+B} < |B_{\text{upp}}| - 1$; thus Lemma 8 implies that $i_{A+B} = |B_{\text{upp}}| - 2$, and in turn Proposition 9(ii) and (iii) hold. To prove (iv), we deduce from (35) that

$$b_A + b_B - 6 > 2i_{A+B} = 2|B_{\text{upp}}| - 4.$$

Therefore (36) yields that either $B_{\text{low}} = \emptyset$ and $b_A = b_B$, or

$$b_A + b_B - 4 < 2|B_{\text{upp}}| < b_A + b_B - 2.$$

In particular, $2|B_{\text{upp}}| = b_A + b_B - 3$ in the second case, which is in turn equivalent to $|B_{\text{upp}}| = |A_{\text{upp}}| + |B_{\text{low}}| + 1$ by $|A_{\text{upp}}| = b_A - 2$ and (17). In addition, $|B_{\text{upp}}| = |A_{\text{upp}}| + |B_{\text{low}}| + 1$ implies that $b_B > b_A$. \square

We have now developed enough machinery to prove Theorem 6, which we restate here:

Theorem 6. *If A and B are finite noncollinear sets in \mathbb{R}^2 such that $|A| = b_A$, $|B| = b_B$ and (11) fails to hold, then either $|A| = 3$, and B is a translate of $A + A$, or $|B| = 3$, and A is a translate of $B + B$.*

Proof. We follow Proposition 9, and choose A , B , and v as in that result. For each $x \in A$, we have that if x lies on a corner of $[A]$, there exist vectors $v_{x,l}$ and $v_{x,r}$ such that $x = l_{v_{x,l},A}$ and $x = r_{v_{x,r},A}$. Since $A_{v,\text{low}} = \emptyset$, in the first case

$r_{v_{x,l},A} = r_{v,A}$, and in the second $l_{v_{x,r},A} = l_{v,A}$. Consider one such $x \in A_{v,\text{upp}}$. By Proposition 9, it follows that A is a subset of the triangle T_A formed by $l_{v,A}$, $r_{v,A}$, and x . And, by the same proposition, all the sides of $[B]$ must be parallel to sides in A , so B is a subset of some triangle $T_B = \phi T_A$, where ϕ is the composition of a transposition and scalar multiplication. Then the corners of $[B]$ are $l_{v,B}$, $r_{v,B}$, and some point $y \in B$. We define open line segments

$$\begin{aligned} s_1 &= (l_{v,A}, r_{v,A}), & s_2 &= (l_{v,A}, x), & s_3 &= (x, r_{v,A}), \\ t_1 &= (l_{v,B}, r_{v,B}), & t_2 &= (l_{v,B}, y), & t_3 &= (y, r_{v,B}). \end{aligned}$$

Let $A_i = s_i \cap A$ and $B_i = t_i \cap B$ for $i \in \{1, 2, 3\}$. Note that $A_1 = \emptyset$, and s_i is parallel to t_i , yet $A_i = \emptyset$ or $B_i = \emptyset$.

Assume for contradiction that $|A| > 3$. By Proposition 9, $|B_{v,\text{upp}}| \geq 2$. Thus $B_i \neq \emptyset$ for one $i \in \{2, 3\}$. Assume without loss of generality that $B_3 \neq \emptyset$; then by Proposition 9, $A_3 = \emptyset$ and so $A_2 \neq \emptyset$. Thus, letting $p \in A_2$, since B_1 and B_3 share no nonzero exterior normals with p , and since A_2 and $r_{v,B}$ share no nonzero exterior normals, $B_1 + p$, $A_2 + r_{v,B}$, and $B_3 + p$ are all in $(A + B)_{\text{int}}$. And since $T_B = \phi T_A$, these three sets are pairwise disjoint. So

$$i_{A+B} \geq |B_1 + p| + |A_2 + r_{v,B}| + |B_3 + p| = b_A + b_B - 6, \quad (38)$$

and thus (11) holds, contrary to our assumption. So $|A| = 3$.

By Proposition 9, we have that if $B_{v,\text{low}} = \emptyset$, then $b_A = b_B = 3$. So, $2i_{A+B} \geq b_A + b_B - 6 = 0$, and again (11) holds. Thus, we have that $|B_{v,\text{low}}| \geq 1$, and so

$$|B_{v,\text{upp}}| = |B_{v,\text{low}}| + 2. \quad (39)$$

That is,

$$|B_2| + |B_3| = |B_1| + 1. \quad (40)$$

By the same argument, we get

$$|B_1| + |B_2| = |B_3| + 1, \quad (41)$$

$$|B_1| + |B_3| = |B_2| + 1. \quad (42)$$

It follows that $|B_1| = |B_2| = |B_3| = 1$ and so $b_B = 6$.

Now, $i_{A+B} > 0$, and if $i_{A+B} \geq 2$ then (11) holds, contradicting our assumption. Assuming then that $i_{A+B} = 1$, we let $b_i \in B_i$ for $i \in \{1, 2, 3\}$. Then we see that $x + b_1 = r_{v,A} + b_2 = l_{v,A} + b_3$. And since $T_B = \phi T_A$, B must just be a translated version of $A + A$. And, as was mentioned in the beginning of this section, Conjecture 1 holds for A and B . \square

References

- [Freiman 1973] G. A. Freiman, *Foundations of a structural theory of set addition*, Translations of Math. Monographs **37**, American Mathematical Society, Providence, R. I., 1973. MR 50 #12944 Zbl 0271.10044
- [Freiman 2002] G. A. Freiman, “Structure theory of set addition, II: Results and problems”, pp. 243–260 in *Paul Erdős and his mathematics, I* (Budapest, 1999), edited by M. S. Gábor Halász, László Lóvász and V. T. Sós, Bolyai Soc. Math. Stud. **11**, János Bolyai Math. Soc., Budapest, 2002. MR 2004c:11190 Zbl 1034.11056
- [Ruzsa 2009] I. Z. Ruzsa, “Sumsets and structure”, pp. 87–210 in *Combinatorial number theory and additive group theory*, edited by A. Geroldinger and I. Z. Ruzsa, Birkhäuser, Basel, 2009. MR 2010m:11013 Zbl 1177.11005
- [Tao and Vu 2006] T. Tao and V. Vu, *Additive combinatorics*, Cambridge Studies in Advanced Mathematics **105**, Cambridge University Press, 2006. MR 2008a:11002

Received: 2012-12-28

Revised: 2013-05-31

Accepted: 2013-09-22

carlos@renyi.hu

*Alfréd Rényi Institute of Mathematics, Hungarian Academy
of Sciences, Reáltanoda utca 13-15, Budapest 1053, Hungary*

and

Central European University, 1051 Budapest, Nádor utca 9, Hungary

benjaminshoffman@gmail.com

*Department of Mathematical Sciences, Lewis & Clark College,
615 Palatine Hill Road, Portland, OR 97219, United States*

An exploration of ideal-divisor graphs

Michael Axtell, Joe Stickles, Lane Bloome, Rob Donovan,
Paul Milner, Hailee Peck, Abigail Richard and Tristan Williams

(Communicated by Scott T. Chapman)

Zero-divisor graphs have given some interesting insights into the behavior of commutative rings. Redmond introduced a generalization of the zero-divisor graph called an ideal-divisor graph. This paper expands on Redmond's findings in an attempt to find additional information about the structure of commutative rings from ideal-divisor graphs.

1. Definitions and introduction

Throughout, we assume that R is a finite commutative ring with identity, though in some instances the proofs given can be extended to more general rings. A *zero-divisor* in R is an element x such that there exists a nonzero $y \in R$ with $xy = 0$. The set of all zero-divisors in R is denoted by $Z(R)$. The set of all nonzero zero-divisors is denoted by $Z(R)^*$.

A graph G is defined by a vertex set $V(G)$ and an edge set

$$E(G) \subseteq \{\{a, b\} \mid a, b \in V(G)\}.$$

Two vertices x and y joined by an edge are said to be *adjacent*, denoted $x - y$. A vertex x is said to be *looped* if $x - x$. A *path* between two elements $a_1, a_n \in V(G)$ is an ordered sequence $\{a_1, a_2, \dots, a_n\}$ of distinct vertices of G such that $a_{i-1} - a_i$ for all $1 < i \leq n$. If there exists a path between any two distinct vertices, then the graph is said to be *connected*. A graph is said to be *complete* if every vertex is adjacent to every other vertex, and we denote the complete graph on n vertices by K^n . A graph G is a *finite graph* if $V(G)$ is a finite set.

If the vertices of a graph G can be partitioned into two sets with vertices adjacent only if they are in distinct sets, then G is *bipartite*. If vertices in a bipartite graph are adjacent if and only if they are in distinct vertex sets, then the graph is called *complete bipartite*. We will denote the complete bipartite graph with distinct vertex sets of cardinalities m and n by $K^{m,n}$. A *star graph* is a complete bipartite graph

MSC2010: 13M05.

Keywords: commutative ring with identity, radical ideal, zero-divisor graph, ideal-divisor graph.

such that one of its vertex sets has cardinality one. In general, we say a graph G is a *refinement* of a graph H if $V(G) = V(H)$ and $E(H) \subseteq E(G)$. We note that any graph of radius one is a refinement of a star graph.

For any other terms not defined here, see [Chartrand 1985] for a graph theory reference, and see [Herstein 1990] for a ring theory reference. The figures in this paper were generated by Mathematica using programs originally written by Brendan Kelly, Darrin Weber, and Elisabeth Wilson and modified to suit our needs.

Beck [1988] was the first to define the zero-divisor graph of a commutative ring. However, it was in the seminal paper [Anderson and Livingston 1999] that the structure was first used extensively to reveal ring-theoretic properties. In this paper, the *zero-divisor graph* of R , denoted $\Gamma(R)$, is the simple graph with vertex set $V(\Gamma(R)) = Z(R)^*$ and edge set

$$E(\Gamma(R)) = \{\{a, b\} \mid a, b \in V(\Gamma(R)), ab = 0 \text{ and } a \neq b\}.$$

Redmond [2003] introduced *ideal-divisor graphs*, a generalization of zero-divisor graphs. For I an ideal of R , an element $x \in R$ is an *ideal-divisor* if there exists some $y \in R \setminus I$ such that $xy \in I$. The set of ideal-divisors of R with respect to I is denoted $Z_I(R)$. The *ideal-divisor graph* of a R with respect to an ideal I , denoted $\Gamma_I(R)$, is the simple graph with vertex set $V(\Gamma_I(R)) = Z_I(R)^*$ and edge set

$$E(\Gamma_I(R)) = \{\{x, y\} \mid x, y \in V(\Gamma_I(R)), x \neq y \text{ and } xy \in I\}.$$

Redmond [2003] proved that if I is an ideal of R , then $\Gamma_I(R)$ is connected with $\text{diam}(\Gamma_I(R)) \leq 3$. He proved further that if $\Gamma_I(R)$ contains a cycle, then $g(\Gamma_I(R)) \leq 7$, and he developed an algorithm for constructing the graph of $\Gamma_I(R)$ from $\Gamma(R/I)$.

Redmond, like Anderson and Livingston, did not include looped vertices in his definition of the ideal-divisor graph. The following definitions have therefore been modified to include looped vertices. The *zero-divisor graph* of R (denoted $\Gamma(R)$) has vertex set $V(\Gamma(R)) = Z(R)^*$ and edge set

$$E(\Gamma(R)) = \{\{a, b\} \mid a, b \in V(\Gamma(R)) \text{ and } ab = 0\}.$$

The *ideal-divisor graph* of R with respect to an ideal I , denoted $\Gamma_I(R)$, has vertex set $V(\Gamma_I(R)) = Z_I(R)^*$ and edge set

$$E(\Gamma_I(R)) = \{\{x, y\} \mid x, y \in V(\Gamma_I(R)) \text{ and } xy \in I\}.$$

These modified definitions allow a vertex b in $\Gamma(R)$ or $\Gamma_I(R)$ to be adjacent to itself if and only if $b^2 = 0$ or $b^2 \in I$ for each graph, respectively.

In Sections 2 and 3, we expand upon Redmond's results by examining the structure of $\Gamma_I(R)$. We also consider the relationships between $\Gamma_I(R)$ and $\Gamma(R/I)$. In particular, we establish conditions for $\Gamma_I(R)$ to be finite, demonstrate several

relationships between the cut-sets of $\Gamma(R/I)$ and $\Gamma_I(R)$, and prove a result on the connectivity of $\Gamma_I(R)$. In Section 4, we modify and prove a modification of a proposition presented in [Redmond 2003]. A brief discussion at the end of this paper examines the structure of $\Gamma_I(R)$ when I is a radical, primary, or weakly prime ideal.

The following results are included for reference. Although these results were proven for graphs without loops, it is straightforward to check that they still hold when the graphs are looped.

Theorem 1.1 [Redmond 2003, Theorem 2.5]. *Let I be an ideal of R , and let $x, y \in R \setminus I$. Then:*

- (1) *If $x + I$ is adjacent to $y + I$ in $\Gamma(R/I)$, then x is adjacent to y in $\Gamma_I(R)$.*
- (2) *If x is adjacent to y in $\Gamma_I(R)$ and $x + I \neq y + I$, then $x + I$ is adjacent to $y + I$ in $\Gamma(R/I)$.*
- (3) *If x is adjacent to y in $\Gamma_I(R)$ and $x + I = y + I$, then $x^2, y^2 \in I$.*

Corollary 1.2 [Redmond 2003, Corollary 2.6]. *If x and y are (distinct) adjacent vertices in $\Gamma_I(R)$, then all (distinct) elements of $x + I$ and $y + I$ are adjacent in $\Gamma_I(R)$. If $x^2 \in I$, then all the distinct elements of $x + I$ are adjacent in $\Gamma_I(R)$.*

2. Structure of $\Gamma_I(R)$

In this section, we investigate the relationship between $\Gamma_I(R)$ and $\Gamma(R/I)$, and provide some results about the general structure of $\Gamma_I(R)$.

A few definitions are needed for clarification in this section. Elements of the vertex set of $\Gamma_I(R)$ which are elements of the same coset in R/I form a *column* in $\Gamma_I(R)$ [Redmond 2003, Theorem 2.9]. Corollary 1.2 gives that if a vertex $a + I$ is looped (i.e., $(a + I)^2 = 0 + I$) in $\Gamma(R/I)$, then all the vertices in the corresponding column of $\Gamma_I(R)$ are adjacent to one another. Finally, the *ideal annihilator* of an element $a \in R \setminus I$ with respect to some ideal I is the set $(I : a) = \{b \mid ab \in I \text{ and } b \in R \setminus I\}$.

Proposition 2.1. *Let I be an ideal of R . If $(a + I)$ and $(b + I)$ are distinct vertices in $\Gamma(R/I)$ with $(a + I) - (b + I)$, then the columns corresponding to $a + I$ and $b + I$, taken as a pair, form a subgraph that is a refinement of a complete bipartite graph in $\Gamma_I(R)$. Moreover, for any $a + i \in V(\Gamma_I(R))$, $|(I : a + i)|$ is equal to $k|I|$ for some $k \in \mathbb{N}$.*

Proof. This result follows directly from Theorem 1.1. □

Example 2.2. In Figure 1, each adjacent pair of columns of $\Gamma_{(8)}(\mathbb{Z}_{24})$ is a refinement of a complete bipartite graph.

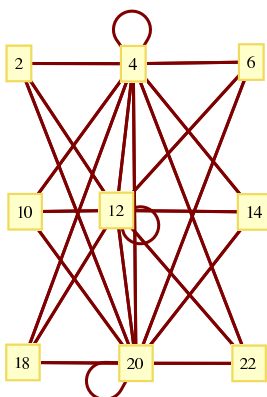


Figure 1. $\Gamma_{(8)}(\mathbb{Z}_{24})$.

Theorem 2.3. *Let S be a commutative ring. Then $\Gamma_I(S)$ is finite if and only if either S is a finite ring or I is a prime ideal. In particular, if $1 \leq |\Gamma_I(S)| < \infty$, then S is a finite ring and I is not a prime ideal.*

Proof. (\Rightarrow) If I is prime, then $\Gamma_I(S) = \emptyset$. So, assume I is not prime.

(1) If I is infinite, then by [Redmond 2003, Corollary 2.7], $\Gamma_I(S)$ is infinite.

(2) If I is finite and S is infinite, then S/I is infinite and not an integral domain, so $\Gamma(S/I)$ is infinite (see [Ganesan 1964]). By [Redmond 2003, Theorem 2.5], since $\Gamma(S/I)$ is isomorphic to a subgraph of $\Gamma_I(S)$, $\Gamma_I(S)$ is also infinite.

(\Leftarrow) Clear. □

3. Cut-sets and connectivity

In a connected graph, a *cut-vertex* is a vertex that, when it and any edges incident to it are removed, separates the graph into two or more connected components. Cut-vertices were introduced into the analysis of zero-divisor graphs in [Axtell et al. 2009] and were further studied in [Axtell et al. 2011]. In [Redmond 2003, Theorem 3.2], Redmond proved that $\Gamma_I(R)$ contains no cut-vertices whenever I is a nonzero proper ideal of R . Cut-sets, a generalization of the cut-vertex, were also introduced into the analysis of zero-divisor graphs in [Coté et al. 2011]. For a connected graph G , a subset $A \subset V(G)$ is a *cut-set* if there exist $c, d \in V(G) \setminus A$ such that every path from c to d contains at least one vertex from A , and no proper subset of A satisfies the same condition. It is easy to show that for a given nonempty set of vertices A , the existence of such c and d is equivalent to the existence of two subgraphs X and Y of G whose (vertexwise and edgewise) union equals G , and whose vertex sets satisfy $V(X) \cap V(Y) = A$, $V(X) \setminus A \neq \emptyset$, and $V(Y) \setminus A \neq \emptyset$. When this happens we say that A *separates* X and Y .

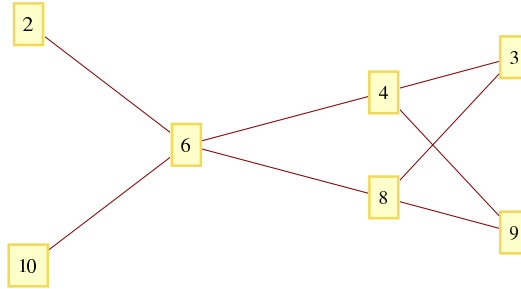


Figure 2. $\Gamma(\mathbb{Z}_{12})$, using Anderson and Livingston’s definition.

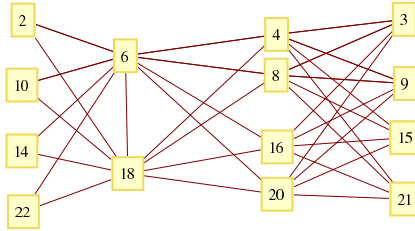


Figure 3. $\Gamma_{(12)}(\mathbb{Z}_{24})$.

Theorem 3.1. *Let I be an ideal of R . If A is a cut-set in $\Gamma_I(R)$, then A is a column or a union of columns.*

Proof. Assume A is a cut-set of $\Gamma_I(R)$. Let $x, y \in V(\Gamma_I(R)) \setminus A$. Let $x - \dots - a + i - \dots - y$ be a path from x to y , where $a + i \in A$. Since $x - \dots - a + \bar{i} - \dots - y$ is also a path from x to y for all $\bar{i} \in I$, we must have $a + I \subseteq A$. \square

As an example, let $R = \mathbb{Z}_{24}$ and let $I = (12)$. Since $R/I \cong \mathbb{Z}_{12}$, we can identify $\Gamma(R/I)$ with Figure 2. We notice that the vertices 4 and 8 form a cut-set. Likewise, looking at Figure 3, in $\Gamma_{(12)}(\mathbb{Z}_{24})$ the set $\{4, 8, 16, 20\}$ is a cut-set. We note that in this figure 4 and 16 form the column associated with $4 + (12)$, while 8 and 20 form the column associated with $8 + (12)$.

Theorem 3.2. *If A is a cut-set in $\Gamma(R/I)$, then $B = \{a + i \mid a + I \in A, i \in I\}$ is a cut-set in $\Gamma_I(R)$.*

Proof. Let X and Y be subgraphs of $\Gamma(R/I)$ separated by the cut-set A . Let $x, y \in V(\Gamma_I(R))$ such that $x + I$ and $y + I$ are vertices of X and Y , respectively. Let $x + I - \dots - y + I$ be a path from $x + I$ to $y + I$. Then since A is a cut-set, this path must contain at least one element from A .

Suppose there exists a path $x - z_1 - \dots - z_n - y$ from x to y that does not contain at least one element from B . From Corollary 1.2, it can be assumed without loss of generality that each z_j is in a distinct column of $\Gamma_I(R)$, where $1 \leq j \leq n$. Thus, by



Figure 4. $\Gamma((\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2)/(\{0\} \times \{0\} \times \mathbb{Z}_2))$.

Theorem 1.1, $(x + I) - (z_1 + I) - \dots - (z_n + I) - (y + I)$ is a path in $\Gamma(R/I)$. This path does not contain at least one element from A , contradicting the fact that A is a cut-set. Therefore, every path between x and y contains at least one element of B .

Suppose B is not the minimal such set. Then there exists some $b \in B$ such that every path from x to y contains at least one vertex from $B \setminus \{b\}$. Then $b + I \in A$, and every path from x to y contains at least one element from $A \setminus \{b + I\}$. This contradicts that A is a cut-set of $\Gamma(R/I)$. \square

The converse is not always true. Consider the graph of

$$\Gamma((\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2)/(\{0\} \times \{0\} \times \mathbb{Z}_2)),$$

shown in Figure 4, which is isomorphic to K^2 .

There are no cut-vertices or cut-sets in the above graph. However, the sets $\{(0, 1, 0), (0, 1, 1)\}$ and $\{(1, 0, 0), (1, 0, 1)\}$ are cut-sets in $\Gamma_{(\{0\} \times \{0\} \times \mathbb{Z}_2)}(\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2)$.

Lemma 3.3. *If $x, y \in \Gamma_I(R)$ are distinct and every path connecting x to y contains a vertex $z \in A \subseteq V(\Gamma_I(R))$, then every path connecting $x + I$ to $y + I$ in $\Gamma(R/I)$ contains an element of $B = \{a + I \mid a \in A\}$.*

Proof. Let $(x + I) - (w_1 + I) - \dots - (w_k + I) - (y + I)$ be a path from $x + I$ to $y + I$ in $\Gamma(R/I)$. If $w_n \notin A$ for all $1 \leq n \leq k$, then there exists a path $x - w_1 - \dots - w_k - y$ in $\Gamma_I(R)$ that does not contain an element of A , a contradiction. \square

Theorem 3.4. *If a cut-set A in $\Gamma_I(R)$ is a union of n columns and $|Z(R/I)^*| - n \geq 2$, then $B = \{a + I \mid a \in A\}$ is a cut-set in $\Gamma(R/I)$.*

Proof. Suppose A is a cut-set in $\Gamma_I(R)$. Since $|Z(R/I)^*| - n \geq 2$, there are $b, c \in \Gamma_I(R) \setminus A$ such that b and c are in different columns, and any path connecting b and c contains an element $a \in A$. To see this, note that if two vertices are in the same column and are separated by A , then these vertices must be isolated when A is removed, because vertices in the same column are adjacent to the same set of vertices by Corollary 1.2. Since there are at least two columns left after the removal of A , we can now choose b and c in different columns that meet the desired conditions.

By Lemma 3.3, any path from $b + I$ to $c + I$ in $\Gamma(R/I)$ must contain $a + I$ for some $a \in A$. The set of all such points is B ; thus, B is a cut-set or contains a cut-set.

Suppose B is not minimal. Then there exists some $a_i + I \in B$ such that $C \subseteq B \setminus \{a_i + I\} = \{a + I \mid a \in A \setminus \{a_i\}\}$ is a cut-set in $\Gamma(R/I)$. Then by Theorem 3.2, $D = \{a + i \mid a + I \in C, i \in I\} \subset A$ is a cut-set in $\Gamma_I(R)$, a contradiction. \square



Figure 5. $\Gamma(\mathbb{Z}_{27}/(9))$.

If $|Z(R/I)^*| - n < 2$, then B would certainly not be a cut-set in $\Gamma(R/I)$. If there was only one column remaining after the removal of A from $\Gamma_I(R)$, then there would only be one coset representative remaining after the removal of B from $\Gamma(R/I)$.

It is proved in [Coté et al. 2011] that if R is not local and if B is a cut-set of $\Gamma(R)$, then $B \cup \{0\}$ is an ideal. A similar theorem for cut-sets in $\Gamma_I(R)$ is provided.

Theorem 3.5. *Let I be an ideal of R such that R/I is nonlocal, let A be a cut-set in $\Gamma(R/I)$, and let $B = \{a + i \mid a + I \in A, i \in I\}$. Then $B \cup I$ is an ideal of R .*

Proof. Let A be a cut-set in $\Gamma(R/I)$. Then $A \cup \{0 + I\}$ is an ideal of R/I by [Coté et al. 2011]. Then $B \cup I = \phi^{-1}(A \cup \{0 + I\})$, where $\phi : R \rightarrow R/I$ is the canonical homomorphism, is an ideal of R . \square

The *connectivity* of a connected graph G , denoted $\kappa(G)$, is the minimum number of vertices that must be removed from G to produce a disconnected graph. It is customary to define the connectivity of the complete graph K^n to be $\kappa(K^n) = n - 1$. In other words, $\kappa(G)$ is the order of the smallest cut-set of G , when G is not isomorphic to K^n . The following result on the connectivity of $\Gamma_I(R)$ is Theorem 3.3 of [Redmond 2003].

Theorem 3.6. *Let I be a nonzero proper ideal of R .*

- (1) *If $\Gamma(R/I)$ is the graph on one vertex, then $\kappa(\Gamma_I(R)) = |I| - 1$.*
- (2) *If $\Gamma(R/I)$ has at least two vertices, then $2 \leq \kappa(\Gamma_I(R)) \leq |I| \cdot \kappa(\Gamma(R/I))$.*
- (3) $|I| - 1 \leq \kappa(\Gamma_I(R))$.

In light of this theorem, consider $\Gamma(\mathbb{Z}_{27}/(9))$, shown in Figure 5. The connectivity of $\kappa(\Gamma(\mathbb{Z}_{27}/(9)))$ is 1. So, by the above theorem, $\kappa(\Gamma_{(9)}(\mathbb{Z}_{27}))$ should be 2 or 3. However, since $\Gamma_{(9)}(\mathbb{Z}_{27})$ (shown in Figure 6) is complete, $\kappa(\Gamma_{(9)}(\mathbb{Z}_{27})) = |\Gamma_{(9)}(\mathbb{Z}_{27})| - 1 = 5$. A reading of the proof of this theorem in [Coté et al. 2011] shows this problem arises only when $\Gamma(R/I)$ is complete. We provide the following modification of this theorem to take into account complete graphs.

Theorem 3.7. *Let I be a nonzero proper ideal of R .*

- (1) *If $\Gamma(R/I)$ is complete on more than two vertices, then*

$$\kappa(\Gamma_I(R)) = |I| \cdot |V(\Gamma(R/I))| - 1.$$

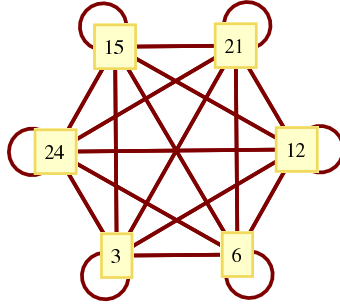


Figure 6. $\Gamma_{(9)}(\mathbb{Z}_{27})$.

(2) If $\Gamma(R/I)$ is the graph on two vertices, then

$$\kappa(\Gamma_I(R)) = |I| \quad \text{or} \quad |I| \cdot |V(\Gamma(R/I))| - 1.$$

(3) If $\Gamma(R/I)$ is not complete and has at least three vertices, then

$$2 \leq \kappa(\Gamma_I(R)) \leq |I| \cdot \kappa(\Gamma(R/I)).$$

(4) $|I| - 1 \leq \kappa(\Gamma_I(R))$.

Proof. Parts 3 and 4 are proved in [Coté et al. 2011].

(1) Suppose $\Gamma(R/I)$ is complete. Then for all $a + I, b + I \in \Gamma(R/I)$, we have $(a + I) - (b + I)$. By [Anderson and Livingston 1999, Theorem 2.8], $Z(R/I)^2 = \{0\}$. Thus, by Theorem 1.1, $\Gamma_I(R)$ is complete. Hence $\kappa(\Gamma_I(R)) = |\Gamma_I(R)| - 1 = |I| \cdot |\Gamma(R/I)| - 1$. (See [Coté et al. 2011, Remark 28].)

(2) Suppose $\Gamma(R/I)$ is the graph on two vertices, $x + I$ and $y + I$. Then by [Anderson and Livingston 1999, Theorem 2.8], either $R/I \cong \mathbb{Z}_2 \times \mathbb{Z}_2$ or $Z(R/I)^2 = \{0\}$. Thus, there are two cases:

- (a) Suppose $x^2 + I = 0 + I = y^2 + I$. Then $\Gamma_I(R)$ is complete. Thus, $\kappa(\Gamma_I(R)) = |I| \cdot |\Gamma(R/I)| - 1$.
- (b) Suppose $x^2 + I \neq 0 + I \neq y^2 + I$. Then $\Gamma_I(R)$ is isomorphic to $K_{|I|,|I|}$. Without loss of generality, let $x \in x + I \subset V(\Gamma_I(R))$. Then x is adjacent to every vertex in $y + I$. Thus, to create a disconnected graph from $\Gamma_I(R)$, every vertex in $y + I$ is removed, i.e., we remove $|I|$ vertices. \square

4. Classifying ideals via ideal-divisor graphs

Let I be an ideal of R . The *radical* of I is the set $\sqrt{I} = \{r \in R \mid r^n \in I \text{ for some } n \in \mathbb{N}\}$. For any ideal I , \sqrt{I} is an ideal of R , and if $\sqrt{I} = I$, then I is called a *radical ideal*. Note that for a radical ideal I of R , if $|I| \geq 2$, then there are no connected columns.

Lemma 4.1. *Let I be an ideal of R and let $a \in Z_I(R)^*$. If no vertex of $\Gamma_I(R)$ is looped, then $a^n \notin I$ for all $n \in \mathbb{N}$.*

Proof. Suppose $a^n \in I$ for some least $n \in \mathbb{N}$. Then, $a^{n-1} \in Z_I(R)^*$ and $(a^{n-1})^2 \in I$. Thus, a^{n-1} is looped, a contradiction. \square

Theorem 4.2. *Let I be an ideal of R . Then I is a radical ideal if and only if no vertex in $\Gamma_I(R)$ is looped (equivalently, $\Gamma_I(R)$ has no connected columns).*

Proof. (\Rightarrow) Consider $a \in V(\Gamma_I(R))$. Since $a \notin I$, $a^n \notin I$ for all $n \in \mathbb{N}$. Thus, $a^2 \notin I$. (\Leftarrow) Let $a \in V(\Gamma_I(R))$. By Lemma 4.1 and the definition of an ideal divisor, $a^n \notin I$ for all $n \in \mathbb{N}$. Hence, if $b^n \in I$ for some $n \in \mathbb{N}$, we must have $b \in I$. Thus, I is a radical ideal. \square

We now move to a classification of primary and weakly prime ideals. Let Q be an ideal of R . We say Q is a *primary ideal* if whenever $ab \in Q$, either $a \in Q$ or $b^n \in Q$ for $n \in \mathbb{N}$. Let P be a proper ideal of R . Then, P is *weakly prime* if $0 \neq ab \in P$ implies $a \in P$ or $b \in P$ (see [Anderson and Smith 2003]).

Lemma 4.3. *Let I be an ideal of R . Let $K = \{k_1, k_2, \dots, k_n\} \subseteq R \setminus I$ such that for each $k_i \in K$, there exists a minimal $m_i \in \mathbb{N}$ such that $k_i^{m_i} \in I$. Then there exists $a \in R \setminus I$ such that $ak_i \in I$ for all $k_i \in K$.*

Proof. There exists a minimal $m_1 \geq 2$ such that $k_1^{m_1} \in I$. Let $a_1 = k_1^{m_1-1}$. Clearly, $a_1 k_1 \in I$ and $a_1 \notin I$. Now, there exists a minimal n_2 with $1 \leq n_2 \leq m_2$ such that $a_1 k_2^{n_2} \in I$. Now let $a_2 = a_1 k_2^{n_2-1}$. Again, $a_2 k_2 \in I$ and $a_2 \notin I$. Continuing in this fashion, there exists an $n_j - 1$ (possibly zero, in which case $a_j = a_{j-1}$) such that $a_j = a_{j-1} k_j^{n_j-1} \notin I$ but $a_j k_j \in I$. Let $a = a_n$. By construction, a is connected to every $k_i \in K$. \square

Theorem 4.4. *Let I be a nonzero ideal of R that is not prime. Then I is a primary ideal if and only if $\Gamma_I(R)$ is a refinement of a star graph.*

Proof. (\Rightarrow) Let $a, b \in V(\Gamma_I(R))$ with $ab \in I$. By definition of $V(\Gamma_I(R))$ and the fact that I is primary, we have $a^r, b^s \in I$ for some $r, s \geq 2$. Therefore, we have $V(\Gamma_I(R)) \subseteq R \setminus I$, and for each $x \in V(\Gamma_I(R))$ there is some $n \in \mathbb{N}$ such that $x^n \in I$. By Lemma 4.3, we have at least one $y \in V(\Gamma_I(R))$ with $xy \in I$ for all $x \in V(\Gamma_I(R))$. That is, the vertex y connects to every other vertex in $\Gamma_I(R)$. Thus, $\Gamma_I(R)$ is a refinement of a star graph.

(\Leftarrow) If $\Gamma_I(R)$ is a refinement of a star graph, then the diameter of $\Gamma_I(R)$ is 2. According to Corollary 2.7 in [Redmond 2003], since I is an ideal of R , $\Gamma_I(R)$ contains a subgraph that is isomorphic to $\Gamma(R/I)$. Using Theorem 1.1 and Corollary 1.2, there exists an element $x + I$ that is connected to every other element, including itself, in $\Gamma(R/I)$. Then, applying Lemma 3.1 in [Axtell et al. 2009] gives us that $Z(R/I)$ is an ideal. If the zero-divisors of a finite ring form an ideal, then that ideal

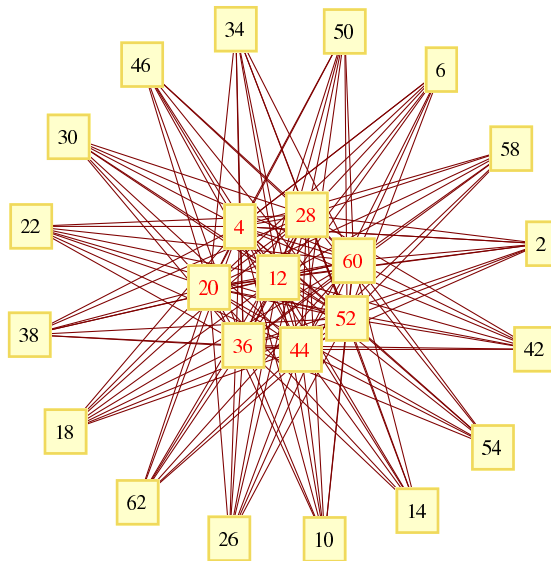


Figure 7. $\Gamma_{(8)}(\mathbb{Z}_{64})$.

is the maximal ideal of the ring, and the ring is local. It is well known that if a ring is local and finite, every zero-divisor is nilpotent. Every zero-divisor in R/I is nilpotent, so I is a primary ideal. \square

Example 4.5. Let $R = \mathbb{Z}_{64}$ and $I = (8)$. We see I is a primary ideal of R , and in the figure below, we see that we have a refinement of a star graph. Note that any of $\{4, 12, 20, 28, 36, 44, 52, 60\}$ could work as our central vertex (see Figure 7).

Note that the condition that I is a nonzero ideal of R in Theorem 4.4 is necessary for the “if” portion on the proof, for if $R = \mathbb{Z}_2 \times F$, where F is a field, and $I = \{(0, 0)\}$, then $\Gamma(R/I)$ is a star graph, but I is not a primary ideal. The issue that arises in this case is that $(1, 0)$ is connected to every other vertex in $\Gamma(R/I)$, but it is not looped.

Lemma 4.6. *Let I be a weakly prime ideal and let $a \in R \setminus I$. If $a^k \in I$ for some $k \in \mathbb{N}$, then $a^k = 0$.*

Proof. Let $a \in R \setminus I$ and assume $a^k \in I^*$. Then $0 \neq a \cdot a^{k-1} \in I$. Since $a \notin I$, we have $a^{k-1} \in I$ because I is weakly prime. Continuing, we obtain $0 \neq a \cdot a \in I$, but $a \notin I$, a contradiction. \square

Theorem 4.7. *Let I be a nonzero ideal of R that is not prime. Then I is weakly prime if and only if $\Gamma_I(R)$ is the induced subgraph of $\Gamma(R)$ on $Z(R) \setminus I$.*

Proof. (\Rightarrow) According to Theorem 7 in [Anderson and Smith 2003], R is not decomposable, so R is either local or a field. Supposing R is local, it is well known

that every zero-divisor is nilpotent. Let $a \in Z(R) \setminus I$. Since a is nilpotent, there exists a minimal $n \in \mathbb{N}$ such that $a^n = 0 \in I$. So, by Lemma 4.6, $a \cdot a^{n-1} \in I$ and $a^{n-1} \notin I$. Hence, $a \in V(\Gamma_I(R))$. Now let $a, b \in V(\Gamma_I(R))$ with $ab \in I$. Since I is weakly prime, $ab = 0$. Hence, $Z(R) \setminus I = V(\Gamma_I(R))$, and $a - b \in \Gamma_I(R)$ if and only if $ab = 0$. Thus, $\Gamma_I(R)$ is the induced subgraph of $\Gamma(R)$ on $Z(R) \setminus I$.

(\Leftarrow) Assume the ideal-divisor graph is the induced subgraph of $\Gamma(R)$ on $Z(R) \setminus I$. Let $a, b \notin I$ and $ab \in I$. Since $\Gamma_I(R)$ is the induced subgraph of $\Gamma(R)$ on $Z(R) \setminus I$, $ab = 0$. Thus, I is a weakly prime ideal. \square

Acknowledgements

This paper is a collaborative effort of two faculty and several undergraduate students. The research of Hailee Peck was funded by the Summer Undergraduate Research Fellowship through Millikin University, as well as the Undergraduate Research Fellow Program for the 2012–2013 year through Millikin University. Lane Bloome was also funded through the Undergraduate Research Fellow Program for the 2012–2013 year through Millikin University. The work of Robert Donovan, Paul Milner, Abigail Richard, and Tristan Williams was the result of undergraduate research performed at the 2010 Wabash College mathematics REU in Crawfordsville, Indiana, which was funded through the National Science Foundation grant DMS-0755260.

The authors would like to thank the referees for their helpful suggestions.

References

- [Anderson and Livingston 1999] D. F. Anderson and P. S. Livingston, “The zero-divisor graph of a commutative ring”, *J. Alg.* **217**:2 (1999), 434–447. MR 2000e:13007 Zbl 0941.05062
- [Anderson and Smith 2003] D. D. Anderson and E. Smith, “Weakly prime ideals”, *Houston J. Math.* **29**:4 (2003), 831–840. MR 2005b:13001 Zbl 1086.13500
- [Axtell et al. 2009] M. Axtell, J. Stickles, and W. Trambachls, “Zero-divisor ideals and realizable zero-divisor graphs”, *Involve* **2**:1 (2009), 17–27. MR 2010b:13011 Zbl 1169.13301
- [Axtell et al. 2011] M. Axtell, N. Baeth, and J. Stickles, “Cut vertices in zero-divisor graphs of finite commutative rings”, *Comm. Algebra* **39**:6 (2011), 2179–2188. MR 2012i:13043 Zbl 1226.13007
- [Beck 1988] I. Beck, “Coloring of commutative rings”, *J. Alg.* **116**:1 (1988), 208–226. MR 89i:13006 Zbl 0654.13001
- [Chartrand 1985] G. Chartrand, *Introductory graph theory*, Dover, New York, 1985. MR 86c:05001
- [Coté et al. 2011] B. Coté, C. Ewing, M. Huhn, C. M. Plaut, and D. Weber, “Cut-sets in zero-divisor graphs of finite commutative rings”, *Comm. Algebra* **39**:8 (2011), 2849–2861. MR 2012i:13014 Zbl 1228.13011
- [Ganesan 1964] N. Ganesan, “Properties of rings with a finite number of zero divisors”, *Math. Ann.* **157** (1964), 215–218. MR 30 #113 Zbl 0135.07704
- [Herstein 1990] I. N. Herstein, *Abstract algebra*, 2nd ed., Macmillan Publishing Company, New York, 1990. MR 92m:00003 Zbl 0841.00003

[Redmond 2003] S. P. Redmond, “An ideal-based zero-divisor graph of a commutative ring”, *Comm. Algebra* **31**:9 (2003), 4425–4443. MR 2004c:13041 Zbl 1020.13001

Received: 2013-02-19 Revised: 2013-06-25 Accepted: 2013-07-03

| | |
|----------------------------|--|
| axte2004@stthomas.edu | <i>Department of Mathematics, University of St. Thomas, St Paul, MN 55105, United States</i> |
| jstickles@millikin.edu | <i>Department of Mathematics, Millikin University, Decatur, IL 62522, United States</i> |
| lbloome@millikin.edu | <i>Department of Mathematics, Millikin University, Decatur, IL 62522, United States</i> |
| rdonovan2@worchester.edu | <i>Department of Mathematics and Computer Science, Worcester State College, Worcester, MA 01602, United States</i> |
| paul.milner89@gmail.com | <i>Department of Mathematics, University of St. Thomas, St. Paul, MN 55105, United States</i> |
| hpeck@millikin.edu | <i>Department of Mathematics, Millikin University, Decatur, IL 62522, United States</i> |
| richarah@miamioh.edu | <i>Department of Mathematics, Miami University, Oxford, OH 45056, United States</i> |
| tristan-williams@uiowa.edu | <i>Department of Mathematics, University of Iowa, Iowa City, IA 52242, United States</i> |

The failed zero forcing number of a graph

Katherine Fetcie, Bonnie Jacob and Daniel Saavedra

(Communicated by Joseph A. Gallian)

Given a graph G , the *zero forcing number* of G , $Z(G)$, is the smallest cardinality of any set S of vertices on which repeated applications of the color change rule results in all vertices joining S . The *color change rule* is: if a vertex v is in S , and exactly one neighbor u of v is not in S , then u joins S in the next iteration.

In this paper, we introduce a new graph parameter, the failed zero forcing number of a graph. The *failed zero forcing number* of G , $F(G)$, is the maximum cardinality of any set of vertices on which repeated applications of the color change rule will never result in all vertices joining the set.

We establish bounds on the failed zero forcing number of a graph, both in general and for connected graphs. We also classify connected graphs that achieve the upper bound, graphs whose failed zero forcing numbers are zero or one, and unusual graphs with smaller failed zero forcing number than zero forcing number. We determine formulas for the failed zero forcing numbers of several families of graphs and provide a lower bound on the failed zero forcing number of the Cartesian product of two graphs.

We conclude by presenting open questions about the failed zero forcing number and zero forcing in general.

1. Introduction

The concept of zero forcing has been explored over the past few years because of its application to minimum rank problems in linear algebra [Barioli et al. 2008; 2010]. For an introduction to minimum rank problems, see [Fallat and Hogben 2007]. While we do not discuss the details of minimum rank problems here, the zero forcing number of a graph provides an upper bound on the maximum nullity of any matrix associated with the graph, which in turn leads to a bound on the minimum rank of these matrices. This has led to active research on zero forcing, particularly on graphs for which the minimum rank is difficult to determine. Programs have been developed to determine the zero forcing number of a graph in Sage [DeLoss et al. 2008].

MSC2010: primary 05C15, 05C78, 05C57; secondary 05C50.

Keywords: zero forcing number, vertex labeling, graph coloring.

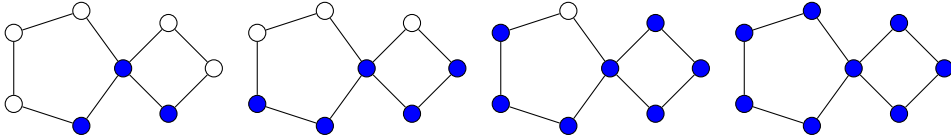


Figure 1. A starting set S and three iterations of the color change rule.

In this paper, we explore the other side of the problem, sets that fail to zero force.

Definitions. Let G be a simple finite graph with vertex set $V(G)$ and edge set $E(G)$. We specify a coloring by choosing a set, usually called S , of vertices. The vertices in the set are filled in, and the others are left blank. Hence, our coloring consists only of two colors: filled, or unfilled. In much of the existing literature, the color black is used to represent filled in, and white is used to represent blank. We simply use *filled* and *unfilled*.

Unlike proper colorings, there are no rules to determine how we choose our initial set or coloring. Instead, we are interested in what happens when we apply the color change rule to our initial set. The standard *color change rule*, as described in [Barioli et al. 2008; 2010] among others, works as follows. Examine each filled vertex, one at a time. If a filled vertex u has exactly one unfilled neighbor, v , then we will fill v at the next iteration. In this case, we say that u *forces* v . Once we have examined all filled vertices, we iterate, and repeat. We repeat this process until no more color changes are possible. In Figure 1 we show a starting set S followed by three iterations of the color change rule.

We use the following term when no more color changes are possible.

Definition 1.1. Let S be a set of vertices in a graph. Suppose that no color changes are possible from S . Then we say that S is *stalled*.

If S is stalled, there are two possible scenarios: either $S = V(G)$ or there are some unfilled vertices that can never be filled. That is, we may be stuck. The two possible conditions under which a set is stalled distinguish a zero forcing set from a failed zero forcing set.

The next two definitions were formalized in [Barioli et al. 2008], although we use slightly different terminology.

Definition 1.2. Let S be a set of vertices in a graph such that repeated applications of the color change rule to S result in all vertices in the graph becoming filled. Then S is a *zero forcing set*.

It is easy to see that $V(G)$ itself is a trivial zero forcing set. The difficult problem is to find the smallest zero forcing set in G . There is considerable work in the literature on this problem, specifically because this parameter provides a bound useful in minimum rank problems [Barioli et al. 2008; 2010].

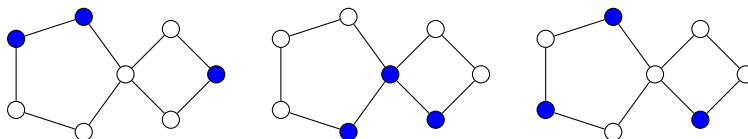


Figure 2. A failed zero forcing set, a zero forcing set, and a stalled failed zero forcing set.

Definition 1.3. The *zero forcing number* of a graph G , denoted $Z(G)$, is the cardinality of a smallest zero forcing set in the graph.

In this paper, we are interested in subsets of a graph's vertex set that are not zero forcing sets. If a set of vertices is not a zero forcing set, then we will call it *failed*.

Definition 1.4. A *failed zero forcing set* is an initial set S of vertices in a graph such that, no matter how many times we apply the color change rule, some vertices in the graph will never be filled.

In Figure 2 we show a failed zero forcing set that is not stalled, a zero forcing set, and a failed zero forcing set that is stalled.

This new concept of failed zero forcing sets is the main topic of this paper. In particular, we are interested in *maximum failed zero forcing sets*, that is, finding failed zero forcing sets of largest cardinality in a graph. We define this parameter.

Definition 1.5. The *failed zero forcing number* of a graph G , denoted $F(G)$, is the maximum cardinality of any failed zero forcing set in the graph.

At times, we will be interested in the concept of *maximal* failed zero forcing sets. Note the difference between maximum and maximal failed zero forcing sets. A maximal failed zero forcing set S is a set of vertices such that adding any other vertex in $V(G)$ to S will change S into a zero forcing set. A maximal failed zero forcing set may not be maximum, but a maximum failed zero forcing set is maximal.

We use the concept of a subgraph, as well as an induced subgraph, in this paper. If G is a graph and G' is a subgraph of G , then $V(G') \subseteq V(G)$, and any two vertices $u, v \in V(G')$ may be adjacent in G' if they are adjacent in $V(G)$, but they may not. If H is an induced subgraph of G , however, then if we have two vertices $u, v \in V(H)$ and $uv \in E(G)$, then $uv \in E(H)$ as well. If $S \subseteq V(G)$, then we use the notation $G[S]$ for the induced subgraph of G with vertex set S .

The concept of a module will be important in this paper.

Definition 1.6. A set X of vertices in $V(G)$ is a *module* if all vertices in X have the same set of neighbors among vertices not in X .

For example, in Figure 3 the set $\{a, b, c\}$ is a module of order 3; $\{b, c\}$ is a module of order 2.

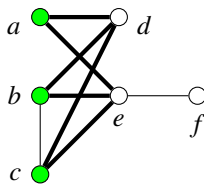


Figure 3. $S = \{a, b, c\}$ is a module.

Throughout the paper, we assume that G is a simple finite graph and use $n = |V(G)|$, unless n is otherwise defined. We now move on to exploring the basic properties of failed zero forcing sets.

Basic properties of failed zero forcing sets. We establish some fundamental observations about failed zero forcing sets. We compare these to known properties of zero forcing sets.

Note that any subset of $V(G)$ is either a zero forcing set or a failed zero forcing set. If S is a zero forcing set of a graph G , then note that any superset of S is also a zero forcing set. We can make a similar statement about failed zero forcing sets.

Observation 1.7. Suppose that G is a graph with failed zero forcing set $S \subseteq V(G)$. Then any subset of S is also a failed zero forcing set.

Next we consider how the color change rule may or may not act on a set of vertices. Let G be a graph, and suppose that S is a proper subset of $V(G)$. If S is a zero forcing set, then a color change must be possible from S . For failed zero forcing sets in general, we cannot make such a statement. For maximum failed zero forcing sets, however, we can.

Observation 1.8. If S is a maximum failed zero forcing set in a graph G , then S is stalled.

To see this, note that since S is a failed zero forcing set, it will not force all vertices in G . Therefore, at some iteration, no more color changes are possible. However, since S is maximum, it must also be maximal. Put simply, if a color change is possible from S , and S is a failed zero forcing set, then clearly, S is not maximum. Hence, any maximum failed zero forcing set S is stalled.

We also note two observations about subgraphs.

Observation 1.9. Let G be a graph with failed zero forcing set S , and let H be a subgraph of G . Then S restricted to H may not be a failed zero forcing set of H .

For example, let $G = P_4$, and let $S = \{v\}$, where v is an internal vertex. If we construct H by deleting the leaf adjacent to v , then S restricted to H is a zero forcing set. However, in a special case, the property is hereditary.

Observation 1.10. Let G be a graph with failed zero forcing set S , and let H be an induced subgraph of G where all vertices in $V(G) \setminus V(H)$ are in S . Then $S \cap V(H)$ is a failed zero forcing set in H .

Goals of this paper. While the zero forcing numbers of many graphs have been determined, the introduction of this relatively new topic has brought with it a large collection of open questions. Note that we can consider zero forcing to be a graph labeling problem with only two labels: *filled* or *unfilled*.

One major difference between zero forcing and other graph labeling problems is that the question of which labelings do *not* work is interesting. In proper coloring, for example, we can construct a failed proper coloring simply by coloring two adjacent vertices the same color. Thus, any graph with an edge has a trivial failed proper coloring. For zero forcing, however, there is no rule to determine how the vertices are labeled. We can choose any starting labeling; whether the labeling is successful or not depends on whether the color change rule leads to all vertices eventually being filled in. Therefore, in general it is not trivial to construct a failed zero forcing set.

Zero forcing opens up a wealth of new problems in graph theory. In this paper, we focus on the failed zero forcing number of different graph families and how these numbers relate to zero forcing numbers. In Section 2, we provide bounds on the failed zero forcing number of a graph, classify graphs with extreme failed zero forcing numbers, such as $F(G) = 0, 1, n - 2$ or $n - 1$, and classify the unusual set of graphs for which $F(G) < Z(G)$. In Section 3, we establish this parameter for several classes of graphs. In Section 4, we explore the failed zero forcing number of the Cartesian product of graphs, including a lower bound in general and determination of the explicit value of the parameter for certain graph families.

We end with a set of open questions about zero forcing in general. While zero forcing numbers have been well studied for their applications to linear algebra, they have also opened up a new area of problems. We list some of these open questions in Section 5.

2. Bounds on failed zero forcing numbers

Whether G is connected or not, there are some fairly immediate bounds on the maximum failed zero forcing number.

Observation 2.1. For any graph G , we have $Z(G) - 1 \leq F(G) \leq n - 1$.

We explain both sides of the inequality here. If $Z(G) - 1 > F(G)$, then any set of order $Z(G) - 1$ forces the graph, contradicting the definition of $Z(G)$ as the minimum order of any zero forcing set. This gives us the lower bound. The upper bound is trivial: if a set S has order $n = V(G)$, then the set is not failed by definition.

It is fairly straightforward to see that $F(G) = n - 1$ if and only if G has an isolated vertex. For the reverse direction, note that if G has an isolated vertex v_0 ,

letting $S = V(G) \setminus \{v_0\}$ makes S a failed zero forcing set. For the forward direction, assume that $F(G) = n - 1$. Then there is some set S of $n - 1$ vertices that does not force the lone vertex $v \in V(G) \setminus S$. If any vertex $u \in S$ is adjacent to v , however, u would force v . Hence, no vertex in S is adjacent to v ; that is, v is an isolated vertex.

Hence, if G is connected, we can improve our bound from Observation 2.1.

Lemma 2.2. *Let G be a connected graph on n vertices where $n \geq 2$. Then*

$$Z(G) - 1 \leq F(G) \leq n - 2.$$

Extreme values. We will show that the upper bound is sharp, that is, that there is a graph G that achieves $F(G) = n - 2$. In fact, we will classify such graphs. First, we prove a related lemma that will help us in classifying graphs with the upper bound.

Lemma 2.3. *Let G be a graph with module X of order $k > 1$. Then $F(G) \geq n - k$.*

Proof. Let $S = V(G) \setminus X$. No vertices in X can be forced by vertices in S since if w is a vertex in S that is adjacent to some vertex $v \in X$, then w is adjacent to all vertices in X , of which there are $k > 1$. Hence, we have found a failed zero forcing set of order $n - k$. \square

Note that if $G[X]$ is connected, we can improve this by letting $S = (V(G) \setminus X) \cup X'$, where X' is a failed zero forcing set of $G[X]$.

We now use Lemma 2.3 to classify connected graphs with failed zero forcing number $n - 2$.

Theorem 2.4. *Let G be connected. Then $F(G) = n - 2$ if and only if G has a module of order 2.*

Proof. Suppose $F(G) = n - 2$. Let S be a maximum failed zero forcing set. Then $V(G) \setminus S = \{u, v\}$ for some vertices u and v . Since neither u nor v can be forced, every neighbor of u in S must also be a neighbor of v , and vice versa. Thus, $\{u, v\}$ is a module of order 2.

The converse follows from Lemmas 2.2 and 2.3. \square

For trees, we can be even more specific.

Corollary 2.5. *Let T be a tree. Then $F(T) = n - 2$ if and only if either T has two leaves adjacent to a single vertex or $T = K_2$.*

Proof. We know by Theorem 2.4 that $F(T) = n - 2$ if and only if T has a module $X = \{u, v\}$ of order 2. If u and v each have two neighbors x and y , then $uxvyu$ forms a cycle; therefore u and v have at most one neighbor. It follows that T has a module $X = \{u, v\}$ if and only if u and v have one or less neighbors. That is, u and v are adjacent to a single common vertex or $T = K_2$. \square

We now examine the lower bound from Lemma 2.2. This is of particular interest because a graph G that achieves this bound has $F(G) < Z(G)$, while we intuitively expect that the failed zero forcing number should be at least as large as the zero forcing number. This property is indeed unusual. Before providing our classification of graphs with $F(G) = Z(G) - 1$, we state two results that will be of use.

Observation 2.6 [Row 2011]. $Z(G) = 1$ if and only if $G = P_n$ for some $n \geq 1$.

Theorem 2.7 [Row 2011]. *Let G be a connected graph with $n = |V(G)| \geq 2$. Then $Z(G) = n - 1$ if and only if $G = K_n$.*

It turns out that complete graphs and their complements are the only graphs with $F(G) < Z(G)$, as we now show.

Theorem 2.8. *For any graph G , $F(G) < Z(G)$ if and only if $G = K_n$ or $G = \bar{K}_n$.*

Proof. We start with the reverse direction. By Theorem 2.7, the zero forcing number of a complete graph is $n - 1$. We also see from Theorem 2.4 that $F(K_n) = n - 2$ since any pair of vertices forms a module. Hence, $F(K_n) = Z(K_n) - 1$. For the null graph (the complement of the complete graph), note that any zero forcing set must consist of the entire vertex set. To fail, we must remove one vertex from this set. Hence, $F(\bar{K}_n) = Z(\bar{K}_n) - 1$.

We now prove the forward direction. Let G be a graph with $F(G) < Z(G)$. Then we know that $F(G) = Z(G) - 1$ by Observation 2.1.

It follows that any set of cardinality $Z(G)$ must be a zero forcing set. Otherwise, we would have a failed zero forcing set of cardinality $Z(G)$, which would contradict our assumption that $F(G) < Z(G)$. Similarly, any set of cardinality $F(G) = Z(G) - 1$ is a failed zero forcing set. Otherwise, we would have a zero forcing set of cardinality $Z(G) - 1$, which contradicts the definition of $Z(G)$.

Let $S \subseteq V(G)$ with $|S| = Z(G)$. If $|S| = 1$, then G is a path P_n by Observation 2.6. By our assumption, any vertex in G is a zero forcing set. But no internal vertex of P_n can force P_n , which means that G has no internal vertices. That is, $n = 2$. Since $P_2 = K_2$, in this case, the proof is complete.

Hence, we assume that $|S| \geq 2$. Now, S is a zero forcing set, which means that either some color change is possible from S or $S = V(G)$. If $S = V(G)$, then by assumption, any set of cardinality $n - 1$ or less fails to force the graph. That is, G must have no edges and is therefore \bar{K}_n , which completes the proof. Otherwise, some color change is possible from S . This means that there exists at least one vertex in S that is adjacent to exactly one vertex in $V(G) \setminus S$. Let S' be this nonempty set of vertices,

$$S' = \{v \in S \mid uv \in E(G) \text{ for exactly one } u \in V(G) \setminus S\}.$$

Let $w \in S$. Note that $S \setminus \{w\}$ is stalled since $|S \setminus \{w\}| = Z(G) - 1 = F(G)$, and we saw above that any set of cardinality $F(G)$ is a maximum failed zero forcing set.

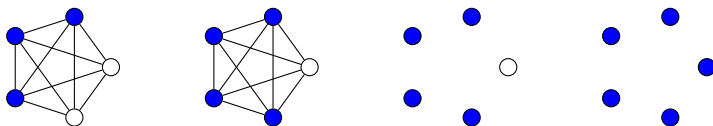


Figure 4. $F(K_5) = 3 < Z(K_5) = 4$; $F(\bar{K}_5) = 4 < Z(\bar{K}_5) = 5$.

Therefore, by Observation 1.8, $S \setminus \{w\}$ is stalled. It follows that w is adjacent to every vertex in S' (except w itself, if $w \in S'$). Hence, every vertex in S is adjacent to every vertex in S' . Additionally, by assumption, any set of cardinality $Z(G) = |S|$ is a zero forcing set. Therefore, we can swap any vertex $u \in V(G) \setminus S$ with any vertex $w \in S$. Therefore, every vertex $u \in V(G) \setminus S$ is adjacent to every vertex in S' .

However, by the definition of S' , each vertex in S' is adjacent to exactly one vertex in $V(G) \setminus S$. Hence, we must have that $|V(G) \setminus S| = 1$. That is, $Z(G) = n - 1$. By Theorem 2.7, $G = K_n$, completing the proof. \square

In Figure 4 we illustrate that the failed zero forcing number of the complete graph on five vertices is less than its zero forcing number and that the failed zero forcing number of the null graph on five vertices is less than its zero forcing number.

Corollary 2.9. *A graph has $F(G) < Z(G)$ if and only if the automorphism group of G is doubly transitive.*

This is a result of the fact that only the complete graph and its complement have doubly transitive automorphism groups [Babai 1995].

Very small values. We have determined which graphs have large failed zero forcing numbers, such as $F(G) = n - 2$ or $n - 1$. We now look at which graphs have very small failed zero forcing numbers.

Theorem 2.10. *Let G be a nonempty graph. Then $F(G) = 0$ if and only if G is either a single vertex or K_2 .*

Proof. The reverse direction is clear: For the case that G is a single vertex v , if we allow v to be in S , then the graph is forced; therefore $F(G) = 0$. For the case that $G = K_2$, allowing either of the vertices to be in S will force the other vertex in the next iteration; therefore $F(G) = 0$.

For the forward case, assume G is a graph with $F(G) = 0$. Then any set $S \subseteq V(G)$ with $|S| = 1$ forces the graph. This means that G consists of a single vertex, or every vertex in G has degree one, and G is connected. But the only connected graph with every vertex of degree one is K_2 . Hence the theorem. \square

Theorem 2.11. *$F(G) = 1$ if and only if G is one of the following graphs: a pair of isolated vertices, K_3 , P_3 or P_4 .*

Proof. The reverse direction is clear: if G is a pair of isolated vertices, then we can pick at most one of them to be in the set, otherwise the graph is trivially forced. If $G = K_3$, then any pair of vertices is a module of order 2, and if $G = P_3$, the end vertices form a module of order 2. Hence, in both cases, $F(G) = n - 2 = 1$. For $G = P_4$, note that a single internal vertex is the largest subset of $V(P_4)$ that is not a zero forcing set.

For the forward direction, assume $F(G) = 1$. If we allow G to be disconnected, it follows that G has at most two maximal connected components because if there are three nonempty components, we can take one vertex each from two of them, and this set will fail to force the third component. Since any pair of vertices in G can force G , each component has at most one vertex because otherwise we could take two vertices in a single component, leaving the other component unforced. Hence, if G is not connected, G is a pair of isolated vertices, and the proof is complete.

We now assume that G is connected. Since $F(G) = 1$, any pair of vertices in G can force. We know that $F(K_n) = n - 2$ for any n ; thus, if $G = K_n$, then $n - 2 = 1$ implies that $G = K_3$, completing the proof. Assume that $G \neq K_n$. Then there is some pair of vertices, u and v , that are not adjacent. Let P be the shortest path from u to v . Since the set $S = \{u, v\}$ forces the graph by assumption, either u or v must force a vertex w in G . Assume without loss of generality that u forces w . Then u is adjacent only to w . Hence, w is the vertex along P that is adjacent to u . The vertex w can force the next vertex along the path (and continue this process) until we reach a vertex w' possibly with $w' = w$, where either w' is adjacent to an unforced vertex not on P in addition to the next vertex on P or the next vertex along P is already forced.

Assume the former. That is, assume that w' is adjacent to the next vertex on P as well as a vertex not on P . Since we assume that S is a zero forcing set, and must therefore eventually force the graph, it follows that one of these two vertices will be forced by some other vertex than w . But so far, the $u - w'$ path is forced, with no vertex except w' adjacent to any other vertex; we also have v forced. Hence, we must have that from v , a sequence of vertices is forced, resulting in one of the two vertices adjacent to w' being forced. Hence, v is only adjacent to a single vertex, and since we assume P is a path from u to v , it must be the vertex along P . By a similar argument, we have that no other vertices along P are adjacent to vertices not on P , except w' . Thus, G consists of P and a set of vertices connected to P only through w' , as in Figure 5, where zigzag lines indicate paths. But then, we could take $S = \{u, w\}$, which will fail to force G , contradicting our assumption.

Similarly, if we assume that u forces a sequence of vertices until reaching w' , which is adjacent to some vertex already forced, we have either $G = P$ or the situation from Figure 5 again, which leads to a contradiction. If $G = P$, then G is a path on n vertices. If $G = P_3$ or $G = P_4$, we're done. Otherwise, we can take any

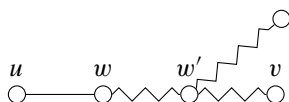


Figure 5. $\{u, v\}$ is a zero forcing set, but $\{u, w\}$ is not.



Figure 6. A maximum failed zero forcing set of P_8 .

pair of nonadjacent vertices of degree 2 in G to be S , contradicting our assumption. Hence, $G = P_3$ or P_4 . \square

There are many examples of graphs with $F(G) = 2$. For example, three isolated vertices, two copies of K_2 , or an isolated vertex and K_2 all have failed zero forcing numbers of 2. Also, any connected graph G on four vertices, except for P_4 , has $F(G) = 2$, as does any connected graph on five vertices that does not have a module of order two, such as P_5 or the house graph. However, there are many such graphs. We stop at $F(G) = 1$ and move on to determining failed zero forcing numbers of different families of graphs.

3. Failed zero forcing numbers of various families of graphs

We have already seen the failed zero forcing numbers of several graphs, including that of complete graphs, $F(K_n) = n - 2$. We now consider several families of graphs, including paths, cycles, complete bipartite graphs, binary trees, wheels, and the Petersen graph. We also give a formula for the failed zero forcing number of graphs with multiple connected components.

Theorem 3.1. *The failed zero forcing number of a path P_n on n vertices is*

$$F(P_n) = \left\lceil \frac{n-2}{2} \right\rceil.$$

Proof. If S is a failed zero forcing set in P_n , then neither end vertex is in S because either end vertex is a zero forcing set. Further, S contains no pairs of adjacent vertices because any pair of adjacent vertices is a zero forcing set. Therefore, S can have at most $\lceil (n-2)/2 \rceil$ vertices in it. We construct such a set by starting with the vertex adjacent to either end vertex in P_n and adding it to S . From there, we take every other vertex until we reach the other end vertex, which we do not add to S . Thus, $|S| = \lceil (n-2)/2 \rceil$, and it does not force the graph because every vertex in S has exactly two neighbors not in S . \square

In Figure 6, the construction of a maximum failed zero forcing set in P_8 is shown.

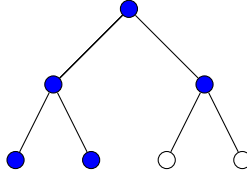


Figure 7. The failed zero forcing number of a binary tree with n vertices is $n - 2$.

Theorem 3.2. *The failed zero forcing number of a cycle C_n on n vertices is*

$$F(C_n) = \left\lfloor \frac{n}{2} \right\rfloor.$$

Proof. Suppose S is a failed zero forcing set. Then there are no adjacent vertices in S since any pair of adjacent vertices forces C_n . Hence, $|S| \leq \lfloor n/2 \rfloor$. We can construct such a set by starting with any vertex in C_n and adding every other vertex to S . Since every vertex in S has two neighbors in $V(G) \setminus S$, the set S will not force the graph. Therefore, $F(C_n) = \lfloor n/2 \rfloor$. \square

We use $K_{m,n}$ to denote the complete bipartite graph with partite sets V_1 and V_2 , where $|V_1| = m \geq 1$ and $|V_2| = n \geq 1$.

Theorem 3.3. *If $m + n \geq 3$, then $F(K_{m,n}) = m + n - 2$.*

Proof. Since $m + n \geq 3$, it follows that $m \geq 2$ or $n \geq 2$. Without loss of generality, assume that $n \geq 2$. Then any pair of vertices in V_2 is a module of order 2 since both vertices have the same sets of neighbors, V_1 . Hence, by Theorem 2.4, $F(K_{m,n}) = m + n - 2$. \square

A full m -ary tree T is a rooted tree whose vertices have m or 0 children, where m is a positive integer of at least 2. Note that if $m = 2$, then T is a full binary tree.

Theorem 3.4. *The failed zero forcing number of a full m -ary tree T with $n > 1$ is $F(T) = n - 2$.*

Proof. Take any two vertices u and v of degree one that have the same parent, w . We know that u and v exist because T is finite and $m \geq 2$. Then, u and v form a module of order two because they each have exactly the same neighbor, w . Hence, by Theorem 2.4, $F(T) = n - 2$. \square

In Figure 7, a binary tree with a maximum failed zero forcing set is shown.

The *join* of graphs G_1 and G_2 , denoted $G_1 \vee G_2$, consists of a copy of G_1 , a copy of G_2 , and an edge between every pair of vertices u and v such that $u \in V(G_1)$ and $v \in V(G_2)$.

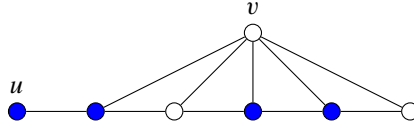


Figure 8. The graph G if $k = 1$.

Lemma 3.5. *Let G be a connected graph, and let $H = G \vee \{v_0\}$. That is, H consists of G and a single vertex v_0 that is adjacent to all vertices in G . Then $F(H) \geq F(G) + 1$.*

Proof. Let $S \subsetneq V(G)$ be stalled. Let $S' \subseteq V(H)$ be defined $S' = S \cup \{v_0\}$.

Since S is a failed zero forcing set in G , there are at least two vertices $u, v \in V(G)$ that are not forced by S . Any vertex in $S' \setminus \{v_0\}$ that is adjacent to v in H must also be adjacent to some other unforced vertex, otherwise it would force v in G . Also, v_0 is adjacent to both v and u , so it will not force v . Hence, S' is a failed zero forcing set of H . Since $|S'| = |S| + 1$, we have that $F(H) \geq F(G) + 1$. \square

For any positive integer k , we can construct a graph G such that $F(G \vee \{v_0\}) \geq F(G) + k$. Let G consist of a path P_l , where $l = 3(k + 1)$, and a vertex v that is adjacent to all vertices in P_l except for one end vertex, u . An example of G for $k = 1$ is shown in Figure 8. We claim that $F(G) \leq \lfloor (2/3)l \rfloor$. First, suppose that S is a maximum failed zero forcing set. If $v \in S$, then no adjacent vertices from the path can be in S and neither end vertex can be in the path. Hence, if $v \in S$, this implies that $F(G) \leq \lfloor l/2 \rfloor + 1$. If $v \notin S$, then no more than two consecutive vertices on the path can be in S because if three are in S , then the middle vertex will force v . Hence, $F(G) \leq \lfloor (2/3)l \rfloor$. Since $l = 3(k + 1) \geq 6$, it follows that $\lfloor l/2 \rfloor + 1 \leq \lfloor (2/3)l \rfloor$. Hence, $F(G) \leq \lfloor (2/3)l \rfloor$.

Letting $H = G \vee \{v_0\}$ for a single vertex v_0 , however, we find that $F(H) \geq l - 1$. Let $S = P_l \setminus \{u\}$. That is, $S = H \setminus \{u, v, v_0\}$. Note that S is stalled because every vertex in S is adjacent to both v and v_0 , which are not in S . Hence, S is a failed zero forcing set. Thus, we have that $F(H) \geq l - 1 = 3(k + 1) - 1 = 3k + 2$, and $F(G) \leq \lfloor (2/3)l \rfloor = 2k + 2$. Thus, $F(H) - F(G) \geq k$.

Therefore, joining an additional vertex to a graph will certainly increase the failed zero forcing number of the graph, and the increase may be large.

We use Lemma 3.5 to examine another graph family. Let W_n be a wheel on $n + 1$ vertices consisting of C_n and an additional vertex v_0 adjacent to all vertices in C_n .

Theorem 3.6. *Let $n \geq 3$. Then $F(W_n) = \lfloor 2n/3 \rfloor$ if $n \neq 4$, and $F(W_4) = 3$.*

Proof. We know by Theorem 3.2 and Lemma 3.5 that $F(W_n) \geq \lfloor n/2 \rfloor + 1$. We construct a failed zero forcing set S on W_n as follows. Starting with any vertex along the cycle, add the vertex and one of its neighbors to S . Continuing around the cycle, leave out the third vertex, add the next two to S , and leave out the next,

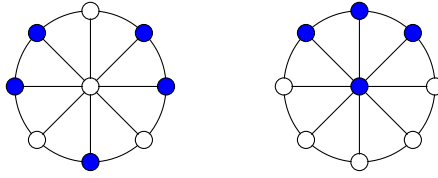


Figure 9. Left: A failed zero forcing set in W_8 . Right: A zero forcing set in W_8 .

as long as vertices are remaining, making sure at the end not to add any three consecutive vertices to S . Also, $v_0 \notin S$. Figure 9 shows this construction for W_8 . Since there are no three consecutive vertices along the cycle in our set and v_0 is not in this set, it follows that every vertex in the set is adjacent to at least two vertices not in the set: v_0 and one vertex along the cycle. Hence, $F(W_n) \geq \lfloor 2n/3 \rfloor$.

First, consider the special case that $n = 4$. Since $F(W_n) \geq \lfloor n/2 \rfloor + 1$, we know that $F(W_4) \geq 3$, but since $|V(W_4)| = 5$, by Lemma 2.2, we know that $F(W_4) \leq 3$. Hence, $F(W_4) = 3$.

We continue with the remaining cases, assuming for the remainder of the proof that $n \neq 4$. If $n \geq 6$, then $\lfloor 2n/3 \rfloor \geq \lfloor n/2 \rfloor + 1$ because

$$\lfloor 2n/3 \rfloor = \lfloor n/2 + n/6 \rfloor \geq \lfloor n/2 \rfloor + 1.$$

Also, for the special cases $n = 3$ and $n = 5$, we see that $\lfloor 2n/3 \rfloor = 2$ and 3 respectively. Similarly, for the same cases of $n = 3$ and $n = 5$, we have $\lfloor n/2 \rfloor + 1 = 2$ and 3 respectively. Hence, if $n \neq 4$, we know that $F(W_n) \geq \lfloor 2n/3 \rfloor \geq \lfloor n/2 \rfloor + 1$.

Before proceeding, note that if at least three consecutive vertices along the cycle are in S and v_0 is in S , then S is a zero forcing set, as shown in Figure 9.

Finally, we show that $F(W_n) \leq \lfloor 2n/3 \rfloor$. Let S be a set of vertices in W_n with $|S| > \lfloor 2n/3 \rfloor$. Then, either $v_0 \in S$ or there is some set of at least three consecutive vertices along the cycle that are in S . Assume that $v_0 \in S$. Since $|S| > \lfloor 2n/3 \rfloor$, there exists at least one pair of adjacent vertices along the cycle. Let u be in one such pair. If both neighbors of u along the cycle are in S , then we know that S is a zero forcing set and we're done. Otherwise, u has exactly one neighbor, w , not in S . Then, u will force w in the next iteration, and S is not a maximum zero forcing set.

The last possibility is that there is some set of three consecutive vertices, v_1, v_2 , and v_3 , along the cycle that are in S . Then v_0 is the only neighbor of v_2 that is not in S . Hence, v_2 forces v_0 in the next iteration, and S is a zero forcing set.

Hence, if $n \neq 4$, we have that $F(W_n) = \lfloor 2n/3 \rfloor$. □

We have found the failed zero forcing numbers of several families of graphs. We now describe how the failed zero forcing number of a disconnected graph can be determined by the failed zero forcing numbers and orders of its components.

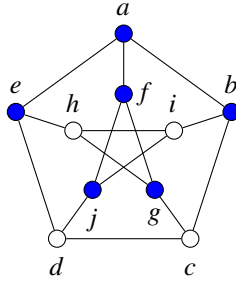


Figure 10. The Petersen graph with a maximum failed zero forcing set.

Theorem 3.7. *Let G be a disconnected graph, that is, a graph with at least two disjoint maximal connected components. Let G_1, G_2, \dots, G_k be the k maximal connected components of the graph. Then*

$$F(G) = \max_k \left(F(G_k) + \sum_{l \neq k} |V(G_l)| \right).$$

Proof. Since the graph is disconnected, if we allow S to consist of all vertices in the graph except those vertices in the component G_k , then clearly, G_k is not forced. We can add any failed zero forcing set of G_k to S , and still not force all vertices in G_k . This will work for any component G_k . Hence, we can pick the component that maximizes the cardinality of the set S . If S' is a set with

$$|S'| > \max_k \left(F(G_k) + \sum_{l \neq k} |V(G_l)| \right),$$

then every component G_l must have $|S' \cap V(G_l)| > F(G_l)$, forcing all components. Hence, S' is a zero forcing set. \square

Theorem 3.8. *Let G be the Petersen graph. Then $F(G) = 6$.*

Proof. We can find a failed zero forcing set of cardinality six: for example, let $S = \{a, b, e, f, g, j\}$, as in Figure 10. This is clearly a failed zero forcing set, since a and f have all three neighbors in S , while all other vertices have exactly two neighbors in $V(G) \setminus S$. Hence, $F(G) \geq 6$.

To prove that $F(G) \leq 6$, suppose S is a maximum failed zero forcing set, and $|S| \geq 7$. By the pigeonhole principle, there are at least four vertices in S that are in the cycle $\{a, b, c, d, e\}$ or in $\{f, g, h, i, j\}$. Since there is an automorphism between these sets, assume without loss of generality that there are at least four from the set $\{a, b, c, d, e\}$ in S . Note that all five vertices cannot be in S because this would force the entire graph. Because of the symmetry, we can assume that $\{a, b, c, d\} \subseteq S$. Since S is a maximum failed zero forcing set, S is stalled. Thus, we must have

$i, g \in S$. Otherwise, b and c would force them. Now, i and g each have exactly two neighbors remaining that we have not assigned to S . These neighbors are f, h and j . If any one of f, h or j is in S , the others will be forced, which will force the graph. Hence, $\{f, h, j\} \subseteq V(G) \setminus S$. But we already know that $e \in V(G) \setminus S$ for the same reason, leaving us with $|S| = 6$. Hence, $F(G) = 6$. \square

4. Cartesian products

We first give a bound on the failed zero forcing number of a Cartesian product graph in terms of the failed zero forcing numbers of the graphs in the product.

Theorem 4.1. *For any graphs G and H ,*

$$F(G \square H) \geq \max\{F(G)|V(H)|, F(H)|V(G)|\}.$$

Proof. Consider the Cartesian product $G \square H$, where $n = |V(G)|$ and $k = |V(H)|$. Label the vertices of G , u_1 through u_n and the vertices of H , w_1 through w_k . We refer to each vertex in $G \square H$ as $v_{i,j}$ where i denotes in which copy of G and j denotes in which copy of H the vertex lies.

Let S be a stalled failed zero forcing set in G . We construct a stalled failed zero forcing set S' in $G \square H$ as follows. Suppose $u_\alpha \in S$. Then for all $i \in \{1, 2, \dots, k\}$, let $v_{i,\alpha} \in S'$. Then $|S'| = |S||V(H)|$. We show that S' is a failed zero forcing set of $G \square H$.

Suppose $v_{i,\alpha}$ is in S' . Then $u_\alpha \in S$ by construction. Since S is a failed zero forcing set in G , then u_α is either adjacent to no vertices in $V(G) \setminus S$ or u_α is adjacent to two or more vertices in $V(G) \setminus S$, u_β and u_γ . In this latter case, it follows that $v_{i,\alpha}$ is adjacent to $v_{i,\beta}$ and $v_{i,\gamma}$ as well. In the former case, if u_α is adjacent to no vertices in $V(G) \setminus S$, then any neighbors of $v_{i,\alpha}$ of the form $v_{i,j}$ for some j are in S' . Since $v_{i,\alpha} \in S'$ for all i by construction, it follows that $v_{i,\alpha}$ has no neighbors in $V(G \square H) \setminus S'$. Thus, S' is a stalled failed zero forcing set.

Since this construction works for any stalled failed zero forcing set in G , and similarly in H , it follows that we can construct in $G \square H$ a failed zero forcing set of cardinality $F(G)|V(H)|$ and similarly a failed zero forcing set of cardinality $F(H)|V(G)|$. Hence the result. \square

Note that the above bound is sharp if $G = P_2$ and $H = K_n$ for $n \geq 4$. Recall that $F(K_n) = n - 2$ and $F(P_2) = 0$. Thus, $\max\{F(G)|V(H)|, F(H)|V(G)|\} = 2(n - 2)$. If we try to construct a failed zero forcing set S of $G \square H$ with more vertices than $2(n - 2)$, by the pigeonhole principle, one copy of K_n must have at least $n - 1$ vertices in S . If one copy has n vertices, then $G \square H$ is forced. Therefore, one copy, H_1 , must have $n - 1$ vertices, and the other, H_2 , then must have at least $n - 2$ in S . So every vertex in $S \cap V(H_1)$ is adjacent to the single vertex v in $V(H_1) \setminus S$. That means that every vertex in $S \cap V(H_1)$ must have at least one neighbor in $V(H_2)$ that is not

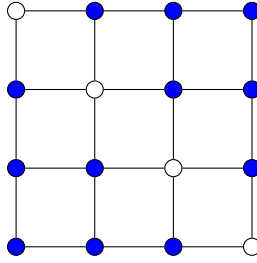


Figure 11. A maximum failed zero forcing set of the square grid $P_4 \square P_4$.

in S . But there are only at most two such vertices, and each has one distinct neighbor in H_1 . Since $n \geq 4$, we know that $n - 1 \geq 3$, which means that there is at least one vertex in $S \cap V(H_1)$ that has only one neighbor v in $V(G \square H) \setminus S$. Thus, v will be forced, which means that $V(H_1)$ will be completely forced, which will in turn force all vertices in the graph. Hence, $F(P_2 \square K_n) = \max\{F(G)|V(H)|, F(H)|V(G)|\}$, showing that our bound from Theorem 4.1 is sharp.

For most cases, the failed zero forcing number of a Cartesian product of graphs is much greater than our bound. The following theorem establishes an exact value for the square grid graph.

Theorem 4.2. *Let $n \geq 2$. The failed zero forcing number of a square grid, $P_n \square P_n$, is $F(P_n \square P_n) = n^2 - n$.*

Proof. We can construct such a failed zero forcing set by putting in the set every vertex in the graph, except those vertices along a single main diagonal. That is, if we label every vertex in the graph $v_{i,j}$, where i denotes the row and j denotes the column of the vertex, we let $v_{i,j}$ be in S if and only if $i \neq j$. See Figure 11 for an example.

We will show that S is indeed a failed zero forcing set. The only vertices that can be forced—because they are not in S —are $v_{i,i}$ for $i = 1, 2, \dots, n$. Take any such $v_{i,i}$. Then $v_{i,i}$ is adjacent to four vertices: $v_{i,i+1}$, $v_{i+1,i}$, $v_{i,i-1}$ and $v_{i-1,i}$. Note that if $i = 1$ or $i = n$, only the first two or the last two vertices (respectively) will be adjacent to $v_{i,i}$.

Now, $v_{i,i+1}$ is also adjacent to $v_{i+1,i+1}$, as is $v_{i+1,i}$. Therefore, neither $v_{i,i+1}$ nor $v_{i+1,i}$ will force $v_{i,i}$. Similarly, $v_{i,i-1}$ and $v_{i-1,i}$ are both also adjacent to $v_{i-1,i-1}$, and therefore do not force $v_{i,i}$. Hence, the set is a failed zero forcing set.

It remains to show that S is a maximum zero forcing set. We will show that any set S' with cardinality $|S'| > n^2 - n$ is not a failed zero forcing set.

By the pigeonhole principle, if $|S'| > n^2 - n$, there must be a column in G , say column \hat{j} , such that $v_{i,\hat{j}} \in S'$ for all $i = 1, 2, \dots, n$. Note that $1 < \hat{j} < n$ because any end column alone would force G . If every vertex in the first row to the right of column \hat{j} is in S' , then the entire graph will be forced; similarly, if all vertices in

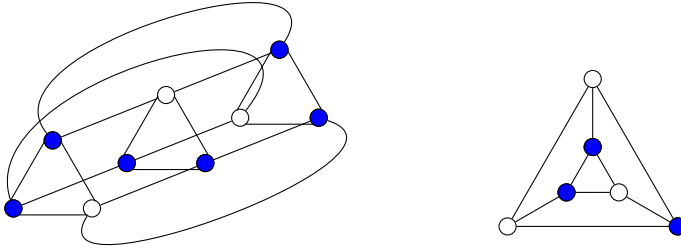


Figure 12. Failed zero forcing sets of $K_3 \square K_3$ and of $K_3 \square K_2$.

any two adjacent rows to the right of column \hat{j} are in S' , then the entire graph will be forced. The same holds for the left of column \hat{j} . Therefore, there are at least $\lceil (n + 1)/2 \rceil$ vertices not in S' on either side of column \hat{j} , or at least $n + 1$ in G , contradicting the assumption that $|S'| > n^2 - n$.

Therefore, $F(P_n \square P_n) = n^2 - n$. □

The same construction works for $P_n \square P_m$ if $m = n + (n - 1)k$ for a positive integer k . However, the construction does not work for rectangular grids in general.

For Cartesian products of complete graphs, $K_n \square K_m$, we have determined $F(K_n \square K_m)$ for all cases. Before providing the general result, we must look at two special cases, $K_3 \square K_2$ and $K_3 \square K_3$. Figure 12 shows failed zero forcing sets for each graph. To see that these are optimal, note that neither graph has a module of order two; therefore $F(K_3 \square K_2) \leq 3$ and $F(K_3 \square K_3) \leq 6$, coinciding with the construction in Figure 12. We now move on to determining $F(K_n \square K_m)$ in general.

Theorem 4.3. *The failed zero forcing number of the rook's graph, $K_n \square K_m$, is $F(K_n \square K_m) = nm - 4$, where $n \geq 4$ and $m \geq 2$.*

Proof. First, we construct a failed zero forcing set S in $G = K_n \square K_m$ with cardinality $nm - 4$. Let each vertex in the graph be labeled $v_{i,j}$, where i denotes in which copy of K_n and j denotes in which copy of K_m the vertex lies. Let all vertices be in S except $v_{1,1}, v_{1,2}, v_{2,1}$ and $v_{2,2}$.

We show that S is a failed zero forcing set. The only vertices in S that are adjacent to the vertices in $V(G) \setminus S$ are vertices $v_{1,k}, v_{2,k}, v_{l,1}$ and $v_{l,2}$, where $3 \leq k \leq n$ and $3 \leq l \leq m$. However, $v_{1,k}$ is adjacent to both $v_{1,1}$ and $v_{1,2}$; $v_{2,k}$ is adjacent to both $v_{2,1}$ and $v_{2,2}$; $v_{l,1}$ is adjacent to both $v_{l,1}$ and $v_{2,1}$; finally, $v_{l,2}$ is adjacent to both $v_{1,2}$ and $v_{2,2}$. Therefore, S is a failed zero forcing set.

We now show that there is no failed zero forcing set larger than S . First, we show that there is no module of order 2 in G . Any two vertices in the same copy of K_n share all the same neighbors in K_n but lie in different copies of K_m and therefore have some distinct neighbors. Similarly, any two vertices in different copies of K_n have different neighbors in their respective copies of K_n . Hence, there is no module of order 2 in G , giving us, by Theorem 2.4, that $nm - 4 \leq F(G) \leq nm - 3$.

Suppose that S' is a set of vertices in G of cardinality $nm - 3$. Let $V(G) \setminus S' = \{x, y, z\}$. If $\{x, y, z\}$ is contained in a single copy of K_n , then take any other copy of K_n . There exist vertices x', y' and z' in this copy such that x' is adjacent to x but not y or z , and similarly for y' and z' . Thus, x' will force x , y' will force y , and z' will force z .

If there exists a copy of K_n , call it H , such that $V(H) \setminus S'$ consists of exactly one vertex, z , then z will be forced by another vertex in H because at most two of the vertices in H can be adjacent to an unforced vertex in any other copy of K_n . Since $\{x, y\}$ is not a module, it will be forced as well. Hence, $F(G) = nm - 4$. \square

5. Conclusion and open questions

In this paper, we have defined a new graph parameter, the failed zero forcing number $F(G)$, and established some properties of this parameter as well as the value of this parameter for several families of graphs. There are many questions about this parameter that remain. More generally, there are many questions that remain about the concept of zero forcing in general. We outline some of these questions here.

As we touched on in the introduction of this paper, the motivation for study of the zero forcing number is minimum rank problems. The maximum nullity of a set of matrices associated with a graph is bounded above by the zero forcing number of the graph. We would like to know if the failed zero forcing number has any such connection to linear algebra.

There are many graph families whose failed zero forcing numbers are unknown. For example, while we found a value of $F(P_n \square P_n)$, we have no formula for $F(P_n \square P_m)$ in general. Also, we have not determined the failed zero forcing number of $C_n \square C_m$ or any Cartesian products of pairs of graphs from different graph families, such as paths and cycles. We know for certain trees — those who have two leaves adjacent to the same vertex — we have $F(T) = n - 2$. Trees in general, however, are open.

We can also look at graphs with failed zero forcing number of 2. We characterized graphs with $F(G) = 0$ or 1, but many more graphs have $F(G) = 2$. While we listed some of these, it would be nice to have a full characterization of all graphs with this property. More generally, given a positive integer k , is there an integer l such that any graph G with $|V(G)| > l$ has $F(G) > k$?

Many graph labeling problems that search for the minimum number of labels required for a given graph are accompanied by a second question: what is the maximum cardinality of any *minimal* labeling? In proper coloring, this is known as the achromatic number [Harary and Hedetniemi 1970]. Failed zero forcing has an analogous problem: the *failed zero forcing number*. The question is: what is the minimum cardinality of any maximal failed zero forcing set?

Finally, while we were able to classify graphs for which $F(G) < Z(G)$, it would be interesting to classify graphs for which $F(G) = Z(G)$, since these graphs seem to be unusual.

References

- [Babai 1995] L. Babai, “Automorphism groups, isomorphism, reconstruction”, Chapter 27, pp. 1447–1540 in *Handbook of combinatorics*, vol. 1, edited by R. L. Graham et al., Elsevier, Amsterdam, 1995. MR 97j:05029 Zbl 0846.05042
- [Barioli et al. 2008] F. Barioli, W. Barrett, S. Butler, S. M. Cioabă, D. Cvetković, S. M. Fallat, C. Godsil, W. Haemers, L. Hogben, R. Mikkelsen, S. Narayan, O. Pryporova, I. Sciriha, W. So, D. Stevanović, H. van der Holst, K. Vander Meulen, and A. Wangsness, “Zero forcing sets and the minimum rank of graphs”, *Linear Algebra Appl.* **428**:7 (2008), 1628–1648. MR 2008m:05166 Zbl 1135.05035
- [Barioli et al. 2010] F. Barioli, W. Barrett, S. M. Fallat, H. T. Hall, L. Hogben, B. Shader, P. van den Driessche, and H. van der Holst, “Zero forcing parameters and minimum rank problems”, *Linear Algebra Appl.* **433**:2 (2010), 401–411. MR 2011g:15002 Zbl 1209.05139
- [DeLoss et al. 2008] L. DeLoss, J. Grout, T. McKay, J. Smith, and G. Tims, “Program for calculating bounds on the minimum rank of a graph using Sage”, preprint, 2008. arXiv 0812.1616
- [Fallat and Hogben 2007] S. M. Fallat and L. Hogben, “The minimum rank of symmetric matrices described by a graph: a survey”, *Linear Algebra Appl.* **426**:2-3 (2007), 558–582. MR 2008f:05114 Zbl 1122.05057
- [Harary and Hedetniemi 1970] F. Harary and S. Hedetniemi, “The achromatic number of a graph”, *J. Combinatorial Theory* **8** (1970), 154–161. MR 40 #7143 Zbl 0195.25702
- [Row 2011] D. D. Row, *Zero forcing number: results for computation and comparison with other graph parameters*, Ph.D. thesis, Iowa State University, Ames, IA, 2011, available at <http://lib.dr.iastate.edu/etd/12003>. MR 2941885

Received: 2013-06-21 Revised: 2013-07-30 Accepted: 2013-08-04

kjf3239@rit.edu *Department of Civil Engineering Technology, Environmental Management and Safety, Rochester Institute of Technology, 1 Lomb Memorial Drive, Rochester, NY 14623, United States*

bonnie.jacob@rit.edu *Science and Mathematics Department, National Technical Institute for the Deaf, Rochester Institute of Technology, 52 Lomb Memorial Drive, Rochester, NY 14623, United States*

dxs6040@rit.edu *Department of Packaging Science, College of Applied Science and Technology, Rochester Institute of Technology, 1 Lomb Memorial Drive, Rochester, NY 14623, United States*

An Erdős–Ko–Rado theorem for subset partitions

Adam Dyck and Karen Meagher

(Communicated by Glenn Hurlbert)

A kl -subset partition, or (k, l) -subpartition, is a kl -subset of an n -set that is partitioned into l distinct blocks, each of size k . Two (k, l) -subpartitions are said to t -intersect if they have at least t blocks in common. In this paper, we prove an Erdős–Ko–Rado theorem for intersecting families of (k, l) -subpartitions. We show that for $n \geq kl$, $l \geq 2$ and $k \geq 3$, the number of (k, l) -subpartitions in the largest 1-intersecting family is at most $\binom{n-k}{k} \binom{n-2k}{k} \cdots \binom{n-(l-1)k}{k} / (l-1)!$, and that this bound is only attained by the family of (k, l) -subpartitions with a common fixed block, known as the *canonical intersecting family of (k, l) -subpartitions*. Further, provided that n is sufficiently large relative to k, l and t , the largest t -intersecting family is the family of (k, l) -subpartitions that contain a common set of t fixed blocks.

1. Introduction

We prove here an Erdős–Ko–Rado theorem for intersecting families of subset partitions. The EKR theorem gives the size and structure of the largest family of intersecting sets, all of the same size, from a base set. This theorem has an interesting history: Erdős [1987] wrote that the work was done in 1938, but due to lack of interest in combinatorics at the time, it wasn't until 1961 that the paper was published. Once the result did appear in the literature, it sparked a great deal of interest in extremal set theory.

To start, we must consider some relevant notation and background information. For any positive integer n , denote $[n] := \{1, \dots, n\}$. A k -set is a subset of size k from $[n]$. Two k -sets A and B are said to *intersect* if $|A \cap B| \geq 1$, and for $1 \leq t \leq k$, they are said to be t -intersecting if $|A \cap B| \geq t$. A *canonical t -intersecting family of k -sets* is one that contains all k -sets with t fixed elements.

EKR theorem. [Erdős et al. 1961] *Let $n \geq k \geq t \geq 1$, and let \mathcal{F} be a t -intersecting family of k -sets from $[n]$. If n is sufficiently large compared to k and t , then*

MSC2010: 05D05.

Keywords: Erdős–Ko–Rado theorem, set partitions.

Meagher is supported by NSERC.

$|\mathcal{F}| \leq \binom{n-t}{k-t}$; further, equality holds if and only if \mathcal{F} is a canonical t -intersecting family of k -sets.

The exact bound on n is known to be $n \geq (t+1)(k-t+1)$ (an elegant proof of this that uses algebraic graph theory is given by Wilson [1984]). If n is smaller than this bound, then there are t -intersecting families that are larger than the canonical t -intersecting family. A complete characterization of the families of maximum size for all values of n is given by Ahlswede and Khachatrian [1997].

From here, many EKR-type theorems have been developed by incorporating other combinatorial objects. Frankl and Wilson [1986] have considered this theorem for vector spaces over a finite field, Rands [1982] for blocks in a design, Cameron and Ku [2003] for permutations, Ku and Leader [2006] for partial permutations, Brunk and Huczynska [2010] for injections, and Ku and Renshaw [2008] for set partitions and cycle-intersecting permutations. All of these cases consider combinatorial objects that are made up of what we shall call *atoms*, and two objects intersect if they contain a common atom and t -intersect if they contain t common atoms. To say that “an EKR-type theorem holds” means that the largest set of intersecting (or t -intersecting) objects is the set of all objects that contain a common atom (or a common t -set of atoms).

In this paper, we shall prove that an EKR-type theorem holds for an object which we call a *subset partition*. We begin by outlining the appropriate notation.

A *uniform l -partition* of $[n]$ is a division of $[n]$ into l distinct, nonempty subsets, known as *blocks*, where each block has the same size and the union of these blocks is $[n]$. Further, a *uniform kl -subset partition* P is a uniform l -partition of a subset of kl elements from $[n]$. We shall also call P a *(k, l) -subpartition*. If P is a (k, l) -subpartition of $[n]$, then $P = \{P_1, \dots, P_l\}$ and $|P_i| = k$ for $i \in \{1, \dots, l\}$, with $|\bigcup_{i=1}^l P_i| = kl$. Let $U_{l,k}^n$ denote the set of all (k, l) -subpartitions from $[n]$, and define

$$U(n, l, k) := |U_{l,k}^n| = \frac{1}{l!} \binom{n}{k} \binom{n-k}{k} \cdots \binom{n-(l-1)k}{k} = \frac{1}{l!} \prod_{i=0}^{l-1} \binom{n-ik}{k}.$$

Two (k, l) -subpartitions $P = \{P_1, \dots, P_l\}$ and $Q = \{Q_1, \dots, Q_l\}$ are said to be *intersecting* if $P_i = Q_j$ for some $i, j \in \{1, \dots, l\}$. Further, for $1 \leq t \leq l$, P and Q are said to be *t -intersecting* if there is an ordering of the blocks such that $P_i = Q_i$ for $i = 1, \dots, t$.

A *canonical t -intersecting family of (k, l) -subpartitions* is a family that contains every (k, l) -subpartition with a fixed set of t blocks. Such a family has size

$$U(n - tk, l - t, k) = \frac{1}{(l-t)!} \prod_{i=t}^{l-1} \binom{n-ik}{k}. \quad (*)$$

In particular, a *canonical intersecting family of (k, l) -subpartitions* has size

$$U(n - k, l - 1, k) = \frac{1}{(l - 1)!} \prod_{i=1}^{l-1} \binom{n - ik}{k}. \tag{**}$$

Finally, note that

$$U(n, l, k) = \frac{1}{l} \binom{n}{k} U(n - k, l - 1, k), \tag{†}$$

and $U(n, 0, 0) = 1$ for $n \geq 0$.

We shall not consider the cases when $k = 1$, as this reduces to the original EKR theorem when $l = 1$, where intersection is trivial, or when $t = l$, where intersection is also trivial.

Theorem 1. *Let n, k, l be positive integers with $n \geq kl, l \geq 2$, and $k \geq 3$. If \mathcal{P} is an intersecting family of (k, l) -subpartitions, then*

$$|\mathcal{P}| \leq \frac{1}{(l - 1)!} \prod_{i=1}^{l-1} \binom{n - ik}{k}.$$

Moreover, this bound can only be attained by a canonical intersecting family of (k, l) -subpartitions.

Theorem 2. *Let n, k, l, t be positive integers with $n \geq n_0(k, l, t)$ and $1 \leq t \leq l - 1$. If \mathcal{P} is a t -intersecting family of (k, l) -subpartitions, then*

$$|\mathcal{P}| \leq \frac{1}{(l - t)!} \prod_{i=t}^{l-1} \binom{n - ik}{k}.$$

Moreover, this bound can only be attained by a canonical t -intersecting family of (k, l) -subpartitions.

Meagher and Moura [2005] introduced Erdős–Ko–Rado theorems for t -intersecting partitions, which fall under the case $n = kl$. Additionally, for the case $k = 2$ with $n > kl$, a (k, l) -subpartition is a partial matching; in their recent paper, Kamat and Misra [2013] presented the corresponding EKR theorems for these objects. They incorporate a very nice Katona-style proof, but interestingly, it does not appear that the Katona method would work very well for (k, l) -subpartitions (it seems that this proof would require an additional lower bound on n). The goal of this work is to complete the work done in both [Meagher and Moura 2005] and [Kamat and Misra 2013] by showing that an EKR-type theorem holds for subpartitions. In this paper, we specifically do not consider the case where $k = 2$ (as this is done in Kamat and Misra’s work). In [Meagher and Moura 2005], the only difficult case is $k = 2$; it is possible that our counting method will work for the partial matchings if some of the tricks used in [loc. cit.] are applied.

2. Three technical lemmas

We shall require results similar to Lemma 3 in [Meagher and Moura 2005] — the proofs of which use similar counting arguments. The first of these, Lemma 3, is just the $t = 1$ case of the third, Lemma 5. We present proofs for both of these lemmas since the proof of Lemma 3 is straight-forward and presenting it first makes the proof of Lemma 5 clearer.

As we shall see, it is worthwhile to consider the size of a canonical t -intersecting family of (k, l) -subpartitions and find when this is an upper bound for the size of any t -intersecting family of (k, l) -subpartitions.

Define a *dominating set* for a family of (k, l) -subpartitions to be a set of blocks, each of size k , that intersects with every (k, l) -subpartition in the family. For the intersecting families being investigated here, each (k, l) -subpartition in the family is also a dominating set. In [Meagher and Moura 2005], dominating sets are called *blocking sets*. We use the term dominating set here because if the blocks in the (k, l) -subpartitions (the k -sets) are considered to be vertices, then each (k, l) -subpartition can be thought of as an edge in an l -uniform hypergraph on these vertices. As a result, a family of (k, l) -subpartitions is a hypergraph, and our definition of a dominating set for a family of (k, l) -subpartitions matches the definition of a dominating set for a hypergraph.

Lemma 3. *Let n, k, l be positive integers with $n \geq kl, l \geq 2$ and let $\mathcal{P} \subseteq U_{l,k}^n$ be an intersecting family of (k, l) -subpartitions. Assume that there does not exist a k -set that occurs as a block in every (k, l) -subpartition in \mathcal{P} . Then*

$$|\mathcal{P}| \leq l^2 U(n - 2k, l - 2, k). \quad (1)$$

Proof. Let $\{P_1, \dots, P_l\}$ be a (k, l) -subpartition in \mathcal{P} , and for $i \in \{1, \dots, l\}$, let \mathcal{P}_i be the set of all (k, l) -subpartitions in \mathcal{P} that contain the block P_i but none of P_1, \dots, P_{i-1} . By assumption, P_i does not appear in every (k, l) -subpartition in \mathcal{P} , so there exists some (k, l) -subpartition Q that does not contain P_i . The subpartitions in \mathcal{P}_i and Q must be intersecting, so each member of \mathcal{P}_i must contain P_i as well as one of the l blocks from Q . Thus, we can bound the size of \mathcal{P}_i by

$$|\mathcal{P}_i| \leq l U(n - 2k, l - 2, k).$$

Further, since $\{P_1, \dots, P_l\}$ is a dominating set for the family of (k, l) -subpartitions, we have that

$$\bigcup_{i \in \{1, \dots, l\}} \mathcal{P}_i = \mathcal{P}.$$

It follows that

$$|\mathcal{P}| \leq l |\mathcal{P}_i| \leq l^2 U(n - 2k, l - 2, k). \quad \square$$

Note that Lemma 3 certainly applies for all $n \geq kl$; however, if the size of n is small enough relative to k and l , then we can improve our bound on such an intersecting family \mathcal{P} . Note that in the case of $n = kl$, we may use the lemma as considered in [Meagher and Moura 2005].

Lemma 4. *Let n, k, l be positive integers with $kl + 1 \leq n \leq k(l + 1) - 1$, $l \geq 2$, and let $\mathcal{P} \subseteq U_{l,k}^n$ be an intersecting family of (k, l) -subpartitions. Assume that there does not exist a k -set that occurs as a block in every (k, l) -subpartition in \mathcal{P} . Then*

$$|\mathcal{P}| \leq l(l - 1)U(n - 2k, l - 2, k). \tag{2}$$

Proof. Under the restriction on the size of n , there are at most $l - 1$ blocks in Q that do not contain an element from P_i . The remainder of the proof follows similarly. \square

We also adapt a similar lemma for the t -intersecting case.

Lemma 5. *Let n, k, l, t be positive integers with $1 \leq t \leq l - 1$, and let $\mathcal{P} \subseteq U_{l,k}^n$ be a t -intersecting family of (k, l) -subpartitions. Assume that there does not exist a k -set that occurs as a block in every (k, l) -subpartition in \mathcal{P} . Then*

$$|\mathcal{P}| \leq (l - t + 1) \binom{l}{t} U(n - (t + 1)k, l - (t + 1), k). \tag{3}$$

Proof. As in the proof of Lemma 3, let $\{P_1, \dots, P_l\}$ be a (k, l) -subpartition in \mathcal{P} , and for $i \in \{1, \dots, l\}$, define the set \mathcal{P}_i similarly. Note that if we order the \mathcal{P}_i sets, then any (k, l) -subpartition in \mathcal{P}_i where $i > l - t + 1$ must contain at least one of the blocks $\{P_1, \dots, P_{l-t+1}\}$ since the (k, l) -subpartitions here must be t -intersecting with $\{P_1, \dots, P_l\}$. The block P_i does not appear in every (k, l) -subpartition in \mathcal{P} , so there exists some (k, l) -subpartition Q that does not contain P_i . Any (k, l) -subpartition $P \in \mathcal{P}_i$ must be t -intersecting with Q , so there are $\binom{l}{t}$ ways to choose the t blocks from Q that are also in P . Thus, we can bound the size of \mathcal{P}_i by

$$|\mathcal{P}_i| \leq \binom{l}{t} U(n - (t + 1)k, l - (t + 1), k).$$

Further, since

$$\bigcup_{i \in \{1, \dots, l-t+1\}} \mathcal{P}_i = \mathcal{P},$$

it follows that

$$|\mathcal{P}| \leq (l - t + 1) \binom{l}{t} U(n - (t + 1)k, l - (t + 1), k). \tag{3} \quad \square$$

3. Proof of Theorem 1

We can use (1) or (2), based on the size of n , and compare these bounds with that of (**). Informally, we may think of these as bounds on the size of *noncanonical* families of (k, l) -subpartitions. If the size of the canonical family is larger than

these bounds, then we know that the canonical families are the largest and that equality holds if and only if the intersecting family is canonical.

Proof of Theorem 1. Let \mathcal{P} be a noncanonical family of intersecting (k, l) -subpartitions. We shall show that

$$|\mathcal{P}| < \frac{1}{l-1} \binom{n-k}{k} U(n-2k, l-2, k). \quad (4)$$

It can be verified from (**) and (†) that the right-hand side of this inequality is the size of a canonical intersecting family of (k, l) -subpartitions; thus, proving this inequality proves Theorem 1.

Case 1: $kl + 1 \leq n \leq k(l+1) - 1$

If we bound n as such, then by (2),

$$|\mathcal{P}| \leq l(l-1)U(n-2k, l-2, k),$$

and using (4), we only need to prove that

$$l(l-1)^2 < \binom{n-k}{k}. \quad (5)$$

Since $n \geq kl + 1$, and using that $k \geq 3$, by Pascal's rule,

$$\binom{n-k}{k} \geq \binom{k(l-1)+1}{k} \geq \binom{3(l-1)+1}{3} = \frac{(3l-2)(3l-3)(3l-4)}{3!}.$$

Thus, (5) can be reduced to checking the inequality

$$l(l-1)^2 < \frac{(3l-2)(3l-3)(3l-4)}{3!}.$$

It can be verified, using the increasing function test, that this holds for all $l \geq 2$.

Case 2: $n \geq k(l+1)$

Similar to the previous case, using (1) and (4), we only need to show that

$$l^2(l-1) < \binom{n-k}{k}. \quad (6)$$

As before, taking $n \geq k(l+1)$, $k \geq 3$, and using Pascal's rule, we find

$$\binom{n-k}{k} \geq \binom{kl}{k} \geq \binom{3l}{3} = \frac{3l(3l-1)(3l-2)}{3!}.$$

So, (6) can be rewritten as

$$l^2(l-1) < \frac{3l(3l-1)(3l-2)}{3!},$$

and we find that this also holds for all $l \geq 2$.

Thus, (4) holds for all values of n , completing the proof of Theorem 1. \square

4. Proof of Theorem 2

Theorem 2 incorporates the t -intersection property, proving a more general EKR-type theorem for (k, l) -subpartitions. Here, the precise lower bound on n for determining when only the canonical families are the largest is unknown — but we shall see that if $k \geq t + 2$, it suffices to take $n \geq k(l + t)$ (though this bound is not optimal).

Proof of Theorem 2. From (*) and (†), the size of a canonical t -intersecting family of (k, l) -subpartitions is

$$U(n - tk, l - t, k) = \frac{1}{l - t} \binom{n - tk}{k} U(n - (t + 1)k, l - (t + 1), k). \tag{7}$$

As before, let \mathcal{P} be a noncanonical family of t -intersecting (k, l) -subpartitions. If there is a block that is contained in every (k, l) -subpartition of \mathcal{P} , then it can be removed from every such subpartition in \mathcal{P} . This does not change the size of the family, but reduces n by k and each of l and t by 1. Now we only need to show that this new family is smaller than the canonical $(t - 1)$ -intersecting family of $(k, l - 1)$ -subpartitions from $[n - k]$ (the size of which is equal to $U(n - (t - 1)k, l - (t - 1), k)$). As such, we may assume that there are no blocks common to every (k, l) -subpartition in \mathcal{P} , and we can apply (3).

To prove this theorem, we need to prove that for n sufficiently large,

$$(l - t + 1)(l - t) \binom{l}{t} < \binom{n - tk}{k}. \tag{8}$$

Clearly, this inequality is strict if n is sufficiently large relative to t, l and k . □

Consider the case where $k \geq t + 2$. If $n \geq k(l + t)$, then (8) holds when

$$(l - t + 1)(l - t) \binom{l}{t} \leq \binom{lk}{k}.$$

Since $k \geq t + 2$, we have that

$$\binom{lk}{k} = \binom{lk}{k} \binom{lk - 1}{k - 2} \binom{lk - 2}{k - 2} > (l - t + 1)(l - t) \binom{l}{t},$$

so (8) holds indeed. We do not attempt to find the function $n_0(k, l, t)$ that produces the exact lower bound on n , but such a lower bound is needed, as shown by the example in [Meagher and Moura 2005, Section 5].

5. Extensions

There are versions of the EKR theorem for many different objects. In this final section, we shall outline how this method can be generalized to these different objects.

In general, when considering an EKR-type theorem, there is a set of objects with some notion of intersection. We shall consider the case when each object is comprised of k atoms, and two objects are intersecting if they both contain a

common atom. If the objects are k -sets, then the atoms are the elements from $\{1, \dots, n\}$, and each k -set contains exactly k atoms. For matchings, the atoms are edges from the complete graph on $2n$ vertices, and a k -matching has k atoms. In this paradigm, if the largest set of intersecting objects is the set of all the objects that contain a fixed atom, then an EKR-type theorem holds.

We can apply the method in this paper to this more general situation. Assume we have a set of objects and that each object contains exactly k distinct atoms from a set of n atoms (there may be many additional rules on which sets of atoms constitute an object). Let $P(n, k)$ be the total number of objects, $P(n - 1, k - 1)$ the number of objects that contain a fixed atom, and $P(n - 2, k - 2)$ the number of objects that contain two fixed atoms.

Using the same argument as in this paper, if for some type of object (as above)

$$k^2 P(n - 2, k - 2) < P(n - 1, k - 1),$$

then an EKR-type theorem holds for these objects. It is very interesting to note that if the ratio between $P(n - 1, k - 1)$ and $P(n - 2, k - 2)$ is sufficiently large, then an EKR-type theorem holds.

For example, this can be applied to k -sets. In this case, the equation is

$$k^2 \binom{n-2}{k-2} < \binom{n-1}{k-1},$$

which holds if and only if

$$k^2(k - 1) + 1 < n.$$

This proves the standard EKR theorem, but with a very bad lower bound on n .

For a second example, consider length- n integer sequences with entries from $\{0, 1, \dots, q - 1\}$. In this case the atoms are ordered pairs (i, a) , where the entry in position i of the sequence is a . Two sequences “intersect” if they have the same entry in the same position. Each sequence contains exactly n atoms, so in this case $k = n$. The values of $P(n - 1, n - 1)$ and $P(n - 2, n - 2)$ are q^{n-1} and q^{n-2} , respectively. Thus an EKR-type theorem for integer sequences holds if $n^2 q^{n-2} < q^{n-1}$, or equivalently if $n^2 < q$. Once again we have a simple proof of an EKR-type theorem, but with an unnecessary bound on n .

Finally, consider the blocks in a t - (n, m, λ) design. The blocks are m -sets, so they are t -intersecting if they contain a common set of t -elements. It is straight-forward to calculate the number of blocks that contain any s -set where $s \leq t$ is

$$\lambda \frac{\binom{n-s}{t-s}}{\binom{m-s}{t-s}}.$$

Thus we have that the EKR theorem holds for intersecting blocks in a t - (n, m, λ)

design if

$$m^2 \frac{\lambda \binom{n}{t} \binom{m}{2}}{\binom{m}{t} \binom{n}{2}} \leq \frac{\lambda \binom{n}{t} \binom{m}{1}}{\binom{m}{t} \binom{n}{1}},$$

which reduces to

$$m^3 - m^2 + 1 < n.$$

This is the same bound found by Rands [1982]. Moreover, this method can be applied to s -intersecting blocks in a design; again we get the same bound as in [Rands 1982].

References

- [Ahlsweide and Khachatryan 1997] R. Ahlsweide and L. H. Khachatryan, “The complete intersection theorem for systems of finite sets”, *European J. Combin.* **18**:2 (1997), 125–136. MR 97m:05251 Zbl 0869.05066
- [Brunk and Huczynska 2010] F. Brunk and S. Huczynska, “Some Erdős–Ko–Rado theorems for injections”, *European J. Combin.* **31**:3 (2010), 839–860. MR 2011d:05380 Zbl 1226.05004
- [Cameron and Ku 2003] P. J. Cameron and C. Y. Ku, “Intersecting families of permutations”, *European J. Combin.* **24**:7 (2003), 881–890. MR 2004g:20003 Zbl 1026.05001
- [Erdős 1987] P. Erdős, “My joint work with Richard Rado”, pp. 53–80 in *Surveys in combinatorics 1987* (New Cross, 1987), edited by C. Whitehead, London Math. Soc. Lecture Note Ser. **123**, Cambridge Univ. Press, 1987. MR 88k:01032 Zbl 0623.01010
- [Erdős et al. 1961] P. Erdős, C. Ko, and R. Rado, “Intersection theorems for systems of finite sets”, *Quart. J. Math. Oxford Ser. (2)* **12** (1961), 313–320. MR 25 #3839 Zbl 0100.01902
- [Frankl and Wilson 1986] P. Frankl and R. M. Wilson, “The Erdős–Ko–Rado theorem for vector spaces”, *J. Combin. Theory Ser. A* **43**:2 (1986), 228–236. MR 87k:05005 Zbl 0609.05055
- [Kamat and Misra 2013] V. Kamat and N. Misra, “An Erdős–Ko–Rado theorem for matchings in the complete graph”, preprint, 2013. arXiv 1303.4061
- [Ku and Leader 2006] C. Y. Ku and I. Leader, “An Erdős–Ko–Rado theorem for partial permutations”, *Discrete Math.* **306**:1 (2006), 74–86. MR 2006j:05205 Zbl 1088.05072
- [Ku and Renshaw 2008] C. Y. Ku and D. Renshaw, “Erdős–Ko–Rado theorems for permutations and set partitions”, *J. Combin. Theory Ser. A* **115**:6 (2008), 1008–1020. MR 2009f:05256 Zbl 1154.05056
- [Meagher and Moura 2005] K. Meagher and L. Moura, “Erdős–Ko–Rado theorems for uniform set-partition systems”, *Electron. J. Combin.* **12** (2005), Research Paper 40. MR 2006d:05178 Zbl 1075.05086
- [Rands 1982] B. M. I. Rands, “An extension of the Erdős, Ko, Rado theorem to t -designs”, *J. Combin. Theory Ser. A* **32**:3 (1982), 391–395. MR 84i:05024 Zbl 0494.05005
- [Wilson 1984] R. M. Wilson, “The exact bound in the Erdős–Ko–Rado theorem”, *Combinatorica* **4**:2-3 (1984), 247–257. MR 86f:05007 Zbl 0556.05039

Received: 2013-10-03

Revised: 2014-04-09

Accepted: 2014-04-12

dyck204a@uregina.ca

Department of Mathematics and Statistics, University of Regina, 3737 Wascana Parkway, S4S 0A4 Regina SK, Canada

karen.meagher@uregina.ca

Department of Mathematics and Statistics, University of Regina, 3737 Wascana Parkway, S4S 0A4 Regina SK, Canada

Nonreal zero decreasing operators related to orthogonal polynomials

Andre Bunton, Nicole Jacobs, Samantha Jenkins,
Charles McKenry Jr., Andrzej Piotrowski and Louis Scott

(Communicated by Michael Dorff)

Laguerre's theorem regarding the number of nonreal zeros of a polynomial and its image under certain linear operators is generalized. This generalization is then used to (1) exhibit a number of previously undiscovered complex zero decreasing sequences for the Jacobi, ultraspherical, Legendre, Chebyshev, and generalized Laguerre polynomial bases and (2) simultaneously generate a basis B and a corresponding complex zero decreasing sequence for the basis B . An extension to transcendental entire functions in the Laguerre–Pólya class is given, which, in turn, gives a new and short proof of a previously known result due to Piotrowski. The paper concludes with several open questions.

1. Introduction

For a function $f : \mathbb{C} \rightarrow \mathbb{C}$ which is not the identically zero function, denote the number (counted according to multiplicity) of real and nonreal zeros of f by $Z_R(f)$ and $Z_C(f)$, respectively. For the identically zero function, define $Z_R(0) = 0$ and $Z_C(0) = 0$. Let $L : \mathbb{R}[x] \rightarrow \mathbb{R}[x]$ be a linear operator. If L has the property that

$$Z_C(L(p)) \leq Z_C(p) \quad (1)$$

for every real polynomial p , then L is called a *complex zero decreasing operator*, or a CZDO. Such an operator L is diagonal with respect to a basis $B = \{b_k\}_{k=0}^{\infty}$ for $\mathbb{R}[x]$ if and only if there are real constants $\{\gamma_k\}_{k=0}^{\infty}$ for which

$$L(b_k(x)) = \gamma_k b_k(x) \quad (k = 0, 1, 2, \dots). \quad (2)$$

MSC2010: 30C15.

Keywords: complex zero decreasing sequences, diagonalizable linear operators, zeros of polynomials, orthogonal polynomials.

This research was partially supported by the MAA through an NREUP grant funded by the NSA (grant H98230-13-1-0270) and the NSF (grant DMS-1156582).

In this case, the sequence $\{\gamma_k\}_{k=0}^{\infty}$ is called a *complex zero decreasing sequence for the basis B* , or a *B -CZDS*.

A theorem of Laguerre demonstrates the existence of CZDSs for the standard basis. We give two versions of his theorem here, the first of which can be found in [Obreschkoff 1963, p. 6] and [Craven and Csordas 2004, p. 23].

Theorem 1 (Laguerre's Theorem). *Let $p(x) = \sum_{k=0}^n a_k x^k$ be an arbitrary real polynomial of degree n . If α lies outside the interval $(-n, 0)$, then*

$$Z_C\left(\sum_{k=0}^n (k + \alpha) a_k x^k\right) \leq Z_C\left(\sum_{k=0}^n a_k x^k\right).$$

In particular, if $\alpha \geq 0$, the sequence $\{k + \alpha\}_{k=0}^{\infty}$ is a CZDS for the standard basis.

With notation as in Theorem 1,

$$xp'(x) + \alpha p(x) = \sum_{k=0}^n (k + \alpha) a_k x^k,$$

and Laguerre's theorem may be restated accordingly.

Theorem 2 (Laguerre's Theorem; Differential Operator Version). *Let $p(x)$ be an arbitrary real polynomial of degree n . If α lies outside the interval $(-n, 0)$, then*

$$Z_C(xp'(x) + \alpha p(x)) \leq Z_C(p(x)).$$

In particular, if $\alpha \geq 0$, then the differential operator $x D + \alpha I$ is a CZDO.

Remark 3. The differentiation operator D defined by $D(p) = p'$ is a CZDO. This is included in Laguerre's theorem as the special case $\alpha = 0$. Indeed, this choice gives

$$Z_C(p'(x)) = Z_C(xp'(x)) \leq Z_C(p(x)).$$

Alternatively, the fact that D is a CZDO can be proved via Rolle's theorem from elementary calculus (see, for example, [Obreschkoff 1963, p. 2–3]).

Laguerre's theorem is easily extended by iteration to sequences of the form $\{h(k)\}_{k=0}^{\infty}$, where h is a real polynomial having only real nonpositive zeros. This, in turn, leads to a further extension via Hurwitz's theorem to sequences of the form $\{\varphi(k)\}_{k=0}^{\infty}$, where φ is an entire function which is the uniform limit on compact subsets of \mathbb{C} of polynomials having only real nonpositive zeros (see, for example, [Craven and Csordas 1995, Theorem 1.4], [Obreschkoff 1963, p. 6], [Pólya 1929]). We have opted to state Laguerre's theorem in its simplest form to ease the comparison of this theorem with some of its generalizations demonstrated below.

In 2007, Piotrowski gave a generalization of Laguerre’s theorem to obtain a class of H -CZDSs, where H denotes the set of Hermite polynomials defined by

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n} e^{-x^2} \quad (n = 0, 1, 2, \dots).$$

Theorem 4 [Piotrowski 2007, p. 57, Proposition 68]. *Suppose $p(x)$ is an arbitrary real polynomial of degree n . If α, β, c, d are real numbers such that $\alpha \geq 0, \beta \geq 0$, and $\alpha + cn \geq 0$, then*

$$Z_C(-\beta p''(x) + (cx + d)p'(x) + \alpha p(x)) \leq Z_C(p(x)).$$

In particular, if α, β , and c are all nonnegative, then $-\beta D^2 + (cx + d)D + \alpha I$ is a CZDO.

Since the Hermite polynomials satisfy the differential equation

$$nH_n(x) = -\frac{1}{2}H_n''(x) + xH_n'(x) \quad (n = 0, 1, 2, \dots)$$

(see, for example, [Rainville 1960, p. 188]), the previous theorem gives, as a special case, the existence of H -CZDSs which can be interpolated by linear polynomials.

Theorem 5 [Piotrowski 2007, p. 87, Theorem 101]. *Let $p(x) = \sum_{k=0}^n a_k H_k(x)$ be an arbitrary real polynomial of degree n . If α lies outside the interval $(-n, 0)$, then*

$$Z_C\left(\sum_{k=0}^n (k + \alpha)a_k H_k(x)\right) \leq Z_C\left(\sum_{k=0}^n a_k H_k(x)\right).$$

In particular, if $\alpha \geq 0$, then the sequence $\{k + \alpha\}_{k=0}^\infty$ is an H -CZDS.

While no complete characterization of CZDSs is currently known for any basis, the characterization of CZDSs which can be interpolated by polynomials has been achieved for both the standard basis and the Hermite basis.

Theorem 6 [Craven and Csordas 1995, p. 13]. *Let $h(x)$ be a real polynomial. Then $\{h(k)\}_{k=0}^\infty$ is a CZDS for the standard basis if and only if either*

- (1) $h(0) \neq 0$ and $h(x)$ has only real negative zeros, or
- (2) $h(0) = 0$ and $h(x)$ is of the form

$$h(x) = x(x - 1)(x - 2) \cdots (x - m + 1) \prod_{k=1}^p (x - b_k), \tag{3}$$

where $m \geq 1$ and $p \geq 0$ are integers and $b_k < m$ for $k = 1, 2, 3, \dots, p$.

The previous theorem remains valid mutatis mutandis if “CZDS for the standard basis” is replaced by “ H -CZDS” (see [Piotrowski 2007, p. 95, Theorem 111]).

The main results of this paper include a generalization of Laguerre’s theorem (Theorem 8), the demonstration of classes of CZDSs for the Jacobi, ultraspherical,

Legendre, Chebyshev, and generalized Laguerre polynomial bases (Proposition 10, Theorem 14, Corollaries 15 and 16, and Theorem 24), a method for simultaneously generating a basis B and a corresponding B -CZDS (Section 4), and the extension of these results to transcendental entire functions in the Laguerre–Pólya class (Section 5.1).

2. A class of complex zero decreasing operators

This section contains two theorems which generalize Laguerre’s theorem.

Theorem 7. *Let p and q be real polynomials, each with degree at least one, and let $\alpha \geq 0$. Then*

$$Z_R(f(x)) \geq Z_R(p(x)) + Z_R(q(x)) - 1,$$

where

$$f(x) = q(x)p'(x) + \alpha q'(x)p(x).$$

Proof. When $\alpha = 0$, we have

$$Z_R(q(x)p'(x)) = Z_R(q(x)) + Z_R(p'(x)) \geq Z_R(p(x)) + Z_R(q(x)) - 1,$$

where the last inequality is a consequence of Rolle’s theorem.

We will now suppose $\alpha > 0$ for the remainder of the proof. Suppose x_0 is a zero of $p(x) \cdot q(x)$ and write

$$p(x) = (x - x_0)^m h_1(x) \quad (h_1(x_0) \neq 0),$$

$$q(x) = (x - x_0)^w h_2(x) \quad (h_2(x_0) \neq 0).$$

Then

$$f(x) = (x - x_0)^{m+w-1} h_3(x),$$

where

$$h_3(x_0) = (m + \alpha w)h_1(x_0)h_2(x_0) \neq 0.$$

That is to say, if x_0 is a zero of $p \cdot q$ of multiplicity $m + w$, then x_0 is a zero of f of multiplicity $m + w - 1$. We will now complete the proof by demonstrating that f must vanish between consecutive real zeros of $p \cdot q$. Define

$$g(x) = \begin{cases} [q(x)]^\alpha & \text{if } q(x) \geq 0, \\ -[-q(x)]^\alpha & \text{if } q(x) < 0, \end{cases}$$

so that

$$|q(x)|^{1-\alpha} \frac{d}{dx} [g(x)p(x)] = q(x)p'(x) + \alpha q'(x)p(x) \quad (x \notin \{z \mid q(z) = 0\}).$$

Let x_1, x_2 be consecutive zeros of $p \cdot q$ with $x_1 < x_2$. Then they are also consecutive zeros of $g \cdot p$, which is continuous on $[x_1, x_2]$ and differentiable on (x_1, x_2) . By

Rolle’s theorem, $(g \cdot p)'$, and therefore $q(x)p'(x) + \alpha q'(x)p(x)$ has a zero in the interval (x_1, x_2) and the conclusion of the theorem holds. \square

We note that Theorem 7 is best possible in the sense that the conclusion does not necessarily hold for any $\alpha < 0$. For example, if $\alpha < 0$, $p(x) = x^n(x^2 + \alpha)$, and $q(x) = x$, then $f(x) = x^n((\alpha + n + 2)x^2 + \alpha(\alpha + n))$. Choosing

$$n = \max\{m \in \mathbb{Z} \mid m \geq 0 \text{ and } \alpha + m < 0\}$$

yields $Z_R(f) = n < n + 2 = Z_R(p) + Z_R(q) - 1$.

Theorem 8. *Let p and q be real polynomials and $\alpha \geq 0$. Then*

$$Z_C(q(x)p'(x) + \alpha q'(x)p(x)) \leq Z_C(p(x)) + Z_C(q(x)).$$

In particular, if q has only real zeros, then $q(x)D + \alpha q'(x)I$ is a CZDO.

Proof. First note that the result is trivial when the function $q(x)p'(x) + \alpha q'(x)p(x)$ is identically zero. Furthermore, if either p or q is a nonzero constant function, then the result follows from Rolle’s theorem as was noted in Remark 3 above. We may, therefore, assume that p and q each have degree at least one. Suppose

$$p(x) = \sum_{k=0}^n a_k x^k \quad \text{and} \quad q(x) = \sum_{k=0}^m b_k x^k.$$

Then the leading term of

$$f(x) = q(x)p'(x) + \alpha q'(x)p(x)$$

is $(n + \alpha m)a_n b_m x^{n+m-1}$, so f has degree $n + m - 1$. Applying Theorem 7, we have

$$\begin{aligned} Z_C(f) &= n + m - 1 - Z_R(f) \\ &\leq n + m - 1 - (Z_R(p) + Z_R(q) - 1) \\ &= n + m - 1 - (n - Z_C(p) + m - Z_C(q) - 1) \\ &= Z_C(p) + Z_C(q). \end{aligned}$$

Therefore, $Z_C(q(x)p'(x) + \alpha q'(x)p(x)) \leq Z_C(p(x)) + Z_C(q(x))$. \square

Note that part of Laguerre’s theorem (Theorem 2) is obtained when we set $q(x) = x$ in Theorem 8.

Remark 9. The two theorems in this section can be extended to any finite number of constants and functions. For example, using the same techniques as above, one can show that

$$Z_C(pqr' + \alpha p'qr + \beta pq'r) \leq Z_C(p) + Z_C(q) + Z_C(r),$$

where α and β are nonnegative real numbers and p, q , and r are polynomials.

3. CZDSs for the Jacobi polynomial basis

3.1. The Jacobi polynomials. We now apply the results of the previous section to demonstrate the existence of CZDSs for the Jacobi polynomial basis. Following [Rainville 1960, p. 257], we define the Jacobi polynomials with parameters $\alpha > -1$ and $\beta > -1$ by

$$P_n^{(\alpha, \beta)}(x) = \frac{(-1)^n (1-x)^{-\alpha} (1+x)^{-\beta}}{2^n n!} \frac{d^n}{dx^n} [(1-x)^{n+\alpha} (1+x)^{n+\beta}].$$

For each nonnegative integer n , the Jacobi polynomials satisfy the differential equation

$$((x^2 - 1)D^2 + [(2 + \alpha + \beta)x + \alpha - \beta]D)P_n^{(\alpha, \beta)}(x) = n(n + 1 + \alpha + \beta)P_n^{(\alpha, \beta)}(x) \quad (4)$$

(see [Rainville 1960, p. 258]).

Proposition 10. *The sequence $\{k(k + 1 + \alpha + \beta)\}_{k=0}^{\infty}$ is a $P^{(\alpha, \beta)}$ -CZDS.*

Proof. Define the linear operator $L : \mathbb{R}[x] \rightarrow \mathbb{R}[x]$ by

$$L(P_k^{(\alpha, \beta)}(x)) = k(k + 1 + \alpha + \beta)P_k^{(\alpha, \beta)}(x) \quad (k = 0, 1, 2, \dots),$$

so that, by linearity,

$$L\left(\sum_{k=0}^n a_k P_k^{(\alpha, \beta)}(x)\right) = \sum_{k=0}^n a_k L(P_k^{(\alpha, \beta)}(x)) = \sum_{k=0}^n a_k k(k + 1 + \alpha + \beta)P_k^{(\alpha, \beta)}(x).$$

Our goal, then, is to show that L is a CZDO. From the differential equation (4), the linear operator L is equal to the differential operator

$$L = ((x^2 - 1)D + [(2 + \alpha + \beta)x + \alpha - \beta]I)D.$$

If, in Remark 9, we take $p(x) = x - 1$, $q(x) = x + 1$, and replace α and β by $\alpha + 1$ and $\beta + 1$, respectively, then we see that

$$(x^2 - 1)D + [(2 + \alpha + \beta)x + \alpha - \beta]I \quad (\alpha, \beta > -1)$$

is a complex zero decreasing operator. Thus, L is the composition of two CZDOs (recall that D is a CZDO as discussed in Remark 3) and so it is a CZDO itself. \square

3.2. Operator identities. In order to extend the preceding result, we will develop a number of operator identities. We consider two operators L_1 and L_2 on $\mathbb{R}[x]$ to be equal if $L_1(p) = L_2(p)$ for every real polynomial p . For example, as a consequence of the product rule for differentiation, $(Dx)p(x) = xp'(x) + p(x)$, and thus we obtain the equality

$$Dx = xD + I. \quad (5)$$

Proposition 11. *Suppose that $\{g_k(x)\}_{k=0}^m$ is a sequence of polynomials satisfying $\deg(g_k) \leq k$ for all k . Then*

$$D^n \sum_{k=0}^m g_k(x) D^k = \left(\sum_{j=0}^m \sum_{k=j}^m \binom{n}{k-j} g_k^{(k-j)}(x) D^j \right) D^n.$$

Proof. We first note that we are following the convention that $\binom{n}{k} = 0$ whenever $k > n$.

Using the fact that the derivative operator is linear, applying Leibniz’s formula for the n -th derivative of a product, and noting our assumption on the degree of the polynomials g_k , we have

$$D^n g_k(x) D^k = \sum_{i=0}^k \binom{n}{i} g_k^{(i)}(x) D^{k+n-i} = \left(\sum_{i=0}^k \binom{n}{i} g_k^{(i)}(x) D^{k-i} \right) D^n.$$

Making the substitution $j = k - i$ and then switching the order of summation gives

$$\begin{aligned} D^n \sum_{k=0}^m g_k(x) D^k &= \left(\sum_{k=0}^m \sum_{j=0}^k \binom{n}{k-j} g_k^{(k-j)}(x) D^j \right) D^n \\ &= \left(\sum_{j=0}^m \sum_{k=j}^m \binom{n}{k-j} g_k^{(k-j)}(x) D^j \right) D^n. \quad \square \end{aligned}$$

In what follows, we will make frequent use of Proposition 11 with $m = 2$, which asserts that if

$$L = D^n (g_2(x) D^2 + g_1(x) D + g_0(x) I), \tag{6}$$

then

$$L = \left(g_2(x) D^2 + (ng_2'(x) + g_1(x)) D + \left(\binom{n}{2} g_2''(x) + ng_1'(x) + g_0(x) \right) I \right) D^n, \tag{7}$$

provided $\deg(g_k) \leq k$ for all k .

3.3. Ultraspherical polynomials. We now focus on the Jacobi polynomials for which $\alpha = \lambda = \beta$, which are called the ultraspherical polynomials (see, e.g., [Rainville 1960, p. 143]). To ease notation, we define

$$P_n^{(\lambda)}(x) = P_n^{(\lambda, \lambda)}(x) \quad (\lambda > -1; n = 0, 1, 2, \dots).$$

With this choice, the differential equation (4) takes on the form

$$[(x^2 - 1) D^2 + (1 + \lambda) 2x D] P_n^{(\lambda)}(x) = n(n + 1 + 2\lambda) P_n^{(\lambda)}(x). \tag{8}$$

Due to the frequent use of the operator involved in the previous equation we define, for any $a \in \mathbb{R}$,

$$\Phi_a = (x^2 - 1) D + (1 + a) 2x I. \tag{9}$$

Lemma 12. *Suppose $\lambda > -1$. Then, for all nonnegative integers n ,*

$$D^n(\Phi_\lambda D - n(n+1+2\lambda)I) = (\Phi_{\lambda+n})D^{n+1},$$

where Φ_a is defined in (9).

Proof. This is an immediate application of (6) and (7). \square

We now use a product notation for composition of operators. Since differential operators need not commute, care is required in using this notation. For a collection of operators L_1, L_2, \dots, L_n on $\mathbb{R}[x]$, we define

$$\left(\prod_{k=1}^n L_k\right)p = (L_1 L_2 \cdots L_n)p = L_1(L_2(\cdots(L_n(p)))) \quad (p \in \mathbb{R}[x]).$$

Proposition 13. *Let w be a positive integer and $\{m_k\}_{k=0}^{w-1} \subset \mathbb{N}$. Then*

$$\prod_{k=0}^{w-1} (\Phi_\lambda D - k(k+1+2\lambda)I)^{m_k} = \left(\prod_{k=0}^{w-1} [(\Phi_{\lambda+k} D)^{m_k-1} \Phi_{\lambda+k}]\right) D^w,$$

where Φ_a is defined by (9).

Proof. We will argue by mathematical induction. The case $w = 1$ is clear. Now suppose that the result is true for some integer $w \geq 1$ and fix natural numbers m_0, m_1, \dots, m_w . Then

$$\prod_{k=0}^w (\Phi_\lambda D - k(k+1+2\lambda)I)^{m_k} = \Theta D^w (\Phi_\lambda D - w(w+1+2\lambda)I)^{m_w}, \quad (10)$$

where

$$\Theta = \prod_{k=0}^{w-1} [(\Phi_{\lambda+k} D)^{m_k-1} (\Phi_{\lambda+k})]. \quad (11)$$

Applying Lemma 12 a total of m_w times, we see that

$$D^w (\Phi_\lambda D - w(w+1+2\lambda)I)^{m_w} = (\Phi_{\lambda+w} D)^{m_w} D^w. \quad (12)$$

Together, (10), (11), and (12) show that

$$\prod_{k=0}^w (\Phi_\lambda D - k(k+1+2\lambda)I)^{m_k} = \left(\prod_{k=0}^w ((\Phi_{\lambda+k} D)^{m_k-1} (\Phi_{\lambda+k}))\right) D^{w+1}. \quad \square$$

We are now in a position to demonstrate the existence of several $P^{(\lambda)}$ -CZDSs for any fixed $\lambda > -1$.

Theorem 14. *If $\lambda > -1$, w is a positive integer, and $\{m_k\}_{k=0}^{w-1} \subset \mathbb{N}$, then the sequence*

$$\left\{ \prod_{k=0}^{w-1} (n(n+1+2\lambda) - k(k+1+2\lambda))^{m_k} \right\}_{n=0}^{\infty} \quad (13)$$

is a $P^{(\lambda)}$ -CZDS, where $P^{(\lambda)}$ is the set of ultraspherical polynomials.

Proof. Let the linear operator $L : \mathbb{R}[x] \rightarrow \mathbb{R}[x]$ be defined by

$$L(P_n^{(\lambda)}(x)) = \left(\prod_{k=0}^{w-1} (n(n+1+2\lambda) - k(k+1+2\lambda))^{m_k} \right) P_n^{(\lambda)}(x).$$

From the differential equation (8), we have

$$L = \prod_{k=0}^{w-1} ((x^2 - 1)D^2 + (1 + \lambda)2xD - k(k + 1 + 2\lambda)I)^{m_k},$$

or, using the notation in (9) and applying Proposition 13,

$$L = \prod_{k=0}^{w-1} (\Phi_\lambda D - k(k + 1 + 2\lambda)I)^{m_k} = \left(\prod_{k=0}^{w-1} ((\Phi_{\lambda+k} D)^{m_k - 1} \Phi_{\lambda+k}) \right) D^w.$$

The operator L is, therefore, a composition of individual operators, each of which is a CZDO. This can be seen by appealing to Theorem 8, which shows that Φ_a is a CZDO whenever $a > -1$. □

3.4. CZDSs for Legendre basis. The polynomials

$$P_n(x) = P_n^{(0)}(x) = P_n^{(0,0)}(x) \quad (n = 0, 1, 2, \dots)$$

are known as the Legendre polynomials (see [Rainville 1960, p. 254]).

In [Blakeman et al. 2012], Open Question (4) conjectures that a certain type of falling factorial sequence is a multiplier sequence for the Legendre basis, or a P -MS. Since every P -CZDS is a P -MS, we can apply the results of the previous section to settle a variation of this question.

Corollary 15. *If w is a positive integer and $\{m_k\}_{k=0}^{w-1} \subset \mathbb{N}$, then the sequence*

$$\left\{ \prod_{k=0}^{w-1} (n(n+1) - k(k+1))^{m_k} \right\}_{n=0}^{\infty} = \left\{ \prod_{k=0}^{w-1} ((n+k+1)(n-k))^{m_k} \right\}_{n=0}^{\infty} \quad (14)$$

is a CZDS for the Legendre basis.

Proof. Apply Theorem 14 with $\lambda = 0$. □

Corollary 15 strengthens and extends some of the results obtained in [Blakeman et al. 2012] by showing that $\{k^2 + k\}_{k=0}^\infty$ is a P -CZDS and by demonstrating the existence of P -CZDSs (and hence P -multiplier sequences) which are not products of quadratic P -multiplier sequences.

3.5. CZDS for the Chebyshev basis. The Chebyshev polynomials $\mathcal{T} = \{T_n(x)\}$ and $\mathcal{U} = \{U_n(x)\}$ of the first and second kind, respectively, can be defined by

$$T_n(x) := \frac{n!}{\left(\frac{1}{2}\right)_n} P_n^{(-1/2)}(x) \quad (n = 0, 1, 2, \dots),$$

$$U_n(x) := \frac{(n+1)!}{\left(\frac{3}{2}\right)_n} P_n^{(1/2)}(x) \quad (n = 0, 1, 2, \dots),$$

where $(a)_n := a(a+1) \cdots (a+n-1)$ is the rising factorial (see [Rainville 1960, p. 301]). In [Piotrowski 2007, Lemma 156] it is shown that a sequence $\{\gamma_k\}_{k=0}^\infty$ is a CZDS for a simple set $Q = \{q_k(x)\}_{k=0}^\infty$ if and only if it is a \widehat{Q} -CZDS, where \widehat{Q} consists of the polynomials

$$\hat{q}_n(x) = c_n q_n(\alpha x + \beta) \quad (\beta \in \mathbb{R}; \alpha, c_n \in \mathbb{R} \setminus \{0\}).$$

Combining this with Theorem 14, we arrive at the following corollary.

Corollary 16. *If w is a positive integer and $\{m_k\}_{k=0}^{w-1} \subset \mathbb{N}$, then*

- (1) *the sequence $\{\prod_{k=0}^{w-1} (n^2 - k^2)^{m_k}\}_{n=0}^\infty$ is a \mathcal{T} -CZDS, and*
- (2) *the sequence $\{\prod_{k=0}^{w-1} (n(n+2) - k(k+2))^{m_k}\}_{n=0}^\infty$ is a \mathcal{U} -CZDS.*

Proof. Apply Theorem 14 with $\lambda = -1/2$ and again with $\lambda = 1/2$. □

4. Simultaneous generation of a basis B and a class of B -CZDSs

Given a basis B and a sequence $\{\gamma_k\}_{k=0}^\infty$, a typical strategy in showing that $\{\gamma_k\}_{k=0}^\infty$ is a B -CZDS is to find a differential operator representation for the diagonal operator which is a CZDO. In this section, we begin with a known CZDO and use it to demonstrate the existence of a basis B and a corresponding B -CZDS. Our results focus on bases which are *simple sets*, i.e., those for which $\deg(b_k) = k$ for all k . In the product notation that follows, we adopt the convention that

$$\prod_{k=0}^n a_k = 1$$

whenever $n < 0$.

Theorem 17. *Let $\alpha \geq 0$ and let*

$$q(x) = c_0 + c_1 x + \cdots + c_r x^r \quad (r \geq 1, c_r \neq 0)$$

be a real polynomial with only real zeros. Then there is a simple set of polynomials $B = \{b_n(x)\}_{n=0}^\infty$ which satisfy the differential equation

$$q(x)b_n^{(r)}(x) + \alpha q'(x)b_n^{(r-1)}(x) = \gamma_n b_n(x) \quad (n = 0, 1, 2, \dots), \tag{15}$$

where

$$\gamma_n = c_r(n + (\alpha - 1)r + 1) \prod_{k=0}^{r-2} (n - k) \quad (n = 0, 1, 2, \dots).$$

Consequently, the sequence $\{\gamma_n\}_{n=0}^\infty$ is a B-CZDS.

Remark 18. We note that, for the case where $r = 1$ and $\alpha \neq 0$, the explicit form of the sequence and the existence of the basis B follow from results contained in the beginning of Section 2 of [Azad et al. 2011] and the beginning of Section II of [Krall and Sheffer 1964]. The proof of the general case is similar, yet different enough to warrant its inclusion here.

Proof of Theorem 17. Consider the differential operator

$$L = q(x)D^r + \alpha q'(x)D^{r-1}.$$

With this notation, the differential equation (15) becomes $L(b_n(x)) = \gamma_n b_n(x)$ and our goal is to find the eigenvalues γ_n of L and show there is a simple set of polynomials consisting of eigenfunctions b_n of L . The matrix representation of L with respect to the standard basis is upper triangular, with eigenvalues on the main diagonal given by the coefficient of x^n in $L(x^n)$. Since

$$L(x^n) = \left(c_r \prod_{k=0}^{r-1} (n - k) + \alpha r c_r \prod_{k=0}^{r-2} (n - k) \right) x^n + h(x),$$

where h is a polynomial of degree less than or equal to $n - 1$, the eigenvalue sequence is given by $\gamma_n = p(n)$ for all n , where

$$p(x) = c_r(x + (\alpha - 1)r + 1) \prod_{k=0}^{r-2} (x - k).$$

Since p has only real zeros, each of which lies in the interval $(-\infty, r - 1]$, we either have

$$0 = \gamma_0 = \gamma_1 = \dots = \gamma_{m-1} < \gamma_m < \gamma_{m+1} < \dots$$

or

$$0 = \gamma_0 = \gamma_1 = \dots = \gamma_{m-1} > \gamma_m > \gamma_{m+1} > \dots,$$

where

$$m = \begin{cases} r - 1 & \text{if } \alpha \neq 0, \\ r & \text{if } \alpha = 0. \end{cases}$$

In either case, all the nonzero eigenvalues must be distinct. Furthermore,

$$L(x^n) \equiv 0 \quad (n = 0, 1, \dots, m-1),$$

so L has the form

$$L = \left[\begin{array}{c|c} 0_{m \times m} & A \\ \hline 0_{\infty \times m} & T \end{array} \right],$$

where T is an upper triangular matrix with distinct nonzero eigenvalues on the main diagonal.

We now show that there is a simple set B consisting of eigenfunctions of the operator L . Indeed, let L_n denote the $n \times n$ truncation of the matrix L . Since $L_m = 0_{m \times m}$, we have complete freedom in choosing our first m eigenfunctions, say

$$b_n(x) = x^n \quad (n = 0, 1, 2, \dots, m-1).$$

For L_{m+1} , there is an eigenfunction corresponding to the (nonzero) eigenvalue γ_m . This eigenfunction is linearly independent from those corresponding to the eigenvalue 0, thus it must be of degree m . Continuing in this fashion, we can construct a simple set B consisting of eigenfunctions of L as desired.

To show that $\{\gamma_n\}_{n=0}^{\infty}$ is a B -CZDS, suppose

$$g(x) = \sum_{k=0}^j d_k b_k(x) \quad (d_j \neq 0)$$

is a real polynomial. Then

$$Z_C(g(x)) \geq Z_C(g^{(r-1)}(x)) \geq Z_C(q(x)g^{(r)}(x) + \alpha q'(x)g^{(r-1)}(x)),$$

where we have made use of Remark 3 and Theorem 8. Since

$$\begin{aligned} q(x)g^{(r)}(x) + \alpha q'(x)g^{(r-1)}(x) &= \sum_{k=0}^j d_k (q(x)b_k^{(r)}(x) + \alpha q'(x)b_k^{(r-1)}(x)) \\ &= \sum_{k=0}^j \gamma_k d_k b_k(x), \end{aligned}$$

the desired result is obtained. \square

As an example, if we choose $q(x) = (x+1)^3$ and $\alpha = 1$, then the corresponding sequence would be $\gamma_n = (n+1)n(n-1)$, and we would need to find a simple set $B = \{b_n(x)\}_{n=0}^{\infty}$ which solves the differential equation

$$(n+1)n(n-1)b_n(x) = (x+1)^3 b_n'''(x) + 3(x+1)^2 b_n''(x) \quad (n = 0, 1, 2, \dots). \quad (16)$$

With some effort, one finds that sets B which solve (16) have the form

$$\begin{aligned} b_0(x) &= r, \\ b_1(x) &= sx + t, \\ b_n(x) &= c_n(x + 1)^n \quad (n = 2, 3, 4, \dots), \end{aligned}$$

where $t \in \mathbb{R}$ and r, s, c_2, c_3, \dots are any (fixed) nonzero real numbers. Thus, the sequence

$$\{(n + 1)n(n - 1)\}_{n=0}^\infty$$

is a B -CZDS for any such basis B .

5. An extension to certain transcendental entire functions

5.1. The Laguerre–Pólya class. A real entire function φ is said to belong to the *Laguerre–Pólya class*, denoted $\varphi \in \mathcal{L}\text{-}\mathcal{P}$, if it can be written in the form

$$\varphi(x) = cx^m e^{-ax^2+bx} \prod_{k=1}^\omega \left(1 + \frac{x}{x_k}\right) e^{-x/x_k}, \tag{17}$$

where $b, c, x_k \in \mathbb{R}$, m is a nonnegative integer, $a \geq 0$, $0 \leq \omega \leq \infty$, and $\sum_{k=1}^\omega x_k^{-2} < \infty$.

An alternate characterization of this class is as follows: $\varphi \in \mathcal{L}\text{-}\mathcal{P}$ if and only if φ is the uniform limit on compact subsets of \mathbb{C} of real polynomials having only real zeros (see, for example, [Levin 1964, Chapter VIII] or [Obreschkoff 1963, Satz 9.2]). This point of view, together with Hurwitz’s theorem (see [Marden 1949, p. 4]), allows us to obtain some useful extensions of results in Section 2.

Theorem 19. *Suppose φ belongs to the class $\mathcal{L}\text{-}\mathcal{P}$, p and q are real polynomials, and $\alpha \geq 0$. Then*

$$Z_C(\varphi qp' + \alpha(\varphi q)'p) \leq Z_C(p) + Z_C(q).$$

Proof. Suppose $\{f_k\}_{k=0}^\infty$ is a sequence of real polynomials with only real zeros which converge uniformly on compact subsets of \mathbb{C} to φ . By Theorem 8,

$$Z_C(f_k qp' + \alpha(f_k q)'p) \leq Z_C(p) + Z_C(q) \quad (k = 0, 1, 2, \dots).$$

Taking into account that $f_k qp' + \alpha(f_k q)'p$ converges uniformly on compact subsets of \mathbb{C} to $\alpha(\varphi q)'p + \varphi qp'$, Hurwitz’s theorem gives the desired result. \square

In order to prove an extension of Laguerre’s theorem related to H -CZDSs (Theorem 4), Piotrowski first proved a special case as a lemma. We now show how to obtain a new proof of this lemma using Theorem 19.

Corollary 20 [Piotrowski 2007, p. 55, Lemma 67]. *Suppose that $p(x)$ is a real polynomial of degree n . If c, d, β are real numbers such that $c \geq 0$ and $\beta \geq 0$, then*

$$Z_C((cx + d)p(x) - \beta p'(x)) \leq Z_C(p(x)).$$

Proof. If $\beta = 0$, the result clearly holds. If $\beta > 0$, we may appeal to Theorem 19 with $\alpha = \beta^{-1}$, $q(x) = 1$, and

$$\varphi(x) = -\exp\left(-\frac{c}{2}x^2 - dx\right) \quad (c \geq 0, d \in \mathbb{R})$$

to obtain the desired result. \square

5.2. CZDSs for the generalized Laguerre polynomial basis. In this section, we combine the results of the previous section with the methods of Section 3.3 to obtain a class of CZDSs for the generalized Laguerre polynomial basis, defined by

$$L_n^{(\alpha)}(x) = \sum_{k=0}^n \binom{n+\alpha}{n-k} \frac{(-x)^k}{k!} \quad (\alpha > -1; n = 0, 1, 2, \dots).$$

The generalized Laguerre polynomials satisfy the differential equation

$$-x \frac{d^2}{dx^2} L_n^{(\alpha)}(x) + (x - (\alpha + 1)) \frac{d}{dx} L_n^{(\alpha)}(x) = n L_n^{(\alpha)}(x) \quad (18)$$

(see, e.g., [Rainville 1960, p. 204]). Just as with the Jacobi basis, we will develop a number of operator identities in order to arrive at a collection of $L^{(\alpha)}$ -CZDSs. We begin by defining, for any $a \in \mathbb{R}$,

$$\Psi_a = -xD + (x - (a + 1))I. \quad (19)$$

Lemma 21. *Suppose $\alpha \in \mathbb{R}$. Then, for all nonnegative integers n ,*

$$D^n(\Psi_\alpha D - nI) = \Psi_{\alpha+n} D^{n+1}.$$

Proof. This is an immediate application of (6) and (7). \square

Proposition 22. *Let w be a positive integer and $\{m_k\}_{k=0}^{w-1} \subset \mathbb{N}$. Then*

$$\prod_{k=0}^{w-1} (\Psi_\alpha D - kI)^{m_k} = \left(\prod_{k=0}^{w-1} [(\Psi_{\alpha+k} D)^{m_k-1} \Psi_{\alpha+k}] \right) D^w,$$

where Ψ_a is defined in (19).

Proof. We will argue by mathematical induction. The case $w = 1$ is clear. Now suppose that the result is true for some integer $w \geq 1$ and fix natural numbers m_0, m_1, \dots, m_w . Then

$$\prod_{k=0}^w (\Psi_\alpha D - kI)^{m_k} = \prod_{k=0}^{w-1} [(\Psi_{\alpha+k} D)^{m_k-1} \Psi_{\alpha+k}] D^w (\Psi_\alpha D - wI)^{m_w}. \quad (20)$$

Applying Lemma 21 a total of m_w times, we see that

$$D^w (\Psi_\alpha D - wI)^{m_w} = (\Psi_{\alpha+w} D)^{m_w} D^w = (\Psi_{\alpha+w} D)^{m_w-1} \Psi_{\alpha+w} D^{w+1}. \quad (21)$$

Together, (20) and (21) give the desired result. \square

In order to use the operator identities above to find a collection of $L^{(\alpha)}$ -CZDSs for any $\alpha > -1$, we will use the result of Section 5.1.

Lemma 23. *For any $a > -1$, the operator*

$$\Psi_a = -xD + (x - (a + 1))I$$

is a CZDO.

Proof. Suppose $a > -1$ and set $c = a + 1$. By Theorem 19, for any real polynomial p ,

$$Z_C \left(c \frac{d}{dx} (-x \exp(-x/c)) p(x) + (-x \exp(-x/c)) p(x) \right) \leq Z_C(p(x)).$$

The smaller quantity above simplifies to

$$Z_C \left((-xp'(x) + (x - c)p(x)) \exp(-x/c) \right).$$

Since the exponential function never vanishes, we have shown that

$$Z_C(\Psi_a p(x)) = Z_C(-xp'(x) + (x - c)p(x)) \leq Z_C(p(x)). \quad \square$$

We now arrive at the main theorem of this section.

Theorem 24. *Fix $\alpha > -1$. If w is a positive integer and $\{m_k\}_{k=0}^{w-1} \subset \mathbb{N}$, then the sequence*

$$\left\{ \prod_{k=0}^{w-1} (n - k)^{m_k} \right\}_{n=0}^{\infty} \quad (22)$$

is an $L^{(\alpha)}$ -CZDS.

Proof. Let the linear operator $\Theta : \mathbb{R}[x] \rightarrow \mathbb{R}[x]$ be defined by

$$\Theta(L_n^{(\alpha)}(x)) = \left(\prod_{k=0}^{w-1} (n - k)^{m_k} \right) L_n^{(\alpha)}(x).$$

Combining the differential equation (18), the notation in (19), and Proposition 22, we have

$$\Theta = \prod_{k=0}^{w-1} (\Psi_\alpha D - kI)^{m_k} = \left(\prod_{k=0}^{w-1} [(\Psi_{\alpha+k} D)^{m_k-1} \Psi_{\alpha+b}] \right) D^w.$$

The operator Θ is, therefore, a composition of individual operators, each of which is a CZDO. This can be seen by appealing to Lemma 23. \square

Theorem 24 is a significant generalization and extension of a theorem due to Forgács and Piotrowski [2013, Theorem 4.4] and a stronger result on a narrower class of sequences than those characterized by Brändén and Ottergren [2014].

6. Open questions

Any sequence of the form

$$\{k(k-1) \cdots (k-(m-1))\}_{k=0}^{\infty}$$

(the “falling-factorial sequence”) is a CZDS for the standard basis. By Corollary 16, any sequence of the form

$$\{k^2(k^2-1) \cdots (k^2-(m-1)^2)\}_{k=0}^{\infty}$$

is a T -CZDS. The similarity of these results leads us to wonder if an analog of Theorem 6 could be obtained for the Chebyshev basis.

Problem 25. Find a complete characterization of polynomials h for which $\{h(k)\}_{k=0}^{\infty}$ is a T -CZDS, where T denotes the Chebyshev basis.

We note that the characterization will be different from that of the standard basis since the sequence $\{k\}_{k=0}^{\infty}$ is not a T -CZDS.

The results on ultraspherical and Laguerre CZDSs also have a falling factorial nature which leads us to consider the more general problem.

Problem 26. For any basis B , find a complete characterization of polynomials h for which $\{h(k)\}_{k=0}^{\infty}$ is a B -CZDS.

Recall that this problem has been solved when the basis is taken to be either the standard basis or the Hermite basis. The result [Piotrowski 2007, Lemma 157] solves the problem for any affine transformation of the standard basis or the Hermite basis. To date, Problem 26 remains unsolved for any other choice of the basis B .

As it was mentioned earlier, no complete characterization of CZDSs for the standard basis is known. In particular, it is not known whether or not every rapidly decreasing sequence (such as $\{\exp(-k^3)\}_{k=0}^{\infty}$) is a CZDS for the standard basis (see [Craven and Csordas 2004, Problem 4.8] for more details). A theorem of Piotrowski gives a connection between these and CZDSs for other bases.

Theorem 27 [Piotrowski 2007, Theorem 159]. *Let $B = \{q_k(x)\}_{k=0}^{\infty}$ be a simple set of polynomials. If the sequence $\{\gamma_k\}_{k=0}^{\infty}$ is a B -CZDS, then the sequence $\{\gamma_k\}_{k=0}^{\infty}$ is a CZDS for the standard basis.*

This prompts us to state a weaker version of Problem 4.8(a) of [Craven and Csordas 2004], which may be easier to settle.

Problem 28. Is there a simple set B for which $\{\exp(-k^3)\}_{k=0}^{\infty}$ is a B -CZDS?

We mention that our methods of simultaneously generating a basis and a CZDS may apply. However, the original operator will have to be modified as all of our methods generated sequences which can be interpolated by polynomials.

7. Acknowledgment

The authors would like to thank the MAA, NSA, and NSF for their financial support of this project, the mathematics program faculty and staff at UAS for their moral and administrative support, and the anonymous referee for referring them to the paper [Azad et al. 2011] which helped to vastly improve Section 4 of this manuscript. Piotrowski would also like to recognize Dr. George Csordas and Dr. Tamás Forgács for their inspiration, encouragement, and helpful suggestions.

References

- [Azad et al. 2011] H. Azad, A. Laradji, and M. T. Mustafa, “Polynomial solutions of differential equations”, *Adv. Difference Equ.* **2011**:58 (2011), 1–12. MR 2891797 Zbl 1273.33008
- [Blakeman et al. 2012] K. Blakeman, E. Davis, T. Forgács, and K. Urabe, “On Legendre multiplier sequences”, *Missouri J. Math. Sci.* **24**:1 (2012), 7–23. MR 2977127
- [Brändén and Ottergren 2014] P. Brändén and E. Ottergren, “A characterization of multiplier sequences for generalized Laguerre bases”, *Constr. Approx.* **39**:3 (2014), 585–596. MR 3207673
- [Craven and Csordas 1995] T. Craven and G. Csordas, “Complex zero decreasing sequences”, *Methods Appl. Anal.* **2**:4 (1995), 420–441. MR 98a:26015 Zbl 0853.30018
- [Craven and Csordas 2004] T. Craven and G. Csordas, “Composition theorems, multiplier sequences and complex zero decreasing sequences”, pp. 131–166 in *Value distribution theory and related topics*, edited by G. Barsegian et al., *Adv. Complex Anal. Appl.* **3**, Kluwer, Boston, 2004. MR 2006f:26024 Zbl 1101.26015
- [Forgács and Piotrowski 2013] T. Forgács and A. Piotrowski, “Multiplier sequences for generalized Laguerre bases”, *Rocky Mountain J. Math.* **43**:4 (2013), 1141–1159. MR 3105315 Zbl 1286.30001
- [Krall and Sheffer 1964] H. L. Krall and I. M. Sheffer, “A characterization of orthogonal polynomials”, *J. Math. Anal. Appl.* **8** (1964), 232–244. MR 29 #2596 Zbl 0125.31404
- [Levin 1964] B. J. Levin, *Distribution of zeros of entire functions*, Amer. Math. Soc., Providence, RI, 1964. MR 28 #217 Zbl 0152.06703
- [Marden 1949] M. Marden, *The geometry of the zeros of a polynomial in a complex variable*, *Mathematical Surveys* **3**, Amer. Math. Soc., New York, 1949. MR 11,101i Zbl 0038.15303
- [Obreschkoff 1963] N. Obreschkoff, *Verteilung und Berechnung der Nullstellen reeller Polynome*, VEB Deutscher Verlag der Wissenschaften, Berlin, 1963. MR 29 #1302 Zbl 0156.28202
- [Piotrowski 2007] A. Piotrowski, *Linear operators and the distribution of zeros of entire functions*, Ph.D. thesis, University of Hawai’i at Manoa, 2007, Available at <http://search.proquest.com/docview/304847146>. MR 2710244
- [Pólya 1929] G. Pólya, “Über einen Satz von Laguerre”, *Jahresber. Dtsch. Math.-Ver.* **38** (1929), 161–168. JFM 55.0777.02
- [Rainville 1960] E. D. Rainville, *Special functions*, Macmillan, New York, 1960. MR 21 #6447 Zbl 0092.06503

Received: 2013-12-22 Revised: 2014-03-29 Accepted: 2014-04-07

andre.bunton@hotmail.com *University of Alaska Southeast, 11120 Glacier Highway,
Juneau, AK 99801, United States*

najacobs@uas.alaska.edu *University of Alaska Southeast, 11120 Glacier Highway,
Juneau, AK 99801, United States*

red_mystic333@yahoo.com *University of Alaska Southeast, 11120 Glacier Highway,
Juneau, AK 99801, United States*

mckenry90@gmail.com *University of Alaska Southeast, 11120 Glacier Highway,
Juneau, AK 99801, United States*

apiotrowski@uas.alaska.edu *University of Alaska Southeast, 11120 Glacier Highway,
Mail Stop SOB1, Juneau, 99801, United States*

lscott28@uas.alaska.edu *University of Alaska Southeast, 11120 Glacier Highway,
Juneau, AK 99801, United States*

Path cover number, maximum nullity, and zero forcing number of oriented graphs and other simple digraphs

Adam Berliner, Cora Brown, Joshua Carlson, Nathanael Cox,
Leslie Hogben, Jason Hu, Katrina Jacobs, Kathryn Manternach,
Travis Peters, Nathan Warnberg and Michael Young

(Communicated by Chi-Kwong Li)

An oriented graph is a simple digraph obtained from a simple graph by choosing exactly one of the two arcs (u, v) or (v, u) to replace each edge $\{u, v\}$. A simple digraph describes the zero-nonzero pattern of off-diagonal entries of a family of (not necessarily symmetric) matrices. The minimum rank of a simple digraph is the minimum rank of this family of matrices; maximum nullity is defined analogously. The simple digraph zero forcing number and path cover number are related parameters. We establish bounds on the range of possible values of all these parameters for oriented graphs, establish connections between the values of these parameters for a simple graph G , for various orientations \vec{G} and for the doubly directed digraph of G , and establish an upper bound on the number of arcs in a simple digraph in terms of the zero forcing number.

1. Introduction

The maximum nullity and the zero forcing number of simple digraphs are studied in [Hogben 2010] and [Berliner et al. 2013]. We study connections between these parameters and path cover number, and we study all of these parameters for special types of digraphs derived from graphs, including oriented graphs and doubly directed graphs. Section 2 considers oriented graphs. We establish a bound on the difference of the parameters path cover number, maximum nullity, and zero forcing number for two orientations of one graph and determine the range of values of these parameters for orientations of paths and cycles and some of the possible values for tournaments.

MSC2010: 05C50, 05C20, 15A03.

Keywords: zero forcing number, maximum nullity, minimum rank, path cover number, simple digraph, oriented graph.

Research of Carlson and Manternach supported by ISU Holl funds. Research of Young supported by NSF DMS 0946431. Research of Cox, Jacobs, Hu, and Brown supported by NSF DMS 0750986.

We establish connections between these parameters for a simple graph and its doubly directed digraph in Section 3. In Section 4, we establish an upper bound on the number of arcs of a simple digraph in terms of the zero forcing number. We also show that several results for simple graphs fail for oriented graphs, including the graph complement conjecture and Sinkovic's theorem that maximum nullity is at most the path cover number for outerplanar graphs.

All graphs and digraphs are taken to be simple. We use $G = (V(G), E(G))$ to denote a graph and $\Gamma = (V(\Gamma), E(\Gamma))$ to denote a digraph, often using V and E when G or Γ is clear. For a digraph Γ and $R \subseteq V$, the *induced subdigraph* $\Gamma[R]$ is the digraph with vertex set R and arc set $\{(v, w) \in E : v, w \in R\}$; an analogous definition is used for graphs. The subdigraph induced by the complement \bar{R} is also denoted by $\Gamma - R$, or in the case where R is a single vertex v , by $\Gamma - v$. A digraph $\Gamma = (V, E)$ is *transitive* if for all $u, v, w \in V$, $(u, v), (v, w) \in E$ implies $(u, w) \in E$.

For a digraph $\Gamma = (V, E)$ having $v, u \in V$ and $(v, u) \in E$, u is an *out-neighbor* of v and v is an *in-neighbor* of u . The *out-degree* of v , denoted by $\deg^+(v)$, is the number of out-neighbors of v in Γ ; *in-degree* is defined analogously and denoted by $\deg^-(v)$. Define $\delta^+(\Gamma) = \min\{\deg^+(v) : v \in V\}$ and $\delta^-(\Gamma) = \min\{\deg^-(v) : v \in V\}$. For a digraph Γ , the *reversal* Γ^T is obtained from Γ by reversing all the arcs.

Let G be a graph. A *path* in G is a subgraph $P = (\{v_1, \dots, v_k\}, E(P))$, where $E(P) = \{\{v_i, v_{i+1}\} : 1 \leq i \leq k-1\}$; this path is often denoted by (v_1, \dots, v_k) and its length is $k-1$. We say that a path in G is an *induced path* if it is an induced subgraph of G . A *path cover* of a graph G is a set of vertex-disjoint induced paths that includes all vertices of G .

Now suppose Γ is a digraph. A *path* in Γ is a subdigraph $P = (\{v_1, \dots, v_k\}, E(P))$, where $E(P) = \{(v_i, v_{i+1}) : 1 \leq i \leq k-1\}$; this path is often denoted by (v_1, \dots, v_k) , its length is $k-1$, and the arcs of $E(P)$ are called *path arcs*. If (v_1, \dots, v_k) is a path in Γ , v_1 is called the *initial vertex* and v_k is the *terminal vertex*. We say vertex u has *access* to v in Γ if there is a path from u to v . A path (v_1, \dots, v_k) in Γ is an *induced path* if E does not contain any arc of the form (v_i, v_j) with $j > i+1$ or $i > j+1$. We note this does not necessarily imply that the path subdigraph is induced because any of the arcs in $\{(v_{i+1}, v_i) : 1 \leq i \leq k-1\}$ are permitted. A path (v_1, \dots, v_k) in Γ is *Hessenberg* if E does not contain any arc of the form (v_i, v_j) with $j > i+1$. Any induced path is Hessenberg but not vice versa. A *path cover* for Γ is a set of vertex-disjoint Hessenberg paths that includes all vertices of Γ [Hogben 2010].

For graphs G and digraphs Γ , the *path cover number* $P(G)$ or $P(\Gamma)$ is the minimum number of paths in a path cover (induced for a graph, Hessenberg for a digraph) and a *minimum path cover* is a path cover with this minimum number of paths.

Zero forcing was introduced in [AIM 2008] for (simple) graphs. We define zero forcing for (simple) digraphs as in [Hogben 2010]. Let Γ be a digraph with each

vertex colored either white or blue¹. The *color change rule* is: if u is a blue vertex of Γ and exactly one out-neighbor v of u is white, then change the color of v to blue. In this situation, we say that u forces v and write $u \rightarrow v$. Given a coloring of Γ , the *final coloring* is the result of applying the color change rule until no more changes are possible. A *zero forcing set* for Γ is a subset of vertices B such that if initially the vertices of B are colored blue and the remaining vertices are white, the final coloring of Γ is all blue. The *zero forcing number* $Z(\Gamma)$ is the minimum of $|B|$ over all zero forcing sets $B \subseteq V(\Gamma)$.

For a given zero forcing set B for Γ , we create a *chronological list of forces* by constructing the final coloring, listing the forces in the order in which they were performed. Although for a given set of vertices B the final coloring is unique, B need not have a unique chronological list of forces. Suppose Γ is a digraph and \mathcal{F} is a chronological list of forces for a zero forcing set B . A *forcing chain* is an ordered set of vertices (w_1, w_2, \dots, w_k) , where $w_j \rightarrow w_{j+1}$ is a force in \mathcal{F} for $1 \leq j \leq k - 1$. A *maximal forcing chain* is a forcing chain that is not a proper subset of another forcing chain. The following results will be used.

Lemma 1.1 [Hogben 2010]. *Suppose Γ is a digraph and \mathcal{F} is a chronological list of forces of a zero forcing set B . Then, every maximal forcing chain is a Hessenberg path that starts with a vertex in B .*

For a fixed chronological list of forces \mathcal{F} of a zero forcing set B of Γ , the *chain set* is the set of all maximal forcing chains. By Lemma 1.1, the chain set of \mathcal{F} is a path cover, called a *zero forcing path cover*, and the maximal forcing chains are also called *forcing paths*.

Proposition 1.2 [Hogben 2010]. *For any digraph Γ , we have $P(\Gamma) \leq Z(\Gamma)$.*

A cycle of length $k \geq 3$ in a graph G or digraph Γ is a sub(di)graph consisting of a path (v_1, \dots, v_k) and the additional edge or arc $\{v_k, v_1\}$ or (v_k, v_1) .

Lemma 1.3. *Suppose $P = (v_1, \dots, v_k)$ is a Hessenberg path in a digraph Γ . Then P is an induced path or $\Gamma[V(P)]$ contains a (digraph) cycle of length at least 3.*

Proof. Suppose P is not an induced path. Then Γ must contain an arc of the form (v_i, v_j) with $j > i + 1$ or $i > j + 1$. Since P is Hessenberg, Γ does not contain an arc of the form (v_i, v_j) with $j > i + 1$. Thus Γ must contain an arc of the form (v_i, v_j) with $i > j + 1$. Then $(v_j, v_{j+1}, \dots, v_i, v_j)$ is a (digraph) cycle in $\Gamma[V(P)]$. \square

Let F be a field. For a square matrix $A = [a_{ij}] \in F^{n \times n}$, the *digraph of A* , denoted $\Gamma(A) = (V, E)$, is the (simple) digraph described by the off-diagonal zero-nonzero pattern of the entries: the set of vertices is $V = \{1, 2, \dots, n\}$ and the set of arcs is $E = \{(i, j) : a_{ij} \neq 0, i \neq j\}$. Note that the value of the diagonal entries of A does not affect $\Gamma(A)$.

¹The early literature uses the color black rather than blue.

Conversely, given any simple digraph Γ (along with an ordering of the vertices), we may associate with Γ a family of matrices $\mathcal{M}^F(\Gamma) = \{A \in F^{n \times n} : \Gamma(A) = \Gamma\}$. The *minimum rank* over F of a digraph Γ is $\text{mr}^F(\Gamma) = \min\{\text{rank } A : A \in \mathcal{M}^F(\Gamma)\}$ and the *maximum nullity* over F of Γ is $\text{M}^F(\Gamma) = \max\{\text{null } A : A \in \mathcal{M}^F(\Gamma)\}$. It is immediate that $\text{mr}^F(\Gamma) + \text{M}^F(\Gamma) = n$.

Similarly, symmetric matrices and undirected graphs are associated. For a symmetric matrix $A = [a_{ij}] \in F^{n \times n}$, the *graph of A* is the (simple) graph $\mathcal{G}(A) = (V, E)$ with $V = \{1, 2, \dots, n\}$ and $E = \{\{i, j\} : i \neq j \text{ and } a_{ij} \neq 0\}$. The family of symmetric matrices associated with G is $\mathcal{S}^F(G) = \{A \in F^{n \times n} : A^T = A, \mathcal{G}(A) = G\}$, and minimum rank and maximum nullity are similarly defined for undirected graphs.

For the much of this paper, we let $F = \mathbb{R}$ and we write $\mathcal{S}(G)$, $\mathcal{M}(\Gamma)$, $\text{M}(\Gamma)$, and $\text{mr}(\Gamma)$ rather than $\mathcal{S}^{\mathbb{R}}(G)$, $\mathcal{M}^{\mathbb{R}}(\Gamma)$, $\text{M}^{\mathbb{R}}(\Gamma)$, and $\text{mr}^{\mathbb{R}}(\Gamma)$, etc. If a graph or digraph parameter that depends on matrices does not change regardless of the field F , then we say that parameter is *field independent*; in the case that M is field independent, $\text{M}^F(\Gamma) = \text{M}(\Gamma)$ for every field F .

Remark 1.4. Clearly $\text{mr}^F(\Gamma^T) = \text{mr}^F(\Gamma)$, and $\text{Z}(\Gamma^T) = \text{Z}(\Gamma)$ is known [Berliner et al. 2013]. Because the reversal of a Hessenberg path is a Hessenberg path, $\text{P}(\Gamma^T) = \text{P}(\Gamma)$.

2. Oriented graphs

In this section, we establish results for minimum rank, maximum nullity, zero forcing number, and path cover number of oriented graphs. Given a graph G , an orientation \vec{G} of G is a digraph obtained by replacing each edge $\{u, v\}$ by exactly one of the arcs (u, v) and (v, u) (so a graph G has $2^{|E(G)|}$ orientations, some of which may be isomorphic to each other).

Range over orientations. We consider the range of values of $\beta(\vec{G})$ over all possible orientations for the parameters $\beta = \text{mr}, \text{M}, \text{Z}, \text{P}$.

Theorem 2.1. *Suppose β is a positive-integer-valued digraph parameter with the following properties for every oriented graph \vec{G} :*

- (1) $\beta(\vec{G}^T) = \beta(\vec{G})$.
- (2) *If $(u, v) \in E(\vec{G})$ and \vec{G}_0 is obtained from \vec{G} by replacing (u, v) by (v, u) (i.e., reversing the orientation of one arc), then $|\beta(\vec{G}_0) - \beta(\vec{G})| \leq 1$.*

Then for any two orientations \vec{G}_1 and \vec{G}_2 of the same graph G ,

$$|\beta(\vec{G}_2) - \beta(\vec{G}_1)| \leq \left\lfloor \frac{|E(G)|}{2} \right\rfloor.$$

Furthermore, every integer between $\beta(\vec{G}_2)$ and $\beta(\vec{G}_1)$ is attained as $\beta(\vec{G})$ for some orientation \vec{G} of G .

Proof. Without loss of generality, $\beta(\vec{G}_2) \geq \beta(\vec{G}_1)$. Let $e = |E(G)|$. Because \vec{G}_1 and \vec{G}_2 share the same underlying graph, it is possible to obtain \vec{G}_2 from \vec{G}_1 by reversing some of the arcs of \vec{G}_1 . Let ℓ be the number of arcs we need to reverse to obtain \vec{G}_2 from \vec{G}_1 . By hypothesis, reversing the direction of one arc changes the value of β by at most one, so $\beta(\vec{G}_2) - \beta(\vec{G}_1) \leq \ell$. The number of arcs that must be reversed to obtain \vec{G}_2^T from \vec{G}_1 is $e - \ell$, so $\beta(\vec{G}_2^T) - \beta(\vec{G}_1) \leq e - \ell$. By hypothesis, $\beta(\vec{G}_2^T) = \beta(\vec{G}_2)$, so $\beta(\vec{G}_2) - \beta(\vec{G}_1) \leq \lfloor e/2 \rfloor$. The last statement follows from hypothesis (2) and the fact that we can go from \vec{G}_1 to \vec{G}_2 by reversing one arc at a time. \square

Corollary 2.2. *If \vec{G}_1 and \vec{G}_2 are both orientations of the graph G , then*

$$\begin{aligned} |\text{mr}(\vec{G}_2) - \text{mr}(\vec{G}_1)| &\leq \left\lfloor \frac{E(G)}{2} \right\rfloor, & |\text{M}(\vec{G}_2) - \text{M}(\vec{G}_1)| &\leq \left\lfloor \frac{E(G)}{2} \right\rfloor, \\ |\text{Z}(\vec{G}_2) - \text{Z}(\vec{G}_1)| &\leq \left\lfloor \frac{E(G)}{2} \right\rfloor, & \text{and} \quad |\text{P}(\vec{G}_2) - \text{P}(\vec{G}_1)| &\leq \left\lfloor \frac{E(G)}{2} \right\rfloor. \end{aligned}$$

Furthermore, every integer between $\beta(\vec{G}_2)$ and $\beta(\vec{G}_1)$ is attained as $\beta(\vec{G})$ for some orientation \vec{G} of G when β is any of the parameters mr , M , Z , P .

Proof. The first hypothesis of Theorem 2.1, $\beta(\vec{G}^T) = \beta(\vec{G})$, is established for these parameters in Remark 1.4. To show that these parameters satisfy the second hypothesis of Theorem 2.1, suppose arc (u, v) of \vec{G} is reversed to obtain \vec{G}_0 from \vec{G} . In each case, the process is reversible, so it suffices to prove $\beta(\vec{G}_0) \leq \beta(\vec{G}) + 1$.

For minimum rank, suppose $\Gamma(A) = \vec{G}$ and $\text{rank } A = \text{mr}(\vec{G})$. Define B by $b_{uu} = b_{vv} = b_{uv} = b_{vu} = -a_{uv}$ and $b_{ij} = 0$ for all other entries of B . Then $\Gamma(A + B) = \vec{G}_0$ and $\text{rank}(A + B) \leq \text{rank } A + 1$. Thus $\text{mr}(\vec{G}_0) \leq \text{mr}(\vec{G}) + 1$. The statement for maximum nullity is equivalent.

For zero forcing number, choose a minimum zero forcing set B and chronological list of forces \mathcal{F} of \vec{G} . If the force $u \rightarrow v$ is in \mathcal{F} , then $B \cup \{v\}$ is a zero forcing set for \vec{G}_0 . If $u \not\rightarrow v$ and for some w , $v \rightarrow w$ is in \mathcal{F} , then $B \cup \{u\}$ is a zero forcing set for \vec{G}_0 . If v does not perform a force and $u \rightarrow v$ is not in \mathcal{F} , then B is a zero forcing set for \vec{G}_0 . Thus, $\text{Z}(\vec{G}_0) \leq \text{Z}(\vec{G}) + 1$.

For path cover number, suppose $\mathcal{P} = \{P^{(1)}, P^{(2)}, \dots, P^{(k)}\}$ is a path cover of \vec{G} and $|\mathcal{P}| = \text{P}(\vec{G})$. If (u, v) is not an arc in one of the paths in \mathcal{P} , then \mathcal{P} is a path cover for \vec{G}_0 and $\text{P}(\vec{G}_0) \leq \text{P}(\vec{G})$. So suppose (u, v) is an arc in some path $P^{(\ell)}$. Then we construct a path cover for \vec{G}_0 by replacing $P^{(\ell)}$ by the two paths resulting from deleting the arc (u, v) . Thus, $\text{P}(\vec{G}_0) \leq \text{P}(\vec{G}) + 1$. \square

Hierarchal orientation. We establish a method for finding an orientation \vec{G} of a graph G for which $\text{P}(\vec{G}) = \text{P}(G)$. Let $\mathcal{P} = \{P^{(1)}, P^{(2)}, \dots, P^{(k)}\}$ be any path cover of a graph G . A *rooted path cover* of G , $\mathcal{R} = \{R^{(1)}, R^{(2)}, \dots, R^{(k)}\}$, is obtained from \mathcal{P} by choosing one endpoint as the *root* of $P^{(i)}$ for each $i = 1, \dots, k$. A set \mathcal{R} is a *minimum rooted path cover* if $|\mathcal{R}| = \text{P}(G)$. In the case that \mathcal{P} is a zero

forcing path cover of a zero forcing set B , the root of $P^{(i)}$ is automatically chosen to be the unique element of B that is a vertex of $P^{(i)}$. A rooted path cover obtained from \mathcal{P} naturally orders $V(P^{(i)})$, starting with the root, and we denote this order by $R^{(i)} = (r_1^{(i)}, r_2^{(i)}, \dots, r_{s_i}^{(i)})$ where $s_i - 1$ is the length of $P^{(i)}$. Observe that if a rooted path cover is formed from a zero forcing path cover of a zero forcing set, the ordering within each rooted path coincides with the forcing order in that path.

Definition 2.3. Given a rooted path cover \mathcal{R} of a graph G , the *hierarchical orientation* $\vec{G}_{\mathcal{R}}$ of G resulting from \mathcal{R} is defined by orienting G as follows:

- (1) Orient each $R^{(i)}$ as $r_1^{(i)} \rightarrow r_2^{(i)} \rightarrow \dots \rightarrow r_{s_i}^{(i)}$; that is, replace the edge $\{r_j^{(i)}, r_{j+1}^{(i)}\}$ by the arc $(r_j^{(i)}, r_{j+1}^{(i)})$ for $j = 1, \dots, s_i - 1$.
- (2) For any edge between $R^{(i)}$ and $R^{(j)}$ with $i < j$, orient as $i \rightarrow j$; that is, if $i < j$, replace the edge $\{r_{\ell_i}^{(i)}, r_{\ell_j}^{(j)}\}$ by the arc $(r_{\ell_i}^{(i)}, r_{\ell_j}^{(j)})$.

Since by definition, the paths in a path cover of a graph are induced, all the edges of G have been oriented by these two rules.

Observation 2.4. For any rooted path cover \mathcal{R} of G , \mathcal{R} is a path cover of $\vec{G}_{\mathcal{R}}$ (with each path originating at its root), so $P(\vec{G}_{\mathcal{R}}) \leq |\mathcal{R}|$.

Proposition 2.5. An oriented graph \vec{G} is the hierarchical orientation $\vec{G}_{\mathcal{R}}$ of G for some rooted path cover \mathcal{R} of G (not necessarily minimum) if and only if \vec{G} does not contain a digraph cycle.

Proof. Suppose $\mathcal{R} = \{R^{(1)}, \dots, R^{(k)}\}$ is rooted path cover of G . Since each path $R^{(i)}$ is induced, in order for $\vec{G}_{\mathcal{R}}$ to have a digraph cycle, $V(\vec{G}_{\mathcal{R}})$ would have to include vertices from at least two paths $R^{(i)}$ and $R^{(j)}$ with $i < j$. But by the definition of $\vec{G}_{\mathcal{R}}$, there are no arcs from vertices in $R^{(j)}$ to vertices in $R^{(i)}$.

Suppose that \vec{G} does not contain a digraph cycle. Then we may order the vertices $\{v_1, \dots, v_n\}$ such that v_j does not have access to v_i whenever $j > i$. Then if $V(R^{(i)}) = \{v_i\}$, $\mathcal{R} = \{R^{(1)}, \dots, R^{(n)}\}$ is a rooted path cover and $\vec{G} = \vec{G}_{\mathcal{R}}$. \square

Theorem 2.6. Suppose $\mathcal{R} = \{R^{(1)}, \dots, R^{(k)}\}$ is a rooted path cover of G and $\vec{G}_{\mathcal{R}}$ is the hierarchical orientation of G resulting from \mathcal{R} . Then any path cover for $\vec{G}_{\mathcal{R}}$ is a path cover for G . If \mathcal{R} is a minimum rooted path cover, then $P(G) = P(\vec{G}_{\mathcal{R}})$.

Proof. Let P be a Hessenberg path in $\vec{G}_{\mathcal{R}}$. By Proposition 2.5, $\vec{G}_{\mathcal{R}}$ does not contain a digraph cycle, so by Lemma 1.3, P is an induced path. Thus, any path cover for $\vec{G}_{\mathcal{R}}$ is a path cover for G , and this implies $P(G) \leq P(\vec{G}_{\mathcal{R}})$. If \mathcal{R} is a minimum rooted path cover of G , then $P(\vec{G}_{\mathcal{R}}) \leq |\mathcal{R}| = P(G) \leq P(\vec{G}_{\mathcal{R}})$, so $P(G) = P(\vec{G}_{\mathcal{R}})$. \square

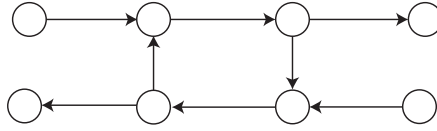


Figure 1. An oriented graph \vec{G} that is not a hierarchal orientation but has $P(\vec{G}) = P(G)$.

Example 2.7. Not every orientation \vec{G} having $P(\vec{G}) = P(G)$ is a hierarchal orientation. The oriented graph \vec{G} shown in Figure 1 has $P(\vec{G}) = 2 = P(G)$, but \vec{G} is not a hierarchal orientation because \vec{G} contains a digraph cycle.

Although for any graph G , we can find an orientation so that $P(\vec{G}) = P(G)$, this is not always the case for zero forcing number or maximum nullity.

Example 2.8. Consider K_4 , the complete graph on four vertices. It is well known that $M(K_4) = Z(K_4) = 3$, whereas we show that for any orientation \vec{K}_4 of K_4 , $2 \geq Z(\vec{K}_4) \geq M(\vec{K}_4)$. If \vec{K}_4 contains a directed 3-cycle, then any one vertex on the 3-cycle and the remaining vertex form a zero forcing set. If \vec{K}_4 has no directed 3-cycle, then we may order the vertices $\{u_1, u_2, u_3, u_4\}$, where u_j does not have access to u_i whenever $j > i$. Then, $\{u_1, u_3\}$ is a zero forcing set.

Observation 2.9. If $\mathcal{R} = \{R^{(1)}, \dots, R^{(k)}\}$ is a rooted path cover for G , then the set of roots $\{r_1^{(1)}, \dots, r_1^{(k)}\}$ is a zero forcing set of the digraph $\vec{G}_{\mathcal{R}}$, as zero forcing can be done in path order along $R^{(k)}$, followed by $R^{(k-1)}$, etc.

Theorem 2.10. Suppose G is a graph and \mathcal{R} is a minimum rooted path cover of G . Then $Z(\vec{G}_{\mathcal{R}}) = P(\vec{G}_{\mathcal{R}}) = P(G)$.

Proof. From Theorem 2.6, Proposition 1.2, Observation 2.9, and the hypotheses, $P(G) = P(\vec{G}_{\mathcal{R}}) \leq Z(\vec{G}_{\mathcal{R}}) \leq |\mathcal{R}| = P(G)$. □

Whenever $P(G) = Z(G)$, we can use a minimum rooted path cover to find an orientation of G realizing $Z(G)$ as its zero forcing number.

Corollary 2.11. Suppose G is a graph such that $P(G) = Z(G)$ and \mathcal{R} is a minimum rooted path cover of G . Then $Z(\vec{G}_{\mathcal{R}}) = Z(G)$.

Because $P(T) = Z(T)$ for every (simple undirected) tree T [AIM 2008], we have the following corollary.

Corollary 2.12. If T is a tree, then there exists an orientation \vec{T} of T such that $Z(\vec{T}) = Z(T)$.

If we allow a path cover that is not a minimum path cover, it is not difficult to find a graph and rooted path cover \mathcal{R} with $P(\vec{G}_{\mathcal{R}}) < Z(\vec{G}_{\mathcal{R}})$ (in fact, $P(\vec{G}_{\mathcal{R}}) < M(\vec{G}_{\mathcal{R}})$).

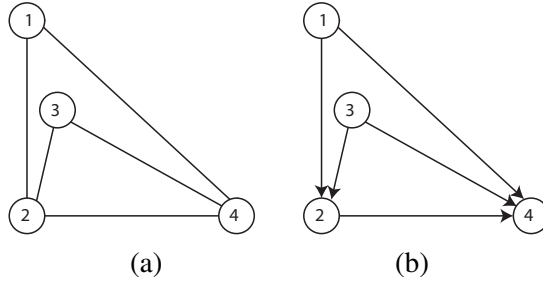


Figure 2. A hierarchal orientation $\vec{G}_{\mathcal{R}}$ having $P(\vec{G}_{\mathcal{R}}) < M(\vec{G}_{\mathcal{R}}) = Z(\vec{G}_{\mathcal{R}})$.

Example 2.13. Let G be the double triangle graph shown in Figure 2(a) and consider the rooted path cover of G defined by $\mathcal{R} = \{R^{(1)}, R^{(2)}, R^{(3)}\}$ where $V(R^{(1)}) = \{1\}$, $V(R^{(2)}) = \{3\}$, and $V(R^{(3)}) = \{2, 4\}$ with 2 as the root of $R^{(3)}$. The hierarchal orientation $\vec{G}_{\mathcal{R}}$ is shown in Figure 2(b).

Then $P(\vec{G}_{\mathcal{R}}) = 2$ because paths $(1, 2)$ and $(3, 4)$ cover all vertices, and the vertices 1 and 3 must each be initial vertices of any path they are in. Let

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

Then $\Gamma(A) = \vec{G}_{\mathcal{R}}$ and nullity $A = 3$. The set $\{1, 2, 3\}$ is a zero forcing set for $\vec{G}_{\mathcal{R}}$, so $3 \leq M(\vec{G}_{\mathcal{R}}) \leq Z(\vec{G}_{\mathcal{R}}) \leq 3$.

Every graph G we have examined has $M(\vec{G}_{\mathcal{R}}) = P(\vec{G}_{\mathcal{R}})$ for minimum rooted path covers \mathcal{R} , but these examples all involve a small number of vertices.

Question 2.14. Does $M(\vec{G}_{\mathcal{R}}) = P(\vec{G}_{\mathcal{R}})$ if \mathcal{R} is a minimum rooted path cover of G ?

Tournaments. A tournament is an orientation of the complete graph K_n . In this section we consider the possible values of path cover number, maximum nullity, and zero forcing number for tournaments.

Example 2.15. We create an orientation of K_n by labeling the vertices $\{1, \dots, n\}$ and by orienting the edges $\{u, v\}$ as (u, v) if and only if $v < u - 1$ or $v = u + 1$. The resulting orientation is called the *Hessenberg tournament* of order n , denoted $\vec{K}_n^{(H)}$. This is the Hessenberg path on n vertices containing all possible arcs except those of the form $(u + 1, u)$ for $1 \leq u \leq n - 1$. Since the zero forcing number of any Hessenberg path is one, $P(\vec{K}_n^{(H)}) = M(\vec{K}_n^{(H)}) = Z(\vec{K}_n^{(H)}) = 1$. Observe that $\vec{K}_n^{(H)}$ is self-complementary as a digraph.

Example 2.16. Label the vertices of K_n by $\{1, \dots, n\}$ and orient the edges $\{u, v\}$ as (u, v) if and only if $u < v$. The resulting orientation is the *transitive tournament*, denoted $\vec{K}_n^{(T)}$. We show that $P(\vec{K}_n^{(T)}) = Z(\vec{K}_n^{(T)}) = M(\vec{K}_n^{(T)}) = \lceil n/2 \rceil$ for any n . Let A be the adjacency matrix of $\vec{K}_n^{(T)}$, and let $D = \text{diag}(0, 1, 0, 1, \dots)$. Then $\Gamma(A + D) = \vec{K}_n^{(T)}$ and $\text{nullity}(A + D) = \lceil n/2 \rceil$ because $A + D$ has $\lfloor n/2 \rfloor$ duplicate rows and, if n is odd, an additional row of zeros. The set of odd numbered vertices $B = \{1, 3, \dots\}$ is a zero forcing set. Thus, $\lceil n/2 \rceil \leq M(\vec{K}_n) \leq Z(\vec{K}_n) \leq \lceil n/2 \rceil$. Furthermore, from the definition of $\vec{K}_n^{(T)}$, no more than 2 vertices can be on the same Hessenberg path.

Proposition 2.17. For any tournament \vec{K}_n , $1 \leq P(\vec{K}_n) \leq \lceil n/2 \rceil$, and for every integer k with $1 \leq k \leq \lceil n/2 \rceil$, there is an orientation \vec{K}_n having $P(\vec{K}_n) = k$.

For every integer k with $1 \leq k \leq \lceil n/2 \rceil$, there is an orientation \vec{K}_n having $Z(\vec{K}_n) = k$.

Proof. For both P and Z , $\vec{K}_n^{(H)}$ (Example 2.15) realizes the lower bound and $\vec{K}_n^{(T)}$ (Example 2.16) realizes the upper bound. For the upper bound on attainable path cover numbers, partition the vertices of \vec{K}_n into $\lceil n/2 \rceil$ sets of size two or one. Each pair of vertices and the arc between them forms a path. The assertion that all values for P and Z between 1 and $\lceil n/2 \rceil$ are possible follows from Corollary 2.2. \square

For $n \leq 7$, the transitive tournament $\vec{K}_n^{(T)}$ achieves the highest zero forcing number; that is, $Z(\vec{K}_n) \leq \lceil n/2 \rceil$ for all orientations \vec{K}_n . (This has been verified using the program [Warnberg 2014], written in Sage.) But for $n = 8$, there exists a tournament having maximum nullity greater than that of the transitive tournament, as in the next example.

Example 2.18. Let \vec{K}_8 be the tournament shown in Figure 3, left (see next page). Observe that $\{1, 2, 3, 4, 8\}$ is a zero forcing set for \vec{K}_8 , so $Z(\vec{K}_8) \leq 5$. The matrix

$$A = \begin{bmatrix} 0 & 1 & 2 & 1 & 2 & 3 & 1 & 2 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}$$

has rank 3 and $\Gamma(A) = \vec{K}_8$, so $5 \leq M(\vec{K}_8)$. Thus, $Z(\vec{K}_8) = M(\vec{K}_8) = 5 > 4 = \lceil 8/2 \rceil$. We also show that $P(\vec{K}_8) = 3$. Since $\{(2, 4, 8), (3, 5, 7), (1, 6)\}$ is a path cover, $P(\vec{K}_8) \leq 3$. There are no induced paths of length greater than two in \vec{K}_8 , so by Lemma 1.3, any path of length three or more must have a cycle. Thus vertices 1 and 6 must be in paths of length at most two. If they are in separate paths in a path

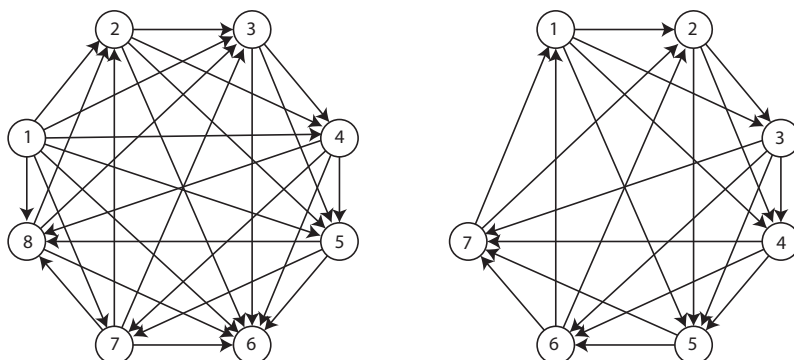


Figure 3. Left: a tournament \vec{K}_8 having $M(\vec{K}_8) = Z(\vec{K}_8) = 5 > \lceil \frac{8}{2} \rceil$. Right: A tournament \vec{K}_7 having $M(\vec{K}_7) = 3 < 4 = Z(\vec{K}_7)$.

cover \mathcal{P} , then $|\mathcal{P}| \geq 3$. So assume $(1, 6)$ is a path in \mathcal{P} . Since $\vec{K}_8 - \{1, 6\}$ is not a (Hessenberg) path, $|\mathcal{P}| \geq 3$.

There are also examples of tournaments \vec{K}_n for which $M(\vec{K}_n) < Z(\vec{K}_n)$.

Proposition 2.19. *The tournament \vec{K}_7 , shown in Figure 3, right, has $P(\vec{K}_7) = 2$, $M(\vec{K}_7) = 3$, and $Z(\vec{K}_7) = 4$.*

Proof. Because $\{(4, 6, 1, 3), (2, 5, 7)\}$ is a path cover for \vec{K}_7 , and \vec{K}_7 is not a Hessenberg path, $P(\vec{K}_7) = 2$.

Next we show $M(\vec{K}_7) \leq 3$. Suppose $\Gamma(A) = \vec{K}_7$. The nonzero pattern of A is

$$\begin{bmatrix} ? & * & * & * & * & 0 & 0 \\ 0 & ? & * & * & * & 0 & 0 \\ 0 & 0 & ? & * & * & * & * \\ 0 & 0 & 0 & ? & * & * & * \\ 0 & 0 & 0 & 0 & ? & * & * \\ * & * & 0 & 0 & 0 & ? & * \\ * & * & 0 & 0 & 0 & 0 & ? \end{bmatrix},$$

where $*$ denotes a nonzero entry and $?$ may have any real value. By considering columns 2, 3, and 6, we see that rows 1, 5, and 7 are necessarily linearly independent.

For A to achieve nullity 4, we must have $\text{rank } A = 3$ and thus all the remaining rows must be in the span of rows 1, 5, and 7. We show this is impossible, implying that $\text{mr}(\vec{K}_7) \geq 4$ and $M(\vec{K}_7) \leq 3$. Once that is done, we can construct a matrix A with $\Gamma(A) = \vec{K}_7$ and $\text{rank } A = 4$ by setting all nonzero off-diagonal entries to 1 and setting the diagonal entries as $a_{ii} = 0$ for i odd and $a_{ii} = 1$ for i even, so $M(\vec{K}_7) = 3$.

If $a_{11} = 0$, then row 3 cannot be expressed as a linear combination of rows 1, 5, and 7: By considering column 1, the coefficient of row 7 must be zero, which

implies that the coefficient of row 1 must be zero by considering column 2. But, by considering column 4, row 3 is not a multiple of row 5. Thus $a_{11} \neq 0$.

If $a_{77} \neq 0$, then row 2 cannot be expressed as a linear combination of rows 1, 5, and 7: By considering column 6, the coefficient of row 5 must be zero, which implies that the coefficient of row 7 must be zero by considering column 7. But, by considering column 1, row 2 is not a multiple of row 1 (because $a_{11} \neq 0$). Thus $a_{77} = 0$.

If $a_{55} = 0$, then row 4 cannot be expressed as a linear combination of rows 1, 5, and 7: By considering column 3, the coefficient of row 1 must be zero, which implies that the coefficient of row 7 must be zero by considering column 1. But row 4 is not a multiple of row 5 (because $a_{55} = 0$). Thus $a_{55} \neq 0$.

Now row 6 cannot be expressed as a linear combination of rows 1, 5, and 7: By considering column 3, the coefficient of row 1 must be zero. By considering column 5, the coefficient of row 5 must be zero. Now by considering column 7, row 6 is not a scalar multiple of row 7. Therefore row 6 is not a linear combination of rows 1, 5, and 7, and thus $\text{rank } A \geq 4$ and $\text{mr}(\vec{K}_7) \geq 4$.

Finally we show that $Z(\vec{K}_7) = 4$. Observe that any zero forcing set must contain a vertex from the set $\{1, 2\}$: if 1 and 2 are initially colored white, the only vertices that can force them are 6 and 7, but 1 and 2 are both out-neighbors of 6 and 7. Observe that any zero forcing set must contain a vertex from the set $\{6, 7\}$: if 6 and 7 are initially colored white, the only vertices that can force them are 3, 4, and 5, but 6 and 7 are both out-neighbors of 3, 4, and 5. Observe that any zero forcing set must contain a vertex from the set $\{3, 4\}$: if 3 and 4 are initially colored white, the only vertices that can force them are 1 and 2, but 3 and 4 are both out-neighbors of 1 and 2. Observe that any zero forcing set must contain a vertex from the set $\{4, 5\}$: if 4 and 5 are initially colored white, the only vertices that can force them are 1, 2, and 3, but 4 and 5 are both out-neighbors of 1, 2, and 3. Hence, a zero forcing set must contain at least four vertices, unless vertex 4 is the only vertex from $\{3, 4\}$ and $\{4, 5\}$ selected. However, by inspection the sets $\{1, 4, 6\}$, $\{1, 4, 7\}$, $\{2, 4, 6\}$, $\{2, 4, 7\}$ are not zero forcing sets. The set $\{1, 2, 4, 6\}$ is a zero forcing set for \vec{K}_7 , and so $Z(\vec{K}_7) = 4$. \square

Orientations of paths. In this section we consider the possible values of path cover number, maximum nullity, and zero forcing number for orientations of paths.

Example 2.20. Starting with the path P_n , label the vertices in path order by $\{1, \dots, n\}$ and orient the edge $\{i, i + 1\}$ as arc $(i, i + 1)$ for $i = 1, \dots, n - 1$. The resulting orientation is the *path orientation* of P_n , denoted $\vec{P}_n^{(H)}$. Then $P(\vec{P}_n^{(H)}) = M(\vec{P}_n^{(H)}) = Z(\vec{P}_n^{(H)}) = 1$.

Example 2.21. Starting with the path P_n , label the vertices in path order by $\{1, \dots, n\}$ and orient the edges as follows: Orient $\{1, 2\}$ as $(1, 2)$. For $i = 1, \dots, \lfloor n/2 \rfloor - 1$, orient $\{2i + 1, 2i\}$ and $\{2i + 1, 2i + 2\}$ as $(2i + 1, 2i)$ and $(2i + 1, 2i + 2)$. If n is odd, orient $\{n - 1, n\}$ as $(n, n - 1)$. The resulting orientation

is the *alternating orientation* of P_n , denoted $\vec{P}_n^{(A)}$. Note that all odd-numbered vertices have in-degree zero. So $P(\vec{P}_n^{(A)}) = M(\vec{P}_n^{(A)}) = Z(\vec{P}_n^{(A)}) = \lceil n/2 \rceil$ because the odd vertices form a minimum zero forcing set, there is no directed path of length greater than one, and the adjacency matrix A of $\vec{P}_n^{(A)}$ has rank $\lfloor n/2 \rfloor$.

Proposition 2.22. *For any oriented path \vec{P}_n , we have $1 \leq P(\vec{P}_n) \leq \lceil n/2 \rceil$, $1 \leq M(\vec{P}_n) \leq \lceil n/2 \rceil$, and $1 \leq Z(\vec{P}_n) \leq \lceil n/2 \rceil$. For every integer k with $1 \leq k \leq \lceil n/2 \rceil$, there are (possibly three different) orientations \vec{P}_n having $P(\vec{P}_n) = k$, $M(\vec{P}_n) = k$, and $Z(\vec{P}_n) = k$.*

Proof. The proof of the second statement follows from Examples 2.20 and 2.21 and Corollary 2.2. To complete the proof, we show that $Z(\vec{P}_n) \leq \lceil n/2 \rceil$ for every orientation \vec{P}_n . Apply Corollary 2.2 to the given orientation \vec{P}_n and to $\vec{P}_n^{(H)}$. Then $Z(\vec{P}_n) - Z(\vec{P}_n^{(H)}) \leq \lfloor (n-1)/2 \rfloor$, so $Z(\vec{P}_n) \leq \lfloor (n-1)/2 \rfloor + 1$. If n is odd, $\lfloor (n-1)/2 \rfloor + 1 = (n-1)/2 + 1 = \lceil n/2 \rceil$. If n is even, $\lfloor (n-1)/2 \rfloor + 1 = (n/2 - 1) + 1 = \lceil n/2 \rceil$. Therefore $Z(\vec{P}_n) \leq \lceil n/2 \rceil$. \square

Orientations of cycles. In this section we consider the possible values of path cover number, maximum nullity, and zero forcing number for orientations of cycles of length at least 4 (since a cycle of length 3 is a complete graph).

Example 2.23. Starting with the cycle C_n , label the vertices in cycle order by $\{1, \dots, n\}$ and orient the edge $\{i, i+1\}$ as arc $(i, i+1)$ for $i = 1, \dots, n$ (where $n+1$ is interpreted as 1). The resulting orientation is the *cycle orientation* of C_n , denoted $\vec{C}_n^{(H)}$. Then $P(\vec{C}_n^{(H)}) = M(\vec{C}_n^{(H)}) = Z(\vec{C}_n^{(H)}) = 1$.

Example 2.24. Starting with C_n , label the vertices in cycle order by $\{1, \dots, n\}$ and orient the edges as follows: Orient $\{1, 2\}$ and $\{1, n\}$ as $(1, 2)$ and $(1, n)$. For $i = 1, \dots, \lfloor n/2 \rfloor - 1$, orient $\{2i+1, 2i\}$ and $\{2i+1, 2i+2\}$ as arcs $(2i+1, 2i)$ and $(2i+1, 2i+2)$. If n is odd, orient the edge $\{n-1, n\}$ as $(n, n-1)$. The resulting orientation is the *alternating orientation* of C_n , denoted $\vec{C}_n^{(A)}$. If n is odd, there is one path of length 2, so $P(\vec{C}_n^{(A)}) = \lfloor n/2 \rfloor$. Let S be the set of odd-numbered vertices (with the exception of vertex n if n is odd), so every vertex in S has in-degree zero and $|S| = \lfloor n/2 \rfloor$. Clearly $S \subseteq B$ for any zero forcing set, and every vertex in S has two out-neighbors not in S , so every zero forcing set must have cardinality at least $\lfloor n/2 \rfloor + 1$. Since $S \cup \{2\}$ is a zero forcing set, $Z(\vec{C}_n^{(A)}) = \lfloor n/2 \rfloor + 1$. We can construct a matrix $A \in \mathcal{M}(\vec{C}_n^{(A)})$ of nullity $\lfloor n/2 \rfloor + 1$, showing that $M(\vec{C}_n^{(A)}) = \lfloor n/2 \rfloor + 1$. Any matrix in $\mathcal{M}(\vec{C}_n^{(A)})$ has two nonzero off-diagonal entries in every odd row (except n if n is odd) and no nonzero off-diagonal entries in every even row. Define a matrix $A = [a_{ij}]$ with $\Gamma(A) = \vec{C}_n^{(A)}$ by setting $a_{ii} = 0$, with the exception that $a_{nn} = -1$ if n is odd, and in each odd row the first nonzero entry is 1 and the second is -1 . Then A has nullity $\lfloor n/2 \rfloor + 1$.

Proposition 2.25. *Let \vec{C}_n be any orientation of C_n ($n \geq 4$). Then $1 \leq P(\vec{C}_n) \leq \lfloor n/2 \rfloor$ and for every integer k with $1 \leq k \leq \lfloor n/2 \rfloor$, there is an orientation \vec{C}_n having $P(\vec{C}_n) = k$. For any orientation of a cycle \vec{C}_n , we have $1 \leq M(\vec{C}_n) \leq Z(\vec{C}_n) \leq \lfloor n/2 \rfloor + 1$ and for every integer k with $1 \leq k \leq \lfloor n/2 \rfloor + 1$, there are (possibly two different) orientations \vec{C}_n having $M(\vec{C}_n) = k$ and $Z(\vec{C}_n) = k$.*

Proof. The proof of the second part of each statement follows from Examples 2.23 and 2.24 and Corollary 2.2. To complete the proof, we show that $P(\vec{C}_n) \leq \lfloor n/2 \rfloor$ and $Z(\vec{C}_n) \leq \lfloor n/2 \rfloor + 1$ for all orientations \vec{C}_n by exhibiting a path cover and zero forcing set of cardinality not exceeding this bound.

For path cover number: If n is even, choose two adjacent vertices, cover them with one path, and delete them, leaving a path on $n - 2$ vertices which is an even number. By Proposition 2.22, there is a path cover of these $n - 2$ vertices with $n/2 - 1$ paths, so there is a path cover of \vec{C}_n having $n/2 = \lfloor n/2 \rfloor$ paths. If n is odd, then for any orientation \vec{C}_n , there is a path on 3 vertices. Cover these vertices with that path and delete them, leaving a path on $n - 3$ vertices (again an even number), which can be covered by $(n - 3)/2$ paths, and there is a path cover of \vec{C}_n having $(n - 3)/2 + 1 = \lfloor n/2 \rfloor$ paths.

For zero forcing number: Delete any one vertex v , leaving a path on $n - 1$ vertices, which has a zero forcing set B with $|B| = \lceil (n - 1)/2 \rceil$ by Proposition 2.22. Then the set $B' := B \cup \{v\}$ is a zero forcing set for \vec{C}_n and $|B'| = \lceil (n - 1)/2 \rceil + 1$. If n is even, $\lceil (n - 1)/2 \rceil + 1 = n/2 + 1 = \lfloor n/2 \rfloor + 1$. If n is odd, $\lceil (n - 1)/2 \rceil + 1 = (n - 1)/2 + 1 = \lfloor n/2 \rfloor + 1$. □

3. Doubly directed graphs

Given a graph G , the *doubly directed graph* $\vec{\vec{G}}$ of G is the digraph obtained by replacing each edge $\{u, v\}$ by both of the arcs (u, v) and (v, u) . In this section we establish results for minimum rank, maximum nullity, zero forcing number, and path cover number of doubly directed graphs.

Proposition 3.1. $P(G) = P(\vec{\vec{G}})$ for any graph G .

Proof. Now, $P(G)$ is the minimum number of induced paths of G and $P(\vec{\vec{G}})$ is the minimum number of Hessenberg paths in $\vec{\vec{G}}$. It is enough to show that all Hessenberg paths in $\vec{\vec{G}}$ are induced. Suppose P is a Hessenberg path in $\vec{\vec{G}}$ that is not induced. Then there exists some arc $(v_i, v_j) \in E(\vec{\vec{G}})$, where $i > j + 1$. But because the digraph is doubly directed, $(v_j, v_i) \in E(\vec{\vec{G}})$, which contradicts the definition of a Hessenberg path. Therefore, all Hessenberg paths must be induced. Thus, $P(G) = P(\vec{\vec{G}})$. □

Proposition 3.2. For any graph G , we have $Z(G) = Z(\vec{\vec{G}})$.

Proof. The color change rule for graphs is that a blue vertex v can force a white vertex w if w is the only white neighbor of v . The color change rule for digraphs is that a blue vertex v can force a white vertex w if w is the only white out-neighbor of v . If G is a graph then for any vertex $v \in V(G)$, w is a neighbor of v in G if and only if w is an out-neighbor of v in \vec{G} . This means that v forces w in G if and only if v forces w in \vec{G} . Thus B is a zero forcing set in G if and only if B is a zero forcing set in \vec{G} and $Z(G) = Z(\vec{G})$. \square

Observation 3.3. For any graph G , we have $M(G) \leq M(\vec{G})$, since $\mathcal{S}(G) \subseteq \mathcal{M}(\vec{G})$.

Corollary 3.4. If G is a graph such that $M(G) = Z(G)$, then $M(G) = M(\vec{G})$.

Proof. $Z(G) = Z(\vec{G}) \geq M(\vec{G}) \geq M(G) = Z(G)$. \square

It was established in [AIM 2008] that for every tree T , $P(T) = M(T) = Z(T)$, giving the following corollary.

Corollary 3.5. If T is a tree, then $P(\vec{T}) = M(\vec{T}) = Z(\vec{T})$.

As the following example shows, it is possible to have $M(\vec{G}) > M(G)$.

Example 3.6. The complete tripartite graph on three sets of three vertices $K_{3,3,3}$ has $V_1 = \{1, 2, 3\}$, $V_2 = \{4, 5, 6\}$, $V_3 = \{7, 8, 9\}$, $V(K_{3,3,3}) = V_1 \dot{\cup} V_2 \dot{\cup} V_3$ and $E(K_{3,3,3})$ equal to the set of all edges with one vertex in V_i and the other in V_j ($i \neq j$). It is well known that $\text{mr}(K_{3,3,3}) = 3$. Let J_3 be the 3×3 matrix with all entries equal to 1, 0_3 be the 3×3 matrix with all entries equal to 0, and let

$$A = \begin{bmatrix} 0_3 & J_3 & -J_3 \\ J_3 & 0_3 & J_3 \\ J_3 & J_3 & 0_3 \end{bmatrix}.$$

Then $\Gamma(A) = \vec{K}_{3,3,3}$ and $\text{rank } A = 2$. Thus, $M(\vec{K}_{3,3,3}) = 7 > 6 = M(K_{3,3,3})$.

The pentasun H_5 graph shown in Figure 5, left, has $M(H_5) = 2 < 3 = P(H_5)$ [Barioli et al. 2004], establishing the noncomparability of M and P (because there are many examples of graphs G with $P(G) < M(G)$). The same is true for the doubly directed pentasun.

Example 3.7. Theorem 2.8 of [Berliner et al. 2013] describes the cut-vertex reduction method for calculating M for directed graphs with a cut-vertex. We compute $M(\vec{H}_5) = 2$ by applying the cut-vertex reduction method to vertex v , using the notation found in [ibid.]. Because $M(H_5 - w) = Z(H_5 - w) = 2$ and $M(H_5 - \{v, w\}) = Z(H_5 - \{v, w\}) = 2$, we have

$$M(\vec{H} - w) = Z(\vec{H} - w) = 2 \quad \text{and} \quad M(\vec{H} - \{v, w\}) = Z(\vec{H} - \{v, w\}) = 2,$$

so $r_v(\vec{H} - w) = 1$. Clearly $\text{mr}(\vec{H}[\{v, w\}]) = 1$ and $\text{mr}(\vec{H}[\{v, w\}] - v) = 0$, so $r_v(\vec{H}[\{v, w\}]) = 1$. The *type* of the cut-vertex v of a digraph Γ , denoted $\text{type}_v(\Gamma)$,

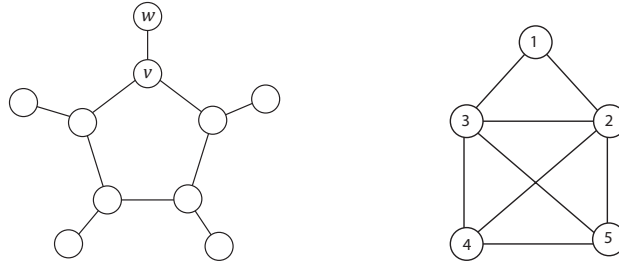


Figure 5. Left: the pentasun H_5 . Right: the full-house graph.

is a subset of $\{C, R\}$, where $C \in \text{type}_v(\Gamma)$ if there exists a matrix $A' \in M(\Gamma - v)$ with $\text{rank } A' = \text{mr}(\Gamma - v)$ and a vector z in $\text{range } A'$ that has the in-pattern of v , and similarly for rows. Thus, $\text{type}_v(\vec{H}[\{v, w\}]) = \emptyset$, so by [ibid., Theorem 2.8], $r_v(\vec{H}) = 2$. So,

$$\text{mr}(\vec{H}) = \text{mr}(\vec{H} - \{v, w\}) + \text{mr}(\vec{H}[\{w\}]) + 2 = 6 + 0 + 2 = 8,$$

and $M(\vec{H}) = 2$. Since $P(\vec{H}) = P(H_5) = 3$, $P(\vec{H}) > M(\vec{H})$. It is easy to find an example of a digraph Γ with $P(\Gamma) < M(\Gamma)$ (e.g., Example 2.13), so M and P are noncomparable.

Proposition 3.8. *Suppose that both G and \vec{G} have field independent minimum rank. Then $\text{mr}(G) = \text{mr}(\vec{G})$ and $M(G) = M(\vec{G})$.*

Proof. Since both G and \vec{G} have field independent minimum rank, $\text{mr}(G) = \text{mr}^{\mathbb{Z}_2}(G)$ and $\text{mr}^{\mathbb{Z}_2}(\vec{G}) = \text{mr}(\vec{G})$. Furthermore, $\mathcal{S}^{\mathbb{Z}_2}(G) = \mathcal{M}^{\mathbb{Z}_2}(\vec{G})$, so

$$\text{mr}(G) = \text{mr}^{\mathbb{Z}_2}(G) = \text{mr}^{\mathbb{Z}_2}(\vec{G}) = \text{mr}(\vec{G}). \quad \square$$

The converse of Proposition 3.8 is not true, however.

Example 3.9. Let G be the full-house graph, shown in Figure 5, right. It is well known that $\text{mr}^{\mathbb{Z}_2}(G) = 3$, yet $\text{mr}(G) = 2 = \text{mr}(\vec{G})$.

4. Digraphs in general

In this section, we present some minimum rank, maximum nullity, and zero forcing results for digraphs in general, where any pair of vertices may or may not have an arc in either direction. We begin with two (undirected) graph properties that do not extend to digraphs.

Sinkovic [2010] has shown that for any outerplanar graph G , $M(G) \leq P(G)$. This is not true for digraphs, because it was shown in Example 2.13 that the outerplanar digraph \vec{G}_R has $M(\vec{G}_R) = Z(\vec{G}_R) = 3 > 2 = P(\vec{G}_R)$, and \vec{G}_R is outerplanar (although Figure 2 is not drawn that way).

The *complement* of a graph $G = (V, E)$ (or digraph $\Gamma = (V, E)$) is the graph $\bar{G} = (V, \bar{E})$ (or digraph $\bar{\Gamma} = (V, \bar{E})$), where \bar{E} consists of all two element sets of vertices (or all ordered pairs of distinct vertices) that are not in E . The graph complement conjecture (GCC) is equivalent to the statement that for any graph G , $M(G) + M(\bar{G}) \geq |G| - 2$. This statement is generalized in [Barioli et al. 2012]: For a graph parameter β related to maximum nullity, the graph complement conjecture for β , denoted GCC_β , is $\beta(G) + \beta(\bar{G}) \geq |G| - 2$. With this notation, the GCC can be denoted GCC_M . The graph complement conjecture for zero forcing number, $Z(G) + Z(\bar{G}) \geq |G| - 2$, denoted GCC_Z , is actually the graph complement theorem for zero forcing [Ekstrand et al. 2012]. However, as the following example shows, the GCC_Z does not hold for digraphs, and since for any digraph $M(\Gamma) \leq Z(\Gamma)$, the GCC_M does not hold for digraphs. A tournament provides a counterexample.

Example 4.1. For the Hessenberg tournament of order n , denoted $\vec{K}_n^{(H)}$, we have $Z(\vec{K}_n^{(H)}) = 1$ because \vec{H}_n is a Hessenberg path. Because $\vec{K}_n^{(H)}$ is self-complementary,

$$Z(\vec{K}_n^{(H)}) + Z(\overline{\vec{K}_n^{(H)}}) = 2,$$

but for $n \geq 5$, we have $n - 2 = |\vec{K}_n^{(H)}| - 2 \geq 3$.

Some properties of minimum rank for graphs do remain true for digraphs. For a graph G , it is well known that if K_r is a subgraph of G then $M(G) \geq r - 1$ (see, for example, [Barioli et al. 2013] and the references therein). An analogous result holds true for digraphs, although the proof is different than those usually given for graphs.

Theorem 4.2. *Suppose \vec{K}_r is a subgraph of a digraph Γ . Then $M(\Gamma) \geq r - 1$.*

Proof. First, we order the vertices of Γ so that the subdigraph induced on the vertices $1, 2, \dots, r$ is \vec{K}_r . We will construct $L \in \mathcal{M}(\Gamma)$ with $\text{rank } L \leq n - r + 1$, where L is partitioned as $L = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$ and $A \in \mathcal{M}(K_r)$. We may choose $D \in \mathcal{M}(\Gamma[\{r + 1, \dots, n\}])$ so that $\text{rank } D = n - r$. We now choose C to be any matrix with the correct zero-nonzero pattern. Denote the i -th column of C by \mathbf{c}_i and the j -th column of D by \mathbf{d}_j . Since D has full rank, there exist coefficients $d_{i,1}, \dots, d_{i,n-r}$ such that $\mathbf{c}_i = d_{i,1}\mathbf{d}_1 + \dots + d_{i,n-r}\mathbf{d}_{n-r}$ for $1 \leq i \leq r$.

Now, we choose B to be any matrix with the correct zero-nonzero pattern and denote the j -th column of B by \mathbf{b}_j . Then define E to be the $r \times r$ matrix whose i -th column is equal to $d_{i,1}\mathbf{b}_1 + \dots + d_{i,n-r}\mathbf{b}_{n-r}$. Therefore, the matrix $L' = \begin{bmatrix} E & B \\ C & D \end{bmatrix}$ has $\text{rank } L' = n - r$. Let p be a real number greater than the absolute value of every entry of E . Define $A := E + pJ_r$, where J_r is the $r \times r$ matrix with all entries equal to 1, so $A \in \mathcal{M}(K_r)$, $L = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \in \mathcal{M}(\Gamma)$ and $\text{rank } L \leq n - r + 1$. \square

In [Butler and Young 2013], the maximum number of edges in a graph G of order n with a prescribed zero forcing number k is shown to be $kn - \binom{k+1}{2}$. We

similarly seek to bound the number of arcs that a digraph Γ of order n may possess given $Z(\Gamma) = k$.

Theorem 4.3. *Suppose Γ is a digraph of order n with $Z(\Gamma) = k$. Then,*

$$|E(\Gamma)| \leq \binom{n}{2} - \binom{k}{2} + k(n - 1). \tag{1}$$

Proof. We prove that (1) holds for a digraph Γ of order $n \geq k + 1$ whenever $Z(\Gamma) \leq k$ (since for a graph Γ of order n , $Z(\Gamma) \leq n - 1$). The proof is by induction on n , for a fixed positive integer k . The base case is $n = k + 1$, or $k = n - 1$, and the inequality (1) reduces to

$$|E(\Gamma)| \leq \binom{n}{2} - \binom{n-1}{2} + (n-1)^2 = \frac{n(n-1)}{2} - \frac{(n-1)(n-2)}{2} + (n-1)^2 = n(n-1).$$

Any digraph Γ of order n has at most $2\binom{n}{2} = n^2 - n$ arcs and thus the claim holds.

Now assume (1) holds true for any digraph of order $n - 1$ that has a zero forcing set of cardinality k . Let Γ be a digraph of order n with $Z(\Gamma) \leq k$. Let B be a zero forcing set of Γ , where $|B| = k$. Suppose \mathcal{F} is a chronological list of forces for B and that the first force of \mathcal{F} occurs on the arc (v, w) . Then, $(B \setminus \{v\}) \cup \{w\}$ is a zero forcing set of cardinality k for $\Gamma - v$ and the induction hypothesis applies to $E(\Gamma - v)$. In order to determine an upper bound on $|E(\Gamma)|$, we determine the maximum number of arcs incident with v in Γ . Since v forces w first in \mathcal{F} , $(v, x) \in E(\Gamma)$ implies $x \in (B \setminus \{v\}) \cup \{w\}$. Furthermore, $E(\Gamma)$ contains at most $n - 1$ arcs of the form (x, v) , one for each vertex $x \neq v$. Therefore,

$$\begin{aligned} |E(\Gamma)| &\leq |E(\Gamma - v)| + k + (n - 1) \leq \left(\binom{n-1}{2} - \binom{k}{2}\right) + k(n - 2) + k + (n - 1) \\ &= \binom{n}{2} - \binom{k}{2} + k(n - 1). \quad \square \end{aligned}$$

In the paper [Butler and Young 2013], the edge bound is used to show that the zero forcing number must be at least half the average degree. However, a Hessenberg tournament (see Example 2.15) has half of all possible arcs and $Z(\vec{H}_n) = 1$, so the analogous result is not true for digraphs, and any correct result of this type for digraphs is not likely to be useful.

For a digraph Γ , where $Z(\Gamma) = k$, Theorem 4.3 gives an upper bound for the number of arcs Γ may possess. However, the proof also suggests that equality is achievable in (1) when $n > k$. The following provides a construction of a class of digraphs for which (1) is sharp.

Theorem 4.4. *Let k be a fixed positive integer. Then for each integer $n > k$ and each partition $\pi = (n_1, \dots, n_k)$ of n , there exists a digraph $\Gamma_{n,k,\pi}$ of order n for which $Z(\Gamma_{n,k,\pi}) = k$, the forcing chains of $\Gamma_{n,k,\pi}$ have lengths n_1, n_2, \dots, n_k*

respectively, and

$$|E(\Gamma_{n,k,\pi})| = \binom{n}{2} - \binom{k}{2} + k(n-1).$$

Proof. For $1 \leq i \leq k$, let Γ_i be a full Hessenberg path on n_i vertices. Among all the Γ_i , there are a total of $\sum_{i=1}^k [(n_i - 1) + \binom{n_i}{2}]$ arcs. Within each Γ_i , we denote the initial vertex of the Hessenberg path by b_i and the terminal vertex of the Hessenberg path by t_i . We define $B = \{b_i : 1 \leq i \leq k\}$ and $T = \{t_i : 1 \leq i \leq k\}$, and note that B and T will intersect if $n_i = 1$ for some i . To create $\Gamma_{n,k,\pi}$, we start with $\bigcup_{i=1}^k \Gamma_i$ and add arcs between the Γ_i in the following manner:

- (1) For $1 \leq j < i \leq k$, we add all arcs from vertices in Γ_i to vertices in Γ_j . This adds a total of $\sum_{i < j} n_i n_j$ arcs.
- (2) Add all arcs from vertex t_i to vertices in other Γ_j . For each i , this adds $\sum_{j \neq i} n_j = n - n_i$ arcs. Over all i , this adds $kn - \sum_{i=1}^k n_i = (k-1)n$ total arcs.

Some arcs have been double-counted, which must be reflected in the overall total. In particular, arcs from t_i to all vertices in Γ_j (for $j < i$) have been double-counted. For an arc from t_i to a vertex v of Γ_j , where $v \neq t_j$, we replace the double-counted arc by an arc from v to b_i . Therefore, we need only remove from the total count the number of arcs from t_j to t_i for $1 \leq i < j \leq k$. There are a total of $\binom{k}{2}$ such arcs. Thus, we have

$$\begin{aligned} |E| &= \sum_{i=1}^k \left[(n_i - 1) + \binom{n_i}{2} \right] + \sum_{1 \leq i < j \leq k} n_i n_j + (k-1)n - \binom{k}{2} \\ &= n - k + \left(\sum_{i=1}^k \binom{n_i}{2} + \sum_{1 \leq i < j \leq k} n_i n_j \right) + kn - n - \binom{k}{2} \\ &= \binom{n}{2} - \binom{k}{2} + k(n-1). \end{aligned}$$

By Theorem 4.3, we know that $Z(\Gamma_{n,k,\pi}) \geq k$. We claim that B is a zero forcing set for $\Gamma_{n,k,\pi}$ and that a chronological list of forces exists for which the forcing chains have lengths n_1, \dots, n_k respectively. Assume that each vertex of B is blue. Now Γ_1 is a Hessenberg path and the only arcs coming from vertices of Γ_1 point to vertices in B , with the exception of arcs coming from t_1 . Thus, forcing may occur along Γ_1 , where t_1 is the last vertex forced. We then proceed to Γ_2 and so on through all Γ_i . When we get to Γ_i , the only arcs coming from vertices of Γ_i point to vertices in B or to the already blue vertices of Γ_h , where $h < i$, with the exception of arcs coming from t_i (which is not used to perform a force). So, forcing may occur along Γ_i until all vertices are blue. Therefore B is a zero forcing set for $\Gamma_{n,k,\pi}$ and $Z(\Gamma_{n,k,\pi}) = k$. □

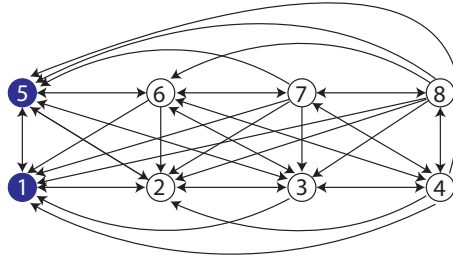


Figure 6. A digraph with the maximum number of arcs for its zero forcing number.

Although the digraphs $Z(\Gamma_{n,k,\pi})$ achieve equality in the bound (1), these are not the only digraphs that do so.

Example 4.5. Let Γ be the digraph of order $n = 8$ in Figure 6. Note that $\{1, 5\}$ is a zero forcing set of Γ . Since $\deg^+(v) \geq 2$ for all vertices v , we have $k = Z(\Gamma) = 2$. Also note that Γ has 41 arcs, the same number of arcs as each digraph $\Gamma_{8,2,\pi}$.

In all of the digraphs $\Gamma_{n,k,\pi}$ constructed in Theorem 4.4, the forcing process may be completed one forcing chain at a time, and we show that this is not true for Γ . Since $\deg^+(v) \geq 3$ for all vertices v other than vertex 1, vertex 1 must be contained in any minimum zero forcing set of Γ along with one of the two out-neighbors of vertex 1. Therefore, the only minimum zero forcing sets are $B_1 = \{1, 5\}$ and $B_2 = \{1, 2\}$. If we color the vertices of B_1 blue, then the first three forces must occur along the arcs $(1, 2)$, $(2, 3)$, and $(5, 6)$, in that order. If we color the vertices of B_2 blue, then the first three forces must occur along the arcs $(1, 5)$, $(2, 3)$, and $(5, 6)$, in that order. In either case, neither of the two forcing chains is completely blue before the forcing process must begin on the other. Therefore, Γ is not equal to any of the $\Gamma_{n,2,\pi}$ constructed in Theorem 4.4.

Although $M(\Gamma)$ does not necessarily equal $Z(\Gamma)$ for all digraphs Γ , we get equality for all $\Gamma_{n,k,\pi}$ constructed in the proof of Theorem 4.4.

Proposition 4.6. *If k and n are positive integers where $k < n$ and $\Gamma_{n,k,\pi}$ is one of the digraphs constructed in the proof of Theorem 4.4, then $M(\Gamma_{n,k,\pi}) = Z(\Gamma_{n,k,\pi})$.*

Proof. We adopt the notation and definitions used in the proof of Theorem 4.4. By construction, each of the k vertices in T has an arc to every other vertex of $\Gamma_{n,k,\pi}$. So for all vertices $v \notin T$, we have $\deg^-(v) \geq k$. Now we consider $t_i \in T$. If $t_i \neq b_i$, then there is an arc to t_i from another vertex of Γ_i . There are also arcs to t_i from all other vertices of T , and therefore $\deg^-(t_i) \geq k$. We now consider the case where $t_i = b_i$. By construction, the subgraph induced on B is \vec{K}_k and thus there is an arc to t_i from each of the other $k - 1$ vertices of B . Furthermore, since $n > k$, there is a vertex $t_j \in T$ for which $t_j \neq b_j$. By construction, there is also

an arc from t_j to t_i , and therefore $\deg^-(t_i) \geq k$. This implies that $\delta^-(\Gamma_{n,k,\pi}) \geq k$. Since $M(\Gamma_{n,k,\pi}) \geq \max\{\delta^-(\Gamma_{n,k,\pi}), \delta^+(\Gamma_{n,k,\pi})\}$ [Berliner et al. 2013], we have $k \leq M(\Gamma_{n,k,\pi}) \leq Z(\Gamma_{n,k,\pi}) \leq k$, and so $M(\Gamma_{n,k,\pi}) = Z(\Gamma_{n,k,\pi})$. \square

For $k = n - 1$, $\Gamma_{n,k,\pi}$ is the digraph \vec{K}_n , so for $n \geq 4$, we have $P(\Gamma_{n,k,\pi}) = \lfloor n/2 \rfloor < n - 1 = Z(\Gamma_{n,k,\pi})$.

References

- [AIM 2008] AIM Minimum Rank – Special Graphs Work Group, “Zero forcing sets and the minimum rank of graphs”, *Linear Algebra Appl.* **428**:7 (2008), 1628–1648. MR 2008m:05166 Zbl 1135.05035
- [Barioli et al. 2004] F. Barioli, S. Fallat, and L. Hogben, “Computation of minimal rank and path cover number for certain graphs”, *Linear Algebra Appl.* **392** (2004), 289–303. MR 2005i:05115 Zbl 1052.05045
- [Barioli et al. 2012] F. Barioli, W. Barrett, S. M. Fallat, H. T. Hall, L. Hogben, and H. van der Holst, “On the graph complement conjecture for minimum rank”, *Linear Algebra Appl.* **436**:12 (2012), 4373–4391. MR 2917415 Zbl 1241.05064
- [Barioli et al. 2013] F. Barioli, W. Barrett, S. M. Fallat, H. T. Hall, L. Hogben, B. Shader, P. van den Driessche, and H. van der Holst, “Parameters related to tree-width, zero forcing, and maximum nullity of a graph”, *J. Graph Theory* **72**:2 (2013), 146–177. MR 3010007 Zbl 1259.05112
- [Berliner et al. 2013] A. Berliner, M. Catral, L. Hogben, M. Huynh, K. Lied, and M. Young, “Minimum rank, maximum nullity, and zero forcing number of simple digraphs”, *Electron. J. Linear Algebra* **26** (2013), 762–780. MR 3141806 Zbl 1282.05095
- [Butler and Young 2013] S. Butler and M. Young, “Throttling zero forcing propagation speed on graphs”, *Australas. J. Combin.* **57** (2013), 65–71. MR 3135948 Zbl 1293.05220
- [Ekstrand et al. 2012] J. Ekstrand, C. Erickson, D. Hay, L. Hogben, and J. Roat, “Note on positive semidefinite maximum nullity and positive semidefinite zero forcing number of partial 2-trees”, *Electron. J. Linear Algebra* **23** (2012), 79–87. MR 2889573 Zbl 1252.05118
- [Hogben 2010] L. Hogben, “Minimum rank problems”, *Linear Algebra Appl.* **432**:8 (2010), 1961–1974. MR 2011b:15002 Zbl 1213.05036
- [Sinkovic 2010] J. Sinkovic, “Maximum nullity of outerplanar graphs and the path cover number”, *Linear Algebra Appl.* **432**:8 (2010), 2052–2060. MR 2011e:05155 Zbl 1201.05061
- [Warnberg 2014] N. Warnberg, “Zero forcing number of simple digraphs”, 2014, https://github.com/warnberg/zero_forcing.

Received: 2013-12-31 Accepted: 2014-04-30

| | |
|---------------------|--|
| berliner@stolaf.edu | Department of Mathematics, Statistics, and Computer Science, St. Olaf College, Northfield, MN 55057, United States |
| brow3138@umn.edu | Department of Mathematics, Carleton College, Northfield, MN 55057, United States |
| jmsdg7@iastate.edu | Department of Mathematics, Iowa State University, Ames, IA 50011, United States |
| coxn42@gmail.com | Department of Mathematics, Statistics, and Computer Science, St. Olaf College, Northfield, MN 55057, United States |

| | |
|----------------------------|--|
| hogben@aimath.org | <i>Department of Mathematics, Iowa State University, Ames, IA 50011, and American Institute of Mathematics, 600 East Brokaw Road, San Jose, CA 95112 United States</i> |
| jason.hu@berkeley.edu | <i>Department of Mathematics, University of California, Berkeley, Berkeley, CA 94720, United States</i> |
| klj02011@mymail.pomona.edu | <i>Department of Mathematics, Pomona College, 610 North College Avenue, Claremont, CA 91711, United States</i> |
| manternachk1@central.edu | <i>Department of Mathematics and Computer Science, Central College, Pella, IA 50219, United States</i> |
| tpeters319@gmail.com | <i>Natural and Mathematical Science Division, Culver-Stockton College, Canton, MO 63435, United States</i> |
| nwarnberg@uwlax.edu | <i>Department of Mathematics, Iowa State University, Ames, IA 50011, United States</i> |
| myoung@iastate.edu | <i>Department of Mathematics, Iowa State University, Ames, IA 50011, United States</i> |

Braid computations for the crossing number of Klein links

Michael Bush, Danielle Shepherd, Joseph Smith,
Sarah Smith-Polderman, Jennifer Bowen and John Ramsay

(Communicated by Colin Adams)

Klein links are a nonorientable counterpart to torus knots and links. It is shown that braids representing a subset of Klein links take on the form of a very positive braid after manipulation. Once the braid has reached this form, its number of crossings is the crossing number of the link it represents. Two formulas are proven to calculate the crossing number of $K(m, n)$ Klein links, where $m \geq n \geq 1$. In combination with previous results, these formulas can be used to calculate the crossing number for any Klein link with given values of m and n .

1. Introduction

A key aspect in the classification of distinct knots and links is the crossing number, a link invariant. The crossing number of a link A , denoted $c(A)$, is the minimum number of crossings that can occur in any projection of the link [Adams 2004]. Through the use of Alexander–Briggs notation, prime links are placed into finite sets based on both their crossing number and number of components [Adams 2004; Rolfsen 1976]. This paper will use Alexander–Briggs notation, specifically corresponding to the labels given by Rolfsen [1976], where the 4_1^2 link has four crossings, two components, and is the first link listed with these invariant values. Braid relations are used to simplify the general braid word for Klein links, which allows us to find their minimal number of crossings.

2. Torus links and Klein links

A torus link is a link that can be placed on the surface of a torus such that it does not cross over itself [Adams 2004]. Torus links are denoted $T(m, n)$, where m is the number of times the link wraps around the longitude of the torus, and n is the number of times it wraps around the meridian. Torus links are a commonly studied

MSC2010: 57M25, 57M27.

Keywords: knots and links in S^3 , invariants of knots and 3-manifolds.

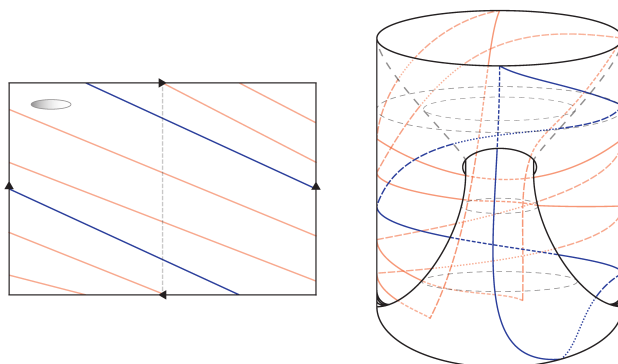


Figure 1. $K(5, 3)$ on the identified rectangular representation of a Klein bottle and on the equivalent once punctured Klein bottle. Dashed lines represent portions of the link that lie on hidden surfaces of the Klein bottle.

class of links and formulas that can be used to determine many of their invariants are known. Given the values of m and n , the crossing number can be computed with the formula $c(T(m, n)) = m(n - 1)$, where $m \geq n$ [Murasugi 1991; Williams 1988].

Similarly, Klein links are links that can be placed on the surface of a once punctured Klein bottle so that they do not intersect themselves. One method used to form this set of Klein links begins with the identified rectangular representation of the Klein bottle seen in Figure 1. For these Klein links, $K(m, n)$, the m strands originating on the left side of the rectangular diagram are placed to remain entirely below the “hole” representing the self-intersection of the once punctured Klein bottle, and the n strands originating from the top remain entirely above the hole [Bowen et al. 2014; Catalano et al. 2010; Shepherd et al. 2012; Freund and Smith-Polderman 2013]. After a link is formed, the Klein bottle is removed and the link is classified based on its invariants.

3. Braids

Braids are a useful technique for representing and classifying links since all links can be represented by braids [Adams 2004]. A braid is a set of strings connected between a top and bottom bar such that each string always progresses downwards as it crosses above or below the other strings [Adams 2004; Shepherd et al. 2012]. The strings of an n -braid are numbered from 1 to n , going from the leftmost to the rightmost string. A closed braid representation of a link is formed when these top and bottom bars are connected and the corresponding strings are attached. When describing braids, *braid words* are commonly used due to their simplicity and

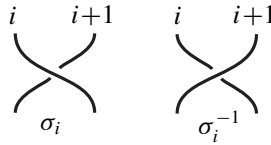


Figure 2. Braid generators [Freund and Smith-Polderman 2013].

usefulness. Each crossing is labeled using σ_i^ϵ , where i represents the i -th strand of the braid crossing over or under the $(i+1)$ -st strand, as illustrated in Figure 2. When the i -th strand crosses over the $(i+1)$ -st strand, $\epsilon = 1$ and when it crosses under, $\epsilon = -1$.

Braids are commonly used to study Klein links and torus links since the corresponding braids are known for given values of m and n . The properties of these braids are exploited to find new properties of the links.

Proposition 1 [Adams 2004]. *A general braid word for a torus link is given by $(\sigma_1\sigma_2 \cdots \sigma_{n-1})^m$ when $m \geq 1$ and $n \geq 2$.*

Proposition 2 [Shepherd et al. 2012; Freund and Smith-Polderman 2013]. *A general braid word for a $K(m, n)$ Klein link composes the general braid word of a torus link with the half twist $\prod_{i=1}^{n-1} (\sigma_{n-1}^{-1}\sigma_{n-2}^{-1} \cdots \sigma_i^{-1})$, shown in Figure 3, which gives*

$$K(m, n) = (\sigma_1\sigma_2 \cdots \sigma_{n-1})^m \prod_{i=1}^{n-1} (\sigma_{n-1}^{-1}\sigma_{n-2}^{-1} \cdots \sigma_i^{-1}).$$

Unlike the general braid word for torus links, the general braid word for Klein links can be manipulated with braid relations to reduce the number of crossings in the braid [Murasugi 1991; Williams 1988].

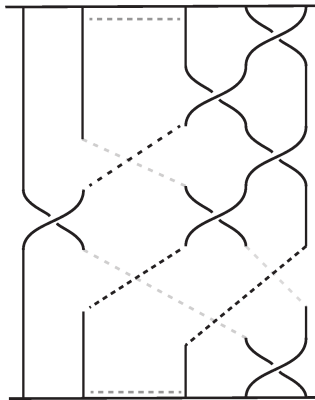


Figure 3. A half twist on an n -strand braid [Shepherd et al. 2012].

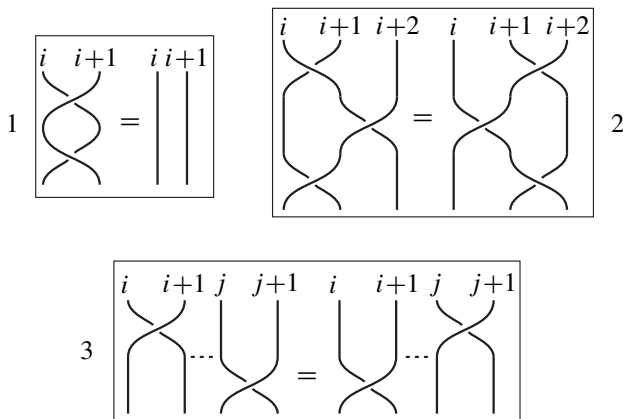


Figure 4. Braid moves 1, 2, and 3 [Freund and Smith-Polderman 2013].

Definition 3 [Adams 2004; Freund and Smith-Polderman 2013]. *Braid relations*, corresponding to the Reidemeister moves for links, allow a braid to be transformed between equivalent forms without altering the link that the closed braid represents. The first three braid moves are shown in Figure 4, and conjugation and stabilization are shown in Figure 5.

Move 1: $\sigma_i \sigma_i^{-1} = 1 = \sigma_i^{-1} \sigma_i$

Move 2: $\sigma_i \sigma_{i+1} \sigma_i = \sigma_{i+1} \sigma_i \sigma_{i+1}$

Move 3: For $|i - j| > 1$, $\sigma_i \sigma_j = \sigma_j \sigma_i$

Conjugation: For an n -string braid word z , we have $z = \sigma_i z \sigma_i^{-1} = \sigma_i^{-1} z \sigma_i$ for i from 1 to $n - 1$.

Stabilization: For an n -string braid word z , we have $z = z \sigma_n$ or $z = z \sigma_n^{-1}$, resulting in an $(n + 1)$ -string braid word. Also for an $(n + 1)$ -string braid word z , assuming z does not contain σ_n or σ_n^{-1} , stabilization allows $z \sigma_n = z$ or $z = z \sigma_n^{-1}$, resulting in an n -string braid word.

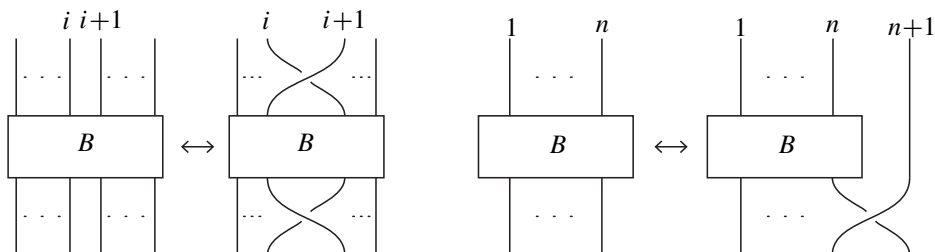


Figure 5. Left: conjugation. Right: stabilization. See [Freund and Smith-Polderman 2013].

When $m \geq n$, a generalized sequence of the first and third braid moves is used to manipulate the general braid word of a Klein link into a form w that untangles the negative half-twist in Lemma 4 below.

Lemma 4. For $K(m, n)$ where $m \geq n$, a simplified version of the braid word, w , is

$$w = (\sigma_1\sigma_2 \cdots \sigma_{n-1})^{m-n+1}(\sigma_1\sigma_2 \cdots \sigma_{n-2})(\sigma_1\sigma_2 \cdots \sigma_{n-3}) \cdots \sigma_1.$$

Proof. For $m \geq n$, a standard $K(m, n)$ braid can be simplified using the following sequence of braid moves 1 and 3:

$$\begin{aligned} K(m, n) &= (\sigma_1\sigma_2 \cdots \sigma_{n-1})^m (\sigma_{n-1}^{-1}\sigma_{n-2}^{-1} \cdots \sigma_1^{-1})(\sigma_{n-1}^{-1}\sigma_{n-2}^{-1} \cdots \sigma_2^{-1}) \cdots \sigma_{n-1}^{-1} \\ &= (\sigma_1\sigma_2 \cdots \sigma_{n-1})^{m-1} (\sigma_{n-1}^{-1}\sigma_{n-2}^{-1} \cdots \sigma_2^{-1})(\sigma_{n-1}^{-1}\sigma_{n-2}^{-1} \cdots \sigma_3^{-1}) \cdots \sigma_{n-1}^{-1} \\ &= (\sigma_1\sigma_2 \cdots \sigma_{n-1})^{m-2} (\sigma_{n-1}^{-1}\sigma_{n-2}^{-1} \cdots \sigma_3^{-1})(\sigma_{n-1}^{-1}\sigma_{n-2}^{-1} \cdots \sigma_4^{-1}) \cdots \sigma_{n-1}^{-1}\sigma_1 \\ &= (\sigma_1\sigma_2 \cdots \sigma_{n-1})^{m-3} (\sigma_{n-1}^{-1}\sigma_{n-2}^{-1} \cdots \sigma_4^{-1}) \cdots \sigma_{n-1}^{-1}\sigma_1\sigma_2\sigma_1 = \cdots \\ &= (\sigma_1\sigma_2 \cdots \sigma_{n-1})^{m-n+2} \sigma_{n-1}^{-1} (\sigma_1\sigma_2 \cdots \sigma_{n-3})(\sigma_1\sigma_2 \cdots \sigma_{n-4}) \cdots \sigma_1 \\ &= (\sigma_1\sigma_2 \cdots \sigma_{n-1})^{m-n+1} (\sigma_1\sigma_2 \cdots \sigma_{n-2})(\sigma_1\sigma_2 \cdots \sigma_{n-3}) \cdots \sigma_1 \quad \square \end{aligned}$$

In this braid word w , all crossings are positive ($\epsilon = 1$ for all σ_i^ϵ), which means it is classified as a homogeneous braid and a positive braid, as defined below.

Definition 5 [Murasugi 1991]. A braid $\gamma = \sigma_{i_1}^{\epsilon_1} \cdots \sigma_{i_k}^{\epsilon_k}$ is a *homogeneous braid* if $\epsilon_j = \epsilon_i$ ($\epsilon_i = \pm 1$) whenever $i_j = i_i$.

Definition 6. A homogeneous braid a , is a *positive braid* if $\epsilon_j = \epsilon_i$ for all σ_i .

The following definitions and properties provide important information about another class of braids, very positive braids.

Definition 7 [Franks and Williams 1987]. A braid with r strands has a *full twist* (Δ^2) if the braid word contains $(\sigma_1\sigma_2\sigma_3 \cdots \sigma_{r-1})^r$.

Note that a full twist can occur at any point within a braid as shown in Figure 6.

Definition 8 [Franks and Williams 1987]. A positive braid with a full twist is a *very positive braid*.

Definition 9. The link invariant *braid index*, denoted $b(L)$, is the minimum number of strands needed to represent a link L as a braid.

Proposition 10 [Franks and Williams 1987; Williams 1988]. When a braid p is a very positive braid, $b(p) = s$, where s is the number of strands in the very positive braid representation of p .

Theorem 11 [Murasugi 1991]. A homogeneous n -braid h , where $b(h) = n$, has the minimal number of crossings for the link it represents.

These properties are combined to form an important crossing number result for very positive braids.

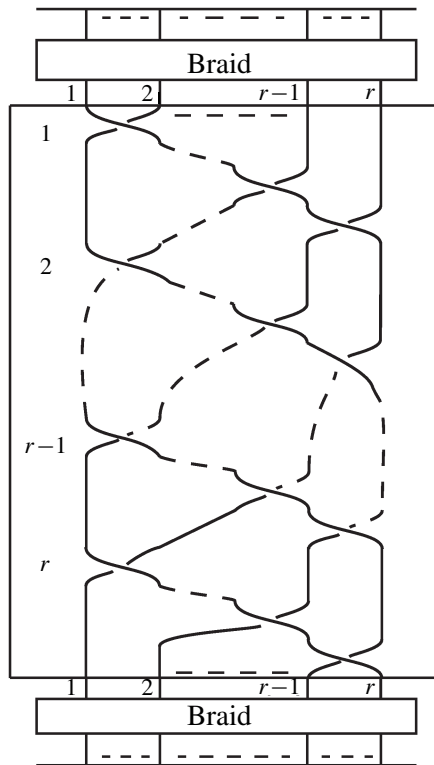


Figure 6. A full twist on an r -strand braid.

Lemma 12. *A very positive braid representation of a link has minimal crossings for that link.*

Proof. Let p be a very positive braid. By Proposition 10, we know $b(p)$ is equal to the number of strands in p and p is a homogeneous braid by Definition 5. Thus, by Theorem 11, a very positive braid contains exactly the number of crossings as the crossing number of the link it represents. \square

Very positive braids are useful for determining properties of links since invariants including the crossing number and braid index can be found from braids in this form. For certain values of m and n , w is already in this form and in other cases, the braid word can be simplified into this form. In determining the crossing number for these links, it is useful to know the number of crossings contained within the half-twist of the Klein link braid word.

Lemma 13. *The number of crossings in a half-twist of an n -braid is*

$$\sum_{i=1}^{n-1} i = \frac{n^2 - n}{2}.$$

Proof. The half-twist $\prod_{i=1}^{n-1} (\sigma_{n-1}^{-1} \sigma_{n-2}^{-1} \cdots \sigma_i^{-1})$, illustrated in Figure 3, has a crossing for each σ term in the product, or $(n-1) + (n-2) + \cdots + 2 + 1$. \square

4. Crossing number theorem

For certain values of m and n , w is a very positive braid, which means that the crossing number for the corresponding Klein link can be easily determined.

Theorem 14. For $m \geq n \geq 1$ and $m \geq 2n - 1$,

$$c(K(m, n)) = m(n-1) - \frac{n^2 - n}{2}.$$

Proof. Consider the simplified version of the braid word of $K(m, n)$ from Lemma 4,

$$w = (\sigma_1 \sigma_2 \cdots \sigma_{n-1})^{m-n+1} (\sigma_1 \sigma_2 \cdots \sigma_{n-2}) (\sigma_1 \sigma_2 \cdots \sigma_{n-3}) \cdots \sigma_1.$$

This braid word contains the same number of crossings as $c(T(m, n)) - (n^2 - n)/2$ due to the reduction process in Lemma 4, which removed one crossing from the torus braid for each crossing in the Klein link half-twist corresponding to the use of braid move 1. Referring to Definition 7, this braid word contains a full twist when $m - n + 1 \geq n$ since $\sigma_1 \sigma_2 \sigma_3 \cdots \sigma_{r-1}$ must occur at least r times and $r = n$. Thus, when $m \geq 2n - 1$, the simplified braid word will be very positive, and by Section 3, will have the minimal number of crossings. \square

5. Finding very positive braid representations

For other values of m and n , a full twist is not contained within w , so only an upper bound on the crossing number is initially known. Since w is a positive braid, stabilization is the only braid relation that can remove crossings. The following example illustrates how braid relations reduce the $K(6, 5)$ to a very positive braid. For simplicity, *subwords* will be specific patterns of consecutive σ_i terms within a braid word.

Example. Let us demonstrate the stabilization process to obtain a full twist on a $K(6, 5)$. First we will consider the reduced braid word w of the $K(6, 5)$,

$$\sigma_1 \sigma_2 \sigma_3 \sigma_4 \sigma_1 \sigma_2 \sigma_3 \sigma_4 \sigma_1 \sigma_2 \sigma_3 \sigma_1 \sigma_2 \sigma_1.$$

One can see that there are two subwords of $\sigma_1 \sigma_2 \sigma_3 \sigma_4$ and three subwords of $\sigma_1 \sigma_2 \sigma_3$, but these do not satisfy the requirements of a full twist. Thus, when reexamining the braid word, one can see that there are at least three subwords of $\sigma_1 \sigma_2$, satisfying the requirements of a full twist if put in order (on a three strand braid):

$$\underline{\sigma_1 \sigma_2} \sigma_3 \sigma_4 \underline{\sigma_1 \sigma_2} \sigma_3 \sigma_4 \underline{\sigma_1 \sigma_2} \sigma_3 \underline{\sigma_1 \sigma_2} \sigma_1.$$

Using braid moves (noted before they are applied) with two stabilizations, we will manipulate the braid word to obtain a full twist, where we use $\bar{\sigma}_i$ to indicate σ_i^{-1} :

$K(6, 5)$

$$\begin{aligned}
&= \sigma_1 \sigma_2 \sigma_3 \underline{\sigma_4 \sigma_1 \sigma_2 \sigma_3} \sigma_4 \sigma_1 \sigma_2 \sigma_3 \sigma_1 \sigma_2 \sigma_1 && \text{(braid move 3)} \\
&= [\sigma_1 \sigma_2 \sigma_3 \sigma_1 \sigma_2 \underline{\sigma_4 \sigma_3 \sigma_4} \sigma_1 \sigma_2 \sigma_3 \sigma_1 \sigma_2 \sigma_1] && \text{(braid move 2, conjugation)} \\
&= \sigma_3 \sigma_1 \sigma_2 \sigma_3 \sigma_1 \sigma_2 \sigma_1 [\sigma_1 \sigma_2 \sigma_3 \sigma_1 \sigma_2 \sigma_3 \sigma_4 \underline{\sigma_3 \sigma_1 \sigma_2 \sigma_3 \sigma_1 \sigma_2 \sigma_1}] \bar{\sigma}_1 \bar{\sigma}_2 \bar{\sigma}_1 \bar{\sigma}_3 \bar{\sigma}_2 \bar{\sigma}_1 \bar{\sigma}_3 && \text{(braid move 1)} \\
&= \underline{\sigma_3 \sigma_1 \sigma_2 \sigma_3 \sigma_1 \sigma_2 \sigma_1 \sigma_1 \sigma_2 \sigma_3 \sigma_1 \sigma_2 \sigma_3} \underline{\sigma_4} && \text{(braid move 3, first stabilization)} \\
&= \sigma_1 \underline{\sigma_3 \sigma_2 \sigma_3} \sigma_1 \sigma_2 \sigma_1 \sigma_1 \sigma_2 \underline{\sigma_3 \sigma_1} \sigma_2 \sigma_3 && \text{(braid move 2, braid move 1)} \\
&= [\sigma_1 \sigma_2 \sigma_3 \sigma_2 \sigma_1 \sigma_2 \sigma_1 \sigma_1 \sigma_2 \sigma_1 \sigma_3 \sigma_2 \sigma_3] && \text{(braid move 2, braid move 2, conjugation)} \\
&= \sigma_3 \sigma_2 [\sigma_1 \sigma_2 \sigma_3 \sigma_1 \sigma_2 \sigma_1 \sigma_1 \sigma_2 \sigma_1 \sigma_2 \underline{\sigma_3 \sigma_2}] \bar{\sigma}_2 \bar{\sigma}_3 && \text{(braid move 1)} \\
&= \sigma_3 \underline{\sigma_2 \sigma_1 \sigma_2} \sigma_3 \sigma_1 \sigma_2 \sigma_1 \sigma_1 \sigma_1 \sigma_2 \sigma_1 \sigma_2 && \text{(braid move 2)} \\
&= \underline{\sigma_3 \sigma_1 \sigma_2 \sigma_1} \underline{\sigma_3 \sigma_1 \sigma_2 \sigma_1 \sigma_1 \sigma_1 \sigma_2 \sigma_1 \sigma_2} && \text{(braid move 3, braid move 3)} \\
&= [\sigma_1 \underline{\sigma_3 \sigma_2 \sigma_3} \sigma_1 \sigma_1 \sigma_2 \sigma_1 \sigma_1 \sigma_1 \sigma_2 \sigma_1 \sigma_2] && \text{(braid move 2, conjugation)} \\
&= \bar{\sigma}_2 \bar{\sigma}_1 [\sigma_1 \sigma_2 \sigma_3 \sigma_2 \sigma_1 \sigma_1 \sigma_2 \sigma_1 \sigma_1 \sigma_1 \sigma_2 \sigma_1 \sigma_2 \sigma_1 \sigma_2] && \text{(braid move 1)} \\
&= \underline{\sigma_3 \sigma_2 \sigma_1 \sigma_1 \sigma_2 \sigma_1 \sigma_1 \sigma_1 \sigma_2 \sigma_1 \sigma_2 \sigma_1 \sigma_2} && \text{(second stabilization)} \\
&= \sigma_2 \sigma_1 \sigma_1 \sigma_2 \sigma_1 \sigma_1 \underline{\sigma_1 \sigma_2 \sigma_1 \sigma_2 \sigma_1 \sigma_2} .
\end{aligned}$$

This positive braid contains a full twist after two stabilization moves. Note that this is one way to obtain a full twist, and the full twist may not always appear at the beginning or end of the braid word.

This process of finding the number of stabilization moves needed to find a very positive form of the Klein link is generalized in Theorem 16 below. The set S in Lemma 15 is used to help determine the number of stabilization moves needed to manipulate the braid into a very positive form.

Lemma 15. *The set S , defined as*

$$S = \{k \in \mathbb{Z}^+ \mid \sigma_1 \sigma_2 \cdots \sigma_{k-1} \text{ occurs at least } k \text{ times in } w\},$$

is nonempty and finite for $K(m, n)$ when $1 \leq n \leq m < 2n - 1$.

Proof. There will always be at least two σ_1 terms in w from Lemma 4, since $m \geq n$ and $m - n + 1 \geq 1$. Thus, because at least the first term and the last term of the braid word must each be σ_1 , we have $2 \in S$ and S is nonempty. The set S is finite because there are exactly n strands in w ; thus if $j > n$, then $j \notin S$. \square

Theorem 16. For $1 \leq n \leq m < 2n - 1$,

$$c(K(m, n)) = m(n - 1) - \frac{n^2 - n}{2} - \left\lfloor \frac{2n - m}{2} \right\rfloor.$$

Proof. Consider the simplified version of the braid word of $K(m, n)$ from Lemma 4,

$$w = (\sigma_1\sigma_2 \cdots \sigma_{n-1})^{m-n+1}(\sigma_1\sigma_2 \cdots \sigma_{n-2})(\sigma_1\sigma_2 \cdots \sigma_{n-3}) \cdots \sigma_1.$$

Referring to the definition of a full twist, one can see that this braid word (before manipulation using braid moves) will never contain a full twist because $m - n + 1 < n$. Since there is not a full twist, the braid is positive, but not very positive and the braid index and crossing number remain unknown.

In order to become a very positive braid, a braid representing a Klein link must be transformed so that it is a positive braid with a full twist. Referring to Lemma 15 with $m < 2n - 1$, one can identify the presence of at least k subwords of the form $\sigma_1\sigma_2 \cdots \sigma_{k-1}$, where k is a positive integer. Lemma 15 shows S to be nonempty and finite; let $r = \max(S)$. Therefore, the subword $\sigma_1\sigma_2 \cdots \sigma_{r-1}$ occurs at least r times in w .

If a subword $\sigma_1\sigma_2 \cdots \sigma_{k-1}$ occurs exactly $k + 1$ times in a braid word w , then r must equal k . This means the subword $\sigma_1\sigma_2 \cdots \sigma_k$ must occur k times due to the form of w . Assume $k \neq r$, then $k + 1 \in S$, since $k \neq \max(S)$. Since the subword $\sigma_1\sigma_2 \cdots \sigma_k$ does not occur $k + 1$ times, $k + 1 \notin S$; this is a contradiction, and therefore $k = r = \max(S)$.

Assume there are $r + 2$ subwords of the form $\sigma_1\sigma_2 \cdots \sigma_{r-1}$. This implies that there exist $r + 1$ subwords of the form $\sigma_1\sigma_2 \cdots \sigma_r$ as seen from the simplified braid word w . This implies that $r + 1 \in S$ and therefore $r \neq \max(S)$, which is a contradiction. This means there will not be $r + 2$ subwords of the form $\sigma_1\sigma_2 \cdots \sigma_{r-1}$ when $r = \max(S)$. Similarly, when there exist more than $r + 2$ subwords of the form $\sigma_1\sigma_2 \cdots \sigma_{r-1}$, there is a value $k \in S$ such that $k > r$; so $r \neq \max(S)$, which is a contradiction. Therefore, only r or $r + 1$ subwords of the form $\sigma_1\sigma_2 \cdots \sigma_{r-1}$ can exist in the simplified braid word of a Klein link where $m < 2n - 1$. We consider these two cases separately.

Case 1. This case examines these simplified braids with r subwords of the form $\sigma_1\sigma_2 \cdots \sigma_{r-1}$. From the simplified braid word form w , it is known that there are $m - n + 1$ subwords of the form $\sigma_1\sigma_2 \cdots \sigma_{n-1}$, where n represents the initial number of strands in the braid. For each stabilization, the number of strands in the braid is decreased by one, and the number of subwords of $\sigma_1\sigma_2 \cdots \sigma_{n'-1}$, where n' is the number of strands in the braid, is increased by one since the maximum index $n' - 1$ is decreased with each stabilization. If x is equal to the number of stabilizations that must be used to obtain a full twist, then this relationship gives

$$(m - n + 1) + x = n - x.$$

Solving this equation for x yields

$$x = \frac{2n - m - 1}{2}.$$

Case 2. Now this case will examine when $r + 1$ subwords of the form $\sigma_1\sigma_2 \cdots \sigma_{r-1}$ are present in the simplified braid word of a Klein link. Similar to Case 1, it is known that there are $m - n + 1$ subwords of $\sigma_1\sigma_2 \cdots \sigma_{n-1}$, and each stabilization decreases the number of strands in the braid by one. However, specific to this case, it is known that there is one additional $\sigma_1\sigma_2 \cdots \sigma_{n-1}$ subword that is unnecessary in the formation of the full twist. Thus, where x is still the number of stabilizations needed,

$$(m - n + 1) - 1 + x = n - x.$$

Solving this equation for x yields

$$x = \frac{2n - m}{2}.$$

If the two cases are compared, it can be seen that the values for x only differ by $\frac{1}{2}$. Thus, they can be combined with the relationship

$$x = \left\lfloor \frac{2n - m}{2} \right\rfloor.$$

These stabilizations, which reduce the number of strands in the braid, each correspond to the elimination of one crossing from the reduced braid word. Since the resulting braid word contains a full twist and is positive, the braid is very positive, and by Section 3, has a minimum number of crossings. Thus,

$$c(K(m, n)) = m(n - 1) - \frac{n^2 - n}{2} - \left\lfloor \frac{2n - m}{2} \right\rfloor. \quad \square$$

6. Conclusion

These theorems increase our knowledge of Klein links [Bowen et al. 2014; Catalano et al. 2010; Shepherd et al. 2012; Freund and Smith-Polderman 2013], while providing new properties that can be used to find additional connections between torus links and Klein links. With previous results regarding the crossing number for $K(m, n)$ with $m \leq n$ and for $m = 0$ or $n = 0$, the crossing number for any Klein link in this set can be calculated [Catalano et al. 2010; Shepherd et al. 2012]. Through the use of these theorems, we have completed a catalog of Klein links that lists the crossing number, number of components, and complete Alexander–Briggs notation (if available) for all Klein links between $K(1, 0)$ and $K(8, 8)$ [Bowen et al. 2014].

References

- [Adams 2004] C. C. Adams, *The knot book: An elementary introduction to the mathematical theory of knots*, Amer. Math. Soc., Providence, RI, 2004. MR 2005b:57009 Zbl 1065.57003
- [Bowen et al. 2014] J. Bowen, M. Bush, and J. Smith, “Klein link research”, 2014, Available at <http://discover.wooster.edu/jbowen/research/klein-links/>.
- [Catalano et al. 2010] L. Catalano, D. Freund, R. Ruzvidzo, J. Bowen, and J. Ramsay, “A preliminary study of Klein knots”, pp. 10–17 in *Proceedings of the Midstates Conference for Undergraduate Research in Computer Science and Mathematics* (Wittenberg University, OH, 2010), 2010.
- [Franks and Williams 1987] J. Franks and R. F. Williams, “Braids and the Jones polynomial”, *Trans. Amer. Math. Soc.* **303**:1 (1987), 97–108. MR 88k:57006 Zbl 0647.57002
- [Freund and Smith-Polderman 2013] D. Freund and S. Smith-Polderman, “Klein links and braids”, *Rose-Hulman Undergrad. Math J.* **14**:1 (2013), 71–84. MR 3071244
- [Murasugi 1991] K. Murasugi, “On the braid index of alternating links”, *Trans. Amer. Math. Soc.* **326**:1 (1991), 237–260. MR 91j:57009 Zbl 0751.57008
- [Rolfsen 1976] D. Rolfsen, *Knots and links*, Mathematics Lecture Series **7**, Publish or Perish, Berkeley, CA, 1976. MR 58 #24236 Zbl 0339.55004
- [Shepherd et al. 2012] D. Shepherd, J. Smith, S. Smith-Polderman, J. Bowen, and J. Ramsay, “The classification of a subset of Klein links”, pp. 38–47 in *Proceedings of the Midstates Conference for Undergraduate Research in Computer Science and Mathematics* (Ohio Wesleyan University, OH, 2012), 2012.
- [Williams 1988] R. F. Williams, “The braid index of an algebraic link”, pp. 697–703 in *Braids* (Santa Cruz, CA, 1986), edited by J. S. Birman and A. Libgober, *Contemp. Math.* **78**, Amer. Math. Soc., Providence, RI, 1988. MR 90c:57006 Zbl 0673.57003

Received: 2014-01-29 Revised: 2014-05-29 Accepted: 2014-05-31

| | |
|-------------------------|--|
| mbush16@wooster.edu | The College of Wooster, Wooster, OH 44691, United States |
| dshepherd14@wooster.edu | The College of Wooster, Wooster, OH 44691, United States |
| jsmith15@wooster.edu | The College of Wooster, Wooster, OH 44691, United States |
| litlbup19@aol.com | The College of Wooster, Wooster, OH 44691, United States |
| jbowen@wooster.edu | The College of Wooster, Wooster, OH 44691, United States |
| jramsay@wooster.edu | The College of Wooster, Wooster, OH 44691, United States |

Guidelines for Authors

Authors may submit manuscripts in PDF format on-line at the Submission page at the *Involve* website.

Originality. Submission of a manuscript acknowledges that the manuscript is original and 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 $\text{Bib}\TeX$ is preferred but not required. Tags will be converted to the house format, however, for submission you may use the format of your choice. Links will be provided to all literature with known web locations and authors are encouraged to provide their own links in addition to those supplied in the editorial process.

Figures. Figures must be of publication quality. After acceptance, you will need to submit the original source files in vector graphics format for all diagrams in your manuscript: vector EPS or vector PDF files are the most useful.

Most drawing and graphing packages (Mathematica, Adobe Illustrator, Corel Draw, MATLAB, etc.) allow the user to save files in one of these formats. Make sure that what you are saving is vector graphics and not a bitmap. If you need help, please write to graphics@msp.org with details about how your graphics were generated.

White space. Forced line breaks or page breaks should not be inserted in the document. There is no point in your trying to optimize line and page breaks in the original manuscript. The manuscript will be reformatted to use the journal's preferred fonts and layout.

Proofs. Page proofs will be made available to authors (or to the designated corresponding author) at a Web site in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

involve

2015

vol. 8

no. 1

| | |
|---|-----|
| Efficient realization of nonzero spectra by polynomial matrices NATHAN MCNEW AND NICHOLAS ORMES | 1 |
| The number of convex topologies on a finite totally ordered set TYLER CLARK AND TOM RICHMOND | 25 |
| Nonultrametric triangles in diametral additive metric spaces TIMOTHY FAVER, KATELYNN KOCHALSKI, MATHAV KISHORE MURUGAN, HEIDI VERHEGGEN, ELIZABETH WESSON AND ANTHONY WESTON | 33 |
| An elementary approach to characterizing Sheffer A-type 0 orthogonal polynomial sequences DANIEL J. GALIFFA AND TANYA N. RISTON | 39 |
| Average reductions between random tree pairs SEAN CLEARY, JOHN PASSARO AND YASSER TORUNO | 63 |
| Growth functions of finitely generated algebras ERIC FREDETTE, DAN KUBALA, ERIC NELSON, KELSEY WELLS AND HAROLD W. ELLINGSEN, JR. | 71 |
| A note on triangulations of sumsets KÁROLY J. BÖRÖCZKY AND BENJAMIN HOFFMAN | 75 |
| An exploration of ideal-divisor graphs MICHAEL AXTELL, JOE STICKLES, LANE BLOOME, ROB DONOVAN, PAUL MILNER, HAILEE PECK, ABIGAIL RICHARD AND TRISTAN WILLIAMS | 87 |
| The failed zero forcing number of a graph KATHERINE FETCIE, BONNIE JACOB AND DANIEL SAAVEDRA | 99 |
| An Erdős–Ko–Rado theorem for subset partitions ADAM DYCK AND KAREN MEAGHER | 119 |
| Nonreal zero decreasing operators related to orthogonal polynomials ANDRE BUNTON, NICOLE JACOBS, SAMANTHA JENKINS, CHARLES MCKENRY JR., ANDRZEJ PIOTROWSKI AND LOUIS SCOTT | 129 |
| Path cover number, maximum nullity, and zero forcing number of oriented graphs and other simple digraphs ADAM BERLINER, CORA BROWN, JOSHUA CARLSON, NATHANAEL COX, LESLIE HOGBEN, JASON HU, KATRINA JACOBS, KATHRYN MANTERNACH, TRAVIS PETERS, NATHAN WARNBERG AND MICHAEL YOUNG | 147 |
| Braid computations for the crossing number of Klein links MICHAEL BUSH, DANIELLE SHEPHERD, JOSEPH SMITH, SARAH SMITH-POLDERMAN, JENNIFER BOWEN AND JOHN RAMSAY | 169 |



1944-4176(201501)8:1;1-5