

# involve

a journal of mathematics

## Editorial Board

Kenneth S. Berenhaut, *Managing Editor*

Colin Adams	David Larson
John V. Baxley	Suzanne Lenhart
Arthur T. Benjamin	Chi-Kwong Li
Martin Bohner	Robert B. Lund
Nigel Boston	Gaven J. Martin
Amarjit S. Budhiraja	Mary Meyer
Pietro Cerone	Emil Minchev
Scott Chapman	Frank Morgan
Jem N. Corcoran	Mohammad Sal Moslehian
Toka Diagana	Zuhair Nashed
Michael Dorff	Ken Ono
Sever S. Dragomir	Timothy E. O'Brien
Behrouz Emamizadeh	Joseph O'Rourke
Joel Foisy	Yuval Peres
Errin W. Fulp	Y.-F. S. Pétermann
Joseph Gallian	Robert J. Plemmons
Stephan R. Garcia	Carl B. Pomerance
Anant Godbole	Bjorn Poonen
Ron Gould	Józeph 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



# involve

msp.org/involve

## INVOLVE YOUR STUDENTS IN RESEARCH

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

### MANAGING EDITOR

Kenneth S. Berenhaut Wake Forest University, USA

### BOARD OF EDITORS

Colin Adams	Williams College, USA	Suzanne Lenhart	University of Tennessee, USA
John V. Baxley	Wake Forest University, NC, USA	Chi-Kwong Li	College of William and Mary, USA
Arthur T. Benjamin	Harvey Mudd College, USA	Robert B. Lund	Clemson University, USA
Martin Bohner	Missouri U of Science and Technology, USA	Gaven J. Martin	Massey University, New Zealand
Nigel Boston	University of Wisconsin, USA	Mary Meyer	Colorado State University, USA
Amarjit S. Budhiraja	U of North Carolina, Chapel Hill, USA	Emil Minchev	Ruse, Bulgaria
Pietro Cerone	La Trobe University, Australia	Frank Morgan	Williams College, USA
Scott Chapman	Sam Houston State University, USA	Mohammad Sal Moslehian	Ferdowsi University of Mashhad, Iran
Joshua N. Cooper	University of South Carolina, USA	Zuhair Nashed	University of Central Florida, USA
Jem N. Corcoran	University of Colorado, USA	Ken Ono	Emory University, USA
Toka Diagana	Howard University, USA	Timothy E. O'Brien	Loyola University Chicago, USA
Michael Dorff	Brigham Young University, USA	Joseph O'Rourke	Smith College, USA
Sever S. Dragomir	Victoria University, Australia	Yuval Peres	Microsoft Research, USA
Behrouz Emamizadeh	The Petroleum Institute, UAE	Y.-F. S. Pétermann	Université de Genève, Switzerland
Joel Foisy	SUNY Potsdam, USA	Robert J. Plemmons	Wake Forest University, USA
Errin W. Fulp	Wake Forest University, USA	Carl B. Pomerance	Dartmouth College, USA
Joseph Gallian	University of Minnesota Duluth, USA	Vadim Ponomarenko	San Diego State University, USA
Stephan R. Garcia	Pomona College, USA	Bjorn Poonen	UC Berkeley, USA
Anant Godbole	East Tennessee State University, USA	James Propp	U Mass Lowell, USA
Ron Gould	Emory University, USA	József H. Przytycki	George Washington University, USA
Andrew Granville	Université Montréal, Canada	Richard Rebarber	University of Nebraska, USA
Jerold Griggs	University of South Carolina, USA	Robert W. Robinson	University of Georgia, USA
Sat Gupta	U of North Carolina, Greensboro, USA	Filip Saidak	U of North Carolina, Greensboro, USA
Jim Haglund	University of Pennsylvania, USA	James A. Sellers	Penn State University, USA
Johnny Henderson	Baylor University, USA	Andrew J. Sterge	Honorary Editor
Jim Hoste	Pitzer College, USA	Ann Trenk	Wellesley College, USA
Natalia Hritonenko	Prairie View A&M University, USA	Ravi Vakil	Stanford University, USA
Glenn H. Hurlbert	Arizona State University, USA	Antonia Vecchio	Consiglio Nazionale delle Ricerche, Italy
Charles R. Johnson	College of William and Mary, USA	Ram U. Verma	University of Toledo, USA
K. B. Kulasekera	Clemson University, USA	John C. Wierman	Johns Hopkins University, USA
Gerry Ladas	University of Rhode Island, USA	Michael E. Zieve	University of Michigan, USA

### PRODUCTION

Silvio Levy, Scientific Editor


Cover: Alex Scorpan

See inside back cover or [msp.org/involve](http://msp.org/involve) for submission instructions. The subscription price for 2018 is US \$190/year for the electronic version, and \$250/year (+\$35, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues and changes of subscriber address should be sent to MSP.

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

Involve peer review and production are managed by EditFLOW<sup>®</sup> from Mathematical Sciences Publishers.

PUBLISHED BY

 **mathematical sciences publishers**  
nonprofit scientific publishing

<http://msp.org/>

© 2018 Mathematical Sciences Publishers

# A mathematical model of treatment of cancer stem cells with immunotherapy

Zachary J. Abernathy and Gabrielle Epelle

(Communicated by Kenneth S. Berenhaut)

Using the work of Shelby Wilson and Doron Levy (2012), we develop a mathematical model to study the growth and responsiveness of cancerous tumors to various immunotherapy treatments. We use numerical simulations and stability analysis to predict long-term behavior of passive and aggressive tumors with a range of antigenicities. For high antigenicity aggressive tumors, we show that remission is only achieved after combination treatment with TGF- $\beta$  inhibitors and a peptide vaccine. Additionally, we show that combination treatment has limited effectiveness on low antigenicity aggressive tumors and that using TGF- $\beta$  inhibition or vaccine treatment alone proves generally ineffective for all tumor types considered. A key feature of our model is the identification of separate cancer stem cell and tumor cell populations. Our model predicts that even with combination treatment, failure to completely eliminate the cancer stem cell population leads to cancer recurrence.

## 1. Introduction

Cancer is a leading cause of death in the world today. Although an enormous amount of resources have been spent in search of a cure, much is still unknown about the dynamics of how cancer cells are created and destroyed. The general consensus is that cancer is caused by mutated cells, which are unable to die and thus grow uncontrollably, and that cancer requires many mutations to transform normal cells into cancer cells [Li and Neaves 2006]. However, another theory of cancer development, which states that cancer arises from stem cells, is steadily gaining recognition.

---

*MSC2010:* 34D05, 34D20, 92B05, 92C37.

*Keywords:* cancer stem cells, immunotherapy, recurrence, ordinary differential equations, stability.

**1.1. *Stem cells and cancer.*** The cancer stem cell hypothesis originated in 1855 when German pathologist Rudolf Virchow theorized that cancers arise from the activation of inactive embryonic-like cells found in mature tissue [Huntly and Gilliland 2005]. In 1994, John Dick's lab showed the presence of leukemia-inducing stem cells in the blood of mice with acute myeloid leukemia. In 2003 and 2004, Michael Clarke's and Peter Dirks' labs showed the presence of cancer stem cells in breast and brain cancer respectively [Li and Neaves 2006].

Cancer stem cells differ from other tumor cells in their potential for growth, development and differentiation. Unlike other cells, cancer stem cells have the ability to self-renew. A cancer stem cell divides to produce two daughter cells. One daughter remains a stem cell while the other mutates and undergoes further differentiation. Cancer stem cells also have a higher potential for proliferation and a longer life span than other cells [Li and Neaves 2006].

**1.2. *Treatment of cancer stem cells.*** Another difference between cancer stem cells and other tumor cell types is their resistance to radiation and chemotherapy. Although these treatments are able to destroy the differentiated tumor cells, they are relatively ineffective against cancer stem cells, which have mechanisms for repairing DNA and resisting cytotoxic drugs [Deonarain et al. 2009]. Even if such treatments cause the patient to go into remission, in many cases the cancer relapses months or years later due to the presence of cancer stem cells [Cripe et al. 2009]. Further complicating matters is the fact that chemotherapy and radiation have a greater effect on normal cells than cancerous cells. Research shows that chemotherapy and radiation cause normal hematopoietic stem cells, but not cancer stem cells, to undergo senescence or premature aging. This gives the cancer cells a growth advantage over normal cells, especially after several rounds of treatment [Jordan et al. 2006].

**1.2.1. *Immunotherapy.*** Immunotherapy is a form of treatment that aims to improve the ability of the immune system to fight cancer cells [Stewart and Smyth 2011]. One of the major advantages of immunotherapy over traditional cancer treatments, such as radiation and chemotherapy, is that the immune system is much more discriminatory in its actions, targeting only cancer cells and leaving the majority of the healthy tissues of the body unharmed [Joshi et al. 2009]. This lessens the competitive advantage of cancer stem cells over normal stem cells after successive rounds of treatment. Paul Ehrlich, an immunologist in the early 20th century, was the first person to conceive the idea that the immune system is capable of scanning for and eradicating the tumors that arise in our bodies before they become clinically manifested [Malmberg 2004]. Although this idea was controversial at first, experimental evidence has shown that when cancer cells proliferate to a detectable number within the human body, the body's immune system is activated into a "search and destroy" mode. This spontaneous immune response is possible only if

the cancer cells have unique surface markers called tumor specific antigens. Tumor cells that possess these antigens are known as immunogenic cancers [Nani and Freedman 2000]. The recognition of cancerous cells by the immune system is called immune surveillance, and cancer progression occurs when this process fails [Stewart and Smyth 2011].

**1.2.2. TGF- $\beta$ : an agent of both tumor suppression and progression.** Transforming growth factor- $\beta$  (TGF- $\beta$ ) is a protein that acts as a strong inhibitor of cell growth and an inducer of programmed cell death or apoptosis [Akhurst and Derynck 2001]. TGF- $\beta$  is present in both normal and tumor cells. It plays a beneficial role in wound healing, inflammation, and angiogenesis (i.e., new blood vessel formation) [Arciero et al. 2004]. At early stages of tumorigenesis, for example, when the tumor is still benign, TGF- $\beta$  acts directly on cancer cells to suppress tumor growth [Akhurst and Derynck 2001]. However, as time elapses, genetic changes allow TGF- $\beta$  to stimulate tumor progression by its activities on both the cancerous and nonmalignant structural cell types of the tumor. Experimental evidence has shown that small tumors produce little or no TGF- $\beta$ , while large tumors produce large amounts of TGF- $\beta$  and rely heavily on its angiogenesis-promoting and immunosuppressive effects.

The discovery of TGF- $\beta$ 's immunosuppressive effects has led scientists to implement new forms of treatment aimed to inhibit TGF- $\beta$  production. Unfortunately, several studies demonstrate that TGF- $\beta$  inhibition alone is not enough to eliminate tumors. For instance, in [Terabe et al. 2009], the authors examined whether the inhibition of TGF- $\beta$  can enhance immune responses caused by a peptide vaccine. Their goal was to ascertain under which conditions this enhanced tumor response slows down or stops tumor growth in mice. They found that treatment with only anti-TGF- $\beta$  had no impact on tumor growth, but anti-TGF- $\beta$  did greatly enhance the effects of the peptide vaccine. Shelby Wilson and Doron Levy [2012] then developed a mathematical model in order to quantitatively study the results of Terabe et al. Our model modifies the Wilson–Levy model in order to study the effects of TGF- $\beta$  inhibition and vaccine combination treatment on cancer stem cells.

## 2. The Wilson–Levy model

We first present the original Wilson–Levy model for proper context. The model follows the size of a tumor represented by  $T(t)$ , the concentration of TGF- $\beta$  represented by  $G(t)$ , the number of effector cells represented by  $E(t)$ , the number of regulatory T cells represented by  $R(t)$ , and the number of additional T cells in a vaccine represented by  $V(t)$ . Note that we relabel the constant  $d$  from their paper as  $d_0$  to avoid confusion with the differential operator. Wilson and Levy's model

[2012] is written as the following system of ordinary differential equations:

$$\frac{dT}{dt} = a_0T(1 - c_0T) - \delta_0 \frac{ET}{1 + c_1B} - \delta_0TV, \quad (1)$$

$$\frac{dB}{dt} = a_1 \frac{T^2}{c_2 + T^2} - d_0B, \quad (2)$$

$$\frac{dE}{dt} = \frac{fET}{1 + c_3TB} - rE - \delta_0RE - \delta_1E, \quad (3)$$

$$\frac{dR}{dt} = rE - \delta_1R, \quad (4)$$

$$\frac{dV}{dt} = g(t) - \delta_1V. \quad (5)$$

Equation (1) describes the growth rate of the tumor measured in  $\text{mm}^2$ . The tumor is assumed to grow logistically with a growth rate of  $a_0$  and a carrying capacity of  $1/c_0$ . The second term of (1) represents the rate at which the effector cells are able to destroy tumor cells. The term  $1 + c_1B$  represents the negative effect that TGF- $\beta$  production has on the effector cells' ability to attack the tumor cells. The last term represents the action of the vaccine on the tumor cells.

Equation (2) represents the rate of change in the concentration of TGF- $\beta$  measured in ng/ml. The switch in the amount of TGF- $\beta$  production between small and large tumors is modeled by the first term in (2). The constant  $c_2$  represents the tumor cell population at which the switch occurs and  $a_1$  is the maximum rate of TGF- $\beta$  production [Arciero et al. 2004]. The decay rate for TGF- $\beta$  is given by  $d_0$ .

Equation (3) represents the rate of change of the number of effector cells in the system. The first term represents the rate at which effector cells are recruited to attack the tumor. The expression  $1 + c_3TB$  represents the negative effect of both TGF- $\beta$  production and tumor growth on the effector cells' ability to proliferate. The constant  $f$  represents the tumor's antigenicity and it measures the degree that the tumor is able to stimulate an immune response. The number  $r$  represents the rate at which effector cells differentiate into regulatory T cells. The effector cells are also removed when interacting with regulatory T cells at a rate of  $\delta_0$ .

Equation (4) represents the number of regulatory T cells in the system. This model assumes that only CD8+ effector cells become regulatory T cells.

Equation (5) represents the rate of change of the vaccine, which is modeled as an addition of 5000 activated T cells at day 3. If the vaccine is given,  $g(t)$  is a constant multiple of a Dirac delta function centered at  $t = 3$ , i.e.,  $g(t) = g_0\delta(t - 3)$ , where  $g_0 = 5000$ . If the vaccine is withheld,  $g(t)$  is identically 0. Finally, the effector cells, regulatory T cells, and activated T cells in the vaccine are all assumed to share a natural death rate of  $\delta_1$ .

### 3. The modified model

We modify Wilson and Levy’s equations by modeling the rate of change of cancer stem cells and tumor cells separately in order to better understand how the proposed treatments affect each population. To highlight the role that TGF- $\beta$  plays in tumor growth and immunosuppression, we follow the example of [Arciero et al. 2004] and choose to consider two scenarios of tumor development, namely:

- passive tumors that do not produce TGF- $\beta$ ,
- aggressive tumors that produce TGF- $\beta$ .

**3.1. Passive tumor model.** In the passive tumor model, we follow the size of the cancer stem cell population represented by  $C(t)$ , the size of the tumorous cell population represented by  $T(t)$ , the number of effector cells represented by  $E(t)$ , and the number of T cells in the vaccine represented by  $V(t)$ . Our model is written as the following system of ordinary differential equations:

$$\frac{dC}{dt} = kC \left( 1 - \frac{C}{M_1} \right) - hEC - hCV, \tag{6}$$

$$\frac{dT}{dt} = kC \frac{C}{M_1} \left( 1 - \frac{T}{M_2} \right) - hET - hTV - d_1T, \tag{7}$$

$$\frac{dE}{dt} = fET - rE - d_3E, \tag{8}$$

$$\frac{dV}{dt} = g(t) - d_3V. \tag{9}$$

Note that  $G = 0$  in the passive tumor case, as these tumors do not produce TGF- $\beta$ . Equation (6) describes the growth rate of the cancer stem cells of the tumor, which are assumed to follow logistic growth with a growth rate of  $k$  and a carrying capacity of  $M_1$ . The term  $hEC$  represents the rate at which effector cells attack the  $C$  stem cells.

Equation (7) represents the growth rate of the tumor cells. The fraction of  $C$  stem cells that differentiate into  $T$  tumor cells is represented by  $C/M_1$ , and we assume that the tumor cells are nondividing. Hence, if  $C < M_1$ , then some of the stem cells will produce more stem cells, while other stem cells will produce tumor cells. If  $C = M_1$ , then all of the stem cells will produce tumor cells. This behavior reflects normal stem cell dynamics [Soltysova et al. 2005]. The carrying capacity of tumor cells is given by  $M_2$ , and we assume the tumor cells have a small natural death rate of  $d_1$ . The term  $hET$  represents the rate at which effector cells attack the tumor cells. We assume that the effector cells are able to attack the  $C$  and  $T$  cells at the same rate. Similarly, the terms  $hCV$  and  $hTV$  represent the detrimental effect that

the vaccine has on both the  $C$  and  $T$  cells, and we assume that the vaccine is equally effective against  $C$  and  $T$  cells.

Equations (8) and (9) model the effector cells and vaccine and follow directly from equations (3) and (5), where we have ignored the contributions of regulatory T cells as they only slightly increase the rate of decay of the effector cells and thus do not greatly affect the dynamics of the model. We relabel the death rate of the effector cells and vaccine as  $d_3$ .

**3.2. Aggressive tumor model.** Our model of aggressive tumors is represented by the following system of ordinary differential equations:

$$\frac{dC}{dt} = kC \left( 1 - \frac{C}{M_1} \right) - h \frac{EC}{1 + c_1 B} - hCV, \quad (10)$$

$$\frac{dT}{dt} = kC \frac{C}{M_1} \left( 1 - \frac{T}{M_2} \right) - h \frac{ET}{1 + c_1 B} - hTV - d_1 T, \quad (11)$$

$$\frac{dB}{dt} = a \frac{C^2}{c_2 + C^2} - d_2 B, \quad (12)$$

$$\frac{dE}{dt} = \frac{fET}{1 + c_3 TB} - rE - d_3 E, \quad (13)$$

$$\frac{dV}{dt} = g(t) - d_3 V. \quad (14)$$

Equations (10), (11), (13) and (14) follow directly from the passive tumor model, with the corresponding adjustments made to the interaction terms involving the effector cells  $E$  in accordance with the Wilson–Levy model. Equation (12) represents the rate at which TGF- $\beta$  is produced by the tumor. We assume that TGF- $\beta$  is only produced by cancer stem cells. There is a growing body of medical evidence that shows the link between TGF- $\beta$  production and cancer stem cells [Dreesen and Brivanlou 2007; Tang et al. 2008; Mishra et al. 2005].

#### 4. Simulations

In order to better understand the behavior of our models, we perform numerical simulations using Mathematica 9's `NDSolve` command. The code used will be made available upon request. All of our simulations are measured in days. We simulate the growth of four types of tumors, namely:

- (1) low antigenicity passive tumors (LAPTs),
- (2) high antigenicity passive tumors (HAPTs),
- (3) low antigenicity aggressive tumors (LAATs),
- (4) high antigenicity aggressive tumors (HAATs).



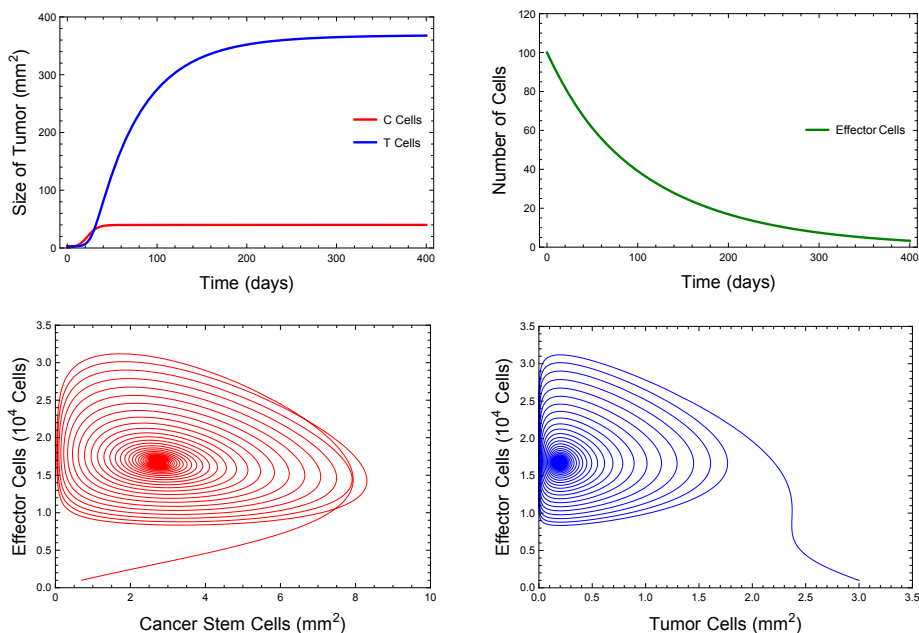
	value	units	description
$k$	0.18	days <sup>-1</sup>	tumor growth rate
$h$	10 <sup>-5</sup>	# <sup>-1</sup> days <sup>-1</sup>	vaccine/effector cell-induced tumor death rate
$M_1$	40	mm <sup>2</sup>	cancer stem cell carrying capacity
$M_2$	369	mm <sup>2</sup>	tumor cell carrying capacity
$d_1$	10 <sup>-9</sup>	days <sup>-1</sup>	death rate of tumor cells
$f$	low: 5 · 10 <sup>-6</sup> high: 0.05	(mm <sup>2</sup> ) <sup>-1</sup> days <sup>-1</sup>	tumor antigenicity
$r$	0.01	days <sup>-1</sup>	effector cell removal rate to regulatory T cells
$d_3$	10 <sup>-5</sup>	days <sup>-1</sup>	vaccine/effector cell death rate
$g_0$	5000	# days <sup>-1</sup>	additional T cells provided by vaccine

**Table 1.** Parameters for passive tumor model.

**4.1. Simulation of the passive tumor model.** Table 1 lists the values of the parameters used in the passive tumor model. All parameter values are taken from Wilson and Levy with the exception of  $f$  and  $k$ , which are taken from Kirschner and Panetta, and  $M_1$  and  $d_1$ , which are estimated based on the expected low ratio of cancer stem cells to tumor cells and slow natural death rate of tumor cells. A parameter sensitivity analysis is conducted in Section 7 to assess sensitivity of the model to these parameter values. Following the example of Wilson and Levy, we assume that there are 100 effector T cells present at the initial time point in all cases except for the high antigenicity passive tumors, in which we assume that there are 1000 effector T cells present (see discussion below). We choose 0.7 and 3 mm<sup>2</sup> as our initial stem cell and tumor cell sizes, respectively. The simulations for passive tumor growth with no treatment for both low and high antigenicities are presented in Figure 1.

In Figure 1, the graphs in the top row show the low antigenicity of the tumor does not prompt a response from the effector cells, and thus both the  $C$  stem cells and  $T$  tumor cells grow to their respective carrying capacities while the number of  $E$  effector cells at the tumor site decays over time. In contrast, the graphs in the bottom row of Figure 1 show the behavior of a passive tumor with high antigenicity. In this case, the effector cells undergo an oscillatory response and begin to restrict the tumor’s growth. There is biological evidence to support these oscillations in cancers such as chronic myeloid leukemia [Kirschner and Panetta 1998]. While both the  $C$  and  $T$  cell populations continue to persist, the effector cells reduce the steady-state population size of each cancer cell type to very minute levels.

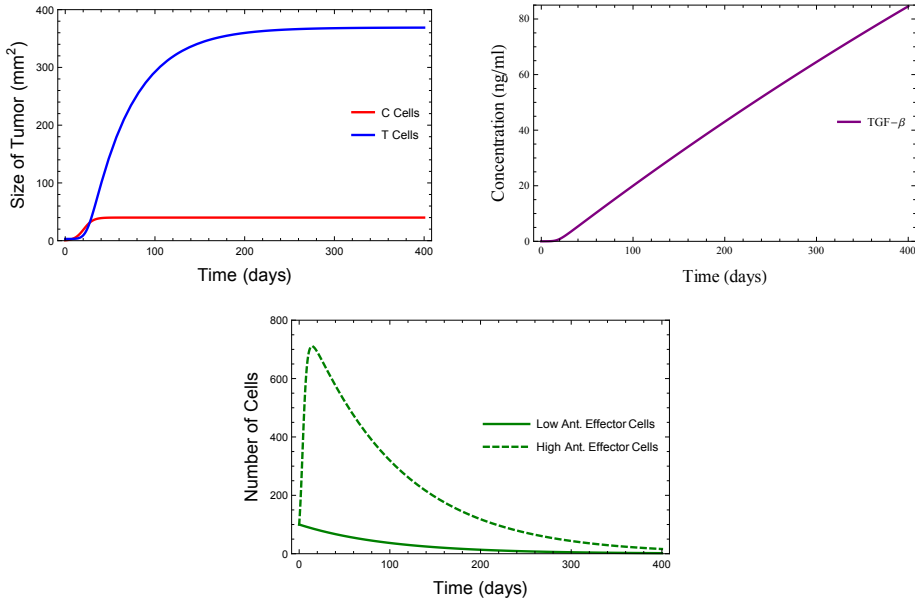
**4.2. Simulation of the aggressive tumor model.** We next present the behavior of our aggressive tumor model for both low and high antigenicities. In Table 2, we introduce the new parameter values used in the aggressive tumor model. As before, all



**Figure 1.** Passive tumor simulations. The graphs in the top row model a tumor with low antigenicity and those in the bottom row model one with high antigenicity.

	value	units	description
$k$	0.1946	days <sup>-1</sup>	tumor growth rate
$M_1$	40	mm <sup>2</sup>	cancer stem cell carrying capacity
$M_2$	369	mm <sup>2</sup>	tumor cell carrying capacity
$h$	10 <sup>-5</sup>	# <sup>-1</sup> days <sup>-1</sup>	vaccine/effector cell-induced tumor death rate
$c_1$	100	ml/ng	TGF- $\beta$ inhibition of effector cell-induced tumor death
$d_1$	10 <sup>-9</sup>	days <sup>-1</sup>	death rate of tumor cells
$a$	0.3	days <sup>-1</sup> ng/ml	maximum rate of TGF- $\beta$ production
$c_2$	300	(mm <sup>2</sup> ) <sup>2</sup>	steepness coefficient of TGF- $\beta$ production
$d_2$	7 · 10 <sup>-4</sup>	days <sup>-1</sup>	rate of degradation of TGF- $\beta$
$f$	low: 5 · 10 <sup>-6</sup> high: 0.62	(mm <sup>2</sup> ) <sup>-1</sup> days <sup>-1</sup>	tumor antigenicity
$c_3$	300	ml/(ng mm <sup>2</sup> )	tumor cell and TGF- $\beta$ inhibition of effector cell activation
$r$	0.01	# <sup>-1</sup>	effector cell removal rate to regulatory T cells
$d_3$	10 <sup>-5</sup>	days <sup>-1</sup>	vaccine/effector cell death rate
$g_0$	5000	# days <sup>-1</sup>	additional T cells provided by vaccine

**Table 2.** Parameters for aggressive tumor model.



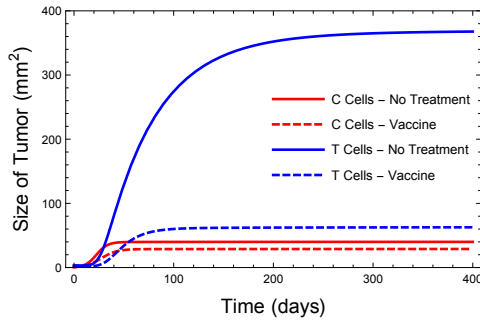
**Figure 2.** Aggressive tumor simulations.

values are taken from Wilson and Levy (including the slightly higher tumor growth rate  $k$  and antigenicity  $f$ ) with the exception of  $M_1$  and  $d_1$ , which are estimated as stated above. The initial conditions for the cancer stem cell, tumor cell, and effector cell populations remain unchanged, and we use 0.0035 ng/ml as the initial concentration of TGF- $\beta$  produced by the tumor. Figure 2 shows the results of our simulations for the aggressive tumor model with no treatment. In Figure 2, top left, the  $C$  cancer stem cells and  $T$  tumor cells grow uninterrupted to their carrying capacities for both low and high antigenicity levels. Similarly, Figure 2, top right, shows the concentration of TGF- $\beta$  produced by the  $C$  stem cells steadily increases regardless of antigenicity level. The only discernible difference with respect to antigenicity occurs with the effector cell population in Figure 2, bottom, where an initial spike in the number of effector cells is seen in the high antigenicity case. However, due to the inhibitory effect of TGF- $\beta$  on the effector cell population, this increase is short-lasting and the effector cell population decays over time, failing to halt tumor progression.

### 5. Treatment

Following the example of [Wilson and Levy 2012], we divide treatment into three cases in order to test their relative effectiveness on both the  $C$  and  $T$  cells, namely:

- vaccine treatment,
- TGF- $\beta$  inhibition,
- combination treatment.



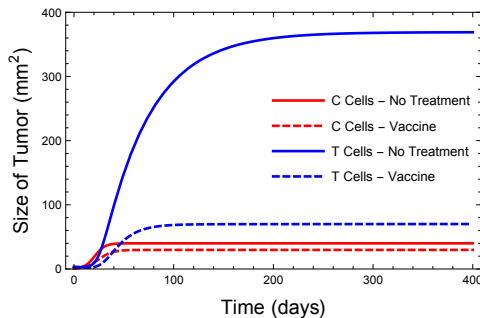
**Figure 3.** Vaccine treatment for LAPTs.

**5.1. Treatment of passive tumors.** Since passive tumors do not produce  $\text{TGF-}\beta$ , we consider only vaccine treatment, which is modeled by the introduction of 5000 effector cells on day 3 of simulation. Figure 3 shows the results of the vaccine treatment for a low antigenicity passive tumor.

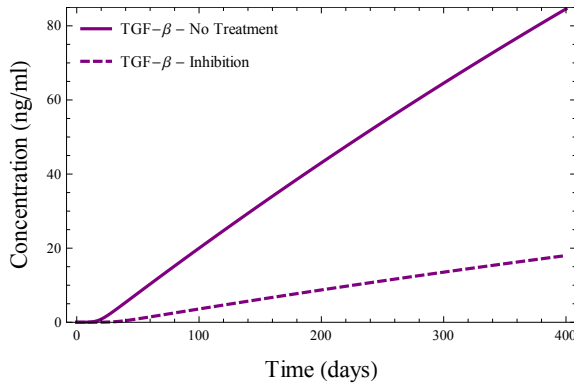
The vaccine treatment is successful in reducing the final steady states of both the cancer stem cells and tumor cells, but cannot clear the tumor entirely. The evolution of the effector cell population is unaffected by the vaccine treatment, and the effector cells decay as in Figure 1, top right. For high antigenicity passive tumors, the vaccine treatment produces no noticeable difference in either the  $C$ ,  $T$ , or  $E$  cell dynamics over time, leading to simulations identical to those found in the graphs in the bottom row of Figure 1. The large oscillatory response of the effector cells dominates any contribution from the vaccine in diminishing the cancer cell populations.

**5.2. Treatment of aggressive tumors.** With the inclusion of  $\text{TGF-}\beta$  production by cancer stem cells in aggressive tumors, we now have all three treatment options to consider.

**5.2.1. Vaccine treatment.** We begin by repeating the vaccine treatment simulation for low and high antigenicity aggressive tumors, shown in Figure 4.



**Figure 4.** Vaccine treatment for LAATs and HAATs.



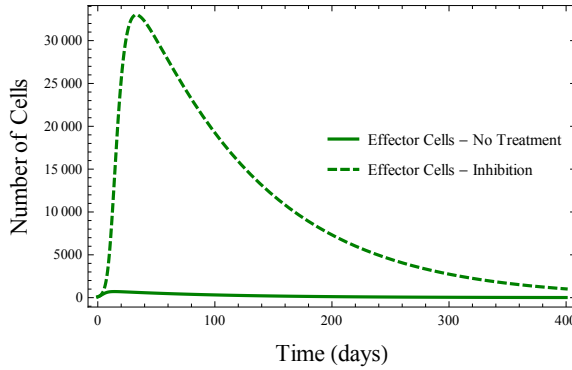
**Figure 5.** TGF- $\beta$  Inhibition for LAATs and HAATs.

For both antigenicity levels, the steady states of the cancer stem cells and tumor cells are diminished by the vaccine, similar to the vaccine treatment of the low antigenicity passive tumor. No appreciable difference is observed in the effector cells or TGF- $\beta$  concentration of the aggressive tumor model under the vaccine. For a more accurate comparison between the effects of treatment in Figures 3 and 4, the size of the  $C$  cell population at day 400 for the low antigenicity passive tumor is approximately  $28.9 \text{ mm}^2$  and for the low/high antigenicity aggressive tumor is approximately  $29.7 \text{ mm}^2$ . Similarly, the size of the  $T$  cell population at day 400 is  $62.7 \text{ mm}^2$  for the low antigenicity passive tumor and  $70.1 \text{ mm}^2$  for the low/high antigenicity aggressive tumor.

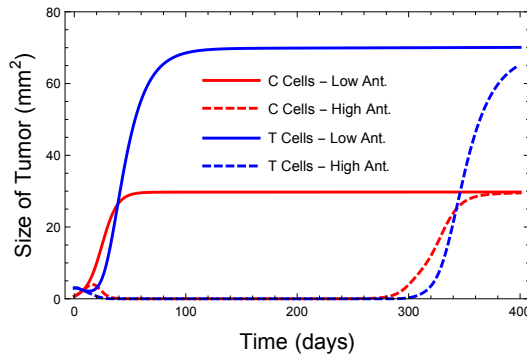
**5.2.2. TGF- $\beta$  inhibition.** As in [Wilson and Levy 2012], TGF- $\beta$  inhibition is modeled as an increase of  $c_2$  from 300 to 7000. For both low and high antigenicity tumors, TGF- $\beta$  inhibition has a nearly negligible effect on the final tumor size, with the  $C$  and  $T$  cells growing to their carrying capacities as in Figure 2, top left. While the treatment succeeds in slowing down TGF- $\beta$  production by the tumor (see Figure 5), it fails to lead to any measurable reduction in cancer growth.

As a final note, while the effector cells continue to decay normally for low antigenicity tumors (as in Figure 2, bottom), the TGF- $\beta$  inhibition induces a large initial response of the effector cells for high antigenicity tumors (see Figure 6). Nevertheless, the effector cells have little impact on tumor growth in this case.

**5.2.3. Combination treatment.** Again following [Wilson and Levy 2012], combination treatment is modeled by both the increase in  $c_2$  from 300 to 7000 and the administration of the vaccine. For low antigenicity tumors, the combination treatment reduces the  $C$  and  $T$  steady-state populations to the same levels as the vaccine treatment alone, with no noticeable benefit from adding the TGF- $\beta$  inhibition. For high antigenicity tumors, on the other hand, the combination treatment is highly effective, reducing both the  $C$  and  $T$  populations to nearly zero by approximately



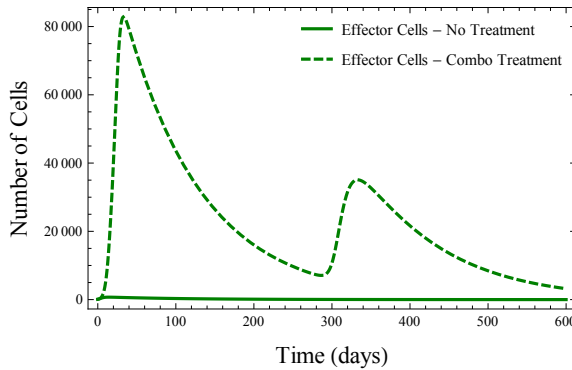
**Figure 6.** Effector cells under TGF- $\beta$  inhibition for HAATs.



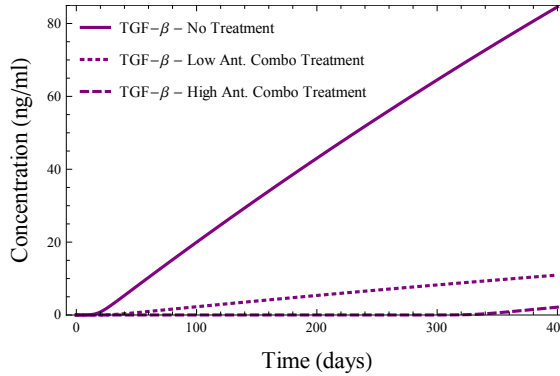
**Figure 7.** Combination treatment for LAATs and HAATs.

day 30. However, the remission is temporary and both cancer cell populations start growing again shortly before day 300 (see Figure 7).

In Figure 8, the effector cell population also displays new behavior under the combination treatment for high antigenicity tumors. The treatment produces a



**Figure 8.** Effector cells under combination treatment for HAATs.



**Figure 9.** TGF-β concentration under combination treatment for LAATs and HAATs.

significant initial spike in effector cells, helping send the cancer into its temporary remission. Once the cancer begins to recur around day 300, the effector cells produce a second smaller response that is unable to slow the cancer’s growth. Finally, the suppression of TGF-β production by the *C* cancer stem cells under combination treatment is shown in Figure 9. For high antigenicity tumors in particular, the concentration of TGF-β is greatly diminished for the first year of simulated time.

**5.2.4. Summary of simulations.** Table 3 provides a summary of the behavior of the treatment outcomes on all four types of tumors. In general, our simulations reveal

	no treatment	vaccine	TGF-β inhibition	combination
LAPT	<i>C, T</i> cells grow to CC; no <i>E</i> cell response	$(C, T) = (72\%, 17\%)$ of CC at end	not applicable	not applicable
HAPT	<i>C, T</i> cells reduced to minute levels; <i>E</i> cells produce large oscillatory response		not applicable	not applicable
LAAT	<i>C, T</i> cells grow to CC; no <i>E</i> cell response	$(C, T) = (74\%, 19\%)$ of CC at end	<i>C, T</i> cells grow to CC; no <i>E</i> cell response	$(C, T) = (74\%, 19\%)$ of CC at end
HAAT	<i>C, T</i> cells grow to CC; small initial <i>E</i> cell response	$(C, T) = (74\%, 19\%)$ of CC at end	<i>C, T</i> cells grow to CC; large initial <i>E</i> cell response	<i>C, T</i> cells reduced to nearly 0; recurrence by day 300; very large initial <i>E</i> cell response with secondary response when cancer recurs

**Table 3.** Summary of treatment outcomes for four types of tumors.

that the vaccine treatment is overall more effective than TGF- $\beta$  inhibition at combating cancer growth. The cancer stem cells appear more resilient to the additional effector cells provided by the vaccine, with a smaller reduction in carrying capacity when compared with the tumor cells. Also, the largest effect of TGF- $\beta$  inhibition is seen when combined with the vaccine against high antigenicity aggressive tumors, sending the cancer into remission for an extended period of time.

## 6. Stability analysis

**6.1. Dimensionless models.** To reduce the number of parameters in the model and ease calculations, we follow the example of [Kirschner and Panetta 1998; Arciero et al. 2004] and nondimensionalize our equations using the following scaling:

$$\begin{aligned} x &= \frac{C}{M_1}, & y &= \frac{T}{M_2}, & z &= c_1 B, & w &= \frac{hE}{r}, & v &= \frac{d_3 V}{g_0}, \\ \tau &= kt, & \rho &= \frac{r}{k}, & \eta &= \frac{hg_0}{kd_3}, & \mu &= \frac{M_1}{M_2}, & \alpha &= \frac{ac_1}{k}, \\ \beta &= \frac{c_2}{M_1^2}, & \gamma &= \frac{M_2 f}{k}, & \sigma &= \frac{M_2 c_3}{c_1}, & \delta_1 &= \frac{d_1}{k}, & \delta_2 &= \frac{d_2}{k}, & \delta_3 &= \frac{d_3}{k}. \end{aligned}$$

This results in the following scaled system of differential equations for the passive tumor model:

$$\frac{dx}{d\tau} = x(1-x) - \rho wx - \eta xv, \quad (15)$$

$$\frac{dy}{d\tau} = \mu x^2(1-y) - \rho wy - \eta yv - \delta_1 y, \quad (16)$$

$$\frac{dw}{d\tau} = \gamma wy - \rho w - \delta_3 w, \quad (17)$$

$$\frac{dv}{d\tau} = \delta_3 \delta (\tau - 3k) - \delta_3 v. \quad (18)$$

Similarly, the scaled aggressive tumor model is given by:

$$\frac{dx}{d\tau} = x(1-x) - \rho \frac{wx}{1+z} - \eta xv, \quad (19)$$

$$\frac{dy}{d\tau} = \mu x^2(1-y) - \rho \frac{wy}{1+z} - \eta yv - \delta_1 y, \quad (20)$$

$$\frac{dz}{d\tau} = \frac{\alpha x^2}{\beta + x^2} - \delta_2 z, \quad (21)$$

$$\frac{dw}{d\tau} = \gamma \frac{wy}{1 + \sigma yz} - \rho w - \delta_3 w, \quad (22)$$

$$\frac{dv}{d\tau} = \delta_3 \delta (\tau - 3k) - \delta_3 v. \quad (23)$$



**6.2. Stability of passive tumor model.** In order to assess the stability of the passive tumor model, we first find equilibrium solutions by setting (15)–(18) equal to 0 and solving the resulting nonlinear algebraic system of equations. Note that the long-term behavior of the vaccine is clearly exponential decay to zero, so we set  $v = 0$  for the remainder of the analysis to simplify the calculation of the other equilibrium populations. Mathematica 9 produces five equilibrium points for the remaining  $x$ ,  $y$ , and  $w$  populations, one of which contains a negative component and is thus not biologically meaningful, and two interior equilibrium points whose closed form is too complex to analyze. The other two equilibria that we are able to study are

$$P_1 : (x, y, w) = (0, 0, 0),$$

$$P_2 : (x, y, w) = \left(1, \frac{\mu}{\delta_1 + \mu}, 0\right).$$

Next, we calculate the Jacobian matrix with  $v = 0$ :

$$\begin{pmatrix} \partial f/\partial x & \partial f/\partial y & \partial f/\partial w \\ \partial g/\partial x & \partial g/\partial y & \partial g/\partial w \\ \partial h/\partial x & \partial h/\partial y & \partial h/\partial w \end{pmatrix},$$

where

$$f(x, y, w) = x(1 - x) - \rho wx,$$

$$g(x, y, w) = \mu x^2(1 - y) - \rho wy - \delta_1 y,$$

$$h(x, y, w) = \gamma wy - \rho w - \delta_3 w.$$

By substituting each equilibrium point into the above Jacobian, a quick calculation of the eigenvalues of the resulting matrix reveals that the origin  $P_1$  is always unstable, while the second equilibrium point  $P_2$  is stable if and only if

$$\rho + \delta_3 > \frac{\gamma\mu}{\delta_1 + \mu}. \tag{24}$$

Biologically, this inequality indicates that if the removal rate of the effector cells is too high relative to the tumor’s antigenicity and size, then the effector cells will provide an insufficient response to halt tumor growth and will eventually decay to zero. Testing the parameters for the passive tumor model in Table 1, we find that for low antigenicity passive tumors,

$$\rho + \delta_3 = 0.0556, \quad \frac{\gamma\mu}{\delta_1 + \mu} = 0.01025.$$

Hence inequality (24) is satisfied and  $P_2$  is stable, supporting the behavior observed in Figure 1, top row. On the other hand, for high antigenicity passive tumors we have

$$\rho + \delta_3 = 0.0556, \quad \frac{\gamma\mu}{\delta_1 + \mu} = 102.5.$$

Thus  $P_2$  is unstable in this case. To further investigate the long-term behavior of high antigenicity passive tumors, we may substitute the parameters from Table 1 into the symbolically intractable interior equilibrium points. We find that there is indeed a third positive equilibrium point  $P_3$ , namely  $(x, y, w) = (0.0683, 0.00054, 16.7705)$ , corresponding to steady-state populations of  $C = 2.732$ ,  $T = .2002$ , and  $E = 16770.5$ . Additionally, all three eigenvalues of the Jacobian matrix for this equilibrium have negative real part, two of which come in a complex conjugate pair. Hence  $P_3$  is stable, and the complex-valued eigenvalues provide evidence for the oscillatory behavior seen in Figure 1, bottom row.

**6.3. Stability of aggressive tumor model.** Following the procedure of the previous section, we set (19)–(23) equal to 0 to search for steady-state solutions of the aggressive tumor model. Mathematica 9 returns seven equilibrium solutions, but due to the highly nonlinear nature of the model, again only two permit a local stability analysis:

$$A_1 : (x, y, z, w) = (0, 0, 0, 0),$$

$$A_2 : (x, y, z, w) = \left(1, \frac{\mu}{\delta_1 + \mu}, \frac{\alpha}{(1 + \beta)\delta_2}, 0\right).$$

The Jacobian matrix of the system with  $v = 0$  now has the form

$$\begin{pmatrix} \partial f/\partial x & \partial f/\partial y & \partial f/\partial z & \partial f/\partial w \\ \partial g/\partial x & \partial g/\partial y & \partial g/\partial z & \partial g/\partial w \\ \partial h/\partial x & \partial h/\partial y & \partial h/\partial z & \partial h/\partial w \\ \partial j/\partial x & \partial j/\partial y & \partial j/\partial z & \partial j/\partial w \end{pmatrix},$$

where

$$f(x, y, z, w) = x(1 - x) - \rho \frac{wx}{1 + z},$$

$$g(x, y, z, w) = \mu x^2(1 - y) - \rho \frac{wy}{1 + z} - \delta_1 y,$$

$$h(w, x, y, z) = \frac{\alpha x^2}{\beta + x^2} - \delta_2 z,$$

$$j(x, y, z, w) = \gamma \frac{wy}{1 + \sigma yz} - \rho w - \delta_3 w.$$

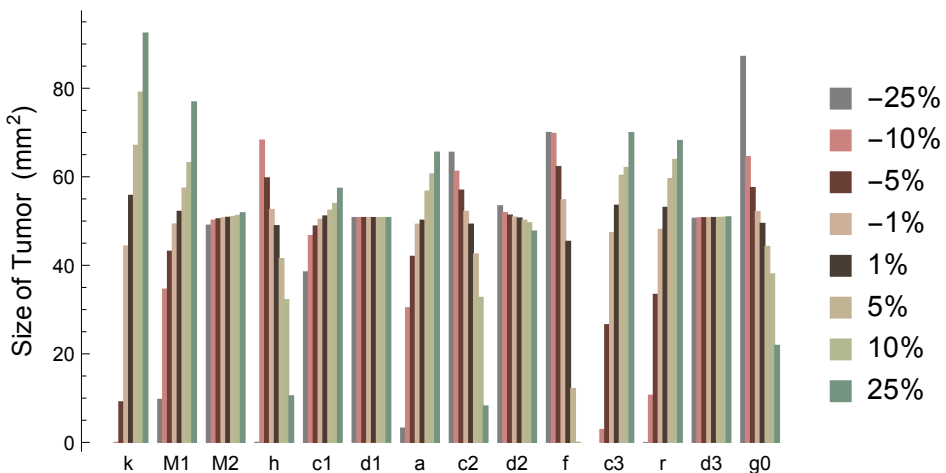
By calculating the eigenvalues of the Jacobian at each equilibrium point, it is easily seen that the origin  $A_1$  is again unstable, while the second equilibrium point  $A_2$  is stable if and only if

$$\rho + \delta_3 > \frac{(1 + \beta)\gamma\mu\delta_2}{\alpha\mu\sigma + (1 + \beta)(\delta_1 + \mu)\delta_2}. \tag{25}$$

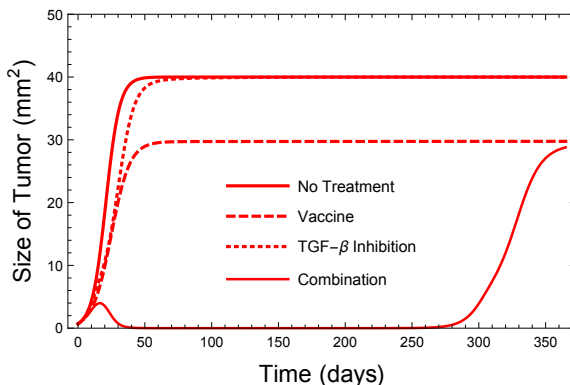
This inequality establishes a threshold for the removal rate of effector cells in terms of tumor antigenicity, size, and TGF- $\beta$  production that, if exceeded, results in exponential decay of the effector cells and growth of the cancer stem cell and tumor cell populations to their carrying capacities. Using the parameters found in Table 2 for the aggressive tumor model, a quick calculation as before reveals that inequality (25) is satisfied for both low and high antigenicity aggressive tumors. Hence  $A_2$  is stable in both cases, matching our earlier observations in Figure 2.

### 7. Sensitivity analysis

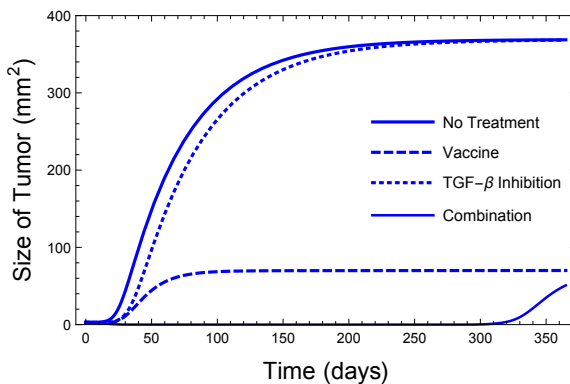
In order to assess the sensitivity of our model to changes in parameters, we conduct a sensitivity analysis for combination treatment of high antigenicity aggressive tumors. More specifically, we vary each parameter over a range of percentages centered around a baseline for 365 simulated days while leaving all other parameters fixed and observe the effects on the resulting  $T$  tumor cell population. The results are presented in Figure 10. In contrast with the findings in [Wilson and Levy 2012], while the antigenicity  $f$  ranked high among the most sensitive parameters, we find that there are three more sensitive parameters: the cancer growth rate  $k$ , the initial injection of T cells by the vaccine  $g_0$ , and the carrying capacity of the cancer stem cells  $M_1$ . It will thus be crucial to obtain highly accurate biological estimates for these parameters to increase the applicability of the model.



**Figure 10.** Sensitivity analysis for HAATs. Baseline values:  $k = 0.1946$ ,  $M_1 = 40$ ,  $M_2 = 369$ ,  $h = 10^{-5}$ ,  $c_1 = 100$ ,  $d_1 = 10^{-9}$ ,  $a = 0.3$ ,  $c_2 = 7000$ ,  $d_2 = 7 \cdot 10^{-4}$ ,  $f = 0.62$ ,  $c_3 = 300$ ,  $r = 0.01$ ,  $d_3 = 10^{-5}$ ,  $g_0 = 5000$ .



**Figure 11.** Response of cancer stem cells to treatment.

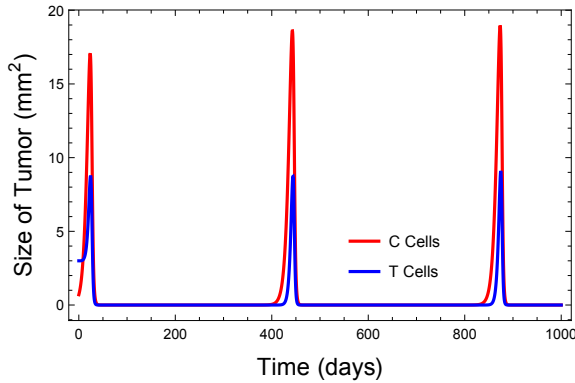


**Figure 12.** Response of tumor cells to treatment.

## 8. Results

Figures 11 and 12 show the relative effectiveness of the vaccine, TGF- $\beta$  inhibition, and combination treatments against the  $C$  and  $T$  cell populations of a high antigenicity aggressive tumor, respectively.

Although TGF- $\beta$  inhibition moderately slows down tumor growth in both cases, the  $C$  stem cells are able to reach their carrying capacity by approximately day 60, while the  $T$  tumor cells reach their carrying capacity by day 250. Alternatively, in the vaccine treatment case, the vaccine is able to reduce the tumor cell population from its carrying capacity of  $369 \text{ mm}^2$  to  $70.1 \text{ mm}^2$  (a reduction of 81%), while it is only able to reduce the stem cell population from its carrying capacity of  $40 \text{ mm}^2$  to  $29.7 \text{ mm}^2$  (a reduction of 16%). Although remission is achieved in our simulations of combination treatment, from Figure 11 we can see that the stem cell population is not completely destroyed and as a result, the cancer stem cells reemerge by day 250 and prompt renewed growth of the tumor cells by day 300. Our results agree with



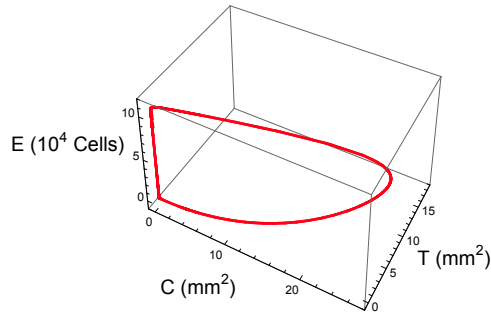
**Figure 13.** Cancer cells approaching limit cycle for HAPT.

studies that show that unless cancer treatment is specifically directed toward cancer stem cells, the cancer can still recur, even if there is a significant reduction in tumor size after treatment [Jordan et al. 2006].

Conversely, in our simulation of low antigenicity aggressive tumors we show that although combination treatment succeeds in reducing the size of the tumor, it is unable to eliminate either the *C* or *T* cell populations. Furthermore, in our simulations of treatment of passive tumors, we find that the vaccine produces a similar outcome for low antigenicity tumors. The effector cell response for high antigenicity passive tumors is sufficient to significantly reduce final tumor size, and the vaccine treatment produces no noticeable benefit for this type of tumor.

The oscillatory behavior seen in our passive tumor model deserves further mention. In [Kirschner and Panetta 1998], the authors find that in the no treatment case, as they increase tumor antigenicity, the long-term dynamics of their model transition from a stable node to a stable limit cycle to a stable spiral. It is interesting to observe that in our model, the progression of these dynamics as the antigenicity increases occurs in a somewhat different manner. Keeping all other values in Table 1 fixed, for  $f < 2.71 \cdot 10^{-5}$ , the equilibrium point  $P_2$  is a stable node, as in Figure 1, top row. Biologically speaking, this implies that extremely low antigenicity tumors are able to effectively escape immunosurveillance and grow to carrying capacity. For  $2.71 \cdot 10^{-5} < f < 1.13 \cdot 10^{-4}$ ,  $P_2$  becomes unstable and one of the interior equilibria becomes a stable node. Next, for  $1.14 \cdot 10^{-4} < f < 0.0856$ , the positive interior equilibrium transitions to a stable spiral. Thus all cell populations begin to oscillate, and the effector cells reduce the size of the tumor to nearly zero before the oscillations eventually dampen out. Finally, for  $f > 0.0857$ , the interior equilibrium becomes an unstable spiral and the cell populations oscillate without bound.

Moreover, for  $f > 0.0445$ , a stable limit cycle is created. Thus for the high antigenicity value used in our passive tumor model,  $f = 0.05$ , the long-term



**Figure 14.** Limit cycle in phase space for HAPT.

dynamics either result in damped or sustained oscillations, depending on the initial conditions. For example, if we let  $f = 0.05$  and the initial population of effector cells satisfy  $E(0) = 100$ , then the cell populations indeed approach the stable limit cycle. A plot of the cancer cell populations in this case is shown in Figure 13, and the limit cycle in phase space is presented in Figure 14.

However, if  $E(0) = 1000$ , we find that the populations approach the interior stable spiral. This behavior was demonstrated in Figure 1, bottom row, for our high antigenicity passive tumors.

## 9. Discussion

The mathematical model presented in this paper describes the dynamics of cancer stem cells, tumor cells, and effector cells under one or more treatment protocols designed to elicit a larger than normal response from the body's natural immune system. The antigenicity of the tumor as well as the aggressiveness of the tumor via TGF- $\beta$  production play a crucial role in predicting the success of such techniques. We find that a vaccine delivering additional effector cells is able to diminish the size of highly antigenic tumors, and pairing the vaccine with a TGF- $\beta$  inhibitor can lead to at least temporary clearance of aggressive tumors. As expected, low antigenic tumors are able to better evade immunosurveillance and persist in the face of immunotherapy techniques, with aggressive tumors of this type being particularly resistant to treatment. For these tumors, other treatment options such as chemotherapy and radiation therapy should be explored.

Qualitatively, the behavior of our model for high antigenicity aggressive tumors agrees with the results of the Wilson–Levy model, with remission only being achieved after combination treatment. However, our model is additionally able to show how each of the various treatments affect the cancer stem cell and tumor cell populations individually. We show that the cancer stem cells are more resistant to the vaccine and experience a smaller reduction in carrying capacity when compared to the tumor cells. In addition the re-emergence of the tumor in all cases of

treatment of high antigenicity aggressive tumors also agrees with the stability analysis presented by Wilson and Levy [2012], which predicts that all treatment scenarios will eventually lead to a nonzero tumor equilibrium.

Furthermore, our simulations of passive tumors agree strongly with the results of Arciero et al., with low antigenic tumors escaping the immune response and growing to carrying capacity, while increasing antigenicity leads to damped oscillations that stabilize into a small persistent tumor. The behavior of aggressive tumors with low and high antigenicity in both models is also similar. The Arciero et al. model [2004] simulates siRNA treatment designed to suppress TGF- $\beta$  expression in tumor cells, and as with our TGF- $\beta$  inhibition strategy, they find that such a strategy alone is insufficient to clear aggressive tumors.

Future research will include further study of the global behavior of the model, including stability analysis of internal equilibria and identification of the basins of attraction for various equilibria and limit cycles. The parameters of the model should additionally be fit to experimental data to obtain a more biologically realistic time-scale for the dynamics predicted by the model. Lastly, the model suggests that inclusion of treatment methods that specifically target cancer stem cells could potentially lead to tumor clearance, even for aggressive low antigenic tumors. This possibility warrants further research by both mathematicians and biologists alike.

## References

- [Akhurst and Derynck 2001] R. J. Akhurst and R. Derynck, "TGF- $\beta$  signaling in cancer: a double-edged sword", *Trends Cell Bio.* **11**:11 (2001), S44–S51.
- [Arciero et al. 2004] J. C. Arciero, T. L. Jackson, and D. E. Kirschner, "A mathematical model of tumor-immune evasion and siRNA treatment", *Discrete Contin. Dyn. Syst. Ser. B* **4**:1 (2004), 39–58. MR Zbl
- [Cripe et al. 2009] T. P. Cripe, P.-Y. Wang, P. Marcato, Y. Y. Mahller, and P. W. K. Lee, "Targeting cancer-initiating cells with oncolytic viruses", *Molecular Therapy* **17**:10 (2009), 1677–1682.
- [Deonarain et al. 2009] M. P. Deonarain, C. A. Kousparou, and A. A. Epenetos, "Antibodies targeting cancer stem cells: a new paradigm in immunotherapy?", *Mabs* **1**:1 (2009), 12–25.
- [Dreesen and Brivanlou 2007] O. Dreesen and A. H. Brivanlou, "Signaling pathways in cancer and embryonic stem cells", *Stem Cell Rev.* **3**:1 (2007), 7–17.
- [Huntly and Gilliland 2005] B. J. P. Huntly and D. G. Gilliland, "Leukaemia stem cells and the evolution of cancer-stem-cell research", *Nat. Rev. Cancer* **5**:4 (2005), 311–321.
- [Jordan et al. 2006] C. T. Jordan, M. L. Guzman, and M. Noble, "Cancer stem cells", *New Eng. J. Med.* **355**:12 (2006), 1253–1261.
- [Joshi et al. 2009] B. Joshi, X. Wang, S. Banerjee, H. Tian, A. Matzavinos, and M. A. J. Chaplain, "On immunotherapies and cancer vaccination protocols: a mathematical modelling approach", *J. Theor. Biol.* **259**:4 (2009), 820–827. MR
- [Kirschner and Panetta 1998] D. Kirschner and J. C. Panetta, "Modeling immunotherapy of the tumor-immune interaction", *J. Math. Biol.* **37**:3 (1998), 235–252. Zbl

- [Li and Neaves 2006] L. Li and W. B. Neaves, “Normal stem cells and cancer stem cells: the niche matters”, *Cancer Res.* **66**:9 (2006), 4553–4557.
- [Malmberg 2004] K.-J. Malmberg, “Effective immunotherapy against cancer”, *Cancer Immunology, Immunotherapy* **53**:10 (2004), 879–892.
- [Mishra et al. 2005] L. Mishra, K. Shetty, Y. Tang, A. Stuart, and S. W. Byers, “The role of TGF- $\beta$  and Wnt signaling in gastrointestinal stem cells and cancer”, *Oncogene* **24**:37 (2005), 5775–5789.
- [Nani and Freedman 2000] F. Nani and H. I. Freedman, “A mathematical model of cancer treatment by immunotherapy”, *Math. Biosci.* **163**:2 (2000), 159–199. MR Zbl
- [Soltysova et al. 2005] A. Soltysova, V. Altanerova, and C. Altaner, “Cancer stem cells”, *Neoplasma* **52**:6 (2005), 435–440.
- [Stewart and Smyth 2011] T. J. Stewart and M. J. Smyth, “Improving cancer immunotherapy by targeting tumor-induced immune suppression”, *Cancer Metastasis Rev.* **30**:1 (2011), 125–140.
- [Tang et al. 2008] Y. Tang, K. Kitisin, W. Jogunoori, C. Li, C.-X. Deng, S. C. Mueller, H. W. Ransom, A. Rashid, A. R. He, J. S. Mendelson, J. M. Jessup, K. Shetty, M. Zasloff, B. Mishra, E. P. Reddy, L. Johnson, and L. Mishra, “Progenitor/stem cells give rise to liver cancer due to aberrant TGF- $\beta$  and IL-6 signaling”, *Proc. Nat. Acad. Sci.* **105**:7 (2008), 2445–2450.
- [Terabe et al. 2009] M. Terabe, E. Ambrosino, S. Takaku, J. J. O’Konek, D. Venzon, S. Lonning, J. M. McPherson, and J. A. Berzofsky, “Synergistic enhancement of CD8+ T cell-mediated tumor vaccine efficacy by an anti-transforming growth factor- $\beta$  monoclonal antibody”, *Clin. Cancer Res.* **15**:21 (2009), 6560–6569.
- [Wilson and Levy 2012] S. Wilson and D. Levy, “A mathematical model of the enhancement of tumor vaccine efficacy by immunotherapy”, *Bull. Math. Biol.* **74**:7 (2012), 1485–1500. MR Zbl

Received: 2014-09-02    Revised: 2016-04-21    Accepted: 2017-06-27

abernathy@winthrop.edu    *Department of Mathematics, Winthrop University,  
Rock Hill, SC, United States*

epelleg2@winthrop.edu    *Department of Mathematics, Winthrop University,  
Rock Hill, SC, United States*



# RNA, local moves on plane trees, and transpositions on tableaux

Laura Del Duca, Jennifer Tripp, Julianna Tymoczko and Judy Wang

(Communicated by Ann N. Trenk)

We define a collection of functions  $s_i$  on the set of plane trees (or standard Young tableaux). The functions are adapted from transpositions in the representation theory of the symmetric group and almost form a group action. They were motivated by *local moves* in combinatorial biology, which are maps that represent a certain unfolding and refolding of RNA strands. One main result of this study identifies a subset of local moves that we call  $s_i$ -local moves, and proves that  $s_i$ -local moves correspond to the maps  $s_i$  acting on standard Young tableaux. We also prove that the graph of  $s_i$ -local moves is a connected, graded poset with unique minimal and maximal elements. We then extend this discussion to functions  $s_i^C$  that mimic reflections in the Weyl group of type  $C$ . The corresponding graph is no longer connected, but we prove it has two connected components, one of symmetric plane trees and the other of asymmetric plane trees. We give open questions and possible biological interpretations.

## 1. Introduction

This paper analyzes a combinatorial question inspired by biology, specifically the mathematical structure of RNA. RNA has primary structure (a sequence of letters A, U, C, and G), secondary structure (a partial matching of the letters in the primary structure, indicating how the RNA strand has folded and bonded to itself), and a tertiary structure (how this folding occurs in 3-dimensional space). All of these structures contribute to the function of the RNA strand in ways that are still being uncovered. While our mathematical model of RNA is motivated by biology, this paper focuses on the model's combinatorial properties rather its direct relationship to biology.

---

*MSC2010:* 92E10, 05A05, 05C40.

*Keywords:* plane trees, RNA, Young tableaux, connected components, permutation.

The authors gratefully acknowledge helpful comments from the referee, and the National Science Foundation (DMS-1143716) and Smith College for their support of the Center for Women in Mathematics. Tymoczko was also partially supported by NSF grant DMS-1248171.

There are many combinatorial models for the secondary structure of RNA, including plane trees and standard Young tableaux of shape  $(n, n)$ . We will compare two important operations on these combinatorial objects, one from biological applications and the other from representation theory.

The first operation is called a *local move*. Defined by Condon, Heitsch, and Hoos (and in Definition 3.2), local moves model unfolding an RNA strand and refolding it differently [Heitsch 2006]. Heitsch [ibid.] described key combinatorial statistics of the graph whose vertices are plane trees on  $n$  edges and whose edges are local moves; she also showed how this graph is related to other important graphs, like an analogous graph whose vertices are noncrossing partitions.

The second operation comes from constructions of representations of the symmetric group  $S_n$ . One classical construction of representations of  $S_n$  uses Young diagrams, which are staircase-shaped collections of boxes. The symmetric group acts naturally on the set of all fillings of a Young diagram with the integers  $1, 2, \dots, n$  (without repeating numbers) just by permuting the numbers. It turns out that this action on filled Young diagrams gives rise to irreducible representations of  $S_n$ ; see, e.g., [Fulton 1997; Sagan 2001] for more.

We restrict our attention to “standard” Young tableaux, which are fillings that increase along both rows and columns. These tableaux are known to index bases for the irreducible representations of  $S_n$ , as well as other quantities of combinatorial interest. It is therefore natural to ask whether the symmetric group can be modified to also act on standard Young tableaux. The answer is yes and no. In Section 2 we define a collection of maps that act on standard Young tableaux and agree as much as possible with the action of the simple transpositions  $(i, i + 1)$  on arbitrary fillings of Young diagrams. More precisely, the map corresponding to the simple reflection  $(i, i + 1)$  simply exchanges  $i$  and  $i + 1$  in the tableau when doing so makes sense. The maps do not induce a group action of  $S_n$  because composition of functions does not agree with multiplication in  $S_n$ . Thus these maps cannot directly give information about  $S_n$ -representations. However, the maps are involutions, as we confirm in Proposition 2.4. Moreover, similar maps arise in other parts of combinatorial representation theory, including Vogan’s generalized tau invariants [Vogan 1979; Housley et al. 2015].

We further restrict our study to the standard Young tableaux corresponding to the partition  $(n, n)$ . This partition is an especially important one in applications from geometry [Fung 2003] to knot theory [Khovanov 2004], as well as the biological applications discussed here. In Theorem 3.6 we prove that our maps actually correspond to certain local moves, whose defining conditions are shown in Figure 3. We call the local moves that arise in this way  *$s_i$ -local moves*.

Note that not all local moves correspond to the action of permutations of the form  $(i, i + 1)$ . In particular the graph  $G^A$  whose vertices are plane trees and whose

edges correspond to  $s_i$ -local moves is different from the graph whose edges are *all* local moves. The graph of *all* local moves is a connected graded poset for which the cardinalities of the ranks form a symmetric, unimodal sequence; see, e.g., [Heitsch 2006]. Section 4 proves that the graph  $G^A$  is still a

- connected (Proposition 4.1),
- graded poset (Proposition 4.6),
- with a unique minimal element and a unique maximal element (Proposition 4.8).

However, the grading of the graph of  $s_i$ -local moves does not coincide with that of the graph of all local moves, nor does the graph of  $s_i$ -local moves satisfy the symmetry of ranks that the graph of local moves does (see Remark 4.7).

Our  $s_i$ -local moves were constructed by analogy with the symmetric group  $S_n$ . Thus we finish by extending the analogy to Weyl groups of other classical types, which we can do by considering these groups as subgroups of  $S_n$ . Our main focus is Weyl groups of type  $C$ , which give rise to type- $C$  local moves. Like Heitsch for local moves, we find that the plane tree model is particularly natural for type- $C$  local moves. Indeed we prove in Corollary 5.8 that the graph  $G^C$  of plane trees under type- $C$  local moves contains exactly two connected components: one consisting of symmetric plane trees and the other consisting of asymmetric plane trees.

We conclude with a brief discussion of extending  $s_i$ -local moves to types  $D$  and  $B$ , as well as possible biological interpretations of all the local moves we describe. We give open questions throughout the manuscript.

Throughout this manuscript  $Y$  denotes standard Young tableaux and  $T$  denotes plane trees.

## 2. Maps on tableaux corresponding to simple transpositions

In this section we describe a set of involutions on the set of standard Young tableaux of shape  $(n, n)$  that are indexed by simple reflections. Our maps are inspired by a well-known  $S_n$ -action from classical representation theory that gives all irreducible representations of the symmetric group. Our maps do not generate a group action, as we show in Remark 2.6. However, because they are involutions, our maps induce a graph whose vertices are the set of standard Young tableaux of shape  $(n, n)$  and whose edges correspond to the image under each map. We define this graph in this section. In subsequent sections we study combinatorial properties of the graph, prove that these maps agree with operations on plane trees from combinatorial biology, and discuss how to change the Lie type of our maps.

To begin we recall the definition of Young tableaux and sketch their relationship to the representation theory of the symmetric group  $S_n$ .

**Definition 2.1.** Let  $\lambda$  be a partition of  $n$ . A Young diagram of shape  $\lambda$  is a collection of  $\lambda_1$  boxes in the top row,  $\lambda_2$  boxes in the second row, and so on, aligned on the top and the left. A standard Young tableau  $Y$  of shape  $\lambda$  is a filling of the Young diagram with the integers  $\{1, 2, \dots, n\}$  without repetition so that each row increases left-to-right and each column increases top-to-bottom.

The *Specht module* for a partition  $\lambda$  is generated as a complex vector space by vectors  $v_T$  indexed by standard tableaux  $T$  of shape  $\lambda$ . The dimension of the irreducible representation of  $S_n$  corresponding to  $\lambda$  is also the number of standard Young tableaux of shape  $\lambda$ . A reasonable question arises: is there an action of  $S_n$  on standard Young tableaux under which the Young tableaux themselves can be the basis for the irreducible representation? Sadly the answer is generally no: the vectors  $v_Y$  in the Specht module are linear combinations of terms corresponding to different fillings of  $\lambda$ . (See [Fulton 1997; Sagan 2001] for more.) The problem is that  $S_n$  “should” act by permuting the entries of  $Y$  but permuting the entries of  $Y$  usually doesn’t produce another standard tableau.

In the following family of maps, we modify the permutation action on all fillings so that it produces standard tableaux. We define the maps on standard Young tableaux for arbitrary partitions; in later sections we specialize to the case when  $\lambda = (n, n)$  and the maps correspond to elements of  $S_{2n}$ .

**Definition 2.2.** Suppose that  $Y$  is a standard Young tableau with  $n$  boxes and  $s_i = (i, i + 1)$ , where  $i = 1, \dots, n - 1$ , is a simple reflection. If  $i, i + 1$  are not in the same row or in the same column of  $Y$  then define  $s_i(Y)$  to be the tableau with  $i$  and  $i + 1$  exchanged. If  $i, i + 1$  are in the same row or in the same column of  $Y$  then define  $s_i(Y)$  to be  $Y$ . Define an arbitrary word  $s_{i_1}s_{i_2} \cdots s_{i_k}(Y)$  to be the tableau obtained by composition of maps.

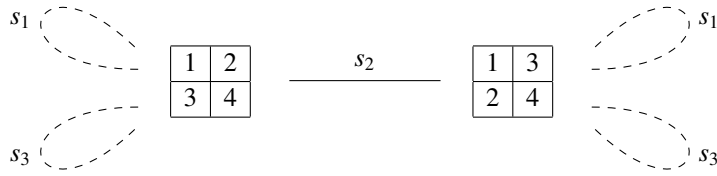
Others have considered an analogous action on 3-row tableaux [Housley et al. 2015] that comes from Vogan’s generalized tau invariant [1979].

The next result shows that these operations always give well-defined maps on standard tableaux (of arbitrary but fixed shape).

**Proposition 2.3.** *For each  $i = 1, \dots, n - 1$ , the map  $s_i$  is well-defined and the image  $s_i(Y)$  is a standard Young tableau of the same shape as  $Y$ .*

*Proof.* By construction,  $s_i$  preserves the shape of  $Y$ . By definition, the boxes containing  $i$  and  $i + 1$  inside the standard Young tableau  $Y$  have numbers less than  $i$  to the left and above and have numbers greater than  $i + 1$  to the right and below. Hence if  $s_i$  exchanges  $i$  and  $i + 1$  then the result  $s_i(Y)$  is also a standard Young tableau.  $\square$

Moreover, these maps have a convenient property.



**Figure 1.** Graph of  $s_i = (i, i + 1)$  on standard Young tableaux of shape  $(2, 2)$ .

**Proposition 2.4.** *Definition 2.2 produces a well-defined involution on the set of standard Young tableaux of shape  $\lambda$ .*

*Proof.* We check that for all  $i$  we have  $s_i^2 = e$  using two cases:

(1) If  $i$  and  $i + 1$  are in the same row then by definition  $s_i(Y) = Y$  so the claim holds.

(2) If  $i$  and  $i + 1$  are in different rows then  $s_i$  swaps the positions of  $i$  and  $i + 1$ . Applying  $s_i$  twice brings  $i$  and  $i + 1$  back to their original positions.  $\square$

This leads us to construct a graph whose vertices are standard Young tableaux of shape  $\lambda$  and whose edges describe the maps  $s_i$ . The edges are undirected precisely because the maps  $s_i$  are involutions for each  $i$ .

**Definition 2.5.** Let  $G_\lambda = (V, E)$  be the edge-labeled graph whose vertices  $V$  are the set of standard Young tableaux of shape  $\lambda$ . An edge labeled  $s_i$  connects tableaux  $Y$  and  $Y'$  when  $s_i(Y) = Y'$ . We call  $G_\lambda$  the graph of  $s_i$ -local moves for  $\lambda$ .

As an example, the graph  $G_{(2,2)}$  corresponding to the partition  $(2, 2)$  is shown in Figure 1.

**Remark 2.6.** Note that the maps  $s_i$  do not induce a group action of the symmetric group on the standard Young tableaux even for the shape  $(n, n)$ . For a counterexample, inspect Figure 1. On the one hand  $s_2s_3s_2(Y) = Y$  for each standard tableau  $Y$  of shape  $(2, 2)$ . On the other hand  $s_3s_2s_3(Y)$  is the opposite tableau of shape  $(2, 2)$ . Since  $s_2s_3s_2 = s_3s_2s_3$  in the symmetric group, we conclude that the maps  $s_i$  do not define a group action.

**Remark 2.7.** We typically omit all edges corresponding to fixed points  $Y = s_i(Y)$  (represented in Figure 1 as dashed self-edges) from our drawings of  $G_\lambda$ . In later sections we restrict to the case  $\lambda = (n, n)$  and so omit  $\lambda$  from our notation. We will also modify the maps  $s_i$  that define the edges, so we often write  $G^A$  to denote the graph with the precise edges in Definition 2.5 or write  $G^C$  to denote the modified graph in Section 5.

**Question 2.8.** In subsequent sections we analyze the graph  $G_{(n,n)}$ . What can be said about the graph  $G_\lambda$  for arbitrary partitions?

### 3. $S_n$ -action and local moves on plane trees

This section relates the functions defined in the previous section to an operation on plane trees called *local moves*. Condon, Heitsch, and Hoos defined local moves to represent an unfolding-and-refolding process on a strand of RNA. Heitsch [2006] then proved many combinatorial properties of a graph whose vertices are plane trees and whose edges come from local moves, for instance that the graph is symmetric and unimodal. In the same paper, she also showed that under one natural modification to the edges, we obtain the graph whose vertices are noncrossing partitions and whose edges come from Kreweras complementation.

We extend these results in a different direction, showing that many local moves correspond naturally to the action of the maps  $s_i$  on standard Young tableaux. Since we specialize to Young diagrams of shape  $(n, n)$ , we also specialize to the permutations  $S_{2n}$  in this section.

We begin by recalling the definitions of plane trees and local moves.

**Definition 3.1.** A plane tree is a rooted tree whose subtrees at any vertex are linearly ordered.

Our convention for a plane tree is that the root is at the top and that the subtrees are linearly ordered from left to right. In figures, the root is drawn with an open circle and ordinary vertices are drawn with solid circles.

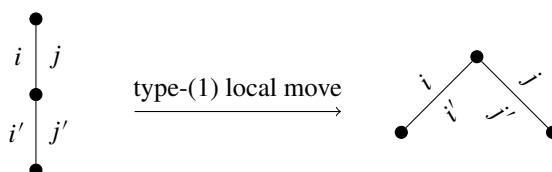
Plane trees are related to Young diagrams, noncrossing matchings, and other fundamental combinatorial objects that are also counted by Catalan numbers. To see this, we interpret each edge of a plane tree with  $n$  edges as a pair of two half-edges, each of which is indexed with one of the integers from 1 to  $2n$ . The half-edges are labeled in increasing order counterclockwise from the root. We write  $e(i, j)$  to denote the edge whose left half-edge is labeled  $i$  and whose right half-edge is labeled  $j$ . Given this setup, the half-edges  $i$  and  $j$  in the edge  $e(i, j)$  satisfy many constraints, including  $i < j$ .

The next definition describes *local moves*, which are operations on plane trees that are central to this paper. We denote the collection of plane trees with  $n$  edges by  $\mathcal{T}_n$ .

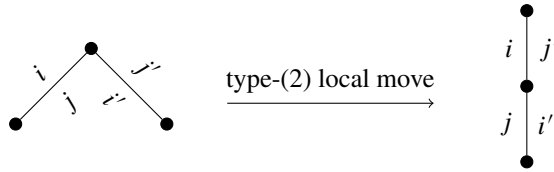
**Definition 3.2.** A local move on a plane tree  $T \in \mathcal{T}_n$  converts a pair of adjacent edges in one of two ways:

- (1) If  $i < i' < j' < j$  then replace  $e(i, j)$  and  $e(i', j')$  with  $e(i, i')$  and  $e(j', j)$ .

This is a local move of type (1):



- (2) If  $i < j < i' < j'$  then replace  $e(i, j)$  and  $e(i', j')$  with  $e(i, j')$  and  $e(j, i')$ . This is a local move of type (2):



The following map provides a natural bijection between plane trees with  $n$  edges and standard Young tableaux of shape  $(n, n)$ .

**Definition 3.3.** Let  $Y_{(n,n)}$  denote the set of standard Young tableaux of shape  $(n, n)$ . Define a map  $\phi : \mathcal{T}_n \rightarrow Y_{(n,n)}$  by the rule that for each  $T \in \mathcal{T}_n$  the Young tableau  $\phi(T)$  has the labels of the left half-edges of  $T$  on its top row and the labels of the right half-edges of  $T$  on its bottom row.

The following proposition confirms that the map  $\phi$  is bijective. Both the image and the domain are sets that are known to index the Catalan numbers [Stanley 1999, Chapter 6, Problem 19(e) and (ww)]; we include the following proof to confirm that the specific map  $\phi$  is a direct bijection.

**Proposition 3.4.** *The map  $\phi : \mathcal{T}_n \rightarrow Y_{(n,n)}$  is a well-defined bijection.*

*Proof.* The half-edges of a plane tree are labeled counterclockwise, so for each  $k$  there are at least as many left half-edges  $i$  with  $i \leq k$  as right half-edges  $j$  with  $j \leq k$ . Thus if  $i$  is above  $j$  in a column of the Young tableau  $\phi(T)$  then  $i < j$ . It follows that  $\phi$  is well-defined.

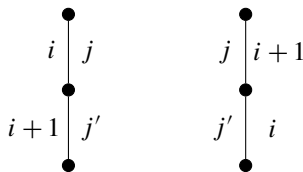
If  $\phi(T) = \phi(T')$  then both  $T$  and  $T'$  have the same set of left half-edges and the same set of right half-edges. Since by definition every subtree of a plane tree is linearly ordered, the indexing of the half-edges determines the plane tree. So  $\phi$  is injective.

The sets  $\mathcal{T}_n$  and  $Y_{(n,n)}$  have the same cardinality so the map  $\phi$  is a bijection.  $\square$

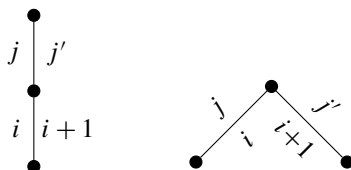
In order to prove our main result, we need more precise information about the fragments of a plane tree that correspond to the boxes filled with  $i$  and  $i + 1$  in a standard Young tableau. The next lemma compiles this information.

**Lemma 3.5.** *Consider a standard Young tableau  $Y$  of shape  $(n, n)$  and its preimage  $\phi^{-1}(Y)$  under the bijection in Definition 3.3. The half-edges corresponding to  $i$  and  $i + 1$  are in one of the following relative positions:*

- (i) *The numbers  $i$  and  $i + 1$  are on the same row in  $Y$  if and only if  $i$  and  $i + 1$  label left half-edges of  $\phi^{-1}(Y)$  in one of the ways shown in Figure 2.*
- (ii) *The numbers  $i$  and  $i + 1$  are on opposite rows in  $Y$  if and only if in  $\phi^{-1}(Y)$  either  $i$  and  $i + 1$  label a leaf (Figure 3, left) or the interior of a peak (Figure 3, right).*



**Figure 2.**  $i$  and  $i + 1$  are on the same row:  $i$  and  $i + 1$  label left half-edges (left) or  $i$  and  $i + 1$  label right half-edges (right).



**Figure 3.**  $i$  and  $i + 1$  are on different rows:  $i$  and  $i + 1$  are incident to the same leaf (left) or  $i$  and  $i + 1$  label the interior of a peak (right).



**Figure 4.**  $i$  and  $i + 1$  are on the same column.

(iii) *The numbers  $i$  and  $i + 1$  are on the same column in  $Y$  if and only if  $i$  and  $i + 1$  label a leaf incident to the root in  $\phi^{-1}(Y)$ , shown in Figure 4.*

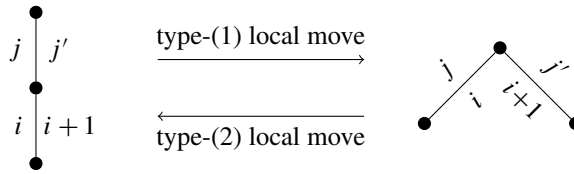
*In no case is there an additional half-edge incident to the vertex between  $i$  and  $i + 1$ .*

*Proof.* By convention, plane trees are labeled counterclockwise from the root. Hence there can be no edges or half-edges on the vertex incident to both  $i$  and  $i + 1$ . We think of each edge  $e(i, j)$  as having a left half-edge labeled  $i$  and a right half-edge labeled  $j$ .

(i) Consider the case where the numbers  $i$  and  $i + 1$  are on the same row in  $Y$ . By the definition of  $\phi$ , the top row of the Young tableau has the labels on the left half-edges of the corresponding plane tree, while the bottom row has the labels on the right half-edges. Suppose  $i$  and  $i + 1$  are on the top row of the Young tableau. Then  $i$  and  $i + 1$  are left half-edges and must be in the configuration shown in Figure 2, left. Suppose  $i$  and  $i + 1$  are on the bottom row of the Young tableau. Then  $i$  and  $i + 1$  are right half-edges and must be in the configuration shown in Figure 2, right.

(ii) Consider the case where the numbers  $i$  and  $i + 1$  are on different rows in  $Y$ . Suppose  $i$  is on the top row and  $i + 1$  is on the bottom row. Then  $i$  is a left half-edge





**Figure 5.** Edges of plane tree under local moves.

and  $i + 1$  is a right half-edge. That means these two numbers will label the same leaf in the tree, as shown in Figure 3, left. Now suppose  $i + 1$  is in the top row and  $i$  is in the bottom row of  $Y$ . Then  $i$  labels a right half-edge and  $i + 1$  labels a left half-edge. In a plane tree, this configuration must be a peak with  $i$  and  $i + 1$  labeling the interior, as shown in Figure 3, right.

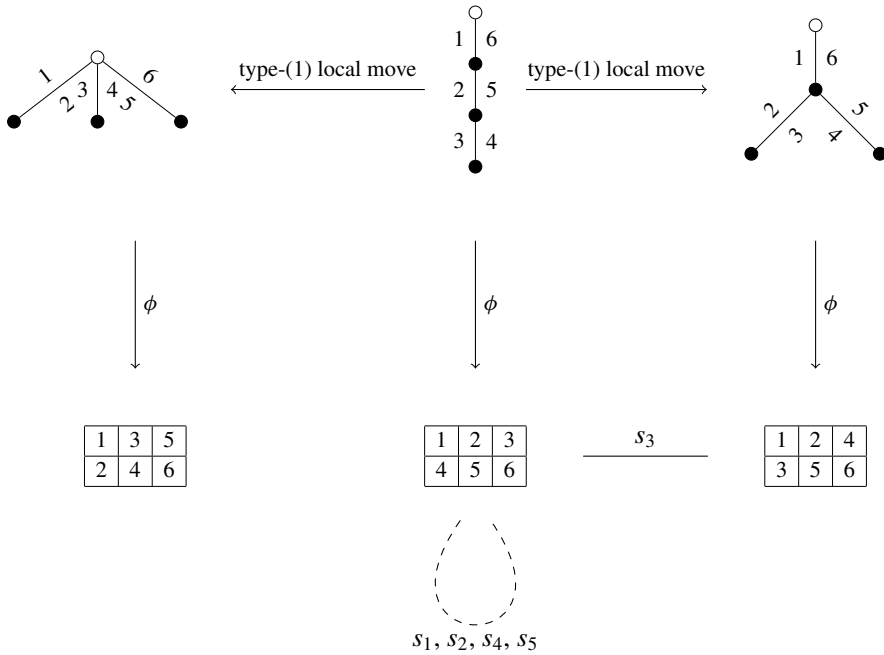
(iii) The numbers  $i$  and  $i + 1$  are on the same column of  $Y$  if and only if the first  $\frac{i-1}{2}$  columns of  $Y$  form a standard Young tableau of size  $(\frac{i-1}{2}, \frac{i-1}{2})$  and filled with the numbers  $1, 2, \dots, i - 1$ . By restricting  $\phi$  to plane trees on  $\frac{i-1}{2}$  edges we note that the first  $\frac{i-1}{2}$  edges of the plane tree  $\phi^{-1}(Y)$  form a subtree with the same root as  $\phi^{-1}(Y)$ . This is equivalent to saying that  $i - 1$  labels the right half of an edge incident to the root, which is true if and only if  $i$  and  $i + 1$  label the half-edges of a leaf incident to the root, as shown in Figure 4.  $\square$

In the next theorem we use Lemma 3.5 to show that if  $i$  and  $i + 1$  are in different rows (but not in the same column) of a standard Young tableau then the action of the map  $s_i$  on the tableau corresponds to a local move on the corresponding plane tree. Henceforth the maps  $s_i$  vary from  $i = 1$  to  $i = 2n - 1$  since there are  $2n$  boxes in the Young diagram.

**Theorem 3.6.** *Consider a plane tree  $T$  and its image  $Y = \phi(T)$  under the bijection in Definition 3.3. The half-edges in  $T$  labeled  $i$  and  $i + 1$  are in one of the two relative positions in Figure 3 if and only if the local move on edges with half-edges  $j < i < i + 1 < j'$  produces the plane tree  $\phi^{-1}(s_i(Y))$ .*

*Proof.* Lemma 3.5 showed that  $i$  and  $i + 1$  are on different rows and different columns exactly when  $i$  and  $i + 1$  are in the configurations in Figure 3. In fact, local moves exchange these two configurations because  $j < i < i + 1 < j'$ , as shown in Figure 5.

Let  $T'$  denote the image of  $T$  under the allowed local move on half-edges  $j < i < i + 1 < j'$  and let  $Y' = \phi(T')$ . Comparing  $T$  and  $T'$  in Figure 5 shows that  $i$  and  $i + 1$  change from a left half-edge to a right half-edge or vice versa. Thus  $i$  is on the opposite row in  $Y$  as it is in  $Y'$  and similarly for  $i + 1$ . By inspection of Figure 5, both  $j$  and  $j'$  stay on the same respective halves of their shared edge in  $T$  and  $T'$ . By definition, a local move changes only the two edges involved in the local



**Figure 6.** A local move that does not correspond to a permutation  $s_i$ .

move. Thus all other numbers remain on the same rows in the corresponding Young tableau, and so every other integer is in the same row in  $Y$  as it is in  $Y'$ . Finally  $i$  and  $i + 1$  are on opposite rows in  $Y$  by the hypotheses of the theorem together with Lemma 3.5. Thus  $Y' = s_i(Y)$ .

Conversely suppose there is a local move involving the half-edges  $j < i < i + 1 < j'$ . The configurations in Figure 5 are the only possibilities listed in Lemma 3.5 that satisfy these inequalities. The claim follows.  $\square$

**Remark 3.7.** Not every local move corresponds to one of the maps  $s_i$ . If  $i$  and  $i + 1$  are on the same row or column of a tableau then  $s_i$  fixes the tableau. Otherwise  $s_i$  describes the local moves in Figure 5. But when  $n > 2$ , a local move may be described by a transposition between  $i$  and  $i + k$  with  $1 < k < 2n - i$  in the tableau. Figure 6 gives an example. The original tableau  $Y$  has 1, 2, 3 along its top row, so every transposition except  $s_3$  fixes  $Y$ . However, the associated plane tree has a local move affecting the half-edges 1, 2, 5, 6 that corresponds to exchanging 2 and 5 in the tableau. The tableau resulting from this local move differs both from the original tableau  $Y$  and from  $s_3(Y)$ .

To avoid this ambiguity we have the following definition.

**Definition 3.8.** Suppose  $T$  is a plane tree with  $n$  edges whose associated standard Young tableau is  $\phi(T) = Y$ . An  $s_i$ -local move is a local move that is consistent with

one of the maps  $s_i$  in the sense that the local move sends  $T$  to  $\phi^{-1}(s_i(Y))$  for some  $s_i$  with  $i = 1, 2, \dots, 2n - 1$ . An  $s_i$ -local move is trivial if  $s_i(Y) = Y$ .

We conclude this section with an open question.

**Question 3.9.** What other types of transpositions  $(i, j)$  can also be interpreted as local moves on plane trees?

#### 4. The graph of $s_i$ -local moves in type $A$

Theorem 3.6 showed that the graph whose vertices are plane trees with  $n$  edges and whose edges are  $s_i$ -local moves is isomorphic to the graph in Definition 2.5 for the partition  $(n, n)$ . Remark 3.7 demonstrated that this graph is a subgraph (proper subgraph for  $n > 2$ ) of the graph of plane trees under *all* local moves.

Heitsch [2006] studied the graph of plane trees under *all* local moves and compared it to similar graphs for other combinatorial objects enumerated by Catalan numbers. However, when we remove edges from these graphs, many of Heitsch's properties no longer hold. We explore the statistics of these modified graphs in this section. We restrict our attention to the partition  $(n, n)$  and denote the graph from Definition 2.5 by  $G^A$ . We refer to  $G^A$  as the graph of  $s_i$ -local moves in type  $A$ . Note that the permutations whose corresponding maps  $s_i$  are defined on this partition are in  $S_{2n}$  rather than  $S_n$ . (In later sections we look at local moves corresponding to other Weyl groups.)

We begin by proving that the graph of  $s_i$ -local moves is still connected in type  $A$ .

**Proposition 4.1.** *The graph  $G^A$  is connected.*

*Proof.* We describe a way to construct a path between any two standard Young tableaux  $Y$  and  $Y'$  that both have shape  $(n, n)$ . If  $Y = Y'$  then the path is trivial. We now induct on the minimum number  $i$  that lies on opposite rows in  $Y$  and  $Y'$ . Suppose that  $i$  is the smallest number whose row in  $Y$  is different from that in  $Y'$ . Suppose further that  $i, i + 1, i + 2, \dots, i + k$  are all on the same row and  $i + k + 1$  is on the opposite row in  $Y$ . (We allow  $k$  to be zero.)

We first prove that in  $Y$  the number  $i + k + 1$  is not in the same column as any of  $i, i + 1, \dots, i + k$ . Indeed if  $i$  is on the bottom row then  $i + k + 1$  must be in a column to the right of  $i + k$  in order for  $Y$  to be standard. Now suppose that  $i$  is on the top row of  $Y$  and thus on the bottom row of  $Y'$ . In  $Y'$  we know that  $i$  is directly below one of  $1, 2, \dots, i - 1$  in order for  $Y'$  to be standard. Both  $Y$  and  $Y'$  have  $1, 2, \dots, i - 1$  in the same positions, so  $Y$  has an empty box in the bottom row below one of  $1, 2, \dots, i - 1$ . This must be the box occupied by  $i + k + 1$ .

Now consider the standard tableau  $s_i s_{i+1} s_{i+2} \cdots s_{i+k-1} s_{i+k}(Y)$ . It is connected to  $Y$  in the graph  $G^A$  by construction. The numbers  $1, 2, \dots, i - 1$  are in the same positions in  $s_i s_{i+1} \cdots s_{i+k}(Y)$  as in  $Y$ . Furthermore the number  $i$  occupies opposite

rows in  $s_i s_{i+1} \cdots s_{i+k}(Y)$  and  $Y$ . Thus the first  $i$  numbers are on the same rows in  $s_i s_{i+1} \cdots s_{i+k}(Y)$  as in  $Y'$ . If  $1, 2, \dots, 2n - 1$  are all on the same rows in  $Y$  as in  $Y'$  then  $2n$  must also be on the same row in  $Y$  and  $Y'$ . (Indeed  $2n$  is on the bottom row for all standard tableaux.) By induction we can find a path from  $Y$  to  $Y'$  in  $G^A$  as desired.  $\square$

The graph of plane trees under local moves has the structure of a graded poset. This is true for  $G^A$  as well, but for a different rank function. The next two results describe *total distance* and *total number of descendants*, two functions that rank  $G^A$ . Like Heitsch, we find that the language of plane trees characterizes the ranking more naturally than tableaux. In particular, we show that  $s_i$ -local moves change both the total distance and the total number of descendants by exactly 1.

**Proposition 4.2.** *Fix a plane tree  $T$  with root  $v_0$ :*

- *The total distance of the plane tree  $d_T$  is defined as*

$$d_T = \sum_{v \in V(T)} \text{dist}(v, v_0).$$

- *If  $T'$  is obtained from  $T$  by an  $s_i$ -local move of type (1) then  $d_T - 1 = d_{T'}$ . If  $T'$  is obtained from  $T$  by an  $s_i$ -local move of type (2) then  $d_T + 1 = d_{T'}$ .*

*Proof.* The proof follows by comparing the distances in the schematics in Figure 7. An  $s_i$ -local move does not change the distance between the root and the vertices in the subtrees  $a, b, c, d$ , and  $e$ , each of which can be empty. In the tree to the left, the leaf between half-edges  $i$  and  $i + 1$  has no descendants. Moreover, this vertex is one edge farther from the root than both “ankles” of the tree to the right are, changing the total distance by exactly 1.  $\square$

**Proposition 4.3.** *Fix a plane tree  $T$  with root  $v_0$ :*

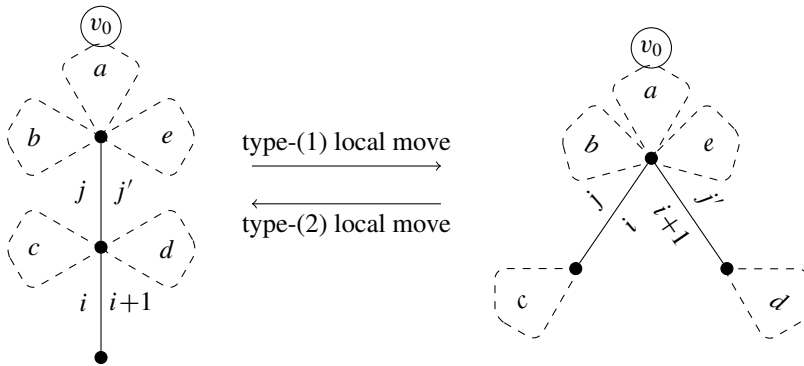
- *The total number of descendants in  $T$  is defined as*

$$\text{des}_T = \sum_{v \in V(T)} |\{\text{descendants of } v\}|.$$

- *If  $T'$  is obtained from  $T$  by an  $s_i$ -local move of type (1) then  $\text{des}_T - 1 = \text{des}_{T'}$ . If  $T'$  is obtained from  $T$  by an  $s_i$ -local move of type (2) then  $\text{des}_T + 1 = \text{des}_{T'}$ .*

*Proof.* Consider again the schematic in Figure 7. The number of descendants of the root, as well as all vertices in the subtrees  $a, b, c, d, e$ , remain the same after each  $s_i$ -local move. However, the length-2 path to the left has a total of three descendants, while the peak to the right has a total of only two.  $\square$

The previous proofs were similar in part because they turn out to count the same quantities, as we prove next.



**Figure 7.** Edges with subtrees under  $s_i$ -local moves.

**Proposition 4.4.** *Let  $T$  be a plane tree with root  $v_0$ . The total distance equals the total number of descendants; namely*

$$d_T = \text{des}_T .$$

*Proof.* Vertex  $v$  of plane tree  $T$  has distance  $k$  from the root exactly when the unique path between  $v$  and the root has  $k + 1$  vertices on it. The  $k$  vertices on this path other than  $v$  are precisely the vertices in  $T$  with  $v$  as a descendant. Thus each vertex  $v$  contributes exactly  $k$  to  $d_T$  and exactly  $k$  to  $\text{des}_T$ .  $\square$

**Remark 4.5.** The notions of total distance and of total number of descendants can be useful in different contexts. For one example, see the proof of Proposition 4.8. For another example, note that each descendant in a plane tree corresponds to a nesting of arcs in the associated noncrossing matching. Thus the total number of descendants in a plane tree corresponds to the total number of nestings within a noncrossing matching. (We do not discuss noncrossing matchings in detail in this manuscript; for more, see, e.g., [Russell 2011; Russell and Tymoczko 2011].)

The next proposition is a direct result of the previous propositions.

**Proposition 4.6.** *Both total distance and total number of descendants partition the vertices of  $G^A$  into the same subsets of plane trees.*

*Direct the graph  $G^A$  according to the rule that each edge is directed  $T \rightarrow T'$  if  $T'$  is obtained from  $T$  by a local move of type (1). This turns  $G_A$  into a graded poset. Moreover, we can impose a rank function  $\rho(T) = d_T$  on this graded poset, whose ranks are characterized by the subsets of plane trees with total distance  $k$  (respectively total number of descendants  $k$ ).*

*Proof.* The first claim is an immediate corollary of the fact that  $d_T = \text{des}_T$  for each plane tree  $T$ .

The directed graph  $G^A$  is acyclic, and thus a poset, because if  $T_1 \rightarrow T_2 \rightarrow \dots \rightarrow T_k$  is any directed path then  $d_{T_1} > d_{T_2} > \dots > d_{T_k}$  and so the endpoint cannot coincide with the initial point of the path.

Finally a function is a rank function if the following two conditions are met:

- (1) The function is compatible with the partial order; namely if there is a path  $T_1 \rightarrow T_2 \rightarrow \dots \rightarrow T_k$  then  $\rho(T_1) > \rho(T_k)$ . We just confirmed this for total distance (respectively total number of descendants).
- (2) If  $T_1 \rightarrow T_2$  is an edge in the graph then  $\rho(T_1) = \rho(T_2) + 1$ . This is the content of Proposition 4.2 (respectively Proposition 4.3 for total number of descendants).

The final claim follows by definition of the rank function. □

**Remark 4.7.** The graph  $G^A$  does not satisfy the same kind of symmetries as the graph for all local moves does. For instance, Heitsch proved that the number of plane trees of rank  $k$  agrees with those of rank  $n - k + 1$  for each  $k = 1, \dots, n$ . That is clearly false here: for instance, the sequence of the number of tableaux of shape  $(3, 3)$  of each rank in increasing order is  $(1, 2, 1, 1)$ , as shown in Figure 8. (Note too that this is not the same rank function that Heitsch uses, as our graded poset has fewer edges than hers.)

However, we can prove the following.

**Proposition 4.8.** *There is a unique element of maximal rank and a unique element of minimal rank.*

*Proof.* Consider the graph  $G^A$  whose vertex set is the set  $\mathcal{T}_n$  of plane trees with  $n$  edges. If  $T$  is a plane tree in  $\mathcal{T}_n$  then its root must have  $n$  descendants since any other vertex in the graph is a descendant of the root. So the minimal total number of descendants is  $n$ . This is achieved by the star graph in Figure 9, left.

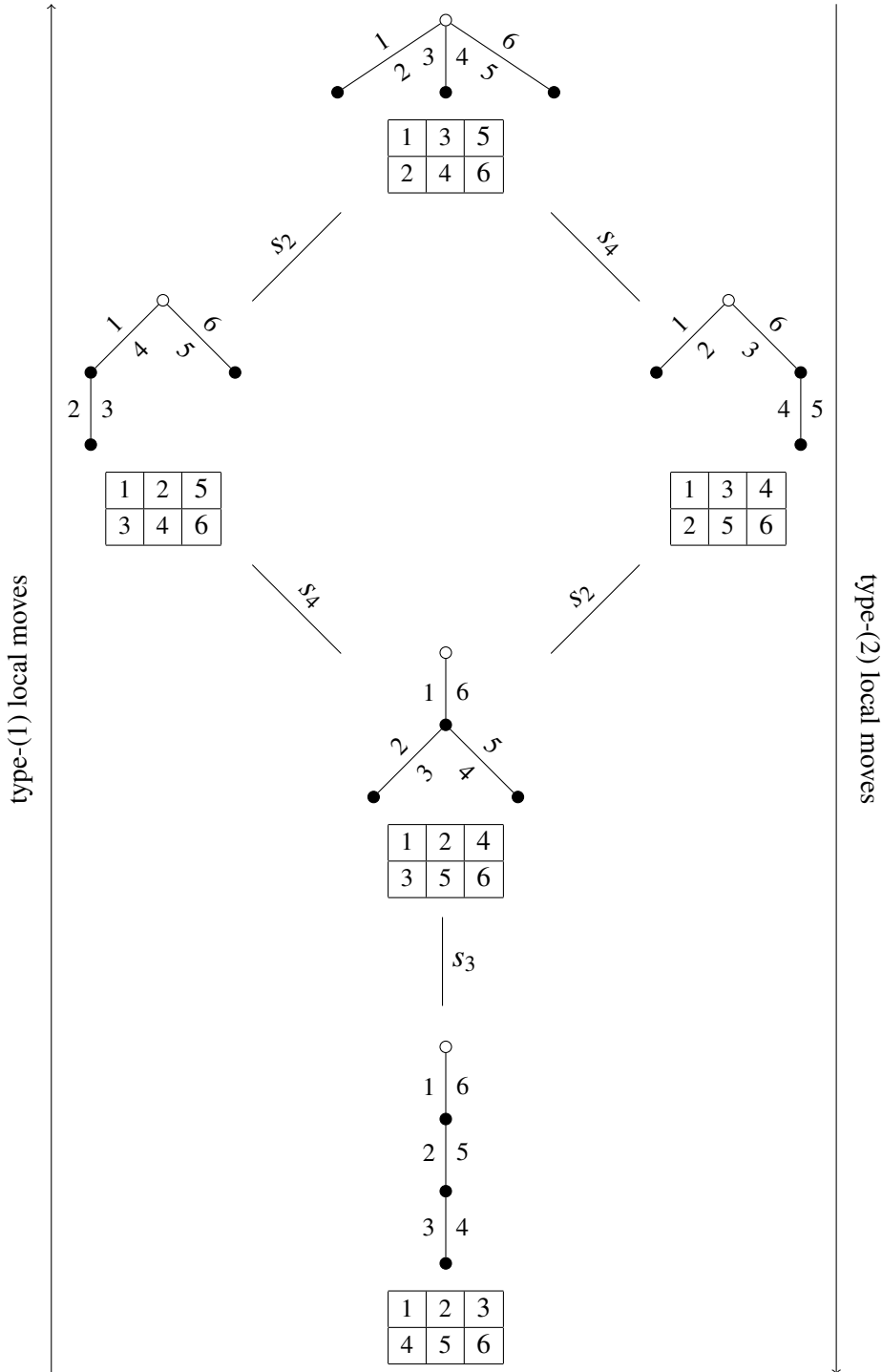
The plane tree  $T$  is connected so there is at least one vertex of each possible distance from the root. The path graph in Figure 9, right, has just one vertex at each distance from the root and therefore maximizes the total distance. □

**Corollary 4.9.** *For the partition  $(n, n)$ , the number of ranks in the graded poset obtained from  $G^A$  and ranked by the total distance function  $d_T$  is  $\binom{n+1}{2} - n + 1$ .*

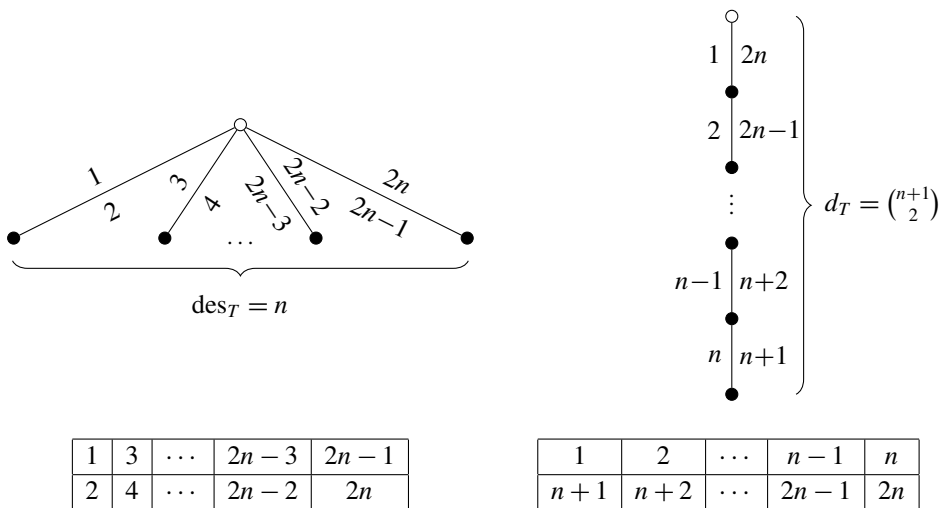
*Proof.* The total distance of the path graph is the binomial coefficient  $\binom{n+1}{2}$ . The total distance of the star graph is  $n$ . There is at least one plane tree of each rank between these because  $G^A$  is connected and each edge changes rank by exactly 1. □

Again we close with an open question.

**Question 4.10.** Is the rank sequence of  $G^A$  unimodal for every  $n$ ?



**Figure 8.** Graded poset obtained from  $G^A$  when  $n = 3$ .



**Figure 9.** Minimal tree which is a star graph (left) and a maximal tree which is a path graph (right) with associated Young tableaux.

### 5. The graph of $s_i$ -local moves in type C

In our description of  $s_i$ -local moves so far, we relied on an analogy with the generators of the symmetric group. We now extend the analogy to define maps  $s_i^C$  corresponding to the generators of the Weyl group of type C. Intuitively the Weyl group of type C plays the same role for the complex symplectic group  $\text{Sp}(2n, \mathbb{C})$  that the permutation matrices play for  $n \times n$  invertible matrices  $\text{GL}(n, \mathbb{C})$ . We will represent the Weyl group of type C as a subgroup of the permutations in  $S_{2n}$  using generators that we describe below.

In this section we show that we can easily define maps  $s_i^C$  on the standard tableaux of shape  $(n, n)$  even when there are no analogous local moves on the corresponding plane trees. Nonetheless, the geometry of the plane trees is the best way to describe key properties of these maps. More precisely we prove that restricting to type-C  $s_i$ -local moves identifies symmetry within the plane trees. The main theorem of this section shows that within the graph whose vertices are plane trees and whose edges are type-C  $s_i$ -local moves, there are precisely two connected components: one composed of *symmetric* plane trees and one composed of *asymmetric* plane trees.

We define functions analogous to the maps  $s_i$  for type C instead of type A. The reader who is not familiar with Weyl groups can take this as a definition of the Weyl group of type C. Like in our earlier treatment, the maps  $s_i$  and  $s_i^C$  are both permutations in  $S_{2n}$ . However, note that in type A we have maps  $s_i$  for each  $i \in \{1, 2, \dots, 2n - 1\}$ , while in type C we only have  $s_i^C$  for  $i \in \{1, 2, \dots, n\}$ .



**Definition 5.1.** The maps of type  $C$  are the involutions on standard tableaux defined by

$$\begin{aligned}
 s_1^C &= s_1 s_{2n-1} && \text{corresponding to the reflection } (1, 2)(2n - 1, 2n), \\
 s_2^C &= s_2 s_{2n-2} && \text{corresponding to the reflection } (2, 3)(2n - 2, 2n - 1), \\
 &\vdots \\
 s_{n-1}^C &= s_{n-1} s_{n+1} && \text{corresponding to the reflection } (n - 1, n)(n + 1, n + 2), \\
 s_n^C &= s_n && \text{corresponding to the reflection } (n, n + 1).
 \end{aligned}$$

Using the bijection  $\phi : \mathcal{T}_n \rightarrow \{\text{standard Young tableaux of size } (n, n)\}$  we also define maps  $s_i^C$  on plane trees according to the rule

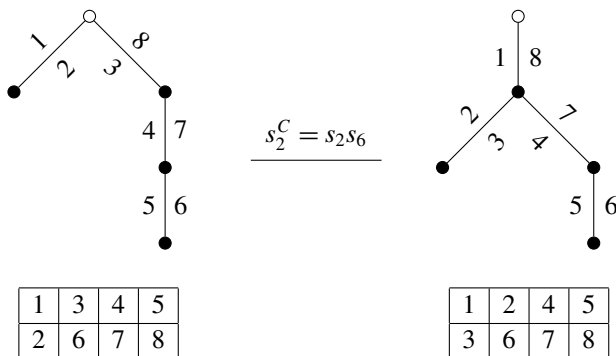
$$s_i^C(T) = \phi^{-1}(s_i^C(\phi(T))).$$

Generally the simple reflections of type  $C$  exchange disjoint pairs of integers according to the product  $s_i s_{2n-i}$  of type- $A$  reflections. However, note that  $s_n$  exchanges just the two integers  $n$  and  $n + 1$ . It is the only simple reflection of type  $C$  that exchanges integers between the sets  $\{1, 2, \dots, n\}$  and  $\{n + 1, n + 2, \dots, 2n\}$ .

**Remark 5.2.** Note that while we use terminology from earlier in the paper, the maps  $s_i^C$  are no longer local moves in the strict sense. Except for the case when  $i = n$ , the maps  $s_i^C = s_i s_{2n-i}$  corresponding to the reflections  $(i, i + 1)(2n - i, 2n - i + 1)$  are in fact *pairs* of  $s_i$ -local moves of type  $A$ . We can perform a pair of  $s_i$ -local moves on a standard tableau  $Y$  simultaneously because the pairs of integers are disjoint: if  $i$  and  $i + 1$  are in the same row or column then  $s_i$  does nothing; otherwise  $s_i$  exchanges the positions of  $i$  and  $i + 1$  leaving all the other numbers in their original positions. The same dynamic holds for  $s_{2n-i}$  with respect to  $2n - i$  and  $2n - i + 1$ . So the standard tableau  $s_i^C(Y)$  is always defined.

Our definition for the plane tree  $s_i^C(T)$  uses the action on the corresponding tableau  $\phi(T)$ . This is because often a pair of  $s_i$ -local moves that would act on a plane tree is *not defined* on that plane tree. Figure 10 provides an example in which the map  $s_2^C$  involves one nontrivial  $s_2$ -local move and one trivial  $s_6$ -local move. The heuristic for determining  $s_i^C(T)$  directly is to perform all of the local moves  $s_i$  and  $s_{2n-i}$  that are nontrivial.

We stress that even though it appears unnatural to define local moves of type  $C$  on plane trees (given that the constituent local moves of type  $A$  are not necessarily well-defined), the maps  $s_i^C$  characterize key geometric properties of the plane trees. Indeed we think it is a theme of this field that different characterizations of standard tableaux (plane trees, noncrossing matchings, etc.) provide valuable and often complementary information.



**Figure 10.** Map  $s_2^C$  involving a nontrivial  $s_2$ -local move and a trivial  $s_6$ -local move.

The maps  $s_i^C$  define a graph  $G^C$  in the same way that the maps  $s_i$  defined a graph  $G^A$ .

**Definition 5.3.** The graph  $G^C$  is the graph whose vertices are plane trees. An edge connects plane trees  $T$  and  $T'$  precisely when  $T' = s_i^C(T)$  for a map  $s_i^C$ . We call  $G^C$  the graph of plane trees under  $s_i^C$ -local moves (read  $s_i$ -local moves of type  $C$ ).

The following definition formalizes our notion of symmetric and asymmetric plane trees.

**Definition 5.4.** Let  $T$  be a plane tree. We say that  $T$  is symmetric if and only if for each edge  $e(i, j)$  in  $T$  the mirror image  $e(2n - j + 1, 2n - i + 1)$  is also an edge in  $T$ . A plane tree is asymmetric if it is not symmetric.

We will prove that the graph of plane trees  $G^C$  under the  $s_i^C$ -local moves has two connected components: one consisting of symmetric plane trees and one consisting of asymmetric plane trees. Our proof uses several steps. First we show that no connected component contains both a symmetric plane tree and an asymmetric plane tree.

**Lemma 5.5.** Each connected component of  $G^C$  consists either entirely of symmetric plane trees or entirely of asymmetric plane trees.

*Proof.* We will show that if  $s_i^C$  is a generator of the Weyl group of type  $C$  and  $T$  is a symmetric plane tree then  $s_i^C(T)$  is also symmetric. It follows that the connected component of  $G^C$  containing any symmetric plane tree consists entirely of other symmetric plane trees. Since every tree is either symmetric or asymmetric, it follows further that the connected component of  $G^C$  containing any asymmetric plane tree must consist entirely of other asymmetric plane trees.

Given a subtree  $T'$  of symmetric plane tree  $T$  we call the edges in  $T$  that are symmetric to  $T'$  the mirror image of  $T$ .

Consider the half-edges labeled by  $i$  and  $i + 1$ . A priori there are four possibilities: they could both be left-half-edges, they could both be right-half-edges, they could form a leaf, or they could form the interior of a peak.

Table 1 shows these four possibilities, the mirror image of these possibilities, the  $s_i^C$ -local move on the original and its mirror image in each case, and the mirror image of the  $s_i^C$ -local move on the original in each case. Note that in the first two possibilities,  $i$  and  $i + 1$ , as well as  $2n - i$  and  $2n - i + 1$ , will stay in their respective rows of the corresponding tableaux after the  $s_i^C$ -local move, which consequently does not alter the either  $T'$  or its mirror image. We inspect columns three and five in Table 1 and observe that they are the same. So if two edges were part of a symmetric tree before we perform an  $s_i^C$ -local move on them, then they will still be part of a symmetric tree after the  $s_i^C$ -local move.

Since these are the only edges changed by the local move, all the other edges will still satisfy the symmetry condition. We conclude that  $s_i^C(T)$  is symmetric whenever  $T$  is symmetric. The result follows.  $\square$

Next we prove there is exactly one connected component of symmetric plane trees in  $G^C$  by showing that each symmetric plane tree can be transformed via  $s_i^C$ -local moves to one with the leaf  $e(1, 2)$  and then using induction.

**Theorem 5.6.** *If  $T$  and  $T'$  are symmetric plane trees then there is a finite sequence of  $s_i^C$ -local moves that transforms  $T$  into  $T'$ .*

*Proof.* The proof is by induction on the total number  $n$  of edges in a plane tree.

There are two base cases. The case when  $n = 2$  was addressed in Figure 1 since  $s_2^C = s_2$  in that setting; it is reproduced in type- $C$  notation in Figure 11, left. The case when  $n = 3$  has three symmetric plane trees as shown in Figure 11, right: the top and the middle are connected by the edge  $s_3^C = s_3$ , while the middle and the bottom are connected by  $s_2^C = s_2s_4$ .

For the induction step, assume that any two symmetric plane trees with at most  $n - 1$  edges can be transformed into each other by a sequence of  $s_i^C$ -local moves. Now consider a symmetric plane tree with  $n$  edges.

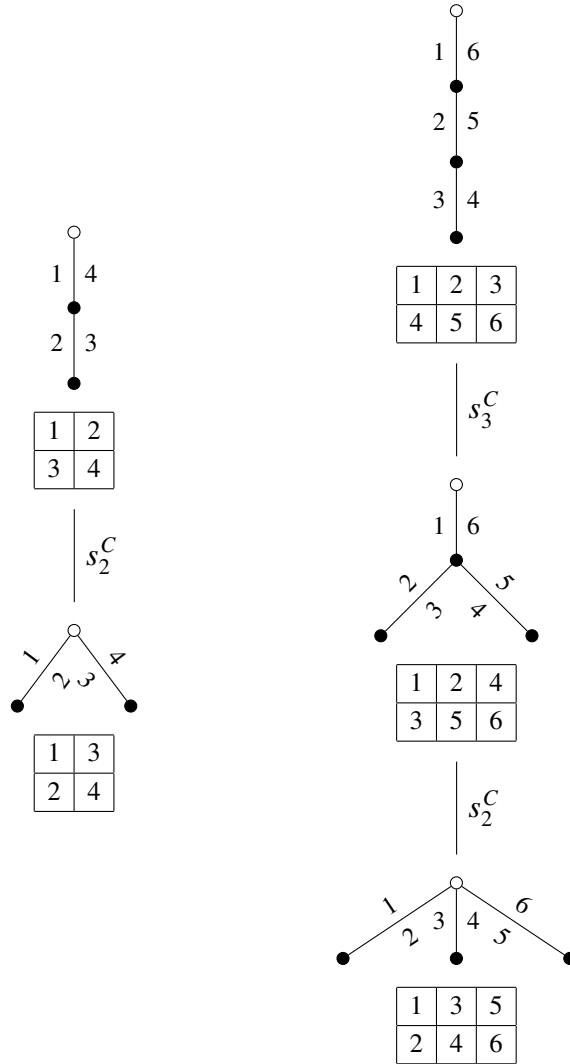
First we show that there is a path of  $s_i^C$ -local moves from each plane tree  $T$  to a plane tree containing the edge  $e(1, 2)$ . If  $T$  does not have the edge  $e(1, 2)$  then it has the edge  $e(1, j)$  for some  $j \geq 3$ . This means that 1 and 2 are both in the top row of the tableau  $\phi(T)$ . Let  $k$  be the first integer *not* in the top row of  $\phi(T)$ . Since  $\phi(T)$  has shape  $(n, n)$ , we know that  $k \leq n + 1$ . Proposition 4.1 showed that the standard tableau  $s_2s_3 \cdots s_{k-1}(\phi(T))$  has 2 in the bottom row by way of  $s_i$ -local moves of type  $A$ . We now confirm that  $s_2^Cs_3^C \cdots s_{k-1}^C$  also moves 2 to the bottom row. If  $k = n + 1$  then the top row of the tableau is filled with the integers from 1 through  $n$ , and  $s_{k-1}^C = s_n$  simply exchanges  $n$  and  $n + 1$ . For  $j \leq n$  we know that  $s_2^Cs_3^C \cdots s_{j-1}^C$  permutes numbers within the disjoint sets  $\{1, 2, \dots, j\}$

$T'$	mirror image of $T'$	$s_i^C$ -local move on mirror image of $T'$	$s_i^C$ -local move on $T'$	mirror image of $s_i^C$ -local move on $T'$

**Table 1.** Identical results from  $s_i^C$ -local move on mirror image of  $T'$  and mirror image of  $s_i^C$ -local move on  $T'$ .

and  $\{j + 1, \dots, 2n\}$  independently. So the tableau  $s_2^C s_3^C \dots s_{k-1}^C(\phi(T))$  has 2 on the bottom row for all  $k \leq n + 1$ . We therefore conclude that  $T$  is in the same connected component of  $G^C$  as a plane tree with the edge  $e(1, 2)$ .

We next show that all symmetric plane trees are in the same connected component of  $G^C$ . Suppose  $T$  and  $T'$  are both symmetric plane trees. By the previous argument, we can assume that they each contain the leaf  $e(1, 2)$  and hence by symmetry the leaf  $e(2n - 1, 2n)$ . Since these edges are both leaves, they can be erased without



**Figure 11.** Type-C base cases  $n = 2$  (left) and  $n = 3$  (right) for symmetric plane trees.

disconnecting the two trees. Consider the subtrees  $T_1$  and  $T'_1$  consisting respectively of all the edges of  $T$  and  $T'$  except  $e(1, 2)$  and  $e(2n - 1, 2n)$ . The two subtrees are still symmetric but have only  $n - 2$  edges. By the inductive hypothesis we can transform  $T_1$  into  $T'_1$  with a sequence of  $s_i^C$ -local moves, which also transforms  $T$  into  $T'$ . By induction the claim is proven.  $\square$

The proof for asymmetric plane trees is somewhat similar but more subtle.

**Theorem 5.7.** *If  $T$  and  $T'$  are asymmetric plane trees then there is a finite sequence of  $s_i^C$ -local moves that transforms  $T$  into  $T'$ .*

*Proof.* The proof is by induction on the total number of edges  $n$  in a plane tree.

The base cases for asymmetric plane trees occur when  $n = 3$  and when  $n = 4$ . There are two asymmetric plane trees with three edges, and these trees are related by  $s_2^C = s_2 s_4$ , as shown in Figure 12, left. There are eight asymmetric plane trees with four edges, and these trees are related by the  $s_i^C$ -local moves shown in Figure 12, right.

For the induction step, let  $n \geq 4$  and assume that any two asymmetric plane trees with at most  $n - 1$  edges can be transformed into each other by a sequence of  $s_i^C$ -local moves.

Let  $T$  be an arbitrary asymmetric plane tree with  $n$  edges. We describe an algorithm to obtain a sequence of  $s_i^C$ -local moves from  $T$  to a plane tree with only the edge  $e(1, 2n)$  incident to the root. (Note the special case of plane trees with three edges, for which there are no asymmetric trees containing the edge  $e(1, 2n) = e(1, 6)$ .) Figure 13 gives a schematic of  $T$  with notation for the half-edges  $j_1 < j_1 + 1 < j_2 < j_2 + 1 < \dots < j_{k-1} + 1 < j_k$  and the possibly empty subtrees  $a_i$ .

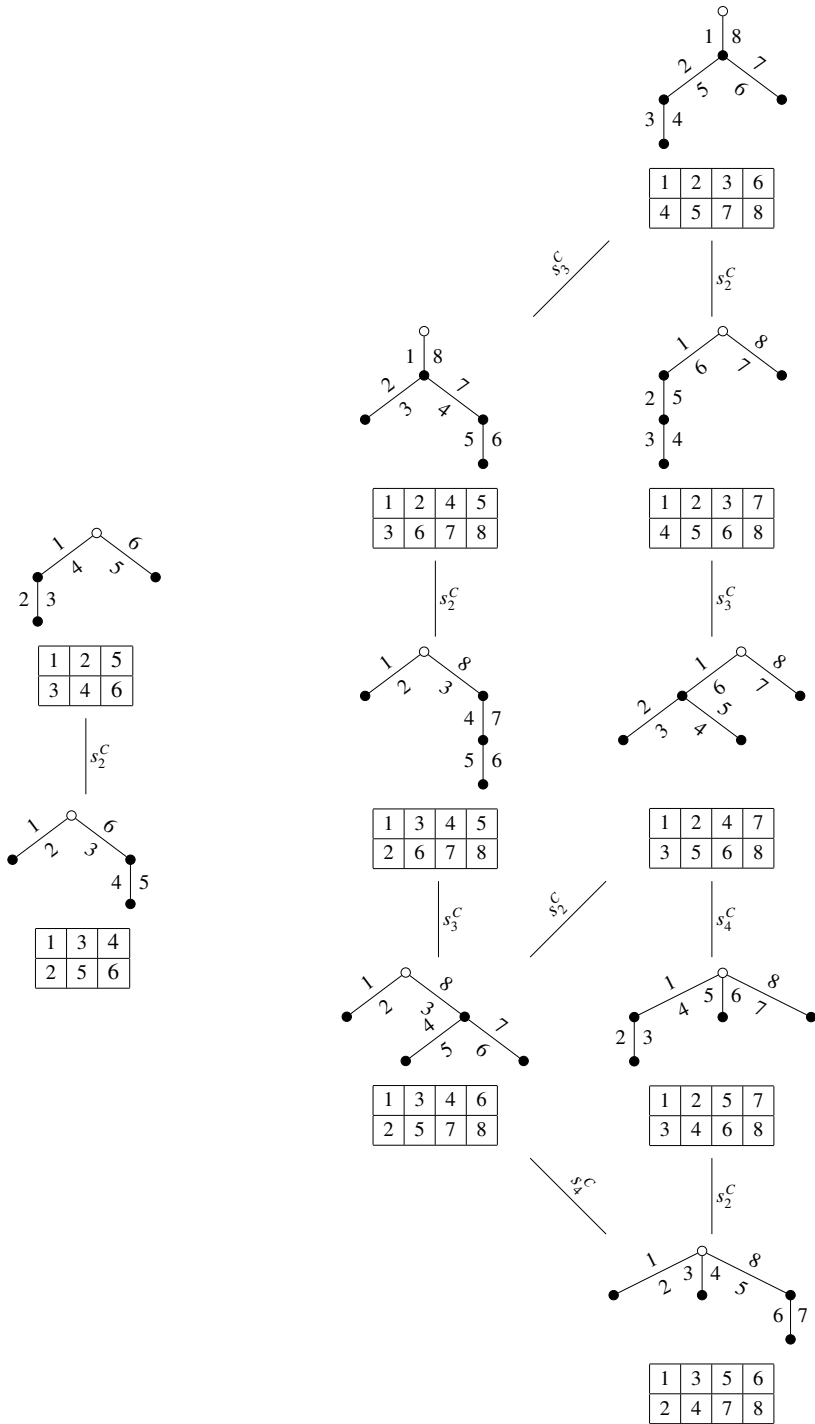
Since  $s_{j_1}^C = s_{j_1} s_{2n-j_1}$ , we can use an  $s_{j_1}^C$ -local move on edges  $e(1, j_1)$  and  $e(j_1 + 1, j_2)$  to form edges  $e(1, j_2)$  and  $e(j_1, j_1 + 1)$ . Repeat this process for each  $j_p$  with  $j_p \leq n$ .

We can continue this process for any edge  $e(1, j)$  with  $j > n$  as long as we are not in the case of Figure 14. The problem in that case is that the local move that collapses  $2n - j_p$  and  $2n - j_p + 1$  simultaneously triggers a type-(1) local move on half-edges  $j_p$  and  $j_p + 1$  and reinserts a lower-indexed branch into the root. (Note that  $j_p < n < 2n - j_p$  by our convention on the labeling of the half-edges in the plane tree.)

To address the case in Figure 14, we apply the sequence  $s_{j_p-1}^C s_{j_{q-1}}^C \dots s_{j_2}^C s_{j_1}^C$  of  $s_i^C$ -local moves. Since  $j_p < n$ , the sequence of local moves permutes indices in the sets  $\{j_1', \dots, j_p\}$  and  $\{2n - j_p + 1, 2n - j_p + 2, \dots, 2n - j_1' + 1\}$  independently. Thus after applying those  $s_i^C$ -local moves, the tree contains both of the edges  $e(1, 2n - j_p)$  and  $e(2, j_p + 1)$ . Applying  $s_{j_p}^C$  to that tree results in a plane tree with edge  $e(1, k)$  for  $k \geq 2n - j_p + 2$  as desired. Continuing this process, we obtain in all cases a sequence of  $s_i^C$ -local moves that transforms an arbitrary asymmetric plane tree to one containing the edge  $e(1, 2n)$ .

Finally we show that all asymmetric plane trees are in the same connected component of  $G^C$ . Suppose  $T$  and  $T'$  are both asymmetric plane trees with at least four edges. By the previous argument, we can assume that they each contain the edge  $e(1, 2n)$ . Consider the subtrees  $T_1$  and  $T'_1$  consisting of all the edges of  $T$  and respectively  $T'$  except  $e(1, 2n)$ . The two subtrees are still asymmetric but have only  $n - 1$  edges. By the inductive hypothesis we can transform  $T_1$  into  $T'_1$  with a sequence of  $s_i^C$ -local moves, which also transforms  $T$  into  $T'$ . By induction the claim is proven.  $\square$

The main result is a simple corollary of the previous results.



**Figure 12.** Type-C base cases  $n = 3$  (left) and  $n = 4$  (right) for asymmetric plane trees.

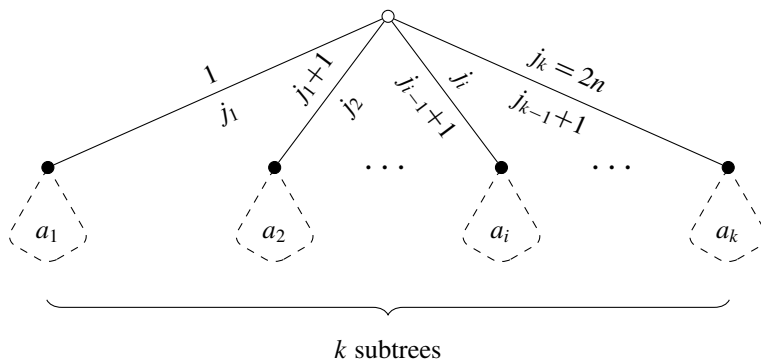


Figure 13. Edges incident to the root in  $T$ .

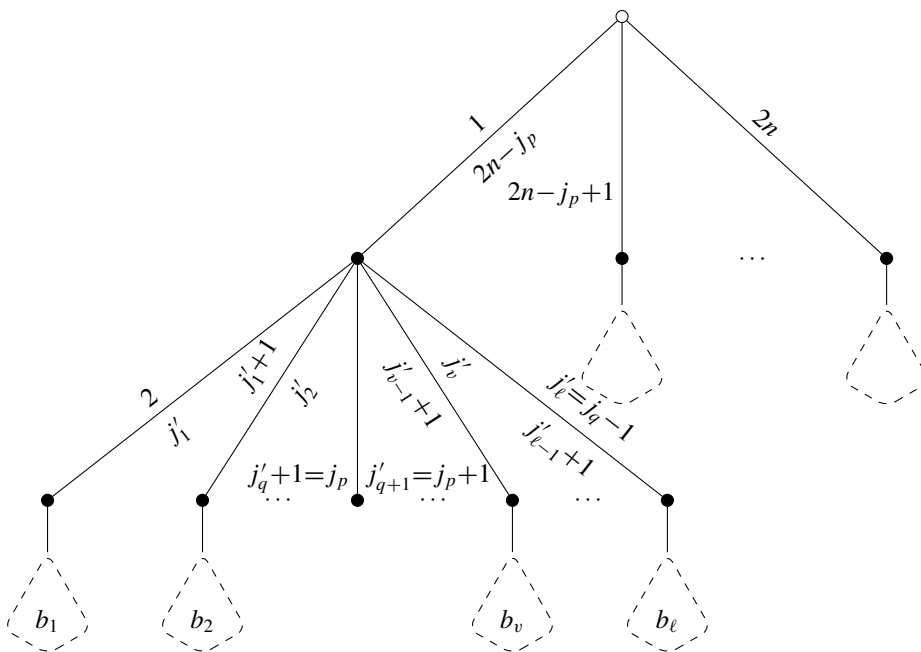


Figure 14. Problematic case.

**Corollary 5.8.** *The graph  $G^C$  has exactly two connected components: one containing exactly the symmetric plane trees and the other containing exactly the asymmetric plane trees.*

Appendix A in the arXiv version of this paper gives examples of Corollary 5.8 for  $n = 5, 6, 7$ . The *Mathematica* notebook that generates these orbits is publicly available online at <http://github.com/jujuwoman/RNA-combinatorics>.



We conclude with a formula for the size of each connected component in  $G^C$ , namely the number of symmetric plane trees and the number of asymmetric plane trees.

**Proposition 5.9.** *Given  $\mathcal{T}_n$  the number of symmetric plane trees is*

$$r = \sum_m \sum_{k_1+k_2+\dots+k_{m+1}=\frac{n-m}{2}} \prod_{j=1}^{m+1} C_{k_j},$$

where  $m$  varies over odd numbers between 0 and  $n$  when  $n$  is odd and over even numbers between 0 and  $n$  when  $n$  is even. The number of asymmetric plane trees is  $C_n - r$ .

*Proof.* We use the fact that the total number of plane trees with  $n$  edges is the Catalan number  $C_n = \frac{1}{n+1} \binom{2n}{n}$ .

Define the *middle path graph* of a symmetric plane tree in  $\mathcal{T}_n$  to be the maximal set of edges of the form  $e(i, 2n + 1 - i)$  for some  $i$  with  $1 \leq i \leq n$ . Let  $m$  be the number of edges in the middle path graph of a symmetric plane tree. To be a symmetric plane tree, any descendants to the left of a vertex in the middle path graph have their mirror image to the right of the same vertex. Thus the set of all symmetric plane trees can be constructed by all possible ways to attach plane trees to the left of the middle path graph, together with the mirror images on the right. There are  $m + 1$  vertices in the middle path graph; suppose that for each  $i$  with  $1 \leq i \leq m + 1$  the  $i$ -th vertex from the root in the middle path graph has a plane tree with  $k_i$  edges to its left. The sum  $k_1 + k_2 + \dots + k_{m+1}$  must satisfy

$$k_1 + k_2 + \dots + k_{m+1} = \frac{n - m}{2}$$

since there are  $n$  total edges in the tree,  $m$  edges on the middle path graph, and another  $k_1 + k_2 + \dots + k_{m+1}$  edges in the mirror images of the subtrees to the left of the middle path graph. By examining parity, we see that  $m$  varies over odd numbers from 0 to  $n$  if  $n$  is odd and over even numbers from 0 to  $n$  if  $n$  is even. For any such partition  $k_1 + k_2 + \dots + k_{m+1}$ , we can independently take any of the  $C_{k_i}$  plane trees on  $k_i$  vertices to attach to the left of the  $i$ -th vertex on the middle path graph, with its mirror image on the right. Thus the total number of symmetric plane trees with  $n$  edges is

$$\sum_m \sum_{k_1+k_2+\dots+k_{m+1}=\frac{n-m}{2}} \prod_{j=1}^{m+1} C_{k_j}$$

as desired. The number of asymmetric plane trees is simply the number of all plane trees minus the number of symmetric plane trees. □

**Question 5.10.** Do other properties of  $G^A$  hold for the components of  $G^C$  as well? For instance, is each component of  $G^C$  graded by a function with a straightforward description?

**6. Remarks on classical types  $B$  and  $D$  and possible biological interpretations**

We conclude this paper with remarks and questions in two directions. First we discuss whether  $s_i$ -local moves could be reasonably extended to Weyl groups of other classical Lie types. At the end we discuss speculative connections to biology.

*Extending  $s_i$ -local moves combinatorially to other classical Lie types.* There are two other Weyl groups of classical types, namely the Weyl groups of type  $B$  and type  $D$ . Both can be described as a subgroup of a sufficiently large permutation group.

We think the Weyl group of type  $D$  is unlikely to extend fruitfully to the setting of plane trees. The problem is that the generators of the Weyl group of type  $D$  cannot be written as a product of disjoint simple transpositions  $(i, i + 1)$ . Indeed, one generator must contain a transposition like  $(n, n + 2)$ . Within the permutation group, that transposition equals

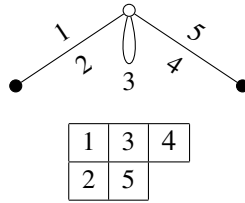
$$(n, n + 1)(n + 1, n + 2)(n, n + 1) = (n + 1, n + 2)(n, n + 1)(n + 1, n + 2).$$

However, the  $s_i$ -local moves do not form a group action; as we discussed in Remark 2.6 there is no consistent way to define  $(n, n + 2)$ .

By contrast the Weyl group of type  $B$  may lead to meaningful biological and combinatorial implications. The maps of type  $B$  are the involutions defined by

$$\begin{aligned} s_1^B &= s_1 s_{2n} && \text{corresponding to the reflection } (1, 2)(2n, 2n + 1), \\ s_2^B &= s_2 s_{2n-1} && \text{corresponding to the reflection } (2, 3)(2n - 1, 2n), \\ &\vdots && \\ s_{n-1}^B &= s_{n-1} s_{n+2} && \text{corresponding to the reflection } (n - 1, n)(n + 2, n + 3), \\ s_n^B &= s_n && \text{corresponding to the reflection } (n, n + 2). \end{aligned}$$

Note that  $s_n^B$  is different from the other permutations, much like  $s_n^C$ . (Also like the Weyl group of type  $C$ , we only have  $s_i^B$  for  $i \in \{1, 2, \dots, n\}$ .) Though it is not a simple transposition, the fact that  $n + 1$  is fixed by all of the other generators  $s_i^B$  means that we can define an unambiguous action on standard tableaux of shape  $(n + 1, n)$ . In this action, the map  $s_n^B$  exchanges  $n$  and  $n + 2$  and the other maps  $s_i^B$  act as the corresponding product of type- $A$   $s_i$ -local moves.



**Figure 15.** Type-B model when  $n = 2$ .

The type- $B$  involutions do not act on plane trees since plane trees must have an even number of half-edges. However, they do act on objects like plane trees that have  $n$  whole edges and an unpaired half-edge labeled  $n + 1$ . This half-edge forms a small loop or bulge between the half-edges labeled  $n$  and  $n + 2$ . Figure 15 gives an example. As with the maps  $s_i^B$  on tableaux, the action on these modified plane trees always fixes the bulge  $n + 1$ .

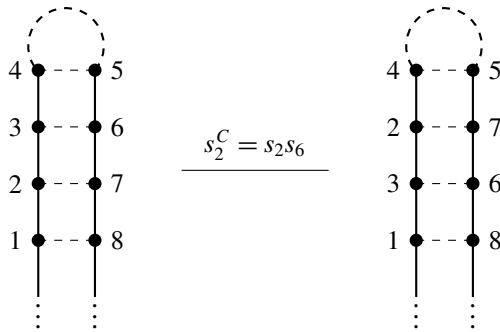
We leave these investigations for future research, for instance in the following questions.

**Question 6.1.** What are the orbits of the action of involutions  $s_i^B$ ? What is a natural collection of involutions to represent mutations on strands with several bulges (namely fixing several integers)?

*Speculative connections between Weyl groups of classical types and biology.* We extended local moves combinatorially from  $S_n$  to other Weyl groups of classical types. We end with speculative comments and questions about whether the maps we defined are observed in any biological contexts.

We begin with possible biological interpretations of type- $C$  local moves. The product of DNA transcription, messenger RNA (mRNA) carries genetic information contained in DNA from the cell nucleus to the cytoplasm, where protein synthesis takes place. During the normal process of translation, a ribosome reads an mRNA strand from the 5' end of the base sequence to the 3' end, decoding three bases into one amino acid molecule at a time. Whereas type- $A$  local moves act by twisting RNA strands at a particular location, we think of a type- $C$  local move as exchanging two triples of base pairs at some point in the translation process, a development that may completely change the sequence of amino acids.

We conjecture that the type- $C$  local moves may correspond to certain RNA mutations. When  $i = n$ , the map  $s_i^C$  replaces stacked bases with their Watson–Crick complement; otherwise, the maps  $s_i^C$  exchange adjacent sets of stacked bases while preserving their bonds. Figure 16 illustrates an example of the twisting mechanism when  $s_2^C$  is applied for  $n = 4$ . (Applying  $s_4^C$  in this example would exchange 4 and 5 instead.)



**Figure 16.** The map  $s_i^C$  for an element of the Weyl group of type  $C$  acting on RNA base pairs.

Like the Weyl group of type  $C$ , the elements of the Weyl group of type  $B$  correspond to mutations on an RNA strand. But the type- $B$  model is different because the stacked bases now contain a bulge, namely the sequence of unmatched nucleotides corresponding to the half-edge  $n + 1$ .

**Question 6.2.** Are any processes like this observed biologically?

### References

[Fulton 1997] W. Fulton, *Young tableaux*, London Mathematical Society Student Texts **35**, Cambridge Univ. Press, 1997. MR Zbl

[Fung 2003] F. Y. C. Fung, “On the topology of components of some Springer fibers and their relation to Kazhdan–Lusztig theory”, *Adv. Math.* **178**:2 (2003), 244–276. MR Zbl

[Heitsch 2006] C. E. Heitsch, “Combinatorics on plane trees, motivated by RNA secondary structure configurations”, preprint, 2006, available at <http://tinyurl.com/heitsch>.

[Housley et al. 2015] M. Housley, H. M. Russell, and J. Tymoczko, “The Robinson–Schensted correspondence and  $A_2$ -web bases”, *J. Algebraic Combin.* **42**:1 (2015), 293–329. MR Zbl

[Khovanov 2004] M. Khovanov, “Crossingless matchings and the cohomology of  $(n, n)$  Springer varieties”, *Commun. Contemp. Math.* **6**:4 (2004), 561–577. MR Zbl

[Russell 2011] H. M. Russell, “A topological construction for all two-row Springer varieties”, *Pacific J. Math.* **253**:1 (2011), 221–255. MR Zbl

[Russell and Tymoczko 2011] H. M. Russell and J. Tymoczko, “Springer representations on the Khovanov Springer varieties”, *Math. Proc. Cambridge Philos. Soc.* **151**:1 (2011), 59–81. MR Zbl

[Sagan 2001] B. E. Sagan, *The symmetric group: representations, combinatorial algorithms, and symmetric functions*, 2nd ed., Graduate Texts in Mathematics **203**, Springer, 2001. MR Zbl

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

[Vogan 1979] D. A. Vogan, Jr., “A generalized  $\tau$ -invariant for the primitive spectrum of a semisimple Lie algebra”, *Math. Ann.* **242**:3 (1979), 209–224. MR Zbl

<a href="mailto:lauraseegerer@gmail.com">lauraseegerer@gmail.com</a>	<i>Department of Mathematics and Statistics, Smith College, Northampton, MA, United States</i>
<a href="mailto:jentripp@comcast.net">jentripp@comcast.net</a>	<i>Department of Mathematics and Statistics, Smith College, Northampton, MA, United States</i>
<a href="mailto:jtymoczko@smith.edu">jtymoczko@smith.edu</a>	<i>Department of Mathematics and Statistics, Smith College, Northampton, MA, United States</i>
<a href="mailto:jwang@alumnae.smith.edu">jwang@alumnae.smith.edu</a>	<i>Department of Mathematics and Statistics, Smith College, Northampton, MA, United States</i>



# Six variations on a theme: almost planar graphs

Max Lipton, Eoin Mackall, Thomas W. Mattman, Mike Pierce,  
Samantha Robinson, Jeremy Thomas and Ilan Weinschelbaum

(Communicated by Joel Foisy)

A graph is *apex* if it can be made planar by deleting a vertex, that is, there exists  $v$  such that  $G - v$  is planar. We also define several related notions; a graph is *edge apex* if there exists  $e$  such that  $G - e$  is planar, and *contraction apex* if there exists  $e$  such that  $G/e$  is planar. Additionally we define the analogues with a universal quantifier: for all  $v$ ,  $G - v$  is planar; for all  $e$ ,  $G - e$  is planar; and for all  $e$ ,  $G/e$  is planar. The graph minor theorem of Robertson and Seymour ensures that each of these six notions gives rise to a finite set of obstruction graphs. For the three definitions with universal quantifiers we determine this set. For the remaining properties, apex, edge apex, and contraction apex, we show there are at least 36, 55, and 82 obstruction graphs respectively. We give two similar approaches to almost nonplanar (there exists  $e$  such that  $G + e$  is nonplanar, and for all  $e$ ,  $G + e$  is nonplanar) and determine the corresponding minor minimal graphs.

## 1. Introduction

Kuratowski [1930] showed that the set of planar graphs is determined by two obstructions.

**Theorem 1.1** [Kuratowski 1930; Wagner 1937]. *A graph is planar if and only if it has neither  $K_5$  nor  $K_{3,3}$  as a minor.*

We give the formulation in terms of minors due to Wagner [1937] to make the connection with Robertson and Seymour's [2004] graph minor theorem. We say  $H$  is a *minor* of graph  $G$  if it can be obtained by contracting edges in a subgraph of  $G$ . We can state the graph minor theorem as follows.

**Theorem 1.2** [Robertson and Seymour 2004]. *In any infinite set of graphs, there is a pair such that one is a minor of the other.*

---

*MSC2010:* primary 05C10; secondary 57M15.

*Keywords:* apex graphs, planar graphs, forbidden minors, obstruction set.

Research supported in part by an NSF REUT grant, as well as the Provost and Math Department of CSU, Chico.

This has two useful consequences. We say  $G$  is *minor minimal*  $\mathcal{P}$  (or  $\text{MMP}$ ) if  $G$  has property  $\mathcal{P}$  but no proper minor does.

**Corollary 1.3.** *For any graph property  $\mathcal{P}$ , there is a corresponding finite set of minor minimal  $\mathcal{P}$  graphs.*

**Corollary 1.4.** *Let  $\mathcal{P}$  be a graph property that is closed under taking minors. Then there is a finite set of minor minimal non- $\mathcal{P}$  graphs  $S$  such that for any graph  $G$ ,  $G$  satisfies  $\mathcal{P}$  if and only if  $G$  has no minor in  $S$ .*

When  $\mathcal{P}$  is minor closed, we say that  $S$  is the *Kuratowski set* for  $\mathcal{P}$ . For example,  $\{K_5, K_{3,3}\}$  is the Kuratowski set for planarity.

The graph minor theorem is not constructive, so there are only a few graph properties  $\mathcal{P}$  for which we know the finite set of  $\text{MMP}$  graphs. In particular, there are several graph properties closely related to planarity for which this set is unknown. Our goal in this paper is to investigate the minor minimal sets for the following eight graph properties.

**Definition 1.5.** A planar graph is *almost nonplanar* (AN) if there exist two nonadjacent vertices such that adding an edge between the vertices yields a nonplanar graph. A planar graph is *completely almost nonplanar* (CAN) if it is not complete and adding an edge between any pair of nonadjacent vertices yields a nonplanar graph.

Let  $G - v$  denote the graph resulting from deletion of vertex  $v$  and its edges in  $G$ , let  $G - e$  denote the graph resulting from the deletion of edge  $e$  in  $G$ , and let  $G/e$  denote the graph resulting from the contraction of edge  $e$  in  $G$ .

**Definition 1.6.** A graph is *not apex* (NA) if, for all vertices  $v$ ,  $G - v$  is nonplanar. Similarly, a graph is *not edge apex* (NE) if, for all edges  $e$ ,  $G - e$  is nonplanar and *not contraction apex* (NC) if, for all edges  $e$ ,  $G/e$  is nonplanar.

**Definition 1.7.** A graph  $G$  is *incompletely apex* (IA) if there is a vertex  $v$  such that  $G - v$  is nonplanar, *incompletely edge apex* (IE) if there is an edge  $e$  such that  $G - e$  is nonplanar, and *incompletely contraction apex* (IC) if there is an edge  $e$  such that  $G/e$  is nonplanar.

We call these last three properties “incomplete” in contrast to their negations. For example, we think of a graph as “completely” apex if  $G - v$  is planar for every vertex  $v$ . Table 1 gives a summary of our eight definitions.

We summarize our results in Table 2. Four of the properties give Kuratowski sets (as their negation generates a minor closed set) and with the exception of NA, NE, and NC, we determine the finite set of  $\text{MMP}$  graphs. For the remaining three properties we give a lower bound, which is simply the number of  $\text{MMP}$  graphs we have found, so far.

Our paper is organized as follows. Below we conclude this introduction with a survey of the literature and provide some preliminary notions used throughout the



property	definition
AN	$\exists e$ such that $G + e$ is nonplanar, where $G$ is planar
CAN	$\forall e$ , $G + e$ is nonplanar, where $G$ is planar, not complete
NA	$\forall v$ , $G - v$ is nonplanar
NE	$\forall e$ , $G - e$ is nonplanar
NC	$\forall e$ , $G/e$ is nonplanar
IA	$\exists v$ such that $G - v$ is nonplanar
IE	$\exists e$ such that $G - e$ is nonplanar
IC	$\exists e$ such that $G/e$ is nonplanar

**Table 1.** Comparison of the eight definitions.

graph property $\mathcal{P}$	AN	CAN	NA	NE	NC	IA	IE	IC
Is (not $\mathcal{P}$ ) minor closed?	no	no	yes	no	no	yes	yes	yes
number of $\text{MM}\mathcal{P}$ graphs	2	1	$\geq 36$	$\geq 55$	$\geq 82$	2	5	7

**Table 2.** Results for the eight graph properties.

paper. In Section 2 we determine the MMAN and MMCAN graphs and show that neither is a Kuratowski set. In Section 3 we give our classification of the MMIA, MMIE, and MMIC graphs, all three of which we show are Kuratowski. In Section 4 we give an overview of the MMNA graphs, which is a Kuratowski set. We classify graphs in this family of connectivity at most 1. For graphs of connectivity 2, with  $\{a, b\}$  a 2-cut, we classify those for which  $ab \in E(G)$ , as well as those for which a component of  $G - a, b$  is nonplanar. We also prove that an MMNA graph has connectivity at most 5. In total, we give explicit constructions for 36 MMNA graphs. Finally, in Section 5 we discuss MMNE and MMNC graphs, first showing these are not Kuratowski. We classify graphs of connectivity at most 1 in these two families and discuss computer searches, complete through graphs of order 9 or size 19, that yielded 55 MMNE and 82 MMNA graphs.

Apex graphs are well-studied, including results on MMNA graphs in [Ayala 2014; Barsotti and Mattman 2016; Pierce 2014]. Note that [Pierce 2014] reports on a computer search that yields 157 MMNA graphs, including all graphs through order 10 or size 21 and most of the 36 graphs we describe here. Different authors have used terms like “almost planar” or “near planar” in various ways. Here is how our definitions relate to others in the literature. Cabello and Mohar [2013] say that a graph is *near-planar* if it can be obtained from a planar graph by adding an edge. This corresponds to our definition of edge apex. Wagner [1967] defined *nearly planar* (*Fastplättbare*), which corresponds to our idea of completely apex

or not IA. Two further notions of almost planar are not directly related to the properties we have defined. For Gubser [1996], a graph  $G$  is *almost planar* if for every edge  $e$ , either  $G - e$  or  $G/e$  is planar. In characterizing graphs with no  $K^{\aleph_0}$ , Diestel, Robertson, Seymour, and Thomas say a graph  $G$  is *nearly planar* if deleting a bounded number of vertices makes  $G$  planar except for a subgraph of bounded linear width sewn onto the unique cuff of  $S^2 - 1$ ; see [Diestel 2010, Section 12.4]. Finally, our notion of CAN is also known as *maximally planar*; see [Diestel 2010].

We conclude this introductory section with some notation and definitions, as well as a lemma, used throughout. For us, graphs are simple (no loops or double edges) and undirected. We use  $V(G)$  and  $E(G)$  to denote the vertices and edges of a graph. The *order* of a graph is  $|V(G)|$  and  $|E(G)|$  is its *size*. We use  $\delta(G)$  to denote the *minimum degree* of all the vertices in  $G$ .

As mentioned earlier,  $G - v$ ,  $G - e$ , and  $G/e$  denote the results of vertex deletion, edge deletion, and edge contraction, respectively. For  $v, w \in V(G)$ , the graph  $G - v, w$  is the result of deleting two vertices and their edges. Similarly, for  $e, f \in E(G)$ , we define as  $G - e, f$  the result of deleting two edges and  $G/e, f$  the result of contracting two edges. Note that the order of deletion or contraction is arbitrary. Contracting an edge may result in a double edge. We will assume that one of the doubled edges is deleted so that  $G/e$  is again a simple graph. We use  $G_1 \sqcup G_2$  to denote the disjoint union of two graphs and  $G_1 \dot{\cup} G_2$  for the union identified on a single vertex. Similarly,  $G_1 \ddot{\cup} G_2$  denotes the union of two graphs identified on two vertices.

In light of Kuratowski's theorem, we call  $K_5$  and  $K_{3,3}$  the *Kuratowski graphs* and also refer to them as minor minimal nonplanar or MMNP. A *Kuratowski subgraph* or *K-subgraph* of  $G$  is one homeomorphic to a Kuratowski graph. A *cut set* of graph  $G$  is a set  $U \subset V(G)$  such that deleting the vertices of  $U$  and their edges results in a disconnected graph. If  $|U| = k$ , we call  $U$  a *k-cut*. We say  $G$  has *connectivity*  $k$  and write  $\kappa(G) = k$  if  $k$  is the largest integer such that  $|V(G)| > k$  and  $G$  has no  $l$ -cut for  $l < k$ . In particular,  $\kappa(K_n) = n - 1$ .

We conclude this introduction with a useful lemma. In the case that  $\kappa(G) = 2$ , we have  $G - a, b = G'_1 \sqcup G'_2$ , where  $\{a, b\}$  is a 2-cut. We will use  $G_i$  to denote the induced subgraph on  $V(G'_i) \cup \{a, b\}$ . In the literature, e.g., [Mohar and Thomassen 2001], the pair  $(G_1, G_2)$  is called a separation of order 2 (since  $|G_1 \cap G_2| = 2$ ).

**Lemma 1.8.** *If  $G$  is homeomorphic to  $K_5$  or  $K_{3,3}$  with cut set  $\{a, b\}$  such that  $G - a, b = G'_1 \sqcup G'_2$ , then one of  $G_1$  and  $G_2$  is an  $a$ - $b$ -path.*

*Proof.* Since,  $\kappa(K_5) = 4$  and  $\kappa(K_{3,3}) = 3$ ,  $G$  must be a proper subdivision of a Kuratowski graph and, since they disconnect the graph,  $a$  and  $b$  are vertices on a subdivided edge of the underlying  $K_5$  or  $K_{3,3}$ . This means that one of the components is simply an  $a$ - $b$ -path.  $\square$

## 2. Almost nonplanar: MMAN and MMCAN graphs

In this section we classify the MMAN and MMCAN graphs. Let  $K_5 - e$  denote the complete graph on five vertices with an edge deleted and  $K_{3,3} - e$  the result of deleting an edge in the complete bipartite graph  $K_{3,3}$ . The unique MMCAN graph is  $K_5 - e$  and there are two MMAN graphs,  $K_5 - e$  and  $K_{3,3} - e$ . Neither of these are Kuratowski sets, since, for example,  $K_5$  is a nonplanar graph (hence neither AN nor CAN) that contains the MMAN and MMCAN graph  $K_5 - e$  as a minor.

Our classification of the minor minimal CAN graphs makes use of a theorem due to Mader.

**Theorem 2.1** [Mader 1998]. *Any graph with  $n$  vertices and at least  $3n - 5$  edges contains a subdivision of  $K_5$ .*

In [Diestel 2010], CAN is called *maximally planar*, and it is proved equivalent to a graph admitting a plane triangulation in Proposition 4.2.8 of that text.

**Theorem 2.2.** *Every plane triangulation with at least five vertices has  $K_5 - e$  as a minor.*

*Proof.* Let  $G$  be a plane triangulation on at least five vertices. By Euler's formula,  $|E(G)| = 3(|V(G)| - 2)$ . Let  $G'$  be a nonplanar graph obtained by adding edge  $ab$  to  $G$ . Then  $|E(G')| = |E(G)| + 1 = 3|V(G)| - 5$ . By Mader's theorem  $G'$  has a subgraph  $H$  homeomorphic to  $K_5$ . Note that we must have  $ab \in E(H)$ , else  $H$  would be planar. Since  $H$  is homeomorphic to  $K_5$ , contracting appropriate edges in  $H - ab$  will result in  $K_5 - e$ , showing that  $K_5 - e$  is a minor of  $G$ .  $\square$

**Corollary 2.3.** *The only MMCAN graph is  $K_5 - e$ .*

**Theorem 2.4.** *The MMAN graphs are  $K_5 - e$  and  $K_{3,3} - e$ .*

*Proof.* First note that these two graphs are MMAN. Let  $G$  be AN and let  $ab$  be the edge that is added to form the nonplanar  $G'$ . By Kuratowski's theorem  $G'$  contains a subdivision  $H$  of  $K_5$  or  $K_{3,3}$  and  $ab \in E(H)$ . By contracting edges,  $H$  gives  $K_5 - e$  or  $K_{3,3} - e$  as a minor of  $G$ . So  $G$  is MMAN only if it is one of these two.  $\square$

## 3. Incomplete properties: MMIA, MMIE, and MMIC graphs

In this section we classify the MMIA, MMIE, and MMIC graphs. Note that each is a Kuratowski set since the corresponding "complete" property is minor closed. In the case of the IA graphs, for example, suppose  $G$  is not IA and let  $H$  be a subgraph of  $G$ . Then for any  $v \in V(H)$ , the graph  $H - v$  is planar since it is a subgraph of the planar graph  $G - v$ . Similarly if  $G$  is not IA, let  $H = G/f$  for some  $f \in E(G)$ . Then for any  $v \in V(H)$ , the graph  $H - v$  is planar since it is a minor of the planar graph  $G - v$ . This shows that the property *not* IA (also known as the completely apex property) is minor closed. Similar arguments show that *not* IE and *not* IC are also minor closed.

We next show there are exactly two MMIA graphs,  $K_1 \sqcup K_5$  and  $K_1 \sqcup K_{3,3}$ . We begin by classifying the disconnected graphs.

**Theorem 3.1.** *If  $G$  is not connected and MMIA, then  $G = K_1 \sqcup G_2$ , where  $G_2 \in \{K_5, K_{3,3}\}$ .*

*Proof.* Note that both  $K_1 \sqcup K_5$  and  $K_1 \sqcup K_{3,3}$  are MMIA. If  $G = G_1 \sqcup G_2$  is nonplanar with neither component empty, then  $K_5$ , or  $K_{3,3}$  is a minor of one of  $G_1$  and  $G_2$ . By minor minimality this means one of  $G_1$  and  $G_2$  is a Kuratowski graph, and, again by minimality, the other component can have no nontrivial proper minors, so must be simply a vertex.  $\square$

**Theorem 3.2.** *There are no connected MMIA graphs.*

*Proof.* Suppose instead that  $G$  is a connected MMIA graph. Then there is a vertex,  $v$ , such that  $G - v$  is nonplanar. However, since  $G$  is connected,  $v$  must have at least one edge,  $e$ . Since when deleting a vertex we also delete all of its edges,  $G - e$  must be a proper, nonplanar minor of  $G$ . However, deleting  $v \in V(G - e)$  is again nonplanar so that  $G - e$  is IA. This contradicts the property that  $G$  is MMIA and therefore cannot happen.  $\square$

**Corollary 3.3.** *There are two MMIA graphs:  $K_1 \sqcup K_5$  and  $K_1 \sqcup K_{3,3}$ .*

Next we show there are five MMIE graphs. We begin with the disconnected examples. Note that if  $G$  has distinct edges  $e, e'$  such that  $G - e, e'$  is nonplanar, then  $G$  is not MMIE. Indeed,  $G - e$  is an IE proper minor.

**Theorem 3.4.** *If  $G$  is not connected and MMIE, then  $G = K_2 \sqcup G_2$ , where  $G_2 \in \{K_5, K_{3,3}\}$ .*

*Proof.* The proof is similar to that of Theorem 3.1, but now the planar component is minor minimal among graphs with edges, so  $K_2$ .  $\square$

Recall that  $G_1 \dot{\cup} G_2$  denotes the union of  $G_1$  and  $G_2$  with one vertex in common.

**Theorem 3.5.** *If  $G$  is connected, MMIE, and has a cut vertex, then  $G = K_2 \dot{\cup} G_2$ , where  $G_2 \in \{K_5, K_{3,3}\}$ .*

*Proof.* Let  $G$  be a connected MMIE graph such that  $G - v = G'_1 \sqcup G'_2$ . Let  $G_i$  denote the induced subgraph on  $V(G'_i) \cup \{v\}$ . If both  $G_1$  and  $G_2$  are nonplanar, then  $G$  would not be MMIE since, for example, there are two distinct edges  $e, e' \in E(G_2)$  such that  $G - e, e'$  contains  $G_1$  and is therefore nonplanar. If both subgraphs were planar, then  $G$  would also be planar and therefore not MMIE. So one of  $G_1$  and  $G_2$  is nonplanar, say  $G_1$ , and the other,  $G_2$ , is planar.

By minor minimality of  $G$ , the nonplanar  $G_2$  is, in fact, a Kuratowski graph, and the planar  $G_1$  is minimal among graphs with edges, i.e.,  $K_2$ .  $\square$

**Theorem 3.6.** *If  $G$  is MMIE, then there is a unique edge  $e$  such that  $G - e$  is nonplanar.*

*Proof.* Assume, for the sake of contradiction, that there are  $e, e' \in E(G)$  such that  $e \neq e'$  but  $G - e$  and  $G - e'$  are nonplanar. If  $G - e$  is nonplanar, then there is a subgraph of  $G - e$ , say  $H$ , with  $e \notin E(H)$ , that has a  $K_5$  or  $K_{3,3}$  minor. Likewise, if  $G - e'$  is nonplanar, then it has a nonplanar subgraph  $H'$  with  $e' \notin E(H')$ . If  $H' = H$ , then  $e' = e$ . Otherwise,  $G - e, e'$  would be nonplanar, contradicting that  $G$  is MMIE. So  $H' \neq H$ . If  $e \notin H'$ , then  $G - e, e'$  contains  $H'$  and will be nonplanar, contradicting that  $G$  is MMIE.

So,  $e \in H'$  and, similarly,  $e' \in H$ . If  $H$  and  $H'$  have empty intersection, then let  $e_1, e_2 \in E(H')$ . This means  $G - e_1, e_2$  contains  $H$  and is nonplanar. This contradicts that  $G$  is MMIE. So,  $H$  and  $H'$  have nonempty intersection. If their intersection is nonplanar, then removing  $e$  and  $e'$  will not change this intersection, and  $G$  is not MMIE. If their intersection is planar, then there must be more than one edge in  $H'$  that is not in  $H$  besides  $e$ . But, if  $H'$  has more edges besides  $e$  that are not in  $H$  it would be possible to remove another edge,  $f \neq e$ , without changing  $H$ . This means that  $G - f, e$  is nonplanar, and contradicts that  $G$  is MMIE.

Therefore, if  $G$  is MMIE, then there is a unique edge  $e$  such that  $G - e$  is nonplanar.  $\square$

Recall that a  $K$ -subgraph is one homeomorphic to  $K_5$  or  $K_{3,3}$ .

**Theorem 3.7.** *If  $G$  is MMIE, then the edge  $e$  such that  $G - e$  is nonplanar is not in a  $K$ -subgraph. Furthermore,  $G - e$  is  $K_5$  or  $K_{3,3}$ .*

*Proof.* Assume, for the sake of contradiction, that  $e$  is in a  $K$ -subgraph,  $H$ . Since no graph homeomorphic to  $K_5$  or  $K_{3,3}$  is IE,  $G - e$  is planar unless  $G - e$  contains some other  $K$ -subgraph,  $H'$ . However, if  $G$  contains two  $K$ -subgraphs  $H$  and  $H'$  with empty intersection, then  $G - e$  will leave  $H'$  unchanged. One could then remove a second edge,  $f \in E(H)$ , leaving  $H'$  unchanged so that  $G - e, f$  is nonplanar. This means that  $G$  cannot be MMIE since  $G$  would have an IE minor  $G - e$ . So,  $H$  and  $H'$  have nonempty intersection. But  $H \neq H'$  since  $e$  cannot be an edge in the only  $K$ -subgraph, otherwise  $G - e$  is planar.

Next, observe that any proper subgraph of a  $K$ -subgraph is planar. This means that for the  $K$ -subgraph,  $H'$ , with  $H \neq H'$ , there must be an edge,  $g \neq e$ , with  $g \in E(H')$  and  $g \notin E(H)$ . Then  $G - g$  contains  $H$  and is nonplanar. This contradicts the uniqueness of the edge  $e$  and shows  $e$  is not in a  $K$ -subgraph.

Following the same argument as above,  $G$  cannot contain more than one  $K$ -subgraph. Indeed, if there were distinct  $K$ -subgraphs  $H$  and  $H'$ , then either the intersection is empty or it is not, and we achieve similar contradictions as in the previous argument. So,  $G$  contains exactly one  $K$ -subgraph.

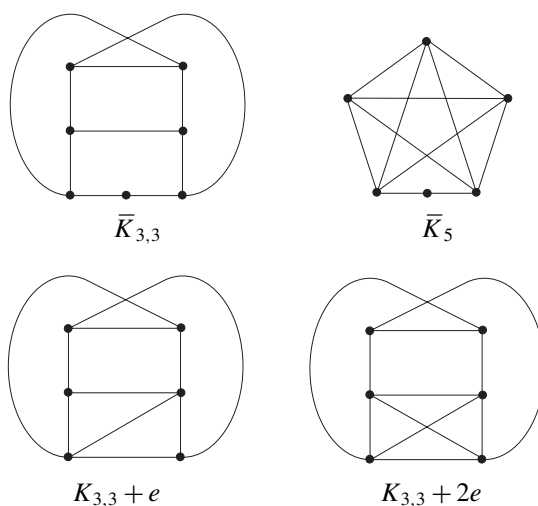
Finally, the only possible K-subgraph contained in  $G$ , call it  $N$ , must contain all edges besides  $e$ . If not, then there is an edge  $e' \neq e$  such that  $G - e'$  is nonplanar. This contradicts the uniqueness of  $e$ . Also, the K-subgraph  $N$  in  $G - e$  must be either  $K_5$  or  $K_{3,3}$ . If not, then  $N$  would be a subdivision of either  $K_5$  or  $K_{3,3}$ . But, then there is a proper minor,  $G'$ , of  $G$ , by contracting an edge,  $e_1 \in E(N)$ , which contains a K-subgraph as well. Provided  $e$  remains as an edge of  $G'$ , the graph  $G' - e$  is nonplanar, contradicting that  $G$  is minor minimal. On the other hand, if contracting  $e_1$  removes  $e$ , then there must be another edge  $e_2$  incident to  $e_1$ , with  $e_2 \in E(N)$ , such that  $e$  is incident to both  $e_1$  and  $e_2$ . Since  $N$  is a subdivision of  $K_5$  or  $K_{3,3}$  and  $G/e_1$  is nonplanar,  $e_1$  and  $e_2$  must be in a path of  $N$  formed by subdividing an edge of the underlying Kuratowski graph. Since  $e$  is incident to both  $e_1$  and  $e_2$ , there exists  $N'$ , another K-subgraph of  $G$  with  $e \in E(N')$ . This contradicts that there is only one K-subgraph of  $G$ .

So, if  $G$  is MMIE then it is made up of either  $K_5$  or  $K_{3,3}$  and an edge that is not in this K-subgraph. □

Aside from the disconnected and connectivity-1 examples above, a final way to add an edge to a K-subgraph is the graph  $K_{3,3} + e$  of Figure 1, formed by adding an edge to the bipartite graph  $K_{3,3}$ .

**Corollary 3.8.** *There are five MMIE graphs:  $K_{3,3} + e$  and  $K_2 \sqcup G_2$ ,  $K_2 \dot{\cup} G_2$ , where  $G_2 \in \{K_5, K_{3,3}\}$ .*

Let  $\bar{K}_5$  and  $\bar{K}_{3,3}$  denote the graphs obtained from  $K_5$  and  $K_{3,3}$  by subdividing a single edge, as in Figure 1. We denote as  $K_{3,3} + 2e$  the graph given by adding two edges to  $K_{3,3}$ , as in Figure 1.



**Figure 1.** MMIE and MMIC graphs.

**Theorem 3.9.** *There are seven MMIC graphs:  $K_{3,3} + 2e$  and  $\bar{K}$ ,  $K_2 \sqcup K$ , and  $K_2 \dot{\cup} K$  with  $K \in \{K_5, K_{3,3}\}$ .*

*Proof.* Observe that these seven graphs are MMIC. If  $G$  is MMIC and disconnected, then  $G$  is  $K_2 \sqcup K$ , with  $K$  a Kuratowski graph. We omit the proof, which is similar to that for MMIE. Note that the remaining five graphs are precisely the graphs that result when a vertex of a Kuratowski graph is split.

Suppose  $G$  is MMIC and connected. Then there is an edge  $e$  such that  $G/e$  is nonplanar. Since contracting an edge will not disconnect the graph,  $G/e$  is also connected and has a  $K$ -subgraph  $H$ . If  $H$  is not a Kuratowski graph, then it has  $\bar{K}_5$  or  $\bar{K}_{3,3}$  as a minor, contradicting  $G$  being minor minimal. Therefore,  $H$  is Kuratowski.

If  $V(H) \neq V(G/e)$ , then since  $G/e$  is connected, considering any vertex in  $G/e$  beyond those in  $H$ , along with one of its edges, shows that  $G/e$  contains  $K_2 \sqcup K$  or  $K_2 \dot{\cup} K$ , with  $K$  Kuratowski, contradicting  $G$  being minor minimal. So,  $V(H) = V(G/e)$ .

Now  $G$  is obtained from  $G/e$  by a vertex split. The corresponding vertex split on  $H$  gives rise to a graph  $H'$ , which is one of the five graphs  $K_{3,3} + 2e$ ,  $\bar{K}$ , or  $K_2 \dot{\cup} K$ . Since  $G$  is minor minimal,  $G = H'$  and is one of these five, and hence one of the seven.  $\square$

#### 4. MMNA graphs

In this section we describe several partial results toward a classification of the MMNA graphs, with a focus on graph connectivity. In all, we describe 36 MMNA graphs, including all those of connectivity at most 1 ( $\kappa(G) \leq 1$ ). For graphs with  $\kappa(G) = 2$ , where  $\{a, b\}$  is a 2-cut, we classify the MMNA graphs having  $ab \in E(G)$ , as well as those for which a component of  $G - a, b$  is nonplanar. We also show that  $\kappa(G) \leq 5$  for MMNA graphs, which is a sharp bound. Since the family of apex graphs is minor closed, the MMNA graphs are a Kuratowski set.

We first bound the minimum degree,  $\delta(G)$ , of an MMNA graph and then classify the examples with  $\kappa(G) \leq 1$ .

**Theorem 4.1.** *The minimum vertex degree in an MMNA graph is at least 3.*

*Proof.* The addition or deletion of an isolated vertex or vertex of degree 1 in a planar graph will again result in a planar graph. Similarly, contracting an edge adjacent to a degree-2 vertex will not affect planarity. So if  $G$  is NA with  $\delta(G) < 3$ , then removing a vertex of small degree will result in a NA graph; hence  $G$  is not MMNA.  $\square$

**Theorem 4.2.** *There are three disconnected MMNA graphs:  $K_5 \sqcup K_5$ ,  $K_5 \sqcup K_{3,3}$ , and  $K_{3,3} \sqcup K_{3,3}$ .*

*Proof.* First observe that these three graphs are all MMNA. On the other hand, if  $G = G_1 \sqcup G_2$  is MMNA, both components must be nonplanar. Otherwise if  $G_1$  is planar, then  $G_2$  must be NA and is a proper minor of  $G$ , contradicting  $G$  being MMNA. So each component  $G_i$  has a  $K_5$  or  $K_{3,3}$  minor and  $G$  has one of the three candidates as a minor. Since  $G$  is minor minimal, it must be one of the three candidates.  $\square$

**Theorem 4.3.** *There are no MMNA graphs of connectivity 1.*

*Proof.* Suppose instead  $G$  is MMNA with cut vertex  $a$ . Then  $G - a = G'_1 \sqcup G'_2$ . If both  $G'_1$  and  $G'_2$  are planar, then  $G - a$  is planar, contradicting that  $G$  is NA. If both are nonplanar, then  $G$  has one of the disconnected MMNA graphs as a proper minor and is not minor minimal. So, one of  $G'_1$  and  $G'_2$ , say  $G'_1$ , is planar, and the other,  $G'_2$ , is not. Let  $G_i$  denote the induced graph on  $V(G'_i) \cup \{a\}$ . If  $G_1$  is nonplanar, then together with  $G'_2$  this gives one of the three disconnected MMNA graphs as a proper minor of  $G$ , contradicting that  $G$  is minor minimal. So  $G_1$  is planar. But then  $G_2$  must be NA, which again contradicts  $G$  being minor minimal.  $\square$

We can also give an upper bound on the connectivity of an MMNA graph. We first bound the minimum degree  $\delta(G)$ .

**Theorem 4.4.** *If  $G$  is MMNA, then  $\delta(G) \leq 5$ .*

*Proof.* Suppose  $G$  is MMNA and, for a contradiction, that  $\delta(G) \geq 6$ . Let  $D$  be the largest integer so that there are two vertices  $a, b \in V(G)$  both of degree at least  $D$ . Surely,  $D \geq 6$ . We will argue that there are two vertices with degree at least  $D + 2$ , contradicting our choice of  $D$ . Let  $v = |V(G)|$  be the number of vertices of  $G$ . There will be  $v - 2$  vertices of degree at least 6 and two vertices of degree at least  $D$ . A lower bound on the number of edges of  $G$  is then  $(6(v - 2) + 2D)/2 = 3v - 6 + D$ .

Since  $G$  is MMNA, we can form a planar graph by deleting an edge (to get a proper minor) and then an apex vertex, which is not adjacent to the deleted edge. For if it were adjacent to the edge, the vertex deletion would also remove the edge, making  $G$  apex, a contradiction.

After deleting an edge,  $G - e$  has at least  $3v - 7 + D$  edges. Next delete a vertex,  $a \in V(G)$  of degree  $d$ . Then the lower bound on the number of edges in the resulting planar graph is  $3v - 7 + D - d$ . As this graph is planar on  $v - 1$  vertices, an upper bound on the number of edges is  $3(v - 1) - 6$ , the number of edges in a triangulation. Thus  $3v - 7 + D - d \leq 3(v - 1) - 6$ , which implies  $d \geq D + 2$ .

This means the degree of  $a$  is at least  $D + 2$ . However, following the argument above, if we first delete an edge incident to  $a$ , we deduce that there is a second vertex  $b$  that is again of degree at least  $D + 2$ . This is a contradiction since  $D$  was assumed to be the maximum such that two vertices have degree at least  $D$ . Therefore, if  $\delta(G) \geq 6$ , then  $G$  is not MMNA.  $\square$



Since  $\kappa(G) \leq \delta(G)$ , we have a bound on connectivity as an immediate corollary.

**Corollary 4.5.** *If  $G$  is MMNA, then  $\kappa(G) \leq 5$ .*

Note that  $K_6$  is an MMNA graph of connectivity 5, so this bound is sharp. Indeed,  $K_6$  is part of the Petersen family, a family of seven graphs shown to be MMNA by Barsotti and Mattman [2016]. Other graphs in this family provide examples of graphs of connectivity 4 ( $K_{3,3,1}$ ) and connectivity 3 ( $K_{4,4} - e$  and the Petersen graph) and the computer search of [Pierce 2014] unearthed numerous further examples with connectivity greater than 2.

Nonetheless, in the remainder of this section, we restrict attention to MMNA graphs of connectivity 2. Let us fix some notation for this situation. For  $G$  MMNA with cut set  $\{a, b\}$ , we have  $G - a, b = G'_1 \sqcup G'_2$ . Let  $G_i$  denote the induced subgraph on  $V(G'_i) \cup \{a, b\}$  so that  $(G_1, G_2)$  is a separation of order 2.

**Theorem 4.6.** *Let  $G$  be an MMNA graph where  $\kappa(G) = 2$ , with cut set  $\{a, b\}$ . If  $G - a, b = G'_1 \sqcup G'_2$ , then  $G'_1$  and  $G'_2$  are not both nonplanar.*

*Proof.* Let  $c_a$  be an apex of  $G - a$ . By the assumption that  $G$  is MMNA,  $G - a, c_a$  is planar. If  $c_a = b$ , we are done because  $G'_1 \sqcup G'_2 = G - a, b = G - a, c_a$ , which would imply both  $G'_1$  and  $G'_2$  are planar.

Without loss of generality, assume  $c_a \in V(G'_1)$ . Since none of the edges of  $G'_2$  are in  $G'_1$  and  $a, c_a \notin V(G'_2)$ , it follows that  $G'_2$  is a subgraph of the planar graph  $G - a, c_a$ . Thus,  $G'_2$  is planar.  $\square$

**Theorem 4.7.** *If  $G$  is MMNA and  $\kappa(G) = 2$  such that  $G - a, b = G'_1 \sqcup G'_2$ , then, up to relabeling,  $G'_1 + a, G'_1 + b$  are planar, and  $G'_2 + a, G'_2 + b$  are nonplanar.*

We prove this with two lemmas.

**Lemma 4.8.**  *$G'_1 + a$  and  $G'_2 + a$  cannot both be planar.*

*Proof.* Let  $G$  be as described. Suppose both  $G'_1 + a$  and  $G'_2 + a$  are planar. Since  $G'_1$  and  $G'_2$  are otherwise disjoint,  $G - b = (G'_1 + a) \cup (G'_2 + a)$  is the union of two planar graphs at only one vertex, with no new edges. Thus,  $G - b$  is planar, which is a contradiction. So it cannot be that both  $G'_1 + a$  and  $G'_2 + a$  are planar. A similar argument could be made for  $b$ .  $\square$

**Lemma 4.9.**  *$G'_1 + a$  and  $G'_2 + b$  cannot both be nonplanar (up to relabeling).*

*Proof.* Let  $G$  be as described. Suppose both  $G'_1 + a$  and  $G'_2 + b$  are nonplanar. Let  $e$  be an edge between a vertex in  $G'_1$  and the vertex  $b$ . Since  $G$  is MMNA,  $G - e$  is apex. So there is a vertex  $v$  such that  $(G - e) - v$  is planar. If  $v = a$  then  $G'_2 + b$  is a subgraph of  $(G - e) - v$ , which is a contradiction since  $G'_2 + b$  is nonplanar. If  $v \in V(G'_1)$  then again  $G'_2 + b$  is a subgraph of  $(G - e) - v$ , which is a contradiction since  $G'_2 + b$  is nonplanar. If  $v = b$  then  $(G - e) - v = G - v$ , which implies  $(G - e) - v$  is nonplanar since  $G$  is NA, so this is a contradiction. If  $v \in V(G'_2)$

then  $G'_1 + a$  is a subgraph of  $(G - e) - v$ , which is a contradiction since  $G'_1 + a$  is nonplanar. Therefore there is no apex for  $G - e$ , which is a contradiction. So our assumption was wrong and one of  $G'_1 + a$  and  $G'_2 + b$  must be planar.  $\square$

We can now prove Theorem 4.7.

*Proof of Theorem 4.7.* Let  $G$  be as described. By the first lemma we know that at least one of  $G'_1 + a$  and  $G'_2 + a$  must be nonplanar. Without loss of generality suppose  $G'_2 + a$  is nonplanar. Since  $G'_2 + a$  is nonplanar, we know that  $G'_1 + b$  must be planar by the second lemma. Since  $G'_1 + b$  is planar, by the first lemma we know that  $G'_2 + b$  is nonplanar. By the second lemma this implies that  $G'_1 + a$  must be planar. Therefore, up to relabeling,  $G'_1 + a$  and  $G'_1 + b$  are both planar, and  $G'_2 + a$  and  $G'_2 + b$  are both nonplanar.  $\square$

Going forward, we adopt the convention suggested by Theorem 4.7 and label  $G'_1$  and  $G'_2$  such that  $G'_1 + a$ ,  $G'_1 + b$  are planar and  $G'_2 + a$ ,  $G'_2 + b$  are not. Let  $G$  be MMNA with cut set  $\{a, b\}$ . Our next goal is to classify such graphs in the case that  $ab$  is an edge of the graph.

**Theorem 4.10.** *If  $G$  is MMNA and  $\kappa(G) = 2$  with cut set  $\{a, b\}$  such that  $ab \in E(G)$ , then  $G_1$  and  $G_2$  are nonplanar.*

*Proof.* Let  $G_i$  denote the induced subgraph on  $V(G'_i) \cup \{a, b\}$ . By Theorem 4.7,  $G_2$  is nonplanar. For the sake of contradiction, assume  $G_1$  is planar. Since  $G_2$  is a proper subgraph of  $G$ , there is a vertex  $v \in V(G_2)$  such that  $G_2 - v$  is planar. But this means  $G - v$  is planar and contradicts that  $G$  is NA.

So if  $G$  is MMNA with cut set  $\{a, b\} \subset V(G)$  such that  $ab \in E(G)$ , then  $G_1$  and  $G_2$  are nonplanar.  $\square$

**Theorem 4.11.** *If  $G$  is MMNA and  $\kappa(G) = 2$  with cut set  $\{a, b\}$  such that  $ab \in E(G)$ , then  $G'_1$  and  $G'_2$  are both planar.*

*Proof.* By Theorem 4.10,  $G_1$  is nonplanar. By Theorem 4.6, without loss of generality,  $G'_1$  is planar. Suppose  $G'_2$  is nonplanar. Then  $G_1 \sqcup G'_2$  is a proper subgraph of  $G$ . Since  $G_1$  and  $G'_2$  are both nonplanar,  $G_1 \sqcup G'_2$  has a disconnected MMNA minor, contradicting that  $G$  is minor minimal.  $\square$

**Theorem 4.12.** *If  $G$  is MMNA with cut set  $\{a, b\}$  such that  $ab \in E(G)$ , then  $G_1 \in \{K_5, K_{3,3}\}$ .*

*Proof.* First observe that for any  $e \in E(G_1)$ , the graph  $G_1 - e$  must be planar. Suppose instead that there is  $e' \in E(G_1)$  such that  $G_1 - e'$  is nonplanar. Since  $G - e'$  is apex, there is a vertex  $v \in V(G)$  such that  $(G - e') - v$  is planar. However,  $v \notin \{a, b\}$  since  $G'_2 + a$  and  $G'_2 + b$  are nonplanar by Theorem 4.7. If  $v \in V(G'_1)$ , then  $G_2$  is a subgraph of  $(G - e') - v$ . By Theorem 4.10, since  $G_2$  is nonplanar,  $(G - e') - v$  is also nonplanar. If  $v \in V(G'_2)$ , then  $G_1 - e'$  is a subgraph of

$(G - e') - v$ , and since  $G_1 - e'$  is nonplanar,  $(G - e') - v$  is nonplanar. So we have a contradiction and deduce that for all  $e \in E(G_1)$ , the graph  $G_1 - e$  must be planar.

Since  $G_1$  is nonplanar by Theorem 4.10, and since  $G_1 - e$  is planar for all  $e \in G_1$ , it follows that  $G_1$  consists of a  $K$ -subgraph along with some number (possibly zero) of isolated vertices. However, if  $G_1$  is anything other than  $K_5$  or  $K_{3,3}$ , then  $G_1$  has a proper minor  $N \in \{K_5, K_{3,3}\}$  formed by deleting isolated vertices or contracting edges in the  $K$ -subgraph. Then  $G$  has a proper minor  $G'$  such that  $N$  is a subgraph of  $G'$ . Since  $G$  is MMNA, there exists vertex  $v \in V(G')$  that is an apex. Since  $N$  and  $G_2$  are subgraphs of  $G'$  and both  $N$  and  $G_2$  are nonplanar, we have that  $v \in V(N) \cap V(G_2) \subset \{a, b\}$ . However,  $G_2 - a = G'_2 + b$  and  $G_2 - b = G'_2 + a$  are both nonplanar (Theorem 4.7) and therefore  $G$  has a proper NA minor. This contradicts  $G$  being minor minimal.

Therefore if  $G$  is MMNA with cut set  $\{a, b\}$  such that  $ab \in E(G)$ , then  $G_1 \in \{K_5, K_{3,3}\}$ .  $\square$

**Theorem 4.13.** *If  $G$  is MMNA with cut set  $\{a, b\}$  such that  $ab \in E(G)$ , then there is a vertex  $c \in V(G'_2)$  such that every  $a$ - $b$ -path in  $G_2 - ab$  passes through  $c$ .*

*Proof.* Assume for the sake of contradiction that there is no such vertex  $c$ . Since  $G$  is MMNA,  $G - ab$  must have some apex  $v$ . If  $v \in \{a, b\}$ , then  $(G - ab) - v = G - v$ . This would mean that  $G$  has an apex, and contradicts that  $G$  is NA. If  $v \in V(G'_1)$ , then  $(G - ab) - v$  is nonplanar as it contains  $G'_2 + a$ , which is nonplanar by Theorem 4.7. So it must be that  $v \in V(G'_2)$ . Then  $G_1 - ab$  is a subgraph of  $(G - ab) - v$ . Note that  $G_1 - ab \in \{K_5 - e, K_{3,3} - e\}$  since  $G_1 \in \{K_5, K_{3,3}\}$  by Theorem 4.12.

Since there is no  $c$  vertex as described in the statement of the theorem, there remains an  $a$ - $b$ -path in  $(G_2 - ab) - v$ . Together with  $G_1 - ab$ , this constitutes a nonplanar subgraph of  $(G - ab) - v$ , contradicting the definition of  $v$  as an apex for  $G - ab$ . Thus, if  $G$  is MMNA with  $ab \in E(G)$ , then there is a vertex  $c$  such that every  $a$ - $b$ -path of  $G_2 - ab$  passes through  $c$ .  $\square$

**Theorem 4.14.** *Let  $G$  be MMNA with cut set  $\{a, b\}$  and  $ab \in E(G)$  and let  $c \in V(G_2)$  be such that every  $a$ - $b$ -path of  $G_2 - ab$  passes through  $c$ . Then  $\{a, c\}$  and  $\{b, c\}$  are also cut sets.*

*Proof.* First we show there exists some  $v_2 \in V(G'_2)$  such that  $v_2 \neq c$ , but  $v_2$  is adjacent to  $a$ . Suppose instead that  $c$  is the only vertex in  $G'_2$  adjacent to  $a$ . Since  $G'_2$  is planar by Theorem 4.11, and since  $G'_2 + a$  has only one more edge than  $G'_2$ ,  $G'_2 + a$  is also planar. However, this contradicts Theorem 4.7, where  $G'_2 + a$  is shown to be nonplanar.

So let  $v_2$  be a vertex of  $G'_2$  that is adjacent to  $a$ , but is not  $c$ , and take  $v_1 \in V(G'_1)$ . We demonstrate there is no  $v_1$ - $v_2$ -path in  $G - a, c$ . Since any path from a vertex in  $G'_1$  to a vertex in  $G'_2$  must pass through  $a$  or  $b$  by assumption, the supposed path

from  $v_1$  to  $v_2$  must pass through  $b$ , since  $a$  has been deleted. However, there cannot be a path from  $b$  to  $v_2$  that does not pass through  $c$ . Otherwise we would be able to find a path from  $b$  to  $v_2$  and finally to  $a$  without passing through  $c$ , violating our assumption on  $c$ . We conclude that  $G - a, c$  is disconnected. By an analogous argument,  $\{b, c\}$  is also a cut set for  $G$ .  $\square$

In order to classify connectivity-2 MMNA graphs with  $ab \in E(G)$ , we need to describe  $G_1$  in the case that  $ab \notin E(G)$ .

**Theorem 4.15.** *If  $G$  is MMNA with cut set  $\{a, b\}$  such that  $ab \notin E(G)$ , then  $G_1 \in \{K_5 - e, K_{3,3} - e, K_{3,3}\}$  and  $G_1 + ab$  is nonplanar.*

*Proof.* Let  $G - a, b = G'_1 \sqcup G'_2$  and let  $G_i$  denote the subgraph induced by vertices  $V(G'_i) \cup \{a, b\}$ . If  $G_1$  is nonplanar, then  $G_1$  has a K-subgraph  $N$ . Form a new graph,  $H$ , by replacing  $G_1$  with  $N$ . It is clear that  $a, b \in V(N)$  because if not, then  $G$  contains two disjoint K-subgraphs ( $G'_2 + a$  and  $G'_2 + b$  are nonplanar, Theorem 4.7) and therefore has a proper MMNA minor.

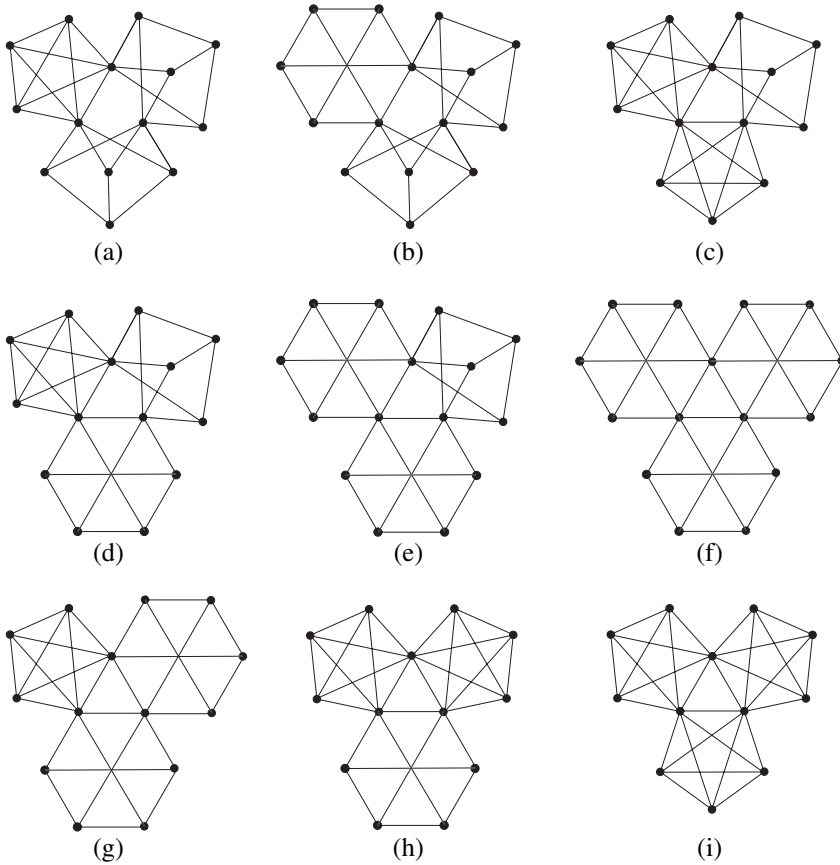
We can see that  $H$  is NA. Take  $v \in V(H)$ . If  $v \in V(N - a, b)$ , then  $G'_2 + a$  is a subgraph of  $H - v$  so  $H - v$  is nonplanar. If  $v \in V(G'_2)$ , then  $N$  is a subgraph of  $H - v$  so  $H - v$  is nonplanar. And if  $v \in \{a, b\}$ , then either  $G'_2 + a$  or  $G'_2 + b$  is a subgraph of  $H - v$  and therefore  $H - v$  is nonplanar. Thus,  $H$  is NA. Since  $G$  is minor minimal,  $G_1 = N$ . As  $G$  is MMNA it has no degree-2 vertices and since  $ab \notin E(G)$ , we have  $G_1 = K_{3,3}$  in this case.

Suppose next that  $G_1$  is planar. Assume for the sake of contradiction  $G_1 + ab$  is planar and replace  $G_1$  with the edge  $ab$  to form a new graph  $H'$ . Equivalently,  $H' = G_2 + ab$ . We observe that for every  $v \in V(H')$ , the graph  $H' - v$  is nonplanar. If  $v \in \{a, b\}$ , then  $G'_2 + a$  or  $G'_2 + b$  is a subgraph of  $H' - v$ , which is then nonplanar. On the other hand if  $v \in V(G'_2)$ , then since  $G$  is NA,  $G - v$  has a K-subgraph  $M$ . However, if  $|\{a, b\} \cap V(M)| < 2$ , then since  $G_1$  is planar,  $M$  lies wholly in  $G_2$  and we may delete  $G'_1$  without changing  $M$ . That is,  $M$  is a subgraph of  $H' - v$ . If  $|\{a, b\} \cap V(M)| = 2$ , then by Lemma 1.8,  $a$  and  $b$  are vertices in a path of  $M$ . Since  $G_1 + ab$  is planar, we may replace  $G_1$  by  $ab$  to create a new K-subgraph  $B$  in  $H' - v$ . Therefore  $H'$  is NA. However, as  $H'$  is a proper minor of  $G$ , this is a contradiction. We conclude  $G_1 + ab$  is nonplanar.

Finally, observe that  $G_1 + ab$  is a K-subgraph. Otherwise, we may replace it with a K-subgraph contained in  $G_1 + ab$  to get a proper minor of  $G$  that is NA. Since an MMNA graph cannot have vertices of degree 2 or less,  $G_1 + ab \in \{K_5, K_{3,3}\}$ .

This shows if  $G$  is MMNA with cut set  $\{a, b\}$  such that  $ab \notin E(G)$ , then we have  $G_1 \in \{K_5 - e, K_{3,3} - e, K_{3,3}\}$ .  $\square$

**Theorem 4.16.** *If  $G$  is MMNA,  $\kappa(G) = 2$  with cut set  $\{a, b\}$ , and  $ab \in E(G)$ , then  $G$  is one of the nine graphs shown in Figure 2.*

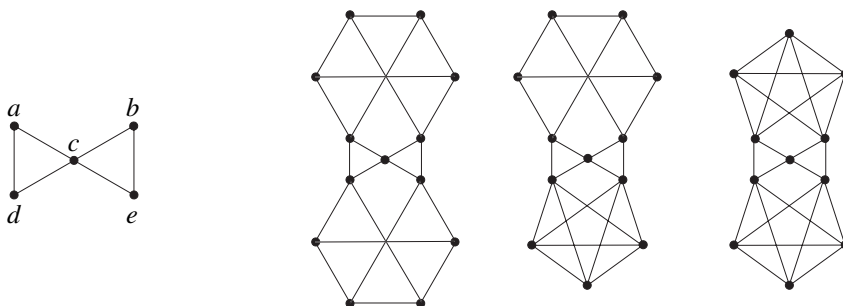


**Figure 2.** The nine MMNA graphs with  $ab \in E(G)$ .

*Proof.* It is straightforward to verify that the nine graphs are MMNA. Let  $G$  be MMNA,  $\kappa(G) = 2$  with cut set  $\{a, b\}$ , and  $ab \in E(G)$ . By Theorems 4.13 and 4.14, there exists a vertex  $c$  such that  $\{a, c\}$  and  $\{b, c\}$  are also 2-cuts for  $G$ . Let  $H_1$  play the role of  $G_1$  for the  $\{a, c\}$  cut set. That is,  $G - a, c = H'_1 \sqcup J'_1$  with  $H'_1 + a$  and  $H'_1 + c$  planar (see Theorem 4.7). Similarly, let  $H_2$  be the  $G_1$  for the  $\{b, c\}$  cut set. By Theorem 4.12,  $G_1 \in \{K_{3,3}, K_5\}$  and by that theorem and Theorem 4.15,  $H_i \in \{K_{3,3}, K_{3,3} - e, K_5, K_5 - e\}$ .

Note that, if  $H_1$  is  $K_{3,3} - e$  or  $K_5 - e$ , then  $G - b$  is planar and similarly for  $H_2$ . Thus,  $H_1, H_2 \in \{K_{3,3}, K_5\}$ . There are three cases depending on whether  $ac, bc \in E(G)$  or not.

First suppose that  $ab$  is the only one of  $ab, bc$ , and  $ac$  present in the graph. As above,  $G_1, H_1$  and  $H_2$  are each either  $K_{3,3}$  or  $K_5$ . However, by Theorem 4.15, this means  $H_1 = H_2 = K_{3,3}$ . So, there are exactly two graphs (graphs (a) and (b) in Figure 2) of this type, depending on whether  $G_1$  is  $K_5$  or  $K_{3,3}$ .



**Figure 3.** Bowtie graphs.

Next suppose that exactly one of  $ac$  and  $bc$ , say  $ac$ , is in the graph. As in the previous case  $H_2$  must be  $K_{3,3}$ . There are three graphs (graphs (c), (d), and (e) of Figure 2) of this type as  $\{G_1, H_1\}$  is either  $\{K_5, K_5\}$ ,  $\{K_5, K_{3,3}\}$ , or  $\{K_{3,3}, K_{3,3}\}$ .

Finally, suppose all three edges  $ab, ac$  and  $bc$  are in the graph. Then, as above,  $G_1, H_1$ , and  $H_2$  are each either  $K_{3,3}$ , or  $K_5$ . There are four graphs of this type, shown as graphs (f) through (i) of Figure 2. For example, such a graph has between zero and three  $K_5$ 's.

This shows that the nine graphs of Figure 2 are the graphs where  $G$  is MMNA,  $\kappa(G) = 2$  with cut set  $\{a, b\}$ , and  $ab \in E(G)$ . □

Henceforth, we can assume  $ab \notin E(G)$ . By Theorem 4.15, this means  $G_1 \in \{K_5 - e, K_{3,3} - e, K_{3,3}\}$ . We will say that  $G$  is a *bowtie* if the neighborhood of  $a, b$  in  $G_2$  is as shown in Figure 3 (left). That is,  $a$  and  $b$  have degree 2 in  $G_2$  and  $c$  has degree 4. Although  $d$  and  $e$  have additional neighbors in  $G_2$  besides  $\{a, c\}$  and  $\{b, c\}$  respectively,  $de \notin E(G_2)$ .

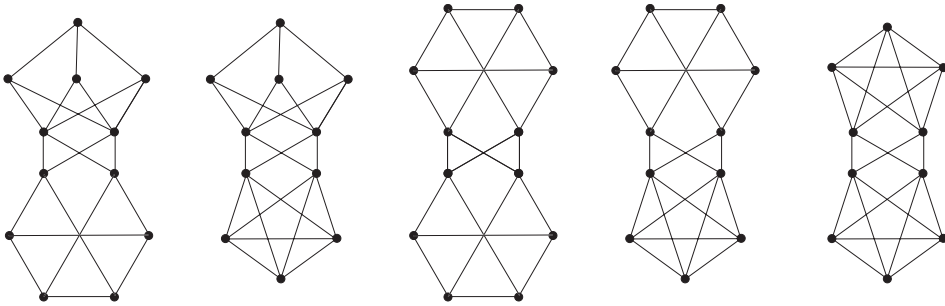
**Theorem 4.17.** *If  $G$  is a bowtie MMNA graph, then  $G$  is one of the three graphs shown in Figure 3 (right).*

*Proof.* It is straightforward to verify that the three graphs in the figure are MMNA. Let  $G$  be a bowtie MMNA graph. Then, referring to Figure 3 (left),  $\{d, e\}$  is a cut set as well. Let  $H_1$  play the role of the  $G_1$  for the  $\{d, e\}$  cut set. By Theorem 4.15,  $G_1$  and  $H_1$  are both drawn from  $\{K_{3,3}, K_{3,3} - e, K_5 - e\}$ .

We will argue that neither  $G_1$  nor  $H_1$  is  $K_{3,3}$ . For the sake of contradiction, assume instead  $G_1 = K_{3,3}$ . Notice  $G_1$  and  $G'_2$  are disjoint, and nonplanar. So,  $G$  has a proper NA minor,  $G_1 \sqcup G'_2$ , which contradicts that  $G$  is to be minor minimal.

So,  $G_1$  and  $H_1$  are both in  $\{K_{3,3} - e, K_5 - e\}$ , where  $ab$  is the missing edge,  $e$ , and the only possibilities are the three graphs shown in Figure 3 (right). □

Let  $G$  be MMNA with cut set  $\{a, b\}$  such that  $ab \notin E(G)$ . We say  $G$  is of  $(2, 2, c)$  type if, in  $G_2$ , the vertices  $a$  and  $b$  are of degree 2 and have  $c$  common neighbors. For example, a bowtie graph is of  $(2, 2, 1)$  type.



**Figure 4.** Graphs of type  $(2, 2, 2)$ .

**Theorem 4.18.** *If  $G$  is MMNA and of  $(2, 2, 2)$  type, then  $G$  is one of the five graphs shown in Figure 4.*

*Proof.* It is straightforward to verify that the five graphs are MMNA. Let  $G$  be MMNA with cut set  $\{a, b\}$  and of  $(2, 2, 2)$  type. Let  $\{c, d\}$  be the common neighbors of  $a$  and  $b$  in  $G_2$ . Note that  $cd \notin E(G)$ , as otherwise  $G$  must be one of the nine graphs of Theorem 4.16 and none of those are  $(2, 2, 2)$  type.

By Theorem 4.15, and using symmetry,  $G_1, G'_2 \in \{K_{3,3}, K_{3,3} - e, K_5 - e\}$ . However, they cannot both be  $K_{3,3}$ , as otherwise  $G_1 \sqcup G'_2$  is a proper NA subgraph, which contradicts that  $G$  is minor minimal. So at most one of the subgraphs can be  $K_{3,3}$ . This leaves the five possibilities shown in Figure 4. □

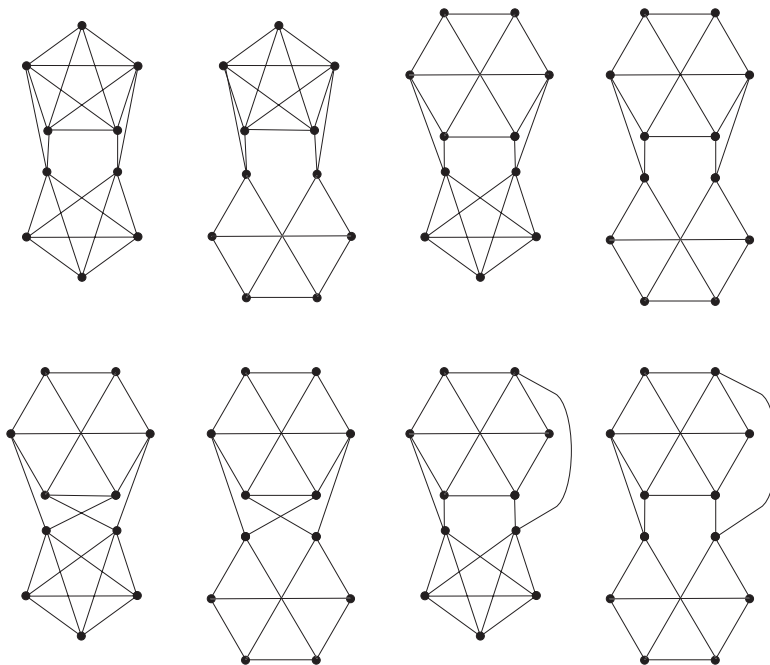
**Theorem 4.19.** *Suppose  $G$  is MMNA and of connectivity 2 with  $G_1 \in \{K_5 - e, K_{3,3} - e\}$ . Then there is no vertex, other than  $a$  and  $b$ , common to all  $a$ - $b$ -paths in  $G_2$ .*

*Proof.* Assume, for the sake of contradiction, that  $G_1 \in \{K_5 - e, K_{3,3} - e\}$  and there is a vertex  $c \in V(G'_2)$  that lies on every  $a$ - $b$ -path in  $G_2$ . Then, as in Theorem 4.14,  $\{a, c\}$  and  $\{b, c\}$  are 2-cuts for  $G$ , and as in the proof of Theorem 4.16, we can let  $H_1$  play the role of the  $G_1$  for the  $\{a, c\}$  cut and similarly  $H_2$  for the  $\{b, c\}$  cut and, by Theorems 4.12 and 4.15, both  $H_1$  and  $H_2$  are drawn from  $\{K_5, K_{3,3}, K_5 - e, K_{3,3} - e\}$ . Then  $G - c$  is planar, contradicting that  $G$  is NA.

Therefore, if  $G$  is MMNA, of connectivity 2 with  $G_1 \in \{K_5 - e, K_{3,3} - e\}$ , then there is no vertex, other than  $a$  and  $b$ , common to all  $a$ - $b$ -paths in  $G_2$ . □

**Theorem 4.20.** *Let  $G$  be MMNA with  $\kappa(G) = 2$  and  $ab \notin E(G)$ , where  $\{a, b\}$  is a 2-cut. If  $G'_2$  is nonplanar, then there are independent  $a$ - $b$ -paths in  $G_2$ .*

*Proof.* By Theorem 4.15,  $G_1 \in \{K_5 - e, K_{3,3}, K_{3,3} - e\}$ . However, if  $G_1 = K_{3,3}$  then, together with  $G'_2$ , this constitutes a pair of disjoint  $K$ -subgraphs, which would mean  $G$  has a proper disconnected NA minor, a contradiction. So  $G_1 \in \{K_5 - e, K_{3,3} - e\}$  and we can apply Menger's theorem and Theorem 4.19. □



**Figure 5.** Graphs of type  $(2, 2, 0)$ .

**Theorem 4.21.** *If  $G$  is MMNA of  $(2, 2, 0)$  type and  $G'_2 \in \{K_5, K_{3,3}\}$ , then  $G$  is one of the eight graphs in Figure 5.*

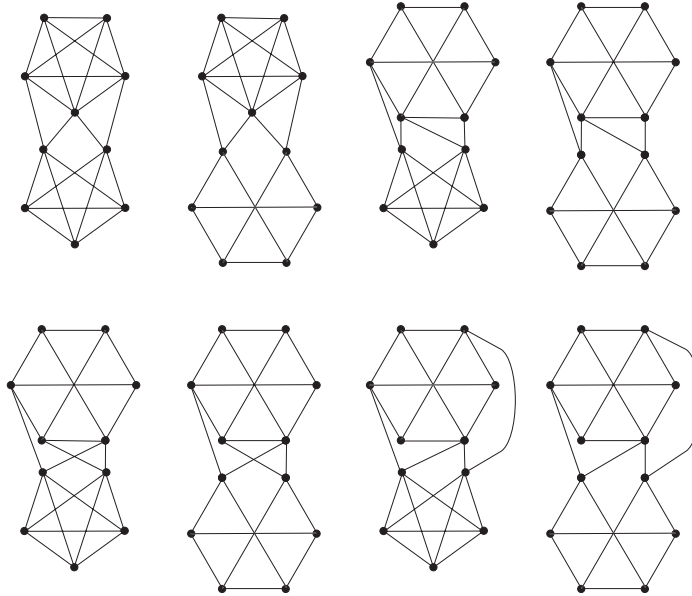
*Proof.* Notice that the eight graphs in the figure are MMNA. Suppose  $G$  is MMNA of  $(2, 2, 0)$  type with  $G'_2$  a Kuratowski graph. By Theorem 4.15,  $G_1 \in \{K_5 - e, K_{3,3}, K_{3,3} - e\}$ . However,  $G_1$  cannot be  $K_{3,3}$  because then, together with  $G'_2$  it forms a disconnected MMNA minor of  $G$ . We continue by examining the ways to construct  $G_2$ .

To construct  $G_2$ , we consider how to add the vertices  $a$  and  $b$  to  $G'_2$ . Let  $a$  have neighbors  $v_1, v_2 \in V(G'_2)$  and let  $v_3, v_4 \in V(G'_2)$  be the neighbors of  $b$ . Since  $G$  is of  $(2, 2, 0)$  type,  $\{v_1, v_2\} \cap \{v_3, v_4\} = \emptyset$ . Up to symmetry, there is only one way to attach  $a$  and  $b$  to  $K_5$ . This gives two of the graphs in the figure, as  $G_1$  is either  $K_5 - e$  or  $K_{3,3} - e$ .

In  $K_{3,3}$ , the vertices are split into two parts  $A$  and  $B$ , each of three vertices. Then the four vertices  $v_i, i = 1, \dots, 4$ , are either divided with two in each part, or else with three in one part and the fourth in the other. In the first case, there are two subcases: either  $\{v_1, v_2\} \subset A$  (and  $\{v_3, v_4\} \subset B$ ) or else  $|\{v_1, v_2\} \cap A| = |\{v_1, v_2\} \cap B| = 1$  (and similarly for  $\{v_3, v_4\}$ ). These three choices for  $G_2$  along with the two choices for  $G_1$ , either  $K_5 - e$  or  $K_{3,3} - e$ , account for the remaining six graphs in Figure 5.  $\square$

**Theorem 4.22.** *If  $G$  is MMNA of  $(2, 2, 1)$  type and  $G'_2 \in \{K_5, K_{3,3}\}$ , then  $G$  is one of the eight graphs of Figure 6.*





**Figure 6.** Graphs of type (2, 2, 1).

*Proof.* The proof is similar to that for (2, 2, 0) type. If  $G'_2$  is a Kuratowski graph, then  $G_1$  cannot be  $K_{3,3}$ , as that would result in a proper NA minor. So  $G_1 \in \{K_5 - e, K_{3,3} - e\}$ . If  $G'_2 = K_5$ , up to symmetry there is only one way to form  $G_2$  and this gives two graphs in the figure, as  $G_1$  is either  $K_5 - e$  or  $K_{3,3} - e$ .

If  $G'_2 = K_{3,3}$ , there are three ways to form  $G_2$ . Together,  $a$  and  $b$  have three neighbors in  $G'_2$ , which can either all lie in one part or else be split with a single vertex in one part and the remaining two in the other. In this second case, there are two further subcases since the vertex that is alone in its part can either be the common neighbor or not. Together with these three choices for  $G_2$ , there are two choices for  $G_1$ , either  $K_5 - e$  or  $K_{3,3} - e$ . This gives the remaining six graphs of Figure 6.  $\square$

We conclude this section with a classification of the MMNA graphs of connectivity 2, with 2-cut  $\{a, b\}$  such that  $G - a, b$  has a nonplanar component. By Theorem 4.11 we must have  $ab \notin E(G)$ , and by Theorem 4.7,  $G'_1$  is planar. In other words, if there is a nonplanar component, it must be  $G'_2$ . So far, we have constructed 21 graphs with nonplanar  $G'_2$ , the three bowtie graphs of Theorem 4.17, two of the (2, 2, 2) graphs (the two to the left of Figure 4), and eight each of (2, 2, 0) type (Theorem 4.21) and (2, 2, 1) type (Theorem 4.22). This is in fact a complete listing of the graphs with  $G'_2$  nonplanar, as we now show.

**Theorem 4.23.** *Let  $G$  be MMNA with  $\kappa(G) = 2$  and 2-cut  $\{a, b\}$  such that  $G - a, b$  has a nonplanar component. Then  $G$  is of (2, 2,  $c$ ) type with  $c = 0, 1,$  or  $2$  and appears in one of Figures 3 (right), 4, 5, or 6.*

*Proof.* Assume the hypothesis. As remarked above, if  $\{a, b\}$  is a 2-cut, this implies  $ab \notin E(G)$  and  $G'_2$  is nonplanar. Let  $H'_2$  be a K-subgraph of  $G'_2$ . Since  $ab \notin E(G)$ , combining Theorems 4.15 and 4.2, we have  $G_1 \in \{K_5 - e, K_{3,3} - e\}$ . By Theorem 4.20 there are independent  $a$ - $b$ -paths in  $G_2$ , call them  $P_1$  and  $P_2$ . Since, by Theorem 4.15,  $G_1 + ab$  is nonplanar,  $P_1$  and  $P_2$  each have vertices in common with  $H'_2$ . (Otherwise,  $G$  has disjoint nonplanar subgraphs and therefore a disconnected NA minor, by Theorem 4.2, contradicting  $G$  being minor minimal.) By contracting edges if necessary, we have a minor of  $G$  for which the vertices of  $P_i$  are  $a, a_i, \dots, b_i, b$  with  $a_i, b_i \in V(H_2)$ ,  $i = 1, 2$ . Then there are several cases that correspond to  $(2, 2, c)$  type, where  $c = 0, 1, 2$ .

Suppose first that  $a_1 = b_1$  and  $a_2 = b_2$  so that  $G$  is of  $(2, 2, 2)$  type. By contracting edges in  $H'_2$  if needed, we recognize that  $G$  has one of the five graphs of Theorem 4.18 as a minor. Since  $G$  is MMNA,  $G$  is one of these five graphs and since  $G'_2$  is nonplanar,  $G$  must be one of the two graphs with  $G'_2 = K_{3,3}$  (i.e., the two to the left of Figure 4). In other words  $G$  is of  $(2, 2, 2)$  type and appears in one of the figures, as required.

The rest of the argument is a little technical and we introduce some notation to simplify the exposition. The K-subgraph  $H'_2$  is a subdivision of  $K_5$  or  $K_{3,3}$  and, along with vertices of degree 2, has five or six vertices of higher degree that we will call *branch vertices*. Corresponding to the edges of  $K_5$  or  $K_{3,3}$ , the branch vertices are connected by paths that we call *2-paths* whose internal vertices are all of degree 2.

To continue the argument, suppose next that, say,  $a_1 = b_1$ , but  $a_2 \neq b_2$ . By contracting edges in  $H'_2$  if necessary, we can arrange that at least two of the three vertices  $a_1, a_2$ , and  $b_2$  become branch vertices of the K-subgraph. If all three can be made branch vertices, then, by further edge contractions, if necessary, we see that one of the eight  $(2, 2, 1)$  graphs of Theorem 4.22 is a minor of  $G$ . Since  $G$  is MMNA, this means  $G$  is one of the  $(2, 2, 1)$  graphs, with  $G'_2 \in \{K_5, K_{3,3}\}$  appearing in Figure 6, as required. If not, we can assume that it is  $a_1$  that remains as a degree-2 vertex of  $H'_2$ . For, if it is  $a_2$  or  $b_2$  that remains, we can contract edges to make  $a_2 = b_2$  and return to the previous case. With  $a_1$  as a degree-2 vertex in  $G'_2$ , we recognize that, perhaps by further edge contractions,  $G$  has a bowtie graph as a minor. Since  $G$  is MMNA,  $G$  is a bowtie graph. That is  $G$  is of  $(2, 2, 1)$  type and appears in Figure 3 (right), as required.

Finally, suppose  $a_1 \neq b_1$  and  $a_2 \neq b_2$ . If all four can be made distinct branch vertices by edge contractions in  $H'_2$ , then  $G$  has a  $(2, 2, 0)$  minor, so  $G$  is a  $(2, 2, 0)$  graph with  $G'_2 \in \{K_5, K_{3,3}\}$  appearing in Figure 5, as required.

Next, suppose at most three can be made into branch vertices and, without loss of generality, suppose it is  $a_1$  that remains as a degree-2 vertex in  $H'_2$ . This means  $a_1$  lies on a 2-path between two of  $b_1, a_2$ , and  $b_2$ . If the path ends at  $b_1$ , by further

edge contractions in  $H'_2$ , we can realize  $a_1 = b_1$  as a branch vertex and return to an earlier case. So, we can assume that  $a_1$  is on a 2-path between  $a_2$  and  $b_2$ . Use the part of the 2-path between  $a_1$  and  $b_2$  to form a new  $a$ - $b$ -path  $P'_1$  (i.e.,  $a'_1 = a_1$  and  $b'_1 = b_2$ ) and use a path in  $H'_2$  between the branch vertices  $a_2$  and  $b_1$  that avoids the branch vertex  $b_2$  to construct an independent  $a$ - $b$ -path  $P'_2$  (i.e.,  $P'_2$  has  $a'_2 = a_2$  and  $b'_2 = b_1$ ). Now we can contract edges in  $P'_1$  to identify  $a'_1 = a_1$  and  $b'_1 = b_2$  to return to the earlier case where  $a_1 = b_1$ . This completes the argument when at most three of the vertices can be moved to branch vertices.

Finally, suppose that at most two of the vertices can be made into branch vertices of  $H'_2$  by contracting edges, if needed. There are two subcases. If  $a_1$  and  $b_1$  are the branch vertices, then  $a_2$  and  $b_2$  are degree-2 vertices on a 2-path between  $a_1$  and  $b_1$ . Here we can further contract edges in  $H'_2$  to identify  $a_2$  and  $b_2$ , which returns us to an earlier case. In the second subcase, without loss of generality, it is  $a_1$  and  $a_2$  that are the branch vertices of  $H'_2$ . Assuming we cannot easily contract edges to identify  $a_1$  and  $b_1$  or  $a_2$  and  $b_2$ , it must be that the 2-path from  $a_1$  to  $a_2$  passes first through  $b_2$  and then through  $b_1$ . In this case, we replace  $P_1$  and  $P_2$  by the independent paths  $P'_1$ , which uses the 2-path from  $a_1$  to  $b_2$  (so  $a'_1 = a_1$  and  $b'_1 = b_2$ ), and  $P'_2$ , which uses the 2-path from  $a_2$  to  $b_1$  (then  $a'_2 = a_2$  and  $b'_2 = b_1$ ). By further edge contractions, we return to our first case where  $a_1 = b_1$  and  $a_2 = b_2$ .  $\square$

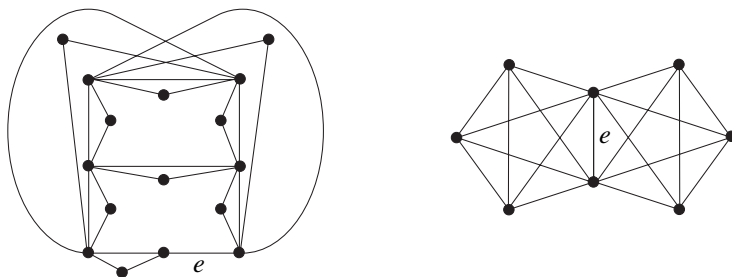
Together, the three bowtie graphs and the eight of Figure 6 give eleven MMNA graphs of  $(2, 2, 1)$  type. In total we have found three disconnected MMNA graphs, nine where  $ab \in E(G)$ , as well as eight, eleven, and five, respectively when  $G$  is of type  $(2, 2, c)$  for  $c = 0, 1, 2$ , respectively. This gives a total of 36 MMNA graphs.

## 5. MMNE and MMNC graphs

In this section we classify MMNE and MMNC graphs of connectivity,  $\kappa(G)$ , at most 1. For MMNE graphs we also show  $\kappa(G) \leq 5$  and determine the graphs with  $\kappa(G) = 2$  and minimum degree at least 3. We conclude the section by describing a computer search that found 55 MMNE and 82 MMNC graphs.

We begin by observing that the MMNE and MMNC graphs are not Kuratowski sets as the opposite properties are not minor closed. Recall that NE is an abbreviation for not edge apex. The opposite property is *edge apex*, meaning there is an  $e \in E(G)$  so that  $G - e$  is planar. We call such an edge an *apex edge*. Similarly, the opposite of NC is *contraction apex*, meaning there is an edge  $e$  such that  $G/e$  is planar. We call  $e$  a *contraction apex*.

**Theorem 5.1.** *Deleting an edge of an edge apex graph results in an edge apex graph. Contracting an edge of an edge apex graph results in an edge apex graph unless the edge that is contracted is the only apex edge.*



**Figure 7.** Examples showing that the sets of MMNE and MMNC graphs are not Kuratowski sets.

*Proof.* Suppose that  $G$  is edge apex, so it contains an edge  $e$  such that  $G - e$  is planar. Let  $G'$  be the result of deleting some edge  $f$  in  $G$ . If  $f \neq e$ , consider  $G' - e$  and note that  $G' - e = G - e, f$ , which is a minor of  $G - e$ . Graph  $G - e$  is planar, so  $G' - e$  is also planar, and  $e$  is an apex edge for  $G'$ , which is therefore edge apex. Otherwise, if  $f = e$ , then  $G'$  would be planar and so would also be edge apex.

Now suppose that  $G$  contains at least two edges  $e_1$  and  $e_2$  ( $e_1 \neq e_2$ ) such that both  $G - e_1$  and  $G - e_2$  are planar. Let  $f$  be an arbitrary edge in  $G$  and let  $G''$  be the result of contracting edge  $f$  in  $G$ . Without loss of generality, suppose that  $f \neq e_1$ . Consider the graph  $G'' - e_1$ , where if  $e_1$  is incident to  $f$  in  $G$  then  $e_1$  is incident to the vertex formed by contracting  $f$  in  $G''$ . Note that this graph  $G'' - e_1$  is a minor of  $G - e_1$ . But  $G - e_1$  is planar, and since planarity is closed under taking minors, the graph  $G'' - e_1$  is planar. So edge  $e_1$  is an apex edge of  $G''$ .  $\square$

**Theorem 5.2.** *The set of graphs that are edge apex is not closed under taking minors.*

*Proof.* Let  $G$  be the graph in Figure 7 (left). This graph can be described as  $K_{3,3}$  with all but one edge replaced by a triangle, and with that one edge subdivided into an edge  $e$  and another edge to be replaced by a triangle. This graph is edge apex with  $e$  as the unique apex edge. However,  $G/e$  is  $K_{3,3}$  with every edge replaced by a triangle, so  $G/e$  is not edge apex.  $\square$

**Theorem 5.3.** *Contracting an edge of a contraction apex graph results in a contraction apex graph. Deleting an edge of a contraction apex graph results in a contraction apex graph unless the edge that is deleted is the **only** contraction apex.*

*Proof.* Suppose that  $G$  is contraction apex, so it contains an edge  $e$  such that  $G/e$  is planar. Let  $G'$  be the result of contracting some edge  $f$  in  $G$ . If  $f \neq e$ , consider  $G'/e$  and note that  $G'/e = G/e, f$ , which is a minor of  $G/e$ . Graph  $G/e$  is planar, so  $G'/e$  is also planar, and  $e$  is a contraction apex for  $G'$ , which is therefore a contraction apex graph. Otherwise, if  $f = e$ , then  $G'$  would be planar and so would also be contraction apex.

Now suppose that  $G$  contains at least two edges  $e_1$  and  $e_2$  ( $e_1 \neq e_2$ ) such that both  $G/e_1$  and  $G/e_2$  are planar. Let  $f$  be an arbitrary edge in  $G$  and let  $G''$  be the result of deleting edge  $f$  in  $G$ . Without loss of generality, suppose that  $f \neq e_1$ . Consider the graph  $G''/e_1$  and note that it is a minor of  $G/e_1$ . But  $G/e_1$  is planar, and since planarity is closed under taking minors, the graph  $G''/e_1$  is planar. So edge  $e_1$  is a contraction apex of  $G''$ .  $\square$

**Theorem 5.4.** *The set of graphs that are contraction apex is not closed under taking minors.*

*Proof.* Define the graph  $G$  as two copies of  $K_5$  that share a common edge  $e$ ; see Figure 7 (right). We show that  $G$  is contraction apex, but has a minor that is NC. Indeed, contracting the common edge,  $G/e = K_4 \dot{\cup} K_4$ , which is planar. Note that this is the unique contraction apex of  $G$ .

Now define the subgraph  $G'$  as  $G - e$ . Label the two subgraphs isomorphic to  $K_5 - e$  as  $G'_1$  and  $G'_2$ . Without loss of generality, suppose we contract an edge  $f$  in  $G'_2$ . Notice that we are left with  $G'_1 = K_5 - e$ , and a path through  $G'_2$  that connects the two degree-3 vertices of  $G'_1$ . Thus,  $G'/f$  has a subgraph homeomorphic to  $K_5$  and is nonplanar. By symmetry, whatever edge  $f \in E(G')$  we choose,  $G'/f$  is nonplanar. Thus  $G'$  is NC.  $\square$

We next classify the disconnected and connectivity-1 MMNE and MMNC graphs, which turn out to be the same sets.

**Theorem 5.5.** *The disconnected MMNE graphs are  $K_5 \sqcup K_5$ ,  $K_5 \sqcup K_{3,3}$ , and  $K_{3,3} \sqcup K_{3,3}$ .*

*Proof.* First observe that these three graphs are MMNE. Let  $G$  be MMNE and disconnected. Suppose one of  $G_1, G_2$  is planar, say  $G_1$ . Then let  $e_1 \in E(G_1)$ , and note that  $G - e_1$  is not NE and nonplanar. Let  $e_2$  be the edge whose removal from  $G - e_1$  gives a planar graph. Since  $G_1$  is planar, it must be that  $e_2$  is in  $E(G_2)$ . But, since  $G_1$  is planar, this means that removing  $e_2$  from  $G$  gives the disconnected union of the planar  $G_1$  and a planar minor of  $G_2$ . So, this graph,  $G - e_2$ , is planar, which is a contradiction since  $G$  is NE. So it must be that  $G_1$  and  $G_2$  are both nonplanar. Thus one of the graphs generated by  $G_1 \sqcup G_2$ , where  $G_1, G_2 \in \{K_5, K_{3,3}\}$ , must be a minor of  $G$ . Since  $G$  is minor minimal,  $G$  must be one of these three graphs.  $\square$

**Theorem 5.6.** *The disconnected MMNC graphs are  $K_5 \sqcup K_5$ ,  $K_5 \sqcup K_{3,3}$ , and  $K_{3,3} \sqcup K_{3,3}$ .*

*Proof.* First observe that these three graphs are MMNC. Let  $G$  be MMNC and disconnected. Suppose one of  $G_1, G_2$  is planar, say  $G_1$ . Then let  $e_1 \in E(G_1)$ , and note that  $G - e_1$  is not NC and nonplanar. Then there is an edge  $e_2 \in E(G - e_1)$  such that  $(G - e_1)/e_2$  is planar. Since  $G_1$  is planar, it must be that  $e_2$  is in  $E(G_2)$ . But, since  $G_1$  is planar, this means that contracting  $e_2$  in  $G$  gives the disconnected

union of the planar  $G_1$  and a planar minor of  $G_2$ . This graph  $G/e_2$  is planar, which is a contradiction since  $G$  is NC. So it must be that  $G_1$  and  $G_2$  are both nonplanar. Then one of the graphs  $G = G_1 \sqcup G_2$ , with  $G_i \in \{K_5, K_{3,3}\}$ , is a minor of  $G$ . Since  $G$  is minor minimal, it is one of those three graphs.  $\square$

**Corollary 5.7.** *Let  $G$  be disconnected. The following are equivalent:  $G$  is MMNA;  $G$  is MMNE;  $G$  is MMNC.*

Recall that  $G_1 \dot{\cup} G_2$  is the union of  $G_1$  and  $G_2$  with one vertex identified.

**Theorem 5.8.** *If  $G$  is MMNE and  $\kappa(G) = 1$  then  $G = G_1 \dot{\cup} G_2$ , where  $G_1, G_2 \in \{K_5, K_{3,3}\}$ , and they share exactly one vertex.*

*Proof.* First observe that these three graphs are MMNE. Let  $G = G_1 \dot{\cup} G_2$  and suppose for the sake of contradiction that one of  $G_1$  and  $G_2$ , say  $G_1$ , is planar. Let  $e$  be an edge of  $G_1$ . Then  $G - e$  is not NE and nonplanar. Let  $f$  be the apex edge of  $G - e$ . Since  $G_1$  is planar,  $f$  must lie in  $E(G_2)$ . Since  $G_2 - f$  is a subgraph of the planar  $G - e, f$ , it must itself be planar. Note that  $G - f = G_1 \cup (G_2 - f)$  is the union of two planar graphs that share at most one vertex, which is clearly planar. This is a contradiction, since  $G$  is NE. So both  $G_1$  and  $G_2$  are nonplanar. So  $G$  has one of the graphs  $G_1 \dot{\cup} G_2$ ,  $G_1, G_2 \in \{K_5, K_{3,3}\}$  as a minor. Since these graphs are NE and  $G$  is minor minimal,  $G$  must be one of these three graphs.  $\square$

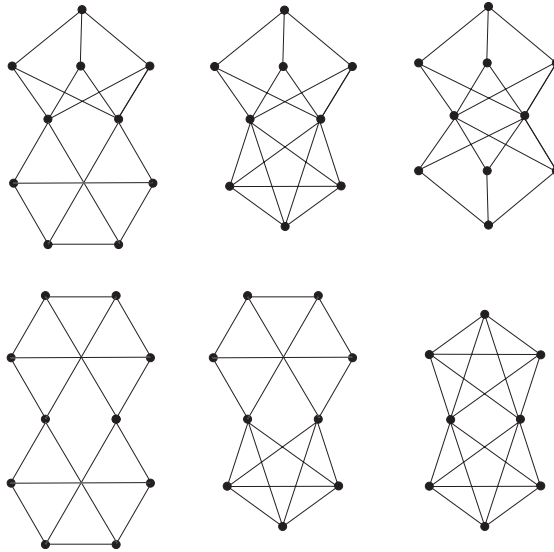
**Theorem 5.9.** *If  $G$  is MMNC and  $\kappa(G) = 1$  then  $G = G_1 \dot{\cup} G_2$ , where  $G_1, G_2 \in \{K_5, K_{3,3}\}$ , and they share exactly one vertex.*

*Proof.* First observe that these three graphs are MMNC. Let  $G = G_1 \dot{\cup} G_2$  and suppose for the sake of contradiction that one of  $G_1$  and  $G_2$ , say  $G_1$ , is planar. Let  $e$  be an edge of  $G_1$ . Then  $G - e$  is not NC and nonplanar. Let  $f \in E(G - e)$  be the contraction apex of  $G - e$ ; that is,  $(G - e)/f$  is planar. Since  $G_1$  is planar,  $f$  must lie in  $G_2$ . Since  $G_2/f$  is a subgraph of the planar  $(G - e)/f$ , it must itself be planar. Note that  $G/f = G_1 \cup (G_2/f)$  is the union of two planar graphs that share at most one vertex, which is clearly planar. This is a contradiction, since  $G$  is NC.

Thus, both  $G_1$  and  $G_2$  are nonplanar. So  $G$  has one of the graphs  $G_1 \dot{\cup} G_2$  with  $G_1, G_2 \in \{K_5, K_{3,3}\}$  as a minor. Since these graphs are NC and  $G$  is minor minimal,  $G$  must be one of these three graphs.  $\square$

**Corollary 5.10.** *Let  $G$  have connectivity 1. Then  $G$  is MMNE if and only if it is MMNC.*

Recall that there are no MMNA graphs of connectivity 1. In particular, for each of  $K_5 \dot{\cup} K_5$ ,  $K_5 \dot{\cup} K_{3,3}$ , and  $K_{3,3} \dot{\cup} K_{3,3}$ , the cut vertex is an apex. We next classify the MMNE graphs of connectivity 2 under the assumption that the minimum degree,  $\delta(G)$ , is at least 3. We will argue that there are exactly six such graphs and we begin with the observation that those graphs are indeed MMNE. As discussed at the end



**Figure 8.** The six MMNE graphs of connectivity 2 with  $\delta(G) \geq 3$ .

of this section, based on a computer search, these again coincide with the MMNC examples of connectivity 2 with  $\delta(G) \geq 3$ . In addition to being both MMNE and MMNC, these 12 graphs with  $\kappa(G) \leq 2$  are exactly the obstructions, of connectivity at most 2, to embedding a graph in the projective plane; see [Mohar and Thomassen 2001, Section 6.5].

**Theorem 5.11.** *The six graphs of Figure 8 are MMNE.*

Note that these graphs are of the form  $G_1 \dot{\cup} G_2$  with  $G_i \in \{K_5 - e, K_{3,3}, K_{3,3} - e\}$ , i.e., the union of  $G_1$  and  $G_2$  identified on two vertices.

*Proof.* Let  $G$  be one of the six graphs and  $e$  denote an arbitrary edge of  $G$ . It is easy to verify that each  $G - e$  is nonplanar, so  $G$  is NE. We must also show that no minor of  $G$  is NE. We first observe that for each choice of  $e$ , there is another edge  $f$  such that  $G - e, f$  is planar. That is,  $G - e$  is not NE. Also, there is an edge  $g$  such that  $(G/e) - g$  is planar, which shows  $G/e$  is not NE.

By Theorem 5.1, deleting or contracting further edges continues to give minors of  $G$  that are not NE, so long as we do not contract the unique apex edge in a graph. Working around this obstacle is not difficult as we very quickly come to planar minors. Planarity is closed under taking minors and a planar graph is not NE.  $\square$

A key step in the classification is the observation that  $ab$  is not an edge of  $G$ .

**Lemma 5.12.** *If  $G$  is MMNE,  $\kappa(G) = 2$  with cut set  $\{a, b\}$ , and  $\delta(G) \geq 3$ , then  $ab$  is not an edge in  $G$ .*

*Proof.* Let  $G$  be as described. Let  $G - a, b = G'_1 \sqcup G'_2$  and let  $G_i$  be the induced subgraph of  $G$  on the vertices  $V(G'_i) \cup \{a, b\}$ . For a contradiction, suppose that  $ab$  is an edge in  $G$ . There are three cases to consider depending on which of  $G_1$  and  $G_2$  is planar. If both are planar, then  $G$  is the union of two planar graphs that share an edge and therefore is planar. This contradicts  $G$  being MMNE.

Next suppose exactly one of  $G_1$  and  $G_2$  is planar, say  $G_1$ . If  $e \in E(G_2)$  is an edge other than  $ab$ , then  $G_2 - e$  must be nonplanar. For otherwise,  $G - e$ , the union of two planar graphs,  $G_1$  and  $G_2 - e$  along  $ab$ , is planar contradicting  $G$  being NE. If  $G_2 - ab$  is also nonplanar, then  $G_2$  is a proper subgraph that is NE, which contradicts  $G$  being minor minimal. So,  $G_2 - ab$  is planar.

This means that  $G - ab$  is the union of the planar  $G_1 - ab$  and the planar  $G_2 - ab$ , joined at two vertices. However, since  $G$  is NE,  $G - ab$  is nonplanar, so it has a subgraph homeomorphic to  $K_5$  or  $K_{3,3}$ . Using Lemma 1.8, we know that the subgraph must use only a path through one of  $G_1$ ,  $G_2$ , and nothing else in that component. This means that one of  $G_i^*$  is an edge away from containing a K-subgraph, where  $G_i^*$  denotes  $G_i - ab$ . Since  $G_1$  is planar, it must be  $G_2^*$  that contains a subdivision of  $K_5$  or  $K_{3,3}$  with an edge removed. Thus,  $G_2$  has a subgraph homeomorphic to  $K_5$  or  $K_{3,3}$  that uses the edge  $ab$ .

Replace  $G_1^*$  by the path of Lemma 1.8 to form a subgraph  $H$  of  $G$ . We claim that  $H$  is NE. Indeed, deleting  $e \in E(G_2^*)$  leaves  $H - e$  with the nonplanar subgraph  $G_2 - e$ . Deleting  $ab$  or an edge in the  $G_1^*$  path leaves an  $a$ - $b$ -path that completes a K-subgraph in  $G_2^*$ . Since  $G$  is minor minimal,  $G$  must be  $H$ . However,  $H$  has at least one degree-2 vertex, contradicting  $\delta(G) \geq 3$ .

Finally, we have the case where  $G_1$  and  $G_2$  are both nonplanar. Here there are three subcases to consider depending on which of  $G_1^* = G_1 - ab$  and  $G_2^* = G_2 - ab$  is planar.

Suppose first that both  $G_1^*$  and  $G_2^*$  are planar. In this case, each of  $G_1$  and  $G_2$  has a K-subgraph that contains  $ab$ . It follows that one of the graphs of Theorem 5.11 is a proper minor of  $G$ , contradicting the minor minimality of  $G$ .

In the subcase where both  $G_1^*$  and  $G_2^*$  are nonplanar, let  $e$  be the apex edge of  $G - ab$ . Since the only edge common to  $G_1^*$  and  $G_2^*$  is  $ab$ , the edge  $e$  is in exactly one of  $G_1^*$  and  $G_2^*$ . Whichever it is not in will constitute a nonplanar subgraph of  $G - ab, e$ , which is a contradiction.

Finally, assume exactly one of  $G_1^*$  and  $G_2^*$  is planar, say  $G_1^*$ . As above,  $G_1^*$  planar and  $G_1$  not implies  $G_1$  contains a K-subgraph including  $ab$  as an edge. On the other hand, since  $G_2^*$  is nonplanar, it has a K-subgraph  $H$ . Let  $M = G_1 \cup H$  and, for a contradiction, suppose that  $M$  is a proper minor. Then  $M$  must have an apex edge. However, if we remove an edge  $e$  from  $G_1$ , then  $H$  remains, meaning  $M - e$  is nonplanar. If we remove  $e$  from  $H$  (which shares no edges with  $G_1$  since it is a subgraph of  $G_2^*$ ), then  $G_1$  remains, meaning  $M - e$  is still nonplanar. Therefore,



no matter what edge we remove from  $M$ , we cannot make it planar and  $M$  is NE. However,  $M$  is a minor of  $G$ , so this contradicts  $G$  being MMNE. Therefore,  $H$  is not a proper minor of  $G_2^*$ , so  $G_2^*$  is a subdivision of  $K_5$  or  $K_{3,3}$ . A similar argument (replacing  $H$  by  $K_5$  or  $K_{3,3}$ ) shows, in fact,  $G_2^*$  is  $K_5$  or  $K_{3,3}$  and not just a subdivision. However, since  $ab$  is not an edge of  $G_2^*$ , then  $G_2^*$  must be  $K_{3,3}$ .

Thus  $G_2^* = K_{3,3}$  and  $G_1$  contains a subdivision of  $K_{3,3}$  or  $K_5$  that includes  $ab$  as an edge. This means  $G$  includes one of the graphs of Theorem 5.11 as a proper minor and is not minor minimal.

This completes the last subcase of the last case and shows that  $ab$  is not an edge of  $G$ . □

For  $G$  of connectivity 2 with cut set  $\{a, b\}$ , we have  $G - a, b = G'_1 \sqcup G'_2$ . We will use  $G_i$  to denote the induced subgraph on  $V(G'_i) \cup \{a, b\}$ .

**Lemma 5.13.** *If  $G$  is MMNE,  $\kappa(G) = 2$ , and  $G_1$  and  $G_2$  are both nonplanar, then  $G_1 = G_2 = K_{3,3}$ .*

*Proof.* Let  $G$  be as described. First suppose for the sake of contradiction that  $G_1$  is nonplanar but not  $K_{3,3}$ . Note that  $G_1$  cannot be  $K_5$  because  $ab \notin E(G)$  by Lemma 5.12. So  $G_1$  has some nonplanar proper minor  $H$ , and  $H \cup G_2$  is a proper minor of  $G$ . Since there are no edges between  $H$  and  $G_2$ , the apex edge of  $H \cup G_2$  must be in exactly one of  $H$  or  $G_2$ . Whichever one does not contain the apex edge will be a nonplanar subgraph even when the edge is removed, contradicting the fact that  $G$  is MMNE. Therefore  $G_1 = K_{3,3}$ . A symmetrical argument can be made for  $G_2$ . □

**Lemma 5.14.** *If  $G$  is MMNE,  $\kappa(G) = 2$ , with cut set  $\{a, b\}$ ,  $\delta(G) \geq 3$ , and both  $G_1$  and  $G_2$  are planar, then  $G_i \in \{K_5 - e, K_{3,3} - e\}$  with  $ab$  as the missing edge.*

*Proof.* Let  $G$  be as described. For a contradiction, assume that  $G_1 + ab$  is planar. Since  $G$  is NE, for every  $e \in E(G)$ , the graph  $G - e$  is nonplanar and, therefore, has a K-subgraph,  $H$ . By Lemma 1.8 and our assumption that  $G_1 + ab$  is planar,  $H \cap G_1$  is an  $a$ - $b$ -path. In particular  $G_2 + ab$  is nonplanar.

Note that there are edge-disjoint  $a$ - $b$ -paths  $P_1$  and  $P_2$  in  $G_1$ . If not, say every  $a$ - $b$ -path goes through the edge  $e'$ . Then  $G - e'$  must be planar as, by Lemma 1.8, a K-subgraph of  $G - e'$  would either use a path in  $G_1$ , which is not possible as all such paths pass through  $e'$ , or else use a path in  $G_2$ , which is not possible since  $G_1 + ab$  is planar. The contradiction shows there are edge-disjoint paths  $P_1$  and  $P_2$ .

This means we can construct a proper minor  $M$  of  $G$  by adding a triangle on  $ab$ . That is,  $V(M) = V(G_2) \cup \{c\}$  and  $E(M) = E(G_2) \cup \{ab, bc, ac\}$ . Since  $G$  is NE, for any  $e \in E(G_2)$ , the graph  $G - e$  is nonplanar with a K-subgraph that uses only a path in  $G_1$ . So,  $M - e$  is also nonplanar. On the other hand, if we delete any  $e$  in  $\{ab, ac, bc\}$ , we are left with a subgraph of  $M - e$  homeomorphic to  $G_2 + ab$ . So  $M - e$  is again nonplanar. Then  $M$  is a proper NE minor of  $G$  contradicting  $G$  being minor minimal.

We conclude  $G_1 + ab$  is nonplanar. A similar argument shows  $G_2 + ab$  is nonplanar as well. Then  $G$  must have one of the NE graphs  $G_1 \dot{\cup} G_2$  with  $G_i \in \{K_5 - e, K_{3,3} - e\}$  as a minor. Since  $G$  is minor minimal,  $G$  is a graph of this form.  $\square$

**Lemma 5.15.** *If  $G$  is MMNE,  $\kappa(G) = 2$ ,  $\delta(G) \geq 3$ ,  $G_1$  is planar, and  $G_2$  is nonplanar, then  $G_1 \in \{K_5 - e, K_{3,3} - e\}$ , sharing two vertices and no edges with  $G_2 = K_{3,3}$ .*

*Proof.* Let  $G$  be as described. For a contradiction, suppose  $G_1 + ab$  is planar. Then  $G_2 + ab$  must be NE. Indeed, if we delete  $ab$ , we are left with the nonplanar  $G_2$ . Let  $e \in E(G_2)$ . Since  $G$  is NE,  $G - e$  is nonplanar and has a  $K$ -subgraph  $K$ . If  $K$  uses at most one of  $\{a, b\}$ , then  $K$  lies entirely in  $G_2$  and avoids  $e$ . So,  $(G_2 + ab) - e$  is nonplanar in this case. On the other hand, if  $\{a, b\} \subset V(K)$ , then, by Lemma 1.8 and since  $G_1 + ab$  is planar, the part of  $K$  in  $G_1$  is an  $a$ - $b$ -path. So using edge  $ab$  instead,  $K$  remains as a  $K$ -subgraph of  $(G_2 + ab) - e$ , which is again nonplanar. However,  $G_2 + ab$  being NE contradicts  $G$  being minor minimal. We conclude  $G_1 + ab$  is nonplanar.

This means  $G_1$  has one of  $K_5 - e$  and  $K_{3,3} - e$  as a minor with the missing edge corresponding to  $ab$ . Replace  $G_1$  by its minor  $K_5 - e$  or  $K_{3,3} - e$ , call it  $H$ , to form  $M = H \cup G_2$ , a minor of  $G$ . We claim  $M$  is again NE. Indeed, if we delete  $e \in E(H)$ , the graph  $G_2$  shows  $M - e$  is nonplanar. For  $e \in E(G_2)$ , we know  $G - e$  has a  $K$ -subgraph  $K$ . If  $K$  sees at most one of  $a$  and  $b$ , it must lie entirely in  $G_2$  (since  $H$  is planar) and  $M - e$  is nonplanar. If  $\{a, b\} \subset V(K)$ , then, by Lemma 1.8,  $K$  is simply a path on one side of the 2-cut. If  $K$  is a path in  $G_1$ , then replace that by a path in  $H$  to recognize  $K$  as a subgraph of  $M - e$ , which is therefore nonplanar. On the other hand, if  $K$  is a path in  $G_2$ , this path avoids  $e$ . So, we can use  $H$  along with that path to again find a nonplanar subgraph of  $M - e$ . Since  $G$  is minor minimal,  $G = M$  and  $G_1 \in \{K_5 - e, K_{3,3} - e\}$  as required.

Now,  $G_2$  being nonplanar has a  $K$ -subgraph  $K$ . Also, there must be an  $a$ - $b$ -path  $P$  in  $G_2$ , as otherwise  $G$  has connectivity 1. Moreover, both  $K$  and  $G_1 \cup P$  are nonplanar, and so they must overlap, as otherwise  $G$  has a proper disconnected MMNE minor. This means  $P$  passes through  $K$  and, by contracting edges in  $P$  if necessary, we can assume  $G$  has a minor with  $\{a, b\} \subset V(K)$ . From this, form the minor  $M = G_1 \cup K$ . If  $K$  is a subdivision of  $K_5$ , Then  $M$  and hence  $G$  has the MMNA graph  $G_1 \dot{\cup} (K_5 - e)$  as a proper minor, which is a contradiction. So,  $K$  is a subdivision of  $K_{3,3}$ . After contracting edges,  $G$  either has the MMNA  $G_1 \dot{\cup} (K_{3,3} - e)$  as a proper minor, which is a contradiction, or else  $G$  has  $G_1 \dot{\cup} K_{3,3}$  as a minor, where  $a$  and  $b$  are in the same part of  $K_{3,3}$ . Since  $G$  was minor minimal, we conclude  $G = G_1 \dot{\cup} K_{3,3}$ . In other words, as required,  $G_2 = K_{3,3}$ , sharing two vertices and no edge with  $G_1 \in \{K_5 - e, K_{3,3} - e\}$ .  $\square$

**Theorem 5.16.** *If  $G$  is MMNE,  $\kappa(G) = 2$ , and  $\delta(G) \geq 3$ , then  $G$  is one of the six graphs of Figure 8.*

*Proof.* We showed that these six graphs are MMNE in Theorem 5.11. Lemma 5.13 immediately gives that if  $G_1$  and  $G_2$  are both nonplanar, then they are both  $K_{3,3}$ . Lemmas 5.14 and 5.15 complete the other parts of the proof. In total, these account for six graphs: one from Lemma 5.13, three from Lemma 5.14, and two from Lemma 5.15.  $\square$

The restriction on the minimum degree in the last theorem is necessary. Indeed, there are many MMNE graphs with  $\delta(G) = 2$  (meaning  $\kappa(G) \leq 2$ ). For example, contracting edge  $e$  of Figure 7 (left) results in an MMNE graph that is formed by replacing each edge of  $K_{3,3}$  with a triangle. Similarly, replacing each edge of  $K_5$  with a triangle also yields an MMNE graph. Further examples of MMNE graphs with a degree-2 vertex are the first seven listed in Section A.1 of the Appendix.

We remark that these examples arise in part due to our insistence that edge contraction lead to a simple graph. Contracting an edge of a degree-2 vertex in a triangle gives a (multi)graph with a doubled edge. Our convention is to delete one of the doubled edges to return to a simple graph.

We next show that  $\delta(G) = 2$  is the minimum for MMNE graphs.

**Theorem 5.17.** *The minimum vertex degree in an MMNE graph is at least 2.*

*Proof.* The addition or deletion of an isolated vertex or vertex of degree 1 in a planar graph will again result in a planar graph. So if  $G$  is NE with  $\delta(G) < 2$ , then removing a vertex of degree 0 or 1 will result in a NE graph; hence  $G$  is not MMNE.  $\square$

Although we cannot completely classify the  $\delta(G) = 2$  MMNE graphs, we show that degree-2 vertices must occur as part of a triangle.

**Theorem 5.18.** *In an MMNE graph, the neighbors of a degree-2 vertex are themselves neighbors.*

*Proof.* Let  $G$  be an NE graph with a degree-2 vertex  $v$  with neighbors  $a$  and  $b$ . For a contradiction, suppose  $ab$  is not an edge of  $G$ . Perhaps  $G$  is MMNE so that every proper minor of  $G$  is not NE. Let  $H = G/av$  be the graph that results from contracting edge  $av$  in  $G$ . Since  $G$  is MMNE, there must be some edge  $e$  in  $H$  such that  $H - e$  is planar. Note that  $e$  cannot be the newly formed edge  $ab$  in  $H$ , else, since degree-1 vertices have no impact on the planarity of a graph,  $G - av$  would also be planar, contradicting  $G$  being MMNE. Consider the graph  $G - e$ . Note that  $G - e$  and  $H - e$  are homeomorphic, so since  $H - e$  is planar,  $G - e$  is also planar. But this contradicts  $G$  being MMNE.  $\square$

If graph  $G$  has a triangle  $abc$ , a  $\nabla Y$  move on  $G$  means forming a new graph  $G'$  with one additional vertex  $v$  (i.e.,  $V(G') = V(G) \cup \{v\}$ ) and replacing the edges

$ab$ ,  $ac$ , and  $bc$  with  $va$ ,  $vb$ ,  $vc$ . So,  $G'$  has the same number of edges as  $G$  and one additional vertex. Pierce [2014] shows that  $\nabla Y$  often preserves NA, as was originally observed by Barsotti in unpublished work. (The bowtie graphs of Figure 3 are examples where  $\nabla Y$  does not preserve NA.) Here we give a similar result for NE graphs.

**Theorem 5.19.** *Given an NE graph  $G$  with triangle  $t$ , let  $G'$  be the result of performing a  $\nabla Y$  move on triangle  $t$  in  $G$ , and let  $v$  be the vertex added in  $G'$ . Graph  $G'$  is NE if and only if  $G' - e_i$  is nonplanar for each  $e_i$  incident to  $v$ .*

*Proof.* If  $G'$  is NE, then  $G' - e_i$  is nonplanar by definition. Conversely suppose that  $G' - e_i$  is nonplanar for each  $e_i$  incident to  $v$ . Perhaps  $G'$  is not NE, so there is  $e \in E(G')$  such that  $G' - e$  is planar. Note that  $e$  cannot be incident to  $v$ . Since  $e$  is not part of triangle  $t$ , performing a  $\nabla Y$  move on  $G - e$  will result in  $G' - e$ , so  $\nabla Y$  on  $G - e$  is also planar. Note that undoing the  $\nabla Y$  transform on this graph will preserve its planarity. However, graph  $G - e$  being planar contradicts  $G$  being NE.  $\square$

We next give an upper bound on the connectivity of MMNE graphs. We first observe that the minimum degree  $\delta(G)$  is bounded by 5.

**Theorem 5.20.** *If  $G$  is MMNE, then  $\delta(G) \leq 5$ .*

*Proof.* Suppose  $G$  is MMNE with  $\delta(G) \geq 6$  and let  $n = |V(G)|$ . We can assume  $n \geq 6$ , as  $G$  must be nonplanar and the only nonplanar graph with five or fewer vertices is  $K_5$ , which is not MMNE. Since  $\delta(G) \geq 6$ , a lower bound on  $|E(G)|$  is  $6n/2 = 3n$ . Now since  $G$  is MMNE, there exist two edges  $e$  and  $f$  such that  $G - e, f$  is a planar graph with at least  $3n - 2$  edges. However, a planar graph on  $n$  vertices can have no more than  $3n - 6$  edges, the number of edges in a planar triangulation. The contradiction shows there is no MMNE graph with  $\delta(G) \geq 6$ .  $\square$

As  $\kappa(G) \leq \delta(G)$ , we have a bound on the connectivity as an immediate corollary.

**Corollary 5.21.** *If  $G$  is MMNE, then  $\kappa(G) \leq 5$ .*

Finally, we observe a nice connection between MMNE and MMNA graphs.

**Theorem 5.22.** *If  $G$  is MMNE, then  $G$  is MMNA or apex.*

*Proof.* Suppose  $G$  is MMNE and NA. We will argue that  $G$  is in fact MMNA. For this, let  $H$  be a proper minor. Since  $G$  is MMNE,  $H$  is edge apex. This means either  $H$  is already planar, or else there is an edge  $e$  such that  $H - e$  is planar. In the latter case, if  $v$  is a vertex of  $e$ , then  $H - v$  is again planar. This shows that  $H$  is apex, as required.  $\square$

**Results of computer searches.** In addition to the results above, we have found other examples of MMNE and MMNC graphs through brute-force computer searches. Our code is available at <https://github.com/mikepierce/MMGraphFunctions/tree/master/brute-force-search>. See the file `Brute-Force-Search.nb` for documentation.

The algorithms underlying the searches are fairly straightforward. First we generate a list of all the graphs that we are going to search using the tools that are available with the `nauty` and `Traces` graph theory software [McKay and Piperno 2014]. Specifically, we use the tools `geng` and `planarg` to produce all connected, nonplanar graphs of minimum vertex degree at least 2 that either have fewer than 20 edges or that have fewer than 10 vertices. The commands used to generate these graphs in `bash` are the following:

```
$ for i in {6..9}; do
geng -c -d2 ${i} | planarg -v > ${i}v.txt
done
$ for i in {10..16}; do
geng -c -d2 ${i} 0:17 | planarg -v > ${i}v,(0-17)e.txt
geng -c -d2 ${i} 18 | planarg -v > ${i}v,(18)e.txt
geng -c -d2 ${i} 19 | planarg -v > ${i}v,(19)e.txt
done
```

This brute force search was carried out on a standard laptop computer with 4 GB of memory and an Intel Core i3-350M 2.266 GHz processor. The graphs to be searched were split among many different files so that the search could be run in more manageable segments and so that we did not overflow the laptop's memory. We chose to limit our search to graphs with fewer than 20 edges or fewer than 10 vertices due to time constraints. There are a total of 158 505 connected, nonplanar graphs that have 9 vertices and a minimum vertex degree of at least 2. Searching these graphs took about five hours. Since there are 9 229 423 such graphs on 10 vertices, searching these would take more than ten days. Similarly it took about three days to search all 7 753 990 connected, nonplanar graphs that have 19 edges and a minimum vertex degree of at least 2, so searching all 44 858 715 similar graphs on 20 edges is not feasible.

Next we reformat these graphs in each file produced to be read into Wolfram Mathematica. Then we use Mathematica functions to iterate over this list of graphs one file at a time and pull out any that are found to be either MMNE or MMNC. The code in Mathematica was run on a single Mathematica kernel (no attempt was made to parallelize the search in Mathematica). An overview of the method of testing if a graph  $G$  is MMNE is as follows, and an analogous method is used to test if a graph is MMNC:

- (1) For each  $e \in E(G)$ , if  $G - e$  is planar return false.
- (2) Build all the simple minors of  $G$  (the graphs in  $\{G - e, G/e \mid e \in E(G)\}$ ) and remove any duplicates (under isomorphism). If for any of these graphs there is no edge  $f$  such that  $G - f$  is planar, return false.
- (3) Take  $S = \{G\} \cup \{G - e \mid e \in E(G)\}$ . While  $S \neq \emptyset$ :

- (a) Reset  $S$  to the result of contracting each edge of each graph in  $S$ .
  - (b) Remove all planar graphs and duplicate graphs from  $S$ .
  - (c) If there exists  $G \in S$  such that  $G - e$  is nonplanar for each  $e \in E(G)$  then return false.
- (4) Return true.

We need step (3) explicitly because both of the properties edge apex and contraction apex are *not* closed under taking graph minors as shown in Theorems 5.2 and 5.4.

In addition to the 12 MMNE graphs that have been considered in this section, the brute-force search has found 15 more examples of MMNE graphs (listed in Section A.1 of the Appendix). Notable graphs in this list are  $K_{4,3}$ ,  $K_6 - e$ , the rook’s graph on 9 vertices, and some examples of MMNE graphs with degree-2 vertices. The brute-force search also found new examples of MMNC graphs in addition to the six graphs considered in this section. In particular, the computer demonstrated that the six MMNE graphs of connectivity 2 in Figure 8 are also MMNC. Along with these graphs there are 69 other MMNC graphs on 19 or fewer edges or 9 or fewer vertices. Section A.2 of the Appendix is an abridged list of these graphs (those on 17 or fewer edges or 9 or fewer vertices).

Beyond a simple brute-force search, we also conducted a more intelligent graph search using the knowledge that performing  $\nabla Y$  and  $Y \nabla$  moves on a graph has the potential to preserve the NE or NC property of that graph; see Theorem 5.19. The idea is that the  $\nabla Y$  or  $Y \nabla$  families of an MMNE or MMNC graph may contain new MMNE or MMNC graphs. The details of the methodology of this search, as well as the Mathematica code, can be found in [Pierce 2014]. In total, we have found 55 MMNE graphs and 82 MMNC graphs, and we suspect that there are many more of each. Tables 3 and 4 below give a classification of the MMNE and MMNC graphs we have found organized by graph size.

graph size ( $ E(G) $ )	$\leq 11$	12	13	14	15	16	17	18	19	20
number of MMNE graphs	0	1	0	2	0	2	3	11	6	$\geq 2$

graph size ( $ E(G) $ )	21	22	23	24	25	26	27	28	29	30
number of MMNE graphs	$\geq 13$	$\geq 7$	$\geq 4$	$\geq 2$	$\geq 0$	$\geq 0$	$\geq 1$	$\geq 0$	$\geq 0$	$\geq 1$

**Table 3.** The number of MMNE graphs we have found grouped by size. Note that this is a complete classification based on graph size up to and including size 19.

graph size ( $ E(G) $ )	$\leq 11$	12	13	14	15	16	17	18	19	20
number of MMNC graphs	0	1	0	0	1	6	14	32	25	$\geq 3$

**Table 4.** The number of MMNC graphs we have found grouped by size. Note that this is a complete classification based on graph size with the exception of size 20.

### Appendix: Edge lists of graphs found through computer searches

**A.1. MMNE graphs.** The following 15 MMNE graphs are the result of a computer search conducted on the set of graphs that have 19 or fewer edges or 9 or fewer vertices, and that all have a minimum vertex degree of at least 2. These graphs, together with eleven other graphs considered explicitly in the paper (i.e., all but  $K_5 \sqcup K_5$ , which has order 10 and size 20) make up all 26 MMNE graphs on 19 or fewer edges or on 9 or fewer vertices. (Note that Table 3 gives 25 graphs of size 19 or less. Adding the graph  $K_5 \dot{\cup} K_5$ , of order 9 and size 20, is what brings the total to 26.)

{(1, 8), (1, 9), (2, 4), (2, 7), (2, 8), (3, 6), (3, 7), (3, 8), (4, 5), (4, 6), (4, 8), (5, 6), (5, 7), (5, 9), (6, 7), (6, 9), (7, 9), (8, 9)}

{(1, 6), (1, 7), (2, 5), (2, 7), (3, 7), (3, 8), (3, 9), (4, 5), (4, 6), (4, 8), (4, 9), (5, 7), (5, 8), (5, 9), (6, 7), (6, 8), (6, 9), (8, 9)}

{(1, 8), (1, 9), (2, 6), (2, 7), (2, 9), (3, 5), (3, 7), (3, 9), (4, 5), (4, 6), (4, 9), (5, 6), (5, 7), (5, 8), (6, 7), (6, 8), (7, 8), (8, 9)}

{(1, 8), (1, 9), (2, 7), (2, 10), (3, 6), (3, 8), (3, 10), (4, 6), (4, 7), (4, 9), (5, 6), (5, 7), (5, 8), (6, 9), (6, 10), (7, 8), (7, 10), (8, 9), (9, 10)}

{(1, 9), (1, 10), (2, 7), (2, 8), (2, 10), (3, 7), (3, 8), (3, 9), (4, 6), (4, 8), (4, 10), (5, 6), (5, 7), (5, 9), (6, 7), (6, 8), (7, 10), (8, 9), (9, 10)}

{(1, 6), (1, 9), (2, 7), (2, 8), (3, 6), (3, 7), (3, 10), (4, 5), (4, 6), (4, 7), (4, 10), (5, 8), (5, 9), (5, 10), (6, 9), (7, 8), (8, 9), (8, 10), (9, 10)}

{(1, 8), (1, 10), (2, 4), (2, 8), (2, 9), (3, 4), (3, 5), (3, 9), (4, 5), (4, 6), (5, 7), (5, 10), (6, 7), (6, 8), (6, 9), (7, 9), (7, 10), (8, 10), (9, 10)}

{(1, 6), (1, 7), (1, 9), (2, 7), (2, 8), (2, 9), (3, 6), (3, 8), (3, 9), (4, 5), (4, 8), (4, 9), (5, 6), (5, 7), (5, 9), (6, 8), (7, 8)}

{(1, 7), (1, 8), (1, 9), (2, 6), (2, 8), (2, 9), (3, 6), (3, 7), (3, 9), (4, 6), (4, 7), (4, 8), (5, 6), (5, 7), (5, 8), (5, 9)}

{(1, 6), (1, 7), (1, 8), (2, 5), (2, 7), (2, 8), (3, 4), (3, 7), (3, 8), (4, 5),  
(4, 6), (4, 7), (4, 8), (5, 6), (5, 7), (5, 8), (6, 7), (6, 8)}

{(1, 6), (1, 7), (1, 9), (2, 5), (2, 7), (2, 8), (3, 7), (3, 8), (3, 9), (4, 5),  
(4, 6), (4, 8), (4, 9), (5, 7), (5, 9), (6, 7), (6, 8), (8, 9)}

{(1, 4), (1, 7), (1, 8), (2, 3), (2, 7), (2, 8), (3, 5), (3, 6), (4, 5), (4, 6),  
(5, 7), (5, 8), (6, 7), (6, 8)}

{(1, 5), (1, 6), (1, 7), (2, 5), (2, 6), (2, 7), (3, 5), (3, 6), (3, 7), (4, 5),  
(4, 6), (4, 7)}

{(1, 6), (1, 7), (1, 8), (1, 9), (2, 4), (2, 5), (2, 8), (2, 9), (3, 4), (3, 5),  
(3, 6), (3, 7), (4, 7), (4, 9), (5, 6), (5, 8), (6, 9), (7, 8)}

{(1, 3), (1, 4), (1, 5), (1, 6), (2, 3), (2, 4), (2, 5), (2, 6), (3, 4), (3, 5),  
(3, 6), (4, 5), (4, 6), (5, 6)}

**A.2. MMNC graphs.** The following 22 MMNC graphs are the result of a computer search conducted on the set of graphs that have 17 or fewer edges or 9 or fewer vertices, and that all have a minimum vertex degree of at least 2.

{(1, 9), (1, 12), (2, 8), (2, 11), (3, 6), (3, 7), (4, 5), (4, 10), (5, 11), (5, 12),  
(6, 9), (6, 11), (7, 8), (7, 12), (8, 10), (9, 10)}

{(1, 6), (1, 10), (2, 5), (2, 9), (3, 4), (3, 6), (3, 8), (4, 5), (4, 7), (5, 10),  
(6, 9), (7, 9), (7, 11), (8, 10), (8, 11), (9, 11), (10, 11)}

{(1, 6), (1, 10), (2, 7), (2, 8), (2, 9), (3, 6), (3, 8), (3, 9), (4, 7), (4, 9),  
(4, 10), (5, 7), (5, 8), (5, 10), (6, 7), (8, 10), (9, 10)}

{(1, 9), (1, 10), (2, 3), (2, 6), (2, 7), (3, 4), (3, 5), (4, 7), (4, 10), (5, 6),  
(5, 9), (6, 8), (6, 10), (7, 8), (7, 9), (8, 9), (8, 10)}

{(1, 9), (1, 11), (2, 9), (2, 10), (3, 4), (3, 6), (3, 11), (4, 5), (4, 10), (5, 8),  
(5, 9), (6, 7), (6, 9), (7, 10), (7, 11), (8, 10), (8, 11)}

{(1, 9), (1, 11), (2, 9), (2, 10), (3, 5), (3, 6), (3, 7), (4, 5), (4, 6), (4, 9),  
(5, 11), (6, 10), (7, 8), (7, 9), (8, 10), (8, 11), (10, 11)}

{(1, 4), (1, 11), (2, 6), (2, 9), (3, 5), (3, 6), (3, 7), (4, 5), (4, 9), (5, 10),  
(6, 11), (7, 9), (7, 10), (8, 9), (8, 10), (8, 11), (10, 11)}

{(1, 9), (1, 11), (2, 4), (2, 5), (2, 6), (3, 5), (3, 6), (3, 7), (4, 8), (4, 9),  
(5, 11), (6, 10), (7, 9), (7, 10), (8, 10), (8, 11), (10, 11)}

{(1, 10), (1, 11), (2, 3), (2, 7), (2, 9), (3, 6), (3, 8), (4, 5), (4, 9), (4, 10),  
(5, 8), (5, 11), (6, 7), (6, 11), (7, 10), (8, 10), (9, 11)}



{(1, 8), (1, 9), (2, 6), (2, 12), (3, 5), (3, 11), (4, 11), (4, 12), (5, 7), (5, 9),  
(6, 7), (6, 8), (7, 10), (8, 11), (9, 12), (10, 11), (10, 12)}

{(1, 9), (1, 11), (2, 5), (2, 12), (3, 4), (3, 12), (4, 8), (4, 9), (5, 7), (5, 9),  
(6, 7), (6, 8), (6, 11), (7, 10), (8, 10), (10, 12), (11, 12)}

{(1, 4), (1, 8), (1, 9), (2, 3), (2, 8), (2, 9), (3, 4), (3, 6), (3, 9), (4, 5),  
(4, 8), (5, 6), (5, 7), (5, 9), (6, 7), (6, 8), (7, 8), (7, 9)}

{(1, 4), (1, 8), (1, 9), (2, 4), (2, 7), (2, 9), (3, 4), (3, 6), (3, 9), (5, 6),  
(5, 7), (5, 8), (5, 9), (6, 7), (6, 8), (7, 8)}

{(1, 5), (1, 6), (1, 8), (2, 3), (2, 4), (2, 7), (3, 6), (3, 10), (4, 5), (4, 10),  
(5, 9), (6, 9), (7, 9), (7, 10), (8, 9), (8, 10)}

{(1, 5), (1, 6), (1, 8), (2, 3), (2, 4), (2, 7), (3, 6), (3, 10), (4, 5), (4, 9),  
(5, 10), (6, 9), (7, 9), (7, 10), (8, 9), (8, 10)}

{(1, 2), (1, 9), (1, 10), (2, 7), (2, 8), (3, 8), (3, 9), (3, 10), (4, 7), (4, 9),  
(4, 10), (5, 7), (5, 8), (5, 10), (6, 7), (6, 8), (6, 9)}

{(1, 2), (1, 4), (1, 10), (2, 3), (2, 9), (3, 4), (3, 7), (4, 8), (5, 7), (5, 8),  
(5, 10), (6, 7), (6, 8), (6, 9), (7, 10), (8, 9), (9, 10)}

{(1, 5), (1, 6), (1, 7), (2, 5), (2, 6), (2, 7), (3, 5), (3, 6), (3, 7), (4, 5),  
(4, 6), (4, 7)}

{(1, 2), (1, 4), (1, 7), (1, 9), (2, 3), (2, 6), (2, 8), (3, 5), (3, 6), (3, 9),  
(4, 5), (4, 7), (4, 8), (5, 8), (5, 9), (6, 8), (6, 9), (7, 8), (7, 9)}

{(1, 6), (1, 7), (1, 8), (1, 9), (2, 4), (2, 5), (2, 8), (2, 9), (3, 4), (3, 5),  
(3, 6), (3, 7), (4, 7), (4, 9), (5, 6), (5, 8), (6, 9), (7, 8)}

{(1, 5), (1, 6), (1, 7), (1, 8), (2, 3), (2, 4), (2, 7), (2, 8), (3, 4), (3, 6),  
(3, 8), (4, 5), (4, 8), (5, 6), (5, 7), (6, 7)}

{(1, 2), (1, 3), (1, 4), (1, 5), (1, 6), (2, 3), (2, 4), (2, 5), (2, 6), (3, 4),  
(3, 5), (3, 6), (4, 5), (4, 6), (5, 6)}

### Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant Number 1156612. We received additional support through a Research and Creativity Award from the Provost's office at CSU, Chico as well as Math Summer Research Internships from the Math Department. We thank Ramin Naimi and Bojan Mohar for helpful conversations.

## References

- [Ayala 2014] H. Ayala, *MMNA graphs on eight vertices or fewer*, master's thesis, California State University, Chico, 2014, available at <http://www.csuchico.edu/~tmattman/HAThesis.pdf>.
- [Barsotti and Mattman 2016] J. Barsotti and T. W. Mattman, "Graphs on 21 edges that are not 2-apex", *Involve* **9**:4 (2016), 591–621. MR Zbl
- [Cabello and Mohar 2013] S. Cabello and B. Mohar, "Adding one edge to planar graphs makes crossing number and 1-planarity hard", *SIAM J. Comput.* **42**:5 (2013), 1803–1829. MR Zbl
- [Diestel 2010] R. Diestel, *Graph theory*, 4th ed., Graduate Texts in Mathematics **173**, Springer, 2010. MR Zbl
- [Gubser 1996] B. S. Gubser, "A characterization of almost-planar graphs", *Combin. Probab. Comput.* **5**:3 (1996), 227–245. MR Zbl
- [Kuratowski 1930] C. Kuratowski, "Sur le problème des courbes gauches in topologie", *Fund. Math* **15**:1 (1930), 271–283. JFM
- [Mader 1998] W. Mader, " $3n - 5$  edges do force a subdivision of  $K_5$ ", *Combinatorica* **18**:4 (1998), 569–595. MR Zbl
- [McKay and Piperno 2014] B. D. McKay and A. Piperno, "Practical graph isomorphism, II", *J. Symbolic Comput.* **60** (2014), 94–112. MR Zbl
- [Mohar and Thomassen 2001] B. Mohar and C. Thomassen, *Graphs on surfaces*, Johns Hopkins University Press, Baltimore, MD, 2001. MR Zbl
- [Pierce 2014] M. Pierce, *Searching for and classifying the finite set of minor-minimal non-apex graphs*, honor's thesis, California State University, Chico, 2014, available at <http://www.csuchico.edu/~tmattman/mpthesis.pdf>.
- [Robertson and Seymour 2004] N. Robertson and P. D. Seymour, "Graph minors, XX: Wagner's conjecture", *J. Combin. Theory Ser. B* **92**:2 (2004), 325–357. MR Zbl
- [Wagner 1937] K. Wagner, "Über eine Eigenschaft der ebenen Komplexe", *Math. Ann.* **114**:1 (1937), 570–590. MR JFM
- [Wagner 1967] K. Wagner, "Fastplättbare Graphen", *J. Combinatorial Theory* **3** (1967), 326–365. MR Zbl

Received: 2015-02-28      Revised: 2016-08-09      Accepted: 2017-05-22

ml2437@cornell.edu	<i>Department of Mathematics, Cornell University, Ithaca, NY, United States</i>
eoinmackall@yahoo.com	<i>Department of Mathematics and Statistics, California State University, Chico, CA, United States</i>
tmattman@csuchico.edu	<i>Department of Mathematics and Statistics, California State University, Chico, CA, United States</i>
mpierce9@mail.csuchico.edu	<i>Department of Mathematics and Statistics, California State University, Chico, CA, United States</i>
mrsrobinsonmath@gmail.com	<i>Etna High School, Etna, CA, United States</i>
jthomas72@mail.csuchico.edu	<i>Department of Mathematics and Statistics, California State University, Chico, CA, United States</i>
iweinschelba@wesleyan.edu	<i>Department of Mathematics and Computer Science, Wesleyan University, Middletown, CT, United States</i>

# Nested Frobenius extensions of graded superrings

Edward Poon and Alistair Savage

(Communicated by Kenneth S. Berenhaut)

We prove a nesting phenomenon for twisted Frobenius extensions. Namely, suppose  $R \subseteq B \subseteq A$  are graded superrings such that  $A$  and  $B$  are both twisted Frobenius extensions of  $R$ ,  $R$  is contained in the center of  $A$ , and  $A$  is projective over  $B$ . Our main result is that, under these assumptions,  $A$  is a twisted Frobenius extension of  $B$ . This generalizes a result of Pike and the second author, which considered the case where  $R$  is a field.

## 1. Introduction

Frobenius extensions, which are a natural generalization of Frobenius algebras, appear frequently in many areas of mathematics, from topological quantum field theory to categorical representation theory. Several generalizations of Frobenius extensions have been introduced since their inception. In particular, Nakayama and Tsuzuku [1960] introduced Frobenius extensions of the second kind. These were further generalized to the concept of  $(\alpha, \beta)$ -Frobenius extensions in [Morita 1965], where  $\alpha$  and  $\beta$  are automorphisms of the rings involved. The corresponding theory for graded superrings was then developed in [Pike and Savage 2016], where they were called *twisted Frobenius extensions*.

In the literature, one finds that many examples of (twisted) Frobenius extensions arise from certain types of subobjects. For instance, if  $H$  is a finite-index subgroup of  $G$ , then the group ring  $R[G]$  is a Frobenius extension of  $R[H]$ , where  $R$  is a commutative base ring. This example dates back to the original paper [Kasch 1954] on Frobenius extensions. Another example comes from the theory of Hopf algebras. In particular, it was shown in [Schneider 1992, Corollary 3.6(1)] that if  $K$  is a Hopf subalgebra of  $H$ , then  $H$  is a Frobenius extension of  $K$  of the second kind. Yet another example comes from Frobenius algebras themselves. Namely, it was shown (in the more general graded super setting) in [Pike and Savage 2016, Corollary 7.4] that if  $A$  is a Frobenius algebra over a field,  $B$  is a subalgebra of  $A$  that is also a Frobenius algebra, and  $A$  is projective over  $B$ , then  $A$  is a twisted Frobenius extension of  $B$ .

---

MSC2010: 17A70, 16W50.

Keywords: Frobenius extension, Frobenius algebra, graded superring, graded superalgebra.

The goal of the current paper is to shed more light on this “nesting” phenomenon. Namely, we consider the situation where we have graded superrings  $R \subseteq B \subseteq A$  such that  $A$  and  $B$  are both twisted Frobenius extensions of  $R$ , and  $R$  is contained in the center of  $A$ . We call these *nested Frobenius extensions*. We first prove that these assumptions imply  $A$  and  $B$  are *untwisted* Frobenius extensions of  $R$  (see Corollary 3.2). Then, our main result (Theorem 3.8) is that, provided  $A$  is projective over  $B$ , it follows that  $A$  is a twisted Frobenius extension of  $B$ . The twisting is given in terms of the Nakayama automorphisms of  $A$  and  $B$ . In particular, even though  $A$  and  $B$  are untwisted Frobenius extensions of  $R$ ,  $A$  can be a nontrivially twisted Frobenius extension of  $B$ . This result can be viewed as a generalization of [Pike and Savage 2016, Corollary 7.4] to the setting of arbitrary supercommutative ground rings.

The organization of the paper is as follows. We begin in Section 2 by recalling the definition of twisted Frobenius extensions of graded superrings, together with some related results. In Section 3, we examine nested Frobenius extensions  $R \subseteq B \subseteq A$ , where  $R$  is contained in the center of  $A$ . We begin by proving that  $A$  and  $B$  are, in fact, *untwisted* Frobenius extensions of  $R$  (Corollary 3.2). Then, after establishing several important lemmas, we prove our main result (Theorem 3.8), that  $A$  is a twisted Frobenius extension of  $B$ , provided  $A$  is projective over  $B$ . We conclude in Section 4 with several applications of our main result. In particular, we explain how the aforementioned examples of group rings and Hopf algebras can be deduced from our main theorem. We also give an example arising from nilcoxeter rings.

**Note on the arXiv version.** For the interested reader, the tex file of the arXiv version of this paper includes hidden details of some straightforward computations and arguments that are omitted in the pdf. These details can be displayed by switching the `details` toggle to true in the tex file and recompiling.

## 2. Twisted Frobenius extensions

In this section we recall the definition of twisted Frobenius extensions, together with some of their properties that will be used in this paper. We refer the reader to [Pike and Savage 2016] for further details.

Fix an abelian group  $\Lambda$  and by *graded*, we mean  $\Lambda$ -graded. In particular, a graded superring is a  $\Lambda \times \mathbb{Z}_2$ -graded ring. In other words, if  $A$  is a graded superring, then

$$A = \bigoplus_{\lambda \in \Lambda, \pi \in \mathbb{Z}_2} A_{\lambda, \pi}, \quad A_{\lambda, \pi} A_{\lambda', \pi'} \subseteq A_{\lambda + \lambda', \pi + \pi'}, \quad \lambda, \lambda' \in \Lambda, \pi, \pi' \in \mathbb{Z}_2.$$

We denote the multiplicative unit of  $A$  by  $1_A$ . To avoid repeated use of the modifiers “graded” and “super”, from now on we will use the term *ring* to mean graded superring and *subring* to mean graded subsuperring. Similarly, by an automorphism

of a ring, we mean an automorphism as graded superrings (homogeneous of degree zero).

We will use the term *module* to mean graded supermodule. In particular, a left  $A$ -module  $M$  is a  $\Lambda \times \mathbb{Z}_2$ -graded abelian group with a left  $A$ -action such that

$$A_{\lambda,\pi} M_{\lambda',\pi'} \subseteq M_{\lambda+\lambda',\pi+\pi'}, \quad \lambda, \lambda' \in \Lambda, \pi, \pi' \in \mathbb{Z}_2,$$

and similarly for right modules. If  $v$  is a homogeneous element in a ring or module, we will denote by  $|v|$  its  $\Lambda$ -degree and by  $\bar{v}$  its  $\mathbb{Z}_2$ -degree. Whenever we write an expression involving degrees of elements, we will implicitly assume that such elements are homogeneous.

For  $M, N$  two  $\Lambda \times \mathbb{Z}_2$ -graded abelian groups, we define a  $\Lambda \times \mathbb{Z}_2$ -grading on the space  $\text{HOM}_{\mathbb{Z}}(M, N)$  of all  $\mathbb{Z}$ -linear maps by setting  $\text{HOM}_{\mathbb{Z}}(M, N)_{\lambda,\pi}$ ,  $\lambda \in \Lambda$ ,  $\pi \in \mathbb{Z}_2$ , to be the subspace of all homogeneous maps of degree  $(\lambda, \pi)$ . That is,

$$\begin{aligned} \text{HOM}_{\mathbb{Z}}(M, N)_{\lambda,\pi} \\ = \{f \in \text{HOM}_{\mathbb{Z}}(M, N) \mid f(M_{\lambda',\pi'}) \subseteq N_{\lambda+\lambda',\pi+\pi'} \text{ for all } \lambda' \in \Lambda, \pi' \in \mathbb{Z}_2\}. \end{aligned}$$

For  $A$ -modules  $M$  and  $N$ , we define the  $\Lambda \times \mathbb{Z}_2$ -graded abelian group

$$\text{HOM}_A(M, N) = \bigoplus_{\lambda \in \Lambda, \pi \in \mathbb{Z}_2} \text{HOM}_A(M, N)_{\lambda,\pi},$$

where the homogeneous components are defined by

$$\begin{aligned} \text{HOM}_A(M, N)_{\lambda,\pi} \\ = \{f \in \text{HOM}_{\mathbb{Z}}(M, N)_{\lambda,\pi} \mid f(am) = (-1)^{\pi \bar{a}} af(m) \text{ for all } a \in A, m \in M\}. \end{aligned}$$

We let  $A\text{-mod}$  denote the category of left  $A$ -modules, with set of morphisms from  $M$  to  $N$  given by  $\text{HOM}_A(M, N)_{0,0}$ . Similarly, we have the category of right  $A$ -modules with morphisms from  $M$  to  $N$  given by

$$\{f \in \text{HOM}_{\mathbb{Z}}(M, N)_{0,0} \mid f(ma) = f(m)a \text{ for all } m \in M, a \in A\}.$$

We will call elements of  $\text{HOM}_A(M, N)_{\lambda,\pi}$  *homomorphisms of degree  $(\lambda, \pi)$*  and, if they are invertible, *isomorphisms of degree  $(\lambda, \pi)$* . Note that they are not morphisms in the category  $A\text{-mod}$  unless they are of degree  $(0, 0)$ . We use similar terminology for right modules.

If  $M$  is a left  $A$ -module, we let  ${}^{\ell}a$  denote the operator given by the left action by  $a$ ; that is,

$${}^{\ell}a(m) = am, \quad a \in A, m \in M. \tag{2-1}$$

If  $M$  is a right  $A$ -module, then for each homogeneous  $a \in A$ , we define a  $\mathbb{Z}$ -linear operator

$${}^r a: M \rightarrow M, \quad {}^r a(m) = (-1)^{\bar{a}\bar{m}} ma, \quad a \in A, m \in M. \tag{2-2}$$

If  $A_1$  and  $A_2$  are rings, then, by definition, an  $(A_1, A_2)$ -bimodule  $M$  is both a left  $A_1$ -module and a right  $A_2$ -module such that the left and right actions commute:

$$(a_1 m) a_2 = a_1 (m a_2) \quad \text{for all } a_1 \in A_1, a_2 \in A_2, m \in M.$$

If  $M$  is an  $(A_1, A_2)$ -bimodule and  $N$  is a left  $A_1$ -module, then  $\text{HOM}_{A_1}(M, N)$  is a left  $A_2$ -module via the action

$$a \cdot f = (-1)^{\bar{a}\bar{f}} f \circ {}^r a, \quad a \in A_2, f \in \text{HOM}_{A_1}(M, N), \tag{2-3}$$

and  $\text{HOM}_{A_1}(N, M)$  is a right  $A_2$ -module via the action

$$f \cdot a = (-1)^{\bar{a}\bar{f}} ({}^r a) \circ f, \quad a \in A_2, f \in \text{HOM}_{A_1}(N, M). \tag{2-4}$$

For  $\lambda \in \Lambda$ ,  $\pi \in \mathbb{Z}_2$ , and an  $A$ -module  $M$ , we let  $\{\lambda, \pi\}M$  denote the  $\Lambda \times \mathbb{Z}_2$ -graded abelian group that has the same underlying abelian group as  $M$ , but a new grading given by  $(\{\lambda, \pi\}M)_{\lambda', \pi'} = M_{\lambda' - \lambda, \pi' - \pi}$ . Abusing notation, we will also sometimes use  $\{\lambda, \pi\}$  to denote the map  $M \rightarrow \{\lambda, \pi\}M$  that is the identity on elements of  $M$ . We define a left action of  $A$  on  $\{\lambda, \pi\}M$  by  $a \cdot \{\lambda, \pi\}m = (-1)^{\pi\bar{a}} \{\lambda, \pi\}am$ . In this way,  $\{\lambda, \pi\}$  defines a functor from the category of  $A$ -modules to itself that leaves morphisms unchanged.

Suppose  $M$  is a left  $A$ -module,  $N$  is a right  $A$ -module, and  $\alpha$  is a ring automorphism of  $A$ . Then we can define the twisted left  $A$ -module  ${}^\alpha M$  and twisted right  $A$ -module  $N^\alpha$  to be equal to  $M$  and  $N$ , respectively, as graded abelian groups, but with actions given by

$$a \cdot m = \alpha(a)m, \quad a \in A, m \in {}^\alpha M, \tag{2-5}$$

$$n \cdot a = n\alpha(a), \quad a \in A, n \in N^\alpha, \tag{2-6}$$

where juxtaposition denotes the original action of  $A$  on  $M$  and  $N$ . If  $\alpha$  is a ring automorphism of  $A$ , and  $B$  is a subring of  $A$ , then we will also use the notation  ${}^\alpha_B A_A$  to denote the  $(B, A)$ -bimodule equal to  $A$  as a graded abelian group, with right action given by multiplication, and with left action given by  $b \cdot a = \alpha(b)a$  (where here juxtaposition is multiplication in the ring  $A$ ), even though  $\alpha$  is not necessarily a ring automorphism of  $B$ . We use  ${}_A A_B^\alpha$  for the obvious right analogue. By convention, when we consider twisted modules as above, operators such as  ${}^r a$  and  ${}^l a$  defined in (2-1) and (2-2) involve the right and left action (respectively) in the *original* (i.e., untwisted) module.

**Definition 2.1** (twisted Frobenius extension). Suppose  $B$  is a subring of a ring  $A$ , that  $\alpha$  is a ring automorphism of  $A$ , and that  $\beta$  is a ring automorphism of  $B$ . Furthermore, suppose  $\lambda \in \Lambda$  and  $\pi \in \mathbb{Z}_2$ . We call  $A$  an  $(\alpha, \beta)$ -Frobenius extension of  $B$  of degree  $(-\lambda, \pi)$  if  $A$  is finitely generated and projective as a left  $B$ -module,

and there is a morphism of  $(B, B)$ -bimodules

$$\text{tr}: {}_B^\beta A_B^\alpha \rightarrow \{\lambda, \pi\}_B B_B$$

satisfying the following two conditions:

(T1) If  $\text{tr}(Aa) = 0$  for some  $a \in A$ , then  $a = 0$ .

(T2) For every  $\varphi \in \text{HOM}_B({}_B^\beta A, \{\lambda, \pi\}_B B)$ , there exists an  $a \in A$  such that  $\varphi = \text{tr} \circ r a$ .

The map  $\text{tr}$  is called a *trace map*. We will often view it as a map  ${}_B^\beta A_A^\alpha \rightarrow {}_B B_B$  that is homogeneous of degree  $(-\lambda, \pi)$ . If  $A$  is an  $(\alpha, \beta)$ -Frobenius extension of  $B$  for some  $\alpha$  and  $\beta$ , we say that  $A$  is a *twisted Frobenius extension* of  $B$ . If  $A$  is an  $(\text{id}_A, \text{id}_B)$ -Frobenius extension of  $B$ , we call it a *Frobenius extension* or *untwisted Frobenius extension* (when we wish to emphasize that the twistings are trivial).

**Remark 2.2.** We say the extension is of degree  $(-\lambda, \pi)$  since that is the degree of the trace map. If  $A$  and  $B$  are concentrated in degree  $(0, 0)$ , then  $(\alpha, \beta)$ -Frobenius extensions were defined in [Morita 1965, p. 41]. In particular, an  $(\text{id}_A, \beta)$ -Frobenius extension is sometimes called a  $\beta^{-1}$ -*extension*, or a *Frobenius extension of the second kind*; see [Nakayama and Tsuzuku 1960].

If  $B$  is a subring of a ring  $A$ , then we define the *centralizer* of  $B$  in  $A$  to be the subring of  $A$  given by

$$C_A(B) = \{a \in A \mid ab = (-1)^{\bar{a}\bar{b}} ba \text{ for all } b \in B\}. \tag{2-7}$$

If  $A$  is an  $(\alpha, \beta)$ -Frobenius extension of  $B$ , then we have the associated *Nakayama isomorphism* (an isomorphism of rings)

$$\psi: C_A(B) \rightarrow C_A(\alpha(B)),$$

which is the unique map satisfying

$$\text{tr}(ca) = (-1)^{\bar{a}\bar{c}} \text{tr}(a\psi(c)) \quad \text{for all } a \in A, c \in C_A(B). \tag{2-8}$$

**Proposition 2.3.** *The ring  $B$  is an untwisted Frobenius extension of  $R$  of degree  $(-\lambda, \pi)$  if and only if there exists a homomorphism of  $(R, R)$ -bimodules  $\text{tr}: B \rightarrow R$  of degree  $(-\lambda, \pi)$ , and finite subsets  $\{x_1, \dots, x_n\}, \{y_1, \dots, y_n\}$  of  $B$  such that  $(|y_i|, \bar{y}_i) = (\lambda - |x_i|, \pi - \bar{x}_i)$  for  $i = 1, \dots, n$ , and*

$$b = (-1)^{\pi\bar{b}} \sum_{i=1}^n (-1)^{\pi\bar{x}_i} \text{tr}(by_i)x_i = \sum_{i=1}^n y_i \text{tr}(x_i b) \quad \text{for all } b \in B. \tag{2-9}$$

We call the sets  $\{x_1, \dots, x_n\}$  and  $\{y_1, \dots, y_n\}$  dual sets of generators of  $B$  over  $R$ .

*Proof.* This is a special case of [Pike and Savage 2016, Proposition 4.9], where the twistings are trivial. □

### 3. Nested Frobenius extensions

In this section, we introduce our main object of study, nested Frobenius extensions, and prove our main result (Theorem 3.8). We begin with a simplification result.

**Lemma 3.1.** *Suppose  $A$  is an  $(\alpha, \beta)$ -Frobenius extension of  $R$  of degree  $(-\lambda, \pi)$ , with trace map  $\text{tr}$  and Nakayama isomorphism  $\psi$ . Furthermore, suppose  $C_A(R) = A$ . Then  $\alpha|_R = \beta$  and  $\psi|_R = \text{id}_R$ .*

*Proof.* For all  $r \in R$  and  $a \in {}^\beta_R A_R^\alpha$ , we have

$$\begin{aligned} \text{tr}(ar) &= \text{tr}(a)\alpha^{-1}(r) = (-1)^{\bar{r}(\pi+\bar{a})}\alpha^{-1}(r)\text{tr}(a) \\ &= (-1)^{\bar{r}\bar{a}}\text{tr}(\beta(\alpha^{-1}(r))a) = \text{tr}(a\beta(\alpha^{-1}(r))), \end{aligned}$$

where the second and fourth equalities follow from the fact that  $C_A(R) = A$ . Since the trace map is linear, this implies

$$\text{tr}(a(r - \beta(\alpha^{-1}(r)))) = 0 \quad \text{for all } a \in {}^\beta_R A_R^\alpha.$$

By (T1), we have  $\beta(\alpha^{-1}(r)) = r$  for all  $r \in R$ . It follows that  $\alpha|_R = \beta$ .

Similarly, for all  $r \in R$  and  $a \in {}^\beta_R A_R^\alpha$ , we have

$$\text{tr}(ar) = (-1)^{\bar{r}\bar{a}}\text{tr}(ra) = \text{tr}(a\psi(r)),$$

and so  $\psi|_R = \text{id}_R$  by (T1). □

**Corollary 3.2.** *If  $A$  is a twisted Frobenius extension of  $R$  and  $C_A(R) = A$ , then  $A$  is in fact an **untwisted** Frobenius extension of  $R$  of the same degree.*

*Proof.* Suppose  $A$  is an  $(\alpha, \beta)$ -Frobenius extension of  $R$  of degree  $(-\lambda, \pi)$ , with trace map  $\text{tr}$  and Nakayama isomorphism  $\psi$ . Furthermore, suppose that  $C_A(R) = A$ . Then, by Lemma 3.1,  $A$  is an  $(\alpha, \alpha)$ -Frobenius extension of  $R$  and  $\alpha(R) = \beta(R) = R$ . The result then follows immediately from [Pike and Savage 2016, Corollary 3.6]. □

For the remainder of this paper, we fix rings

$$R \subseteq B \subseteq A, \quad \text{with } C_A(R) = A.$$

This implies  $C_B(R) = B$  and  $C_R(R) = R$ . In particular,  $R$  is supercommutative, and so we do not distinguish between left and right  $R$ -modules. In light of Corollary 3.2, we suppose that  $A$  and  $B$  are untwisted Frobenius extensions of  $R$  of degrees  $(-\lambda_A, \pi_A)$  and  $(-\lambda_B, \pi_B)$ , respectively. We denote their trace maps by  $\text{tr}_A$  and  $\text{tr}_B$  and their Nakayama isomorphisms by  $\psi_A$  and  $\psi_B$ , respectively. We call  $A$  and  $B$  *nested Frobenius extensions* of  $R$ .

**Remark 3.3.** The assumption that  $C_A(R) = A$  implies that  $\psi_A$  and  $\psi_B$  are ring automorphisms of  $A$  and  $B$ , respectively. In fact, this is precisely why we assume  $C_A(R) = A$ .



For an  $R$ -module  $M$ , we define

$$M^\vee = \text{HOM}_R(M, R).$$

If, in addition,  $M$  is a  $(B, A)$ -bimodule, then it is straightforward to verify that  $M^\vee$  is an  $(A, B)$ -bimodule with action given by

$$a \cdot f \cdot b = (-1)^{\bar{a}\bar{f}} f \circ {}^r a \circ {}^\ell b = (-1)^{\bar{a}\bar{f} + \bar{a}\bar{b}} f \circ {}^\ell b \circ {}^r a, \quad a \in A, b \in B, f \in M^\vee. \quad (3-1)$$

Note that  $B$  is naturally a  $(B, B)$ -bimodule via left and right multiplication. We denote this bimodule by  ${}_B B_B$  to emphasize the actions. Therefore, if  $M$  is a  $(B, A)$ -bimodule,  $\text{HOM}_B(M, {}_B B_B)$  is an  $(A, B)$ -bimodule via the actions (2-3) and (2-4).

**Lemma 3.4.** *For any  $(B, A)$ -bimodule  $M$ , the map*

$$\text{HOM}_B({}^{\psi_B} M, {}_B B_B) \rightarrow M^\vee, \quad f \mapsto \text{tr}_B \circ f, \quad (3-2)$$

*is a homomorphism of  $(A, B)$ -bimodules of degree  $(-\lambda_B, \pi_B)$ .*

*Proof.* By Lemma 3.1, we have  $\psi_B(r) = r$  for all  $r \in R$ . Thus, any element  $f \in \text{HOM}_B({}^{\psi_B} M, B)$  is also an element of  $\text{HOM}_R(M, B)$ , and hence  $\text{tr}_B \circ f \in M^\vee$ . The map (3-2) is also clearly of degree  $(-\lambda_B, \pi_B)$ , since  $\text{tr}_B$  is.

It remains to show that (3-2) is a homomorphism of  $(A, B)$ -bimodules. It is clearly a homomorphism of abelian groups. For  $a \in A$  and  $f \in \text{HOM}_B({}^{\psi_B} M, B)$ , we have

$$\text{tr}_B \circ (a \cdot f) = (-1)^{\bar{a}\bar{f}} \text{tr}_B \circ f \circ {}^r a = (-1)^{\bar{a}\pi_B} a \cdot (\text{tr}_B \circ f).$$

Thus, (3-2) is a homomorphism of left  $A$ -modules. Now let  $b \in B$  and  $y \in {}^{\psi_B} M_A$ . Then

$$\begin{aligned} \text{tr}_B \circ (f \cdot b)(y) &= (-1)^{\bar{b}\bar{f}} \text{tr}_B \circ ({}^r b \circ f)(y) \\ &= (-1)^{\bar{b}\bar{f}} \text{tr}_B \circ ({}^r b(f(y))) \\ &= (-1)^{\bar{b}\bar{y}} \text{tr}_B(f(y)b) \\ &= (-1)^{\bar{b}\bar{f}} \text{tr}_B(\psi_B^{-1}(b)f(y)) \\ &= \text{tr}_B(f(by)) = \text{tr}_B \circ f \circ {}^\ell b(y) = ((\text{tr}_B \circ f) \cdot b)(y). \end{aligned}$$

Thus the map (3-2) is also a homomorphism of right  $B$ -modules. □

Let

$$\{x_i\}_{i=1}^n \quad \text{and} \quad \{y_i\}_{i=1}^n$$

be dual sets of generators of  $B$  over  $R$ , where  $|x_i| + |y_i| = \lambda_B$  and  $\bar{x}_i + \bar{y}_i = \pi_B$  for each  $i = 1, \dots, n$  (see Proposition 2.3).

**Proposition 3.5.** *If  $M$  is a  $(B, A)$ -bimodule, then the map*

$$M^\vee \rightarrow \text{HOM}_B({}^\psi B M, {}_B B_B), \quad (3-3)$$

$$\theta \mapsto \left( m \mapsto (-1)^{\pi_B(\bar{\theta}+\bar{m})} \sum_{i=1}^n (-1)^{\bar{y}_i(\pi_B+\bar{m})} \theta(y_i m) x_i \right),$$

is a homomorphism of  $(A, B)$ -bimodules of degree  $(\lambda_B, \pi_B)$ . Moreover, the maps (3-2) and (3-3) are mutually inverse isomorphisms of  $(A, B)$ -bimodules.

*Proof.* The map

$$m \mapsto (-1)^{\pi_B(\bar{\theta}+\bar{m})} \sum_{i=1}^n (-1)^{\bar{y}_i(\pi_B+\bar{m})} \theta(y_i m) x_i \quad (3-4)$$

is clearly a homomorphism of abelian groups. Now let  $b \in B$  and  $m \in {}^\psi B M$ . Then  $b \cdot m = \psi_B(b)m$  maps to

$$\begin{aligned} & (-1)^{\pi_B(\bar{\theta}+\bar{b}+\bar{m})} \sum_{i=1}^n (-1)^{\bar{y}_i(\pi_B+\bar{b}+\bar{m})} \theta(y_i \psi_B(b)m) x_i \\ \stackrel{(2-9)}{=} & (-1)^{\pi_B(\bar{\theta}+\bar{b}+\bar{m})} \sum_{i=1}^n (-1)^{\bar{y}_i(\pi_B+\bar{b}+\bar{m})} \theta \left( \sum_{j=1}^n y_j \text{tr}_B(x_j y_i \psi_B(b)) m \right) x_i \\ = & (-1)^{\pi_B(\bar{\theta}+\bar{b}+\bar{m})} \sum_{i,j=1}^n (-1)^{\bar{y}_i(\pi_B+\bar{b}+\bar{m})+\bar{y}_j(\pi_B+\bar{x}_j+\bar{y}_i+\bar{b})} \theta(\text{tr}_B(x_j y_i \psi_B(b)) y_j m) x_i \\ = & (-1)^{\pi_B(\bar{\theta}+\bar{b}+\bar{m})} \sum_{i,j=1}^n (-1)^{\bar{y}_i(\pi_B+\bar{b}+\bar{m})+(\bar{y}_j+\bar{\theta})(\pi_B+\bar{x}_j+\bar{y}_i+\bar{b})} \text{tr}_B(x_j y_i \psi_B(b)) \theta(y_j m) x_i \\ = & (-1)^{\pi_B(\bar{\theta}+\bar{b})} \sum_{i,j=1}^n (-1)^{\bar{y}_i(\pi_B+\bar{b})+\bar{m}(\bar{x}_j+\bar{b})} \theta(y_j m) \text{tr}_B(x_j y_i \psi_B(b)) x_i \\ = & (-1)^{\pi_B(\bar{\theta}+\bar{b})} \sum_{i,j=1}^n (-1)^{\bar{y}_i \pi_B + \bar{m}(\bar{x}_j + \bar{b}) + \bar{b} \bar{x}_j} \theta(y_j m) \text{tr}_B(b x_j y_i) x_i \\ \stackrel{(2-9)}{=} & (-1)^{\pi_B(\bar{\theta}+\bar{b})} \sum_{j=1}^n (-1)^{\pi_B \bar{y}_j + \bar{m}(\bar{x}_j + \bar{b}) + \bar{b} \bar{y}_j} \theta(y_j m) b x_j \\ = & (-1)^{\bar{b}(\bar{\theta}+\pi_B) + \pi_B \bar{\theta}} \sum_{j=1}^n (-1)^{\pi_B \bar{y}_j + \bar{m} \bar{x}_j} b \theta(y_j m) x_j \\ = & (-1)^{\bar{b}(\bar{\theta}+\pi_B)} b \left( (-1)^{\pi_B(\bar{\theta}+\bar{m})} \sum_{j=1}^n (-1)^{\bar{y}_j(\pi_B+\bar{m})} \theta(y_j m) x_j \right). \end{aligned}$$

Thus (3-4) is a homomorphism of left  $B$ -modules of degree  $(\lambda_B, \pi_B)$ . Since the (set-theoretic) inverse of a bimodule homomorphism is also a bimodule homomorphism, it remains to show that (3-2) and (3-3) are mutually inverse.

Let  $f \in \text{HOM}_B({}^{\psi_B}M, {}_B B_B)$ . The map (3-2) followed by (3-3) sends  $f$  to the map

$$\begin{aligned} m \mapsto & (-1)^{\pi_B(\pi_B + \bar{f} + \bar{m})} \sum_{i=1}^n (-1)^{\bar{y}_i(\pi_B + \bar{m})} \text{tr}_B(f(y_i m)) x_i \\ &= (-1)^{\pi_B(\pi_B + \bar{f} + \bar{m})} \sum_{i=1}^n (-1)^{\bar{y}_i(\pi_B + \bar{m} + \bar{f})} \text{tr}_B(\psi_B^{-1}(y_i) f(m)) x_i \\ &= (-1)^{\pi_B(\bar{f} + \bar{m})} \sum_{i=1}^n (-1)^{\pi_B \bar{x}_i} \text{tr}_B(f(m) y_i) x_i \stackrel{(2-9)}{=} f(m). \end{aligned}$$

Thus (3-3) is left inverse to (3-2).

Now let  $\theta \in M^\vee$ . The map (3-3) followed by the map (3-2) sends  $\theta$  to the map

$$\begin{aligned} m \mapsto & (-1)^{\pi_B(\bar{\theta} + \bar{m})} \sum_{i=1}^n (-1)^{\bar{y}_i(\pi_B + \bar{m})} \text{tr}_B(\theta(y_i m)) x_i \\ &= \sum_{i=1}^n (-1)^{\bar{y}_i \bar{m}} \theta(y_i m) \text{tr}_B(x_i) = \sum_{i=1}^n (-1)^{\bar{y}_i(\bar{\theta} + \bar{y}_i)} \text{tr}_B(x_i) \theta(y_i m) \\ &= \sum_{i=1}^n (-1)^{\bar{y}_i} \theta(\text{tr}_B(x_i) y_i m) = \sum_{i=1}^n \theta(y_i \text{tr}_B(x_i) m) = \theta\left(\sum_{i=1}^n y_i \text{tr}_B(x_i) m\right) \stackrel{(2-9)}{=} \theta(m). \end{aligned}$$

Hence (3-3) is also right inverse to (3-2). □

We will let

$$\kappa : ({}_B A_A^{\psi_A})^\vee \xrightarrow{\cong} \text{HOM}_B({}^{\psi_B} A_A^{\psi_A}, {}_B B_B)$$

be the special case of the isomorphism (3-3) of  $(A, B)$ -bimodules where one takes  $M$  to be  ${}_B A_A^{\psi_A}$ .

**Proposition 3.6.** *The map*

$$\varphi_A : {}_A A_B \rightarrow ({}_B A_A^{\psi_A})^\vee, \quad \varphi_A(a) = \text{tr}_A \circ {}^r \psi_A(a),$$

*is an isomorphism of  $(A, B)$ -bimodules of degree  $(-\lambda_A, \pi_A)$ .*

*Proof.* The map  $\varphi_A$  is clearly a homomorphism of abelian groups. Let  $r \in R$ ,  $a \in A$ , and  $x \in {}_A A_B^{\psi_A}$ . Then

$$\begin{aligned} \varphi_A(a)(rx) &= \text{tr}_A \circ {}^r \psi_A(a)(rx) = (-1)^{\bar{a}(\bar{r} + \bar{x})} \text{tr}_A(rx \psi_A(a)) \\ &= (-1)^{\bar{a}(\bar{r} + \bar{x}) + \pi_A \bar{r}} r \text{tr}_A(x \psi_A(a)) = (-1)^{\bar{r}(\bar{a} + \pi_A)} r \text{tr}_A \circ {}^r \psi_A(a)(x) \\ &= (-1)^{\bar{r}(\bar{a} + \pi_A)} r \varphi_A(a)(x). \end{aligned}$$

Thus,  $\varphi_A(a) \in ({}_B A_A^{\psi_A})^\vee$ .

Now, for  $a, a' \in A$  and  $x \in {}_A A_B^{\psi_A}$ , we have

$$\begin{aligned} \varphi_A(a'a)(x) &= \text{tr}_A \circ {}^r\psi_A(a'a)(x) \\ &= (-1)^{\bar{x}(\bar{a}'+\bar{a})} \text{tr}_A(x\psi_A(a'a)) \\ &= (-1)^{\bar{x}(\bar{a}'+\bar{a})} \text{tr}_A(x\psi_A(a')\psi_A(a)) \\ &= (-1)^{\bar{a}'(\bar{x}+\bar{a})} \text{tr}_A \circ {}^r\psi_A(a)(x\psi_A(a')) \\ &= (-1)^{\bar{a}'(\bar{x}+\bar{a})} \varphi_A(a)(x\psi_A(a')) \\ &= (-1)^{\bar{a}'\bar{a}} \varphi_A(a) \circ {}^r\psi_A(a')(x) \\ &= (-1)^{\bar{a}'\pi_A} (a' \cdot \varphi_A(a))(x). \end{aligned}$$

Thus  $\varphi_A$  is a homomorphism of left  $A$ -modules of degree  $(-\lambda_A, \pi_A)$ .

On the other hand, for  $a \in A$ ,  $b \in B$ , and  $x \in {}_A A_B^{\psi_A}$ , we have

$$\begin{aligned} \varphi_A(ab)(x) &= \text{tr}_A \circ {}^r\psi_A(ab)(x) \\ &= (-1)^{(\bar{a}+\bar{b})\bar{x}} \text{tr}_A(x\psi_A(ab)) \\ &= (-1)^{(\bar{a}+\bar{b})\bar{x}} \text{tr}_A(x\psi_A(a)\psi_A(b)) \\ &= (-1)^{\bar{a}(\bar{x}+\bar{b})} \text{tr}_A(bx\psi_A(a)) \\ &= \text{tr}_A \circ {}^r\psi_A(a)(bx) = \text{tr}_A \circ {}^r\psi_A(a) \circ {}^\ell b(x) = (\varphi_A(a) \cdot b)(x). \end{aligned}$$

Thus  $\varphi_A$  is a homomorphism of right  $B$ -modules.

It remains to show that  $\varphi_A$  is an isomorphism. Suppose  $\varphi(a) = \varphi(a')$  for some  $a, a' \in A$ . This implies  $\bar{a} = \bar{a}'$ . Then, for all  $x \in {}_A A_A^{\psi_A}$ , we have

$$\begin{aligned} \varphi(a)(x) = \varphi(a')(x) &\implies \text{tr}_A \circ {}^r a(x) = \text{tr}_A \circ {}^r a'(x) \\ &\implies (-1)^{\bar{a}\bar{x}} \text{tr}_A(x\psi_A(a)) = (-1)^{\bar{a}'\bar{x}} \text{tr}_A(x\psi_A(a')) \\ &\implies 0 = (-1)^{\bar{a}\bar{x}} \text{tr}_A(x(\psi_A(a) - \psi_A(a'))). \end{aligned}$$

It thus follows from (T1) that  $\psi_A(a) = \psi_A(a')$ , and hence  $a = a'$ . Thus  $\varphi_A$  is injective.

Now, every element  $\varphi \in ({}_B A_A^{\psi_A})^\vee$  can be viewed as an element of  $\text{HOM}_R({}_R A, R)$ . Then, by (T2), there exists an  $a \in A$  such that  $\varphi = \text{tr}_A \circ {}^r a$ . Since  $\psi_A$  is a ring isomorphism, we have

$$\text{tr}_A \circ {}^r \psi_A(\psi_A^{-1}(a)) = \text{tr}_A \circ {}^r a = \varphi.$$

Thus,  $\varphi_A$  is surjective. □

**Proposition 3.7.** *The map*

$$\kappa \circ \varphi_A: {}_A A_B \rightarrow \text{HOM}_B({}_B A_A^{\psi_B \psi_A}, {}_B B_B)$$

is an isomorphism of  $(A, B)$ -bimodules of degree  $(\lambda_B - \lambda_A, \pi_A + \pi_B)$ . Moreover, the map

$$\text{tr}: {}_B^{\psi_B} A {}_B^{\psi_A} \rightarrow {}_B B_B, \quad \text{tr}(a) = (-1)^{\pi_B(\pi_A + \bar{a})} \sum_{i=1}^n (-1)^{\bar{y}_i(\pi_B + \bar{a})} \text{tr}_A(y_i a) x_i$$

is a trace map; i.e., it satisfies conditions (T1) and (T2).

*Proof.* Since  $\kappa \circ \varphi_A$  is a composition of  $(A, B)$ -bimodule isomorphisms, it too is an  $(A, B)$ -bimodule isomorphism. Now, for  $a \in {}_A A_B$ , we have

$$\begin{aligned} (\kappa \circ \varphi_A)(1_A)(a) &= (\kappa(\text{tr}_A \circ {}^r\psi_A(1_A)))(a) = (\kappa(\text{tr}_A))(a) \\ &= (-1)^{\pi_B(\pi_A + \bar{a})} \sum_{i=1}^n (-1)^{\bar{y}_i(\pi_B + \bar{a})} \text{tr}_A(y_i a) x_i. \end{aligned}$$

Then by [Pike and Savage 2016, Proposition 4.1],  $\text{tr}$  is left trace map. □

**Theorem 3.8.** *Let  $A$  be a ring extension of  $B$ , and  $B$  be a ring extension of  $R$ , with  $C_A(R) = A$ . Suppose that  $A$  is a Frobenius extension of  $R$  of degree  $(-\lambda_A, \pi_A)$ , with Nakayama automorphism  $\psi_A$ , and that  $B$  is a Frobenius extension of  $R$  of degree  $(-\lambda_B, \pi_B)$ , with Nakayama automorphism  $\psi_B$ . If  $A$  is projective as a left  $B$ -module, then  $A$  is a  $(\psi_A, \psi_B)$ -Frobenius extension of  $B$  of degree  $(\lambda_B - \lambda_A, \pi_B + \pi_A)$ . Moreover, the induction functor  ${}_A A_B \otimes_B -$  is right adjoint to the shifted twisted restriction functor  $\{\lambda_B - \lambda_A, \pi_B + \pi_A\}_B^{\psi_B} A {}_A^{\psi_A} \otimes_A -$ .*

*Proof.* Since  $A$  is a Frobenius extension of  $R$ , it is finitely generated as an  $R$ -module, and hence also finitely generated as a left  $B$ -module. Moreover, by Proposition 3.7, there is a trace map satisfying (T1) and (T2). Thus  $A$  is an  $(\psi_A, \psi_B)$ -Frobenius extension of  $B$ . The final assertion follows from [Pike and Savage 2016, Theorem 6.2]. □

**Remark 3.9.** Recall that, by Corollary 3.2, we gain no generality in Theorem 3.8 by allowing for  $A$  and  $B$  to be twisted Frobenius extensions of  $R$ . In the case that  $R$  is a field, concentrated in degree  $(0, 0)$ , Theorem 3.8 recovers [Pike and Savage 2016, Corollary 7.4].

### 4. Applications

In this final section, we give several examples that illustrate Theorem 3.8. In particular, we see that a number of results that have appeared in the literature follow immediately from this theorem.

**Example 4.1** (group rings). Let  $R$  be a supercommutative ring,  $G$  a finite group, and  $H$  a subgroup of  $G$ . Consider the following group rings over  $R$ :

$$R \cong R[\{e\}] \subseteq R[H] \subseteq R[G],$$

where  $e$  is the identity element of  $G$ . By construction,  $R[H]$  and  $R[G]$  are free as  $R$ -modules. It is easy to verify that the map

$$\text{tr}: R[G] \rightarrow R, \quad \text{tr}\left(\sum_{g \in G} r_g g\right) = r_e$$

satisfies (T1) and (T2) with  $\alpha$  and  $\beta$  both the identity map. Thus  $R[G]$  and  $R[H]$  are both untwisted Frobenius extensions of  $R$  and their Nakayama automorphisms are the identity automorphisms. The ring  $R$  clearly lies in the center of  $R[G]$  and  $R[G]$  is free as a left  $R[H]$ -module, with basis given by a set of left coset representatives. Therefore, by Theorem 3.8,  $R[G]$  is an untwisted Frobenius extension of  $R[H]$ . In the case that  $R$  is concentrated in degree  $(0, 0)$ , this recovers the well-known result that a finite group ring is a Frobenius extension of a subgroup ring.

**Example 4.2** (Hopf algebras). Let  $R$  be an unique factorization domain, let  $H$  be a Hopf algebra over  $R$  that is finitely generated and projective as an  $R$ -module, and let  $K$  be a Hopf subalgebra of  $H$ . Then  $H$  and  $K$  are both untwisted Frobenius extensions of  $R$  by [Pareigis 1971, Corollary 1]. Let  $\psi_H$  and  $\psi_K$  denote their respective Nakayama automorphisms. If  $H$  is projective as a left  $K$ -module (this condition is automatically satisfied when  $R$  is a field by [Nichols and Zoeller 1989, Theorem 7]), then  $H$  is a  $(\psi_H, \psi_K)$ -Frobenius extension of  $K$  by Theorem 3.8. Moreover, we have that  $H$  is an  $(\text{id}_H, \psi_K \circ \psi_H^{-1})$ -Frobenius extension of  $K$  by applying [Pike and Savage 2016, Proposition 3.4] with  $u = 1_H$ . That is, it is a Frobenius extension of the second kind. Thus we recover the result [Schneider 1992, Corollary 3.6(1)].

**Example 4.3** (nilcoxeter rings). Let  $R$  be a supercommutative ring and fix a non-negative integer  $n$ . The nilcoxeter ring  $N_n$  over  $R$  is generated by the elements  $u_1, \dots, u_{n-1}$  with the relations

$$\begin{aligned} u_i^2 &= 0 && \text{for } 1 \leq i \leq n-1, \\ u_i u_j &= u_j u_i && \text{for } 1 \leq i, j \leq n-1 \text{ such that } |i-j| > 1, \\ u_i u_{i+1} u_i &= u_{i+1} u_i u_{i+1} && \text{for } 1 \leq i < n-1. \end{aligned}$$

As an  $R$ -module,  $N_n$  has the basis  $\{u_w \mid w \in S_n\}$ , where  $S_n$  is the symmetric group on  $n$  elements. Multiplication of basis elements is given by

$$u_v u_w = \begin{cases} u_{vw} & \text{if } \ell(v+w) = \ell(v) + \ell(w), \\ 0 & \text{if } \ell(v+w) \neq \ell(v) + \ell(w), \end{cases}$$

where  $\ell$  is the length function of the symmetric group. So  $N_n$  is free and thus projective as an  $R$ -module. Now consider the  $R$ -linear function determined by

$$\text{tr}_n: N_n \rightarrow R, \quad \text{tr}_n(u_w) = \begin{cases} 1 & \text{if } w = w_0 \in S_n, \\ 0 & \text{if } w \neq w_0 \in S_n, \end{cases}$$

where  $w_0$  denotes the permutation of maximal length in  $S_n$ . It can be shown that  $N_n$  is an untwisted Frobenius extension of  $R$  of degree  $(-\binom{n}{2}, \binom{n}{2})$  with trace map  $\text{tr}_n$ , and the Nakayama automorphism associated to  $\text{tr}_n$  is given by  $\psi_n(u_i) = u_{n-i}$ ; see [Pike and Savage 2016, Lemma 8.2], where one replaces  $\mathbb{F}$  with  $R$ . Although the author of [Khovanov 2001, Proposition 4] works over the field  $\mathbb{Q}$ , his proof that  $N_n$  is projective as a left  $N_{n-1}$ -module still holds over  $R$ . It is clear that  $C_{N_n}(R) = N_n$ . Therefore, by Theorem 3.8,  $N_n$  is a  $(\psi_n, \psi_{n-1})$ -Frobenius extension of  $N_{n-1}$  of degree  $(\binom{n-1}{2} - \binom{n}{2}, \binom{n}{2} + \binom{n-1}{2})$ .

### Acknowledgements

This paper is the result of a research project completed in the context of an Undergraduate Student Research Award from the Natural Sciences and Engineering Research Council of Canada (NSERC), received by Poon. Savage was supported by an NSERC Discovery Grant.

### References

- [Kasch 1954] F. Kasch, “Grundlagen einer Theorie der Frobeniusweiterungen”, *Math. Ann.* **127** (1954), 453–474. MR Zbl
- [Khovanov 2001] M. Khovanov, “Nilcoxeter algebras categorify the Weyl algebra”, *Comm. Algebra* **29**:11 (2001), 5033–5052. MR Zbl
- [Morita 1965] K. Morita, “Adjoint pairs of functors and Frobenius extensions”, *Sci. Rep. Tokyo Kyoiku Daigaku Sect. A* **9** (1965), 40–71. MR Zbl
- [Nakayama and Tsuzuku 1960] T. Nakayama and T. Tsuzuku, “On Frobenius extensions, I”, *Nagoya Math. J.* **17** (1960), 89–110. MR Zbl
- [Nichols and Zoeller 1989] W. D. Nichols and M. B. Zoeller, “A Hopf algebra freeness theorem”, *Amer. J. Math.* **111**:2 (1989), 381–385. MR Zbl
- [Pareigis 1971] B. Pareigis, “When Hopf algebras are Frobenius algebras”, *J. Algebra* **18** (1971), 588–596. MR Zbl
- [Pike and Savage 2016] J. Pike and A. Savage, “Twisted Frobenius extensions of graded superrings”, *Algebr. Represent. Theory* **19**:1 (2016), 113–133. MR Zbl
- [Schneider 1992] H.-J. Schneider, “Normal basis and transitivity of crossed products for Hopf algebras”, *J. Algebra* **152**:2 (1992), 289–312. MR Zbl

Received: 2016-05-05    Revised: 2017-06-27    Accepted: 2017-06-28

*Department of Mathematics and Statistics, University of Ottawa*

epoon061@uottawa.ca

*Department of Mathematics and Statistics,  
University of Ottawa, Ottawa, Canada*

alistair.savage@uottawa.ca

*Department of Mathematics and Statistics,  
University of Ottawa, Ottawa, Canada*





# On $G$ -graphs of certain finite groups

Mohammad Reza Darafsheh and Safoora Madady Moghadam

(Communicated by Kenneth S. Berenhaut)

The notion of  $G$ -graph was introduced by Bretto et al. and has interesting properties. This graph is related to a group  $G$  and a set of generators  $S$  of  $G$  and is denoted by  $\Gamma(G, S)$ . In this paper, we consider several types of groups  $G$  and study the existence of Hamiltonian and Eulerian paths and circuits in  $\Gamma(G, S)$ .

## 1. Introduction

Let  $G$  be a finitely generated group with a generating set  $S = \{s_1, s_2, \dots, s_n\}$ . The left transversal of the left cosets of the subgroup  $\langle s_i \rangle$  in  $G$  is denoted by  $T_{\langle s_i \rangle}$ . This means that  $\{\langle s_i \rangle x \mid x \in T_{\langle s_i \rangle}\}$  is the set of all the distinct left cosets of  $\langle s_i \rangle$  in  $G$ . A simple graph  $\Gamma(G, S)$  is defined as follows: the vertex set of  $\Gamma(G, S)$  is the set  $\{\langle s_i \rangle x_j \mid x_j \in T_{\langle s_i \rangle}\}$ , and two distinct vertices  $\langle s_i \rangle x_j$  and  $\langle s_k \rangle x_l$  are joined by an edge if  $\langle s_i \rangle x_j \cap \langle s_k \rangle x_l \neq \emptyset$ .

The  $G$ -graphs were introduced in [Bretto and Faisant 2005] to study the group isomorphism problem. They also defined a similar graph  $\bar{\Gamma}(G, S)$ , which differs from  $\Gamma(G, S)$  by the fact that there are  $p$  edges between  $\langle s_i \rangle x_j$  and  $\langle s_k \rangle x_l$  if  $|\langle s_i \rangle x_j \cap \langle s_k \rangle x_l| = p$ . In this paper, we are more concerned with the simple graph  $\Gamma(G, S)$ . For more information on the subject see, for example, [Bretto et al. 2007; Bretto and Gillibert 2005]. By [Bretto et al. 2007], if  $S$  is a generating set of  $G$ , then  $\Gamma(G, S)$  is a connected graph. We always choose  $S$  such that  $G = \langle S \rangle$ .

The existence of Hamiltonian paths and circuits in  $\Gamma(G, S)$  was the main interest of [Bretto and Faisant 2011]. In [Bauer et al. 2008] the authors considered various classes of finite groups  $G$  and studied the Eulerianness and Hamiltonicity of the graph  $\Gamma(G, S)$ . For instance, they studied the Hamiltonicity of certain  $G$ -graphs on the groups  $Z_m \times Z_n$  and  $D_{2n}$ , the dihedral group of order  $2n$ . In this paper we will consider the groups  $Z_{n_1} \times Z_{n_2} \times \dots \times Z_{n_k}$  such that  $n_1 \mid n_2 \mid \dots \mid n_k$ , the dicyclic group  $T_{4n}$  of order  $4n$  with presentation

$$T_{4n} = \langle a, b \mid a^{2n} = e, a^n = b^2, b^{-1}ab = a^{-1} \rangle,$$

*MSC2010:* primary 05C25, 20F05; secondary 05C45.

*Keywords:*  $G$ -graphs, finite group, Hamiltonian circuit, graphs, paths, circuits.

$V_{8n}$ , a group of order  $8n$  with presentation

$$V_{8n} = \langle a, b \mid a^{2n} = b^4 = e, ba = a^{-1}b^{-1}, b^{-1}a = a^{-1}b \rangle,$$

and obtain the conditions under which  $\Gamma(G, S)$  is Eulerian or Hamiltonian.

## 2. Preliminaries

Let  $S = \{s_1, s_2, \dots, s_n\}$  be a generating set for the group  $G$ . Let

$$V_{s_i} = \{\langle s_i \rangle x_j \mid x_j \in T_{\langle s_i \rangle}\}, \quad 1 \leq i \leq n,$$

where  $T_{\langle s_i \rangle}$  is a complete set of left transversals of  $\langle s_i \rangle$  in  $G$ . Then by definition the vertex set of  $\Gamma(G, S)$  is  $V(\Gamma(G, S)) = \bigsqcup_{i=1}^n V_{s_i}$ . The graph  $\Gamma(G, S)$  is connected and  $n$ -partite. We recall some results which will be used in this paper.

**Result 1** [Bondy and Murty 1976]. Let  $\Gamma$  be a nontrivial connected graph. Then:

- (a)  $\Gamma$  has an Eulerian circuit if and only if every vertex of  $\Gamma$  has even degree.
- (b)  $\Gamma$  has an Eulerian path if and only if  $\Gamma$  has exactly two vertices of odd degree. Furthermore, the path begins at one of the vertices of odd degree and terminates at the other one.

**Result 2** [Bauer et al. 2008]. Let  $G$  be a group with a generating set given by  $S = \{s_1, s_2, \dots, s_n\}$ . Let  $S_{ij} = |\langle s_i \rangle \cap \langle s_j \rangle|$ . Then the degree of the vertex  $\langle s_i \rangle$  in the graph  $\Gamma(G, S)$  is equal to  $\deg(\langle s_i \rangle) = \sum_{j=1}^n (o(s_i)/S_{ij}) - 1$ , where  $o(s_i)$  denotes the order of the element  $s_i \in G$ . Note that for all elements  $x_j \langle s_i \rangle$  in  $V_i$  we have  $\deg(x_j \langle s_i \rangle) = \deg(\langle s_i \rangle)$ .

**Result 3** [Bauer et al. 2008]. Let  $G = Z_n \times Z_m$  and  $S = \{(1, 0), (0, 1)\}$ . Then  $\Gamma(G, S)$  has a Hamiltonian path if and only if  $|m - n| \leq 1$ .

In the following we generalize Result 3 to obtain a necessary condition for a Hamiltonian circuit of  $\Gamma(G, S)$ .

**Theorem 2.1.** *Let  $G = \langle a, b \rangle$ ,  $S = \{a, b\}$  and  $X = |G|/o(a)$  and  $Y = |G|/o(b)$ . If  $\Gamma(G, S)$  has a Hamiltonian path, then  $|X - Y| \leq 1$ .*

*Proof.* Let  $V_a = \{a_1, a_2 \cdots a_X\}$  and  $V_b = \{b_1, b_2 \cdots b_Y\}$ .

Case 1: Assume that the Hamiltonian path begins from a vertex in  $V_a$ . Call this vertex  $a_{i_1}$ . The next vertex can't be from  $V_a$ . Thus it is from  $V_b$ . Call this vertex  $b_{i_1}$ . In this way, the Hamiltonian path can be represented as  $a_{i_1}, b_{i_1}, a_{i_2}, b_{i_2}, \dots$ .

If this Hamiltonian path ends with a vertex from  $V_a$ , it is represented as

$$a_{i_1}, b_{i_1}, a_{i_2}, b_{i_2}, \dots, a_{i_{X-1}}, b_{i_{X-1}}, a_{i_X}.$$

Now notice that  $b_{i_1}, b_{i_2}, \dots, b_{i_{X-1}}$  should exhaust all the vertices of  $V_b$  exactly once. So  $\{b_{i_1}, b_{i_2}, \dots, b_{i_{X-1}}\} = \{b_1, b_2, \dots, b_Y\}$ ; hence  $X - 1 = Y$ , which implies

$X - Y = 1$ . But if this path ends with a vertex of  $V_b$ , it is represented as  $a_{i_1}, b_{i_1}, a_{i_2}, b_{i_2}, \dots, a_{i_X}, b_{i_X}$ . Similarly,  $\{b_{i_1}, b_{i_2}, \dots, b_{i_X}\} = \{b_1, b_2, \dots, b_Y\}$ , so  $X = Y$ .

Case 2: Assume that the Hamiltonian path begins with a vertex from  $V_b$ . In the same manner as above, this path can be represented as  $b_{i_1}, a_{i_1}, b_{i_2}, a_{i_2}, \dots$

If this path ends with a vertex from  $V_a$ , it is represented by  $b_{i_1}, a_{i_1}, b_{i_2}, a_{i_2}, \dots, b_{i_Y}, a_{i_Y}$ . Notice that  $a_{i_1}, a_{i_2}, \dots, a_{i_Y}$  should exhaust all the vertices of  $V_a$  exactly once, so  $\{a_{i_1}, a_{i_2}, \dots, a_{i_Y}\} = \{a_1, a_2, \dots, a_X\}$ ; hence  $Y = X$ . But if this path, ends with a vertex from  $V_b$ , it is represented by  $b_{i_1}, a_{i_1}, b_{i_2}, a_{i_2}, \dots, b_{i_{Y-1}}, a_{i_{Y-1}}, b_{i_Y}$ . Similarly,  $\{a_{i_1}, a_{i_2}, \dots, a_{i_{Y-1}}\} = \{a_1, a_2, \dots, a_X\}$ , so  $Y - 1 = X$ , implying  $Y - X = 1$ .

Thus in the general case the inequality  $|X - Y| \leq 1$  holds. □

**Result 4.** Let  $G = Z_n \times Z_m$  and  $S = \{(1, 0), (0, 1)\}$ . Then  $\Gamma(G, S)$  has a Hamiltonian circuit if and only if  $m = n$ .

A generalization of Result 4 for the existence of a Hamiltonian circuit is given in the following theorem.

**Theorem 2.2.** Let  $G = \langle a, b \rangle$ ,  $S = \{a, b\}$  and  $X = |G|/o(a)$  and  $|G|/o(b)$ . If  $\Gamma(G, S)$  has Hamiltonian circuit, then  $X = Y$ .

*Proof.* Let  $V_a = \{a_1, a_2, \dots, a_X\}$  and  $V_b = \{b_1, b_2, \dots, b_Y\}$ , and assume this circuit starts from a vertex in  $V_a$ , which is called  $a_{i_1}$ . The next vertex can't be from  $V_a$ , so it should be from  $V_b$ ; call this vertex  $b_{i_1}$ . Therefore this circuit can be represented by  $a_{i_1}, b_{i_1}, a_{i_2}, b_{i_2}, \dots, a_{i_X}, b_{i_X}, a_{i_1}$ . Now notice that  $b_{i_1}, b_{i_2}, \dots, b_{i_X}$  should exhaust all the vertices of  $V_b$  exactly once. So  $\{b_{i_1}, b_{i_2}, \dots, b_{i_X}\} = \{b_1, b_2, \dots, b_Y\}$ ; hence  $X = Y$ . □

### 3. Finite abelian groups

From [Rotman 1995] it's well known that every finite abelian group  $G$  is isomorphic to a direct product of cycle groups, say  $G \cong Z_{n_1} \times Z_{n_2} \times \dots \times Z_{n_k}$ , where  $n_1 | n_2 | \dots | n_k$ . We choose

$$S = \{(1, 0, 0, \dots, 0), (0, 1, 0, \dots, 0), \dots, (0, 0, 0, \dots, 1)\}$$

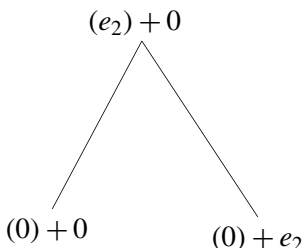
as a generating set of  $G$ . The vector  $(0, \dots, 1, \dots, 0)$  with 1 in the  $i$ -th position is denoted by  $e_i$ , and the zero vector is denoted by  $0 = (0, 0, \dots, 0)$ .

We are going to generalize the results of Section 3 in [Bauer et al. 2008] and obtain necessary and sufficient conditions in order that  $\Gamma(G, S)$  contains an Eulerian path or circuit.

**Theorem 3.1.** Let  $G$  be a finite abelian group which can be represented by  $G \cong Z_{n_1} \times Z_{n_2} \times \dots \times Z_{n_k}$ , where  $n_1 | n_2 | \dots | n_k$ . Let  $S = \{e_1, e_2, \dots, e_k\}$ . Then  $\Gamma(G, S)$  has an Eulerian circuit if and only if  $k$  is odd or  $n_1$  is even. Furthermore  $\Gamma(G, S)$  has an Eulerian path if and only if  $G \cong Z_1 \times Z_1$  or  $G \cong Z_1 \times Z_2$ .



**Figure 1.**  $\Gamma(Z_1 \times Z_1, S)$ .



**Figure 2.**  $\Gamma(Z_1 \times Z_2, S)$ .

*Proof.* Let us check the vertices  $\langle e_i \rangle + 0$  ( $1 \leq i \leq k$ ) of  $\Gamma(G, S)$ :

$$\begin{aligned} \langle e_1 \rangle + 0 &= (0, e_1, 2e_1, \dots, (n_1 - 1)e_1), \\ \langle e_2 \rangle + 0 &= (0, e_2, 2e_2, \dots, (n_2 - 1)e_2), \\ &\vdots \\ \langle e_k \rangle + 0 &= (0, e_k, 2e_k, \dots, (n_k - 1)e_k). \end{aligned}$$

For all  $i, j$  such that  $1 \leq i, j \leq k$ ,  $i \neq j$ , we have  $(\langle e_i \rangle + 0 \cap \langle e_j \rangle + 0) = 0$ , so  $|\langle e_i \rangle + 0 \cap \langle e_j \rangle + 0| = 1$ . Thus for all  $\langle e_i \rangle + x$  and  $\langle e_j \rangle + y$  such that  $\langle e_i \rangle + x \in V_{e_i}$  and  $\langle e_j \rangle + y \in V_{e_j}$ , if  $|\langle e_i \rangle + 0 \cap \langle e_j \rangle + 0| \neq 0$ , then  $|\langle e_i \rangle + 0 \cap \langle e_j \rangle + 0| = 1$ . So in the simple graph  $\Gamma(G, S)$ , we have  $\text{deg}(\langle e_i \rangle + x) = (k - 1)n_i$  for every  $\langle e_i \rangle + x$  from vertices of  $\Gamma(G, S)$  (Result 2). Now consider the following cases:

Case 1: If  $k$  is odd, then the degree of every vertex of  $\Gamma(G, S)$  is even. On the other hand,  $G = \langle (1, 0, 0, \dots, 0), (0, 1, 0, \dots, 0), \dots, (0, 0, 0, \dots, 1) \rangle$ . Thus  $\Gamma(G, S)$  is connected, so it has an Eulerian circuit but it doesn't have any Eulerian paths (Result 1).

Case 2: Assume that  $k$  is even:

Case 2.1: If  $n_1$  is even, then  $n_i$  is even for each  $1 \leq i \leq k$ , because  $n_1 | n_2 | \dots | n_k$ . So the degree of every vertex of  $\Gamma(G, S)$  is even; thus it has an Eulerian circuit but it doesn't have any Eulerian paths (Result 1).

Case 2.2: If  $n_1$  is odd and  $G \cong Z_1 \times Z_1$ , then  $\Gamma(G, S)$  is given in Figure 1. It has an Eulerian path, but it doesn't have any Eulerian circuits (Result 1).

Case 2.3: If  $n_1$  is odd and  $G \cong Z_1 \times Z_2$ , then  $\Gamma(G, S)$  is given in Figure 2. It has an Eulerian path, but it doesn't have any Eulerian circuits.

Case 2.4: If  $n_1$  is odd,  $n_1 \geq 3$  and  $G = Z_{n_1} \times Z_{n_2}$ , then  $n_1 \mid n_2$ , so  $n_2 \geq 3$ . On the other hand, the number of vertices of  $V_{e_1}$  is  $|G|/o(e_1) = n_2$ . So  $\Gamma(G, S)$  has at least three vertices of odd order. Thus it doesn't have any Eulerian paths or circuits (Result 1).

Case 2.5: If  $G = Z_{n_1} \times Z_{n_2} \times \cdots \times Z_{n_k}$  such that  $n_1$  is odd and  $k > 2$ , then  $\Gamma(G, S)$  doesn't have any Eulerian paths or circuits: the number of vertices of  $V_{e_1}$  is  $|G|/o(e_1) = \prod_{j=2}^k n_{i_j}$ .

If  $\prod_{j=2}^k n_{i_j} = 1$ , then  $G = Z_1 \times \cdots \times Z_1 \times Z_1$ , so  $\Gamma(G, S)$  has  $k$  vertices of odd degree (the degree is  $k - 1$ ). Thus  $\Gamma(G, S)$  has at least four vertices of odd degree, and hence it doesn't have any Eulerian paths or circuits (Result 1).

If  $\prod_{j=2}^k n_{i_j} = 2$ , then  $G = Z_1 \times \cdots \times Z_1 \times Z_2$ , so

$$\sum_{r=1}^{k-1} |V_{e_r}| = \sum_{r=1}^{k-1} \frac{|G|}{o(e_r)} = 2(k - 1) \geq 6.$$

Thus  $\Gamma(G, S)$  has at least six vertices of odd degree (the degree is  $k - 1$ ), so it doesn't have any Eulerian paths or circuits (Result 1).

If  $\prod_{j=2}^k n_{i_j} \geq 3$ , then  $\Gamma(G, S)$  has at least three vertices of odd degree (the degree is  $n_1(k - 1)$ ), so it doesn't have any Eulerian paths or circuits (Result 1). Therefore the theorem is proved. □

#### 4. Dicyclic group

Let  $G$  be the dicyclic group whose presentation is

$$T_{4n} = \langle a, b \mid a^{2n} = e, a^n = b^2, b^{-1}ab = a^{-1} \rangle, \tag{1}$$

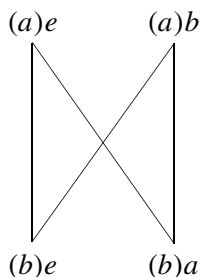
which is a group of order  $4n$ . We want to check the existence of Eulerian and Hamiltonian circuits and paths in the graph  $\Gamma(G, S)$  for a suitable subset  $S$  of  $G$ .

**Theorem 4.1.** *Let  $G$  be the group (1) and  $S = \{a, b\}$ . If  $n$  is even,  $\Gamma(G, S)$  has an Eulerian circuit and doesn't have any Eulerian paths. If  $n$  is odd,  $\Gamma(G, S)$  has an Eulerian path and doesn't have any Eulerian circuits.*

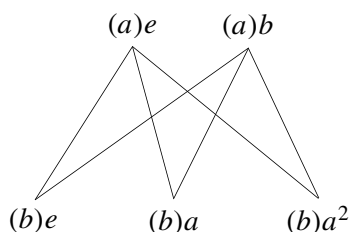
*Proof.* Clearly  $o(b) = 4$ . Now we check the vertices  $(a)e$  and  $(b)e$ , where  $e$  is the identity element of  $G$ :

$$\begin{aligned} (a)e &= (e, a, a^2, \dots, a^{2n-1}), \\ (b)e &= (e, b, b^2, b^3) = (e, b, a^n, a^n b). \end{aligned}$$

So  $(a)e \cap (b)e = \{e, a^n\}$ , and thus  $|(a)e \cap (b)e| = 2$ . Now we know that if  $(a)x \cap (b)y \neq \emptyset$ , then by [Bauer et al. 2008],  $|(a)x \cap (b)y| = 2$ . Notice that the number of vertices of  $V_a$  is  $|G|/o(a) = (4n)/(2n) = 2$ . On the other hand  $o(b) = 4$ , so  $\deg((b)y) = 4$  for every  $(b)y \in V_b$ . Thus every vertex of  $V_b$  has exactly



**Figure 3.**  $\Gamma(T_8, \{a, b\})$ .



**Figure 4.**  $\Gamma(T_{12}, \{a, b\})$ .

two edges to every vertex of  $V_a$ . Also we know that the number of vertices of  $V_b$  is  $|G|/o(b) = 4n/4 = n$ ; thus  $\bar{\Gamma}(G, S)$  is isomorphic to  $K_{n,2}^2$ , so  $\Gamma(G, S) \cong K_{n,2}$ .

Next if  $n$  is even, then  $\deg(v)$  is even for every vertex  $v$  of  $\Gamma(G, S)$ ; hence  $\Gamma(G, S)$  has an Eulerian circuit and it doesn't have any Eulerian paths (Result 1).

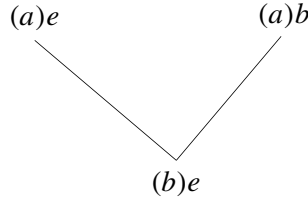
But if  $n$  is odd, then  $\deg(b)y$  is 2 for every  $(b)y$  in  $V_b$ , and  $\deg(a)x$  is  $n$ , which is odd for every  $(a)x$  in  $V_a$ . So  $\Gamma(G, S)$  has exactly two vertices of odd order; thus it has an Eulerian path and it doesn't have any Eulerian circuits (Result 1).  $\square$

**Theorem 4.2.** *Let  $G$  be the group (1) and  $S = \{a, b\}$ . If  $n = 2$ , then  $\Gamma(G, S)$  has a Hamiltonian path and circuit. If  $n = 1$  or 3, then  $\Gamma(G, S)$  has Hamiltonian path but it doesn't have any Hamiltonian circuits. If  $n \neq 1, 2, 3$ , then  $\Gamma(G, S)$  doesn't have any Hamiltonian paths or circuits.*

*Proof.* Assume that  $\Gamma(G, S) = K_{n,2}$  has a Hamiltonian path; then  $|n - 2| \leq 1$  (Theorem 2.1). Therefore just one of the following cases happens:

Case 1:  $n = 2$ . So  $\Gamma(G, S)$  is as in Figure 3. Thus its Hamiltonian path is  $(a)e, (b)a, (a)b, (b)e$ , and the Hamiltonian circuit is  $(a)e, (b)a, (a)b, (b)e, (a)e$ .

Case 2:  $(n - 2 = 1) \Rightarrow (n = 3)$ . So  $\Gamma(G, S)$  is as in Figure 4. Thus its Hamiltonian path is  $(b)e, (a)e, (b)a, (a)b, (b)a^2$ , but it doesn't have any Hamiltonian circuits because  $n \neq 2$  (Theorem 2.2).



**Figure 5.**  $\Gamma(T_4, \{a, b\})$ .

Case 3:  $(2-n = 1) \Rightarrow (n = 1)$ . So  $\Gamma(G, S)$  is as in Figure 5. Thus its Hamiltonian path is  $(a)e, (b)e, (a)b$ , but it doesn't have any Hamiltonian circuits because  $n \neq 2$  (Theorem 2.2).

So  $\Gamma(G, S)$  has a Hamiltonian circuit if and only if  $n = 2$ , and it has a Hamiltonian path if and only if  $n = 1$  or  $3$ . □

**Theorem 4.3.** *Let  $G$  be the group (1) and  $S = \{ab, b\}$ . Then  $\Gamma(G, S)$  has Eulerian and Hamiltonian circuits, and the Hamiltonian circuit is just the Eulerian circuit. Also  $\Gamma(G, S)$  has a Hamiltonian path, but it doesn't have any Eulerian paths.*

*Proof.* Clearly  $o(ab) = 4$ . Now let us check the vertices of  $V_b$ :

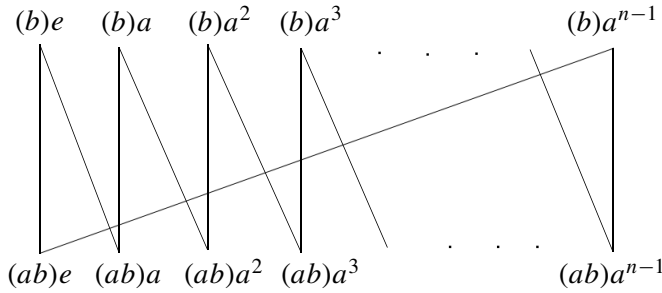
$$\begin{aligned} (b)e &= (e, b, b^2, b^3), \\ (b)a &= (a, ba, b^2, b^3a), \\ (b)a^2 &= (a^2, ba^2, b^2, b^3a^2), \\ &\vdots \\ (b)a^{n-1} &= (a^{n-1}, ba^{n-1}, b^2, b^3a^{n-1}). \end{aligned}$$

Now notice that  $ba^i = a^{2n-i}b$ ,  $(b)^2a^i = a^{n+i}$  and  $(b)^3a^i = a^{n-i}b$ . So

$$\begin{aligned} (b)e &= (e, b, a^n, (a)^nb), \\ (b)a &= (a, a^{2n-1}b, a^{n+1}, (a)^{n-1}b), \\ (b)a^2 &= (a^2, a^{2n-2}b, a^{n+2}, (a)^{n-2}b), \\ &\vdots \\ (b)a^{n-1} &= (a^{n-1}, a^{n+1}b, a^{2n-1}, ab). \end{aligned}$$

Next let us see the vertices of  $V_{ab}$ :

$$\begin{aligned} (ab)e &= (e, ab, (ab)^2, (ab)^3), \\ (ab)a &= (a, aba, (ab)^2a, (ab)^3a), \\ (ab)a^2 &= (a^2, aba^2, (ab)^2a^2, (ab)^3a^2), \\ &\vdots \\ (ab)a^{n-1} &= (a^{n-1}, aba^{n-1}, (ab)^2a^{n-1}, (ab)^3a^{n-1}). \end{aligned}$$



**Figure 6.**  $\Gamma(T_{4n}, \{ab, b\})$ .

Since  $aba^i = a(ba^i) = a^{2n-1+i}$ , we know  $(ab)^2a^i = a_n a^i = a^{n+i}$  and  $(ab)^3a^i = a^{n+1}ba^i = a^{n-i+1}$ . So

$$\begin{aligned} (ab)e &= (e, ab, (a)^n, (a)^{n+1}b), \\ (ab)a &= (a, b, (a)^{n+1}, (a)^n b), \\ (ab)a^2 &= (a^2, a^{2n-1}b, (a)^{n+2}, (a)^{n-1}b), \\ &\vdots \\ (ab)a^{n-1} &= (a^{n-1}, a^{n+2}b, (a)^{2n-1}, (a)^2b). \end{aligned}$$

Thus we have

$$\begin{aligned} (ab)a^i \cap (b)a^i &= \{a^i, a^{n+i}\}, \\ (ab)a^{i+1} \cap (b)a^i &= \{a^{2n-i}, a^{n-i}b\}, \\ (ab)e \cap (b)a^{n-1} &= \{ab, a^{n+1}b\}. \end{aligned}$$

Therefore  $\Gamma(G, S)$  is as shown in Figure 6.

Hence the Eulerian and Hamiltonian circuit is

$$(ab)e, (b)e, (ab)a, (b)a, (ab)a^2, (b)a^2, \dots, (ab)a^{n-1}, (b)a^{n-1}, (ab)e,$$

the Hamiltonian path is

$$(ab)e, (b)e, (ab)a, (b)a, (ab)a^2, (b)a^2, \dots, (ab)a^{n-1}, (b)a^{n-1}$$

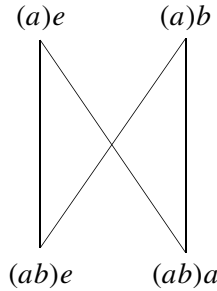
and  $\Gamma(G, S)$  doesn't have any Eulerian paths because the degree of every vertex of  $\Gamma(G, S)$  is even (Result 1). □

**Theorem 4.4.** *Let  $G$  be the group (1) and  $S = \{a, ab\}$ . If  $n$  is even,  $\Gamma(G, S)$  has an Eulerian circuit and it doesn't have any Eulerian paths, and if  $n$  is odd,  $\Gamma(G, S)$  has an Eulerian path and it doesn't have any Eulerian circuits.*

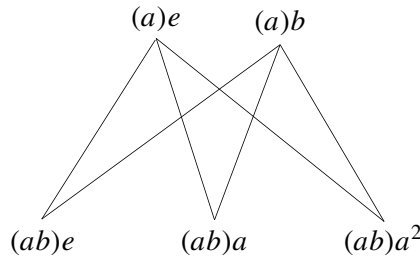
*Proof.* Let us check the vertices  $(a)e$  and  $(ab)e$ :

$$\begin{aligned} (a)e &= (e, a, a^2, \dots, a^{2n-1}), \\ (b)e &= (e, ab, a^n, a^{n+1}b). \end{aligned}$$





**Figure 7.**  $\Gamma(T_8, \{a, ab\})$ .



**Figure 8.**  $\Gamma(T_{12}, \{a, ab\})$ .

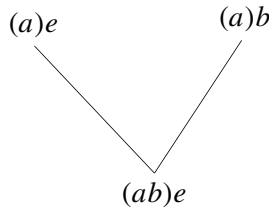
So  $(a)e \cap (ab)e = \{e, a^n\}$ ; thus  $|(a)e \cap (ab)e| = 2$ . We know that for  $(a)x \in V_a$  and  $(ab)y \in V_{ab}$ , if  $(a)x \cap (ab)y \neq \emptyset$ , then by [Bauer et al. 2008],  $|(a)x \cap (ab)y| = 2$ . On the other hand  $o(ab) = 4$  so  $\deg(ab)x = 4$  for every  $(ab)x \in V_{ab}$ , and also we know that the number of vertices of  $V_a$  is  $|G|/o(a) = (4n)/(2n) = 2$ . Thus in  $\Gamma(G, S)$ , every vertex of  $V_b$  has an edge to every vertex of  $V_a$ , so  $\Gamma(G, S)$  is  $K_{n,2}$ . Now if  $n$  is even, the degree of every vertex of  $\Gamma(G, S)$  is even, so it has an Eulerian circuit and doesn't have any Eulerian paths (Result 1).

But if  $n$  is odd,  $\Gamma(G, S)$  has exactly two vertices of odd degree ( $(a)e$  and  $(a)b$ ), so it has an Eulerian path and doesn't have any Eulerian circuits (Result 1).  $\square$

**Theorem 4.5.** *Let  $G$  be the group (1) and  $S = \{a, ab\}$ . If  $n = 2$ , then  $\Gamma(G, S)$  has a Hamiltonian path and circuit, if  $n = 1$  or  $n = 3$ , then  $\Gamma(G, S)$  has a Hamiltonian path and it doesn't have any Hamiltonian circuits, and if  $n \neq 1, 2, 3$ , then  $\Gamma(G, S)$  doesn't have any Hamiltonian paths or circuits.*

*Proof.* The  $G$ -graph  $\Gamma(G, S)$  is isomorphic to  $K_{n,2}$  (as we have already proved). Assume that it has a Hamiltonian path; then  $|n - 2| \leq 1$  (Theorem 2.1). So just one of the following cases happens:

Case 1:  $n = 2$ . So  $\Gamma(G, S)$  is as in Figure 7. Therefore its Hamiltonian path is  $(a)e, (ab)e, (a)b, (ab)a$ , and its Hamiltonian circuit is  $(a)e, (ab)e, (a)b, (ab)a, (a)e$ .



**Figure 9.**  $\Gamma(T_4, \{a, ab\})$ .

Case 2:  $(n-2 = 1) \Rightarrow (n = 3)$ . So  $\Gamma(G, S)$  is as in Figure 8. Therefore its Hamiltonian path is  $(ab)e, (a)e, (ab)a, (a)b, (ab)a^2$ . But it doesn't have any Hamiltonian circuits because  $n \neq 2$  (Theorem 2.2).

Case 3:  $(2-n = 1) \Rightarrow (n = 1)$ . So  $\Gamma(G, S)$  is as in Figure 9. Therefore its Hamiltonian path is  $(a)e, (ab)e, (a)b$ . But it doesn't have any Hamiltonian circuits because  $n \neq 2$  (Theorem 2.2). So  $\Gamma(G, S)$  has a Hamiltonian circuit if and only if  $n = 2$ , and it has a Hamiltonian path if and only if  $n = 1$  or  $3$ .  $\square$

**5. The group  $V_{8n}$  of order  $8n$**

The group  $G = V_{8n}$  has presentation

$$V_{8n} = \langle a, b \mid a^{2n} = b^4 = e, ba = a^{-1}b^{-1}, b^{-1}a = a^{-1}b \rangle. \tag{2}$$

We want to check the existence of Eulerian and Hamiltonian paths and circuits in  $\Gamma(G, S)$ .

**Theorem 5.1.** *Let  $G$  be the group (2) and  $S = \{a, b\}$ . Then  $\Gamma(G, S)$  always has an Eulerian circuit and never has Eulerian paths.*

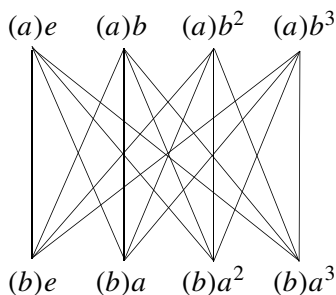
*Proof.* Let us check  $(a)e$  and  $(b)e$ :

$$\begin{aligned} (a)e &= (e, a, a^2, \dots, a^{2n-1}), \\ (b)e &= (e, b, b^2, b^3). \end{aligned}$$

So,  $(a)e \cap (b)e = \{e\}$ ; thus  $|(a)e \cap (b)e| = 1$ . Hence, for every  $(a)x \in V_a$  and  $(b)y \in V_b$ , if  $(a)x \cap (b)y \neq \emptyset$ , then  $|(a)x \cap (b)y| = 1$  [Bauer et al. 2008]. Now notice that  $o(a) = 2n$ , so the number of vertices of  $V_a$  is  $|G|/o(a) = (8n)/(2n) = 4$ . Also we know that  $o(b) = 4$ , so  $\deg(b)y = 4$  for every  $(b)y \in V_b$ . Thus every vertex of  $V_b$  has exactly one edge to every vertex of  $V_a$ . On the other hand, the number of vertices of  $V_b$  is  $|G|/o(b) = 8n/4 = 2n$ , so  $\Gamma(G, S) = K_{2n,4}$ .

Hence the degree of every vertex of  $\Gamma(G, S)$  is even ( $2n$  or  $4$ ), so it has an Eulerian circuit but it doesn't have any Eulerian paths (Result 1).  $\square$

**Theorem 5.2.** *Let  $G$  be the group (2) and  $S = \{a, b\}$ . Then  $\Gamma(G, S)$  has a Hamiltonian circuit if and only if  $n = 2$ .*



**Figure 10.**  $\Gamma(V_{16}, \{a, b\})$ .

*Proof.* The  $G$ -graph  $\Gamma(G, S)$  is isomorphic to  $K_{2n,4}$ . Assume that it has a Hamiltonian path, so  $|2n - 4| \leq 1$  (Theorem 2.1); hence one of the following cases happens:

Case 1:  $(2n = 4) \Rightarrow (n = 2)$ . So  $\Gamma(G, S)$  is as in Figure 10. The Hamiltonian path is  $(a)e, (b)e, (a)b, (b)a, (a)b^2, (b)a^2, (a)b^3, (b)a^3$ , and the Hamiltonian circuit is  $(a)e, (b)e, (a)b, (b)a, (a)b^2, (b)a^2, (a)b^3, (b)a^3, (a)e$ .

Case 2:  $(4 - 2n = 1) \Rightarrow (2n = 3)$ , which is not possible.

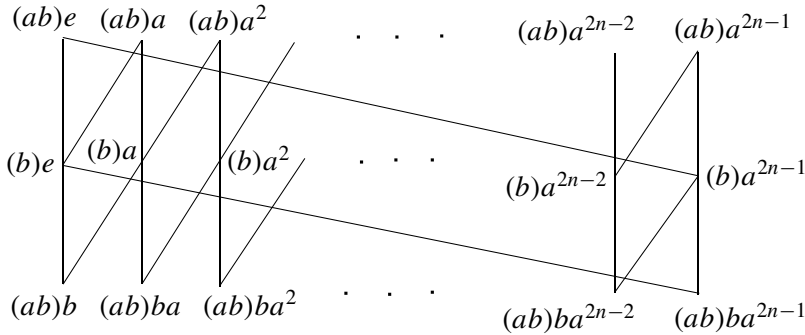
Case 3:  $(2n - 4 = 1) \Rightarrow (2n = 5)$ , which is not possible.

Notice that if  $n \neq 2$ , then  $\Gamma(G, S)$  doesn't have any Hamiltonian circuits (Theorem 2.2). So  $\Gamma(G, S)$  has a Hamiltonian path and circuit if and only if  $n = 2$ .  $\square$

**Theorem 5.3.** *Let  $G$  be the group  $(2)$  and  $S = \{b, ab\}$ . Then  $\Gamma(G, S)$  always has an Eulerian circuit and doesn't have any Eulerian paths.*

*Proof.* Clearly  $o(ab) = 2$ . Now notice that  $aba^i = b^3a^{i-1}$  and  $ab^2a^i = b^2a^{i+1}$ . Next let us check the vertices of  $V_{ab}$ :

$$\begin{aligned}
 (ab)e &= (e, ab) = (e, b^3a^{2n-1}), \\
 (ab)a &= (a, aba) = (a, b^3), \\
 (ab)a^2 &= (a^2, aba) = (a, b^3a), \\
 &\vdots \\
 (ab)a^{2n-1} &= (a^{2n-1}, aba) = (a, b^3a^{2n-2}), \\
 (ab)b &= (b, ab^2) = (b, b^2a), \\
 (ab)ba &= (ba, ab^2a) = (ba, b^2a^2), \\
 (ab)ba^2 &= (ba^2, ab^2a^2) = (ba^2, b^2a^3), \\
 &\vdots \\
 (ab)ba^{2n-1} &= (ba^{2n-1}, ab^2a^{2n-1}) = (ba^{2n-1}, b^2).
 \end{aligned}$$



**Figure 11.**  $\Gamma(V_{8n}, \{b, ab\})$ .

Let us also check those of  $V_b$ :

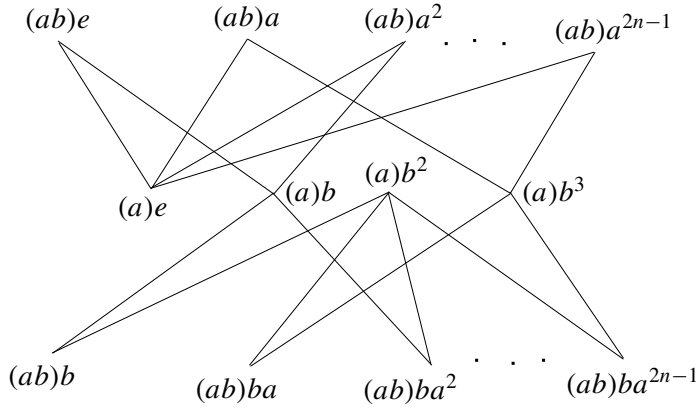
$$\begin{aligned}
 (b)e &= (e, b, b^2, b^3), \\
 (b)a &= (a, ba, b^2a, b^3a), \\
 (b)a^2 &= (a^2, ba^2, b^2a^2, b^3a^2), \\
 &\vdots \\
 (b)a^{2n-1} &= (a^{2n-1}, ba^{2n-1}, b^2a^{2n-1}, b^3a^{2n-1}).
 \end{aligned}$$

So we have  $(ab)a^i \cap (b)a^i = \{a^i\}$  and  $(ab)a^{i+1} \cap (b)a^i = \{b^3a^i\}$  and  $(ab)ba^i \cap (b)a^i = \{ba^i\}$  and  $(ab)ba^{i-1} \cap (b)a^i = \{b^2a^i\}$ . Hence in  $\Gamma(G, S)$ , the degree of every vertex of  $V_{ab}$  is 2, and the degree of every vertex of  $V_b$  is 4. So the degree of every vertex of  $\Gamma(G, S)$  is even. On the other hand  $G = V_{8n} = \langle ab, b \rangle$ , so  $\Gamma(G, S)$  is connected [Bretto et al. 2007]. Thus  $\Gamma(G, S)$  is a connected graph such that the degree of every vertex is even, so it has an Eulerian circuit and it doesn't have any Eulerian paths (Result 1). The Eulerian circuit in  $\Gamma(G, S)$  is

$$\begin{aligned}
 &(b)a^{2n-1}, (ab)e, (b)e, (ab)a, (b)a, (ab)a^2, (b)a^2, \\
 &\dots, (ab)a^{2n-2}, (b)a^{2n-2}, (ab)a^{2n-1}, (b)a^{2n-1}, (ab)ba^{2n-1}, \\
 &\quad (b)e, (ab)e, (b)a, (ab)ba, (b)a^2, (ab)ba^2, \\
 &\quad \dots, (b)a^{2n-2}, (ab)ba^{2n-2}, (b)a^{2n-1}. \quad \square
 \end{aligned}$$

**Theorem 5.4.** *Let  $G$  be the group (2) and  $S = \{b, ab\}$ . Then  $\Gamma(G, S)$  doesn't have any Hamiltonian paths or circuits.*

*Proof.* The number of vertices of  $V_b$  is  $|G|/o(b) = 8n/4 = 2n$ , and the number of vertices of  $V_{ab}$  is  $|G|/o(a) = 8n/2 = 4n$ . Now assume that  $\Gamma(G, S)$  has a Hamiltonian path, so  $|4n - 2n| \leq 1$  (Theorem 2.1). Hence one of the following cases will happen:



**Figure 12.**  $\Gamma(V_{8n}, \{a, ab\})$ .

Case 1:  $(4n = 2n) \Rightarrow (n = 0)$ .

Case 2:  $(4n - 2n = 1) \Rightarrow (2n = 1) \Rightarrow (n = \frac{1}{2})$ .

Case 3:  $(2n - 4n = 1) \Rightarrow (2n = -1) \Rightarrow (n = -\frac{1}{2})$ .

Obviously none of these cases can happen, so  $\Gamma(G, S)$  doesn't have any Hamiltonian paths, and thus it doesn't have any Hamiltonian circuits.  $\square$

**Theorem 5.5.** *Let  $G$  be the group (2) and  $S = \{a, ab\}$ . Then  $\Gamma(G, S)$  has an Eulerian circuit and doesn't have any Eulerian paths.*

*Proof.* Notice that  $o(a) = 2n$  and  $o(ab) = 2$ . Also notice that  $(ab)e = (e, ab)$  and  $(a)e = (e, a, a^2, \dots, a_{2n-1})$ , so  $(ab)e \cap (a)e = \{e\}$ . Thus, for every  $(a)x \in V_a$  and  $(ab)y \in V_{ab}$ , if  $(a)x \cap (ab)y \neq \emptyset$ , then  $|(a)x \cap (ab)y| = 1$  [Bauer et al. 2008]. So the degree of every vertex of  $V_a$  is  $2n$ , and the degree of every vertex of  $V_{ab}$  is 2.

On the other hand  $G = \langle a, ab \rangle$ , so  $\Gamma(G, S)$  is connected [Bretto et al. 2007]. Thus,  $\Gamma(G, S)$  is a connected graph such that the degree of every vertex is even. So it has an Eulerian circuit and doesn't have any Eulerian paths (Result 1).  $\square$

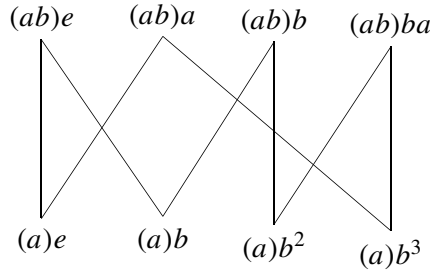
**Theorem 5.6.** *Let  $G$  be the group (2) and  $S = \{a, ab\}$ . Then  $\Gamma(G, S)$  has a Hamiltonian path and circuit if and only if  $n = 1$ .*

*Proof.* The number of vertices of  $V_a$  is  $|G|/o(a) = (8n)/(2n) = 4$ , and the number of vertices of  $V_{ab}$  is  $|G|/o(ab) = 8n/2 = 4n$ . Now assume that  $\Gamma(G, S)$  has a Hamiltonian path, so  $|4n - 4| \leq 1$  (Theorem 2.1). Hence one of the following cases happens:

Case 1:  $(4n - 4 = 1) \Rightarrow (4n = 5)$ , which is impossible.

Case 2:  $(4 - 4n = 1) \Rightarrow (4n = 3)$ , which is impossible.

Case 3:  $(4n - 4 = 0) \Rightarrow (4n = 4) \Rightarrow (n = 1)$ . In this case, the image of  $\Gamma(G, S)$  is shown in Figure 13. Its Hamiltonian path is  $(ab)e, (a)b^3, (ab)ba, (a)b^2, (ab)b,$



**Figure 13.**  $\Gamma(V_8, \{a, ab\})$ .

$(a)b$ ,  $(ab)a$ ,  $(a)e$ , and its Hamiltonian circuit is  $(ab)e$ ,  $(a)b^3$ ,  $(ab)ba$ ,  $(a)b^2$ ,  $(ab)b$ ,  $(a)b$ ,  $(ab)a$ ,  $(a)e$ ,  $(ab)e$ . If  $\Gamma(G, S)$  doesn't have any Hamiltonian paths, then it doesn't have any Hamiltonian circuits; thus  $\Gamma(G, S)$  has a Hamiltonian path and circuit if and only if  $n = 1$ .  $\square$

## 6. Conclusion

In this paper we investigated the existence of Eulerian circuits and paths in the  $G$ -graphs of finite abelian groups. Also we checked the existence of Hamiltonian and Eulerian circuits and paths in the  $G$ -graphs of some nonabelian finite groups. Our method can be applied to other finite groups as well.

## References

- [Bauer et al. 2008] C. M. Bauer, C. K. Johnson, A. M. Rodriguez, B. D. Temple, and J. R. Daniel, "Paths and circuits in  $G$ -graphs", *Involve* **1**:2 (2008), 135–144. MR Zbl
- [Bondy and Murty 1976] J. A. Bondy and U. S. R. Murty, *Graph theory with applications*, Elsevier, New York, 1976. MR Zbl
- [Bretto and Faisant 2005] A. Bretto and A. Faisant, "Another way for associating a graph to a group", *Math. Slovaca* **55**:1 (2005), 1–8. MR Zbl
- [Bretto and Faisant 2011] A. Bretto and A. Faisant, "Cayley graphs and  $G$ -graphs: some applications", *J. Symbolic Comput.* **46**:12 (2011), 1403–1412. MR Zbl
- [Bretto and Gillibert 2005] A. Bretto and L. Gillibert, "Symmetry and connectivity in  $G$ -graphs", *Electron. Notes Disc. Math.* **22** (2005), 481–486. Zbl
- [Bretto et al. 2007] A. Bretto, A. Faisant, and L. Gillibert, " $G$ -graphs: a new representation of groups", *J. Symbolic Comput.* **42**:5 (2007), 549–560. MR Zbl
- [Rotman 1995] J. J. Rotman, *An introduction to the theory of groups*, 4th ed., Graduate Texts in Mathematics **148**, Springer, 1995. MR Zbl

Received: 2016-10-13      Revised: 2017-03-07      Accepted: 2017-07-18

darafsheh@ut.ac.ir

*School of Mathematics, Statistics and Computer Science,  
College of Science, University of Tehran, Tehran, Iran*

madadi.safoora@ut.ac.ir

*School of Mathematics, Statistics and Computer Science,  
University of Tehran, Tehran, Iran*

# The tropical semiring in higher dimensions

John Norton and Sandra Spiroff

(Communicated by Scott T. Chapman)

We discuss the generalization, in higher dimensions, of the tropical semiring, whose two binary operations on the set of real numbers together with infinity are defined to be the minimum and the sum of a pair, respectively. In particular, our objects are closed convex sets, and for any pair, we take the convex hull of their union and their Minkowski sum, respectively, as the binary operations. We consider the semiring in several different cases, determined by a recession cone.

## Introduction

The tropical semiring is  $(\mathbb{R} \cup \{\infty\}, \oplus, \odot)$ , with the two operations defined by

$$x \oplus y = \min(x, y) \quad \text{and} \quad x \odot y = x + y.$$

The fact that this is a semiring comes from the lack of inverses under  $\oplus$ , as the additive neutral object is infinity. The multiplicative neutral object, i.e., under the operation  $\odot$ , is zero. Inspired by [Speyer and Sturmfels 2009, p. 165], we generalize the tropical semiring to higher dimensions. In particular, our elements are polyhedra, or more generally, closed convex sets, in  $\mathbb{R}^n$  with a fixed recession cone, i.e., the directions in which the set recedes, and the two operations are defined by taking the convex hull of the union and by the Minkowski sum. Indeed, when  $n = 1$  and the recession cone is  $\mathbb{R}_+ = \{\xi : \xi \geq 0\}$ , then this definition reduces to the tropical semiring [Maclagan and Sturmfels 2015; Speyer and Sturmfels 2009] as described above: the real numbers  $x$  and  $y$  represent the sets of solutions to the inequalities  $t \geq x$  and  $t \geq y$ , respectively; i.e., they correspond to the polyhedra in  $\mathbb{R}$  given by the positive rays with vertices at  $x, y$ . In particular, for each, the recession cone is the nonnegative ray emanating from the origin, or  $\mathbb{R}_+$ . Clearly, the union of these two sets is represented by the inequality  $t \geq \min(x, y)$  and likewise, the Minkowski sum is given by the inequality  $t \geq x + y$ . Careful consideration must be given to the neutral objects in this setting.

---

*MSC2010:* primary 16Y60, 52B11, 52A20; secondary 52A07.

*Keywords:* tropical semiring, polyhedra, compact subsets.

Spiroff is supported by a grant (#245926) from the Simons Foundation.

As suggested in [Speyer and Sturmfels 2009], the set of convex polyhedra in  $\mathbb{R}^n$  with fixed recession cone will form a semiring. We explore this idea in detail, considering various recession cones. In particular, we first consider the case of bounded polyhedra, i.e., convex polytopes, in  $\mathbb{R}^n$ . In this case, the common recession cone is  $\{0\}$  and the properties follow quite nicely. Furthermore, we can generalize this case to that of compact (convex) sets in  $\mathbb{R}^n$ . These proofs are the content of the second section.<sup>1</sup> Prior to that, we provide the necessary background on recession cones and asymptotic cones, and include examples to demonstrate the possible pathology of  $\oplus$  and  $\odot$  if the recession cone is not fixed. The main portion of the paper is dedicated to establishing the axioms of the various semirings, and most especially, those dealing with the closure of the two operations. The final section of the paper considers unbounded closed convex sets. We demonstrate the semirings of closed convex polyhedra and general convex sets, both with recession cone equal to the nonnegative orthant  $\mathbb{R}_+^n$ .

## 1. Background: polyhedra, recession cones, and asymptotic cones

Some general references for the material in this section are [Rockafellar 1970; Ziegler 1995; Border 1985; 2002].

**Definition 1.1** [Rockafellar 1970, p. 10]. A subset  $P$  of  $\mathbb{R}^n$  is *convex* if it satisfies the following property: for every  $x, y \in P$  and  $\lambda \in \mathbb{R}$ ,  $0 < \lambda < 1$ , the element  $\lambda x + (1 - \lambda)y$  is in  $P$ .

**Fact 1.2** [Rockafellar 1970, §2]. Given a subset  $S$  of  $\mathbb{R}^n$ , the *convex hull* of  $S$ , denoted by  $\text{conv } S$ , is the intersection of all the convex sets containing  $S$ . It is the smallest convex set containing  $S$ . In particular, it is the set of all convex combinations of the elements of  $S$ ; i.e.,

$$\text{conv } S = \{ \lambda_1 s_1 + \cdots + \lambda_k s_k : s_i \in S, \lambda_i \geq 0, \lambda_1 + \cdots + \lambda_k = 1, k \in \mathbb{N} \}.$$

**Definition 1.3** [Rockafellar 1970, p. 61]. Given a nonempty convex set  $P$  in  $\mathbb{R}^n$ , the *recession cone* is the set of all  $y \in \mathbb{R}^n$  such that  $p + y \in P$  for all  $p \in P$ . Denoted by  $0^+ P$ , the recession cone is the set of all directions in which  $P$  recedes, i.e., is unbounded.

**Fact 1.4** [Rockafellar 1970, Theorem 8.4]. A nonempty closed convex set  $P$  in  $\mathbb{R}^n$  is bounded if and only if its recession cone  $0^+ P$  consists of the zero vector alone.

**Example 1.5.** In the case of  $n = 2$ , the following sets have recession cone equal to the first quadrant of the plane  $\mathbb{R}_+^2 = \{ \mathbf{x} = (\xi_1, \xi_2) : \xi_1 \geq 0, \xi_2 \geq 0 \}$ .

$$(1) P = \{ (x, y) : x \geq -5, y \geq -18, y \geq -\frac{5}{3}x + 2 \};$$

<sup>1</sup>Section 2 and part of Section 3 are the basis for Norton's undergraduate thesis for the Honors College at the University of Mississippi.



- (2)  $Q = \{(x, y) : x \geq -3, y \geq -15, y \geq -6x - 16, y \geq -\frac{1}{2}x - 8\}$ ;
- (3) [Rockafellar 1970, Example p. 62]  $\{(x, y) : x > 0, y \geq 1/x\}$ .

**Definition 1.6** [Rockafellar 1970, p. 170; Ziegler 1995, p. 28; Aliprantis and Border 2006, p. 232]. A *polyhedral convex set* in  $\mathbb{R}^n$  is a set which can be expressed as the intersection of some finite collection of closed half spaces; i.e., it is the set of solutions to some finite system of inequalities  $Ax \leq b$ . A *convex polytope* is a bounded polyhedron; i.e., the convex hull of a finite set.

**Fact 1.7** [Ziegler 1995, Proposition 1.12]. If  $P$  is a polyhedral convex set in  $\mathbb{R}^n$ , then  $0^+P$  is the set of solutions to the system of inequalities  $Ax \leq 0$ .

**Definition 1.8** [Rockafellar 1970, p. 162]. A point  $x$  in a convex set  $P$  is an *extreme point* if the only way to express  $x$  as the convex combination  $(1 - \lambda)y + \lambda z$  for  $y, z \in P$  and  $0 < \lambda < 1$  is by taking  $y = z = x$ . Denote the set of extreme points of  $P$  by  $\text{ext}(P)$ .

**Fact 1.9** [Rockafellar 1970, Corollary 19.1.1]. If  $P$  is a polyhedral convex set, then  $\text{ext}(P)$  is finite.

In Example 1.5, the first two sets are polyhedra (see Figure 2), but the third one is not. The finite system of inequalities associated to  $P$  is

$$\begin{bmatrix} -1 & 0 \\ 0 & -1 \\ -\frac{5}{3} & -1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \leq \begin{bmatrix} 5 \\ 18 \\ -2 \end{bmatrix},$$

so that  $0^+P = \{(x, y) : x \geq 0, y \geq 0\}$ , and  $\text{ext}(P) = \{(-5, \frac{31}{3}), (12, -18)\}$ .

For a set that is not convex, there is a generalization of the notion of a recession cone. While we only consider convex sets, this new cone is relevant since the two definitions coincide when the convex set is closed; hence we may apply related results in the literature in our cases. We apply this material in the last section.

**Definition 1.10** [Border 1985, Definition 2.34]. The *asymptotic cone* of a set  $P$  in  $\mathbb{R}^n$ , denoted by  $AP$ , is the set of all possible limits of sequences of the form  $\{\alpha_i x_i\}_i$ , where each  $x_i \in P$ ,  $\alpha_i > 0$ , and  $\alpha_i \rightarrow 0$ .

Some properties of the asymptotic cone will be necessary to our proof:

**Fact 1.11** [Debreu 1959, §1.9; Border 2002, Lemma 4]. The following hold for sets  $E, F$  in  $\mathbb{R}^n$ :

- (1)  $AE$  is a cone.
- (2)  $AE \subseteq AF$  if  $E \subseteq F$ .
- (3)  $0^+E \subseteq AE$ .
- (4)  $AE \subseteq A(E + F)$ .

- (5)  $AE$  is closed.
- (6)  $AE$  is convex if  $E$  is convex.
- (7)  $0^+E = AE$  if  $E$  is closed and convex.
- (8)  $AE + AF \subseteq A(E + F)$  if  $E + F$  is convex.
- (9) A set  $E$  is bounded if and only if  $AE = \{0\}$ .

**Fact 1.12** [Shveidel 2001, proof of Theorem 2.3]. For a set  $P \subseteq \mathbb{R}^n$ , we have  $AP = A\bar{P}$ .

**Example 1.13** [Woo 2013]. In  $\mathbb{R}^2$ , let

$$P = \{(x, y) : 0 \leq x \leq 1, 0 \leq y \leq 1\} \cup \{(x, y) : 0 \leq x < 1, y \geq 1\}.$$

Although  $P$  is unbounded,  $0^+P = \{0\}$ ; however,  $P$  is not closed (see Fact 1.7). On the other hand,  $0^+\bar{P} = \{(0, y) : y \geq 0\} = A\bar{P} = AP$ .

As the above definitions and results are important to establishing the closure of the operation  $\oplus$ , the following definition and result are helpful in establishing the closure of the operation  $\odot$ .

**Fact 1.14** [Schneider 2014, Theorem 1.1.2]. Let  $P, Q$  be convex subsets of  $\mathbb{R}^n$ . Then  $\text{conv}(P) = P$ , and the Minkowski sum  $P + Q$  of  $P$  and  $Q$  is convex. In particular, if  $P, Q$  are nonempty, then  $P + Q = \{p + q : p \in P, q \in Q\}$ , and  $P + \emptyset = \emptyset$ .

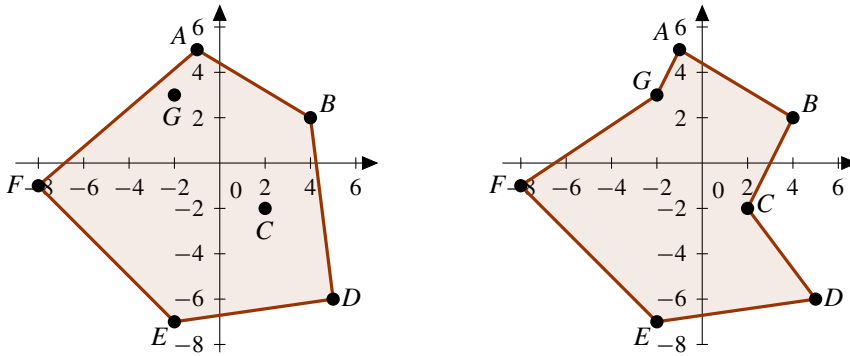
**Definition 1.15** [Debreu 1959, 1.9 m., p. 22]. The cones  $C_1, C_2, \dots, C_k$  in  $\mathbb{R}^n$  are *positively semi-independent* if, for any  $c_i \in C_i$ , the condition  $c_1 + c_2 + \dots + c_k = 0$  implies that each  $c_i = 0$ .

**Fact 1.16** [Border 2002, Theorem 8]. For closed and convex sets  $E, F \subseteq \mathbb{R}^n$  whose asymptotic cones  $AE$  and  $AF$  are positively semi-independent, the Minkowski sum  $E + F$  is closed and  $A(E + F) \subseteq AE + AF$ .

**Example 1.17** [Border 2002, Example 2]. In  $\mathbb{R}^2$ , set  $E = \{(x, y) : x > 0, y \geq 1/x\}$  and  $F = \{(x, y) : x < 0, y \geq -1/x\}$ . Note that both  $E$  and  $F$  are closed sets, but  $E + F = \{(x, y) : y > 0\}$ , which is not closed.

Finally, Carathéodory's theorem (see, e.g., [Schneider 2014, Theorem 1.1.4]) will be helpful when considering the elements of convex sets.

**Carathéodory's theorem.** *If a point  $x$  lies in the convex (hull of a) set  $P \subseteq \mathbb{R}^n$ , then  $x$  can be written as a convex combination of no more than  $n + 1$  points in  $P$ ; i.e., there are  $p_0, p_1, \dots, p_n \in P$  and  $\lambda_i \geq 0$  such that  $\lambda_0 + \lambda_1 + \dots + \lambda_n = 1$  and  $x = \lambda_0 p_0 + \dots + \lambda_n p_n$ .*



**Figure 1.** A convex polytope (left) and a nonconvex set (right) in  $\mathbb{R}^2$ .

**2. The tropical semiring in higher dimensions: the bounded case**

*The semiring of convex polytopes.* Recall that a convex polytope in  $\mathbb{R}^n$  is a bounded polyhedral set; i.e., the convex hull of a finite number of points in  $\mathbb{R}^n$ . See Figure 1. In particular, these sets are those convex polyhedra in  $\mathbb{R}^n$  with recession cone equal to the zero vector.

**Theorem 2.1.** *The set of all convex polytopes  $P, Q$  in  $\mathbb{R}^n$ , with operations shown below, is a semiring:*

$$P \oplus Q = \text{conv}(P \cup Q) \quad P \odot Q = P + Q = \{p + q : p \in P, q \in Q\}. \quad (2-1)$$

*Proof.* Let  $P, Q, R$  be convex polytopes in  $\mathbb{R}^n$ . Note that the empty set satisfies the convexity property vacuously, and as the solution set of any inconsistent system, it is a polytope. In particular, if  $P, Q$  are nonempty, set  $P = \text{conv}(p_1, \dots, p_s)$  and  $Q = \text{conv}(q_1, \dots, q_t)$ .

**Claim 2.1A.** *The set of all convex polytopes in  $\mathbb{R}^n$  under the operation of  $\oplus$  is a commutative monoid.*

- The operation  $\oplus$  is closed; i.e.,  $\text{conv}(P \cup Q)$  is a convex polytope:<sup>2</sup> First of all,  $P \oplus \emptyset = \text{conv}(P \cup \emptyset) = \text{conv}(P) = P$ , as  $P$  is convex, and likewise for  $\emptyset \oplus Q$ . Moreover,  $\emptyset \oplus \emptyset = \emptyset$ . Thus, we may assume that  $P, Q$  are both nonempty. We will show that  $\text{conv}(P \cup Q) = \text{conv}(p_1, \dots, p_s, q_1, \dots, q_t)$ . Let  $z \in \text{conv}(P \cup Q)$ . By Carathéodory’s theorem,  $z = \sum_{i=0}^n \lambda_i y_i$ , where each  $\lambda_i \geq 0$ ,  $\sum_{i=0}^n \lambda_i = 1$  and  $y_i \in P \cup Q$ . For each  $y_i \in P$ , one can write  $y_i = \sum_{j=1}^s \delta_{ij} p_j$ , where  $\delta_{ij} \geq 0$  for

<sup>2</sup>This fact appears in several books without proof. Therefore, we provide an argument, for the benefit of the undergraduate reader. (Likewise, for some other proofs in this section.) For algorithms that compute the convex hull of a finite set of points in the plane, for example, Graham’s scan and Jarvis’s march, see, e.g., [Cormen et al. 2001, Chapter 33, Section 3].

all  $j$  and  $\sum_{j=1}^s \delta_{ij} = 1$ . If all  $y_i \in P$ , then

$$\begin{aligned} z &= \sum_{i=0}^n \left( \lambda_i \sum_{j=1}^s \delta_{ij} p_j \right) \\ &= \sum_{j=1}^s \left( \sum_{i=0}^n \lambda_i \delta_{ij} \right) p_j \in \text{conv}(p_1, \dots, p_s) \subseteq \text{conv}(p_1, \dots, p_s, q_1, \dots, q_t); \end{aligned}$$

it is similar if all  $y_i \in Q$ . Thus, let  $m \in \mathbb{N}$ ,  $m < n$ , such that  $y_0, \dots, y_{m-1} \in P \setminus Q$  and  $y_m, \dots, y_n \in Q$ . Then

$$z = \sum_{i=0}^n \lambda_i y_i = \sum_{i=0}^{m-1} \left( \sum_{j=1}^s \lambda_i \delta_{ij} p_j \right) + \sum_{i=m}^n \left( \sum_{k=1}^t \lambda_i \delta_{ik} q_k \right)$$

is a convex combination of  $\{p_1, \dots, q_t\}$ ; hence,  $\text{conv}(P \cup Q) \subseteq \text{conv}(p_1, \dots, q_t)$ . Since the containment  $\supseteq$  is clear,  $\text{conv}(P \cup Q) = \text{conv}(p_1, \dots, p_s, q_1, \dots, q_t)$ , and the latter, by Definition 1.6, is a polytope.

- The operation  $\oplus$  is associative; i.e.,  $(P \oplus Q) \oplus R = P \oplus (Q \oplus R)$ : Regarding  $(P \oplus Q) \oplus R = P \oplus (Q \oplus R)$ , we wish to prove

$$\text{conv}[\text{conv}(P \cup Q) \cup R] = \text{conv}[P \cup \text{conv}(Q \cup R)]. \tag{2-2}$$

If any one or more of the sets is the empty set, then it is easy to see that the equality holds. Otherwise, it suffices to show that each of these sets is equal to  $\text{conv}(P \cup Q \cup R)$ . Consider the set on the left. Since  $P \cup Q \cup R \subseteq \text{conv}(P \cup Q) \cup R$ , we have  $\text{conv}(P \cup Q \cup R) \subseteq \text{conv}[\text{conv}(P \cup Q) \cup R]$ .

Conversely, as  $\text{conv}(P \cup Q), R \subseteq \text{conv}(P \cup Q \cup R)$ , we have  $\text{conv}(P \cup Q) \cup R \subseteq \text{conv}(P \cup Q \cup R)$ . Take the convex hull of both sides:  $\text{conv}[\text{conv}(P \cup Q) \cup R] \subseteq \text{conv}(P \cup Q \cup R)$ . This establishes that  $\text{conv}(P \cup Q \cup R) = \text{conv}[\text{conv}(P \cup Q) \cup R]$ . The argument for  $\text{conv}[P \cup \text{conv}(Q \cup R)]$  is analogous; hence we have (2-2).

- The operation  $\oplus$  is commutative: order does not matter in unions of sets.
- There exists a neutral object  $\mathcal{O}$  for addition such that for any convex polytope  $P$  in  $\mathbb{R}^n$ ,  $P \oplus \mathcal{O} = \mathcal{O} \oplus P = P$ : take  $\mathcal{O}$  to be the empty set  $\emptyset$ , since  $\text{conv}(P \cup \emptyset) = P$ .

**Claim 2.1B.** *The set of all convex polytopes in  $\mathbb{R}^n$  under the operation of  $\odot$  is a commutative monoid.*

- The operation  $\odot$  is closed; i.e.,  $P + Q$  is a convex polytope: First of all,  $P \odot \emptyset = \emptyset$  since  $P + \emptyset = \emptyset$  in Minkowski addition, and likewise for  $\emptyset \odot Q$ . Moreover,  $\emptyset \odot \emptyset = \emptyset$ . Thus, we may assume that  $P, Q$  are both nonempty. We will show that  $P + Q = \text{conv}(\{p_j + q_k : 1 \leq j \leq s, 1 \leq k \leq t\})$ , as per the hint in [Aliprantis and Border 2006, proof of Lemma 5.124]. Let  $p \in P$  and  $q \in Q$ . Write  $p = \sum_{j=1}^s \lambda_j p_j$

and  $q = \sum_{k=1}^t \mu_k q_k$ , where  $\lambda_j, \mu_k \geq 0$  and  $\sum_{j=1}^s \lambda_j = 1 = \sum_{k=1}^t \mu_k$ . Then,

$$\begin{aligned} p + q &= \sum_{j=1}^s \lambda_j p_j + \sum_{k=1}^t \mu_k q_k \\ &= \left( \sum_{k=1}^t \mu_k \right) \sum_{j=1}^s \lambda_j p_j + \left( \sum_{j=1}^s \lambda_j \right) \sum_{k=1}^t \mu_k q_k = \sum_{j=1}^s \sum_{k=1}^t \lambda_j \mu_k (p_j + q_k) \end{aligned}$$

is a convex combination of  $\{p_j + q_k : 1 \leq j \leq s, 1 \leq k \leq t\}$ .

Conversely, let  $\sum_{i=0}^n \lambda_i (x_i + y_i)$  be a convex combination of  $\{p_j + q_k : 1 \leq j \leq s, 1 \leq k \leq t\}$ ; i.e.,  $x_i = p_j$  for some  $j$  and  $y_i = q_k$  for some  $k$ ,  $\lambda_i \geq 0$ , and  $\sum_{i=0}^n \lambda_i = 1$ . Then

$$\sum_{i=0}^n \lambda_i (x_i + y_i) = \sum_{i=0}^n \lambda_i x_i + \sum_{i=0}^n \lambda_i y_i,$$

where the first sum is in  $\text{conv}(p_1, \dots, p_s)$  and the second sum is in  $\text{conv}(q_1, \dots, q_t)$ . Thus,  $P + Q = \text{conv}(\{p_j + q_k \mid 1 \leq j \leq s, 1 \leq k \leq t\})$ , and the latter, by Definition 1.6, is a convex polytope.

- The operation  $\odot$  is associative: addition in  $\mathbb{R}^n$  is associative.
- The operation  $\odot$  is commutative: addition in  $\mathbb{R}^n$  is commutative.
- There exists a neutral object  $\mathcal{I}$  for multiplication such that for any convex polytope  $P$  in  $\mathbb{R}^n$ ,  $P \odot \mathcal{I} = \mathcal{I} \odot P = P$ : Take  $\mathcal{I}$  to be  $\text{conv}(\{\mathbf{0}\}) = \{\mathbf{0}\}$ , which is a convex polytope by Definition 1.6, and the common recession cone of all nonempty convex polytopes  $P$  in  $\mathbb{R}^n$ . Then  $P + \mathbf{0} = P$ , by definition of  $0^+P$ , and  $\emptyset + \mathbf{0} = \emptyset$ .

**Claim 2.1C.** *The operation  $\odot$  is distributive over  $\oplus$ ; i.e.,*

$$P \odot (Q \oplus R) = (P \odot Q) \oplus (P \odot R).$$

We wish to establish that  $P + \text{conv}(Q \cup R) = \text{conv}[(P + Q) \cup (P + R)]$ . If  $P = \emptyset$  or more than two of the sets are empty, then both expressions equal  $\emptyset$ , and if only  $Q = \emptyset$  or only  $R = \emptyset$ , then both expressions equal  $P + R$  or  $P + Q$  respectively. Thus, assume all three are nonempty.

First of all, take  $p + z$ , where  $p \in P$  and  $z \in \text{conv}(Q \cup R)$ . Then  $z = \sum_{i=0}^n \lambda_i y_i$ , where  $\lambda_i \geq 0$ ,  $\sum_{i=0}^n \lambda_i = 1$ , and  $y_i \in Q \cup R$ . Therefore, we have

$$p + z = 1p + \sum_{i=0}^n \lambda_i y_i = \left( \sum_{i=0}^n \lambda_i \right) p + \sum_{i=0}^n \lambda_i y_i = \sum_{i=0}^n \lambda_i (p + y_i).$$

The elements  $p + y_j$  are in  $P + Q$  or  $P + R$ , and possibly both. Therefore, the last expression is in  $\text{conv}[(P + Q) \cup (P + R)]$ ; i.e.,  $p + z \in \text{conv}[(P + Q) \cup (P + R)]$ . Since  $p$  and  $z$  are arbitrary, we have  $P + \text{conv}(Q \cup R) \subseteq \text{conv}[(P + Q) \cup (P + R)]$ .

Conversely, since  $P + Q, P + R \subseteq P + \text{conv}(Q \cup R)$ , it follows that  $P + \text{conv}(Q \cup R)$  contains  $(P + Q) \cup (P + R)$ . Take the convex hull of both sides:

$$\text{conv}[(P + Q) \cup (P + R)] \subseteq \text{conv}[P + \text{conv}(Q \cup R)] = \text{conv}(P) + \text{conv}[\text{conv}(Q \cup R)],$$

where the equality follows by Fact 1.14. Now since both terms in the last sum are convex, the expression simplifies to  $P + \text{conv}(Q \cup R)$ . This establishes the other inclusion, and therefore,  $P + \text{conv}(Q \cup R) = \text{conv}[(P + Q) \cup (P + R)]$ .

**Claim 2.1D.** *The additive neutral object  $\mathcal{O}$  is an absorbing element for  $\odot$ ; i.e., for any convex polytope  $P$  in  $\mathbb{R}^n$ ,  $\mathcal{O} \odot P = P \odot \mathcal{O} = \mathcal{O}$ .*

This follows from the fact that, in Minkowski addition,  $\emptyset + P = \emptyset$ . □

**The semiring of convex compact sets.** In this section, we generalize the above work with convex polytopes to general convex compact subsets of  $\mathbb{R}^n$ . Of import is the Heine–Borel theorem (see, e.g., [Aliprantis and Border 2006, Theorem 3.19]):

**Heine–Borel theorem.** *Subsets of  $\mathbb{R}^n$  are compact if and only if they are closed and bounded.*

**Proposition 2.2.** *The set of all compact convex sets  $P, Q$  in  $\mathbb{R}^n$ , with the operations as in (2-1), is a semiring.*

*Proof.* We note that the arguments for many of the claims above do not change. In particular, the empty set is compact; hence it remains the neutral element under  $\oplus$ . However, closure of the two operations must be considered. Therefore, let  $P, Q$  be compact convex sets in  $\mathbb{R}^n$ .

- The operation  $\oplus$  is closed; i.e.,  $\text{conv}(P \cup Q)$  is a compact convex set: The union of finitely many compact sets is compact. Thus,  $P \cup Q$  is compact. Next, the convex hull of a compact set in  $\mathbb{R}^n$  remains compact (see, e.g., [Aliprantis and Border 2006, Corollary 5.18]); thus,  $\text{conv}(P \cup Q)$  is a compact convex set.
- The operation  $\odot$  is closed; i.e.,  $P + Q$  is a compact convex set: As per [Border 2002, Corollary 11], the summation of a closed set and a compact set is closed. As such,  $P \odot Q = P + Q$  is closed, and convex. Moreover,  $P + Q$  is bounded since  $P, Q$  are bounded. Apply the Heine–Borel theorem. □

### 3. The tropical semiring in higher dimensions: the unbounded case

**The semiring of convex polyhedra.** We consider the set of convex polyhedra in  $\mathbb{R}^n$  with the operations  $\oplus$  and  $\odot$  as in (2-1). Although convex polyhedra are necessarily closed (see, e.g., [Rockafellar 1970, Theorem 19.1]), the convex hull of the union of two convex polyhedral sets need not be polyhedral or closed, as evinced by Example 3.1 below, that is, if their recession cones do not coincide. Therefore,

we restrict our sets to those with the same recession cone, namely the nonnegative orthant  $\mathbb{R}_+^n = \{\mathbf{x} = (\xi_1, \dots, \xi_n) : \xi_1 \geq 0, \dots, \xi_n \geq 0\}$ . This restriction is a generalization of the nonnegative ray in the tropical semiring when  $n = 1$ .

**Example 3.1** [Rockafellar 1970, p. 177]. In  $\mathbb{R}^2$ , let  $P = \{(-1, 0)\}$  and  $Q = \{(x, y) : x, y \geq 0\}$ . Then  $\text{conv}(P \cup Q) = \{(-1, 0)\} \cup \{(x, y) : -1 < x, 0 \leq y\}$ , which is neither polyhedral nor closed. However,  $0^+P$  is the origin, while  $0^+Q = Q = \mathbb{R}_+^2$ .

**Proposition 3.2.** *Let  $\mathcal{P}$  be the set of all convex polyhedra in  $\mathbb{R}^n$  with recession cone equal to the nonnegative orthant  $\mathbb{R}_+^n$ . Then  $(\mathcal{P} \cup \{\emptyset\}, \oplus, \odot)$ , with operations defined in (2-1), is a semiring.*

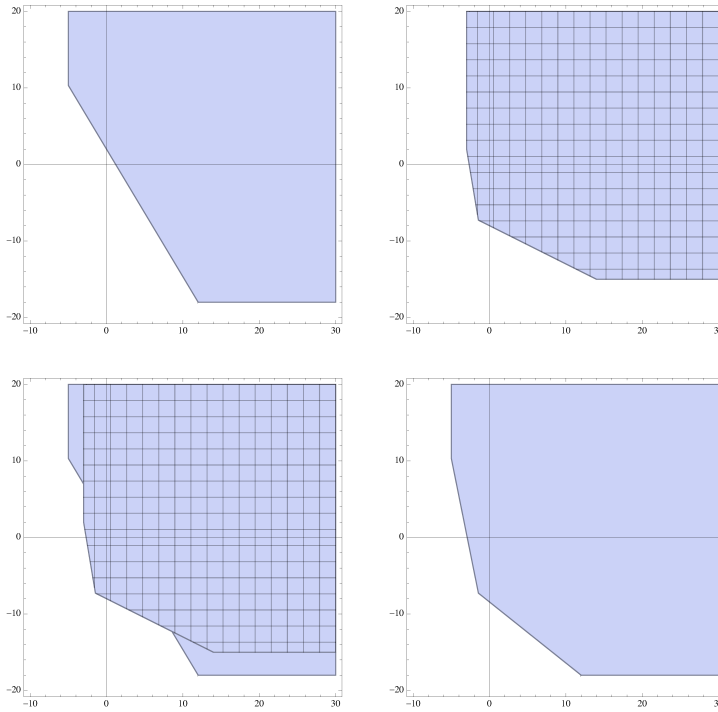
*Proof.* It suffices to address the issues regarding closure of the two operations for convex polyhedra  $P, Q$  in  $\mathbb{R}^n$  with recession cone equal to  $\mathbb{R}_+^n$ , and the multiplicative neutral object, since the earlier arguments for the remaining properties apply here.

- The operation  $\oplus$  is closed; i.e.,  $\text{conv}(P \cup Q)$  is a convex polyhedron in  $\mathbb{R}^n$  with recession cone equal to  $\mathbb{R}_+^n$ : Since  $\text{conv}(P \cup Q)$  is convex, it remains to establish that  $\text{conv}(P \cup Q)$  is polyhedral with a recession cone equal to the nonnegative orthant. The fact that the recession cone of  $\text{conv}(P \cup Q)$  is equal to  $\mathbb{R}_+^n$  follows from [Rockafellar 1970, Theorem 9.8.1]; therefore, it only remains to show that  $\text{conv}(P \cup Q)$  is polyhedral. By Definition 1.6,  $P$  is the irredundant intersection of some finite collection of closed half spaces, including those of the form  $\{\mathbf{x} : \langle \mathbf{x}, (0, \dots, 0, 1, 0, \dots, 0) \rangle \geq a_i\}$  for some  $a_i \in \mathbb{R}$ , i.e.,  $x_i \geq a_i$ , since  $0^+P = \mathbb{R}_+^n$ . Likewise,  $0^+Q = \mathbb{R}_+^n$ ; hence, for each  $i$ , the half-spaces defining  $Q$  include  $x_i \geq c_i$  for some  $c_i \in \mathbb{R}$ . Thus, every element of  $P \cup Q$  satisfies the set of inequalities

$$\{\mathbf{x} : \langle \mathbf{x}, (0, \dots, 0, 1, 0, \dots, 0) \rangle \geq \min(a_i, c_i)\}.$$

Moreover, if  $z \in \text{conv}(P \cup Q) \setminus (P \cup Q)$ , then  $z$  is in the finite region bounded by the (necessarily finite set of) extreme points of  $P$  and  $Q$ . See Figure 2 for an example. Thus,  $\text{conv}(P \cup Q) = \text{conv}(\text{ext}(P) \cup \text{ext}(Q)) + \mathbb{R}_+^n$ , and the latter, by [Ziegler 1995, Theorem 1.2], is polyhedral.

- The operation  $\odot$  is closed; i.e.,  $P + Q$  is a convex polyhedron with recession cone equal to  $\mathbb{R}_+^n$ : By [Rockafellar 1970, Corollary 19.3.2], the Minkowski sum of two polyhedral convex sets in  $\mathbb{R}^n$  is polyhedral, and it is convex. Therefore, it remains to show that  $0^+(P + Q) = \mathbb{R}_+^n$ . Since polyhedral convex sets are closed, their recession cones are equal to their asymptotic cones. Hence by Fact 1.11(8),  $AP + AQ \subseteq A(P + Q)$ . Next, as  $AP = AQ = \mathbb{R}_+^n$ , it follows that if  $y \in AP \setminus \{\mathbf{0}\}$ , then  $-y \notin AQ$ . In other words,  $AP$  and  $AQ$  are positively semi-independent, as per Definition 1.15. Thus, by Fact 1.16,  $A(P + Q) \subseteq AP + AQ$  and the result follows.



**Figure 2.** The graphs of polyhedra  $P$  (top left) and  $Q$  (top right) from Example 1.5, and  $P \cup Q$  (bottom left) and  $\text{conv}(P \cup Q)$  (bottom right).

- There exists a neutral object  $\mathcal{I}$  for multiplication: Take  $\mathcal{I}$  to be  $\mathbb{R}_+^n$ , which is not only an element of  $\mathcal{P}$ , but also the common recession cone of all nonempty polyhedra  $P$  in  $\mathcal{P}$ . Thus  $P + \mathbb{R}_+^n = P$ , by the definition of  $0^+P$ , and  $\emptyset + \mathbb{R}_+^n = \emptyset$ .  $\square$

**Remark 3.3.** While the set of real numbers  $\mathbb{R}^1$  is in one-to-one correspondence with the set of all nonempty closed convex polyhedra in the real number line, the same is not true for  $\mathbb{R}^n$  when  $n \geq 2$ . As mentioned in the Introduction,  $r \leftrightarrow [r, \infty)$ , in the case that  $n = 1$ , but an ordered pair  $(r_1, r_2)$  does not correspond to a unique closed convex polyhedron in  $\mathbb{R}^2$ .

**The semiring of closed convex sets with a fixed recession cone.** Finally, we generalize the above work to closed convex subsets of  $\mathbb{R}^n$  with a fixed recession cone  $C$ . As evinced in Example 1.13, pathology arises if the convex sets are not assumed to be closed. However, despite taking two convex sets that are closed, neither the convex hull of the union nor the Minkowski sum need be closed, as demonstrated by Examples 3.1 and 1.17, respectively, that is, if their recession cones do not coincide. Moreover, our earlier work hints at the possible necessity of taking  $C$  such that  $AC \cap (-AC) = \{\mathbf{0}\}$ .



**Theorem 3.4.** *Let  $S$  be the set of all closed convex sets in  $\mathbb{R}^n$  with fixed recession cone  $C$  satisfying either of the conditions below:*

(1)  $AC \cap (-AC) = \{\mathbf{0}\}$ .

(2)  $C$  is a closed half-space containing the origin.

Then  $\langle S \cup \{\emptyset\}, \oplus, \odot \rangle$ , with operations defined in (2-1), is a semiring.

*Proof.* Again, the earlier arguments for the most of the properties apply here; therefore, we address the issues regarding closure of the two operations for closed convex sets  $P, Q$  of  $\mathbb{R}^n$  with fixed recession cone  $C$  satisfying either of the two conditions. The fact that  $\text{conv}(P \cup Q)$  is a closed convex subset in  $\mathbb{R}^n$  with recession cone  $C$  follows from [Rockafellar 1970, Theorem 9.8.1]. If  $C$  satisfies condition (1), then we may apply our previous argument. If  $C$  satisfies condition (2), then  $P$  and  $Q$  are parallel to  $C$ , and hence so is  $P + Q$ . The result follows.  $\square$

To tie this theorem to our earlier work, we make note of the following:

**Corollary 3.5.** *The empty set, together with the set of all closed convex sets in  $\mathbb{R}^n$  with recession cone equal to  $\mathbb{R}_+^n$ , and operations defined in (2-1), is a semiring.*

**Remark 3.6.** The set of all closed convex sets in  $\mathbb{R}^n$  with recession cone  $C$  equal to  $\mathbb{R}^n$  is the trivial semiring  $\{C\}$ .

### Acknowledgements

The authors thank professor Sam Lisi and graduate student Anastasiia Minenkova at the University of Mississippi, and Marten Wortel, a postdoctoral researcher at North-West University in Potchefstroom, South Africa, for helpful conversations regarding this material. The authors also thank the referee, whose comments have greatly improved this manuscript.

### References

- [Aliprantis and Border 2006] C. D. Aliprantis and K. C. Border, *Infinite dimensional analysis: a hitchhiker's guide*, 3rd ed., Springer, 2006. MR Zbl
- [Border 1985] K. C. Border, *Fixed point theorems with applications to economics and game theory*, Cambridge Univ. Press, 1985. MR Zbl
- [Border 2002] K. C. Border, "Sums of sets, etc.", course notes, California Institute of Technology, 2002, available at <http://people.hss.caltech.edu/~kcb/Notes/AsymptoticCones.pdf>.
- [Cormen et al. 2001] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to algorithms*, 2nd ed., MIT Press, Cambridge, MA, 2001. MR Zbl
- [Debreu 1959] G. Debreu, *Theory of value: an axiomatic analysis of economic equilibrium*, Cowles Foundation Res. Econ. Yale Univ. **17**, Wiley, New York, 1959. MR Zbl
- [Maclagan and Sturmfels 2015] D. Maclagan and B. Sturmfels, *Introduction to tropical geometry*, Graduate Studies in Mathematics **161**, American Mathematical Society, Providence, RI, 2015. MR Zbl

- [Rockafellar 1970] R. T. Rockafellar, *Convex analysis*, Princeton Mathematical Series **28**, Princeton Univ. Press, 1970. MR Zbl
- [Schneider 2014] R. Schneider, *Convex bodies: the Brunn–Minkowski theory*, expanded 2nd ed., Encyclopedia of Mathematics and its Applications **151**, Cambridge Univ. Press, 2014. MR Zbl
- [Shveidel 2001] A. P. Shveidel, “Recession cones of star-shaped and co-star-shaped sets”, pp. 403–414 in *Optimization and related topics* (Ballarat/Melbourne, 1999), edited by A. Rubinov and B. Glover, Applied Optimization **47**, Kluwer, Dordrecht, 2001. MR Zbl
- [Speyer and Sturmfels 2009] D. Speyer and B. Sturmfels, “Tropical mathematics”, *Math. Mag.* **82**:3 (2009), 163–173. MR Zbl
- [Woo 2013] C. Woo, “Recession cone”, web page, 2013, available at <http://planetmath.org/recessioncone>.
- [Ziegler 1995] G. M. Ziegler, *Lectures on polytopes*, Graduate Texts in Mathematics **152**, Springer, 1995. MR Zbl

Received: 2016-12-08    Revised: 2017-05-26    Accepted: 2017-06-13

jmnorton@go.olemiss.edu    *Department of Mathematics, University of Mississippi,  
University, MS, United States*

spiroff@olemiss.edu    *Department of Mathematics, University of Mississippi,  
University, MS, United States*

# A tale of two circles: geometry of a class of quartic polynomials

Christopher Frayer and Landon Gauthier

(Communicated by Michael Dorff)

Let  $\mathcal{P}$  be the family of complex-valued polynomials of the form  $p(z) = (z-1)(z-r_1)(z-r_2)^2$  with  $|r_1| = |r_2| = 1$ . The Gauss–Lucas theorem guarantees that the critical points of  $p \in \mathcal{P}$  will lie within the unit disk. This paper further explores the location and structure of these critical points. For example, the unit disk contains two “desert” regions, the open disk  $\{z \in \mathbb{C} : |z - \frac{3}{4}| < \frac{1}{4}\}$  and the interior of  $2x^4 - 3x^3 + x + 4x^2y^2 - 3xy^2 + 2y^4 = 0$ , in which critical points of  $p$  cannot occur. Furthermore, each  $c$  inside the unit disk and outside of the two desert regions is the critical point of at most two polynomials in  $\mathcal{P}$ .

## 1. Introduction

Given a complex-valued polynomial  $p(z)$ , the Gauss–Lucas theorem guarantees that its critical points lie in the convex hull of its roots. Critical points of polynomials of the form

$$p(z) = (z-1)(z-r_1)(z-r_2)$$

with  $|r_1| = |r_2| = 1$  are studied in [Frayer et al. 2014]. For such a polynomial, a critical point almost always determines  $p$  uniquely, and the unit disk contains a *desert*, the open disk  $\{z \in \mathbb{C} : |z - \frac{2}{3}| < \frac{1}{3}\}$ , in which critical points of  $p$  cannot occur.

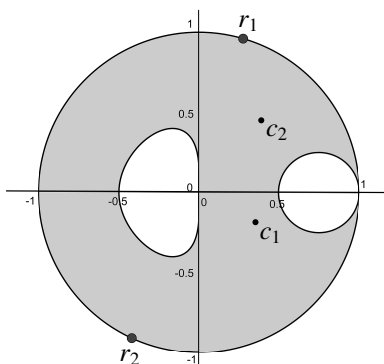
This paper extends the results of [Frayer et al. 2014] to a family of polynomials of the form

$$\mathcal{P} = \{p : \mathbb{C} \rightarrow \mathbb{C} : p(z) = (z-1)(z-r_1)(z-r_2)^2, |r_1| = |r_2| = 1\}.$$

We used GeoGebra to investigate the critical points of  $p(z) = (z-1)(z-r_1)(z-r_2)^2$ . In Figure 1, we set  $r_1$  and  $r_2$  in motion around the unit circle and traced the loci of the critical points with the color gray. Much to our surprise, the unit disk contained

*MSC2010:* 30C15.

*Keywords:* geometry of polynomials, critical points, Gauss–Lucas theorem.



**Figure 1.** Letting the roots vary and tracking the loci of the critical points yields a very surprising result.

two desert regions. In this paper we determine the boundary equations of the desert regions and characterize the critical points of polynomials in  $\mathcal{P}$ .

### 2. Preliminary information

Circles tangent to the line  $x = 1$  will appear frequently throughout this paper. We let  $T_\alpha$  denote the circle of diameter  $\alpha$  passing through  $1$  and  $1 - \alpha$  in the complex plane. That is,

$$T_\alpha = \{z \in \mathbb{C} : |z - (1 - \frac{1}{2}\alpha)| = \frac{1}{2}\alpha\}.$$

For example,  $T_2$  is the unit circle. A key result from [Frayer et al. 2014] will be used to analyze critical points of a polynomial in  $\mathcal{P}$ .

**Theorem 1** [Frayer et al. 2014]. *Suppose  $f(z) = (z - 1)(z - r_1) \cdots (z - r_n)$  with  $|r_k| = 1$  for each  $k$ . Let  $c_1, c_2, \dots, c_n$  denote the critical points of  $f(z)$ , and suppose that  $1 \neq c_k \in T_{\alpha_k}$  for each  $k$ . Then*

$$\sum_{k=1}^n \frac{1}{\alpha_k} = n. \tag{1}$$

An additional fact of interest is related to fractional linear transformations. Functions of the form

$$f(z) = e^{i\theta} \frac{z - \alpha}{\bar{\alpha}z - 1}$$

with  $|\alpha| < 1$  are the only one-to-one analytic mappings of the unit disk onto itself [Saff and Snider 1993, p. 334]. Therefore, the only fractional linear transformations sending the unit circle to the unit circle are of the form  $f(z)$  or  $1/f(z)$ . In either case, writing  $e^{i\theta} = e^{i\theta/2}/e^{-i\theta/2}$  leads to the following result.

**Theorem 2.** *A fractional linear transformation  $T$  sends the unit circle to the unit circle if and only if*

$$T(z) = \frac{\bar{\alpha}z - \bar{\beta}}{\beta z - \alpha}$$

for some  $\alpha, \beta \in \mathbb{C}$  with  $|\alpha/\beta| \neq 1$ .

### 3. Critical points

A polynomial of the form

$$p(z) = (z - 1)(z - r_1)(z - r_2)^2 \in \mathcal{P}$$

has three critical points: one trivial critical point at the repeated root  $r_2$ , and two additional critical points. Differentiation yields

$$p'(z) = (z - r_2)(4z^2 - (3r_1 + 2r_2 + 3)z + r_1r_2 + 2r_1 + r_2).$$

**Definition 3.** We define the *nontrivial* critical points of  $p$  to be the two roots of

$$q(z) = 4z^2 - (3r_1 + 2r_2 + 3)z + r_1r_2 + 2r_1 + r_2.$$

We begin by analyzing a few special cases for future reference.

**Example 4.** Let  $p \in \mathcal{P}$  have a nontrivial critical point at  $z = 1$ . Then  $p$  must have a repeated root at  $z = 1$ . Therefore,  $p \in \mathcal{P}$  has a nontrivial critical point at  $z = 1$  if and only if  $p(z) = (z - 1)^2(z - r)^2$  or  $p(z) = (z - 1)^3(z - r)$  for some  $r \in T_2$ .

Now that we know which polynomials in  $\mathcal{P}$  have a nontrivial critical point at  $c = 1$ , we will assume that  $c \neq 1$  as necessary throughout the remainder of the paper.

**Example 5.** Let  $p \in \mathcal{P}$  have a nontrivial critical point at  $c \in T_2$ , where  $c \neq 1$ . The Gauss–Lucas theorem implies that  $c$  is a root of  $p$ . In order for  $c$  to be a nontrivial critical point,  $p$  must have a triple root at  $c$ . Therefore,  $p \in \mathcal{P}$  has a nontrivial critical point at  $c \in T_2$ , where  $c \neq 1$ , if and only if  $p(z) = (z - 1)(z - c)^3$ . In this case,  $p'(z) = 4(z - 1)^2(z - (\frac{3}{4} + \frac{1}{4}c))$  and the other nontrivial critical point,  $\frac{3}{4} + \frac{1}{4}c \in T_{1/2}$ , lies on the line segment  $1c$ . In fact, whenever  $p$  has two distinct roots, due to repeated roots, then the critical points of  $p$  lie on the line segment between the two roots.

The Gauss–Lucas theorem guarantees that the nontrivial critical points of  $p \in \mathcal{P}$  lie within the unit disk. But we can say more; there is a *desert*, the open disk  $\{z : z \in T_\alpha \text{ with } 0 < \alpha < \frac{1}{2}\}$ , in which critical points of  $p$  cannot occur. This desert corresponds to the white disk in Figure 1.

**Theorem 6.** *No polynomial  $p \in \mathcal{P}$  has a critical point strictly inside  $T_{1/2}$ .*

*Proof.* Let  $c_1, c_2 \neq 1$  be nontrivial critical points of  $p(z) = (z - 1)(z - r_1)(z - r_2)^2$  with  $c_1 \in T_\alpha$  and  $c_2 \in T_\beta$ . As the trivial critical point lies on  $T_2$ , Theorem 1 gives

$$\frac{1}{2} + \frac{1}{\alpha} + \frac{1}{\beta} = 3. \tag{2}$$

Suppose for the sake of contradiction that  $\alpha < \frac{1}{2}$ . Then

$$\frac{1}{\beta} < \frac{5}{2} - 2 = \frac{1}{2}$$

implies  $\beta > 2$ , which violates the Gauss–Lucas theorem. □

A similar analysis leads to the following theorem.

**Theorem 7.** *Let  $c_1, c_2 \neq 1$  be nontrivial critical points of  $p \in \mathcal{P}$ . If  $c_1$  lies on  $T_{4/5}$  so does  $c_2$ . Otherwise,  $c_1$  and  $c_2$  lie on opposite sides of  $T_{4/5}$ .*

*Proof.* Let  $c_1 \in T_\alpha$  and  $c_2 \in T_\beta$ . Then, (2) implies  $1/\alpha + 1/\beta = \frac{5}{2}$ . Therefore,  $\alpha = \frac{4}{5}$  if and only if  $\beta = \frac{4}{5}$  and  $\alpha > \frac{4}{5}$  if and only if  $\beta < \frac{4}{5}$ . □

#### 4. The second desert

Figure 1 suggests the existence of two desert regions in which critical points cannot occur. Methods from [Frayer et al. 2014] quickly identify the desert region  $\{z : z \in T_\alpha \text{ with } 0 < \alpha < \frac{1}{2}\}$ . See Theorem 6. Determining the second desert, the white region enclosed by the “bean”-shaped curve in Figure 1, requires a significant amount of analysis.

To begin this analysis we investigate the relationship between the roots and nontrivial critical points of a polynomial in  $\mathcal{P}$ . Given  $p(z) = (z - 1)(z - r_1)(z - r_2)^2$  with a nontrivial critical point at  $c$ , we have

$$0 = q'(c) = 4c^2 - (3r_1 + 2r_2 + 3)c + r_1r_2 + 2r_1 + r_2.$$

Direct calculations give

$$r_1 = \frac{(1 - 2c)r_2 + 4c^2 - 3c}{-r_2 + 3c - 2} \quad \text{and} \quad r_2 = \frac{(2 - 3c)r_1 + 4c^2 - 3c}{-r_1 + 2c - 1}.$$

**Definition 8.** Given  $c \in \mathbb{C}$ , define

$$f_{1,c}(z) = \frac{(1 - 2c)z + 4c^2 - 3c}{-z + 3c - 2} \quad \text{and} \quad f_{2,c}(z) = \frac{(2 - 3c)z + 4c^2 - 3c}{-z + 2c - 1}$$

and let  $S_1 = f_{1,c}(T_2)$  and  $S_2 = f_{2,c}(T_2)$ .

Observe that  $f_{1,c}$  and  $f_{2,c}$  are fractional linear transformations with  $f_{1,c}(r_2) = r_1$  and  $f_{2,c}(r_1) = r_2$ . We have established the following theorem.

**Theorem 9.** *The polynomial  $p(z) = (z - 1)(z - r_1)(z - r_2)^2 \in \mathcal{P}$  has a nontrivial critical point at  $c \neq 1$  if and only if  $f_{1,c}(r_2) = r_1$  and  $f_{2,c}(r_1) = r_2$ .*

When  $c = 1$ ,

$$f_{1,c}(z) = f_{2,c}(z) = \frac{-z + 1}{-z + 1} = 1.$$

If  $c \neq 1$ , then  $f_{1,c}$  and  $f_{2,c}$  are one-to-one with  $(f_{1,c})^{-1} = f_{2,c}$ . Furthermore,  $f_{1,c}(r_2) = r_1 \in T_2$ , so that  $r_1 \in S_1 \cap T_2$ , and  $f_{2,c}(r_1) = r_2 \in T_2$ , so that  $r_2 \in S_2 \cap T_2$ . We can use these facts to classify the polynomials in  $\mathcal{P}$  having a critical point at  $c \neq 1$  in the closed unit disk. We will show that  $|S_1 \cap T_2| = |S_2 \cap T_2|$  (Lemma 10) and that the cardinality of  $S_1 \cap T_2$  is the number of polynomials in  $\mathcal{P}$  having a nontrivial critical point at  $c$  (Lemma 11).

As fractional linear transformations map circles and lines to circles and lines,  $S_1$  is a circle or line. Therefore,  $S_1 = T_2$  or  $|S_1 \cap T_2| \leq 2$ . We will show that  $S_1 \neq T_2$ . If  $S_1 = T_2$ , then  $f_{1,c}(T_2) = T_2$ . Since

$$f_{1,c}(z) = \frac{(1 - 2c)z + 4c^2 - 3c}{-z + 3c - 2},$$

Theorem 2 implies that  $\overline{1 - 2c} = 2 - 3c$  and  $\overline{4c^2 - 3c} = 1$ . The second equation implies  $4c^2 - 3c = 1$  and it follows that

$$0 = 4c^2 - 3c - 1 = (4c + 1)(c - 1)$$

so that  $c = -\frac{1}{4}$  or  $c = 1$ . However,  $c = -\frac{1}{4}$  does not satisfy the equation  $\overline{1 - 2c} = 2 - 3c$ , and when  $c = 1$ , we know  $f_{1,1}(z) = 1$  does not satisfy the hypothesis of Theorem 2. Therefore,  $S_1 \neq T_2$ . Likewise, as  $(f_{1,c})^{-1} = f_{2,c}$ , there is no  $c$  for which  $S_2 = T_2$ .

**Lemma 10.** *If  $c \neq 1$ , then  $|S_1 \cap T_2| = |S_2 \cap T_2| \in \{0, 1, 2\}$ .*

*Proof.* Without loss of generality, suppose  $|S_1 \cap T_2| = 1$  and  $S_2 \cap T_2 = \{a, b\}$  with  $a \neq b$ . By definition of  $S_2$ , there exist  $a_0, b_0 \in T_2$  with  $f_{2,c}(a_0) = a$ ,  $f_{2,c}(b_0) = b$  and  $a_0 \neq b_0$ . Hence,  $f_{1,c}(f_{2,c}(a_0)) = f_{1,c}(a)$  and  $f_{1,c}(f_{2,c}(b_0)) = f_{1,c}(b)$ , which implies

$$f_{1,c}(a) = a_0 \quad \text{and} \quad f_{1,c}(b) = b_0$$

so that  $|S_1 \cap T_2| > 1$ ; a contradiction. Therefore,  $|S_1 \cap T_2| = |S_2 \cap T_2|$ . □

The following lemma characterizes the three possible cardinalities of  $S_1 \cap T_2$ .

**Lemma 11.** *Suppose  $c \neq 1$ .*

- (1) *If  $S_1$  and  $T_2$  are disjoint, then no  $p \in \mathcal{P}$  has a critical point at  $c$ .*
- (2) *If  $S_1$  and  $T_2$  are tangent, then  $c$  is the nontrivial critical point of exactly one  $p \in \mathcal{P}$ .*
- (3) *If  $S_1$  and  $T_2$  intersect in two distinct points, then  $c$  is the nontrivial critical point of exactly two polynomials in  $\mathcal{P}$ .*

*Proof.* Suppose  $c \neq 1$ . If  $S_1 \cap T_2 = \emptyset$ , then no point in  $\mathbb{C}$  is eligible to be  $r_1$  or  $r_2$  and it follows that no  $p \in \mathcal{P}$  has a critical point at  $c$ . If  $S_1 \cap T_2 = \{a\}$ , it follows from Lemma 10 that  $S_2 \cap T_2 = \{b\}$ . By the definitions of  $S_1$  and  $S_2$ , there exist  $a_0, b_0 \in T_2$  with  $f_{1,c}(a_0) = a$  and  $f_{2,c}(b_0) = b$ . As  $(f_{1,c})^{-1} = f_{2,c}$ , we have

$$a_0 = f_{2,c}(a) \quad \text{and} \quad b_0 = f_{1,c}(b).$$

Therefore  $a_0 = b$  and  $b_0 = a$ . By Theorem 9,  $c$  is a nontrivial critical point of  $p(z) = (z-1)(z-a)(z-b)^2$ . Furthermore, as  $r_1 \in S_1 \cap T_2 = \{a\}$  and  $r_2 \in S_2 \cap T_2 = \{b\}$ , no other  $p \in \mathcal{P}$  has a nontrivial critical point at  $c$ .

If  $S_1 \cap T_2 = \{a, b\}$  with  $a \neq b$ , it follows from Lemma 10 that  $S_2 \cap T_2 = \{d, e\}$  with  $d \neq e$ . By the definition of  $S_1$ , there exist  $a_0, b_0 \in T_2$  with  $f_{1,c}(a_0) = a$ ,  $f_{1,c}(b_0) = b$  and  $a_0 \neq b_0$ . Hence,  $a_0 = f_{2,c}(a)$  and  $b_0 = f_{2,c}(b)$  and it follows that  $\{a_0, b_0\} = \{d, e\}$ . Therefore,  $f_{2,c}(a) = a_0$  and  $f_{1,c}(a_0) = a$ . Theorem 9 implies that  $c$  is a nontrivial critical point of  $p_1(z) = (z-1)(z-a)(z-a_0)^2$ . Likewise,  $f_{2,c}(b) = b_0$  and  $f_{1,c}(b_0) = b$  implies that  $c$  is also a nontrivial critical point of  $p_2(z) = (z-1)(z-b)(z-b_0)^2$ . Moreover, as  $r_1 \in S_1 \cap T_2 = \{a, b\}$ , we have exhausted the potential candidates for  $r_1$  and no other  $p \in \mathcal{P}$  has a nontrivial critical point at  $c$ . When  $|S_1 \cap T_2| = 2$ , there are exactly two polynomials in  $\mathcal{P}$  with a nontrivial critical point at  $c$ . □

In light of Lemmas 10 and 11,  $S_1$  alone is sufficient to characterize the nontrivial critical points of polynomials in  $\mathcal{P}$ .

**4.1. Analyzing  $S_1$ .** To determine the boundary equation of the second desert region, we need to further explore  $S_1$ . Let  $1 \neq c \in \mathbb{C}$ . Since

$$f_{1,c}(z) = \frac{(1-2c)z + 4c^2 - 3c}{-z + 3c - 2}$$

is a fractional linear transformation,  $S_1$  will be a line when there exists  $z \in T_2$  with  $-z + 3c - 2 = 0$ . This occurs when

$$|3c - 2| = |z| = 1 \iff \left|c - \frac{2}{3}\right| = \frac{1}{3}.$$

Therefore,  $S_1$  is a line whenever  $c \in T_{2/3}$ . We now investigate an example for future reference.

**Example 12.** Let  $c \in T_{2/3}$ . Then,  $S_1$  is a line passing through  $f_{1,c}(1) = \frac{1}{3}(4c - 1)$  and  $f_{1,c}(-1) = (4c^2 - c - 1)/(3c - 1)$ . Moreover,

$$f_{1,c}(1) - f_{1,c}(-1) = \frac{4 - 4c}{9c - 3}. \tag{3}$$

Substituting  $c = \frac{2}{3} + \frac{1}{3}e^{i\theta}$  into (3) and simplifying yields  $\text{Re}(f_c(1) - f_c(-1)) = 0$ . When  $c \in T_{2/3}$ , we have  $S_1$  is a vertical line through  $f_{1,c}(1) = \frac{1}{3}(4c - 1) \in T_{8/9}$ .



For  $c \notin T_{2/3}$ , we will determine the center and radius of  $S_1$ . By definition,  $z \in S_1$  if and only if there exists a  $w \in T_2$  with  $f_{1,c}(w) = z$ . That is,  $w = (f_{1,c})^{-1}(z) = f_{2,c}(z) \in T_2$ , which is true if and only if  $|f_{2,c}(z)| = 1$ . Equivalently,

$$\left| \frac{(2 - 3c)(z) + 4c^2 - 3c}{-z + 2c - 1} \right| = 1.$$

Therefore,  $z \in S_1$  if and only if

$$|z - (2c - 1)| = |2 - 3c| \left| z - \frac{3c - 4c^2}{2 - 3c} \right|. \tag{4}$$

For  $k \neq 1$ , the solution set of

$$|z - u| = k|z - v|$$

is a circle with center  $C$  and radius  $R$  satisfying

$$C = v + \frac{v - u}{k^2 - 1} \quad \text{and} \quad R^2 = |C|^2 - \frac{k^2|v|^2 - |u|^2}{k^2 - 1}.$$

Observe that when  $k = |2 - 3c| = 1$ ,

$$\left| \frac{2}{3} - c \right| = \frac{1}{3} \iff c \in T_{2/3}$$

and by Example 12,  $S_1$  is a line. When  $c \in T_\alpha$  with  $\alpha \neq \frac{2}{3}$ , we have  $k = |2 - 3c| \neq 1$  and routine calculations establish the following lemma.

**Lemma 13.** *Suppose  $c \neq 1$  and  $c \in T_\alpha$  with  $\alpha \neq \frac{2}{3}$ . Then,  $S_1$  is a circle with center  $\gamma$  and radius  $r$  given by*

$$\gamma = \frac{4c - 1}{3} + \frac{2\alpha}{9\alpha - 6} \quad \text{and} \quad r = \frac{2\alpha}{3|3\alpha - 2|}.$$

We now study a special case.

**Example 14.** Suppose  $c \in T_2$  with  $c \neq 1$ . Direct calculations give

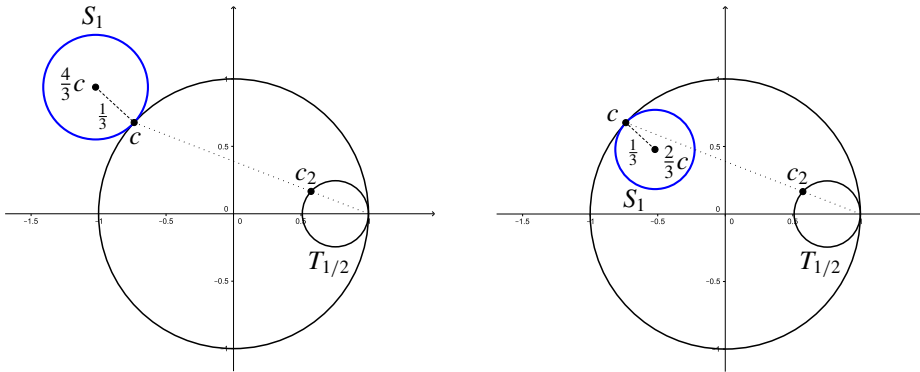
$$f_{1,c}(c) = c, \quad f_{1,c}(1) = \frac{4c - 1}{3} \quad \text{and} \quad f_{1,c}(-1) = \frac{4c^2 - c - 1}{3c - 1},$$

so that

$$|f_{1,c}(z) - \frac{4}{3}c| = \frac{1}{3}$$

for  $z \in \{c, \pm 1\}$ . Therefore, for  $c \in T_2$  with  $c \neq 1$ , we have  $S_1$  is a circle with radius  $\frac{1}{3}$  and center  $\frac{4}{3}c$ , which is externally tangent to  $T_2$  at  $c$ . See Figure 2.

When  $1 \neq c \in T_2$ , it follows from Example 5 that the other nontrivial critical point,  $c_2 = \frac{3}{4} + \frac{1}{4}c \in T_{1/2}$ , lies on the line segment  $\overline{1c}$ . Similar calculations show that for  $c_2 = \frac{3}{4} + \frac{1}{4}c$ , we have  $S_1$  is a circle with radius  $\frac{1}{3}$  and center  $\frac{2}{3}c$ , which is internally tangent to  $T_2$  at  $c$ . See Figure 2.



**Figure 2.** Left: for  $c \in T_2$  with  $c \neq 1$ , the circle  $S_1$  is externally tangent to  $T_2$  at  $c$ . Right: for the corresponding nontrivial critical point,  $c_2$ , the circle  $S_1$  is internally tangent to  $T_2$  at  $c$ .

**4.2. When is  $S_1$  tangent to  $T_2$ ?** Let  $1 \neq c \in \mathbb{C}$ . When  $S_1 \cap T_2 = \emptyset$ , Lemma 11 implies that  $c$  is not the critical point of any  $p \in \mathcal{P}$ . To better understand this case, we will determine when  $|S_1 \cap T_2| = 1$ . That is, for what  $c$  in the unit disk will  $S_1$  and  $T_2$  be tangent? By Example 14, if  $c \in T_{1/2}$ , where  $T_{1/2}$  is the boundary of the first desert region, then  $S_1$  is internally tangent to  $T_2$ . Additionally, if  $c \in T_\alpha$  with  $\alpha < \frac{1}{2}$ , it follows from Theorem 6 that  $S_1$  and  $T_2$  are disjoint.

For  $1 \neq c \in T_\alpha$  with  $\frac{1}{2} \leq \alpha \leq 2$ , if  $S_1$  is internally tangent to  $T_2$ , then

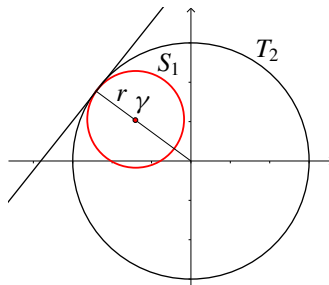
$$|\gamma| + r = 1. \tag{5}$$

See Figure 3. For  $R = 2\alpha/(9\alpha - 6)$ , the circle  $S_1$  has center  $\gamma = \frac{1}{3}(4c - 1) + R$  and radius  $r = |R|$ . Substituting into (5) and setting  $c = x + iy$  gives

$$(4x - 1 + 3R)^2 + 16y^2 = 9(1 - |R|)^2. \tag{6}$$

Since  $R$  depends upon  $\alpha$ , we denote (6) by  $D_\alpha$ .

Since  $r > 0$ , (5) is satisfied if and only if  $S_1$  is internally tangent to  $T_2$  or  $S_1 = T_2$ . Recalling that there is no  $c$  for which  $S_1 = T_2$ , we obtain the following result.



**Figure 3.** When  $|\gamma| + r = 1$ , the circle  $S_1$  will be internally tangent to  $T_2$ .

**Lemma 15.** *Let  $c \neq 1$  and  $\frac{1}{2} \leq \alpha \leq 2$ . Then,  $S_1$  is internally tangent to  $T_2$  if and only if  $c \in T_\alpha \cap D_\alpha$ .*

To apply Lemma 15 we need to determine when and where the circles  $T_\alpha$  and  $D_\alpha$  intersect, that is, the values of  $\alpha$  for which  $T_\alpha \cap D_\alpha \neq \emptyset$ , and the corresponding points of intersection. Because of the  $|R| = |2\alpha/(9\alpha - 6)|$  appearing in (6), we consider three cases:

- (1)  $\frac{1}{2} \leq \alpha < \frac{2}{3}$ ;
- (2)  $\alpha = \frac{2}{3}$ ;
- (3)  $\frac{2}{3} < \alpha \leq 2$ .

In the first case,  $|R| = -R$  and (6) becomes

$$\left(x - \left(1 - \frac{11\alpha - 6}{12\alpha - 8}\right)\right)^2 + y^2 = \left(\frac{11\alpha - 6}{12\alpha - 8}\right)^2.$$

For  $\frac{1}{2} \leq \alpha < \frac{2}{3}$ , circles  $T_\alpha$  and  $D_\alpha$  intersect if and only if  $\alpha = \frac{1}{2}$ . When  $\alpha = \frac{1}{2}$ ,  $T_{1/2} = D_{1/2}$  and by Lemma 15, when  $c \in T_{1/2}$ , we have  $S_1$  is internally tangent to  $T_2$ . See Example 14.

In the second case,  $\alpha = \frac{2}{3}$  and  $D_\alpha$  is undefined. By Example 12, when  $c \in T_{2/3}$ ,  $S_1$  is a vertical line passing through  $f_{1,c}(1) = \frac{1}{3}(4c - 1) \in T_{8/9}$  and  $S_1$  is not tangent to  $T_2$ .

In the third case,  $|R| = R$  and (6) becomes

$$\left(x - \left(-\frac{1}{2} + \frac{7\alpha - 6}{12\alpha - 8}\right)\right)^2 + y^2 = \left(\frac{7\alpha - 6}{12\alpha - 8}\right)^2.$$

For  $\frac{2}{3} < \alpha \leq 2$ , the circles  $T_\alpha$  and  $D_\alpha$  intersect if and only if  $1 \leq \alpha \leq \frac{3}{2}$ . To determine the values of  $c$  where  $S_1$  is internally tangent to  $T_2$ , we need to find the intersection of the circles  $D_\alpha$  and  $T_\alpha$ . Upon simplification, these equations become

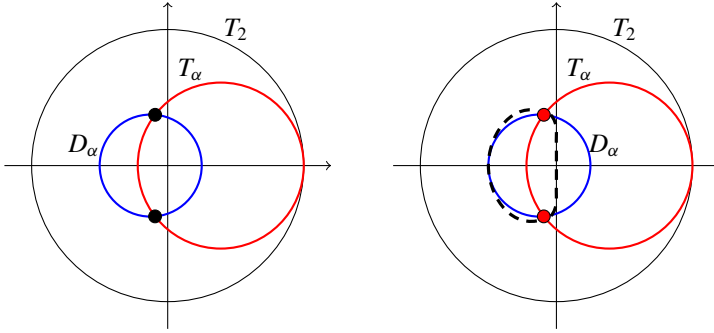
$$\begin{aligned} (4x - 1 + 3R)^2 + 16y^2 &= 9(1 - R)^2, \\ \alpha(1 - x) - (1 - x)^2 &= y^2. \end{aligned}$$

By setting  $R = 2\alpha/(9\alpha - 6)$  and using substitution, we eventually obtain

$$x = \frac{2(\alpha - 1)^2}{(2\alpha - 1)(\alpha - 2)} \quad \text{and} \quad y = \pm \frac{\alpha\sqrt{(3 - 2\alpha)(\alpha - 1)}}{(2\alpha - 1)(\alpha - 2)}. \tag{7}$$

As  $\alpha$  varies from 1 to  $\frac{3}{2}$ , a parametric curve is formed. See Figure 4. For each value of  $c$  on the parametric curve,  $S_1$  is internally tangent to  $T_2$ . Using resultant methods, see [Sederberg et al. 1984], the curve can be implicitized. Substituting  $t = \alpha - 1$  into (7) implies  $0 \leq t \leq \frac{1}{2}$  and

$$x = \frac{2t^2}{2t^2 - t - 1} \quad \text{and} \quad y^2 = \frac{-2t^4 - 3t^3 + t}{4t^4 - 4t^3 - 3t^2 + 2t + 1}, \tag{8}$$



**Figure 4.** Left: when  $1 \leq \alpha \leq \frac{3}{2}$ , we have  $|D_\alpha \cap T_\alpha| = 2$ . Right: as  $\alpha$  varies from 1 to  $\frac{3}{2}$ , parametric equations (7) trace the boundary of the second desert.

so that

$$f = (2x - 2)t^2 + (-x)t + (-x) = 0,$$

$$g = (4y^2 + 2)t^4 + (-4y^2 + 3)t^3 + (-3y^2)t^2 + (2y^2 - 1)t + y^2 = 0.$$

The resultant of  $f$  and  $g$  with respect to  $t$ ,

$$\text{Res}(f, g; t) = \begin{vmatrix} 2x - 2 & -x & -x & 0 & 0 & 0 \\ 0 & 2x - 2 & -x & -x & 0 & 0 \\ 0 & 0 & 2x - 2 & -x & -x & 0 \\ 0 & 0 & 0 & 2x - 2 & -x & -x \\ 4y^2 + 2 & -4y^2 + 3 & -3y^2 & 2y^2 - 1 & y^2 & 0 \\ 0 & 4y^2 + 2 & -4y^2 + 3 & -3y^2 & 2y^2 - 1 & y^2 \end{vmatrix},$$

eliminates the variable  $t$  and is the implicit form of the curve. With the assistance of Mathematica, we find

$$\text{Res}(f, g; t) = 2x^4 - 3x^3 + x + 4x^2y^2 - 3xy^2 + 2y^4$$

and the cartesian representation of (7) is

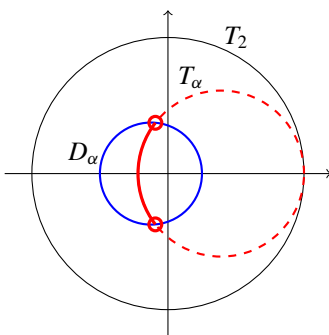
$$2x^4 - 3x^3 + x + 4x^2y^2 - 3xy^2 + 2y^4 = 0. \tag{9}$$

Equation (9) represents the boundary of the second desert region.

**Theorem 16.** *No polynomial in  $\mathcal{P}$  has a critical point strictly inside  $2x^4 - 3x^3 + x + 4x^2y^2 - 3xy^2 + 2y^4 = 0$ .*

*Proof.* Let  $c = x + iy \in T_\alpha$  with  $\alpha \in [1, \frac{3}{2}]$ . Then,  $c$  lies inside  $2x^4 - 3x^3 + x + 4x^2y^2 - 3xy^2 + 2y^4 = 0$  whenever

$$(4x - 1 + 3R)^2 + 16y^2 < 9(1 - R)^2 \quad \text{and} \quad x + iy \in T_\alpha.$$



**Figure 5.** The bold semicircle lies strictly inside the circle  $(4x - 1 + 3R)^2 + 16y^2 = 9(1 - R)^2$  and on  $T_\alpha$ .

See Figure 5. Equivalently, (5) and (6) imply  $|\gamma| + r < 1$  and  $c \in T_\alpha$ . Therefore,  $S_1$  and  $T_2$  are disjoint. By Lemma 11,  $c$  is not the critical point of any  $p \in \mathcal{P}$ .  $\square$

The analysis of the circles  $D_\alpha$  and  $T_\alpha$  has established the following result.

**Lemma 17.** *The circle  $S_1$  is internally tangent to  $T_2$  if and only if  $c = x + iy$  is on  $T_{1/2}$  or  $2x^4 - 3x^3 + x + 4x^2y^2 - 3xy^2 + 2y^4 = 0$ .*

Furthermore, for  $c \in T_\alpha$  with  $\frac{1}{2} \leq \alpha \leq 2$ , the circle  $S_1$  will be externally tangent to  $T_2$  if and only if  $|\gamma| - r = 1$ . A similar, but less involved, analysis leads to the following result.

**Lemma 18.** *The circle  $S_1$  is externally tangent to  $T_2$  if and only if  $c \in T_2$ .*

### 5. Main result

We are now ready to characterize the critical points of a polynomial in  $\mathcal{P}$ . Let  $O$  represent the region strictly inside the closed unit disk and outside of  $T_{1/2}$  and  $2x^4 - 3x^3 + x + 4x^2y^2 - 3xy^2 + 2y^4 = 0$ . That is,  $O$  is the gray shaded region in Figure 1. Denote the closure of  $O$  by  $\bar{O}$ .

**Theorem 19.** *Let  $c \in \mathbb{C}$ .*

- (1) *The polynomial  $p \in \mathcal{P}$  has a nontrivial critical point at  $c = 1$  if and only if  $p(z) = (z - 1)^2(z - r)^2$  or  $p(z) = (z - 1)^3(z - r)$  for some  $r \in T_2$ .*
- (2) *If  $c \notin \bar{O}$ , there is no  $p \in \mathcal{P}$  with a critical point at  $c$ .*
- (3) *If  $1 \neq c \in \bar{O} - O$ , there is a unique  $p \in \mathcal{P}$  with a nontrivial critical point at  $c$ .*
- (4) *If  $c \in O$ , there are exactly two polynomials in  $\mathcal{P}$  with a nontrivial critical point at  $c$ .*

*Proof.* A polynomial  $p \in \mathcal{P}$  has a nontrivial critical point at  $c = 1$  if and only if  $p$  has a repeated root at 1, that is,  $p(z) = (z - 1)^2(z - r)^2$  or  $p(z) = (z - 1)^3(z - r)$  for some  $r \in T_2$ . See Example 4.

Let  $c$  lie strictly inside  $T_{1/2}$ , strictly inside  $2x^4 - 3x^3 + x + 4x^2y^2 - 3xy^2 + 2y^4 = 0$ , or strictly outside  $T_2$ . Then, it follows from Theorems 6, 16 and the Gauss–Lucas theorem respectively, that no  $p \in \mathcal{P}$  has a critical point at  $c$ .

Let  $c \neq 1$  lie on  $T_2$ ,  $T_{1/2}$ , or  $2x^4 - 3x^3 + x + 4x^2y^2 - 3xy^2 + 2y^4 = 0$ . Lemmas 17 and 18 imply that  $S_1$  is tangent to  $T_2$ . Therefore, by Lemma 11, there is exactly one  $p \in \mathcal{P}$  with a nontrivial critical point at  $c$ .

Lastly, we need to show that for  $c \in O$ , we have  $|S_1 \cap T_2| = 2$ . This follows from a “root dragging” argument. Without loss of generality, suppose  $S_1 \cap T_2 = \emptyset$  with  $S_1$  contained inside of  $T_2$ . As we “drag”  $c$  to  $T_2$  along a line segment contained in  $O$ ,  $S_1$  is continuously transformed into a circle externally tangent to  $T_2$ . By continuity, there exists a  $c_0$  on the line segment with  $S_1$  internally tangent to  $T_2$ . As  $c$  never crosses  $T_{1/2}$  or  $2x^4 - 3x^3 + x + 4x^2y^2 - 3xy^2 + 2y^4 = 0$ , this contradicts Lemma 17. Therefore,  $|S_1 \cap T_2| = 2$  and by Lemma 11 there are exactly two polynomials in  $\mathcal{P}$  with a nontrivial critical point at  $c$ .  $\square$

This completes the characterization of critical points of polynomials in  $\mathcal{P}$ . Our results can be extended to polynomials of the form

$$p(z) = (z - 1)^k (z - r_1)^m (z - r_2)^n$$

with  $|r_1| = |r_2| = 1$  and  $\{k, m, n\} \subseteq \mathbb{N}$ . Similar to  $\mathcal{P}$ , when  $m \neq n$ , the unit disk contains two “desert” regions in which critical points cannot occur, and each  $c$  inside the unit disk and outside of the desert regions is the critical point of exactly two such polynomials. However, some questions remain unanswered. For example, if a polynomial has four or more distinct roots, all of which lie on the unit circle, how many desert regions will be in the unit disk?

### Acknowledgment

We are grateful to Dave Boyles for showing us how to use resultant methods to implicitize the boundary equation of the second desert region.

### References

- [Frayer et al. 2014] C. Frayer, M. Kwon, C. Schafhauser, and J. A. Swenson, “The geometry of cubic polynomials”, *Math. Mag.* **87**:2 (2014), 113–124. MR Zbl
- [Saff and Snider 1993] E. B. Saff and A. D. Snider, *Fundamentals of complex analysis for mathematics, science, and engineering*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1993. Zbl
- [Sederberg et al. 1984] T. W. Sederberg, D. C. Anderson, and R. N. Goldman, “Implicit representation of parametric curves and surfaces”, *Comp. Vis. Graph. Image Proc.* **28**:1 (1984), 72–84. Zbl

Received: 2017-02-21      Revised: 2017-06-05      Accepted: 2017-06-13

frayerc@uwplatt.edu

*Department of Mathematics, University of Wisconsin,  
Platteville, WI, United States*

gauthierl@uwplatt.edu

*University of Wisconsin, Platteville, WI, United States*

# Zeros of polynomials with four-term recurrence

Khang Tran and Andres Zumba

(Communicated by Kenneth S. Berenhaut)

Given real numbers  $b, c \in \mathbb{R}$ , we form the sequence of polynomials  $\{H_m(z)\}_{m=0}^\infty$  satisfying the four-term recurrence

$$H_m(z) + cH_{m-1}(z) + bH_{m-2}(z) + zH_{m-3}(z) = 0, \quad m \geq 1,$$

with the initial conditions  $H_0(z) = 1$  and  $H_{-1}(z) = H_{-2}(z) = 0$ . We find necessary and sufficient conditions on  $b$  and  $c$  under which the zeros of  $H_m(z)$  are real for all  $m$ , and provide an explicit real interval on which  $\bigcup_{m=0}^\infty \mathcal{Z}(H_m)$  is dense, where  $\mathcal{Z}(H_m)$  is the set of zeros of  $H_m(z)$ .

## 1. Introduction

Consider the sequence of polynomials  $\{H_m(z)\}_{m=0}^\infty$  satisfying the finite recurrence

$$\sum_{k=0}^n a_k(z)H_{m-k}(z) = 0, \quad m \geq n, \quad (1-1)$$

where  $a_k(z)$ ,  $1 \leq k \leq n$ , are complex polynomials. With certain initial conditions, one may ask for the locations of the zeros of  $H_m(z)$  on the complex plane. There are two common approaches to answering this question. The first describes the asymptotic location of the zeros of the generated polynomials, while the second provides the exact location of these zeros (or at least for the zeros of  $H_m(z)$  for  $m \gg 1$ ). Recent works in the first direction include [Beraha et al. 1975; 1978; Borcea et al. 2006; Boyer and Goh 2007; 2008]. Results using the first approach prove useful when establishing the necessary condition for  $H_m(z)$  to be hyperbolic, as we will see in Section 3.

When considering polynomials satisfying a generic recurrence such as (1-1), the task of finding an explicit curve where the zeros of the  $H_m(z)$  must lie is difficult. For three-term recurrences with degree two and appropriate initial conditions, the curve containing zeros is given in [Tran 2014]. The corresponding curve for a three-term recurrence with degree  $n$  is given in [Tran 2015]. Among all possible

---

*MSC2010:* 30C15, 26C10, 11C08.

*Keywords:* generating functions, hyperbolic polynomials, recursive sequence.

curves containing the zeros of the  $H_m(z)$ , the real line plays an important role. We say that a polynomial is hyperbolic if all of its zeros are real. There are a lot of recent works on hyperbolic polynomials and on linear operators preserving hyperbolicity of polynomials; see for example [Bates and Yoshida 2016; Borcea and Brändén 2009; Bunton et al. 2015; Craven and Csordas 2004]. For studies of sequences of hyperbolic polynomials satisfying finite recurrences, see [Eğecioğlu et al. 2001; Forgács and Tran 2016].

The main result of this paper, Theorem 2, is the identification of necessary and sufficient conditions on  $b, c \in \mathbb{R}$  under which the zeros of the sequence of polynomials  $H_m(z)$  satisfying the recurrence

$$\begin{aligned} H_m(z) + cH_{m-1}(z) + bH_{m-2}(z) + zH_{m-3}(z) &= 0, \quad m \geq 1, \\ H_0(z) &\equiv 1, \\ H_m(z) &\equiv 0, \quad m < 0, \end{aligned} \tag{1-2}$$

are real. We use the convention that the zeros of the constant zero polynomial are real.

**Definition 1.** The set of zeros of  $H_m(z)$  is denoted by  $\mathcal{Z}(H_m)$ .

**Theorem 2.** Suppose  $b, c \in \mathbb{R}$ , and let  $\{H_m(z)\}_{m=0}^\infty$  be defined as in (1-2). The zeros of  $H_m(z)$  are real for all  $m$  if and only if one of the two conditions below holds:

- (i)  $c = 0$  and  $b \geq 0$ .
- (ii)  $c \neq 0$  and  $-1 \leq b/c^2 \leq \frac{1}{3}$ .

In the first case, if  $b > 0$ , then  $\bigcup_{m=0}^\infty \mathcal{Z}(H_m)$  is dense in  $(-\infty, \infty)$ . In the second case,  $\bigcup_{m=0}^\infty \mathcal{Z}(H_m)$  is dense in the interval

$$c^3 \cdot \left( -\infty, \frac{1}{27} \left( -2 + 9b/c^2 - 2\sqrt{(1 - 3b/c^2)^3} \right) \right].$$

Our paper is organized as follows. In Section 2, we prove a sufficient condition for the zeros of all  $H_m(z)$  to be real in the case  $c \neq 0$ . The case  $c = 0$  follows from similar arguments whose key differences will be outlined in Section 3. Finally, in Section 4, we prove the necessary condition for the zeros of  $H_m(z)$  to be real.

### 2. The case $c \neq 0$ and $-1 \leq b/c^2 \leq \frac{1}{3}$

We write the sequence  $\{H_m(z)\}_{m=0}^\infty$  in (1-2) using its generating function

$$\sum_{m=0}^\infty H_m(z)t^m = \frac{1}{1 + ct + bt^2 + zt^3}. \tag{2-1}$$

Substituting  $t \rightarrow t/c$ ,  $b/c^2 \rightarrow a$ , and  $z/c^3 \rightarrow z$ , we will prove the following form of the theorem.



**Theorem 3.** Consider the sequence of polynomials  $\{H_m(z)\}_{m=0}^\infty$  generated by

$$\sum_{m=0}^\infty H_m(z)t^m = \frac{1}{1+t+at^2+zt^3}, \tag{2-2}$$

where  $a \in \mathbb{R}$ . If  $-1 \leq a \leq \frac{1}{3}$  then the zeros of  $H_m(z)$  lie in the real interval

$$I_a = \left(-\infty, \frac{1}{27}(-2+9a-2\sqrt{(1-3a)^3})\right], \tag{2-3}$$

and  $\bigcup_{m=0}^\infty \mathcal{Z}(H_m)$  is dense in  $I_a$ .

We will see later that the density of the union of zeros on  $I_a$  follows naturally from the fact that  $\mathcal{Z}(H_m) \subset I_a$  and thus we focus on proving this claim. We note that each value of  $a \in [-1, \frac{1}{3}]$  generates a sequence  $\{H_m(z, a)\}_{m=0}^\infty$ . The lemma below asserts that it suffices to prove that  $\mathcal{Z}(H_m(z, a)) \subset I_a$  for all  $a$  in a dense subset of  $[-1, \frac{1}{3}]$ .

**Lemma 4.** Let  $S$  be a dense subset of  $[-1, \frac{1}{3}]$ , and let  $m \in \mathbb{N}$  be fixed. If

$$\mathcal{Z}(H_m(z, a)) \subset I_a$$

for all  $a \in S$ , then

$$\mathcal{Z}(H_m(z, a^*)) \subset I_{a^*}$$

for all  $a^* \in [-1, \frac{1}{3}]$ .

*Proof.* Let  $a^* \in [-1, \frac{1}{3}]$  be given. By the density of  $S$  in  $[-1, \frac{1}{3}]$ , we can find a sequence  $\{a_n\}$  in  $S$  such that  $a_n \rightarrow a^*$ . For any  $z^* \notin I_{a^*}$ , we will show that  $H_m(z^*, a^*) \neq 0$ . We note that the zeros of  $H_m(z, a_n)$  lie in the interval  $I_{a_n}$  whose right endpoint approaches the right endpoint of  $I_{a^*}$  as  $n \rightarrow \infty$ . If we let  $z_k^{(n)}$ ,  $1 \leq k \leq \deg H_m(z, a_n)$ , be the zeros of  $H_m(z, a_n)$  then

$$|H_m(z^*, a_n)| = \gamma^{(n)} \prod_{k=1}^{\deg H_m(z, a_n)} |z^* - z_k^{(n)}|,$$

where  $\gamma^{(n)}$  is the leading coefficient of  $H_m(z, a_n)$ . Since  $\deg H_m(z, a_n) \leq \lfloor \frac{1}{3}m \rfloor$  (see Lemma 5), using this product representation and the assumption that  $z^* \notin I_a$ , we conclude that there is a fixed (independent of  $n$ )  $\delta > 0$  so that  $|H_m(z^*, a_n)| > \delta$  for all large  $n$ . Since  $H_m(z^*, a)$  is a polynomial in  $a$  for any fixed  $z^*$ , we conclude that

$$H_m(z^*, a^*) = \lim_{n \rightarrow \infty} H_m(z^*, a_n) \neq 0$$

and the result follows. □

Lemma 4 allows us to ignore some special values of  $a$ . In particular, we may assume  $a \neq 0$ . In our main argument, we count the number of zeros of  $H_m(z)$  on the interval  $I_a$  in (2-3) and show that this number is at least as big as

the degree of  $H_m(z)$ . To count the number of zeros of  $H_m(z)$  on  $I_a$ , we write  $z = z(\theta)$  as a strictly increasing function of a variable  $\theta$  on the interval  $(\frac{2\pi}{3}, \pi)$ . Then we construct a function  $g_m(\theta)$  on  $(\frac{2\pi}{3}, \pi)$  with the property that  $\theta$  is a zero of  $g_m(\theta)$  on  $(\frac{2\pi}{3}, \pi)$  if and only if  $z(\theta)$  is a zero of  $H_m(z)$  on  $I_a$ . From this construction, we count the number of zeros of  $g_m(\theta)$  on  $(\frac{2\pi}{3}, \pi)$ , which will be the same as the number of zeros of  $H_m(z)$  on  $I_a$  by the monotonicity of the function  $z(\theta)$ .

We next obtain an upper bound for the degree of  $H_m(z)$  and provide heuristic arguments for the formulas of  $z(\theta)$  and  $g_m(\theta)$ .

**Lemma 5.** *The degree of the polynomial  $H_m(z)$  defined by (2-2) is at most  $\lfloor \frac{1}{3}m \rfloor$ .*

*Proof.* We rewrite (2-2) as

$$(1 + t + at^2 + zt^3) \sum_{m=0}^{\infty} H_m(z)t^m = 1.$$

By equating the coefficients in  $t$  of both sides, we see that the sequence  $\{H_m(z)\}_{m=0}^{\infty}$  satisfies the recurrence

$$H_{m+3}(z) + H_{m+2}(z) + aH_{m+1}(z) + zH_m(z) = 0$$

and the initial conditions

$$H_0(z) \equiv 1, \quad H_1(z) \equiv -1, \quad \text{and} \quad H_2(z) \equiv 1 - a.$$

The lemma follows by induction. □

**2.1. Heuristic arguments.** We now provide heuristic arguments to motivate the formulas for two functions  $z(\theta)$  and  $g_m(\theta)$  on  $(\frac{2\pi}{3}, \pi)$ . Let  $t_0 = t_0(z)$ ,  $t_1 = t_1(z)$ , and  $t_2 = t_2(z)$  be the three zeros of the denominator  $1 + t + at^2 + zt^3$ . We will show rigorously in Section 2.2 that  $t_0, t_1, t_2$  are nonzero and distinct with  $t_0 = \bar{t}_1$ . We let  $q = t_1/t_0 = e^{2i\theta}$ ,  $\theta \neq 0, \pi$ . We have

$$\sum_{m=0}^{\infty} H_m(z)t^m = \frac{1}{1 + t + at^2 + zt^3} = \frac{1}{z(t - t_0)(t - t_1)(t - t_2)}.$$

We apply partial fractions to rewrite the generating function given above as

$$(z(t - t_0)(t_0 - t_1)(t_0 - t_2))^{-1} + (z(t - t_1)(t_1 - t_0)(t_1 - t_2))^{-1} + (z(t - t_2)(t_2 - t_0)(t_2 - t_1))^{-1},$$

which can be expanded as a series in  $t$  as

$$- \sum_{m=0}^{\infty} \frac{1}{z} \left( ((t_0 - t_1)(t_0 - t_2)t_0^{m+1})^{-1} + ((t_1 - t_0)(t_1 - t_2)t_1^{m+1})^{-1} + ((t_2 - t_0)(t_2 - t_1)t_2^{m+1})^{-1} \right) t^m. \quad (2-4)$$

From this expression, we deduce that  $z$  is a zero of  $H_m(z)$  if and only if

$$\begin{aligned} &((t_0 - t_1)(t_0 - t_2)t_0^{m+1})^{-1} + ((t_1 - t_0)(t_1 - t_2)t_1^{m+1})^{-1} \\ &+ ((t_2 - t_0)(t_2 - t_1)t_2^{m+1})^{-1} = 0. \end{aligned} \quad (2-5)$$

After multiplying the left side of (2-5) by  $t_0^{m+3}$ , we obtain the equality

$$\begin{aligned} &((1 - t_1/t_0)(1 - t_2/t_0))^{-1} + ((t_1/t_0 - 1)(t_1/t_0 - t_2/t_0)(t_1/t_0)^{m+1})^{-1} \\ &+ ((t_2/t_0 - 1)(t_2/t_0 - t_1/t_0)(t_2/t_0)^{m+1})^{-1} = 0. \end{aligned}$$

Setting  $\zeta = t_2/t_0 e^{i\theta}$ , we rewrite the left side as

$$\begin{aligned} &((1 - e^{2i\theta})(1 - \zeta e^{i\theta}))^{-1} + ((e^{2i\theta} - 1)(e^{2i\theta} - \zeta e^{i\theta})(e^{2i\theta})^{m+1})^{-1} \\ &+ ((\zeta e^{i\theta} - 1)(\zeta e^{i\theta} - e^{2i\theta})(\zeta e^{i\theta})^{m+1})^{-1}, \end{aligned}$$

or equivalently

$$\begin{aligned} &(e^{2i\theta}(-2i \sin \theta)(e^{-i\theta} - \zeta))^{-1} + ((2i \sin \theta)(e^{i\theta} - \zeta)(e^{2i\theta})^{m+2})^{-1} \\ &+ ((\zeta - e^{-i\theta})(\zeta - e^{i\theta})(\zeta)^{m+1}(e^{i\theta})^{m+3})^{-1}. \end{aligned}$$

We multiply this expression by  $(\zeta - e^{-i\theta})(\zeta - e^{i\theta})e^{i(m+3)\theta}$  and set the summation equal to zero to arrive at

$$\begin{aligned} 0 &= \frac{(\zeta - e^{i\theta})e^{i(m+1)\theta}}{2i \sin \theta} + \frac{e^{-i\theta} - \zeta}{(2i \sin \theta)e^{i(m+1)\theta}} + \frac{1}{\zeta^{m+1}} \\ &= \frac{(\zeta - e^{i\theta})e^{i(m+1)\theta} - (\zeta - e^{-i\theta})e^{-i(m+1)\theta}}{2i \sin \theta} + \frac{1}{\zeta^{m+1}} \\ &= \frac{\zeta(e^{i(m+1)\theta} - e^{-i(m+1)\theta}) + e^{-i(m+2)\theta} - e^{i(m+2)\theta}}{2i \sin \theta} + \frac{1}{\zeta^{m+1}} \\ &= \frac{\zeta(2i \sin((m+1)\theta)) - 2i \sin((m+2)\theta)}{2i \sin \theta} + \frac{1}{\zeta^{m+1}} \\ &= \frac{2i \zeta \sin((m+1)\theta) - 2i \sin((m+1)\theta) \cos \theta - 2i \cos((m+1)\theta) \sin \theta}{2i \sin \theta} + \frac{1}{\zeta^{m+1}} \\ &= \frac{(\zeta - \cos \theta) \sin((m+1)\theta)}{\sin \theta} - \cos((m+1)\theta) + \frac{1}{\zeta^{m+1}}. \end{aligned} \quad (2-6)$$

The last expression will serve as the definition of  $g_m(\theta)$ ; see (2-15).

We next provide a motivation for the specific form of  $z(\theta)$ . Since  $t_0$ ,  $t_1$ , and  $t_2$  are the zeros of  $D(t, z) = 1 + t + at^2 + zt^3$ , they satisfy the three identities

$$t_0 + t_1 + t_2 = -\frac{a}{z}, \quad t_0 t_1 + t_0 t_2 + t_1 t_2 = \frac{1}{z}, \quad \text{and} \quad t_0 t_1 t_2 = -\frac{1}{z}.$$

If we divide the first equation by  $t_0$ , the second by  $t_0^2$ , and the third by  $t_0^3$  then these identities become

$$1 + e^{2i\theta} + \zeta e^{i\theta} = -\frac{a}{zt_0}, \tag{2-7}$$

$$e^{2i\theta} + \zeta e^{i\theta} + \zeta e^{3i\theta} = \frac{1}{zt_0^2}, \tag{2-8}$$

$$\zeta e^{3i\theta} = -\frac{1}{zt_0^3}. \tag{2-9}$$

We next divide (2-7) by (2-8), and (2-8) by (2-9) to obtain

$$\frac{1 + e^{2i\theta} + \zeta e^{i\theta}}{e^{2i\theta} + \zeta e^{i\theta} + \zeta e^{3i\theta}} = -at_0 \quad \text{and} \quad \frac{e^{2i\theta} + \zeta e^{i\theta} + \zeta e^{3i\theta}}{\zeta e^{3i\theta}} = -t_0,$$

from which we deduce that

$$(1 + e^{2i\theta} + \zeta e^{i\theta})\zeta e^{3i\theta} = a(e^{2i\theta} + \zeta e^{i\theta} + \zeta e^{3i\theta})^2.$$

This equation is equivalent to

$$(e^{-i\theta} + e^{i\theta} + \zeta)\zeta e^{4i\theta} = a e^{4i\theta} (1 + \zeta e^{-i\theta} + \zeta e^{i\theta})^2,$$

or simply

$$(2 \cos \theta + \zeta)\zeta = a(1 + 2\zeta \cos \theta)^2.$$

**Lemma 6.** For any  $a \in [-1, \frac{1}{3}]$  and  $\theta \in (\frac{2\pi}{3}, \pi)$ , the zeros in  $\zeta$  of the polynomial

$$(2 \cos \theta + \zeta)\zeta - a(1 + 2\zeta \cos \theta)^2 \tag{2-10}$$

are real and distinct.

*Proof.* We consider the discriminant of the above polynomial in  $\zeta$ :

$$\Delta = (1 - 4a) \cos^2 \theta + a.$$

There are three possible cases depending on the value of  $a$ . If  $\frac{1}{4} \leq a \leq \frac{1}{3}$ , the inequality  $\Delta > 0$  comes directly from

$$a \geq 4a - 1 > (1 - 4a) \cos^2 \theta.$$

If  $0 \leq a < \frac{1}{4}$ , the claim  $\Delta > 0$  is trivial since  $1 - 4a > 0$ . Finally, if  $a < 0$ , we have

$$\Delta \geq \frac{1}{4}(1 - 4a) + a = \frac{1}{4}.$$

It follows that the zeros of (2-10) are real and distinct for any  $a \in [-1, \frac{1}{3}]$  and  $\theta \in (\frac{2\pi}{3}, \pi)$ . □

To obtain a formula for  $z(\theta)$ , we multiply (2-7) and (2-8) to get

$$(1 + e^{2i\theta} + \zeta e^{i\theta})(e^{2i\theta} + \zeta e^{i\theta} + \zeta e^{3i\theta}) = -\frac{a}{z^2 t_0^3},$$

and divide (2-9) by this equation to arrive at

$$\begin{aligned} z &= \frac{ae^{3i\theta}\zeta}{(1+e^{2i\theta}+\zeta e^{i\theta})(e^{2i\theta}+\zeta e^{i\theta}+\zeta e^{3i\theta})} \\ &= \frac{ae^{3i\theta}\zeta}{e^{3i\theta}(e^{-i\theta}+e^{i\theta}+\zeta)(1+\zeta e^{-i\theta}+\zeta e^{i\theta})} = \frac{a\zeta}{(2\cos\theta+\zeta)(1+2\zeta\cos\theta)}. \end{aligned} \quad (2-11)$$

**2.2. Rigorous proof.** Motivated by Section 2.1, we now rigorously prove Theorem 3. We start by defining the function  $\zeta(\theta)$  according to (2-10).

**Definition 7.** The function  $\zeta(\theta)$  is defined on  $(\frac{2\pi}{3}, \pi)$  as

$$\zeta = \zeta(\theta) = \frac{(2a-1)\cos\theta + \sqrt{(1-4a)\cos^2\theta + a}}{1-4a\cos^2\theta}. \quad (2-12)$$

**Remark 8.** From Lemma 6,  $\zeta(\theta)$  is a real function on  $(\frac{2\pi}{3}, \pi)$  with a possible vertical asymptote at

$$\theta = \cos^{-1}\left(-\frac{1}{2\sqrt{a}}\right) \quad (2-13)$$

when  $\frac{1}{4} < a \leq \frac{1}{3}$ . However, we note that the function  $1/\zeta(\theta)$  is a real continuous function on  $(\frac{2\pi}{3}, \pi)$ .

**Lemma 9.** Let  $\zeta(\theta)$  be defined as in (2-12). Then  $|\zeta(\theta)| > 1$  for every  $a \in (-1, \frac{1}{3})$  and every  $\theta \in (\frac{2\pi}{3}, \pi)$  with  $1-4a\cos^2\theta \neq 0$ .

*Proof.* From (2-10), we note that  $\zeta_+ := \zeta(\theta)$  and

$$\zeta_- := \frac{(2a-1)\cos\theta - \sqrt{(1-4a)\cos^2\theta + a}}{1-4a\cos^2\theta}$$

are the zeros of

$$f(\zeta) := (2\cos\theta + \zeta)\zeta - a(1 + 2\zeta\cos\theta)^2.$$

Note that

$$f(-1)f(1) = (-1 + 2\cos\theta)(1 + 2\cos\theta)(4a^2\cos^2\theta - (a-1)^2).$$

If  $\theta \in (\frac{2\pi}{3}, \pi)$  and  $a \in (-1, \frac{1}{3})$ , this product is negative since

$$4a^2\cos^2\theta - (a-1)^2 \leq 4a^2 - (a-1)^2 = (a+1)(3a-1) < 0.$$

Thus exactly one of the zeros of the quadratic function  $f(\zeta)$  lies outside the interval  $[-1, 1]$ . The claim follows from the fact that  $|\zeta_+| > |\zeta_-|$ .  $\square$

Although one can prove Lemma 9 for the extreme values  $a = -1$  or  $a = \frac{1}{3}$ , that will not be necessary by Lemma 4. Next, motivated by (2-11), we define the real function  $z(\theta)$  as follows.

**Definition 10.** The function  $z(\theta)$  is defined on  $(\frac{2\pi}{3}, \pi)$  as

$$z = z(\theta) := \frac{a\zeta}{(2 \cos \theta + \zeta)(1 + 2\zeta \cos \theta)}. \tag{2-14}$$

Lemma 9 implies  $1 + 2\zeta \cos \theta \neq 0$ , and by (2-10), neither is  $2 \cos \theta + \zeta$ . Dividing the numerator and the denominator of (2-14) by  $\zeta^2(\theta)$  and combining with the fact that  $1/\zeta(\theta)$  is continuous on  $(\frac{2\pi}{3}, \pi)$ , we conclude that the possible discontinuity of  $z(\theta)$  in (2-13) is removable. Finally, motivated by (2-6), we define the function  $g_m(\theta)$  as follows.

**Definition 11.** The function  $g_m(\theta)$  is defined on  $(\frac{2\pi}{3}, \pi)$  as

$$g_m(\theta) := \frac{(\zeta - \cos \theta) \sin((m + 1)\theta)}{\sin \theta} - \cos((m + 1)\theta) + \frac{1}{\zeta^{m+1}}. \tag{2-15}$$

We note that  $g_m(\theta)$  has the same vertical asymptote as that of  $\zeta(\theta)$  in (2-13) when  $\frac{1}{4} < a \leq \frac{1}{3}$ .

From Lemma 9, we see that the sign of the function  $g_m(\theta)$  alternates at values of  $\theta$  where  $\cos(m + 1)\theta = \pm 1$ . Thus by the intermediate value theorem, the function  $g_m(\theta)$  has at least one root on each subinterval whose endpoints are solutions of  $\cos(m + 1)\theta = \pm 1$ . However, in the case  $\frac{1}{4} \leq a \leq \frac{1}{3}$ , one of the subintervals contains the vertical asymptote given in (2-13). The lemma below counts the number of zeros of  $g_m(\theta)$  on such a subinterval.

**Lemma 12.** Let  $g_m(\theta)$  be defined as in (2-15). Suppose  $\frac{1}{4} < a \leq \frac{1}{3}$  and  $m \geq 6$ . Then there exists  $h \in \mathbb{N}$  such that

$$\theta_{h-1} := \frac{h-1}{m+1}\pi < \cos^{-1}\left(-\frac{1}{2\sqrt{a}}\right) \leq \frac{h}{m+1}\pi =: \theta_h,$$

where  $\lfloor \frac{2}{3}(m+1) \rfloor + 1 \leq h-1 < h \leq m+1$ . Furthermore, as long as

$$\cos^{-1}\left(-\frac{1}{2\sqrt{a}}\right) \neq \frac{h}{m+1}\pi, \tag{2-16}$$

the function  $g_m(\theta)$  has at least two zeros on the interval

$$\theta \in \left(\frac{h-1}{m+1}\pi, \frac{h}{m+1}\pi\right) := J_h \tag{2-17}$$

whenever  $h$  is at most  $m$ , and at least one zero when  $h$  is  $m+1$ .

*Proof.* Suppose  $a \in (\frac{1}{4}, \frac{1}{3}]$ . Since the function  $\cos^{-1}(-1/(2\sqrt{x}))$  is decreasing on the interval  $(\frac{1}{4}, \frac{1}{3}]$ , we conclude that

$$\cos^{-1}\left(-\frac{1}{2\sqrt{a}}\right) \geq \frac{5\pi}{6}.$$

The existence of  $h$  now follows directly from

$$\frac{\lfloor \frac{2}{3}(m+1) \rfloor + 1}{m+1} \pi < \frac{5\pi}{6},$$

when  $m \geq 6$ .

The vertical asymptote of  $g_m(\theta)$  at  $\cos^{-1}(-1/(2\sqrt{a}))$  divides the interval  $J_h$  in (2-17) into two subintervals. We will show that each subinterval contains at least one zero of  $g_m(\theta)$  if  $h \leq m$ . In the case  $h = m + 1$ , only the subinterval on the left contains at least one zero of  $g_m(\theta)$ . We analyze these two subintervals in the two cases below.

We consider the first case when  $\theta \in J_h$  and  $\theta < \cos^{-1}(-1/(2\sqrt{a}))$ . By Lemma 9 and (2-15) we see that the sign of  $g_m(\theta_{h-1})$  is  $(-1)^h$ . We now show that the sign of  $g_m(\theta)$  is  $(-1)^{h-1}$  when  $\theta \rightarrow \cos^{-1}(-1/(2\sqrt{a}))$ . From (2-12), we observe that  $\zeta(\theta) \rightarrow +\infty$  as  $\theta \rightarrow \cos^{-1}(-1/(2\sqrt{a}))$ . Since  $\theta \in J_h$ , the sign of  $\sin((m+1)\theta)$  is  $(-1)^{h-1}$  and consequently the sign of  $g_m(\theta)$  is  $(-1)^{h-1}$  when  $\theta \rightarrow \cos^{-1}(-1/(2\sqrt{a}))$  by (2-15). By the intermediate value theorem, we obtain at least one zero of  $g_m(\theta)$  in this case.

Next we consider the case when  $\theta \in J_h$  and  $\theta > \cos^{-1}(-1/(2\sqrt{a}))$ . In this case the sign of  $g_m(\theta_h)$  is  $(-1)^{h-1}$  if  $h \leq m$  by Lemma 9. Since  $\zeta(\theta) \rightarrow -\infty$  as  $\theta \rightarrow \cos^{-1}(-1/(2\sqrt{a}))$  and the sign of  $\sin((m+1)\theta)$  is  $(-1)^{h-1}$ , the sign of  $g_m(\theta)$  is  $(-1)^h$  as  $\theta \rightarrow \cos^{-1}(-1/(2\sqrt{a}))$ . By the intermediate value theorem, we obtain at least one zero of  $g_m(\theta)$  in this case if  $h \leq m$ . □

Note that as a consequence of Lemma 4, we may assume that none of the partitioning points under consideration are the points  $\cos^{-1}(-1/(2\sqrt{a}))$ . From the fact that the sign of  $g_m(\theta)$  in (2-15) alternates when  $\cos((m+1)\theta) = \pm 1$ , we can find a lower bound for the number of zeros of  $g_m(\theta)$  on  $(\frac{2\pi}{3}, \pi)$  by the intermediate value theorem. We will relate the zeros of  $g_m(\theta)$  to the zeros of  $H_m(z)$  by (2-6). However to ensure that the partial fractions procedure preceding (2-6) is rigorous, we need the lemma below.

**Lemma 13.** *Let  $\theta \in (0, \pi)$  be such that  $\theta \neq \cos^{-1}(-1/(2\sqrt{a}))$  whenever  $a > \frac{1}{4}$ . The zeros in  $t$  of  $1 + t + at^2 + z(\theta)t^3$  are*

$$t_0 = -\frac{e^{2i\theta} + \zeta e^{i\theta} + \zeta e^{3i\theta}}{\zeta e^{3i\theta}}, \quad t_1 = t_0 e^{2i\theta} \quad \text{and} \quad t_2/t_0 = \zeta e^{i\theta},$$

where  $\zeta := \zeta(\theta)$  is given in (2-12).

*Proof.* We first note that

$$\begin{aligned} P(t_0) &= 1 + t_0 + at_0^2 + zt_0^3 \\ &= -\frac{1}{\zeta e^{i\theta}} - e^{-2i\theta} + \frac{a}{\zeta^2 e^{2i\theta}} (1 + \zeta e^{-i\theta} + \zeta e^{i\theta})^2 - \frac{z}{\zeta^3 e^{3i\theta}} (1 + \zeta e^{-i\theta} + \zeta e^{i\theta})^3, \end{aligned}$$

where  $\zeta$  is a root of the quadratic equation  $(2 \cos \theta + \zeta)\zeta - a(1 + 2\zeta \cos \theta)^2 = 0$ . We apply the identities

$$(1 + \zeta e^{-i\theta} + \zeta e^{i\theta})^2 = (1 + 2\zeta \cos \theta)^2 = \frac{1}{a}(2 \cos \theta + \zeta)\zeta = \frac{1}{a}(e^{-i\theta} + e^{i\theta} + \zeta)\zeta$$

and

$$z = \frac{a\zeta}{(2 \cos \theta + \zeta)(1 + 2\zeta \cos \theta)} = \frac{\zeta^2}{(1 + 2\zeta \cos \theta)^3} = \frac{\zeta^2}{(1 + \zeta e^{-i\theta} + \zeta e^{i\theta})^3}, \quad (2-18)$$

to conclude that  $P(t_0) = 0$ . Similarly, we have

$$\begin{aligned} P(t_1) &= 1 + t_0 e^{2i\theta} + at_0^2 e^{4i\theta} + zt_0^3 e^{6i\theta} \\ &= -\frac{e^{i\theta}}{\zeta} - e^{2i\theta} + \frac{ae^{2i\theta}}{\zeta^2}(1 + \zeta e^{-i\theta} + \zeta e^{i\theta})^2 - \frac{ze^{3i\theta}}{\zeta^3}(1 + \zeta e^{-i\theta} + \zeta e^{i\theta})^3 \\ &= -\frac{e^{i\theta}}{\zeta} - e^{2i\theta} + \frac{ae^{2i\theta}}{\zeta^2} \frac{(e^{-i\theta} + e^{i\theta} + \zeta)\zeta}{a} - \frac{e^{3i\theta}}{\zeta^3} \zeta^2 = 0. \end{aligned}$$

Finally,

$$\begin{aligned} P(t_2) &= P(\zeta t_0 e^{i\theta}) \\ &= -\zeta e^{-i\theta} - \zeta e^{i\theta} + a(1 + \zeta e^{-i\theta} + \zeta e^{i\theta})^2 - z(1 + \zeta e^{-i\theta} + \zeta e^{i\theta})^3 \\ &= -\zeta e^{-i\theta} - \zeta e^{i\theta} + a \frac{1}{a}(e^{-i\theta} + e^{i\theta} + \zeta)\zeta - \zeta^2 = 0. \quad \square \end{aligned}$$

As a consequence of Lemma 13, if  $\theta \in (\frac{2\pi}{3}, \pi)$ , then the zeros of  $1 + t + at^2 + z(\theta)t^3$  will be distinct and  $t_1 = \bar{t}_0$  since  $\zeta \in \mathbb{R}$  by Lemma 6. Thus we can apply the partial fractions given in the beginning of Section 2.1. From this partial fraction decomposition, we conclude that if  $\theta$  is a zero of  $g_m(\theta)$ , then  $z(\theta)$  will be a zero of  $H_m(z)$ . In fact, we claim that each distinct zero of  $g_m(\theta)$  on  $(\frac{2\pi}{3}, \pi)$  produces a distinct zero of  $H_m(z)$  on  $I_a$ . This is the content of the following two lemmas.

**Lemma 14.** *Let  $\zeta(\theta)$  be defined as in (2-12). The function  $z(\theta)$  defined as in (2-14) is increasing on  $\theta \in (\frac{2\pi}{3}, \pi)$ .*

*Proof.* Lemma 13 gives

$$-z = \frac{1 + t_0 + at_0^2}{t_0^3} = \frac{1 + t_1 + at_1^2}{t_1^3}.$$

We differentiate the three terms and obtain

$$dz = \frac{3 + 2t_0 + at_0^2}{t_0^4} dt_0 = \frac{3 + 2t_1 + at_1^2}{t_1^4} dt_1, \quad (2-19)$$

where

$$dt_1 = d(t_0 e^{2i\theta}) = e^{2i\theta} dt_0 + 2it_0 e^{2i\theta} d\theta.$$



If we set

$$f(t_0) = \frac{3 + 2t_0 + at_0^2}{t_0^4}, \quad f(t_1) = \frac{3 + 2t_1 + at_1^2}{t_1^4},$$

then  $f(t_0) = \overline{f(t_1)}$ , and consequently  $f(t_0)f(t_1) \geq 0$ . Thus (2-19) implies

$$f(t_0)dt_0 = f(t_1)(e^{2i\theta}dt_0 + 2it_0e^{2i\theta}d\theta).$$

After solving this equation for  $dt_0$  and substituting it into (2-19), we obtain

$$\frac{dz}{d\theta} = \frac{2if(t_0)f(t_1)t_0e^{2i\theta}}{f(t_0) - f(t_1)e^{2i\theta}}. \quad (2-20)$$

With  $t_0 = \tau e^{-i\theta}$ ,  $\tau \in \mathbb{R}$ , we have

$$\frac{f(t_0) - f(t_1)e^{2i\theta}}{2it_0e^{2i\theta}} = \frac{f(t_0)e^{-i\theta} - f(t_1)e^{i\theta}}{2it_0e^{i\theta}} = \frac{\Im(f(t_0)e^{-i\theta})}{\tau} = \frac{1}{\tau} \Im\left(\frac{3 + 2t_0 + at_0^2}{t_0^4} e^{-i\theta}\right).$$

We now substitute  $3 = -3t_0 - 3at_0^2 - 3zt_0^3$  and have

$$\begin{aligned} \frac{f(t_0) - f(t_1)e^{2i\theta}}{2it_0e^{2i\theta}} &= \frac{1}{\tau} \Im\left(\frac{-t_0 - 2at_0^2 - 3zt_0^3}{t_0^4} e^{-i\theta}\right) = \frac{1}{\tau^4} \Im(-e^{2i\theta} - 2a\tau e^{i\theta} - 3z\tau^2) \\ &= \frac{1}{\tau^4} (-\sin 2\theta - 2a\tau \sin \theta) = \frac{2 \sin \theta}{\tau^4} (-\cos \theta - a\tau). \end{aligned}$$

In the formula for  $t_0$  in Lemma 13, we substitute  $\tau = -1/\zeta - 2 \cos \theta$  and obtain

$$\frac{f(t_0) - f(t_1)e^{2i\theta}}{2it_0e^{2i\theta}} = \frac{2 \sin \theta}{\tau^4} (-\cos \theta + a/\zeta + 2a \cos \theta). \quad (2-21)$$

We finish this lemma by showing that  $-\cos \theta + a/\zeta + 2a \cos \theta > 0$ . This strict inequality implies that we cannot have  $f(t_0) = f(t_1) = 0$  by (2-21), and the lemma follows from (2-20). To prove the inequality, we expand and divide both sides of (2-10) by  $\zeta$  to get

$$\zeta(1 - 4a \cos^2 \theta) + 2 \cos \theta(1 - 2a) - a/\zeta = 0,$$

or equivalently,

$$\zeta(1 - 4a \cos^2 \theta) + \cos \theta(1 - 2a) = -\cos \theta + 2a \cos \theta + a/\zeta.$$

Finally, using the definition of  $\zeta$  in (2-12) and Lemma 6, we calculate

$$\zeta(1 - 4a \cos^2 \theta) + \cos \theta(1 - 2a) = \sqrt{(1 - 4a) \cos^2 \theta + a} > 0. \quad \square$$

**Lemma 15.** *The function  $z(\theta)$  as defined in (2-14) maps the interval  $(\frac{2\pi}{3}, \pi)$  onto the interior of  $I_a$ .*

*Proof.* Since  $z(\theta)$  is a continuous increasing function on  $(\frac{2\pi}{3}, \pi)$ , we only need to evaluate the limits of  $z(\theta)$  at the endpoints. Since  $|\zeta| > 1$  by Lemma 9, the formula of  $\zeta(\theta)$  in (2-12) implies  $\zeta(\theta) \rightarrow 1^+$  as  $\theta \rightarrow (\frac{2\pi}{3})^+$ . Consequently, (2-18) gives

$$\lim_{\theta \rightarrow (2\pi/3)^+} z(\theta) = -\infty.$$

Finally, the fact that

$$\lim_{\theta \rightarrow \pi} \zeta(\theta) = \frac{1 - 2a + \sqrt{1 - 3a}}{1 - 4a},$$

together with (2-14), implies

$$\begin{aligned} \lim_{\theta \rightarrow \pi} z(\theta) &= \frac{a(1 - 2a + \sqrt{1 - 3a})(1 - 4a)}{(-1 + 6a + \sqrt{1 - 3a})(-1 - 2\sqrt{1 - 3a})} \\ &= \frac{a(-1 + 4a)^2(-2 + 9a) + 2a(-1 + 3a)(-1 + 4a)^2\sqrt{1 - 3a}}{27(1 - 4a)^2a} \\ &= \frac{-2 + 9a - 2\sqrt{(1 - 3a)^3}}{27}, \end{aligned} \tag{2-22}$$

where we obtain (2-22) after multiplication and division by  $(-1 + 6a - \sqrt{1 - 3a})(-1 + 2\sqrt{1 - 3a})$ . □

Before making the final arguments connecting the results of this section, we check the sign of  $g_m(\theta)$  at one of the endpoints.

**Lemma 16.** *If  $-1 \leq a < \frac{1}{4}$ , then the sign of  $g_m(\pi^-)$  is  $(-1)^m$ .*

*Proof.* Since  $-1 \leq a < \frac{1}{4}$ , one can check that

$$\lim_{\theta \rightarrow \pi^-} \zeta(\theta) = \frac{1 - 2a + \sqrt{1 - 3a}}{1 - 4a} \geq 1.$$

The result follows directly from (2-15) and the fact that

$$\lim_{\theta \rightarrow \pi^-} \frac{\sin((m + 1)\theta)}{\sin(\theta)} = (m + 1)(-1)^m. \tag{□}$$

With all the lemmas at our disposal, we produce the final arguments to finish the proof of Theorem 3. We consider the function  $g_m(\theta)$  at the points

$$\theta_h = \frac{h\pi}{m + 1} \in \left(\frac{2\pi}{3}, \pi\right), \quad \lfloor \frac{2}{3}(m + 1) \rfloor + 1 \leq h \leq m.$$

We note that the number of such values of  $h$  is

$$m - \lfloor \frac{2}{3}(m + 1) \rfloor = \lfloor \frac{1}{3}m \rfloor,$$

where the equality can be checked by considering the residue classes of  $m$  modulo 3. From the formula of  $g_m(\theta)$  in (2-15) and Lemma 9, the sign of  $g_m(\theta_h)$  is  $(-1)^{h-1}$ .

By the intermediate value theorem and Lemma 12, there are at least  $\lfloor \frac{1}{3}m \rfloor - 1$  zeros of  $g_m(\theta)$  on  $(\frac{2\pi}{3}, \pi)$ . In fact, we claim that there are at least  $\lfloor \frac{1}{3}m \rfloor$  zeros of  $g_m(\theta)$  on  $(\frac{2\pi}{3}, \pi)$ . In the case  $-1 \leq a < \frac{1}{4}$ , we obtain one more zero of  $g_m(\theta)$  from Lemma 16. On the other hand, if  $\frac{1}{4} < a \leq \frac{1}{3}$ , then we obtain another zero of  $g_m(\theta)$  by Lemma 12. Using Lemmas 14 and 15, we obtain at least  $\lfloor \frac{1}{3}m \rfloor$  zeros of  $H_m(z)$  on  $I_a$ . Since the degree of  $H_m(z)$  is at most  $\lfloor \frac{1}{3}m \rfloor$  by Lemma 5, all the zeros of  $H_m(z)$  lie in  $I_a$ . Recall that we can ignore the case  $a = \frac{1}{4}$  by Lemma 4. The density of  $\bigcup_{m=0}^{\infty} \mathcal{Z}(H_m(z))$  in  $I_a$  comes from the density of  $\bigcup_{m=0}^{\infty} \mathcal{Z}(g_m(\theta))$  in  $(\frac{2\pi}{3}, \pi)$  and from  $z(\theta)$  being a continuous map.

### 3. The case $c = 0$ and $b \geq 0$

It is trivial that if  $c = 0$  and  $b = 0$ , then the zeros of  $H_m(z)$  are real under the convention that the constant zero polynomial is hyperbolic. When  $b > 0$ , we make the substitution  $t \rightarrow t/\sqrt{b}$  and reformulate the claim as follows.

**Theorem 17.** *The zeros of the sequence of polynomials  $\{H_m(z)\}_{m=0}^{\infty}$  generated by*

$$\sum_{m=0}^{\infty} H_m(z)t^m = \frac{1}{1+t^2+zt^3} \tag{3-1}$$

*are real, and the set  $\bigcup_{m=0}^{\infty} \mathcal{Z}(H_m)$  is dense in  $(-\infty, \infty)$ .*

The proof of Theorem 17 follows from a similar procedure as that seen in Section 2. We will point out some key differences. The following lemma comes directly from the recurrence relation

$$H_m(z) + H_{m-2}(z) + zH_{m-3}(z) = 0$$

and induction.

**Lemma 18.** *The degree of the polynomial  $H_m(z)$  generated by (3-1) is at most*

$$\begin{cases} \frac{1}{3}m & \text{if } m \equiv 0 \pmod{3}, \\ \frac{1}{3}(m-4) & \text{if } m \equiv 1 \pmod{3}, \\ \frac{1}{3}(m-2) & \text{if } m \equiv 2 \pmod{3}. \end{cases}$$

We define the following three functions on the interval  $(\frac{\pi}{3}, \frac{\pi}{2})$ :

$$\begin{aligned} \zeta(\theta) &= -\frac{1}{2 \cos \theta}, \\ g_m(\theta) &= \frac{-\sin((m+1)\theta)}{2 \cos \theta \sin \theta} (2 + \cos 2\theta) - \cos((m+1)\theta) + (-2 \cos \theta)^{m+1}, \\ z(\theta) &= \frac{2 \cos \theta}{\sqrt{(1-4 \cos^2 \theta)^3}}. \end{aligned} \tag{3-2}$$

The proof of the lemma below is similar to that of Lemma 13. We leave the detailed computations to the reader.

**Lemma 19.** *Suppose  $\theta \in (\frac{\pi}{3}, \frac{\pi}{2})$ ,  $\zeta = \zeta(\theta)$ , and  $z = z(\theta)$  are defined by (3-2). The three zeros of  $1 + t^2 + z(\theta)t^3$  are*

$$t_0 = -\frac{e^{-i\theta}}{z(2 \cos \theta + \zeta)}, \quad t_1 = t_0 e^{2i\theta}, \quad t_2/t_0 = \zeta e^{i\theta}.$$

Looking at  $z'(\theta)$ , one can check that  $z(\theta)$  is strictly decreasing on the interval  $(\frac{\pi}{3}, \frac{\pi}{2})$ . Using the partial fraction decomposition

$$\sum_{m=0}^{\infty} H_m(z)t^m = \frac{1}{1 + t^2 + zt^3} = \frac{1}{z(t - t_0)(t - t_1)(t - t_2)},$$

we conclude that for each zero of  $g_m(\theta)$  on the interval  $(\frac{\pi}{3}, \frac{\pi}{2})$  we obtain two zeros  $\pm z(\theta)$  of  $H_m(z)$ . We can also check by induction that  $z = 0$  is a simple zero of  $H_m(z)$  if  $m$  is odd, and  $z = 0$  is not a zero of  $H_m(z)$  when  $m$  is even. The formula of  $g_m(\theta)$  implies that the sign of this function alternates when  $\cos((m + 1)\theta) = \pm 1$ , that is, when,

$$(m + 1)\theta = k\pi, \quad \frac{1}{3}(m + 1) < k < \frac{1}{2}(m + 1).$$

Since  $g_m(\theta)$  is continuous on  $(\frac{\pi}{3}, \frac{\pi}{2})$ , we may apply the intermediate value theorem to compute the number of zeros of  $g_m(\theta)$  on  $(\frac{\pi}{3}, \frac{\pi}{2})$  and the corresponding number of zeros of  $H_m(z)$  on  $(-\infty, \infty)$ . We will see that this number is equal to the degree of  $H_m(z)$ , thereby proving Theorem 17. We summarize the six arising cases, where  $\theta^*$  denotes the smallest solution  $(m + 1)\theta = k\pi$  on the interval  $(\frac{\pi}{3}, \frac{\pi}{2})$ .

Case 1:  $m \equiv 1 \pmod{3}$  and  $m$  is even. There are

$$\frac{1}{2}m - \frac{1}{3}(m + 2) = \frac{1}{6}(m - 4)$$

zeros of  $g_m(\theta)$  on  $(\frac{\pi}{3}, \frac{\pi}{2})$ , which give  $\frac{1}{3}(m - 4)$  zeros of  $H_m(z)$  on  $(-\infty, \infty)$ .

Case 2:  $m \equiv 1 \pmod{3}$  and  $m$  is odd. There are

$$\frac{1}{2}(m - 1) - \frac{1}{3}(m + 2) = \frac{1}{6}(m - 7)$$

zeros of  $g_m(\theta)$  on  $(\frac{\pi}{3}, \frac{\pi}{2})$ , which give  $\frac{1}{3}(m - 7)$  nonzero zeros of  $H_m(z)$ . We add a simple zero  $z = 0$  and obtain  $\frac{1}{3}(m - 4)$  zeros of  $H_m(z)$  on  $(-\infty, \infty)$ .

Case 3:  $m \equiv 0 \pmod{3}$  and  $m$  is even. With the observation that  $\lim_{\theta \rightarrow \pi/3} g_m(\theta) = -3 < 0$  and  $g_m(\theta^*) > 0$ , we obtain

$$\frac{1}{2}m - (\frac{1}{3}m + 1) + 1 = \frac{1}{6}m$$

zeros of  $g_m(\theta)$  on  $(\frac{\pi}{3}, \frac{\pi}{2})$ , which give  $\frac{1}{3}m$  zeros of  $H_m(z)$  on  $(-\infty, \infty)$ .

Case 4:  $m \equiv 0 \pmod{3}$  and  $m$  is odd. With the observation that  $\lim_{\theta \rightarrow \pi/3} g_m(\theta) = 3 > 0$  and  $g_m(\theta^*) < 0$ , we obtain

$$\frac{1}{2}(m-1) - \left(\frac{1}{3}(m)+1\right) + 1 = \frac{1}{6}(m-3)$$

zeros of  $g_m(\theta)$  on  $(\frac{\pi}{3}, \frac{\pi}{2})$ , which give  $\frac{1}{3}(m-3)$  nonzero zeros of  $H_m(z)$ . We add a simple zero  $z=0$  and obtain  $\frac{1}{3}m$  zeros of  $H_m(z)$  on  $(-\infty, \infty)$ .

Case 5:  $m \equiv 2 \pmod{3}$  and  $m$  is even. With the observation that  $g_m(\frac{\pi}{3}) = 0$ ,  $g'_m(\frac{\pi}{3}) > 0$ , and  $g_m(\theta^*) < 0$ , we obtain

$$\frac{1}{2}m - \left(\frac{1}{3}(m+1)+1\right) + 1 = \frac{1}{6}(m-2)$$

zeros of  $g_m(\theta)$  on  $(\frac{\pi}{3}, \frac{\pi}{2})$ , which give  $\frac{1}{3}(m-2)$  zeros of  $H_m(z)$  on  $(-\infty, \infty)$ .

Case 6:  $m \equiv 2 \pmod{3}$  and  $m$  is odd. With the observation that  $g_m(\frac{\pi}{3}) = 0$ ,  $g'_m(\frac{\pi}{3}) < 0$ , and  $g_m(\theta^*) > 0$ , we obtain

$$\frac{1}{2}(m-1) - \left(\frac{1}{3}(m+1)+1\right) + 1 = \frac{1}{6}(m-5)$$

zeros of  $g_m(\theta)$  on  $(\frac{\pi}{3}, \frac{\pi}{2})$ , which give  $\frac{1}{3}(m-5)$  nonzero roots of  $H_m(z)$ . We add a simple zero  $z=0$  and obtain  $\frac{1}{3}(m-2)$  zeros of  $H_m(z)$  on  $(-\infty, \infty)$ .

In all cases above the number of zeros of  $H_m(z)$  on  $(-\infty, \infty)$  corresponds to the degree of  $H_m(z)$  and Theorem 17 follows.

#### 4. Necessary condition for the reality of zeros

To prove the necessary condition of Theorem 2, we first show that if  $c=0$  and  $b < 0$  then not all polynomials  $H_m(z)$  are hyperbolic. In fact, with the substitution  $t \rightarrow it$ , we conclude that all the zeros of  $H_m(z)$  will be purely imaginary by Theorem 17.

It remains to consider the sequence  $H_m(z)_{m=0}^{\infty}$  generated by

$$\sum_{m=0}^{\infty} H_m(z)t^m = \frac{1}{1+t+at^2+zt^3},$$

and to show that if  $a \notin [-1, \frac{1}{3}]$  then there is an  $m$  such that not all the zeros of  $H_m(z)$  are real. In fact, we will show if  $a \notin [-1, \frac{1}{3}]$ , then  $H_m(z)$  is not hyperbolic for all large  $m$ . To prove this, let us introduce some definitions, discussed in [Sokal 2004], related to the root distribution of a sequence of functions

$$f_m(z) = \sum_{k=1}^n \alpha_k(z)\beta_k(z)^m,$$

where  $\alpha_k(z)$  and  $\beta_k(z)$  are analytic in a domain  $D$ . We say that an index  $k$  is dominant at  $z$  if  $|\beta_k(z)| \geq |\beta_l(z)|$  for all  $l$  ( $1 \leq l \leq n$ ). Let

$$D_k = \{z \in D : k \text{ is dominant at } z\}.$$

Let  $\liminf \mathcal{Z}(f_m)$  be the set of all  $z \in D$  such that every neighborhood  $U$  of  $z$  has a nonempty intersection with all but finitely many of the sets  $\mathcal{Z}(f_m)$ . Let  $\limsup \mathcal{Z}(f_m)$  be the set of all  $z \in D$  such that every neighborhood  $U$  of  $z$  has a nonempty intersection with all but infinitely many of the sets  $\mathcal{Z}(f_m)$ . We will need the following theorem from Sokal.

**Theorem 20** [Sokal 2004, Theorem 1.5]. *Let  $D$  be a domain in  $\mathbb{C}$ , and let  $\alpha_1, \dots, \alpha_n, \beta_1, \dots, \beta_n$  ( $n \geq 2$ ) be analytic functions on  $D$ , none of which is identically zero. Let us further assume a “no-degenerate-dominance” condition: there do not exist indices  $k \neq k'$  such that  $\beta_k \equiv \omega\beta_{k'}$  for some constant  $\omega$  with  $|\omega| = 1$  and such that  $D_k (= D_{k'})$  has nonempty interior. For each integer  $m \geq 0$ , define  $f_m$  by*

$$f_m(z) = \sum_{k=1}^n \alpha_k(z)\beta_k(z)^m.$$

*Then  $\liminf \mathcal{Z}(f_m) = \limsup \mathcal{Z}(f_m)$ , and a point  $z$  lies in this set if and only if either*

- (i) *there is a unique dominant index  $k$  at  $z$ , and  $\alpha_k(z) = 0$ , or*
- (ii) *there are two or more dominant indices at  $z$ .*

If  $z^* \in \mathbb{C}$  such that the zeros in  $t$  of  $1+t+at^2+z^*t^3$  are distinct then by the partial fractions given in (2-4) and Theorem 20,  $z^*$  will belong to  $\liminf \mathcal{Z}(H_m)$  when the two smallest (in modulus) zeros of  $1+t+at^2+z^*t^3$  have the same modulus. We also note that  $t_0(z)$ ,  $t_1(z)$ , and  $t_2(z)$  are analytic in a neighborhood of  $z^*$  by the implicit function theorem. If we let  $\omega = e^{2i\theta}$ , then the no-degenerate-dominance condition in Theorem 20 comes directly from equations (2-14) and (2-12) since  $\theta$  is a fixed constant (and thus  $z$  is a fixed point which has empty interior).

Suppose  $a \notin [-1, \frac{1}{3}]$ . With the setup in the previous paragraph, our main goal is to find a  $z^* \notin \mathbb{R}$  so that the zeros of  $1+t+at^2+z^*t^3$  are distinct and the two smallest (in modulus) zeros of this polynomial have the same modulus. If we can find such a point, then  $z^* \in \liminf \mathcal{Z}(H_m) = \limsup \mathcal{Z}(H_m)$ . This implies that on a small neighborhood of  $z^*$  which does not intersect the real line, there is a nonreal zero of  $H_m(z)$  for all large  $m$  by the definition of  $\liminf \mathcal{Z}(H_m)$ . Our choice of  $z^* = z(\theta^*)$  comes from (2-14) for a special  $\theta^*$ . Unlike in Section 2,  $\theta^*$  will not belong to  $(\frac{2\pi}{3}, \pi)$  to ensure that  $z^* \notin \mathbb{R}$ . In particular, we consider the two cases  $a < -1$  and  $a > 3$ .

**The case  $a < -1$ .** We select  $0 \ll \theta^* < \frac{\pi}{2}$ . Since

$$\lim_{\theta \rightarrow \frac{\pi}{2}} \zeta(\theta) = i\sqrt{|a|},$$

see (2-12), we can pick  $0 < \theta^* < \frac{\pi}{2}$  sufficiently close to  $\frac{\pi}{2}$  so that  $\zeta := \zeta(\theta^*) \in \mathbb{C} \setminus \mathbb{R}$  and  $|\zeta(\theta^*)| > 1$ . By Lemma 13, we have  $t_2 = \zeta t_0 e^{i\theta^*}$  and  $t_1 = t_0 e^{2i\theta^*}$ . The fact that

$|\zeta| > 1$  and  $\theta^* \neq 0, \frac{\pi}{2}$  implies that the polynomial  $1 + t + at^2 + z(\theta^*)t^3$  has distinct zeros and not all its zeros are real. We will show that  $z(\theta^*) \notin \mathbb{R}$ , by contradiction. Indeed, if  $z(\theta^*) \in \mathbb{R}$  then the zeros of the polynomial  $1 + t + at^2 + z(\theta^*)t^3 \in \mathbb{R}[t]$  satisfy  $t_0 = \bar{t}_1$  and

$$t_2 = t_0 \zeta e^{i\theta^*} \in \mathbb{R}.$$

This gives a contradiction because the first equation implies  $t_0 e^{i\theta^*} \in \mathbb{R}$ , while the second equation implies  $t_0 e^{i\theta^*} \notin \mathbb{R}$  since  $\zeta \notin \mathbb{R}$ .

**The case  $a > \frac{1}{3}$ .** We select  $\beta < \cos \theta^* \ll 1$ , where  $\beta = \sqrt{a/(4a-1)} < 1$ . Once more,

$$\left| \lim_{\cos \theta \rightarrow \beta} \zeta(\theta) \right| = \begin{cases} \left| \frac{\sqrt{4a^2 - a}}{1 - 2a} \right| & \text{if } a \neq \frac{1}{2}, \\ \infty & \text{if } a = \frac{1}{2}, \end{cases}$$

where we can easily check that

$$\left| \frac{\sqrt{4a^2 - a}}{1 - 2a} \right| > 1, \quad a > \frac{1}{3}.$$

Thus if  $\cos \theta^*$  is sufficiently close to  $\beta$ , then  $0 < \theta^* < \frac{\pi}{2}$ ,  $|\zeta(\theta^*)| > 1$ , and  $|\zeta(\theta^*)| \notin \mathbb{R}$ , where the last statement comes from (2-12) and the inequality

$$(1 - 4a) \cos^2 \theta^* + a < 0.$$

With  $0 < \theta^* < \frac{\pi}{2}$ ,  $|\zeta(\theta^*)| > 1$ , and  $|\zeta(\theta^*)| \notin \mathbb{R}$ , we apply the same arguments given in the previous case to complete the proof.

### Acknowledgment

The authors would like to thank Professor T. Forgács for his careful review of the paper and the reviewer for their helpful comments on the paper.

### References

- [Bates and Yoshida 2016] R. Bates and R. Yoshida, "Quadratic hyperbolicity preservers and multiplier sequences", *Rocky Mountain J. Math.* **46**:1 (2016), 51–72. MR
- [Beraha et al. 1975] S. Beraha, J. Kahane, and N. J. Weiss, "Limits of zeroes of recursively defined polynomials", *Proc. Nat. Acad. Sci. U.S.A.* **72**:11 (1975), 4209. MR Zbl
- [Beraha et al. 1978] S. Beraha, J. Kahane, and N. J. Weiss, "Limits of zeros of recursively defined families of polynomials", pp. 213–232 in *Studies in foundations and combinatorics*, edited by G.-C. Rota, Adv. in Math. Suppl. Stud. **1**, Academic Press, New York, 1978. MR Zbl
- [Borcea and Brändén 2009] J. Borcea and P. Brändén, "Pólya–Schur master theorems for circular domains and their boundaries", *Ann. of Math.* (2) **170**:1 (2009), 465–492. MR Zbl
- [Borcea et al. 2006] J. Borcea, R. Bøgvad, and B. Shapiro, "On rational approximation of algebraic functions", *Adv. Math.* **204**:2 (2006), 448–480. MR Zbl

- [Boyer and Goh 2007] R. Boyer and W. M. Y. Goh, “On the zero attractor of the Euler polynomials”, *Adv. in Appl. Math.* **38**:1 (2007), 97–132. MR Zbl
- [Boyer and Goh 2008] R. Boyer and W. M. Y. Goh, “Polynomials associated with partitions: asymptotics and zeros”, pp. 33–45 in *Special functions and orthogonal polynomials*, edited by D. Dominici and R. S. Maier, Contemp. Math. **471**, Amer. Math. Soc., Providence, RI, 2008. MR Zbl
- [Bunton et al. 2015] A. Bunton, N. Jacobs, S. Jenkins, C. McKenry, Jr., A. Piotrowski, and L. Scott, “Nonreal zero decreasing operators related to orthogonal polynomials”, *Involve* **8**:1 (2015), 129–146. MR Zbl
- [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 Zbl
- [Eğecioğlu et al. 2001] O. Eğecioğlu, T. Redmond, and C. Ryavec, “From a polynomial Riemann hypothesis to alternating sign matrices”, *Electron. J. Combin.* **8**:1 (2001), art. id. 36. MR Zbl
- [Forgács and Tran 2016] T. Forgács and K. Tran, “Polynomials with rational generating functions and real zeros”, *J. Math. Anal. Appl.* **443**:2 (2016), 631–651. MR Zbl
- [Sokal 2004] A. D. Sokal, “Chromatic roots are dense in the whole complex plane”, *Combin. Probab. Comput.* **13**:2 (2004), 221–261. MR Zbl
- [Tran 2014] K. Tran, “Connections between discriminants and the root distribution of polynomials with rational generating function”, *J. Math. Anal. Appl.* **410**:1 (2014), 330–340. MR Zbl
- [Tran 2015] K. Tran, “The root distribution of polynomials with a three-term recurrence”, *J. Math. Anal. Appl.* **421**:1 (2015), 878–892. MR Zbl

Received: 2017-03-14    Revised: 2017-06-07    Accepted: 2017-06-19

khangt@csufresno.edu

*Department of Mathematics, California State University,  
Fresno, CA, United States*

andreszumba@mail.fresnostate.edu

*Department of Mathematics, California State University,  
Fresno, CA, United States*



# Binary frames with prescribed dot products and frame operator

Veronika Furst and Eric P. Smith

(Communicated by David Royal Larson)

This paper extends three results from classical finite frame theory over real or complex numbers to binary frames for the vector space  $\mathbb{Z}_2^d$ . Without the notion of inner products or order, we provide an analog of the “fundamental inequality” of tight frames. In addition, we prove the binary analog of the characterization of dual frames with given inner products and of general frames with prescribed norms and frame operator.

## 1. Introduction

Due to applications in signal and image processing, data compression, sampling theory, and other problems in engineering and computer science, frames in finite-dimensional spaces have received much attention from pure and applied mathematicians alike, over the past thirty years; see, for example, Chapter 1 of [Casazza and Kutyniok 2013]. The redundant representation of vectors inherent to frame theory is central to the idea of efficient data storage and transmission that is robust to noise and erasures.

Frames for  $\mathbb{C}^d$  and  $\mathbb{R}^d$  have been extensively studied; see [Christensen 2003] for a standard introduction to frame theory, [Kovačević and Chebira 2007] for applications, and [Han et al. 2007] for an exposition at the undergraduate level. Noting the similarity between frames and error-correcting codes, Bodmann, Le, Reza, Tobin, and Tomforde [Bodmann et al. 2009] introduced the concept of *binary frames*, that is, finite frames for the vector space  $\mathbb{Z}_2^d$ . Binary Parseval frames robust to erasures were characterized in [Bodmann et al. 2014], and their Gramian matrices were studied in [Baker et al. 2018]. A more generalized approach to binary frames was taken in [Hotovy et al. 2015].

We begin with a brief introduction to classical frame theory terminology. Let  $\mathbb{H}^d$  denote the Hilbert space  $\mathbb{C}^d$  or  $\mathbb{R}^d$ .

---

*MSC2010:* 15A03, 15A23, 15B33, 42C15.

*Keywords:* frames, binary vector spaces, frame operators, Gramian matrices.

**Definition 1.1.** A (finite) *frame* for a Hilbert space  $\mathbb{H}^d$  is a collection  $\{x_j\}_{j=1}^K$  of vectors in  $\mathbb{H}^d$  for which there exist finite constants  $A, B > 0$  such that for every  $y \in \mathbb{H}^d$ ,

$$A\|y\|^2 \leq \sum_{j=1}^K |\langle y, x_j \rangle|^2 \leq B\|y\|^2.$$

The constants  $A$  and  $B$  are known as *frame bounds*. An *A-tight frame* is one for which  $A = B$ , and a *Parseval frame* is one for which  $A = B = 1$ .

The vectors  $x_j$  in the above definition need not be orthogonal or even linearly independent. An orthonormal basis is most closely resembled by a Parseval frame, for which we have the (not necessarily unique) reconstruction formula:

**Proposition 1.2.** A collection of vectors  $\{x_j\}_{j=1}^K$  in a finite-dimensional Hilbert space  $\mathbb{H}^d$  is a Parseval frame for  $\mathbb{H}^d$  if and only if

$$y = \sum_{j=1}^K \langle y, x_j \rangle x_j$$

for each  $y \in \mathbb{H}^d$ .

**Definition 1.3.** Let  $\{x_j\}_{j=1}^K$  be a frame for the finite-dimensional Hilbert space  $\mathbb{H}^d$ . The corresponding *frame operator*  $S : \mathbb{H}^d \rightarrow \mathbb{H}^d$  is defined by

$$S(x) = \sum_{j=1}^K \langle x, x_j \rangle x_j.$$

It can be seen as the composition  $S = \Theta\Theta^*$  of the *synthesis operator*  $\Theta : \mathbb{C}^K \rightarrow \mathbb{H}^d$  and its adjoint, the *analysis operator*  $\Theta^* : \mathbb{H}^d \rightarrow \mathbb{C}^K$ , given by the formulas

$$\Theta \left( \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_K \end{bmatrix} \right) = \sum_{j=1}^K c_j x_j \quad \text{and} \quad \Theta^*(x) = \begin{bmatrix} \langle x, x_1 \rangle \\ \langle x, x_2 \rangle \\ \vdots \\ \langle x, x_K \rangle \end{bmatrix}.$$

The frame operator is a bounded, invertible, self-adjoint operator satisfying  $AI_d \leq S \leq BI_d$ . Here and in what follows, we use  $I_d$  to denote the  $d \times d$  identity matrix and  $0_d$  to denote the  $d \times d$  zero matrix. A frame is Parseval if and only if its frame operator is the identity operator.

From both a pure and an applied point-of-view, construction of frames with desired properties has been a central question [Bownik and Jasper 2015]. In particular, much attention has been paid to tight frames with prescribed norms and general frames with both prescribed norms and frame operator. In the case of tight frames, the answer, the so-called “fundamental frame inequality”, was provided by Casazza, Fickus, Kovačević, Leon, and Tremain:

**Theorem 1.4** [Casazza et al. 2006, Corollary 4.11]. *Given real numbers  $a_1 \geq a_2 \geq \dots \geq a_K > 0$ ,  $K \geq d$ , there exists a  $\lambda$ -tight frame  $\{x_j\}_{j=1}^K$  for a  $d$ -dimensional Hilbert space  $\mathbb{H}^d$  with prescribed norms  $\|x_j\|^2 = a_j$  for  $1 \leq j \leq K$  if and only if*

$$\lambda = \frac{1}{d} \sum_{j=1}^K a_j \geq a_1.$$

Casazza and Leon generalized this result to frames with prescribed frame operators (the classical case is when  $S = \lambda I_d$ ):

**Theorem 1.5** [Casazza and Leon 2010, Theorem 2.1]. *Let  $S$  be a positive self-adjoint operator on a  $d$ -dimensional Hilbert space  $\mathbb{H}^d$  with eigenvalues  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d > 0$ . Given real numbers  $a_1 \geq a_2 \geq \dots \geq a_K > 0$ ,  $K \geq d$ , there is a frame  $\{x_j\}_{j=1}^K$  for  $\mathbb{H}^d$  with frame operator  $S$  and  $\|x_j\|^2 = a_j$  for all  $1 \leq j \leq K$  if and only if*

$$\sum_{j=1}^K a_j = \sum_{j=1}^d \lambda_j \quad \text{and} \quad \sum_{j=1}^k a_j \leq \sum_{j=1}^k \lambda_j$$

for every  $1 \leq k \leq d$ .

This can be seen as a consequence of the classical Schur–Horn theorem [Bownik and Jasper 2015]. Cahill, Fickus, Mixon, Poteet, and Strawn [Cahill et al. 2013] introduced a so-called *eigenstep* method for constructing all frames with a given frame operator and set of norms; see also [Fickus et al. 2013], and [Bownik and Jasper 2015] for a survey of the topic.

A different approach was taken by Christensen, Powell, and Xiao [Christensen et al. 2012], extending Theorem 1.4 to the setting of dual frame pairs. Given a frame  $\{x_j\}_{j=1}^K$ , a sequence  $\{y_j\}_{j=1}^K$  is called a *dual frame* if, for every  $y \in \mathbb{H}^d$ ,

$$y = \sum_{j=1}^K \langle y, x_j \rangle y_j = \sum_{j=1}^K \langle y, y_j \rangle x_j.$$

**Theorem 1.6** [Christensen et al. 2012, Theorem 3.1]. *Given a sequence of numbers  $\{a_j\}_{j=1}^K \subset \mathbb{H}$  with  $K > d$ , the following are equivalent:*

- (1) *There exist dual frames  $\{x_j\}_{j=1}^K$  and  $\{y_j\}_{j=1}^K$  for  $\mathbb{H}^d$  such that  $\langle x_j, y_j \rangle = a_j$  for all  $1 \leq j \leq K$ .*
- (2) *There exists a tight frame  $\{x_j\}_{j=1}^K$  and dual frame  $\{y_j\}_{j=1}^K$  for  $\mathbb{H}^d$  such that  $\langle x_j, y_j \rangle = a_j$  for all  $1 \leq j \leq K$ .*
- (3)  $\sum_{j=1}^K a_j = d$ .

The goal of this paper is to extend the theory of frames with prescribed norms (or inner products) from the classical Hilbert spaces of  $\mathbb{C}^d$  and  $\mathbb{R}^d$  to the binary space  $\mathbb{Z}_2^d$ . We provide analogs of Theorems 1.4, 1.5, and 1.6 for binary frames.

The challenge, of course, is the lack of an inner product, positive elements, and guaranteed eigenvalues. Section 2 contains background material on binary frames. In Section 3, we explore dual binary frames and prove the binary versions of Theorems 1.6 and 1.4. In Section 4, we construct binary frames with prescribed “norms” and frame operator, as an analog to Theorem 1.5. We conclude in Section 5 with examples and a catalog.

## 2. Binary frames

Bodmann, Le, Reza, Tobin, and Tomforde [Bodmann et al. 2009] introduced a theory of frames over the  $d$ -dimensional binary space  $\mathbb{Z}_2^d$ , where  $\mathbb{Z}_2^d$  is the direct product  $\mathbb{Z}_2 \oplus \cdots \oplus \mathbb{Z}_2$  having  $d \geq 1$  copies of  $\mathbb{Z}_2$ . The main trouble in defining frames in a binary space stems from the lack of an ordering on  $\mathbb{Z}_2$ . Without an order, there can be no inner product defined for binary space. In spite of this, [Bodmann et al. 2009] establishes the dot product as the analog of the inner product on  $\mathbb{R}^d$  and  $\mathbb{C}^d$ .

**Definition 2.1** [Bodmann et al. 2009]. The dot product on  $\mathbb{Z}_2^d$  is defined as the map  $(\cdot, \cdot) : \mathbb{Z}_2^d \times \mathbb{Z}_2^d \rightarrow \mathbb{Z}_2$  given by

$$(a, b) = \sum_{n=1}^d a[n]b[n].$$

Due to the degenerate nature of the dot product (note that  $(a, a) = 0$  need not imply  $a = 0$ ), it fails to help define a frame in the manner of Definition 1.1. However, when working over finite-dimensional spaces in the classical case, a frame is merely a spanning sequence of vectors. This motivates the definition of a frame in binary space.

**Definition 2.2** [Bodmann et al. 2009]. A frame is a sequence of vectors  $\mathcal{F} = \{f_j\}_{j=1}^K$  in  $\mathbb{Z}_2^d$  such that  $\text{Span}(\mathcal{F}) = \mathbb{Z}_2^d$ .

The synthesis, analysis, and frame operators of  $\mathcal{F}$  are defined similarly to Definition 1.3 and are denoted by  $\Theta_{\mathcal{F}}$ ,  $\Theta_{\mathcal{F}}^*$ , and  $S_{\mathcal{F}}$ , respectively.

**Definition 2.3** [Bodmann et al. 2009]. The synthesis operator of a frame  $\mathcal{F} = \{f_j\}_{j=1}^K$  is the  $d \times K$  matrix whose  $i$ -th column is the  $i$ -th vector in  $\mathcal{F}$ . The analysis operator  $\Theta_{\mathcal{F}}^*$  is the transpose of the synthesis operator. Explicitly,

$$\Theta_{\mathcal{F}} = \begin{bmatrix} | & & | \\ f_1 & \cdots & f_K \\ | & & | \end{bmatrix} \quad \text{and} \quad \Theta_{\mathcal{F}}^* = \begin{bmatrix} - & f_1^* & - \\ \vdots & & \\ - & f_K^* & - \end{bmatrix}.$$

The frame operator is  $S_{\mathcal{F}} = \Theta_{\mathcal{F}} \Theta_{\mathcal{F}}^*$ .

It is demonstrated in [Bodmann et al. 2009] that the spanning property of  $\mathcal{F}$  is necessary and sufficient for  $\mathcal{F}$  to have a reconstruction identity with a dual family  $\mathcal{G}$ .

This fact is summed up in the following theorem and is shown by choosing a basis consisting of  $d$  vectors in  $\mathcal{F}$  (without loss of generality, assumed to be  $f_1, \dots, f_d$ ) and applying the Riesz representation theorem to the linear functionals  $\gamma_i$  defined by  $\gamma_i(f_j) = \delta_{ij}$ .

**Theorem 2.4** [Bodmann et al. 2009, Theorem 2.4]. *The family  $\mathcal{F} = \{f_j\}_{j=1}^K$  in  $\mathbb{Z}_2^d$  is a frame if and only if there exist vectors  $\mathcal{G} = \{g_j\}_{j=1}^K$  such that, for all  $y \in \mathbb{Z}_2^d$ ,*

$$y = \sum_{j=1}^K (y, g_j) f_j. \tag{1}$$

In the proof,  $g_i$  is defined as the unique vector satisfying  $\gamma_i(y) = (y, g_i)$  for every  $y$  for  $1 \leq i \leq d$ , and  $g_i = 0$  for  $d < i \leq K$ . Equation (1) can be rewritten as

$$\Theta_{\mathcal{F}} \Theta_{\mathcal{G}}^* = I_d,$$

which is equivalent to  $\Theta_{\mathcal{G}} \Theta_{\mathcal{F}}^* = I_d$ . Consequently,  $\mathcal{G}$  is a dual frame to  $\mathcal{F}$ . We will refer to the dual frame  $\mathcal{G}$  as a *natural dual* to  $\mathcal{F}$ . Note that this definition is unrelated to the usual definition of the canonical dual in  $\mathbb{C}^d$  or  $\mathbb{R}^d$  as  $\{S^{-1}(f_j)\}$ , where  $S = \Theta \Theta^*$  is the frame operator from Definition 1.3. Although  $S_{\mathcal{F}}$  is no longer necessarily invertible, we still have

$$S_{\mathcal{F}}(g_i) = \sum_{j=1}^K (g_i, f_j) f_j = \sum_{j=1}^K \delta_{ij} f_j = f_i$$

for  $i \leq d$ .

Propositions 2.5 and 2.6 make clear that the natural dual frame is unique, up to permutation, if and only if  $K = d$ .

**Proposition 2.5.** *Let  $\mathcal{F} = \{f_j\}_{j=1}^K$  be a frame for  $\mathbb{Z}_2^d$  with a natural dual frame  $\mathcal{G}$ . Then  $\mathcal{H}$  is a dual frame of  $\mathcal{F}$  if and only if  $\Theta_{\mathcal{H}}^* = \Theta_{\mathcal{G}}^* + C$  for some  $K \times d$  matrix  $C$  satisfying  $\Theta_{\mathcal{F}} C = 0_d$ .*

*Proof.* Given the existence of a matrix  $C$  with  $\Theta_{\mathcal{H}}^* = \Theta_{\mathcal{G}}^* + C$  and  $\Theta_{\mathcal{F}} C = 0_d$ , it is immediate that  $\Theta_{\mathcal{F}} \Theta_{\mathcal{H}}^* = \Theta_{\mathcal{F}} \Theta_{\mathcal{G}}^* = I_d$ . Conversely, if  $\mathcal{H}$  is a dual frame of  $\mathcal{F}$ , then letting  $C = \Theta_{\mathcal{H}}^* - \Theta_{\mathcal{G}}^*$  gives  $\Theta_{\mathcal{F}} C = \Theta_{\mathcal{F}} \Theta_{\mathcal{H}}^* - \Theta_{\mathcal{F}} \Theta_{\mathcal{G}}^* = I_d - I_d = 0_d$ .  $\square$

The following result is well known in  $\mathbb{R}^d$  and  $\mathbb{C}^d$ ; see [Han et al. 2007, Proposition 6.3]. Since the proof in that text uses the invertibility of the frame operator, we provide a modified proof for  $\mathbb{Z}_2^d$  here.

**Proposition 2.6.** *Let  $\mathcal{F} = \{f_j\}_{j=1}^K$  be a frame for  $\mathbb{Z}_2^d$ . Then  $\mathcal{F}$  has a unique dual frame if and only if  $\mathcal{F}$  is a basis.*

*Proof.* Since a frame is a spanning set,  $\mathcal{F}$  is a basis if and only if the vectors  $f_j$  are linearly independent and  $K = d$ . This is equivalent to the only  $K \times d$  matrix  $C$  satisfying  $\Theta_{\mathcal{F}} C = 0_d$  being the zero matrix. By Proposition 2.5, this happens if and only if the (unique choice of) natural dual  $\mathcal{G}$  is the only dual frame of  $\mathcal{F}$ .  $\square$

The diagonal of the *Gramian* matrix  $\Theta_{\mathcal{F}}^* \Theta_{\mathcal{F}}$  is the vector whose  $i$ -th entry is  $(f_i, f_i)$ ; when  $\mathcal{F}$  and  $\mathcal{H}$  are a dual frame pair, the diagonal of the *cross-Gramian* matrix  $\Theta_{\mathcal{F}}^* \Theta_{\mathcal{H}}$  is the vector whose  $i$ -th element is  $(h_i, f_i)$ .

**Definition 2.7** [Bodmann et al. 2009]. A *Parseval frame* for  $\mathbb{Z}_2^d$  is a sequence of vectors  $\mathcal{F} = \{f_j\}_{j=1}^K \subset \mathbb{Z}_2^d$  such that

$$y = \sum_{j=1}^K (y, f_j) f_j$$

for all  $y \in \mathbb{Z}_2^d$ .

Note that a binary Parseval frame must be a binary frame. In matrix notation,  $\mathcal{F}$  is a Parseval frame for  $\mathbb{Z}_2^d$  if and only if  $\Theta_{\mathcal{F}} \Theta_{\mathcal{F}}^* = I_d$ . If a collection of vectors  $\{x_j\} \subset \mathbb{Z}_2^d$  satisfies  $(x_i, x_j) = 0$  for all  $i \neq j$  and  $(x_i, x_i) = 1$  for all  $i$ , we say, through a slight abuse of terminology, that  $\{x_j\}$  is an orthonormal set. An easy, matrix-theoretical consequence of the definitions of frame and Parseval frame is the following proposition:

**Proposition 2.8.** *Let  $\mathcal{F} = \{f_j\}_{j=1}^K$  be a sequence of vectors in  $\mathbb{Z}_2^d$ .*

- (1) *The rows of  $\Theta_{\mathcal{F}}$  are linearly independent if and only if  $\mathcal{F}$  is a frame.*
- (2) *The rows of  $\Theta_{\mathcal{F}}$  are orthonormal if and only if  $\mathcal{F}$  is a Parseval frame.*

In the rest of this paper, unless otherwise noted, all vectors are elements of the binary vector spaces  $\mathbb{Z}_2^d$  or  $\mathbb{Z}_2^K$ . All operations are performed modulo 2; for example,  $\text{Tr}(A)$  represents the usual trace of a matrix  $A$ , but  $\text{Tr}(AB) \equiv \text{Tr}(BA) \pmod{2}$  for two binary matrices  $A$  and  $B$ . All *frames* refer to *binary frames*. Throughout, we denote the standard orthonormal basis in  $\mathbb{Z}_2^d$  by  $\{\varepsilon_1, \varepsilon_2, \dots, \varepsilon_d\}$ .

### 3. Dual and Parseval binary frames

If  $\mathcal{F} = \{f_j\}_{j=1}^K$  is a frame for  $\mathbb{Z}_2^d$  and  $K = d$ , then the vectors  $\{f_j\}$  must be linearly independent and hence a basis with a unique (natural) dual  $\mathcal{G}$ . In this case,  $(f_j, g_j) = \gamma_j(f_j) = 1$  for all  $1 \leq j \leq K$ . In this section, we are largely concerned with the question of which sequences  $\alpha \in \mathbb{Z}_2^K$  satisfy  $\alpha[j] = (f_j, h_j)$  for a dual frame pair  $(\mathcal{F}, \mathcal{H})$ ; so we assume  $K > d$ . We use  $\|\alpha\|_0$  to denote the number of nonzero entries in a vector  $\alpha$ , the parity of which will be fundamental in this paper.

**Lemma 3.1.** *Let  $\mathcal{F} = \{f_j\}_{j=1}^K$  be a frame for  $\mathbb{Z}_2^d$  and let  $\pi$  be a permutation of the set  $\{1, 2, \dots, K\}$ . Denote by  $\mathcal{F}_{\pi}$  the frame  $\{f_{\pi(j)}\}_{j=1}^K$ . Then a frame  $\mathcal{H}$  is a dual frame of  $\mathcal{F}$  if and only if  $\mathcal{H}_{\pi}$  is a dual frame of  $\mathcal{F}_{\pi}$ . Furthermore, if  $\alpha$  is a sequence in  $\mathbb{Z}_2^K$ , then the dual frame pair  $(\mathcal{F}, \mathcal{H})$  satisfies  $(f_j, h_j) = \alpha[j]$  for every  $j$  if and only if the dual frame pair  $(\mathcal{F}_{\pi}, \mathcal{H}_{\pi})$  satisfies  $(f_{\pi(j)}, h_{\pi(j)}) = \alpha[\pi(j)]$  for every  $j$ .*

*Proof.* Suppose  $\Theta_{\mathcal{F}} \Theta_{\mathcal{H}}^* = I_d$  and  $\text{diag}(\Theta_{\mathcal{H}}^* \Theta_{\mathcal{F}}) = \alpha$ . Then

$$\begin{aligned} \Theta_{\mathcal{F}_\pi} \Theta_{\mathcal{H}_\pi}^* &= \begin{bmatrix} | & & | \\ f_{\pi(1)} & \cdots & f_{\pi(K)} \\ | & & | \end{bmatrix} \begin{bmatrix} -h_{\pi(1)}^* - \\ \vdots \\ -h_{\pi(K)}^* - \end{bmatrix} \\ &= \begin{bmatrix} | & & | \\ f_{\pi(1)} & \cdots & f_{\pi(K)} \\ | & & | \end{bmatrix} P^* P \begin{bmatrix} -h_{\pi(1)}^* - \\ \vdots \\ -h_{\pi(K)}^* - \end{bmatrix} \\ &= \begin{bmatrix} | & & | \\ f_1 & \cdots & f_K \\ | & & | \end{bmatrix} \begin{bmatrix} -h_1^* - \\ \vdots \\ -h_K^* - \end{bmatrix} \\ &= \Theta_{\mathcal{F}} \Theta_{\mathcal{H}}^* = I_d, \end{aligned}$$

where

$$P = \begin{bmatrix} -\varepsilon_{\pi^{-1}(1)}^* - \\ \vdots \\ -\varepsilon_{\pi^{-1}(K)}^* - \end{bmatrix},$$

and here we use  $\varepsilon_i$  to indicate the  $i$ -th standard basis vector in  $\mathbb{Z}_2^K$ . Thus  $\mathcal{F}_\pi$  and  $\mathcal{H}_\pi$  are dual frames and  $(f_j, h_j) = \alpha[j]$  for each  $j$  implies  $(f_{\pi(j)}, h_{\pi(j)}) = \alpha[\pi(j)]$ . For the converse statements, let  $\sigma = \pi^{-1}$ .  $\square$

The next theorem and corollary are the analog of Theorem 1.6 in binary space.

**Theorem 3.2.** *Given  $\alpha \in \mathbb{Z}_2^K$ , there exists a dual frame pair  $(\mathcal{F}, \mathcal{H})$  for  $\mathbb{Z}_2^d$  with  $(f_i, h_i) = \alpha[i]$  for every  $i$  if and only if  $\|\alpha\|_0 \equiv d \pmod{2}$ .*

*Proof.* Suppose  $(\mathcal{F}, \mathcal{H})$  is a dual frame pair for  $\mathbb{Z}_2^d$  such that  $(f_i, h_i) = \alpha[i]$  for every  $i$ . Then

$$\|\alpha\|_0 \equiv \text{Tr}(\Theta_{\mathcal{H}}^* \Theta_{\mathcal{F}}) \equiv \text{Tr}(\Theta_{\mathcal{F}} \Theta_{\mathcal{H}}^*) \equiv \text{Tr}(I_d) \equiv d \pmod{2}.$$

Conversely, suppose  $\|\alpha\|_0 \equiv d \pmod{2}$ . We consider three cases.

Case 1:  $\|\alpha\|_0 = d$ . Let  $f_j = \varepsilon_j$  for  $1 \leq j \leq d$  and let  $f_j$  be arbitrary for  $d < j \leq K$ . A natural dual frame is then given by  $g_j = \varepsilon_j$  for  $1 \leq j \leq d$  and  $g_j = 0$  for  $d < j \leq K$ . Define  $\beta \in \mathbb{Z}_2^K$  by  $\beta[j] = 1$  for  $1 \leq j \leq d$  and  $\beta[j] = 0$  for  $d < j \leq K$ , and let  $\pi$  be a permutation of  $\{1, 2, \dots, K\}$  such that  $\beta[j] = \alpha[\pi(j)]$ . It follows that

$$\text{diag}(\Theta_{\mathcal{G}}^* \Theta_{\mathcal{F}}) = \beta.$$

By Lemma 3.1,  $(\mathcal{F}_{\pi^{-1}}, \mathcal{G}_{\pi^{-1}})$  is the desired dual frame pair with  $(f_{\pi^{-1}(j)}, g_{\pi^{-1}(j)}) = \beta[\pi^{-1}(j)] = \alpha[j]$  for each  $j$ .

Case 2:  $\|\alpha\|_0 = d + 2t$  for some positive integer  $t \leq \frac{1}{2}(K - d)$ . Consider the frame  $\mathcal{F}$  defined in Case 1 above, but set  $f_j = \varepsilon_1$  for all  $d + 1 \leq j \leq d + 2t$ . A

natural dual frame  $\mathcal{G}$  of  $\mathcal{F}$  is the same as that defined in Case 1 above. Let  $C$  be the  $K \times d$  matrix whose top  $d \times d$  block is  $0_d$  and rows  $d + 1$  through  $d + 2t$  are equal to  $\varepsilon_1^*$ . The remaining rows of  $C$  are zeros. Due to the introduction of an even number of  $\varepsilon_1^*$ 's, we see that  $\Theta_{\mathcal{F}} C = 0_d$ , and hence  $\mathcal{H}$  given by the rows of  $\Theta_{\mathcal{H}}^* = \Theta_{\mathcal{G}}^* + C$  is a dual frame of  $\mathcal{F}$ , by Proposition 2.5. Since  $\text{diag}(\Theta_{\mathcal{H}}^* \Theta_{\mathcal{F}})$  is the vector composed of  $d + 2t$  ones followed by  $K - (d + 2t)$  zeros, Lemma 3.1 again implies the existence of the desired dual frame pair.

Case 3:  $\|\alpha\|_0 = d - 2t$  for some positive integer  $t \leq \frac{1}{2}d$ . Consider again the frame  $\mathcal{F}$  defined in Case 1, except set  $f_{d+1} = \varepsilon_d + \varepsilon_{d-1} + \dots + \varepsilon_{d-2t+2} + \varepsilon_{d-2t+1}$ . We again take the same natural dual frame  $\mathcal{G}$  as in Case 1 above. Let  $C$  be the  $K \times d$  matrix whose top  $d - 2t$  rows are zeros, rows  $d - 2t + 1$  through  $d + 1$  are  $f_{d+1}^*$ , and the remaining rows are zeros. Then  $\Theta_{\mathcal{F}} C = 0_d$ , and hence  $\mathcal{H}$  defined by  $\Theta_{\mathcal{H}}^* = \Theta_{\mathcal{G}}^* + C$  is a dual frame of  $\mathcal{F}$ . Due to the presence of an even number of ones in  $f_{d+1}$ , we have  $(f_{d+1}, f_{d+1}) = 0$ , while  $(f_j, h_j) = 0$  for  $d - 2t + 1 \leq j \leq d$ . Consequently,  $\text{diag}(\Theta_{\mathcal{H}}^* \Theta_{\mathcal{F}})$  consist of ones in its first  $d - 2t$  entries followed by  $K - (d - 2t)$  zeros. Lemma 3.1 again completes the proof.  $\square$

**Corollary 3.3.** *Given  $\alpha \in \mathbb{Z}_2^K$ , there exists a Parseval frame  $\mathcal{F}$  and a corresponding dual frame  $\mathcal{H}$  for  $\mathbb{Z}_2^d$  with  $(f_i, h_i) = \alpha[i]$  for every  $i$  if and only if  $\|\alpha\|_0 \equiv d \pmod{2}$ .*

*Proof.* The necessity of the condition on  $\|\alpha\|_0$  follows immediately by Theorem 3.2. The sufficiency depends on slight modifications of the frame  $\mathcal{F}$  constructed in the proof of Theorem 3.2. In Case 1, instead of letting  $f_j$  for  $d < j \leq K$  be arbitrary, set each of those vectors to be the zero vector,  $\vec{0}$ , in  $\mathbb{Z}_2^d$ . Proposition 2.8 implies  $\mathcal{F}$  is a Parseval frame. Similarly, the frame built in Case 2 is a Parseval frame if we set  $f_j = \vec{0}$  for  $2t + 1 \leq j \leq K$ . The frame built in Case 3 is not a Parseval frame; however, consider instead the frame  $\mathcal{F}'$  defined as  $f'_j = f_j$  for  $1 \leq j \leq d + 1$ ,  $f'_{d+2} = f_{d+1}$ , and  $f'_j = \vec{0}$  for  $d + 3 \leq j \leq K$ . By Proposition 2.8,  $\mathcal{F}'$  is a Parseval frame. Note that each column of the matrix  $C$  constructed in Case 3 is still a (possibly trivial) dependence relation among the columns of  $\Theta_{\mathcal{F}'}$ , which implies  $\Theta_{\mathcal{F}'} C = 0_d$ . Since the natural dual  $\mathcal{G}$  of  $\mathcal{F}$  constructed in Case 3 is still a natural dual of  $\mathcal{F}'$ , the frame  $\mathcal{H}$  with analysis operator  $\Theta_{\mathcal{H}}^* = \Theta_{\mathcal{G}}^* + C$  is a dual frame of  $\mathcal{F}'$ ; moreover,  $(f'_j, h_j) = (f_j, h_j)$  for all  $j$  since  $h_{d+2} = 0$ .  $\square$

**Remark 3.4.** The Parseval frames built in Cases 1 and 2 of the above corollary, in fact, satisfy  $(f_j, f_j) = \alpha[j]$  for each  $j$  after a suitable permutation, as allowed by Lemma 3.1. However, this is not true in Case 3. By constructing a Parseval frame that satisfies  $(f_j, f_j) = \alpha[j]$  for each  $j$  in the case when  $\|\alpha\|_0 = d - 2t$  for some positive integer  $t \leq \frac{1}{2}d$ , we will prove the binary analog of Theorem 1.4.

**Theorem 3.5.** *Given nonzero  $\alpha \in \mathbb{Z}_2^K$ , there exists a Parseval frame  $\mathcal{F}$  for  $\mathbb{Z}_2^d$  with  $(f_i, f_i) = \alpha[i]$  for every  $i$  if and only if  $\|\alpha\|_0 \equiv d \pmod{2}$ .*



*Proof.* Since a Parseval frame is self-dual, the necessity of the condition on  $\|\alpha\|_0$  follows immediately from Theorem 3.2. For sufficiency, Remark 3.4 implies that we need only construct Parseval frames satisfying  $(f_i, f_i) = \alpha[i]$  for every  $i$  for  $\|\alpha\|_0 < d$ . Since  $\|\alpha\|_0 \equiv d \pmod{2}$  and  $\|\alpha\|_0 \neq 0$ , we must have  $d \geq 3$ .

For  $d = 3$  and  $\|\alpha\|_0 = 1$ , we build the Parseval frame

$$\left\{ \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right\},$$

and note that we may permute the vectors as needed. Moreover, we may insert any number of copies of  $\vec{0}$  to satisfy any  $K > 4$ . By augmenting each vector with a last entry of 0 and inserting the vector  $\varepsilon_4 \in \mathbb{Z}_2^4$ , we construct the Parseval frame

$$\left\{ \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \right\}$$

for  $\mathbb{Z}_2^4$  that, after suitable permutation and inclusion of copies of  $\vec{0}$ , satisfies any  $\|\alpha\|_0 = 2$ .

Given any odd dimension  $d$ , suppose we have constructed, without zero vectors, the Parseval frames  $\mathcal{F}^1, \mathcal{F}^3, \dots, \mathcal{F}^{d-4}$  for  $\mathbb{Z}_2^{d-2}$  corresponding to  $\|\alpha\|_0 = 1, 3, \dots, d-4$ . For each odd  $n$ , create the collection  $\tilde{\mathcal{F}}^{n+2}$  by augmenting each vector of  $\mathcal{F}^n$  with two zero entries and unioning the augmented vectors with  $\varepsilon_{d-1}, \varepsilon_d \in \mathbb{Z}_2^d$ . Then  $\tilde{\mathcal{F}}^3, \tilde{\mathcal{F}}^5, \dots, \tilde{\mathcal{F}}^{d-2}$  are Parseval frames for  $\mathbb{Z}_2^d$  corresponding to  $\|\alpha\|_0 = 3, 5, \dots, d-2$ . Let  $\tilde{\mathcal{F}}^1 = \{\vec{1} + \varepsilon_1, \vec{1} + \varepsilon_2, \dots, \vec{1} + \varepsilon_d, \vec{1}\}$ , where  $\vec{1}$  represents the vector with  $d$  ones in  $\mathbb{Z}_2^d$ . Then  $\tilde{\mathcal{F}}^1$  is a Parseval frame for  $\mathbb{Z}_2^d$  corresponding to  $\|\alpha\|_0 = 1$ .

If the dimension  $d$  is even, we create the Parseval frames  $\tilde{F}^{n+1}$  from  $\tilde{F}^n$  for each  $n = 1, 3, \dots, d-3$ , corresponding to  $\|\alpha\|_0 = 2, 4, \dots, d-2$ : augment each vector of  $\tilde{F}^n$  with a last entry of 0 and insert the vector  $\varepsilon_d \in \mathbb{Z}_2^d$ .

In both cases, permutation of the vectors and possible inclusion of copies of  $\vec{0}$  finishes the proof. □

### 4. Binary frames with prescribed frame operator

In the previous section, we gave a necessary and sufficient condition on  $\alpha \in \mathbb{Z}_2^K$  for the existence of a Parseval frame  $\mathcal{F}$  for  $\mathbb{Z}_2^d$  with  $(f_j, f_j) = \alpha[j]$  for every  $j$ . In classical frame theory over  $\mathbb{R}$  or  $\mathbb{C}$ , the characterization has been broadened to frames with a given frame operator and specified values for  $\|f_j\|$  (the case of a Parseval frame is when  $S = I$ ), as in Theorem 1.5. In the classical case, the frame operator is a symmetric, invertible, positive definite matrix. For a binary frame  $\mathcal{F}$ ,  $S_{\mathcal{F}} = \Theta_{\mathcal{F}} \Theta_{\mathcal{F}}^*$

is not necessarily invertible; for example, the zero matrix is the frame operator of any frame in which every vector occurs twice. Consequently, we must first characterize those binary symmetric matrices that are frame operators of binary frames.

Throughout this section, we rely heavily on the idea of vector parity in  $\mathbb{Z}_2^d$ .

**Definition 4.1.** Describe a vector  $v \in \mathbb{Z}_2^d$  as *even* if  $(v, v) = 0$ . Equivalently, a vector is even if  $\|v\|_0 \equiv 0 \pmod{2}$ . If a vector is not even, then it is *odd*.

**Lemma 4.2.** (1) *The sum of two even vectors is an even vector.*

(2) *The sum of two odd vectors is an even vector.*

(3) *The sum of an odd vector and an even vector is an odd vector.*

*Proof.* This follows from the above definition and the observation that if  $u, v \in \mathbb{Z}_2^d$ , then

$$(u + v, u + v) = (u, u) + (u, v) + (v, u) + (v, v) = (u, u) + (v, v). \quad \square$$

As a consequence of this lemma, we note that a collection of only even vectors cannot span  $\mathbb{Z}_2^d$ .

Given a  $d \times d$  symmetric matrix  $S$ , we call  $A$  a *factor* of  $S$  if  $S = AA^*$ . We say  $A$  is a *minimal factor* if it has the minimum number of columns over all factors of  $S$ . Minimal factorization of symmetric binary matrices also arises in the computation of the covering radius of Reed–Muller codes [Cohen et al. 1997].

**Theorem 4.3** [Lempel 1975, Theorem 1]. *Every binary symmetric matrix  $S$  can be factorized as  $S = AA^*$  for some binary matrix  $A$ . The number of columns of a minimal factor of  $S$  is  $\text{rank}(S)$  if  $\text{diag}(S) \neq \vec{0}$  and  $\text{rank}(S) + 1$  if  $\text{diag}(S) = \vec{0}$ .*

**Proposition 4.4.** *If  $S = AA^*$  for some  $d \times m$  matrix  $A$ , where  $m = \text{rank}(S)$  or  $m = \text{rank}(S) + 1$ , then  $\text{rank}(A) = \text{rank}(S)$ .*

*Proof.* Frobenius' rank inequality and the fact that, for properly sized matrices  $C$  and  $D$ ,  $\text{rank}(CD) \leq \min\{\text{rank}(C), \text{rank}(D)\}$  [Horn and Johnson 1985] imply

$$\text{rank}(A) + \text{rank}(A^*) \leq m + \text{rank}(S) \leq m + \text{rank}(A).$$

If  $m = \text{rank}(S)$ , the above inequalities simplify to

$$\text{rank}(A) \leq \text{rank}(S) \leq \text{rank}(A),$$

and hence  $\text{rank}(A) = \text{rank}(S)$ . If  $m = \text{rank}(S) + 1$ , we instead have

$$2 \text{rank}(A) \leq 2 \text{rank}(S) + 1 \leq 2 \text{rank}(A) + 1.$$

Since  $2 \text{rank}(A) = 2 \text{rank}(S) + 1$  is impossible, we must have  $\text{rank}(A) = \text{rank}(S)$ .  $\square$

We can use factors of a matrix to construct frames with a given frame operator. Minimal factors correspond to *minimal frames*, that is, frames with the fewest number of elements. In the previous section, we disregarded frames  $\{f_j\}_{j=1}^K$  that

were bases since they corresponded to unique duals  $\{g_j\}_{j=1}^K$  with predetermined values for the dot products  $(f_j, g_j)$ . In this section, however, we are concerned with  $(f_j, f_j)$ , so we do not rule the case  $K = d$  out of consideration.

**Theorem 4.5.** *Let  $S$  be a  $d \times d$  symmetric matrix with  $\text{rank}(S) = d$ . There exists a (minimal)  $d$ -element frame  $\mathcal{F}$  (i.e., basis) such that  $S = \Theta_{\mathcal{F}} \Theta_{\mathcal{F}}^*$  if and only if  $\text{diag}(S) \neq \vec{0}$ . If  $\text{diag}(S) = \vec{0}$ , then there exists a (minimal)  $(d+1)$ -element frame  $\mathcal{F}$  such that  $S = \Theta_{\mathcal{F}} \Theta_{\mathcal{F}}^*$ .*

*Proof.* Suppose there is a  $d$ -element frame  $\mathcal{F}$  such that  $S = \Theta_{\mathcal{F}} \Theta_{\mathcal{F}}^*$ . If  $\text{diag}(S) = \vec{0}$ , then  $\Theta_{\mathcal{F}}$  is a  $d \times d$  square matrix with all even rows, which cannot span  $\mathbb{Z}_2^d$ . Hence, the columns of  $\Theta_{\mathcal{F}}$  cannot span  $\mathbb{Z}_2^d$ , contradicting  $\mathcal{F}$  being a frame. Conversely, suppose  $\text{diag}(S) \neq \vec{0}$ . By Theorem 4.3, there exists a  $d \times k$  matrix  $A$  such that  $S = AA^*$  and  $k = \text{rank}(S)$ . By Proposition 4.4,  $\text{rank}(A) = \text{rank}(S) = k = d$ . Thus,  $A$  is a  $d \times d$  matrix whose columns span  $\mathbb{Z}_2^d$ . Defining  $\mathcal{F}$  by  $\Theta_{\mathcal{F}} = A$  constructs the  $d$ -element frame.

Now suppose  $\text{diag}(S) = \vec{0}$ . By Theorem 4.3, there exists a minimal  $d \times k$  factor  $A$  such that  $S = AA^*$  and  $k = \text{rank}(S) + 1$ . By Proposition 4.4,  $\text{rank}(A) = \text{rank}(S) = d$ . Therefore, the columns of  $A$  span  $\mathbb{Z}_2^d$ , and we can construct a  $(d+1)$ -element frame by taking  $\Theta_{\mathcal{F}} = A$ . □

**Remark 4.6.** The columns of  $\Theta_{\mathcal{F}}$  can be augmented by copies of the zero vector without affecting  $S$ , so nonminimal frames can always be constructed from minimal frames by including any number of copies of the zero vector.

Next we construct minimal frames whose prescribed frame operators are not of full rank. Let  $r(A)$  and  $c(A)$  denote the number of rows and columns of a matrix  $A$ , respectively.

**Lemma 4.7** [Lempel 1975, Section 4]. *Let  $S$  be a  $d \times d$  symmetric matrix of  $\text{rank}(S) < d$ . There exists a permutation matrix  $P$  and a nonsingular matrix  $T$  such that*

$$S = P^* \begin{bmatrix} L & M \\ M^* & K \end{bmatrix} P = P^* T \begin{bmatrix} L & 0 \\ 0 & 0 \end{bmatrix} T^* P,$$

where  $L$  is a symmetric matrix with  $r(L) = \text{rank}(L) = \text{rank}(S)$ .

**Corollary 4.8.** *A  $d \times d$  symmetric matrix  $S$  with  $\text{diag}(S) = \vec{0}$  must have even rank.*

*Proof.* It is known that if  $\text{rank}(S) = d$ , then  $d$  must be even; see, for example, [Cohen et al. 1997, Section 9.3]. Suppose  $\text{rank}(S) < d$ . The  $\text{rank}(S) \times \text{rank}(S)$  symmetric matrix  $L$  constructed in Lemma 4.7 has  $\text{rank}(L) = \text{rank}(S)$ . Since the diagonal elements of  $S$  are the diagonal elements of  $L$  and  $K$ ,  $\text{diag}(L) = \vec{0}$ . So  $\text{rank}(S)$  must be even. □

**Theorem 4.9.** *Let  $S$  be a  $d \times d$  symmetric matrix with  $\text{rank}(S) < d$ . There exists a  $k$ -element frame  $\mathcal{F}$  such that  $S = \Theta_{\mathcal{F}} \Theta_{\mathcal{F}}^*$  if and only if  $k \geq 2d - \text{rank}(S)$ .*

*Proof.* Necessity follows from Frobenius’s rank inequality

$$\text{rank}(\Theta_{\mathcal{F}}) + \text{rank}(\Theta_{\mathcal{F}}^*) \leq k + \text{rank}(S)$$

since frames are spanning sets.

Conversely, let  $k$  be an integer such that  $k \geq 2d - \text{rank}(S)$ . Let  $L, P, T$  be the matrices guaranteed by Lemma 4.7, and let  $V = P^*T$ . Suppose  $\text{diag}(L) \neq \vec{0}$ . By Theorem 4.3 there exists a factor  $H$  of  $L$  such that

$$\begin{bmatrix} L & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} H \\ 0 \end{bmatrix} \begin{bmatrix} H^* & 0 \end{bmatrix}$$

and  $r(H) = c(H) = \text{rank}(L)$ . Consider the augmented matrix

$$A = \begin{bmatrix} H & 0 \\ 0 & B \end{bmatrix},$$

where the columns of  $B$  are the standard basis vectors of  $\mathbb{Z}_2^{d-\text{rank}(S)}$ , each repeated twice. Then  $r(A) = d$ ,  $c(A) = 2d - \text{rank}(S)$ , and  $AA^* = \begin{bmatrix} L & 0 \\ 0 & 0 \end{bmatrix}$ . By construction,  $\text{rank}(A) = d$ . Since  $V$  is nonsingular,  $VA$  is a  $d \times (2d - \text{rank}(S))$  matrix of rank  $d$  such that  $S = VAA^*V^*$ . If  $k = 2d - \text{rank}(S)$ , let a minimal frame  $\mathcal{F}$  be the columns of  $VA$ ; if  $k > 2d - \text{rank}(S)$ , augment the columns of  $VA$  with the necessary number of zero vectors.

Now suppose  $\text{diag}(L) = \vec{0}$ . By Corollary 4.8,  $\text{rank}(L) = \text{rank}(S)$  must be even. As above (by Theorem 4.3) we can factorize  $L$  with a matrix  $H$ , but now  $r(H) = \text{rank}(L)$  and  $c(H) = \text{rank}(L) + 1$ . In this case, we build the augmented matrix

$$\tilde{A} = \left[ \begin{array}{c|c} H & 0 \\ \hline r_1 & r_2 \\ \hline 0 & \tilde{B} \end{array} \right],$$

where  $r_1 = [1 \ 1 \ 1 \ \dots \ 1]$  is a vector of length  $c(H)$ ,  $r_2 = [1 \ 0 \ 0 \ 0 \ \dots \ 0]$  has length  $2(d - \text{rank}(S)) - 1$ , and the columns of  $\tilde{B}$  are the zero vector followed by the standard basis vectors of  $\mathbb{Z}_2^{d-(\text{rank}(S)+1)}$ , each repeated twice. Since the  $(i, j)$  entry of the product  $\tilde{A}\tilde{A}^*$  can be viewed as the dot product of the  $i$ -th and  $j$ -th rows of  $\tilde{A}$ , it is easy to see that  $\tilde{A}\tilde{A}^* = \begin{bmatrix} L & 0 \\ 0 & 0 \end{bmatrix}$ . Indeed, since  $\text{diag}(L) = \vec{0}$ , each row of  $H$  is an even vector. Since  $c(H)$  is odd, the vector  $[r_1 \ | \ r_2]$  is even, as is each row of  $\tilde{B}$ . By construction,  $\text{rank}(\tilde{A}) = d$ ,  $r(\tilde{A}) = r(H) + 1 + d - \text{rank}(S) - 1 = d$ , and  $c(\tilde{A}) = c(H) + 1 + 2(d - \text{rank}(S) - 1) = 2d - \text{rank}(S)$ . As above let  $\mathcal{F}$  consist of the columns of  $V\tilde{A}$  if  $k = 2d - \text{rank}(S)$  or these columns together with  $k - 2d + \text{rank}(S)$  copies of the zero vector if  $k > 2d - \text{rank}(S)$ .  $\square$

Theorems 4.5 and 4.9 provide minimal (and nonminimal) frames with frame operator  $S$ , subject only to restrictions based on  $\text{rank}(S)$ . In what follows, sometimes

we will make additional assumptions on  $S$ , which allow the construction of frames with frame operator  $S$  in different, and sometimes more intuitive, ways.

**Definition 4.10.** A  $d \times d$  symmetric matrix  $S$  is said to be *parity indicative* if, for every  $1 \leq i \leq d$ , the diagonal entry  $S_{ii}$  is equal to 1 if and only if the  $i$ -th row of  $S$  is odd.

**Lemma 4.11.** Let  $S$  be a  $d \times d$  symmetric matrix, and suppose  $S = AA^*$  for some  $d \times m$  matrix  $A$ . If every column of  $A$  is odd, then  $S$  is parity indicative. Conversely, if  $S$  is parity indicative and the columns of  $A$  are linearly independent, then every column of  $A$  must be odd.

*Proof.* Assume  $S = AA^* = \sum_{i=1}^m a_i a_i^*$ , where each column  $a_i$  of  $A$  is odd. The  $i$ -th row of  $S$  equals the sum of those  $a_j^*$  satisfying  $a_j[i] = 1$ . Suppose the  $i$ -th row of  $S$  is odd. Then this sum must be composed of an odd number of nonzero terms by Lemma 4.2. That is, there is an odd number of indices  $j$  having  $a_j[i] = 1$ . Consequently, the  $i$ -th row of  $A$  is an odd vector, and  $S_{ii} = \sum_{j=1}^m a_j[i] a_j[i] = 1$ . On the other hand, if the  $i$ -th row of  $S$  is even, then Lemma 4.2 implies that an even number of  $a_j$  have  $a_j[i] = 1$ , resulting in  $S_{ii} = 0$ .

To show the converse, assume that the columns of  $A$  are linearly independent. Suppose some of the columns of  $A$  are even; denote the odd columns by  $\{o_j\}$  and the even columns by  $\{e_j\}$ . If for every index  $l$ , we have  $e_j[l] = 1$  for an even number of the vectors  $\{e_j\}$ , then  $\sum e_j = \vec{0}$ , contradicting the linear independence of the columns of  $A$ . So, assume that there exists an index  $i$  such that the number of the vectors  $\{e_j\}$  that satisfy  $e_j[i] = 1$  is odd. If an odd number of the vectors  $\{o_j\}$  are such that  $o_j[i] = 1$ , then the  $i$ -th row of  $A$  is even, so  $S_{ii} = 0$ ; on the other hand, the  $i$ -th row of  $S$  equals the sum of an odd number of rows  $e_j^*$  plus an odd number of rows  $o_j^*$ , which is odd, by Lemma 4.2. If there are an even number of the vectors  $\{o_j\}$  with  $o_j[i] = 1$ , then the  $i$ -th row of  $A$  is odd, so  $S_{ii} = 1$ ; on the other hand, the  $i$ -th row of  $S$  equals the sum of an odd number of  $e_j^*$  plus an even number of  $o_j^*$ , which is even, by Lemma 4.2. Therefore,  $S$  is not parity indicative.  $\square$

**Lemma 4.12.** Let  $S$  be a  $d \times d$  symmetric matrix, and suppose  $S = AA^*$  for some  $d \times m$  matrix  $A$ . If  $S$  is parity indicative,  $\text{diag}(S) = \vec{0}$ , and  $c(A) = \text{rank}(A) + 1$ , then either every column of  $A$  is even or every column of  $A$  is odd.

*Proof.* Denote the odd columns and even columns of  $A$  by  $\{o_j\}$  and  $\{e_j\}$ , respectively, and assume both sets are nonempty. Since each row of  $S$  is even, for every  $i$ , an even number of the vectors  $\{o_j\}$  must have  $o_j[i] = 1$ , by Lemma 4.2. It follows that  $\sum o_j = \vec{0}$ ; that is,  $Ax = \vec{0}$  where  $x[j] = 1$  if  $j$  is the index of an odd column and  $x[j] = 0$  if  $j$  is the index of an even column. But since every row of  $A$  is even,  $A\vec{1} = \vec{0}$ . Hence  $Ay = \vec{0}$  for  $y = \vec{1} + x$ . Since the nonzero linearly independent vectors  $x$  and  $y$  are both contained in the null space of  $A$ , the rank-nullity theorem implies

$$\text{rank}(A) \leq c(A) - 2 = \text{rank}(A) + 1 - 2 = \text{rank}(A) - 1,$$

a contradiction. Therefore, either  $\{o_j\}$  or  $\{e_j\}$  must be empty. □

One additional useful fact is required before we state our main result.

**Lemma 4.13.** (1) *Suppose  $e, o_1, o_2, o_3 \in \mathbb{Z}_2^d$  are four vectors such that  $e$  is even and  $o_1, o_2, o_3$  are odd. Then there exists three even vectors  $f_1, f_2, f_3$  and an odd vector  $p$  such that*

$$ee^* + o_1o_1^* + o_2o_2^* + o_3o_3^* = f_1f_1^* + f_2f_2^* + f_3f_3^* + pp^*,$$

$$\text{and Span}\{e, o_1, o_2, o_3\} = \text{Span}\{f_1, f_2, f_3, p\}.$$

(2) *Suppose  $e_1, e_2, e_3, o \in \mathbb{Z}_2^d$  are four vectors such that  $e_1, e_2, e_3$  are even and  $o$  is odd. Then there exists an even vector  $f$  and odd vectors  $p_1, p_2, p_3$  such that*

$$e_1e_1^* + e_2e_2^* + e_3e_3^* + oo^* = ff^* + p_1p_1^* + p_2p_2^* + p_3p_3^*,$$

$$\text{and Span}\{e_1, e_2, e_3, o\} = \text{Span}\{f, p_1, p_2, p_3\}.$$

*Proof.* For part (1), let  $f_1 = e + o_1 + o_2$ ,  $f_2 = e + o_1 + o_3$ ,  $f_3 = e + o_2 + o_3$ , and  $p = o_1 + o_2 + o_3$ . By Lemma 4.2,  $f_1, f_2$  and  $f_3$  are even and  $p$  is odd. For part (2), let  $f = e_1 + e_2 + e_3$ , and  $p_1 = e_1 + e_2 + o$ ,  $p_2 = e_1 + e_3 + o$ ,  $p_3 = e_2 + e_3 + o$ . Lemma 4.2 implies  $f$  is even and  $p_1, p_2, p_3$  are odd. Easy computations show that the given equalities are satisfied. □

We are now ready for the binary analog of Theorem 1.5: necessary and sufficient conditions on pairs of matrices  $S$  and vectors  $\alpha$  such that  $S$  is the frame operator of a frame with vector “norms” determined by  $\alpha$ . The necessary condition is easy.

**Theorem 4.14.** *Let  $\mathcal{F} = \{f_i\}_{i=1}^K$  be a frame such that  $S = \Theta_{\mathcal{F}} \Theta_{\mathcal{F}}^*$ . Let  $\alpha$  be the vector in  $\mathbb{Z}_2^K$  defined by  $\alpha[i] = (f_i, f_i)$  for each  $i$ . Then  $\|\alpha\|_0 \equiv \text{Tr}(S) \pmod{2}$ .*

*Proof.*  $\|\alpha\|_0 \equiv \text{Tr}(\Theta_{\mathcal{F}}^* \Theta_{\mathcal{F}}) \equiv \text{Tr}(\Theta_{\mathcal{F}} \Theta_{\mathcal{F}}^*) \equiv \text{Tr}(S) \pmod{2}$ . □

Sufficiency breaks down into three possible scenarios. If  $S$  is parity indicative, then a minimal frame  $\mathcal{F}$  with frame operator  $S$  must consist of only odd vectors or can attain any nonzero vector  $\alpha$  with  $\|\alpha\|_0 \equiv \text{Tr}(S) \pmod{2}$  in the sense that  $(f_i, f_i) = \alpha[i]$  for each  $i$ ; if  $S$  is not parity indicative, a minimal frame must contain at least one even vector. This is shown in Theorems 4.15 and 4.17. Nonminimal frames can be constructed to correspond to any nonzero  $\alpha$  with  $\|\alpha\|_0 \equiv \text{Tr}(S) \pmod{2}$  if  $S$  is parity indicative or to any such  $\alpha$  with at least one zero entry if  $S$  is not parity indicative (Corollary 4.16 and Theorem 4.17).

The frame elements can be permuted in any way without affecting the frame operator. Indeed, if  $\Theta_{\mathcal{F}} \Theta_{\mathcal{F}}^* = S$  and  $\Theta_{\tilde{\mathcal{F}}} = \Theta_{\mathcal{F}} P^*$  for some permutation matrix  $P$ , then

$$\Theta_{\tilde{\mathcal{F}}} \Theta_{\tilde{\mathcal{F}}}^* = \Theta_{\mathcal{F}} P^* P \Theta_{\mathcal{F}}^* = \Theta_{\mathcal{F}} \Theta_{\mathcal{F}}^* = S.$$

Therefore, in what follows, we need only construct frames with the correct number of odd elements, corresponding to  $\|\alpha\|_0$ , in order to attain the dot products prescribed by  $\alpha$ .

**Theorem 4.15.** *Let  $S$  be a  $d \times d$  parity indicative symmetric matrix. Let  $K = 2d - \text{rank}(S)$ , and let  $\alpha \in \mathbb{Z}_2^K$  be a nonzero vector with  $\|\alpha\|_0 \equiv \text{Tr}(S) \pmod{2}$ .*

(1) *Suppose  $\text{diag}(S) \neq \vec{0}$ .*

(a) *If  $\text{rank}(S) = d$ , there exists a (minimal)  $K$ -element frame  $\mathcal{F}$  such that  $S = \Theta_{\mathcal{F}} \Theta_{\mathcal{F}}^*$  and  $(f_i, f_i) = \alpha[i]$  for every  $i$  only if  $\|\alpha\|_0 = K$ .*

(b) *If  $\text{rank}(S) < d$ , there exists a (minimal)  $K$ -element frame  $\mathcal{F}$  such that  $S = \Theta_{\mathcal{F}} \Theta_{\mathcal{F}}^*$  and  $(f_i, f_i) = \alpha[i]$  for every  $i$ .*

(2) *Suppose  $\text{diag}(S) = \vec{0}$ . Then  $\text{rank}(S) < d$ .*

(a) *If  $\text{rank}(S) = d - 1$ , there exists a (minimal)  $K$ -element frame  $\mathcal{F}$  such that  $S = \Theta_{\mathcal{F}} \Theta_{\mathcal{F}}^*$  and  $(f_i, f_i) = \alpha[i]$  for every  $i$  only if  $\|\alpha\|_0 = K$ .*

(b) *If  $\text{rank}(S) < d - 1$ , there exists a (minimal)  $K$ -element frame  $\mathcal{F}$  such that  $S = \Theta_{\mathcal{F}} \Theta_{\mathcal{F}}^*$  and  $(f_i, f_i) = \alpha[i]$  for every  $i$ .*

*Proof.* In case (1a), Theorem 4.5 implies the existence of a  $K$ -element frame  $\mathcal{F}$ , where  $K = d$ , whose frame operator is  $S$ . The  $d$  columns of  $\Theta_{\mathcal{F}}$  must be linearly independent, so every  $f_i$  must be odd, by Lemma 4.11. (As a corollary of Theorem 4.14, we note that  $d \equiv \text{Tr}(S) \pmod{2}$  for any  $d \times d$ , full rank, parity indicative symmetric matrix  $S$  with  $\text{diag}(S) \neq \vec{0}$ .)

For case (1b), instead of using the result of Theorem 4.9, we rely directly on Theorem 4.3 to construct a  $d \times \text{rank}(S)$  matrix  $A$  with  $\text{rank}(A) = \text{rank}(S)$  such that  $AA^* = S$ . Since the columns of  $A$  are linearly independent, Lemma 4.11 implies that they are all odd. As in the proof of Theorem 4.9, we consider an augmented matrix

$$\Theta_{\mathcal{F}} = [A \ B]$$

but with more care taken in the choice of  $B$ . By letting the columns of  $B$  be  $d - \text{rank}(S)$  of the standard basis vectors not in the column space of  $A$ , each repeated twice, we construct a frame  $\mathcal{F}$  for  $\|\alpha\|_0 = 2d - \text{rank}(S)$ . Replacing any identical pair of columns of  $B$ , say  $\{\varepsilon_l, \varepsilon_l\}$ , with  $\{\varepsilon_l + \varepsilon_n, \varepsilon_l + \varepsilon_n\}$  for any other basis vector  $\varepsilon_n \neq \varepsilon_l$ , the columns of  $\Theta_{\mathcal{F}}$  still span,  $\Theta_{\mathcal{F}} \Theta_{\mathcal{F}}^*$  is still equal to  $S$ , but now  $\mathcal{F}$  contains two fewer odd vectors. In this way, we are able to construct a frame  $\mathcal{F}$  with dot products  $(f_i, f_i)$  satisfying any  $\|\alpha\|_0 = \text{rank}(S) + 2m$  for  $0 \leq m \leq d - \text{rank}(S)$ . (Note that, by the proof of Theorem 4.14,  $\text{rank}(S) \equiv \text{Tr}(S) \pmod{2}$ .) By Lemma 4.13, any four vectors consisting of three odds and one even can be substituted by four vectors consisting of three evens and one odd, having the same span and no effect on  $\Theta_{\mathcal{F}} \Theta_{\mathcal{F}}^*$ . Each substitution allows us to increase the number of even vectors by

two, until only two odd vectors remain in  $\mathcal{F}$  if  $\text{rank}(S)$  is even or one odd vector remains if  $\text{rank}(S)$  is odd. Therefore, we can build a frame with  $2d - \text{rank}(S)$  elements, corresponding to any nonzero  $\alpha$  with  $\|\alpha\|_0 \equiv \text{rank}(S) \equiv \text{Tr}(S) \pmod{2}$ .

Now let  $S$  be parity indicative with  $\text{diag}(S) = \vec{0}$ . Since every row of  $S$  is even,  $\text{rank}(S) < d$ . In case (2a), Theorem 4.9 implies the existence of a  $(d+1)$ -element frame  $\mathcal{F}$  whose frame operator is  $S$ . Since  $\mathcal{F}$  is a spanning set, it must contain an odd vector. By Lemma 4.12, every vector in  $\mathcal{F}$  must be odd. (Note that, by Corollary 4.8, case (2a) can only occur if  $\text{rank}(S)$  is even, and hence  $d$  is odd.)

Lastly, we assume in case (2b) that  $\text{rank}(S) \leq d - 2$ . By Theorem 4.3 and Proposition 4.4, there exists a  $d \times (\text{rank}(S) + 1)$  matrix  $A$  with  $\text{rank}(A) = \text{rank}(S)$  such that  $AA^* = S$ . By Lemma 4.12, either every column of  $A$  is even or every column is odd. Since  $S_{ii} = 0$  for every  $i$ , every row of  $A$  is even, and hence the sum of the columns of  $A$  is  $\vec{0}$ . Moreover, since  $\text{diag}(S) = \vec{0}$ , we know that  $\text{rank}(S)$  must be even, by Corollary 4.8. If every column of  $A$  were odd, then the sum of all  $\text{rank}(S) + 1$  columns would have to be odd, by Lemma 4.2, yielding a contradiction. So every column of  $A$  must be even.

Augment  $A$  with a column of zeros and call the resulting matrix  $B$ . Then each column of  $B$  is even, each row of  $B$  is even, and  $B$  has an even number of columns. Consider a row  $b_n^*$  of  $B$  such that  $b_n^* \in \text{Span}\{b_j^* : b_j^* \text{ is a row of } B \text{ and } j \neq n\}$ . Replace  $b_n^*$  by its complement (that is, add  $\vec{1}^*$  for  $\vec{1} \in \mathbb{Z}_2^{\text{rank}(S)+2}$  to  $b_n^*$ ), and call the resulting matrix  $C$ . Then  $CC^* = S$ ,  $\text{rank}(C) = \text{rank}(S) + 1$ , and  $C$  is composed of  $\text{rank}(S) + 2$  odd columns. As in case (1b), we now augment  $C$  with  $d - (\text{rank}(S) + 1)$  of the standard basis vectors not in the column space of  $C$ , each repeated twice, to construct  $\Theta_{\mathcal{F}}$ . In doing so, we construct a frame  $\mathcal{F}$  consisting of  $\text{rank}(S) + 2 + 2(d - (\text{rank}(S) + 1)) = 2d - \text{rank}(S)$  vectors, with frame operator  $S$ , such that every element of  $\mathcal{F}$  is odd. By Theorem 4.14,  $\|\alpha\|_0$  must be even. As in case (1b), we can replace pairs of odd elements of  $\mathcal{F}$  by even vectors until only two odd vectors remain. □

**Corollary 4.16.** *Let  $S$  be a  $d \times d$  parity indicative symmetric matrix. Let  $K > 2d - \text{rank}(S)$ . Let  $\alpha \in \mathbb{Z}_2^K$  be a nonzero vector such that  $\|\alpha\|_0 \equiv \text{Tr}(S) \pmod{2}$ . Then there exists a  $K$ -element frame  $\mathcal{F}$  such that  $S = \Theta_{\mathcal{F}} \Theta_{\mathcal{F}}^*$  and  $(f_i, f_i) = \alpha[i]$  for every  $i$ .*

*Proof.* Since  $S$  is parity indicative, a  $K$ -element frame with  $K > 2d - \text{rank}(S)$  is necessarily nonminimal and can be constructed by augmenting the minimal frames of the previous theorem. Consider first the minimal frame  $\mathcal{F}$  guaranteed by case (1a) of Theorem 4.15. Adding the zero vector to  $\mathcal{F}$  allows us to apply Lemma 4.13 and create frames satisfying  $(f_i, f_i) = \alpha[i]$  for any  $\|\alpha\|_0 \equiv \text{Tr}(S) \pmod{2}$  with  $0 < \|\alpha\|_0 < d$ . Similarly, for case (2a), including the zero vector allows the construction of a frame corresponding to any  $\|\alpha\|_0 = 2, 4, 6, \dots, d + 1$ . In either



case, the addition of two identical copies of odd vectors or two identical copies of even vectors provides frames for any  $\|\alpha\|_0 \equiv \text{Tr}(S) \pmod{2}$  when  $\|\alpha\|_0 \geq d + 2$  or  $\|\alpha\|_0 \geq d + 3$ , corresponding to cases (1a) and (2a), respectively. Similarly in cases (1b) and (2b), the addition of two identical copies of an odd vector or two identical copies of an even vector yield frames for  $2d - \text{rank}(S) < \|\alpha\|_0$ .  $\square$

**Theorem 4.17.** *Let  $S$  be a  $d \times d$  symmetric matrix that is not parity indicative. Let  $K \geq 2d - \text{rank}(S)$  or if  $\text{rank}(S) = d$  and  $\text{diag}(S) = \vec{0}$ , let  $K \geq d + 1$ . Let  $\alpha \in \mathbb{Z}_2^K$  be a nonzero vector such that  $\|\alpha\|_0 \equiv \text{Tr}(S) \pmod{2}$ . Then there exists a  $K$ -element frame  $\mathcal{F}$  such that  $S = \Theta_{\mathcal{F}} \Theta_{\mathcal{F}}^*$  and  $(f_i, f_i) = \alpha[i]$  for every  $i$  only if  $\|\alpha\|_0 \neq K$ .*

*Proof.* We use Theorem 4.5 and Remark 4.6 or Theorem 4.9 to construct a  $K$ -element frame  $\mathcal{F}$  such that  $\Theta_{\mathcal{F}} \Theta_{\mathcal{F}}^* = S$ . By Lemma 4.11,  $\mathcal{F}$  must contain an even vector. Of course,  $\mathcal{F}$  must also contain an odd vector, in order to span. Let  $m \equiv \text{Tr}(S) \pmod{2}$  represent the number of odd elements of  $\mathcal{F}$  and  $K - m$  be the number of even elements. By Lemma 4.13,  $\mathcal{F}$  may be replaced by a frame with two more or two fewer odd vectors. Through repeated applications, we can construct a frame  $\mathcal{F}$  corresponding to any  $\|\alpha\|_0 = 1, 3, 5, \dots, K - 1$  if  $m$  is odd and  $K$  is even, any  $\|\alpha\|_0 = 1, 3, 5, \dots, K - 2$  if  $m$  is odd and  $K$  is odd, any  $\|\alpha\|_0 = 2, 4, 6, \dots, K - 1$  if  $m (\geq 2)$  is even and  $K (> m)$  is odd, or any  $\|\alpha\|_0 = 2, 4, 6, \dots, K - 2$  if  $m (\geq 2)$  is even and  $K (> m)$  is even.  $\square$

### 5. Examples and data

**Examples.** In this subsection we consider two symmetric matrices  $S$  and build frames with various  $\alpha$ 's to illustrate the main result of Section 4. The algorithm for factorizing a matrix as  $S = AA^*$  and for reducing  $A$  into a minimal factor can be found in [Lempel 1975].

**Example 5.1.** Consider the identity matrix

$$S = I_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

and note that it is a symmetric, parity indicative, full-rank matrix. Any frame with frame operator  $S$  is a Parseval frame. By Theorem 4.15, a minimal 4-element such Parseval frame must satisfy  $\|\alpha\|_0 = 4$ , where  $\alpha[i] = (f_i, f_i)$  for each  $i$ ; clearly, this follows from the Parseval frame necessarily being an orthonormal basis. Corollary 4.16 guarantees Parseval frames in  $\mathbb{Z}_2^4$  of length  $K = 5$  with either two or four odd vectors corresponding to any  $\alpha \in \mathbb{Z}_2^5$  with  $\|\alpha\|_0 = 2, 4$ . To begin the construction, factor  $S$  as  $I_4 I_4^*$ . Appending the zero-column to the left factor yields the matrix

$$\Theta_{\mathcal{F}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix},$$

the columns of which constitute a frame with frame operator  $S$  and  $\alpha = (1, 1, 1, 1, 0)$ . To obtain any other  $\alpha \in \mathbb{Z}_2^5$  with  $\|\alpha\|_0 = 4$ , simply permute the columns.

We utilize Lemma 4.13 to reduce the number of odd vectors by two. Let  $e$  be the zero-column and  $o_1, o_2, o_3$  be the first, second, and third columns of  $\Theta_{\mathcal{F}}$ , respectively. Replacing  $e, o_1, o_2, o_3$  with their counterparts constructed in Lemma 4.13 results in

$$\Theta_{\mathcal{F}'} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

Taking the columns of  $\Theta_{\mathcal{F}'}$  as frame vectors builds the frame  $\mathcal{F}'$  satisfying  $\alpha = (0, 0, 0, 1, 1)$ . Again, the columns of  $\mathcal{F}'$  can be permuted to acquire any  $\alpha \in \mathbb{Z}_2^5$  with  $\|\alpha\|_0 = 2$ . Notice that a permutation of  $\mathcal{F}'$  appears in the proof of Theorem 3.5.

**Example 5.2.** Suppose we wish to find a frame for  $\mathbb{Z}_2^3$  of length 7 with frame operator

$$S = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

Since this rank-2, symmetric matrix is not parity indicative, we apply Theorem 4.17. In doing so, we follow the proof of Theorem 4.9 and factorize  $S$  as

$$S = P^* T \begin{bmatrix} L & 0 \\ 0 & 0 \end{bmatrix} T^* P,$$

where  $P$  is the  $3 \times 3$  identity matrix,

$$L = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}, \quad \text{and} \quad T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}.$$

Then

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

and we append three zero-columns to  $TA$  to build the frame  $\mathcal{F}$ :

$$\Theta_{\mathcal{F}} = [TA \ \vec{0} \ \vec{0} \ \vec{0}] = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 \end{bmatrix}.$$

Letting  $(f_i, f_i) = \alpha[i]$  for each  $i$ , we see that  $\mathcal{F}$  satisfies  $\alpha = (1, 0, 1, 1, 0, 0, 0)$ . We increase or decrease the number of odd vectors as desired, by applying Lemma 4.13 first to  $\{f_1, f_3, f_4, f_5\}$  and then to  $\{f_4, f_5, f_6, f_7\}$ . We obtain frames  $\mathcal{F}_1$  and  $\mathcal{F}_2$  satisfying

$$\Theta_{\mathcal{F}_1} = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}, \quad \alpha_1 = (0, 0, 0, 0, 1, 0, 0);$$

$$\Theta_{\mathcal{F}_2} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}, \quad \alpha_2 = (1, 0, 1, 0, 1, 1, 1).$$

By Theorem 4.17,  $\|\alpha\|_0 = 7$  is unattainable.

**Data.** An exhaustive search for frame operators  $\Theta_{\mathcal{F}} \Theta_{\mathcal{F}}^*$  and  $\|\alpha\|_0$  associated with  $\mathcal{F} = \{f_j\}_{j=1}^K$  in  $\mathbb{Z}_2^d$  was performed, using Python 3.6, for various dimensions and frame lengths (i.e., various  $d$ 's and  $K$ 's). The tables contained in this subsection hold information about the number of symmetric matrices that are frame operators and the set of  $\|\alpha\|_0$  that accompany them. We include summaries for dimensions  $d = 2, \dots, 5$ . Because every frame in  $\mathbb{Z}_2^d$  must have at least  $d$  vectors, and because  $2d$  is the minimum number of vectors needed to ensure every symmetric matrix is a frame operator (Theorems 4.5, 4.9), the computations were performed for  $K = d, \dots, 2d$ .

For  $d = 2, \dots, 5$ , in the table containing information about the  $d$ -dimensional binary space, the entry in the row labeled  $\{\alpha_{\min}, \alpha_{\min} + 2, \dots, \alpha_{\min} + 2t\}$  and column labeled  $K = k_0$  shows the number of symmetric matrices  $S$  in  $d$ -dimensional space such that for each  $\alpha \in \{\alpha_{\min}, \alpha_{\min} + 2, \dots, \alpha_{\min} + 2t\}$  there exists a frame  $\mathcal{F} = \{f_j\}_{j=1}^{k_0}$  such that  $S = \Theta_{\mathcal{F}} \Theta_{\mathcal{F}}^*$  and  $\alpha[i] = (f_i, f_i)$  for  $1 \leq i \leq k_0$ .

In each table, the sum of the entries of the last column represents all possible symmetric  $d \times d$  binary matrices. There are  $2^{d(d+1)/2}$  such matrices, which becomes prohibitively large as the dimension  $d$  increases.

$\{\ \alpha\ _0\}$	$K$		
	2	3	4
{1}	2	2	0
{2}	1	3	2
{1, 3}	0	2	4
{2, 4}	0	0	2

**Table 1.** Number of attainable frame operators of frames for  $\mathbb{Z}_2^2$ .

$\{\ \alpha\ _0\}$	$K$			
	3	4	5	6
{1}	12	0	0	0
{2}	12	21	0	0
{3}	4	0	0	0
{4}	0	1	0	0
{1, 3}	0	28	24	0
{2, 4}	0	6	31	24
{1, 3, 5}	0	0	8	32
{2, 4, 6}	0	0	0	8

**Table 2.** Number of attainable frame operators of frames for  $\mathbb{Z}_2^3$ .

$\{\ \alpha\ _0\}$	$K$				
	4	5	6	7	8
{2}	168	0	0	0	0
{4}	28	0	0	0	0
{1, 3}	224	392	0	0	0
{2, 4}	0	420	441	0	0
{1, 3, 5}	0	56	504	448	0
{2, 4, 6}	0	0	63	511	448
{1, 3, 5, 7}	0	0	0	64	512
{2, 4, 6, 8}	0	0	0	0	64

**Table 3.** Number of attainable frame operators of frames for  $\mathbb{Z}_2^4$ .

$\{\ \alpha\ _0\}$	$K$					
	5	6	7	8	9	10
{5}	448	0	0	0	0	0
{6}	0	28	0	0	0	0
{1, 3}	6720	0	0	0	0	0
{2, 4}	6720	13020	0	0	0	0
{1, 3, 5}	0	13888	15120	0	0	0
{2, 4, 6}	0	840	15988	15345	0	0
{1, 3, 5, 7}	0	0	1008	16368	15360	0
{2, 4, 6, 8}	0	0	0	1023	16383	15360
{1, 3, 5, 7, 9}	0	0	0	0	1024	16384
{2, 4, 6, 8, 10}	0	0	0	0	0	1024

**Table 4.** Number of attainable frame operators of frames for  $\mathbb{Z}_2^5$ .

## Acknowledgement

We thank Erich McAlister for his very valuable feedback.

## References

- [Baker et al. 2018] Z. J. Baker, B. G. Bodmann, M. G. Bullock, S. N. Branum, and J. E. McLaney, “What is odd about binary Parseval frames?”, *Involve* **11**:2 (2018), 219–233.
- [Bodmann et al. 2009] B. G. Bodmann, M. Le, L. Reza, M. Tobin, and M. Tomforde, “Frame theory for binary vector spaces”, *Involve* **2**:5 (2009), 589–602. MR Zbl
- [Bodmann et al. 2014] B. G. Bodmann, B. Camp, and D. Mahoney, “Binary frames, graphs and erasures”, *Involve* **7**:2 (2014), 151–169. MR Zbl
- [Bownik and Jasper 2015] M. Bownik and J. Jasper, “Existence of frames with prescribed norms and frame operator”, pp. 103–117 in *Excursions in harmonic analysis, IV*, edited by R. Balan et al., Birkhäuser, Cham, 2015. MR
- [Cahill et al. 2013] J. Cahill, M. Fickus, D. G. Mixon, M. J. Poteet, and N. Strawn, “Constructing finite frames of a given spectrum and set of lengths”, *Appl. Comput. Harmon. Anal.* **35**:1 (2013), 52–73. MR Zbl
- [Casazza and Kutyniok 2013] P. G. Casazza and G. Kutyniok (editors), *Finite frames: theory and applications*, Birkhäuser, New York, 2013. MR Zbl
- [Casazza and Leon 2010] P. G. Casazza and M. T. Leon, “Existence and construction of finite frames with a given frame operator”, *Int. J. Pure Appl. Math.* **63**:2 (2010), 149–157. MR Zbl
- [Casazza et al. 2006] P. G. Casazza, M. Fickus, J. Kovačević, M. T. Leon, and J. C. Tremain, “A physical interpretation of tight frames”, pp. 51–76 in *Harmonic analysis and applications*, edited by C. Heil, Birkhäuser, Boston, 2006. MR Zbl
- [Christensen 2003] O. Christensen, *An introduction to frames and Riesz bases*, Birkhäuser, Boston, 2003. MR Zbl
- [Christensen et al. 2012] O. Christensen, A. M. Powell, and X. C. Xiao, “A note on finite dual frame pairs”, *Proc. Amer. Math. Soc.* **140**:11 (2012), 3921–3930. MR Zbl
- [Cohen et al. 1997] G. Cohen, I. Honkala, S. Litsyn, and A. Lobstein, *Covering codes*, North-Holland Mathematical Library **54**, North-Holland, Amsterdam, 1997. MR Zbl
- [Fickus et al. 2013] M. Fickus, D. G. Mixon, M. J. Poteet, and N. Strawn, “Constructing all self-adjoint matrices with prescribed spectrum and diagonal”, *Adv. Comput. Math.* **39**:3-4 (2013), 585–609. MR Zbl
- [Han et al. 2007] D. Han, K. Kornelson, D. Larson, and E. Weber, *Frames for undergraduates*, Student Mathematical Library **40**, American Mathematical Society, Providence, 2007. MR Zbl
- [Horn and Johnson 1985] R. A. Horn and C. R. Johnson, *Matrix analysis*, Cambridge Univ. Press, 1985. MR Zbl
- [Hotovy et al. 2015] R. Hotovy, D. R. Larson, and S. Scholze, “Binary frames”, *Houston J. Math.* **41**:3 (2015), 875–899. MR Zbl
- [Kovačević and Chebira 2007] J. Kovačević and A. Chebira, “Life beyond bases: the advent of frames, II”, *IEEE Signal Proc. Mag.* **24**:5 (2007), 115–125.
- [Lempel 1975] A. Lempel, “Matrix factorization over  $\text{GF}(2)$  and trace-orthogonal bases of  $\text{GF}(2^n)$ ”, *SIAM J. Comput.* **4**:2 (1975), 175–186. MR Zbl

Received: 2017-05-24

Accepted: 2017-07-17

furst\_v@fortlewis.edu

*Department of Mathematics, Fort Lewis College,  
Durango, CO, United States*

eric.powell.smith@gmail.com

*Department of Mathematics, Fort Lewis College,  
Durango, CO, United States*

## Guidelines for Authors

Submissions in all mathematical areas are encouraged. All manuscripts accepted for publication in *Involve* are considered publishable in quality journals in their respective fields, and include a minimum of one-third student authorship. Submissions should include substantial faculty input; faculty co-authorship is strongly encouraged. Authors may submit manuscripts in PDF format on-line at the Submission page at the Involve website.

**Originality.** Submission of a manuscript acknowledges that the manuscript is original and is not, in whole or in part, published or under consideration for publication elsewhere. It is understood also that the manuscript will not be submitted elsewhere while under consideration for publication in this journal.

**Language.** Articles in *Involve* are usually in English, but articles written in other languages are welcome.

**Required items.** A brief abstract of about 150 words or less must be included. It should be self-contained and not make any reference to the bibliography. If the article is not in English, two versions of the abstract must be included, one in the language of the article and one in English. Also required are keywords and subject classifications for the article, and, for each author, postal address, affiliation (if appropriate), and email address.

**Format.** Authors are encouraged to use L<sup>A</sup>T<sub>E</sub>X but submissions in other varieties of T<sub>E</sub>X, and exceptionally in other formats, are acceptable. Initial uploads should be in PDF format; after the refereeing process we will ask you to submit all source material.

**References.** Bibliographical references should be complete, including article titles and page ranges. All references in the bibliography should be cited in the text. The use of BibT<sub>E</sub>X is preferred but not required. Tags will be converted to the house format, however, for submission you may use the format of your choice. Links will be provided to all literature with known web locations and authors are encouraged to provide their own links in addition to those supplied in the editorial process.

**Figures.** Figures must be of publication quality. After acceptance, you will need to submit the original source files in vector graphics format for all diagrams in your manuscript: vector EPS or vector PDF files are the most useful.

Most drawing and graphing packages (Mathematica, Adobe Illustrator, MATLAB, etc.) allow the user to save files in one of these formats. Make sure that what you are saving is vector graphics and not a bitmap. If you need help, please write to [graphics@msp.org](mailto:graphics@msp.org) with details about how your graphics were generated.

**White space.** Forced line breaks or page breaks should not be inserted in the document. There is no point in your trying to optimize line and page breaks in the original manuscript. The manuscript will be reformatted to use the journal's preferred fonts and layout.

**Proofs.** Page proofs will be made available to authors (or to the designated corresponding author) at a Web site in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

# involve

2018 vol. 11 no. 3

A mathematical model of treatment of cancer stem cells with immunotherapy	361
ZACHARY J. ABERNATHY AND GABRIELLE EPELLE	
RNA, local moves on plane trees, and transpositions on tableaux	383
LAURA DEL DUCA, JENNIFER TRIPP, JULIANNA TYMOCZKO AND JUDY WANG	
Six variations on a theme: almost planar graphs	413
MAX LIPTON, EOIN MACKALL, THOMAS W. MATTMAN, MIKE PIERCE, SAMANTHA ROBINSON, JEREMY THOMAS AND ILAN WEINSCHELBAUM	
Nested Frobenius extensions of graded superrings	449
EDWARD POON AND ALISTAIR SAVAGE	
On $G$ -graphs of certain finite groups	463
MOHAMMAD REZA DARAFSHEH AND SAFOORA MADADY MOGHADAM	
The tropical semiring in higher dimensions	477
JOHN NORTON AND SANDRA SPIROFF	
A tale of two circles: geometry of a class of quartic polynomials	489
CHRISTOPHER FRAYER AND LANDON GAUTHIER	
Zeros of polynomials with four-term recurrence	501
KHANG TRAN AND ANDRES ZUMBA	
Binary frames with prescribed dot products and frame operator	519
VERONIKA FURST AND ERIC P. SMITH	

