

# involve

a journal of mathematics

## Editorial Board

Kenneth S. Berenhaut, *Managing Editor*

Colin Adams  
John V. Baxley  
Arthur T. Benjamin  
Martin Bohner  
Nigel Boston  
Amarjit S. Budhiraja  
Pietro Cerone  
Scott Chapman  
Jem N. Corcoran  
Toka Diagana  
Michael Dorff  
Sever S. Dragomir  
Behrouz Emamizadeh  
Joel Foisy  
Errin W. Fulp  
Joseph Gallian  
Stephan R. Garcia  
Anant Godbole  
Ron Gould  
Andrew Granville  
Jerrold Griggs  
Sat Gupta  
Jim Haglund  
Johnny Henderson  
Jim Hoste  
Natalia Hritonenko  
Glenn H. Hurlbert  
Charles R. Johnson  
K. B. Kulasekera  
Gerry Ladas

David Larson  
Suzanne Lenhart  
Chi-Kwong Li  
Robert B. Lund  
Gaven J. Martin  
Mary Meyer  
Emil Minchev  
Frank Morgan  
Mohammad Sal Moslehian  
Zuhair Nashed  
Ken Ono  
Timothy E. O'Brien  
Joseph O'Rourke  
Yuval Peres  
Y.-F. S. Pétermann  
Robert J. Plemmons  
Carl B. Pomerance  
Bjorn Poonen  
József H. Przytycki  
Richard Rebarber  
Robert W. Robinson  
Filip Saidak  
James A. Sellers  
Andrew J. Sterge  
Ann Trenk  
Ravi Vakil  
Antonia Vecchio  
Ram U. Verma  
John C. Wierman  
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 2017 is US \$175/year for the electronic version, and \$235/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/>

© 2017 Mathematical Sciences Publishers

# Algorithms for finding knight's tours on Aztec diamonds

Samantha Davies, Chenxiao Xue and Carl R. Yerger

(Communicated by Ronald Gould)

A knight's tour is a sequence of knight's moves such that each square on the board is visited exactly once. An Aztec diamond is a square board of size  $2n$  where triangular regions of side length  $n - 1$  have been removed from all four corners.

We show that the existence of knight's tours on Aztec diamonds cannot be proved inductively via smaller Aztec diamonds, and explain why a divide-and-conquer approach is also not promising. We then describe two algorithms that aim to efficiently find knight's tours on Aztec diamonds. The first is based on random walks, a straightforward but limited technique that yielded tours on Aztec diamonds for all  $n \neq 22$  apart from  $n = 17, 21$ . The second is a path-conversion algorithm that finds a solution for all  $n \leq 100$ . We then apply the path-conversion algorithm to random graphs to test the robustness of our algorithm. Online supplements provide source code, output and more details about these algorithms.

## 1. Introduction

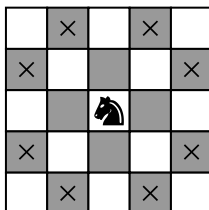
The problem of finding a knight's tour is one of many classes of chess-related problems that have been studied for hundreds of years. An early instance of such a tour was described by al-Adli ar-Rumi from Baghdad around the year 840 [Murray 1913]. Euler [1759; 1782] also studied the problem.

In chess, a *knight's move* on a board moves the piece horizontally by one square and vertically by two squares, or horizontally by two squares and vertically by one square. For clarity, Figure 1 indicates the possible moves of a knight. A *knight's tour* is defined to be a sequence of knight's moves on a board such that the sequence hits every square on the board exactly once. In an *open knight's tour*, there is no restriction on the starting and ending squares, whereas in a *closed knight's tour*, the starting square has to be one knight's move away from the ending square. We will focus mainly on closed knight's tours; if not stated otherwise, a knight's tour will

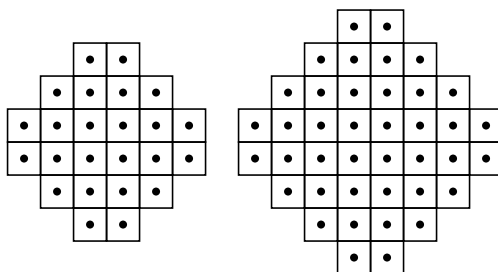
---

*MSC2010:* primary 05C45; secondary 05C57, 97A20.

*Keywords:* knight's tour, Aztec diamond, Hamiltonian, algorithm.



**Figure 1.** Possible knight's moves.



**Figure 2.** Aztec diamonds of radius 3 (left) and radius 4 (right).

refer to a closed knight's tour. We use the term *partial knight's tour* for a tour that visits squares only once, but not necessarily all of them.

Conditions for boards on which a knight's tour exists have been published for rectangular boards and variations thereof such as cylindrical chessboards [Watkins and Hoenigman 1997], toroidal chessboards [Watkins 2000], spherical (pillow) chessboards [Cairns 2002], and boards with deleted squares [Bi et al. 2015; Demaio and Hippchen 2009; Miller and Farnsworth 2013]. Most proofs of the existence of knight's tours on these types of boards involve the expander method, made popular by Schwenk [1991]. With this method, one can take an open knight's tour on a board, add small strips of squares to extend the tour, and then use rotation, symmetries and induction to make a closed knight's tour on a new board.

One board that cannot be constructed by identifying edges of a regular chessboard is the Aztec diamond. We define an *Aztec diamond of radius  $n$*  as a lattice of squares in the  $\mathbb{Z}^2$  coordinate system, whose centers  $(x, y)$  satisfy  $|x| + |y| \leq n$  (these centers are composed of half-integral coordinates). Figure 2 shows Aztec diamonds of radii 3 and 4, with black dots representing the centers of those squares. These black dots are included because each square of the Aztec diamond chessboard corresponds to a vertex in the associated Aztec diamond graph. Two vertices in this associated graph are adjacent if their corresponding squares can be reached via a single knight's move.

In Section 2, we show that an Aztec diamond cannot be partitioned into smaller Aztec diamonds, which suggests that an inductive approach to finding knight's tours

on an Aztec diamond is likely to fail. The symmetry of an Aztec diamond makes a divide-and-conquer algorithm appealing but we find that it is extremely difficult and sometimes impossible to divide an Aztec diamond. In Section 3, we introduce two algorithms for computing knight's tours (hamiltonian cycles): the *random-walk algorithm* and the *path-conversion algorithm*. The random-walk algorithm is deterministic, which means it will determine whether a knight's tour exists or not on a certain Aztec diamond upon completion. Our path-conversion algorithm is nondeterministic and hence cannot be used to disprove the existence of knight's tours on an Aztec diamond. But for Aztec diamonds, this algorithm is much more efficient than the random-walk algorithm in finding a knight's tour. In Section 4, we discuss some possible improvements on the path-conversion algorithm and several open problems in finding knight's tours on Aztec diamonds.

## 2. Theoretical results

**Lemma 1.** *The length of any closed knight's tour is even.*

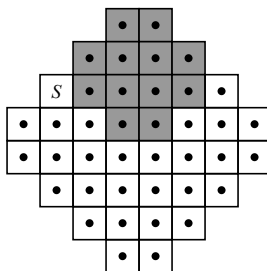
*Proof.* First, we color the board with two colors, say black and white. Two squares are colored differently if they share a boundary. Hence, as shown via Figure 1, a knight can only move to a square that has a different color from its current square. In a closed knight's tour, the last visited square must be adjacent to the starting square, which means they are colored differently. Because the color alternates for every move, the number of squares visited in a closed knight's tour must be even.  $\square$

**Lemma 2.** *An Aztec diamond cannot be dissected into several smaller Aztec diamonds.*

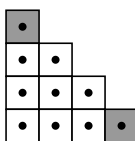
*Proof.* We define the *border* of an Aztec diamond to be the set of squares that have a boundary that is not shared by two squares. By observation, the degree of a vertex on the border of any Aztec diamond can be 3, 4, 6 or 8. However, if we try to cover a top square of an Aztec diamond with a smaller one, we will obtain a square  $s$ , a degree 2 square, on the border of the remaining uncovered graph. Hence, this square cannot be a part of an Aztec diamond, which means that no Aztec diamond can be dissected into smaller Aztec diamonds. Figure 3 is a graphical representation of the arguments made in this proof.  $\square$

A *quadrant* of an Aztec diamond consists of squares whose centers are in the same quadrant of the coordinate system. For example, the first quadrant of an Aztec diamond of radius 4 is shown in Figure 4.

**Lemma 3.** *No knight's tours can be found in a quadrant of an Aztec diamond. In addition, an Aztec diamond of radius  $n$  cannot be dissected into four closed partial knight's tours of the same length, if  $n \equiv 1, 2 \pmod{4}$ .*



**Figure 3.** Visual proof of Lemma 2.



**Figure 4.** First quadrant of an Aztec diamond of radius 4.

*Proof.* The analogue of the two shaded squares in Figure 4 in a quadrant of an arbitrarily sized Aztec diamond have degree 1 and thus cannot be part of a closed knight's tour. Now suppose that  $n = 4k + 1$  or  $n = 4k + 2$  for some  $k \in \mathbb{Z}$ . The total number of squares in such an Aztec diamond is  $n \cdot (n + 1) \cdot 2$ . If it can be dissected into four closed partial knight's tours of the same length, each tour is of length  $\frac{1}{2}n \cdot (n + 1)$ , which can be expressed as  $(4k + 1) \cdot (2k + 1)$  or  $(2k + 1) \cdot (4k + 3)$ , in contradiction with the fact that a closed knight's tour must be even in length by Lemma 1. This completes the proof.  $\square$

### 3. The random-walk algorithm and its variants

Since finding a knight's tour is essentially finding a hamiltonian cycle, we first introduce a brute-force algorithm for finding hamiltonian cycles called the random-walk algorithm. Again, these algorithms are run on the graphs associated with knight's tour moves on the Aztec diamond as described in Section 1. We present this algorithm on a nondeterministic machine because it is more succinct than the deterministic version. Whether a graph has a hamiltonian cycle or not can be determined by the following nondeterministic machine:

$N =$  On input  $\langle G \rangle$ , where  $G$  is a graph:

1. If  $G$  has only one vertex, accept. Else, proceed to Step 2.
2. Pick an arbitrary vertex  $v$  of  $G$  as the starting vertex. Mark  $v$  as visited.
3. Nondeterministically choose an unvisited vertex that is adjacent to the last marked vertex and mark it as visited. If none can be marked, reject.

4. If there exists an unvisited vertex, go to Step 2. Else, proceed to Step 5.
5. If the last marked vertex is adjacent to  $v$ , accept. Otherwise, reject.

To transform this algorithm into its deterministic form, we have to try every possibility when picking a vertex as described in Step 3, and backtrack if the chosen vertex fails to produce a hamiltonian cycle. In order to make this algorithm run faster, we add the following rules:

1. If at any point the starting vertex has no unvisited neighbors but the graph still has unvisited vertices, abort the current branch.
2. If a vertex  $v$  is adjacent to the last marked vertex and has only one unmarked neighbor, choose  $v$  to be the next marked vertex (note: there is a chance that  $v$  is the end of the tour, but this requires that  $v$  is adjacent to the starting vertex and that this situation has not been encountered in the previous steps). If there is more than one vertex that has this property, abort the current branch.

Although these improvements do not affect the asymptotic running time of this algorithm, they do expedite the process significantly in practice. Unfortunately, since the size of this problem grows exponentially as the radius increases, the running time of completing this improved algorithm on a large graph is astronomical. Our implementation of the random-walk algorithm runs for over one week on a Macbook Pro (2GHz Intel Core i7, 4GB memory, 1333MHz DDR3) without completion on an Aztec diamond of radius 5. However, we discovered that the choice of the starting vertex in Step 2 will tremendously affect the time used to find a knight's tour. For example, if we start with a certain vertex on an Aztec diamond of radius 4, the algorithm could run for more than ten hours without giving us a result, whereas with the right choice of starting vertex, we might obtain a cycle in ten seconds. This finding leads us to the next version of the random-walk algorithm.

To modify the existing algorithm, we simply run the algorithm with a given starting vertex on a separate thread. If no result is obtained within a certain amount of time, say ten seconds, we switch to the next starting vertex. The program halts if we find a cycle or if all the vertices have been chosen. An obvious flaw of this algorithm is that it is no longer deterministic because it does not exhaust all possible cycles.

When using this algorithm in practice, we are able to find knight's tours on Aztec diamonds of radius 22 or less except for those of radius 17 and 21, as shown in the table supplement. We could theoretically prolong the search time of each thread to increase the chance of obtaining a hamiltonian cycle, but given the size of our graphs, even an extension of five seconds per thread would lead to an increase of one hour in the total search time. We stop at radius 22 because the amount of time used to complete a search exceeds three hours.

#### 4. The path-conversion algorithm

The idea of this algorithm springs from Parberry [1997]. In this paper, the author divides a large rectangular board into smaller rectangular boards (usually into four pieces), finds structured knight's tours on those smaller boards, and uses these special structures to connect all the disjoint partial tours together. The exact same technique would fail to generate knight's tours on the Aztec diamond because an Aztec diamond cannot be dissected into Aztec diamonds of smaller radii by Lemma 2. The attempt of cutting an Aztec diamond into four equally sized pieces does not seem to be feasible based on our proof when  $n \equiv 1, 2 \pmod{4}$  in Lemma 3.

Now we present our new algorithm, and details about how each step is accomplished will follow in the subsequent paragraphs. The algorithm does the following things:

1. Find an open knight's tour on an Aztec diamond.
2. Cut this open knight's tour into disjoint closed knight's tours.
3. Connect these closed knight's tours together.
4. If there exists a knight's tour that cannot be connected to the rest of the knight's tours, reject. Otherwise, accept.

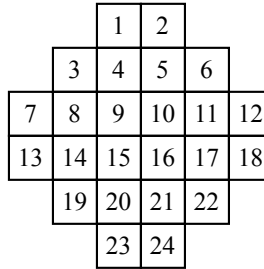
*Step 1: Find an open knight's tour on an Aztec diamond.* In theory, finding an open knight's tour on an Aztec diamond is as hard as finding a closed one since both are NP-complete [West 2001]. In practice, the time spent finding an open knight's tour is significantly less than the time spent finding a closed one because there are more open knight's tours than closed knight's tours. We could still use the random-walk algorithm described earlier without the final checking step (Step 5) to find an open knight's tour. But for the sake of efficiency, we decide to use Warnsdorff's [1823] heuristic rule (as used in [Ganzfried 2004]) to speed up the process. As Ganzfried pointed out, Warnsdorff's rule does not hold true for every open knight's tour and it fails more regularly when the size of the graphs increases. But for the scope of this paper, Warnsdorff's rule (with some slight modifications) proves to be successful.

**Warnsdorff's Rule** [von Warnsdorff 1823]. In picking the next move, always pick an adjacent, unvisited square that has the least number of unvisited neighbors.

There are different rules about tie-breaking if two unvisited squares have the same amount of unvisited neighbors, but we simply use an ordering system of the squares to break ties. We number vertices from top to bottom, moving from left to right along each row (as shown in Figure 5) and pick the square with a smaller number if a tie appears.

In order to present the following steps visually, we assume that an open tour  $ABCDEFGH$  is obtained as shown in Figure 5.



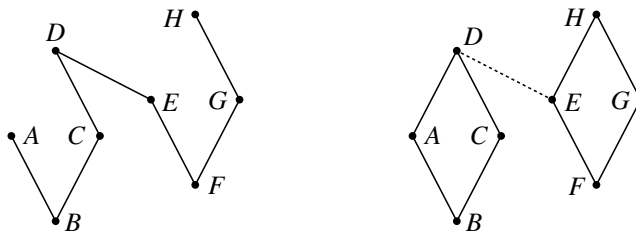


**Figure 5.** The ordering system to break ties in Warnsdorff's rule.

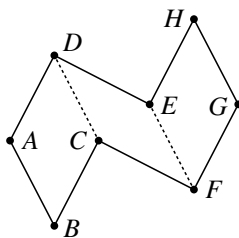
*Step 2: Cut this open knight's tour into disjoint closed knight's tours.* Suppose that we have obtained an open knight's tour in the previous step and that the tour starts at vertex  $v$ . Then label each vertex with a number indicating when it is visited. For example,  $v$  is the 1st vertex and a neighbor of  $v$  is the 2nd vertex (note that the labels in this step are different from the labels in Step 1, which are assigned based upon positions and used only to break ties). If a neighbor of  $v$  is the  $n$ -th vertex, then the vertices with numbers from 1 to  $n$  form a closed partial knight's tour because the piece can move back to  $v$ , so we add an edge between the starting vertex and the  $n$ -th vertex and delete the edge between the  $n$ -th vertex and the  $(n+1)$ -st vertex. Now we are left with a closed partial knight's tour consisting of vertices with numbers from 1 to  $n$ , and an open partial knight's tour consisting of vertices with numbers greater than  $n$ . We now let the  $(n+1)$ -st vertex be our new starting vertex and find another closed knight's tour using the above method. Repeat until every vertex in the original open knight's tour becomes a part of a closed partial knight's tour.

Since we eventually have to join these tours together, it is in our favor to make the number of closed knight's tours as small as possible. To achieve this goal, we use a greedy approach: we always pick the neighbor of the starting vertex with the greatest number to be the last vertex that closes the partial tour.

Figure 6 shows this procedure in action. On the left is an open tour; we first



**Figure 6.** Left: an open tour on eight vertices. Right: the open tour cut into partial closed tours.



**Figure 7.** Joining partial closed tours into one closed tour.

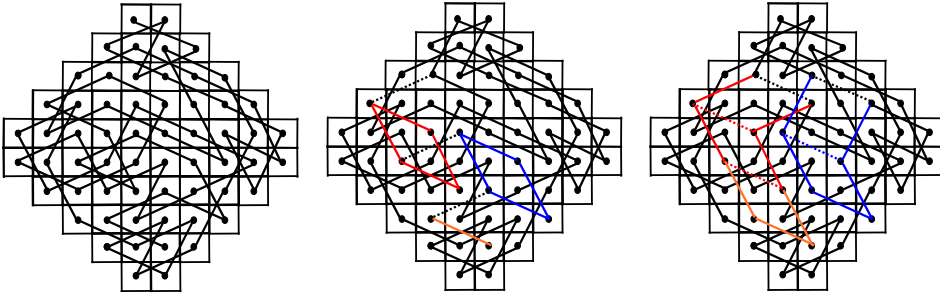
pick  $A$  as the starting vertex and choose the 4th vertex  $D$  to close the partial tour. Then we pick the 5th vertex  $E$  as the next starting vertex and close the partial tour with  $H$ . Now every vertex is a part of a closed partial knight's tour and we move on to the next step.

*Step 3: Connect these closed knight's tours together.* Two knight's tours are able to connect to each other if a pair of adjacent squares in one tour is *parallel* to a pair of adjacent squares in the other tour. That is, the four squares in two parallel pairs must be able to form a closed partial knight's tour. To join these two tours, we delete the original edges between the four vertices and add the other pair of parallel edges. For example,  $AB$  is parallel to  $CD$ , and  $DC$  is parallel to  $EF$  in Figure 6, right. Hence, the partial tour  $ABCD$  can be joined to the partial tour  $EFGH$  as shown in Figure 7. Notice that our edge switching procedure is dependent on the fact that the knight's tour graph on the Aztec diamond has regular symmetrical structure, making easier switching opportunities.

In theory, the order in which we join these closed partial knight's tours matters because each time we join two tours, we change the direction of only two edges in these tours. In practice, however, we conjecture that it is less important overall and do not have a specific heuristic for joining tours. One ordering that may be helpful is joining the shorter knight's tours first before trying longer ones because the probability of whether a tour can be joined depends on the length of the tour. This ordering is hard to implement due to the data structure we use and thus ignored in our case.

*Step 4: If there exists a knight's tour that cannot be connected to the rest of the knight's tours, reject. Otherwise, accept.* This step checks whether all the closed knight's tours obtained in Step 2 are joined together. If so, then we have a complete knight's tour on the entire board. If not, we fail to construct a closed knight's tour with the given open knight's tour in Step 1. Because it is fast to find an open knight's tour with Warnsdorff's rule, we switch to a new open knight's tour if a closed knight's tour cannot be constructed from the previous one.

Figure 8 is an example of the entire algorithm on an Aztec diamond of radius 5. The left picture shows an open tour obtained in Step 1. The middle picture shows



**Figure 8.** An example of the path-conversion algorithm on the radius 5 diamond.

how Step 2 of our algorithm cuts this open tour into four closed partial tours, each with a different color. The right picture shows how the Step 3 joins all the partial tours together.

With this new algorithm, we are able to find knight's tours on all Aztec diamonds of radius 100 or less as well as boards of radius 102, 104, 105, 106, 108, 109 and 111. The performance of this algorithm seems to deteriorate after the radius of an Aztec diamond exceeds 112 as we were only able to find an Aztec diamond of radius 125 via a search of boards having sizes between 112 and 140. This is probably due to the application of Warnsdorff's rule, which has a worse performance on large graphs. Our algorithm finishes in four minutes for a board of radius 100 or less and it grows polynomially as the radius increases because there is a one-to-one correspondence between starting vertices and open tours obtained in Step 1 of the algorithm (note that although the random walk algorithm can provide a knight's tour in 10 seconds, the algorithm takes a long time to finish even for an Aztec diamond of radius 5).

This new algorithm not only works for Aztec diamonds but also any graph because it is essentially an algorithm that transforms open knight's tours (hamiltonian paths) to closed knight's tours (hamiltonian cycles).

### 5. Applications to random graphs

We applied the path conversion algorithm to random graphs to test its robustness. Two questions were asked during this process:

- (1) How many hamiltonian paths can be converted into hamiltonian cycles?
- (2) Of all the graphs that have at least one hamiltonian path, how many have a hamiltonian cycle?

Note that the answers obtained below are not true answers but ones provided by our path conversion algorithm. Evidenced by the classic theorems of Dirac and Ore

[West 2001], sufficient conditions for the existence of hamiltonian cycles often involve degree constraints. By Ganzfried's claim [2004], Warnsdorff's rule fails more regularly when the size of the graph increases. Hence, we controlled for the average degree and the number of vertices when running our algorithm on random graphs.

**5.1. Generating random graphs.** Suppose we want to create a random graph with  $n$  vertices and (expected) average degree  $d$ . The total number of edges is  $\frac{1}{2}dn$ . Since a complete graph with  $n$  vertices has at most  $\frac{1}{2}n(n-1)$  edges, we set the probability of existence of any edge between two vertices to  $d/(n-1)$  so that the expected number of edges is  $\frac{1}{2}dn$ . Such a random graph, however, might not be connected. A disconnected graph has no hamiltonian cycles. Therefore, we ignore all disconnected graphs generated during this process.

**5.2. Results on random graphs.** To explore how average degree and number of vertices affect the answers to the two questions proposed at the beginning of this section, we conducted two sets of experiments. The first set fixed the number of vertices to be 1000 and changed the average degree, while the second fixed the average degree to be 8 and changed the number of vertices. The reason for choosing 1000 vertices for the first set is that it is a number small enough such that we could collect a decent amount of data in a short period of time and large enough such that a brute-force algorithm would take a long time to terminate. The reason for choosing average degree to be 8 is that it gives a rough comparison between the random graphs and the Aztec diamonds. Almost all vertices in an Aztec diamond have a degree of 8 except for those in peripheral areas.

The results are summarized in Table 1 and Table 2. The first column is the controlled variable, which could be either the average degree or the number of vertices; the second column measures, of all hamiltonian paths found using Warnsdorff's rule, how many can be converted into hamiltonian cycles; the last column shows how likely a graph contains a hamiltonian cycle if we know that at least one hamiltonian path can be found using Warnsdorff's rule. The statistics of each row is obtained from performing the path-conversion algorithm on exactly 1000 graphs. For convenience, we call the statistics in the second column the *conversion rate* and those in the third column the *success rate*.

From Table 1 we conclude that as the average degree of a graph goes up, it is more likely that a hamiltonian path can be converted into a hamiltonian cycle. In addition, more hamiltonian paths in total can be found. Therefore, the chance of finding a hamiltonian cycle rises significantly as the average degree increases. Similarly, if we fix the average degree and increase the number of vertices in a graph, we see a drop in the conversion rate. The number of found paths does not change monotonically. It rises first when the number of vertices changes from 100 to 200 and then falls when

degree	cycles/paths	cycles found/paths found
7	6.78%(4/59)	25.00%(1/4)
8	9.02%(1967/21804)	42.01%(92/219)
9	13.66%(19841/145207)	53.91%(338/627)
10	18.74%(75764/410166)	77.09%(700/908)
11	23.57%(141773/601410)	88.13%(861/977)
12	28.53%(212320/744178)	95.36%(946/992)
13	33.33%(272710/818283)	97.90%(979/1000)
14	37.96%(332630/876159)	99.30%(990/997)
15	42.40%(388014/915194)	99.90%(999/1000)

**Table 1.** Performance on random graphs with 1000 vertices.

vertices	cycles/paths	cycles found/paths found
100	69.67%(56676/81354)	87.73%(858/978)
200	50.47%(61280/121407)	76.29%(708/928)
300	37.44%(43992/117502)	64.90%(514/792)
400	28.34%(33034/116568)	60.03%(428/713)
500	21.87%(19129/87478)	48.58%(309/636)
600	18.27%(12085/66157)	44.18%(220/498)
700	14.14%(7323/51784)	46.55%(182/391)
800	12.42%(5068/40801)	44.03%(140/318)
900	10.35%(2868/27703)	37.96%(104/274)

**Table 2.** Performance on random graphs with expected average degree 8.

the number of vertices is further increased. The success rate, however, changes monotonically despite the oscillation in the number of paths found.

An Aztec diamond of radius 100 has 20200 vertices. According to these tables, the probability for finding a hamiltonian cycle on such a huge graph should be really small. Yet we were able to find hamiltonian cycles for all Aztec diamonds up to radius 100. Therefore, we conjecture that the degree distribution also affects how likely a graph has hamiltonian cycles. In the following subsections, we will test the performance of our algorithm on random regular graphs. But first, let us talk about how these graphs are generated.

**5.3. Generating random regular graphs.** To generate random regular graphs, we utilized the following algorithm described by Kim and Vu [2003]. Let  $n$  be the number of vertices in  $G$  and  $d$  be the degree of each vertex:

degree	cycles/paths	cycles found/paths found
7	2.67%(142/5328)	13.66%(134/981)
8	4.13%(1118/28549)	67.20%(672/1000)
9	5.94%(4300/80293)	97.60%(976/1000)
10	8.24%(11087/151781)	99.90%(999/1000)
11	11.07%(23181/241791)	100%(1000/1000)
12	14.28%(39929/323588)	100%(1000/1000)
13	17.91%(60671/401756)	100%(1000/1000)
14	21.86%(85501470826)	100%(1000/1000)
15	25.83%(114657/534245)	100%(1000/1000)

**Table 3.** Performance on random regular graphs with 1000 vertices.

1. Create a graph  $G$  with  $n$  vertices. Label these  $n$  vertices with integers from 0 to  $n - 1$ . Create a list  $L$  with  $d$  copies of each integer.
2. Find two random integers  $i$  and  $j$  from  $L$ . While  $i = j$  or vertex  $i$  and vertex  $j$  are already adjacent, choose another  $j$ . Connect vertex  $i$  and vertex  $j$  once  $i$  and  $j$  are chosen. Remove  $i$  and  $j$  from  $L$ .
3. If  $L$  is not empty, repeat Step 2. Else, output  $G$ .

Note that  $G$  may not be connected. Again we ignore all disconnected graphs. Furthermore, we can get stuck at Step 2 if we are unlucky. For example,  $L$  may contain integers of the same value. To avoid getting trapped in infinite loops, we restart our algorithm if no suitable pair of  $i$  and  $j$  can be found within a certain number of iterations.

**5.4. Results on random regular graphs.** To make comparisons more direct, we choose the same parameters. That is, all graphs generated during this experiment have exactly 1000 vertices. As shown in Table 3, the conversion rate on random regular graphs is much lower than that on random graphs. There are also fewer hamiltonian paths found on random regular graphs when the degree is larger than 8. However, random regular graphs have a better success rate than random graphs with the same parameters except for degree 7 (the statistics on random graphs of degree 7 are unreliable because the algorithm finds only 59 paths). One possible explanation is that the found paths are distributed more evenly on random regular graphs. Another possible explanation is that our algorithm works better for random regular graphs. Although the obtained results are algorithm-specific, it is worth asking whether degree distribution (instead of minimum degree) is related to the probability of finding a hamiltonian cycle.

## 6. Online supplement

Two supplementary files are provided online only. The code supplement contains Java programs implementing the algorithms described in the paper. The table supplement contains two Excel files: one shows all the results obtained from the revised random-walk algorithm on Aztec diamonds of radii from 2 to 20, and the other shows examples of knight's tours on Aztec diamonds of radii 30, 40, 50, 60, 70, and 80. Both folders include documentation.

## 7. Future work

Although our algorithm has worse performance on larger graphs, it can be improved in various ways. First, we can use better tie-breaking rules for Warnsdorff's rule or we can switch to another rule to find an open knight's tour. This will increase the likelihood of finding an open knight's tour, which is essential to the construction of a closed knight's tour. Second, we can have a different heuristic for cutting an open knight's tour into several closed partial knight's tours. A good cutting method will minimize the number of partial tours and possibly the variance in the lengths of these partial tours because it will reduce the probability of having a partial tour that is not able to attach to other tours. Third, it is likely that there exists a better order in which we join the closed partial tours.

**Open Problem.** Is there an Aztec diamond that has no open knight's tour?

**Open Problem.** Is there an Aztec diamond that has an open knight's tour but not a closed knight's tour?

## References

- [Bi et al. 2015] B. Bi, S. Butler, S. DeGraaf, and E. Doebel, "Knight's tours on boards with odd dimensions", *Involve* **8**:4 (2015), 615–627. MR Zbl
- [Cairns 2002] G. Cairns, "Pillow chess", *Math. Mag.* **75**:3 (2002), 173–186. MR
- [Demaio and Hippchen 2009] J. Demaio and T. Hippchen, "Closed knight's tours with minimal square removal for all rectangular boards", *Math. Mag.* **82**:3 (2009), 219–225. Zbl
- [Euler 1759] L. Euler, "Solution d'une question curieuse qui ne paroît soumise à aucune analyse", *Mém. Acad. Roy. Sci. Belles Lett. (Berlin)* **15** (1759), 310–337. Reprinted in *Commentationes arithmeticae* **1** (1849), 337–355, and in *Commentationes algebraicae ad theoriam combinationum et probabilitatum pertinentes*, edited by L. G. Du Pasquier, Opera Omnia (1), **7** (1923), 26–56.
- [Euler 1782] L. Euler, "Recherches sur un nouvelle espèce de quarrés magiques", *Verh. Zeeuwsch Genootsch. Wetensch. Vlissingen* **9** (1782), 85–239. Reprinted in *Commentationes arithmeticae* **2** (1849), 302–361, and in *Commentationes algebraicae ad theoriam combinationum et probabilitatum pertinentes*, edited by L. G. Du Pasquier, Opera Omnia (1), **7** (1923), 291–392.
- [Ganzfried 2004] S. Ganzfried, "A new algorithm for knight's tours", REU proceeding, Oregon State University, 2004, available at [https://www.cs.cmu.edu/~sganzfri/Knights\\_REU04.pdf](https://www.cs.cmu.edu/~sganzfri/Knights_REU04.pdf).

- [Kim and Vu 2003] J. H. Kim and V. H. Vu, “Generating random regular graphs”, pp. 213–222 in *Proceedings of the thirty-fifth annual ACM symposium on theory of computing*, ACM, New York, 2003. MR Zbl
- [Miller and Farnsworth 2013] A. M. Miller and D. L. Farnsworth, “Knight’s tours on  $3 \times n$  chessboards with a single square removed”, *Open J. Discrete Math.* **3**:1 (2013), 56–59.
- [Murray 1913] H. J. R. Murray, *A history of chess*, Oxford University Press, London, 1913.
- [Parberry 1997] I. Parberry, “An efficient algorithm for the knight’s tour problem”, *Discrete Appl. Math.* **73**:3 (1997), 251–260. MR Zbl
- [Schwenk 1991] A. J. Schwenk, “Which rectangular chessboards have a knight’s tour?”, *Math. Mag.* **64**:5 (1991), 325–332. MR Zbl
- [von Warnsdorff 1823] H. C. von Warnsdorff, *Des Rösselsprunges einfachste und allgemeinste Lösung*, Varnhagen, Schmalkalden, 1823.
- [Watkins 2000] J. J. Watkins, “Knight’s tours on cylinders and other surfaces”, *Congr. Numer.* **143** (2000), 117–127. MR Zbl
- [Watkins and Hoenigman 1997] J. J. Watkins and R. L. Hoenigman, “Knight’s tours on a torus”, *Math. Mag.* **70**:3 (1997), 175–184. MR Zbl
- [West 2001] D. B. West, *Introduction to graph theory*, 2nd ed., Prentice Hall, Upper Saddle River, NJ, 2001. Zbl

Received: 2015-07-28      Revised: 2016-07-08      Accepted: 2016-08-21

daviess@uw.edu

*Padelfor Hall, University of Washington, W. Stevens Way NE,  
Seattle, WA 98105, United States*

cx94.main@gmail.com

*Department of Mathematics and Computer Science, Davidson  
College, Box 7129, Davidson, NC 28035, United States*

cayerger@davidson.edu

*Department of Mathematics and Computer Science, Davidson  
College, Box 7059, Davidson, NC 28035, United States*



# Optimal aggression in kleptoparasitic interactions

David G. Sykes and Jan Rychtář

(Communicated by Natalia Hritonenko)

We have created and analyzed a model for kleptoparasitic interactions when individuals decide on the level of aggression in which they want to engage in the contest over a resource item. The more aggressive each individual is relative to an opponent, the higher are the chances of winning the item, but also the higher is the cost of the interaction for that individual. We consider a general class of cost functions and show that for any parameter values, i.e., for any maximal potential level of aggression of the individuals, any value of the resource and any type of the cost function, there is always a unique Nash equilibrium. We identify four possible kinds of Nash equilibria and give precise conditions for when they occur. We find that nonaggressive behavior is not a Nash equilibrium even when the cost function is such that aggressive behavior yields lower payoffs than avoiding the conflict altogether.

## 1. Introduction

Kleptoparasitism is the resource gathering behavior where one animal steals from another. Possible contested resources include territory, mates, and food [Iyengar 2008]. This stealing behavior is exhibited by a wide variety of species, such as seabirds [Spear et al. 1999; Steele and Hockey 1995; Triplet et al. 1999], insects [Jeanne 1972], fish [Grimm and Klinge 1996] and mammals [Kruuk 1972]. Kleptoparasitic interactions manifest in several varieties and are distinguished by the energy invested by the kleptoparasite and the resource owner. Some kleptoparasites display only minor levels of aggression and may be easily dissuaded by a highly invested adversary (e.g., catbirds steal food provisions from digger wasps, but they forgo this foraging strategy when sparrows are present because the prospect of competition with sparrows dissuades them [Benttinen and Preisser 2009]), whereas others are as aggressive as possible (e.g., male southern giant petrels will attack adult king penguins for food despite having low success rates [Hunter 1991]). Similarly, some resource owners are easily convinced to forfeit their resources (e.g., when attacked by turkey vultures, adolescent great blue herons are known to weakly resist

---

*MSC2010:* 91A05, 91A40.

*Keywords:* kleptoparasitism, food stealing, game theory.

the attack by pecking, but if the pecking does not dissuade the vulture then they will disgorge food [Brockmann and Barnard 1979]), whereas others engage in costly attempts to defend their resources (e.g., lapwings will undergo extensive aerial chases to avoid forfeiting food to assailing black-headed gulls [Källander 1977]).

Mathematical models of kleptoparasitism are quite common; see, for example, [Giraldeau and Livoreil 1998; Broom and Ruxton 2003; Broom et al. 2004; 2008; 2010; Broom and Rychtář 2007; 2013; Hadjichrysanthou and Broom 2012; Kokko 2013]. With mathematical modeling, we can determine the conditions under which the benefits of the various kleptoparasitic behaviors observed in nature outweigh the costs. Mathematical modeling also allows us to predict which ecological conditions make the occurrence of kleptoparasitism more likely.

Here we modify a game-theoretical “producer-scrounger” model developed in [Broom et al. 2015]. Producer-scrounger models [Barnard and Sibly 1981; Barnard 1984; Vickery et al. 1991; Caraco and Giraldea 1991; Dubois and Giraldeau 2005] describe interactions where after a kleptoparasite (i.e., the scrounger) encounters an individual with resources (i.e., the producer), the scrounger invests some amount of energy into stealing the resources while the producer attempts to defend them. Many game-theoretical models of two-individual interactions have been developed wherein the individuals have a discrete set of strategies available to them [Smith and Price 1973; Dubois and Giraldeau 2005; Broom et al. 2013], but realistically, individuals competing in kleptoparasitic interactions can invest amounts of energy from a continuous range of possibilities. This possibility is incorporated in our model, as it was in [Broom et al. 2015], where the producer-scrounger conflict is modeled as an extensive form game where the scrounger chooses its strategy first and the producer knows the scrounger’s choice before making its own. Here, we present and analyze the simultaneous version of the game where both individuals have to decide without knowing the opponent’s action.

The organization of our paper is as follows. In Section 2 we give a detailed mathematical description of our model. In Section 3 we analyze our model mathematically; in particular, we find best responses to opponent’s actions in Section 3.1 and give conditions for Nash equilibria in Section 3.2. The results of our analysis are presented in Section 4. In Section 4.1 we show that (for the case  $\alpha > 0$ ) Nash equilibria do not overlap and in Section 4.2 we show that the Nash equilibria exist for any parameter combination. We end our paper in Section 5 where we compare our model and its results to previous work, most notably to [Broom et al. 2015].

## 2. Model

One individual, a scrounger, is searching for resources and encounters another individual, a producer, who has a resource item of value  $v$ . Simultaneously, and

with no knowledge of the choice of the other, they both have to decide how aggressive to be in the contest for the item. The more aggressive each individual is relative to the opponent, the higher are the chances of winning the item, but also the higher is the cost of the interaction for that individual. Let  $P_{\max}$  (or  $S_{\max}$ ) be the maximum level of aggression that a producer (or scrounger, respectively) can display in a fight, and by  $p \in [0, P_{\max}]$  (or  $s \in [0, S_{\max}]$ ) we will denote the actually displayed level of aggression in a particular contest. The producer wins the fight (and the resource) with a probability of  $p/(s+p)$ , while the scrounger wins with probability  $s/(s+p)$ . If no individual fights (i.e.,  $p = s = 0$ ), then the producer is assumed to win and will keep the resource.

We will adopt the model for the fight costs from [Broom et al. 2015]. When no individual fights, the cost is 0. Otherwise, the cost to each individual is  $(s+p)^\alpha$ . Here,  $\alpha$  is a tuning parameter that allows us to consider a broad range of scenarios. If  $\alpha < 1$ , then low aggression is costly relative to no aggression at all, but once the aggression reaches a certain level, increasing the aggression is typically not that costly in relative terms. On the contrary, when  $\alpha > 1$ , low aggression levels are relatively cheap, but escalating the fight (i.e., being a bit more aggressive) is relatively expensive as the function  $x^\alpha$  is concave up. For the rest of the paper, we will assume  $\alpha > 0$  except when we discuss the extreme case  $\alpha = 0$  (when the cost of the fight is constant) separately in Section 4.3.

Assuming the scrounger plays  $s \in [0, S_{\max}]$ , the producer plays  $p \in [0, P_{\max}]$ , and the value of the resource is  $v$ , the payoffs to the producer and scrounger are given by

$$U_{pr}(s, p) = \begin{cases} v & \text{if } s = p = 0, \\ p/(s+p)v - (s+p)^\alpha & \text{if } s+p > 0, \end{cases} \quad (1)$$

$$U_{sc}(s, p) = \begin{cases} 0 & \text{if } s = p = 0, \\ s/(s+p)v - (s+p)^\alpha & \text{if } s+p > 0. \end{cases} \quad (2)$$

### 3. Analysis

**3.1. Best responses.** Here we will determine best responses for the scrounger and producer. A best response is a strategy that maximizes an individual's payoff given that their adversary's strategy is fixed, i.e., for a given  $s \in [0, S_{\max}]$ , we are looking for  $p_{br}(s) \in [0, P_{\max}]$  such that

$$U_{pr}(s, p_{br}(s)) = \max_{p \in [0, P_{\max}]} \{U_{pr}(s, p)\}, \quad (3)$$

and, similarly, for a given  $p \in [0, P_{\max}]$  we are looking for  $s_{br}(p) \in [0, S_{\max}]$  such that

$$U_{sc}(s_{br}(p), p) = \max_{s \in [0, S_{\max}]} \{U_{sc}(s, p)\}. \quad (4)$$

When  $s = 0$ , it immediately follows from (1) that  $p_{br}(0) = 0$ . When  $s > 0$ , we have

$$\frac{\partial}{\partial p} U_{pr}(s, p) = \frac{sv - \alpha(s + p)^{\alpha+1}}{(s + p)^2} \quad \text{for } p > -s. \tag{5}$$

Let us define

$$f(x) = \left(\frac{xv}{\alpha}\right)^{1/(\alpha+1)} - x. \tag{6}$$

It follows from (5) that, for fixed  $s$  and variable  $p$ , the function  $U_{pr}(s, p)$  is increasing on  $(-s, f(s)]$  and decreasing on  $[f(s), +\infty)$ . Therefore,

$$p_{br}(s) = \begin{cases} 0 & \text{if } s = 0, \text{ or } s > 0 \text{ and } f(s) \leq 0, \\ f(s) & \text{if } s > 0 \text{ and } 0 \leq f(s) \leq P_{\max}, \\ P_{\max} & \text{if } s > 0 \text{ and } f(s) \geq P_{\max}. \end{cases} \tag{7}$$

We note that the conditions in (7) are formally not mutually exclusive, but whenever two of the conditions coincide, so does the best response defined by them.

When  $p = 0$ , we have  $U_{sc}(s, 0) = v - s^\alpha$  which increases to  $v > 0$  as  $s$  decreases to 0 while  $U_{sc}(0, 0) = 0$ . Thus, there is no best response for the scrounger in this case. When  $p > 0$ ,

$$\frac{\partial}{\partial s} U_{sc}(s, p) = \frac{pv - \alpha(s + p)^{\alpha+1}}{(s + p)^2} \quad \text{for } s > -p. \tag{8}$$

Consequently, for a fixed  $p$  and variable  $s$ , the function  $U_{sc}(s, p)$  is increasing on  $(-p, f(p)]$  and decreasing on  $[f(p), +\infty)$ . Hence,

$$s_{br}(p) = \begin{cases} \text{does not exist} & \text{if } p = 0, \\ 0 & \text{if } p > 0 \text{ and } f(p) \leq 0, \\ f(p) & \text{if } 0 < f(p) \leq S_{\max}, \\ S_{\max} & \text{if } f(p) \geq S_{\max}. \end{cases} \tag{9}$$

As with  $p_{br}$ , we note that the conditions in (9) are formally not mutually exclusive, but whenever two of the conditions coincide, so does the best response defined by them.

**3.2. Nash equilibria.** Here we will identify all Nash equilibria of our game. A pair of strategies  $(s^*, p^*)$  is a *Nash equilibrium* if  $p^*$  is the producer’s best response to  $s^*$  and  $s^*$  is the scrounger’s best response to  $p^*$ .

By (7), we only need to consider cases when  $p^* = 0$ ,  $p^* = f(s^*)$  and  $p^* = P_{\max}$ ; and by (9), for any of those cases we only need to consider  $s^* = 0$ ,  $s^* = f(p^*)$  and  $s^* = S_{\max}$ .

By (9), no pair  $(s^*, 0)$  is a Nash equilibrium. When  $s^* = 0$ , by (7), we would need  $p^* = 0$  and so no pair  $(0, p^*)$  is a Nash equilibrium either. We will now investigate the remaining types separately. Table 1 summarizes the results.

Nash equilibrium	conditions
$(S_{\max}, P_{\max})$	$P_{\max} \leq f(S_{\max}), S_{\max} \leq f(P_{\max})$
$(S_{\max}, f(S_{\max}))$	$0 < f(S_{\max}) < P_{\max}, S_{\max} \leq f(f(S_{\max}))$
$(f(P_{\max}), P_{\max})$	$0 < f(P_{\max}) < S_{\max}, P_{\max} \leq f(f(P_{\max}))$
$(\frac{1}{2}(v/(2\alpha))^{1/\alpha}, \frac{1}{2}(v/(2\alpha))^{1/\alpha})$	$\frac{1}{2}(v/(2\alpha))^{1/\alpha} < \min(P_{\max}, S_{\max})$

**Table 1.** Nash equilibria and the conditions for their existence. As  $f(x) = (xv/\alpha)^{1/(\alpha+1)} - x$ , the conditions are given in terms of  $P_{\max}, S_{\max}, v$  and  $\alpha$ .

**3.2.1.** *Type  $(S_{\max}, P_{\max})$ .* By (7) and (9),  $(S_{\max}, P_{\max})$  is a Nash equilibrium if and only if

$$P_{\max} \leq f(S_{\max}), \tag{10a}$$

$$S_{\max} \leq f(P_{\max}). \tag{10b}$$

**3.2.2.** *Type  $(S_{\max}, f(S_{\max}))$ .* By (7) and (9),  $(S_{\max}, f(S_{\max}))$  is a Nash equilibrium if and only if

$$0 < f(S_{\max}) < P_{\max}, \tag{11a}$$

$$S_{\max} \leq f(f(S_{\max})). \tag{11b}$$

**3.2.3.** *Type  $(f(P_{\max}), P_{\max})$ .* By (7) and (9),  $(f(P_{\max}), P_{\max})$  is a Nash equilibrium if and only if

$$0 < f(P_{\max}) < S_{\max}, \tag{12a}$$

$$P_{\max} \leq f(f(P_{\max})). \tag{12b}$$

**3.2.4.** *Type  $(p^*, s^*)$  where  $p^* = f(s^*)$  and  $s^* = f(p^*)$ .* Solving  $p^* = f(f(p^*))$  yields a unique solution  $p^* = \frac{1}{2}(v/(2\alpha))^{1/\alpha}$ . Indeed, we have

$$x = f(f(x)) \tag{13}$$

$$= \left(\frac{f(x)v}{\alpha}\right)^{1/(\alpha+1)} - f(x) \tag{14}$$

$$= \left(\frac{f(x)v}{\alpha}\right)^{1/(\alpha+1)} - \left(\frac{xv}{\alpha}\right)^{1/(\alpha+1)} + x, \tag{15}$$

which after simple algebra yields

$$x = f(x) = \left(\frac{xv}{\alpha}\right)^{1/(\alpha+1)} - x, \tag{16}$$

and thus

$$x = \frac{1}{2}\left(\frac{v}{2\alpha}\right)^{1/\alpha}. \tag{17}$$

Consequently, the only candidate for such a type of a Nash equilibrium is

$$\left( \frac{1}{2} \left( \frac{v}{2\alpha} \right)^{1/\alpha}, \frac{1}{2} \left( \frac{v}{2\alpha} \right)^{1/\alpha} \right).$$

By (7) and (9) it is indeed a Nash equilibrium if  $0 < f(p^*) < S_{\max}$  and  $0 < f(s^*) < P_{\max}$ , and since  $p^* = f(s^*)$ ,  $s^* = f(p^*)$ , we get that  $(\frac{1}{2}(v/(2\alpha))^{1/\alpha}, \frac{1}{2}(v/(2\alpha))^{1/\alpha})$  is a Nash equilibrium if and only if

$$\frac{1}{2} \left( \frac{v}{2\alpha} \right)^{1/\alpha} < \min(P_{\max}, S_{\max}). \tag{18}$$

### 4. Results

We have seen that there are only four potential Nash equilibria in this game:

- (1)  $(S_{\max}, P_{\max})$ ,
- (2)  $(S_{\max}, f(S_{\max}))$ ,
- (3)  $(f(P_{\max}), P_{\max})$  and
- (4)  $(\frac{1}{2}(v/(2\alpha))^{1/\alpha}, \frac{1}{2}(v/(2\alpha))^{1/\alpha})$ .

Here, we will show that under any parameter values  $v > 0$ ,  $\alpha > 0$ ,  $S_{\max} > 0$ ,  $P_{\max} > 0$ , there exists one and only one Nash equilibrium.

The conditions (10), (11), (12) and (18) for the equilibria are given in terms of  $f(x) = (xv/\alpha)^{1/(\alpha+1)} - x$ . It is therefore crucial to understand the behavior of  $f$ . The following two equivalencies for  $x \geq 0$  follow easily from simple algebra:

$$x \begin{matrix} \leq \\ \geq \end{matrix} \frac{1}{2} \left( \frac{v}{2\alpha} \right)^{1/\alpha} \quad \text{if and only if} \quad x \begin{matrix} \leq \\ \geq \end{matrix} f(x), \tag{19}$$

and similarly,

$$x \begin{matrix} \leq \\ \geq \end{matrix} \frac{1}{2} \left( \frac{v}{2\alpha} \right)^{1/\alpha} \quad \text{if and only if} \quad x \begin{matrix} \leq \\ \geq \end{matrix} f(f(x)), \tag{20}$$

and they will be useful when determining the existence and uniqueness of Nash equilibria.

**4.1. Nash equilibria do not overlap.** First, it follows from (20) that when (18) holds, one has  $P_{\max} > f(f(P_{\max}))$  and  $S_{\max} > f(f(S_{\max}))$ , i.e., neither (12b) nor (11b) holds. Thus,  $(\frac{1}{2}(v/(2\alpha))^{1/\alpha}, \frac{1}{2}(v/(2\alpha))^{1/\alpha})$  cannot occur at the same time as  $(f(P_{\max}), P_{\max})$  or  $(S_{\max}, f(S_{\max}))$ . By (19),  $f(S_{\max}) < S_{\max}$  and  $f(P_{\max}) < P_{\max}$ . Consequently, either  $f(S_{\max}) < S_{\max} \leq P_{\max}$  or  $f(P_{\max}) < P_{\max} \leq S_{\max}$ , i.e.,  $(\frac{1}{2}(v/(2\alpha))^{1/\alpha}, \frac{1}{2}(v/(2\alpha))^{1/\alpha})$  cannot occur at the same time as  $(S_{\max}, P_{\max})$ .

Second, when (12) holds, then, by (20),  $P_{\max} \leq \frac{1}{2}(v/(2\alpha))^{1/\alpha}$  and thus, by (19),  $P_{\max} \leq f(P_{\max})$  and so  $P_{\max} < S_{\max}$ . By a similar argument, when (11) holds,

$S_{\max} < P_{\max}$ . Consequently,  $(S_{\max}, f(S_{\max}))$  and  $(f(P_{\max}), P_{\max})$  are never Nash equilibria at the same time.

Finally, it is evident that neither (12) nor (11) can hold when (10) does. Consequently, there is always at most one Nash equilibria.

**4.2. Nash equilibrium always exist.** We show that for any  $v > 0$ ,  $\alpha > 0$ ,  $S_{\max} > 0$ ,  $P_{\max} > 0$ , there is a Nash equilibrium. Here we will assume  $P_{\max} < S_{\max}$ , but when  $S_{\max} \leq P_{\max}$ , the proofs are analogous.

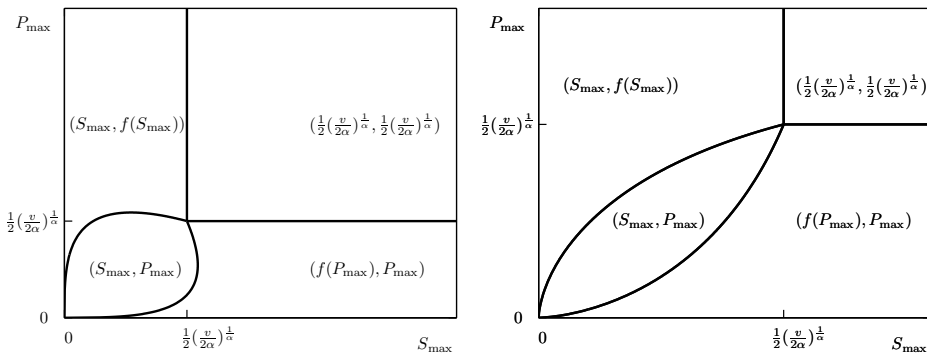
If  $\frac{1}{2}(v/(2\alpha))^{1/\alpha} < P_{\max} < S_{\max}$ , then by (18),  $(\frac{1}{2}(v/(2\alpha))^{1/\alpha}, \frac{1}{2}(v/(2\alpha))^{1/\alpha})$  is a Nash equilibrium.

If  $P_{\max} < S_{\max} < \frac{1}{2}(v/(2\alpha))^{1/\alpha}$ , then by (19),  $P_{\max} < S_{\max} < f(S_{\max})$ , i.e., (10a) holds. Also, by (20),  $P_{\max} < f(f(P_{\max}))$ , i.e., (12b) holds. Consequently, if  $f(P_{\max}) < S_{\max}$ , then  $(f(P_{\max}), P_{\max})$  is a Nash equilibrium (because we assumed  $P_{\max} < \frac{1}{2}(v/(2\alpha))^{1/\alpha}$  and thus, by (19),  $0 < P_{\max} < f(P_{\max})$ , i.e., (12) holds); and, similarly, if  $f(P_{\max}) \geq S_{\max}$ , then  $(S_{\max}, P_{\max})$  is a Nash equilibrium.

If  $P_{\max} < \frac{1}{2}(v/(2\alpha))^{1/\alpha} < S_{\max}$ , then, by (20),  $P_{\max} < f(f(P_{\max}))$ , i.e., (12b) holds. Consequently,

- (a) if  $f(P_{\max}) < S_{\max}$ , then  $(f(P_{\max}), P_{\max})$  is a Nash equilibrium; and
- (b) if  $f(P_{\max}) \geq S_{\max}$  and  $f(S_{\max}) \geq P_{\max}$ , then  $(S_{\max}, P_{\max})$  is a Nash equilibrium.

Since  $P_{\max} < S_{\max}$ , one cannot have  $f(P_{\max}) \geq S_{\max}$  and also  $f(S_{\max}) < P_{\max}$ . Consequently, the above cases are the only two possible cases and thus there is always a Nash equilibrium.



**Figure 1.** Regions of existence of Nash equilibria as  $v = 1$ ,  $S_{\max}$  and  $P_{\max}$  varies and (left)  $\alpha = 2$  and (right)  $\alpha = 0.5$ . Note that the regions do not overlap and the individuals are always aggressive (the level of aggression increases with increasing  $v$  (when  $S_{\max}$  and  $P_{\max}$  are fixed).

Figure 1 shows the Nash equilibria for fixed  $v$  and  $\alpha$  and variable  $S_{\max}$  and  $P_{\max}$ . Figure 2 shows Nash equilibria and payoffs for fixed  $S_{\max}$ ,  $P_{\max}$ ,  $\alpha$  and variable  $v$ . We see that for small  $v$ , individuals play  $\frac{1}{2}(v/(2\alpha))^{1/\alpha}$ ,  $\frac{1}{2}(v/(2\alpha))^{1/\alpha}$ . For large  $v$ , individuals play  $(S_{\max}, P_{\max})$ . For medium  $v$ , individuals play  $(f(P_{\max}), P_{\max})$  when  $S_{\max} > P_{\max}$  and  $(S_{\max}, f(S_{\max}))$  when  $S_{\max} < P_{\max}$ . Note that as  $v$  increases, so does the optimal aggression level; yet with increasing aggression, the relative payoff may decrease, as seen in Figure 2a and c for equilibria of the form  $(S_{\max}, f(S_{\max}))$  and  $(f(P_{\max}), P_{\max})$ . Also note that for  $\alpha < 1$  and small  $v$ , the payoffs are negative for both players; see Figure 2b and d. As  $v$  grows, the payoffs eventually become positive (it happens first for a more aggressive individuals).

**4.3. Case  $\alpha = 0$ .** So far, we have considered only  $\alpha > 0$ . When  $\alpha = 0$ , the cost of a fight is the constant 1 no matter what the exact aggression levels are (as long as at least one individual is aggressive). Thus, for fixed  $s > 0$ ,  $U_{pr}(s, p)$  is increasing in  $p$  and, for fixed  $p > 0$ ,  $U_{sc}(s, p)$  is increasing in  $s$  and the individuals effectively choose between being not aggressive at all or being aggressive at their maximal level. Hence, they play the following bimatrix game where the scrounger's payoff is

$$S \backslash P \quad \begin{array}{cc} 0 & P_{\max} \\ 0 & \begin{pmatrix} 0 & -1 \\ v-1 & \frac{S_{\max}}{S_{\max} + P_{\max}}v-1 \end{pmatrix} \\ S_{\max} & \end{array}, \quad (21)$$

and the producer's payoff is

$$S \backslash P \quad \begin{array}{cc} 0 & P_{\max} \\ 0 & \begin{pmatrix} v & v-1 \\ -1 & \frac{P_{\max}}{S_{\max} + P_{\max}}v-1 \end{pmatrix} \\ S_{\max} & \end{array}. \quad (22)$$

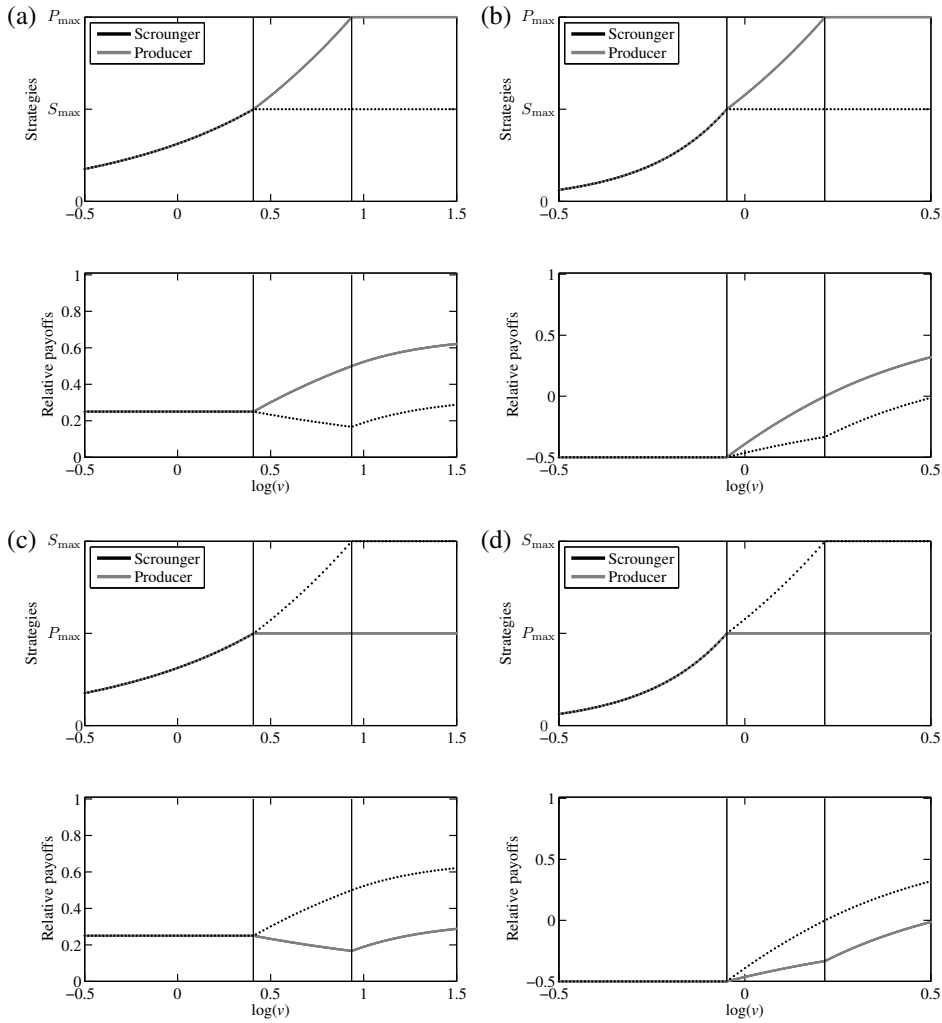
It turns out that this game is a variant of the stag hunt game [Skyrms 2004] for  $v < 1$  and the prisoner's dilemma game for  $v > 1$ .

When the producer plays  $p = P_{\max}$ , the scrounger always prefers  $s = S_{\max}$  over  $s = 0$ . When the scrounger plays  $s = S_{\max}$ , the producer always prefers  $p = P_{\max}$  over  $p = 0$ . Consequently,  $(S_{\max}, P_{\max})$  is always a Nash equilibrium. When the scrounger plays  $s = 0$ , the producer prefers  $p = 0$ . When the producer plays  $p = 0$ , the scrounger prefers  $s = 0$  when  $v < 1$  and prefers  $s = S_{\max}$  when  $v > 1$ . Consequently, when  $v > 1$ ,  $(S_{\max}, P_{\max})$  is the only Nash equilibrium, and when  $v < 1$ , both  $(S_{\max}, P_{\max})$  and  $(0, 0)$  are Nash equilibria.

Note the paradoxical situation in the case when

$$1 < v < \frac{S_{\max} + P_{\max}}{S_{\max}}.$$





**Figure 2.** Nash equilibria and payoffs relative to  $v$  (i.e.,  $U_{sc}(s^*, p^*)/v$  for the scrounger and  $U_{pr}(s^*, p^*)/v$  for the producer) when: (a)  $S_{\max} = 0.4$ ,  $P_{\max} = 0.8$ ,  $\alpha = 2$ , (b)  $S_{\max} = 0.4$ ,  $P_{\max} = 0.8$ ,  $\alpha = 0.5$ ; (c)  $S_{\max} = 0.8$ ,  $P_{\max} = 0.4$ ,  $\alpha = 2$ ; and (d)  $S_{\max} = 0.8$ ,  $P_{\max} = 0.4$ ,  $\alpha = 0.5$ . The vertical lines show the switch between Nash equilibria.

In this case, as in the prisoner's dilemma game,  $(S_{\max}, P_{\max})$  is a Nash equilibrium but the scrounger is getting a negative payoff (and the producer is also getting strictly less than  $v$ ). Hence both individuals would prefer not to engage in an aggressive conflict. Yet  $(0, 0)$  is not a Nash equilibrium because once either of the individuals decides not to be aggressive, the other one will be better off by being as aggressive as possible.

## 5. Discussion

We have created and analyzed a model for kleptoparasitic interactions when individuals decide on the level of aggression in which they want to engage in the contest over a resource item. We show that for any parameter values, there is a unique Nash equilibrium. We have provided explicit formulas for all of four possible types of Nash equilibria and have also derived explicit conditions for their existence.

Our model extends the model of [Broom et al. 2015] where the authors considered sequential decisions whereas we consider simultaneous decisions (or equivalently, a situation when both individuals have to choose the action without knowing the opponent's action). The analysis in our model is more complicated because individuals only know that the opponent will choose the optimal level of aggression, but unlike in the setting of [Broom et al. 2015], what is optimal depends on individual's own action as well. Our results also differ from the sequential setting, as our game does not admit multiple Nash equilibria (when  $\alpha > 0$ ) and it is also optimal to express at least some level of aggression. When  $v$  is small (relative to the maximal potential level of aggression of at least one of the individuals), the sequential model allows individuals to avoid the actual conflict, while they still fight aggressively in our simultaneous model. The difference between the two models is largest for concave down cost functions ( $\alpha < 1$ ) when the individuals would be better off without engaging in any fight (and this is indeed the Nash equilibrium for sequential decisions) but they still end up being aggressive when the decisions need to be made simultaneously. On the other hand, when  $v$  is large, then both the sequential and simultaneous decision models yield the same Nash equilibria.

The fight cost function plays a critical role in the determination of the equilibrium solutions. The fight cost functions that are considered in our model have the form  $(s + p)^\alpha$ , where  $0 \leq \alpha$ . It would be possible to model the fight cost function with greater complexity to increase the model's realism (see, e.g., [Baye et al. 2005; 2012]), but we have worked with the present fight cost formulation because it encompasses several possible fight cost functions without sacrificing the model's tractability. The appropriate setting for fight cost structure (i.e., for  $\alpha$ ) will of course depend on the interactions being modeled. Circumstances that lead to different settings of  $\alpha$  are considered in [Broom et al. 2015]; in particular,  $\alpha > 1$  corresponds to interactions for which the primary cost is risk of injury or lost energy whereas  $\alpha < 1$  corresponds to interactions for which the primary cost is a time cost. Such a time cost can be opportunity cost (i.e., lost time that can otherwise be spent foraging) or it can be the predation risk incurred by prolonged exposure while fighting for resources.

Similarly to [Broom et al. 2015], we assume that all individuals know the values of all parameters; in particular the scrounger knows  $P_{\max}$  and the producer knows

$S_{\max}$  and both individuals know the cost function and  $v$ . In [Broom and Rychtář 2009; 2016; Broom et al. 2013; 2014], the authors study the situation when  $v$  is not known to one of the individuals. However, as shown in Figure 1, different  $P_{\max}$  may not only yield different behavior of producer but may also yield different behavior of the scrounger. Thus, not knowing the opponent's maximum potential level of aggression will potentially influence the choice of individual strategies. Consequently, it would be interesting to model such a scenario.

### Acknowledgements

David Sykes was supported by the Undergraduate Research and Creativity Award from the Undergraduate Research, Scholarship and Creativity Office at UNCG. Jan Rychtář was supported by the Simons Foundation grant 245400. The authors would also like to thank Dr. Michal Johanis for useful discussions about the topic of this paper and anonymous referees for comments and suggestions that helped to improve the manuscript.

### References

- [Barnard 1984] C. J. Barnard, *Producers and scroungers: strategies of exploitation and parasitism*, C. Helm, London, 1984.
- [Barnard and Sibly 1981] C. Barnard and R. Sibly, "Producers and scroungers: a general model and its application to captive flocks of house sparrows", *Anim. Behav.* **29**:2 (1981), 543–550.
- [Baye et al. 2005] M. R. Baye, D. Kovenock, and C. G. de Vries, "Comparative analysis of litigation systems: an auction-theoretic approach", *Econ. J.* **115**:505 (2005), 583–601.
- [Baye et al. 2012] M. R. Baye, D. Kovenock, and C. G. de Vries, "Contests with rank-order spillovers", *Econom. Theory* **51**:2 (2012), 315–350. MR Zbl
- [Benttinen and Preisser 2009] J. Benttinen and E. Preisser, "Avian kleptoparasitism of the digger wasp *Sphex pensylvanicus*", *Can. Entomol.* **141**:6 (2009), 604–608.
- [Brockmann and Barnard 1979] H. J. Brockmann and C. J. Barnard, "Kleptoparasitism in birds", *Anim. Behav.* **27**:2 (1979), 487–514.
- [Broom and Ruxton 2003] M. Broom and G. D. Ruxton, "Evolutionarily stable kleptoparasitism: consequences of different prey types", *Behav. Ecol.* **14**:1 (2003), 23–33.
- [Broom and Rychtář 2007] M. Broom and J. Rychtář, "The evolution of a kleptoparasitic system under adaptive dynamics", *J. Math. Biol.* **54**:2 (2007), 151–177. MR Zbl
- [Broom and Rychtář 2009] M. Broom and J. Rychtář, "A game theoretical model of kleptoparasitism with incomplete information", *J. Math. Biol.* **59**:5 (2009), 631–649. MR Zbl
- [Broom and Rychtář 2013] M. Broom and J. Rychtář, *Game-theoretical models in biology*, CRC Press, Boca Raton, FL, 2013. MR Zbl
- [Broom and Rychtář 2016] M. Broom and J. Rychtář, "A model of food stealing with asymmetric information", *Ecol. Complex.* **26** (2016), 137–142.
- [Broom et al. 2004] M. Broom, R. M. Luther, and G. D. Ruxton, "Resistance is useless?—Extensions to the game theory of kleptoparasitism", *Bull. Math. Biol.* **66**:6 (2004), 1645–1658. MR Zbl

- [Broom et al. 2008] M. Broom, R. M. Luther, G. D. Ruxton, and J. Rychtář, “A game-theoretic model of kleptoparasitic behavior in polymorphic populations”, *J. Theoret. Biol.* **255**:1 (2008), 81–91. MR
- [Broom et al. 2010] M. Broom, M. L. Crowe, M. R. Fitzgerald, and J. Rychtář, “The stochastic modelling of kleptoparasitism using a Markov process”, *J. Theoret. Biol.* **264**:2 (2010), 266–272. MR
- [Broom et al. 2013] M. Broom, J. Rychtář, and D. G. Sykes, “The effect of information on payoff in kleptoparasitic interactions”, pp. 125–134 in *Topics from the 8th annual UCG regional mathematics and statistics conference*, edited by J. Rychtář et al., Springer Proceedings in Mathematics and Statistics **64**, Springer, New York, 2013.
- [Broom et al. 2014] M. Broom, J. Rychtář, and D. Sykes, “Kleptoparasitic interactions under asymmetric resource valuation”, *Math. Model. Nat. Phenom.* **9**:3 (2014), 138–147. MR Zbl
- [Broom et al. 2015] M. Broom, M. Johannis, and J. Rychtář, “The effect of fight cost structure on fighting behaviour”, *J. Math. Biol.* **71**:4 (2015), 979–996. MR Zbl
- [Caraco and Giraldea 1991] T. Caraco and L.-A. Giraldea, “Social foraging: producing and scrounging in a stochastic environment”, *J. Theor. Biol.* **153**:4 (1991), 559–583.
- [Dubois and Giraldeau 2005] F. Dubois and L.-A. Giraldeau, “Fighting for resources: the economics of defense and appropriation”, *Ecology* **86**:1 (2005), 3–11.
- [Giraldeau and Livoreil 1998] L.-A. Giraldeau and B. Livoreil, “Game theory and social foraging”, pp. 16–37 in *Game theory and animal behavior*, edited by L. A. Dugatkin and K. R. Hudson, Oxford University Press, 1998.
- [Grimm and Klinge 1996] M. P. Grimm and M. Klinge, “Pike and some aspects of its dependence on vegetation”, pp. 125–156 in *Pike: biology and exploitation*, edited by J. F. Craig, Springer, Dordrecht, The Netherlands, 1996.
- [Hadjichrysanthou and Broom 2012] C. Hadjichrysanthou and M. Broom, “When should animals share food? Game theory applied to kleptoparasitic populations with food sharing”, *Behav. Ecol.* **23**:5 (2012), 977–991.
- [Hunter 1991] S. Hunter, “The impact of avian predator-scavengers on king penguin *Aptenodytes patagonicus* chicks at Marion Island”, *Ibis* **133**:4 (1991), 343–350.
- [Iyengar 2008] E. V. Iyengar, “Kleptoparasitic interactions throughout the animal kingdom and a re-evaluation, based on participant mobility, of the conditions promoting the evolution of kleptoparasitism”, *Biol. J. Linnean Soc.* **93**:4 (2008), 745–762.
- [Jeanne 1972] R. Jeanne, “Social biology of the neotropical wasp *Mischocyttarus drewseni*”, *Bull. Mus. Comp. Zool.* **144**:3 (1972), 63–150.
- [Källander 1977] H. Källander, “Piracy by black-headed gulls on lapwings”, *Bird Study* **24**:3 (1977), 186–194.
- [Kokko 2013] H. Kokko, “Dyadic contests: modelling fights between two individuals”, pp. 5–32 in *Animal contests*, edited by I. C. W. Hardy and M. Briffa, Cambridge University Press, 2013.
- [Kruuk 1972] H. Kruuk, *The spotted hyena: a study of predation and social behavior*, University of Chicago Press, 1972.
- [Skyrms 2004] B. Skyrms, *The stag hunt and the evolution of social structure*, Cambridge University Press, 2004.
- [Smith and Price 1973] J. M. Smith and G. R. Price, “The logic of animal conflict”, *Nature* **246**:5427 (1973), 15–18.
- [Spear et al. 1999] L. B. Spear, S. N. G. Howell, C. S. Oedekoven, D. Legay, and J. Bried, “Kleptoparasitism by brown skuas on albatrosses and giant-petrels in the Indian Ocean”, *The Auk* **116**:2 (1999), 545–548.

[Steele and Hockey 1995] W. K. Steele and P. A. R. Hockey, “Factors influencing rate and success of intraspecific kleptoparasitism among kelp gulls (*Larus dominicanus*)”, *The Auk* **112**:4 (1995), 847–859.

[Triplet et al. 1999] P. Triplet, R. A. Stillman, and J. D. Goss-Custard, “Prey abundance and the strength of interference in a foraging shorebird”, *J. Anim. Ecol.* **68**:2 (1999), 254–265.

[Vickery et al. 1991] W. L. Vickery, L.-A. Giraldeau, J. J. Templeton, D. L. Kramer, and C. A. Chapman, “Producers, scroungers, and group foraging”, *Am. Nat.* **137**:6 (1991), 847–863.

Received: 2015-08-14

Revised: 2016-07-25

Accepted: 2016-08-07

dgsykes@tamu.edu

*Department of Mathematics, Texas A&M University,  
College Station, TX 77843-3368, United States*

rychtar@uncg.edu

*Department of Mathematics and Statistics,  
The University of North Carolina at Greensboro,  
Greensboro, NC 27412, United States*



# Domination with decay in triangular matchstick arrangement graphs

Jill Cochran, Terry Henderson, Aaron Ostrander and Ron Taylor

(Communicated by Glenn Hurlbert)

We provide results for the exponential dominating numbers and total exponential dominating numbers of a family of triangular grid graphs. We then prove inequalities for these numbers and compare them with inequalities that hold more generally for exponential dominating numbers of graphs.

## 1. Introduction

A *dominating set* of a graph  $G$  is a set  $S \subseteq V(G)$  such that every  $v \in V(G)$  is either in  $S$  or is adjacent to a member of  $S$ . A *total dominating set* of a graph  $G$  is a set  $S \subseteq V(G)$  such that every  $v \in V(G)$  is adjacent to a member of  $S$ . The vertices in  $S$  are called *dominating vertices* or *dominators*, and a vertex adjacent to a dominator is said to be *dominated* by that dominator. In most kinds of domination a dominator is considered to dominate itself, but this is not the case for *total domination* where each dominator must be dominated by another dominator.

When considering domination at a distance, a *k-dominating set* of a graph  $G$  is a set  $S \subseteq V(G)$  such that every  $v \in V(G)$  is either in  $S$  or is a distance of  $k$  or less from any member of  $S$ . More examples of domination at a distance have been investigated in [Erwin 2004; Slater 1976].

In [Dankelmann et al. 2009] the authors introduce *exponential domination*, a variety of distance domination where the dominating power of a vertex decreases exponentially with the distance from that vertex. In this paper, we consider exponential domination and introduce a variation of exponential domination which we call *total exponential domination*. In the rest of the paper we sometimes talk about exponential domination or total exponential domination just in terms of domination when the context is clear.

For a connected graph  $G$  and  $S \subseteq V(G)$  we denote by  $G[S]$  the subgraph of  $G$  induced by  $S$ . For  $u \in S$  and  $v \in V(G) \setminus S$  we define  $d_S(u, v)$  to be the distance

---

*MSC2010:* 05A20, 05C69.

*Keywords:* domination, distance, triangular grid graphs.

Research was supported by the Berry College Department of Mathematics and Computer Science.

between  $u$  and  $v$  in  $G[V(G) \setminus (S \setminus \{u\})]$ ; i.e., minimum length paths do not include other dominators.

For exponential domination we use the same weight function as in [Dankelmann et al. 2009], given by

$$w_S(v) = \begin{cases} \sum_{u \in S} 2^{-d_S(u,v)+1}, & v \notin S, \\ 2, & v \in S. \end{cases}$$

For total exponential domination we use a similar weight function given by

$$w_S^t(v) = \begin{cases} \sum_{u \in S} 2^{-d_S(u,v)+1}, & v \notin S, \\ \sum_{u \in S, u \neq v} 2^{-d_S(u,v)+1}, & v \in S. \end{cases}$$

Note that the only difference between these two weight functions is that  $w_S(u) = 2$  but  $w_S^t(u)$  depends on the distribution of the other dominators for  $u \in S$ .

As in [Dankelmann et al. 2009], if for each  $v \in V(G)$  (or equivalently  $v \in V(G) \setminus S$ ) we have that  $w_S(v) \geq 1$ , then  $S$  is an *exponential dominating set* of  $G$ . The *exponential dominating number* of a graph  $G$ , denoted by  $\gamma_e(G)$ , is the smallest cardinality of an exponential dominating set of  $G$ . Similarly, if for each  $v \in V(G)$  we have that  $w_S^t(v) \geq 1$ , then  $S$  is a *total exponential dominating set* of  $G$ . The *total exponential dominating number* of a graph  $G$ , denoted by  $\gamma_{te}(G)$ , is the smallest cardinality of a total exponential dominating set of  $G$ . For an arbitrary  $S$  and arbitrary  $v \in V(G) \setminus S$ , if  $w_S(v) \geq 1$  or  $w_S^t(v) \geq 1$  then  $v$  is *exponentially dominated* or *totally exponentially dominated* by  $S$ .

We restrict ourselves to a particular family of *triangular grid graphs*. A triangular grid graph is a graph  $G$  such that  $V(G)$  can be put in a correspondence with points  $(x, y) = (\frac{1}{2}a - b, \frac{\sqrt{3}}{2}a)$ , where  $a, b \in \mathbb{Z}$ ; additionally, we require that in this correspondence two vertices can be adjacent only if their corresponding points are separated by unit distance (this is the same definition that is found in [Gordon et al. 2008]). We denote by  $G_n$  the graph whose vertices correspond with the points in

$$\{(\frac{1}{2}a - b, \frac{\sqrt{3}}{2}a) \mid a, b \in \mathbb{Z}, 0 \leq b \leq a \leq n\}$$

and which has as many edges as possible;  $G_n$  is called the *triangular matchstick arrangement graph of side  $n$* . This is the family of graphs which we consider in this paper. The *corners* of  $G_n$  are those vertices corresponding to  $a = b = 0$ ,  $a = b = n$ , and  $a = n, b = 0$ . The *perimeter* of  $G_n$  is the set of vertices and edges that lie on the minimal length paths between the corners. Any one of these minimal length paths is a *perimeter edge*; note that each perimeter edge of  $G_n$  contains  $n$  edges.

In Section 2 we determine the exponential dominating numbers for  $G_n$  up to  $n = 7$ . In Section 3 we provide upper bounds for exponential dominating numbers for arbitrary  $G_n$ . In Section 4 we determine the total exponential dominating numbers for  $G_n$  up to  $n = 5$ . In Section 5 we use arguments similar to those from Section 3 to provide upper bounds for total exponential dominating numbers for arbitrary  $G_n$ .



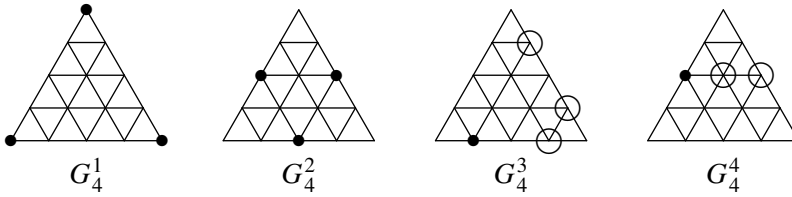


Figure 1. Graphs for Lemma 5.

2. Base cases for exponential domination

We use the following lemmas in proving Theorem 8.

Lemma 1.  $\gamma_e(G_n) \leq \gamma_e(G_{n+1})$ .

Lemma 2. *If there exists an arrangement of dominators that dominates  $G_n$  where a dominator is placed at a corner vertex, then the graph is also dominated by the arrangement of dominators produced by moving the corner dominator to a vertex adjacent to it and leaving the rest of the dominators in their original positions.*

Lemma 3.  $\gamma_e(G_1) = 1$ .

Lemma 4.  $\gamma_e(G_2) = 2$ .

*Proof.* To see that  $\gamma_e(G_2) \leq 2$ , note that picking any two vertices of  $G_2$  to be dominators suffices to dominate the graph.

Suppose  $\gamma_e(G_2) = 1$ . For every vertex in  $V(G_2)$  there is a second vertex that is a distance of 2 away. Thus no matter where the dominator is placed there always is one vertex with only a weight of  $\frac{1}{2}$ , so  $G_2$  is not dominated, which is a contradiction.  $\square$

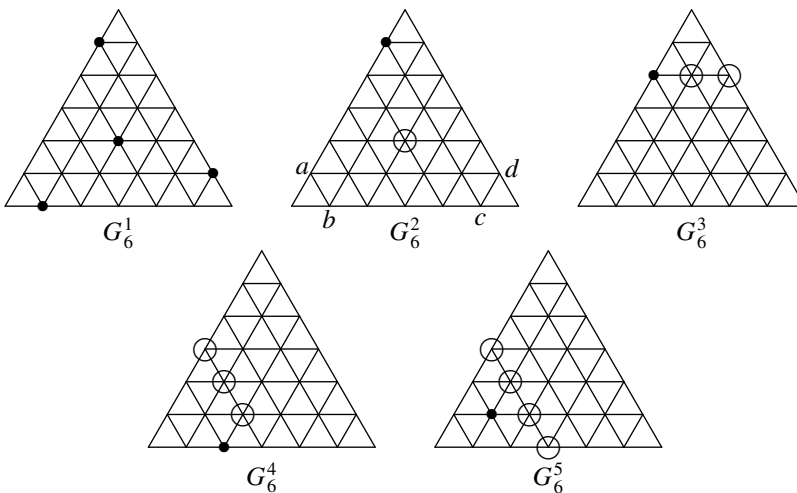
Lemma 5.  $\gamma_e(G_4) = 3$ .

*Proof.* The graphs  $G_4^i$  referred to in this proof are contained in Figure 1. To see that  $\gamma_e(G_4) \leq 3$ , consider  $G_4^1$  or  $G_4^2$  (from now on all vertices appearing as bullet points are dominators). If  $\gamma_e(G_4) < 3$  then we can dominate  $G_4$  with two dominators. We obviously must dominate the corners of  $G_4$ , and by Lemma 2 we can assume that no dominator is in a corner.

Supposing that the two dominators are at a distance of 1 from two corners (to ensure that at the least those corners are dominated), we produce graph  $G_4^3$ , where one of the circled vertices is also a dominator. The graph is not dominated in any of these cases.

Supposing that the two dominators are each at a distance of 2 from a single corner (to ensure that one corner is dominated), we produce  $G_4^4$ , where one of the circled vertices is also a dominator.  $G_4$  is not dominated in either case. This suffices to prove the lemma.  $\square$

Lemma 6.  $\gamma_e(G_6) = 4$ .



**Figure 2.** Graphs for Lemma 6.

*Proof.* The graphs  $G_6^i$  referred to in this proof are contained in Figure 2. To see that  $\gamma_e(G_6) \leq 4$ , consider  $G_6^1$ .

If  $\gamma_e(G_6) < 4$  then we can dominate  $G_6$  with three dominators. We first consider the case where each dominator is a distance of 1 from each corner. Doing so we produce  $G_6^2$ , where one of  $a$  or  $b$  and one of  $c$  or  $d$  is a dominator. In each of these cases the circled vertex is not dominated.

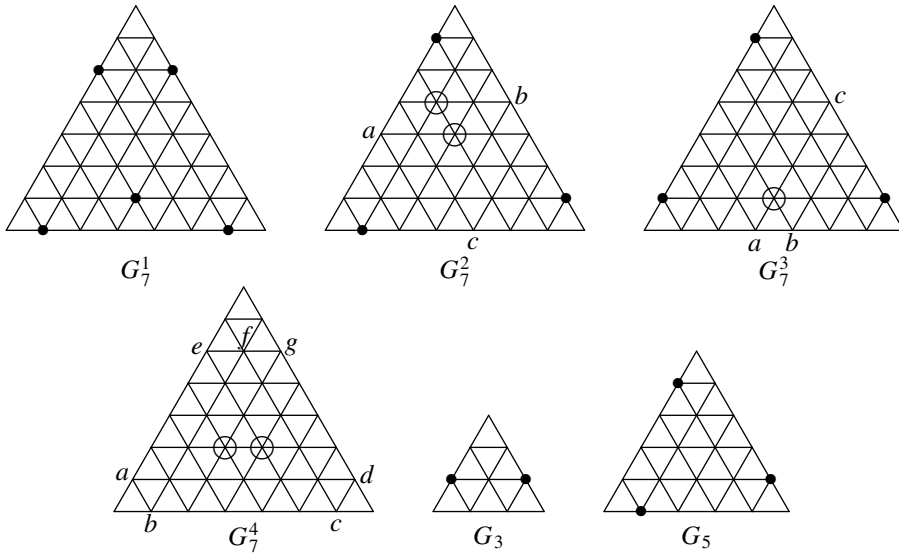
Considering next the case where one of the corners has two dominators at a distance of 2, we must place the third dominator on  $G_6^3$ , where one of the circled vertices is a dominator. We cannot place a third dominator in either of these cases so that all of the corners are dominated.

We now consider placing two dominators at a distance of 3 from a corner and the third dominator at a distance of 2 from the same corner. Doing so, we generate  $G_6^4$  or  $G_6^5$ , where two of the circled vertices are dominators. In any such graph only one of the corners is dominated. This suffices to prove the lemma.  $\square$

**Lemma 7.**  $\gamma_e(G_7) = 5.$

*Proof.* The graphs  $G_7^i$  referred to in this proof are contained in Figure 3. To see that  $\gamma_e(G_7) \leq 5$ , consider  $G_7^1$ . If  $\gamma_e(G_7) < 5$ , then four dominators suffice to dominate the graph. We first try to dominate  $G_7$  by placing three dominators so that each lies at a distance of 1 from each corner. Doing so, we produce  $G_7^2$  or  $G_7^3$ .

Notice that the vertices  $a, b$  and  $c$  in  $G_7^2$  have domination of  $\frac{17}{32}$  due to the first three dominators. So the fourth dominator must be placed at either of the circled vertices so that  $a$  and  $b$  will both have domination greater than 1. However, doing so, the domination of  $c$  is either  $\frac{21}{32}$  or  $\frac{25}{32}$ , so we cannot dominate  $G_7$  with four dominators by starting with  $G_7^2$ .



**Figure 3.** Graphs for Lemma 7 and Theorem 8.

Note that the vertices  $a$  and  $b$  in  $G_7^3$  have domination  $\frac{13}{32}$  due to the first three dominators. So the fourth dominator must be placed at the circled vertex in order for both  $a$  and  $b$  to have domination greater than 1. However, doing so, the domination of  $c$  is  $\frac{25}{32}$ , so we cannot dominate  $G_7$  with four dominators by starting with  $G_7^3$ .

We next try to dominate  $G_7$  by placing two dominators a distance of 1 from two corners and two other dominators a distance of 2 from the third corner. Doing so, we produce  $G_7^4$ , where one of  $a$  and  $b$ , one of  $c$  and  $d$ , and two of  $e$ ,  $f$ , and  $g$  are dominators. In each of these graphs one of the circled vertices fails to be dominated. This exhausts all of the ways that we can ensure that all of the corners are dominated, which suffices to prove the lemma.  $\square$

**Theorem 8.** *The exponential domination numbers for  $G_1$  through  $G_7$  are*

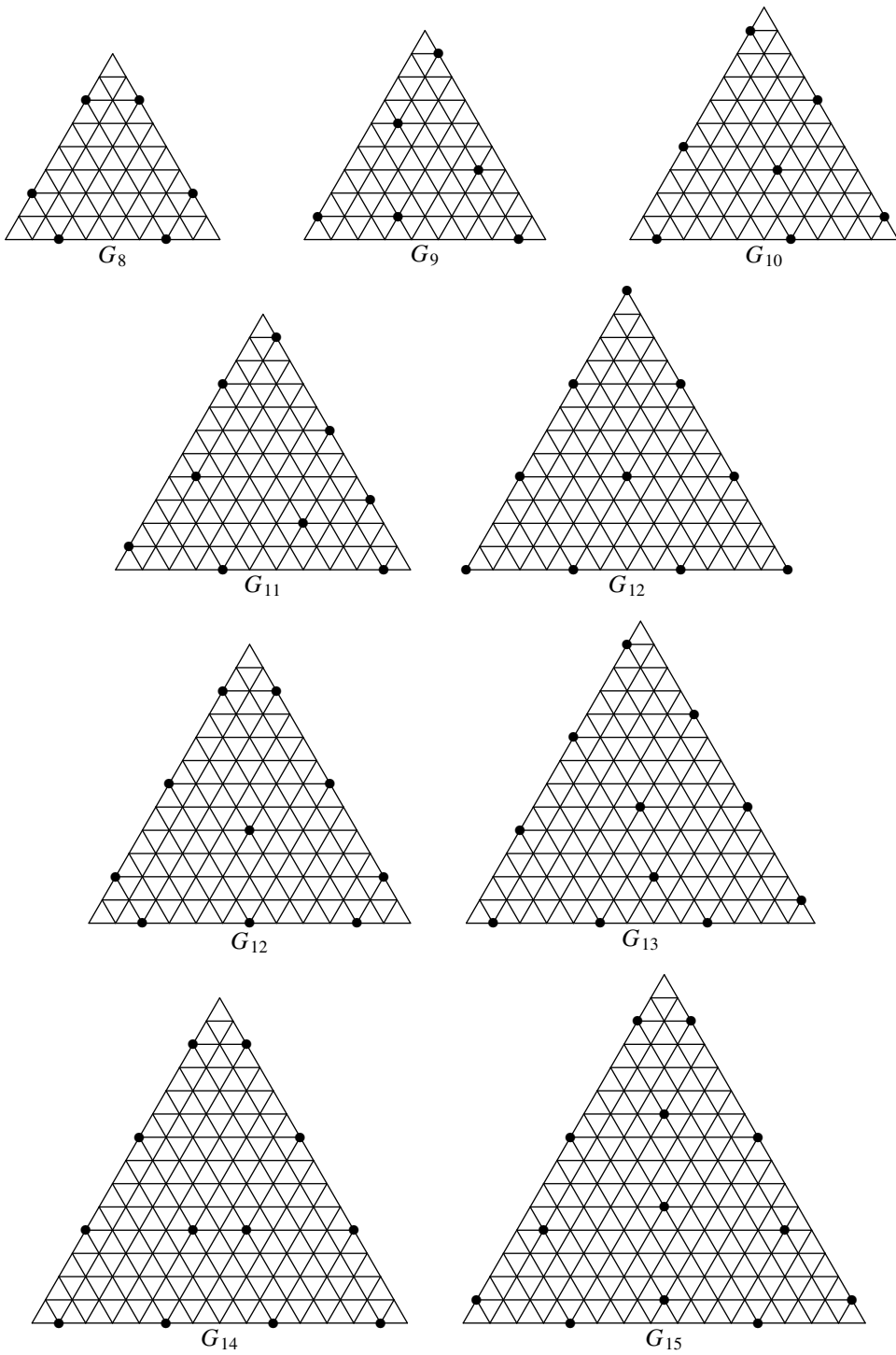
$n$	1	2	3	4	5	6	7
$\gamma_e(G_n)$	1	2	2	3	3	4	5

*Proof.* Lemmas 3–7 provide values for  $\gamma_e(G_n)$  for  $n \in \{1, 2, 4, 6, 7\}$ . To see that  $\gamma_e(G_3) \leq 2$ , consider  $G_3$  in Figure 3; by Lemmas 1 and 4,  $\gamma_e(G_3) = 2$ . To see that  $\gamma_e(G_5) \leq 3$ , consider  $G_5$  in Figure 3; by Lemmas 1 and 5 we see that  $\gamma_e(G_5) = 3$ .  $\square$

**Theorem 9.** *The exponential domination numbers for  $G_{10}$  through  $G_{15}$  are bounded as follows:*

$n$	8	9	10	11	12	13	14	15
$\gamma_e(G_n) \leq$	6	6	7	9	10	11	12	13

*Proof.* Consult Figure 4 for graphs that satisfy these bounds.  $\square$



**Figure 4.** Graphs for Theorem 9.

### 3. Inequalities for exponential domination

We will now determine the total exponential dominating numbers for  $G_n$  up to  $n = 5$ .

**Theorem 10** [Dankelmann et al. 2009]. *If  $G$  is a connected graph of size  $n$  then*

$$\gamma_e(G) \leq \frac{2}{5}(n + 2).$$

Applying this inequality to triangular grid graphs, we have the bound

$$\gamma_e(G_n) \leq \frac{2}{5} \left( \binom{n+2}{2} + 2 \right) = \frac{1}{5}(n^2 + 3n + 6).$$

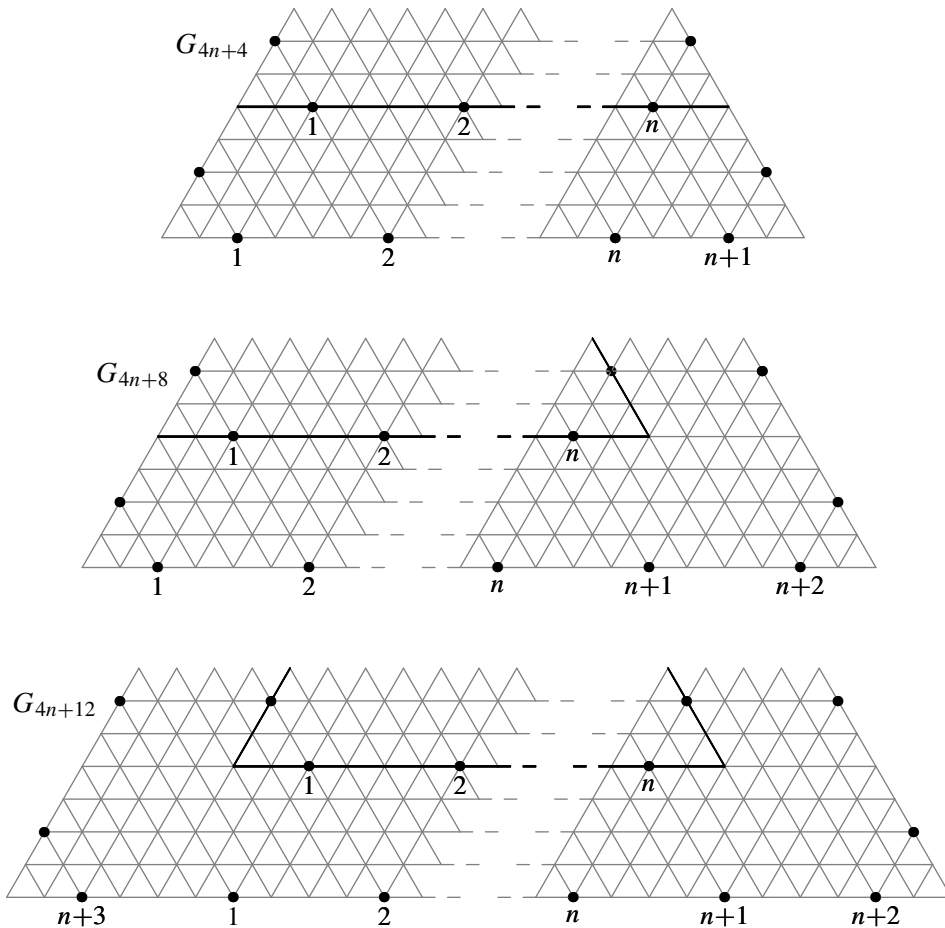
The theorem in [Dankelmann et al. 2009] that bounds  $\gamma_e(G)$  is established by considering an exponentially dominated spanning tree of  $G$  and not  $G$  itself. In establishing our bounds, we make use of the fact that  $G_{n+1}$  can be constructed from  $G_n$  by a set of elementary operations that are dependent on  $n$ . Our strongest bounds arise from considering how we can construct a distribution of dominators that dominates  $G_{n+r}$  based on a distribution of dominators that dominates  $G_n$ .

**Lemma 11.** *Suppose  $G_{4n}$  can be dominated by a set of  $m$  dominators where each of the corners are dominated by two dominators placed on the perimeter at a distance of 2 from each corner:*

- (1) *If dominators are placed along the rest of the perimeter edge between those two corners with a distance of 4 between each dominator, then  $G_{4n+4}$  can be dominated in a similar manner with  $m + n + 3$  dominators.*
- (2) *If dominators are placed along two of the perimeter edges with a distance of 4 between each dominator, then  $G_{4n+8}$  can be dominated in a similar manner by  $m + 2n + 6$  dominators.*
- (3) *If dominators are placed along the rest of the perimeter with a distance of 4 between each dominator, then  $G_{4n+12}$  can be dominated in a similar manner by  $m + 3n + 9$  dominators.*

*Proof.* (1) Consider the labeled  $G_{4n+4}$  implied by Figure 5. A labeled  $G_{4n}$  can be seen by removing the lower four rows of vertices from the  $G_{4n+4}$  with the lower perimeter of  $G_{4n}$ , including the row of vertices labeled  $\{1, 2, \dots, n\}$ . Both  $G_{4n}$  and  $G_{4n+4}$  have dominators placed as described in the hypotheses of the lemma. It can be confirmed that all of the vertices in the additional four rows of  $G_{4n+4}$  are dominated by this arrangement of dominators. If the  $G_{4n}$  is dominated by  $m$  dominators, then we have dominated  $G_{4n+4}$  by adding  $n + 1$  dominators along the lower perimeter and two dominators on the other perimeters; thus we have dominated  $G_{4n+4}$  with  $m + n + 3$  dominators.

(2) Similarly, consider the labeled  $G_{4n+8}$  implied by Figure 5. If  $G_{4n}$  is dominated by  $m$  dominators then we have dominated  $G_{4n+8}$  by adding  $n + 2$  dominators along



**Figure 5.** Graphs for Lemma 11.

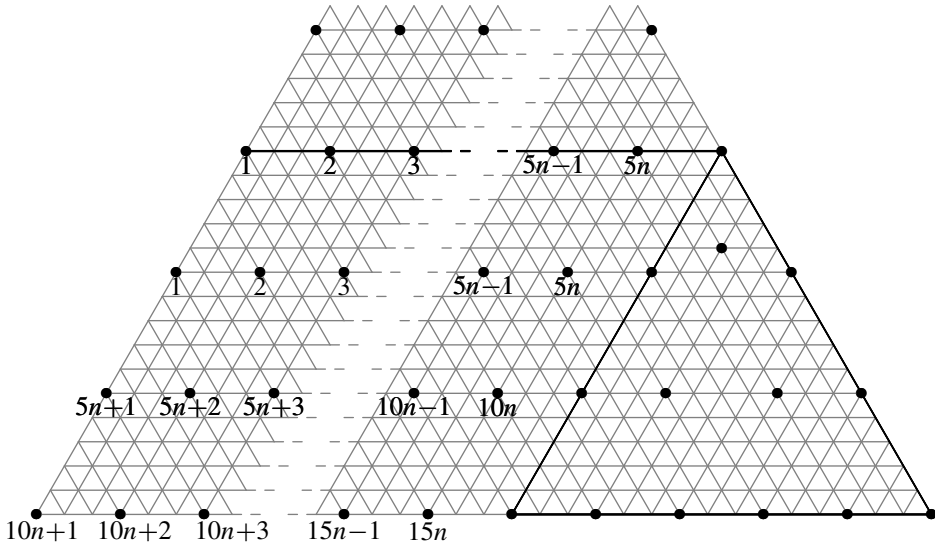
two perimeter edges and two dominators on the other perimeter edge; thus we have dominated  $G_{4n+8}$  with  $m + 2n + 6$  dominators.

(3) Similarly, consider the labeled  $G_{4n+12}$  implied by Figure 5. If  $G_{4n}$  is dominated by  $m$  dominators then we have dominated  $G_{4n+12}$  by adding  $n + 3$  dominators along each perimeter edge; thus we have dominated  $G_{4n+12}$  with  $m + 3n + 9$  dominators.  $\square$

**Lemma 12.** *If  $G_{15n}$  can be dominated by  $m$  dominators where*

- (1) *dominators are placed at two corners, and*
- (2) *along the perimeter edge between those corners dominators are placed with a distance of 3 between them,*

*then  $G_{15n+15}$  can be dominated in a similar manner by  $m + 13 + 15n$  dominators.*



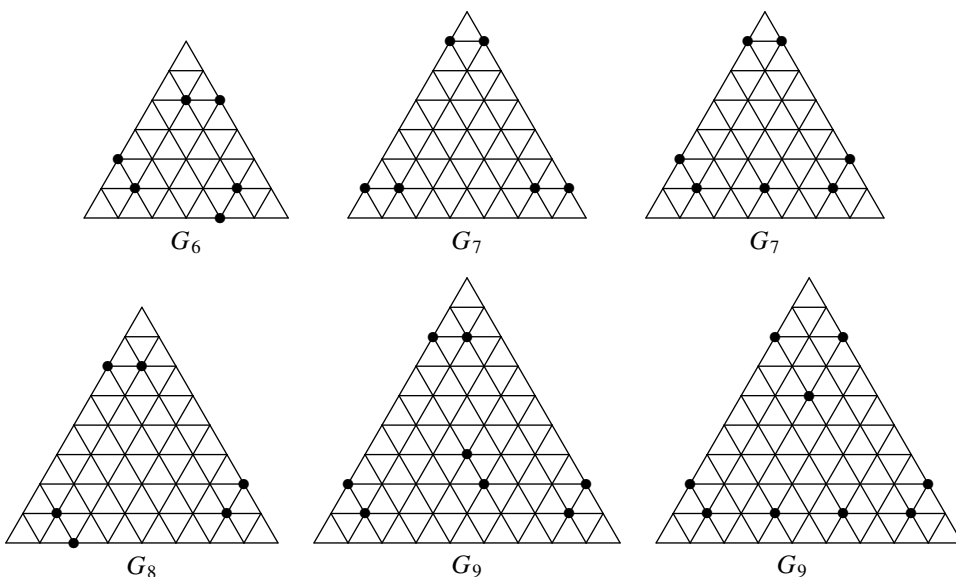
**Figure 6.** Graph for Lemma 12.

*Proof.* Consider the labeled  $G_{15n+15}$  implied by Figure 6. The lower perimeter edge of a  $G_{15n}$  is emboldened, as is part of the perimeter of a  $G_{15}$  dominated by 14 dominators. If dominators are placed in the lower 15 rows of  $G_{15n+15}$ , outside of the bold  $G_{15}$ , as suggested by the placement of the labeled dominators, then both  $G_{15n}$  and  $G_{15n+15}$  have dominator distributions as described in the hypotheses of the lemma. It can be confirmed that  $G_{15n+15}$  is dominated provided that  $G_{15n}$  is dominated. If  $G_{15n}$  is dominated by  $m$  dominators then we have dominated  $G_{15n+15}$  by placing  $15n$  dominators in a regular pattern in the lower 15 rows and 13 dominators in the remaining space; thus we have dominated  $G_{15n+15}$  with  $m + 13 + 15n$  dominators.  $\square$

Note that this lemma makes use of a  $G_{15}$  that is dominated using 14 dominators; however, from Theorem 9 we see that the exponential dominating number is at most 13. We use more dominators than necessary here in order to produce a consistent pattern of dominators along the lower perimeter edge of each subsequently constructed graph.

**Theorem 13.** *The following inequalities hold for  $n \geq 0$ :*

- (1)  $\gamma_e(G_{4n}) \leq \frac{1}{2}(n(n + 5)).$
- (2)  $\gamma_e(G_{4+8n}) \leq 2n^2 + 6n + 3.$
- (3)  $\gamma_e(G_{4+12n}) \leq \frac{1}{2}(9n^2 + 15n + 6).$
- (4)  $\gamma_e(G_{8+4n}) \leq \frac{1}{2}(n^2 + 9n + 12).$
- (5)  $\gamma_e(G_{8n}) \leq 2n(n + 2).$



**Figure 7.** Graphs for Theorem 13.

- (6)  $\gamma_e(G_{8+12n}) \leq \frac{1}{2}(9n^2 + 21n + 12)$ .
- (7)  $\gamma_e(G_{12+4n}) \leq \frac{1}{2}(n^2 + 11n + 20)$ .
- (8)  $\gamma_e(G_{12+8n}) \leq 2n^2 + 10n + 10$ .
- (9)  $\gamma_e(G_{12n}) \leq \frac{1}{2}(9n^2 + 9n + 2)$ .
- (10)  $\gamma_e(G_{15n}) \leq \frac{1}{2}(15n^2 + 11n + 2)$ .

*Proof.* We will prove inequalities (1)–(3) and (10); inequalities (4)–(9) can be proven by means similar to those used to prove (1)–(3) using the same lemmas. The dominated  $G_8$  used for inequalities (4)–(6) and the dominated  $G_{12}$  used for inequalities (7)–(9) can be found in Figure 4; the second  $G_{12}$  that appears in Figure 4 is the one we use because it is the only one that satisfies the hypotheses of Lemma 11.

From Figure 1 we see that  $G_4^2$  satisfies the hypotheses of Lemma 11. Lemma 5 also implies that  $\gamma_e(G_4) \leq 3$ , so we see that inequality (1) holds for the case where  $n = 1$ , and inequalities (2) and (3) hold for the case where  $n = 0$ .

Suppose that inequality (1) holds for all  $n$  up to some  $m$ ; also suppose that  $G_{4m}$  can be dominated in agreement with the hypotheses of Lemma 11 by a number of dominators less than or equal to the bound provided by inequality (1). Then  $\gamma_e(G_{4m}) \leq \frac{1}{2}(m^2 + 5m)$ , and by Lemma 11 we see that  $G_{4m+4}$  can be dominated by

$$\frac{1}{2}(m^2 + 5m) + m + 3 = \frac{1}{2}(m^2 + 7m + 6) = \frac{1}{2}(m + 1)(m + 6)$$

dominators. Thus  $\gamma_e(G_{4(m+1)}) \leq \frac{1}{2}(m + 1)((m + 1) + 5)$ , which proves inequality (1).



Suppose that inequality (2) holds for all  $n$  up to some  $m$ ; also suppose that  $G_{4+8m}$  can be dominated in agreement with the hypotheses of Lemma 11 by a number of dominators less than or equal to the bound provided by inequality (2). Then  $\gamma_e(G_{4+8m}) = \gamma_e(G_{4(1+2m)}) \leq 2m^2 + 6m + 3$ , and by Lemma 11 we see that  $G_{4(1+2m)+8}$  can be dominated by

$$(2m^2 + 6m + 3) + 2(1 + 2m) + 6 = 2m^2 + 10m + 11 = 2(m + 1)^2 + 6(m + 1) + 3$$

dominators. Thus

$$\gamma_e(G_{4(1+2m)+8}) = \gamma_e(G_{4+8(m+1)}) \leq 2(m + 1)^2 + 6(m + 1) + 3,$$

which proves inequality (2).

Suppose that inequality (3) holds for all  $n$  up to some  $m$ ; also suppose that  $G_{4+12m}$  can be dominated in agreement with the hypotheses of Lemma 11 by a number of dominators less than or equal to the bound provided by inequality (3). Then  $\gamma_e(G_{4+12m}) = \gamma_e(G_{4(1+3m)}) \leq \frac{1}{2}(9m^2 + 15m + 6)$ , and by Lemma 11 we see that  $G_{4(1+3m)+12}$  can be dominated by

$$\frac{1}{2}(9m^2 + 15m + 6) + 3(1 + 3m) + 9 = \frac{1}{2}(9m^2 + 33m + 30) = \frac{1}{2}(9(m + 1)^2 + 15(m + 1) + 6)$$

dominators. Thus

$$\gamma_e(G_{4(1+3m)+12}) = \gamma_e(G_{4+12(m+1)}) \leq \frac{1}{2}(9(m + 1)^2 + 15(m + 1) + 6),$$

which proves inequality (3).

From Figure 6 we see that  $G_{15}$  can be dominated by 14 dominators in a way that satisfies the hypotheses of Lemma 12. This implies that  $\gamma_e(G_{15}) \leq 14$ , so we see inequality 10 holds for the case where  $n = 1$ . Suppose that inequality (10) holds for all  $n$  up to some  $m$ ; also suppose that  $G_{15m}$  can be dominated in agreement with the hypotheses of Lemma 12 by a number of dominators less than or equal to the bound provided by inequality (10). Then  $\gamma_e(G_{15m}) \leq \frac{1}{2}(15m^2 + 11m + 2)$ , and by Lemma 12 we see that  $G_{15(m+1)}$  can be dominated by

$$\frac{1}{2}(15m^2 + 11m + 2) + 13 + 15m = \frac{1}{2}(15m^2 + 41m + 28) = \frac{1}{2}(15(m + 1)^2 + 11(m + 1) + 2)$$

dominators. Thus,

$$\gamma_e(G_{15(m+1)}) \leq \frac{1}{2}(15(m + 1)^2 + 11(m + 1) + 2),$$

which proves inequality (10). □

This provides us with the following corollary.

**Corollary 14.** *The following inequalities hold for  $n \in \mathbb{Z}^+$  as specified:*

- (1)  $\gamma_e(G_n) \leq \frac{1}{32}(n^2 + 12n + 32)$ , where  $n \bmod 4 = 0$ .
- (2)  $\gamma_e(G_n) \leq \frac{1}{30}(n^2 + 11n + 30)$ , where  $n \bmod 15 = 0$ .

The first inequality here is implied by inequalities (3), (6), and (9) in Theorem 13, and the second inequality is implied by inequality (10). The other inequalities from Theorem 13 do not provide bounds that are as good as these.

#### 4. Base cases for total exponential domination

The following lemma is the analogue of Lemma 1 for total exponential domination.

**Lemma 15.**  $\gamma_{te}(G_n) \leq \gamma_{te}(G_{n+1})$ .

**Lemma 16.**  $\gamma_{te}(G_1) = 2$ .

*Proof.* To see that  $\gamma_{te}(G_1) \leq 2$ , note that picking any two vertices as dominators suffices to dominate  $G_1$ . To see that  $\gamma_{te}(G_1) \neq 1$ , note that a single dominator can never dominate an entire graph.  $\square$

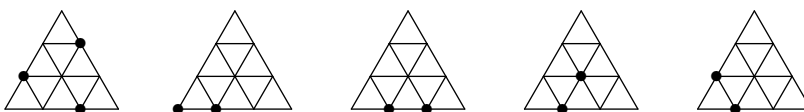
**Lemma 17.**  $\gamma_{te}(G_3) = 3$ .

*Proof.* To see that  $\gamma_{te}(G_3) \leq 3$ , consider the first graph in Figure 8. Suppose that  $\gamma_{te}(G_3) < 3$ . Then the graph is dominated with two dominators. In a graph with only two dominators, the dominators must be adjacent since otherwise both of them will not have weight greater than 1. Any  $G_3$  with two adjacent dominators will be one of the graphs shown in Figure 8, none of which is dominated.  $\square$

**Lemma 18.**  $\gamma_{te}(G_5) = 5$ .

*Proof.* The graphs referred to in this proof appear in Figure 9. To see that  $\gamma_{te}(G_5) \leq 5$ , consider the graph  $G_5$ . Supposing that  $\gamma_{te}(G_5) < 5$ , we can dominate the graph with four dominators. Since each corner must have a weight greater than or equal to 1, we organize this proof according to the ways that corners can be dominated by the fewest number of dominators. Note that 1 can be written as a sum of four or fewer powers of  $\frac{1}{2}$  (not necessarily unique) with numerators of 1 in the following five ways:  $1$ ,  $2(\frac{1}{2})$ ,  $\frac{1}{2} + 2(\frac{1}{4})$ ,  $\frac{1}{2} + \frac{1}{4} + 2(\frac{1}{8})$ ,  $4(\frac{1}{4})$ . We consider each of these possible combinations of weights separately.

**1:** One way for all of the corners to have weight at least 1 is to place dominators at a distance of 1 from each of the corners. Doing so, we produce either  $G_5^1$  or  $G_5^2$ . In these graphs each dominator has weight less than  $\frac{1}{2}$ , so in order for the dominators to be dominated we must place another dominator no more than a distance of 1 away from each. It is not possible to do this with a single dominator, so there is



**Figure 8.** Graphs for Lemma 17.

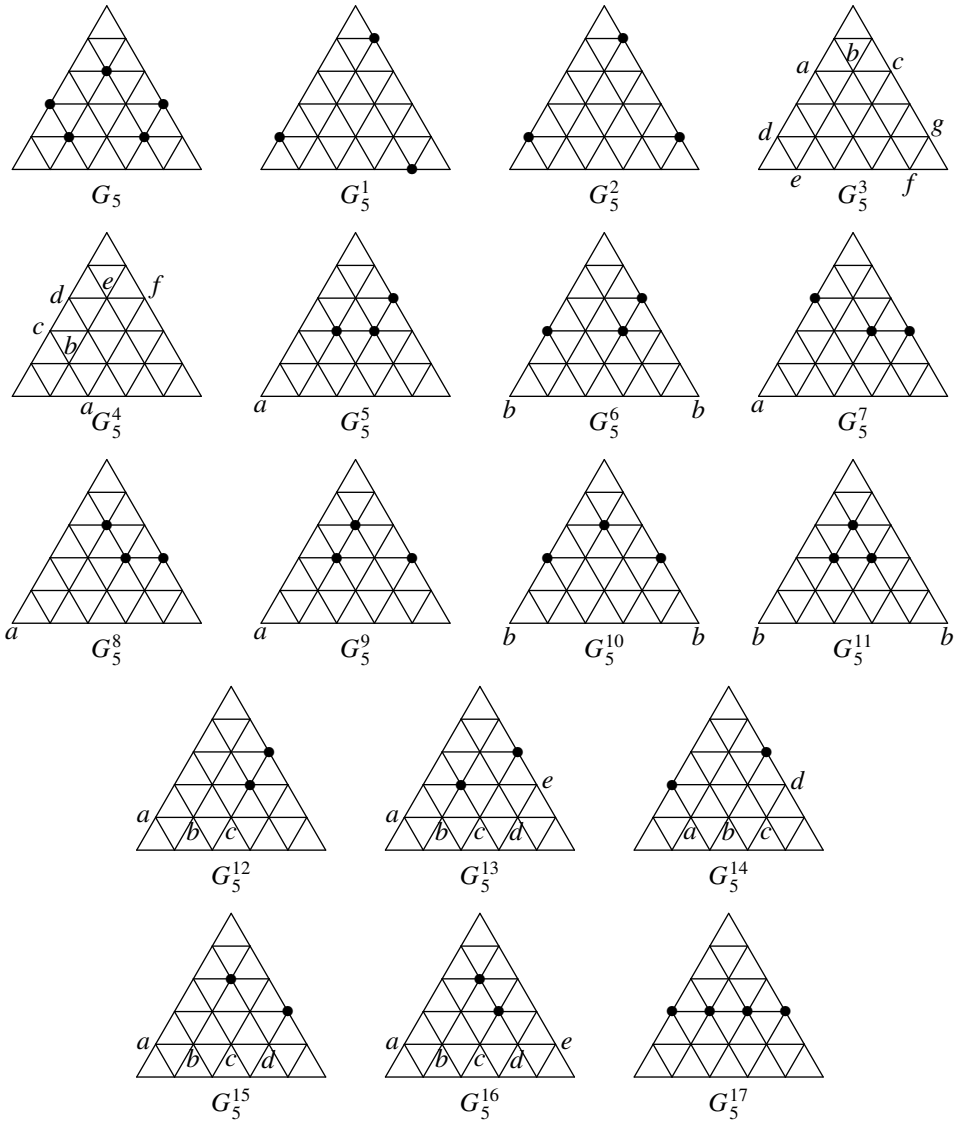


Figure 9. Graphs for Lemma 18.

not a configuration of dominators that dominates the graph where each corner is adjacent to a dominator.

$2(\frac{1}{2})$ : Another method to dominate all of the corners is to place two dominators a distance of 2 away from one corner and to place the other dominators adjacent to the other corners. Doing this we produce  $G_5^3$ , where two of  $a$ ,  $b$ , and  $c$  are dominators, one of  $d$  and  $e$  is a dominator, and one of  $f$  and  $g$  is a dominator. It can be confirmed that in each of these cases the graph fails to be dominated.

The next way to consider having corners dominated is to place two dominators at a distance of 2 from one corner, and to do the same for a second corner. Doing so we produce  $G_5^4$ , where two of  $a, b$ , and  $c$  are dominators and two of  $d, e$ , and  $f$  are dominators. It can be confirmed that in each of these cases the graph fails to be dominated.

$\frac{1}{2} + 2(\frac{1}{4})$ : The third case involves dominating a single corner by placing one dominator at a distance of 2 and two dominators at a distance of 3 in such a way that the dominators don't interfere with one another. If we begin by doing this for the top corner, we make one of the graphs from  $G_5^5$  to  $G_5^{11}$ . In those graphs with vertices labeled  $a$ , each labeled vertex has domination less than  $\frac{1}{2}$ , so a dominator must be placed adjacent to it. It is easy to confirm that doing so will never suffice to dominate the graph by considering the lower corner vertex opposite to the labeled vertex. In those graphs with vertices labeled  $b$ , the weight of each labeled vertex is at least  $\frac{1}{2}$  but less than  $\frac{3}{4}$ , so a dominator must be placed at distance of 2 or closer. Since the labeled vertices are a distance of 5 apart this is not possible.

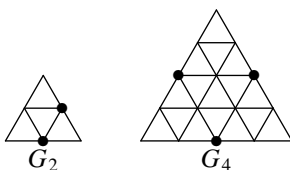
$\frac{1}{2} + \frac{1}{4} + 2(\frac{1}{8})$ : The only other cases that need to be considered are those in which all four dominators are used to dominate a single corner. This can be achieved by placing dominators in a configuration with one dominator at a distance of 2, one dominator at a distance of 3, and two dominators at a distance of 4 (using  $d_5(u, v)$ ). Doing so, we produce one graph from  $G_5^{12}$  to  $G_5^{16}$ , where two of the vertices labeled by letters are dominators. It can be confirmed that in each case the graph fails to be dominated (this can be easily done by considering domination of the other two corners).

$4(\frac{1}{4})$ : If we try to dominate  $G_5$  using four dominators all at a distance of 3 from one of the corners then we produce  $G_5^{17}$ , which is not dominated. □

**Theorem 19.** *The total exponential domination numbers for  $G_1$  through  $G_5$  are*

$n$	1	2	3	4	5
$\gamma_{te}(G_n)$	2	2	3	3	5

*Proof.* Lemmas 16–18 provide  $\gamma_{te}(G_n)$  for  $n \in \{1, 3, 5\}$ . To see that  $\gamma_{te}(G_2) \leq 2$ , consider the graph in Figure 10; by Lemmas 15 and 16 we see that  $\gamma_{te}(G_2) = 2$ . To see that  $\gamma_{te}(G_4) \leq 3$ , consider the graph in Figure 10; by Lemmas 15 and 17 we see that  $\gamma_{te}(G_4) = 3$ . □



**Figure 10.** Graphs for Theorem 19.

**Theorem 20.** *The total exponential domination numbers for  $G_6$  through  $G_9$  are bounded as follows:*

$n$	6	7	8	9
$\gamma_{1e}(G_n) \leq$	6	6	6	8

*Proof.* Consult Figure 7 for graphs that satisfy these bounds. □

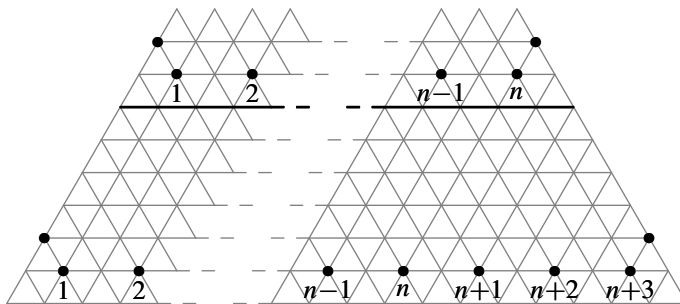
### 5. Inequalities for total exponential domination

**Lemma 21.** *If  $G_{2n+1}$  can be dominated by a set of  $m$  dominators so that there exists a subgraph of  $G_{2n+1}$  isomorphic to  $G_{2n}$  that contains all of the dominators and such that*

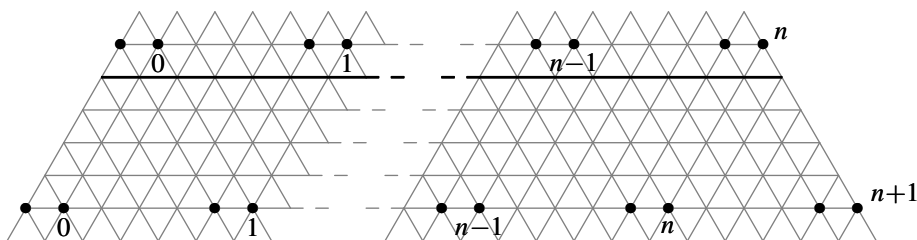
- (1) *two of the corners of the  $G_{2n}$  subgraph are adjacent to two dominators, and*
- (2) *dominators are placed along the rest of perimeter edge between those two corners with a distance of 2 between each dominator,*

*then  $G_{2n+7}$  can be dominated in a similar way by  $m + n + 5$  dominators.*

*Proof.* Consider the labeled  $G_{2n+7}$  implied by Figure 11. The lower perimeter edge of a  $G_{2n+1}$  has been emboldened. The lower perimeter edge of a  $G_{2n+6}$  subgraph corresponds with the second-lowest set of edges and vertices, including the vertices labeled by  $\{1, 2, \dots, n + 2, n + 3\}$ ; this graph has dominators placed as described above. A labeled  $G_{2n}$  can be produced by removing the lower seven rows of vertices from the  $G_{2n+7}$ ; this  $G_{2n}$  has dominators placed as described above and is a subgraph of the  $G_{2n+1}$  whose lower perimeter edge is bold. It can be confirmed that all of the vertices in the lower six rows are dominated by an arrangement of dominators like the one depicted in Figure 11. If  $G_{2n+1}$  is dominated by  $m$  dominators then we have dominated  $G_{2n+7}$  by adding a total of  $n + 5$  dominators, thereby dominating  $G_{2n+7}$  with  $m + n + 5$  dominators. □



**Figure 11.** Graph for Lemma 21.



**Figure 12.** Graph for Lemma 22.

**Lemma 22.** *If  $G_{5n+2}$  can be dominated by a set of  $m$  dominators so that there exists a subgraph of  $G_{5n+2}$  isomorphic to  $G_{5n+1}$  that contains all of the dominators and such that*

- (1) *two of the corners of the subgraph are dominators,*
- (2) *along the perimeter edge of the subgraph between those dominators there are two dominators adjacent to each of the above-mentioned dominators, and*
- (3) *along the rest of the perimeter there occur pairs of adjacent dominators with a distance of 4 between each pair,*

*then  $G_{5n+7}$  can be dominated in a similar way by  $m + 4 + 2n$  dominators.*

*Proof.* Consider the labeled  $G_{5n+7}$  implied by Figure 12. The lower perimeter edge of a  $G_{5n+2}$  subgraph has been emboldened. The lower perimeter edge of a  $G_{5n+6}$  subgraph corresponds with the second-lowest set of edges and vertices, including the vertices labeled by  $\{1, 2, \dots, n, n + 1\}$ ; this graph has dominators placed as described above. A labeled  $G_{5n+1}$  can be produced by removing the lower six rows of vertices from the  $G_{5n+7}$ ; this graph has dominators placed as described above and is a subgraph of the  $G_{5n+2}$  whose lower perimeter edge has been emboldened. It can be confirmed that all of the vertices in the lower five rows are dominated by an arrangement of dominators like the one depicted in Figure 12. If the  $G_{5n+2}$  is dominated by  $m$  dominators then we have dominated  $G_{5n+7}$  by adding a total of  $2n + 4$  dominators, thereby dominating  $G_{5n+7}$  with  $m + 4 + 2n$  dominators.  $\square$

**Theorem 23.** *The following inequalities hold for  $n \geq 0$ :*

- (1)  $\gamma_{te}(G_{5+6n}) \leq \frac{1}{2}(3n^2 + 11n + 10).$
- (2)  $\gamma_{te}(G_{7+6n}) \leq \frac{1}{2}(3n^2 + 13n + 14).$
- (3)  $\gamma_{te}(G_{9+6n}) \leq \frac{1}{2}(3n^2 + 15n + 18).$
- (4)  $\gamma_{te}(G_{2+5n}) \leq (n + 1)(n + 2).$

*Proof.* We will prove inequalities (1) and (4). The proofs for inequalities (2) and (3) are similar to the proof of inequality (1). The dominated  $G_7$  and  $G_9$  used to

prove inequalities (2) and (3) can be found in Figure 7; we use the second  $G_7$  and second  $G_9$  that appear because they are the only ones that satisfy the hypotheses of Lemma 21.

From Figure 9 we see that  $G_5$  can be dominated by five dominators in a way that satisfies the hypotheses of Lemma 21. This implies that  $\gamma_{te}(G_5) \leq 5$ , so we see that inequality (1) is satisfied for the case where  $n = 0$ . Suppose that inequality (1) holds for all  $n$  up to some  $m$ ; also suppose that  $G_{5+6m}$  can be dominated in agreement with the hypotheses of Lemma 21 by a number of dominators less than or equal to the bound provided by inequality (1). Then  $\gamma_{te}(G_{5+6m}) = \gamma_{te}(G_{2(3m+2)+1}) \leq \frac{1}{2}(3m^2 + 11m + 10)$ , and by Lemma 21 we see that  $G_{5+6(m+1)}$  can be dominated by

$$\begin{aligned} \frac{1}{2}(3m^2 + 11m + 10) + (3m + 2) + 5 &= \frac{1}{2}(3m^2 + 17m + 24) \\ &= \frac{1}{2}(3(m + 1)^2 + 11(m + 1) + 10) \end{aligned}$$

dominators. Thus

$$\gamma_{te}(G_{5+6(m+1)}) \leq \frac{1}{2}(3(m + 1)^2 + 11(m + 1) + 10),$$

which proves inequality (1).

From Figure 10 we see that  $G_2$  can be dominated in a way that satisfies the hypotheses of Lemma 22. This implies that  $\gamma_{te}(G_2) \leq 2$ , so we see that inequality (4) holds in the case where  $n = 0$ . For some  $m > 0$  suppose that  $G_{2+5m}$  can be dominated in agreement with the hypotheses of Lemma 22 by a number of dominators less than or equal to the bound provided by inequality (4). Then  $\gamma_{te}(G_{2+5m}) \leq m^2 + 3m + 2$ , and by Lemma 22,  $G_{7+5m}$  can be dominated by

$$(m^2 + 3m + 2) + 4 + 2m = m^2 + 5m + 6 = (m + 1)^2 + 3(m + 1) + 2$$

dominators. Thus

$$\gamma_{te}(G_{7+5m}) = \gamma_{te}(G_{2+5(m+1)}) \leq (m + 1)^2 + 3(m + 1) + 2,$$

which proves inequality (4). □

**Corollary 24.** *The following inequalities hold for  $n$  as specified:*

- (1)  $\gamma_{te}(G_n) \leq \frac{1}{24}(n^2 + 12n + 35)$ , where  $n$  is odd and  $n \geq 5$ .
- (2)  $\gamma_{te}(G_n) \leq \frac{1}{25}(n^2 + 11n + 24)$ , where  $n \bmod 5 = 2$ .

### 6. Conclusion

In this paper we have proven the values of  $\gamma_e(G_n)$  for  $n \leq 7$  and  $\gamma_{te}(G_n)$  for  $n \leq 5$ . We also provided bounds on  $\gamma_e(G_n)$  for  $n \leq 15$  and  $\gamma_{te}(G_n)$  for  $n \leq 9$ . We made use of the regular structure of triangular matchstick arrangement graphs to establish bounds on  $\gamma_e(G_n)$  and  $\gamma_{te}(G_n)$  for arbitrary  $n$ . The constructive methods

we used produced inequalities that are significantly tighter than those found in [Dankelmann et al. 2009]. These techniques are particularly promising since the family of triangular grid graphs is just one family of graphs where  $G_{n+1}$  can be constructed from  $G_n$  by adding edges and vertices in a regularly defined manner. Similar methods could be used with recursively constructible families of graphs (studied in [Noy and Ribó 2004]) and regular  $n$ -gon grid graphs, such as square grid graphs, as in [Gonçalves et al. 2011].

### References

- [Dankelmann et al. 2009] P. Dankelmann, D. Day, D. Erwin, S. Mukwembi, and H. Swart, “Domination with exponential decay”, *Discrete Math.* **309**:19 (2009), 5877–5883. MR Zbl
- [Erwin 2004] D. J. Erwin, “Dominating broadcasts in graphs”, *Bull. Inst. Combin. Appl.* **42** (2004), 89–105. MR Zbl
- [Gonçalves et al. 2011] D. Gonçalves, A. Pinlou, M. Rao, and S. Thomassé, “The domination number of grids”, *SIAM J. Discrete Math.* **25**:3 (2011), 1443–1453. MR Zbl
- [Gordon et al. 2008] V. S. Gordon, Y. L. Orlovich, and F. Werner, “Hamiltonian properties of triangular grid graphs”, *Discrete Math.* **308**:24 (2008), 6166–6188. MR Zbl
- [Noy and Ribó 2004] M. Noy and A. Ribó, “Recursively constructible families of graphs”, *Adv. in Appl. Math.* **32**:1–2 (2004), 350–363. MR Zbl
- [Slater 1976] P. J. Slater, “ $R$ -domination in graphs”, *J. Assoc. Comput. Mach.* **23**:3 (1976), 446–450. MR Zbl

Received: 2015-09-11      Revised: 2016-07-14      Accepted: 2016-07-24

jcochran@berry.edu	<i>Department of Mathematics and Computer Science, Berry College, Mount Berry, GA 30149, United States</i>
kenneth.t.henderson@gmail.com	<i>Department of Mathematics and Statistics, Wake Forest University, Winston-Salem, NC 27109, United States</i>
the.aaron.ostrander@gmail.com	<i>Department of Physics, University of Maryland, College Park, MD 20742, United States</i>
rtaylor@berry.edu	<i>Department of Mathematics and Computer Science, Berry College, Mount Berry, GA 30149, United States</i>



# On the tree cover number of a graph

Chassidy Bozeman, Minerva Catral, Brendan Cook,  
Oscar E. González and Carolyn Reinhart

(Communicated by Anant Godbole)

Given a graph  $G$ , the tree cover number of the graph, denoted  $T(G)$ , is the minimum number of vertex disjoint simple trees occurring as induced subgraphs that cover all the vertices of  $G$ . This graph parameter was introduced in 2011 as a tool for studying the maximum positive semidefinite nullity of a graph, and little is known about it. It is conjectured that the tree cover number of a graph is at most the maximum positive semidefinite nullity of the graph.

In this paper, we establish bounds on the tree cover number of a graph, characterize when an edge is required to be in some tree of a minimum tree cover, and show that the tree cover number of the  $d$ -dimensional hypercube is 2 for all  $d \geq 2$ .

## 1. Introduction

A *simple graph* is a pair  $G = (V, E)$ , where  $V = \{1, 2, \dots, n\}$  is the vertex set, and  $E$ , the edge set, is a set of 2-element subsets (edges) of the vertices. A *multigraph* is a pair  $G = (V, E)$ , where  $V = \{1, 2, \dots, n\}$ , and  $E$  is a multiset of 2-element subsets of the vertices. That is, a multigraph allows multiple edges between a pair of vertices (note that all simple graphs are multigraphs). Two vertices  $u, v \in V(G)$  are said to be *adjacent* if  $\{u, v\} \in E(G)$ . We say that the edge  $\{u, v\} \in E(G)$  is a *simple edge* if  $\{u, v\}$  appears in  $E(G)$  exactly once. If  $\{u, v\}$  appears in  $E(G)$  more than once, then it is a *multiedge*. All graphs in this paper are considered to be multigraphs unless otherwise stated.

For a multigraph  $G$ ,  $\mathcal{S}(G)$  denotes *the set of real valued symmetric  $n \times n$  matrices  $(a_{i,j})$  satisfying:*

- (1)  $a_{i,j} = 0$  if  $i \neq j$  and  $i, j$  are nonadjacent,
- (2)  $a_{i,j} \neq 0$  if  $i \neq j$  and  $i, j$  are adjacent via one edge, and
- (3)  $a_{i,j} \in \mathbb{R}$  if  $i = j$  or  $i, j$  are adjacent via multiple edges.

*MSC2010:* 05C05, 05C50, 05C76.

*Keywords:* tree cover number, hypercube, maximum nullity, minimum rank.

Research supported by NSF DSM-1457443.

The *maximum nullity* of a multigraph  $G$  is defined to be

$$M(G) = \max\{\text{null}(A) : A \in \mathcal{S}(G)\}.$$

The maximum nullity of a simple graph  $G$  is equivalent to the maximum multiplicity of an eigenvalue among all matrices in  $S(G)$ . This graph parameter has connections to many other concepts in linear algebra (as can be seen in [Fallat and Hogben 2007; 2014]), and has been given a significant amount of consideration as it is very difficult to compute.

A related and equally important parameter is the maximum positive semidefinite nullity of a graph. A symmetric  $n \times n$  real matrix  $A$  is said to be *positive semidefinite* if  $x^T Ax \geq 0$  for all  $x \in \mathbb{R}^n$ . The *maximum positive semidefinite nullity* of a multigraph  $G$  is defined to be

$$M_+(G) = \max\{\text{null}(A) : A \in \mathcal{S}_+(G)\},$$

where  $\mathcal{S}_+(G) = \{A \in \mathcal{S}(G) : A \text{ is positive semidefinite}\}$ . It follows that for a multigraph  $G$ ,  $M_+(G) \leq M(G)$ . In some cases, one can use tools such as orthogonal representations (see [Fallat and Hogben 2014]) to compute  $M_+(G)$ , obtaining a lower bound for  $M(G)$ .

The tree cover number of a graph was introduced in [Barioli et al. 2011] as another tool for studying the maximum positive semidefinite nullity of a multigraph.

The (*simple*) *path* on  $n$  vertices, denoted  $P_n$ , is the graph with vertex set  $V(P_n) = \{1, \dots, n\}$  and edge set  $E(P_n) = \{\{i, i + 1\} \mid i \in 1, \dots, n - 1\}$ . A simple graph  $G = (V, E)$  is said to be a *tree* if for every  $u, v \in V(G)$ , there is exactly one path from  $u$  to  $v$ .

Given a graph  $G = (V, E)$ , a *subgraph*  $G' = (V', E')$  is a graph such that  $V(G') \subseteq V(G)$  and  $E(G') \subseteq E(G)$ , i.e., a subgraph of a graph  $G$  can be obtained by deleting edges and vertices (and edges incident to the deleted vertices) of  $G$ . A subgraph  $G' = (V', E')$  of  $G$  is said to be an *induced* subgraph of  $G$  if for each edge  $uv \in E(G)$  with  $u, v \in V(G')$ , it follows that  $uv \in E(G')$ , i.e., an induced subgraph of  $G$  can be obtained by only deleting vertices (and any edges incident to the deleted vertices). For a subset  $S \subseteq V(G)$ , the *graph induced by  $S$* , denoted  $G[S]$ , is the induced subgraph of  $G$  with vertex set  $S$ .

A *tree cover* is a set of vertex disjoint simple trees occurring as induced subgraphs that cover all the vertices of the graph. The *tree cover number* of a graph  $G$ , denoted  $T(G)$ , is defined as

$$T(G) = \min\{|\mathcal{T}| : \mathcal{T} \text{ is a tree cover of } G\}.$$

**Conjecture 1** [Barioli et al. 2011].  $T(G) \leq M_+(G)$ .

This bound has been proven to be true for several families of graphs, including outerplanar graphs and chordal graphs [Barioli et al. 2011]. In fact, in the previous

work, the authors showed that equality holds for outerplanar graphs (and in fact for all graphs of tree-width at most 2, as observed in [Ekstrand et al. 2012]).

In Section 2 we give bounds on the tree cover number, provide an example in which the tree cover number behaves like the maximum positive semidefinite nullity, and provide an example in which the tree cover number does not behave like the maximum positive semidefinite nullity; see [Barioli et al. 2011; Ekstrand et al. 2012] for definitions of outerplanar and tree-width. In Section 3, we characterize when an edge is required to be in some tree of a minimum tree cover. In Section 4, we prove that the tree cover number of the  $d$ -dimensional hypercube is 2 for all  $d \geq 2$ .

**1.1. More notation and terminology.** The *cycle* on  $n$  vertices, denoted  $C_n$ , is the graph with vertex set  $V(C_n) = \{1, \dots, n\}$  and edge set

$$E(C_n) = \{\{i, i + 1\} \mid i \in 1, \dots, n - 1\} \cup \{1, n\}.$$

The *star*  $K_{1,n}$  is the graph with vertex set  $\{1, \dots, n\}$  and edge set  $\{\{1, j\} \mid j \in \{2, \dots, n\}\}$ . The *complete graph*, denoted  $K_n$ , is the graph on  $n$  vertices such that there is an edge between any two vertices.

A graph is said to be *connected* if there is a path from any vertex to any other vertex. If  $G$  is not connected, then it is said to be *disconnected*. Given a graph  $G = (V, E)$ , a *connected component* of  $G$  is a subgraph  $C$ , where  $C$  is connected and no vertex in  $C$  is adjacent to any vertex of  $V(G) \setminus V(C)$ . A graph is said to be a *forest* if each of its connected components is a tree.

If vertices  $u$  and  $v$  are adjacent, we say that they are *neighbors*. The neighborhood of a vertex  $v$ , denoted  $N(v)$ , is the set of neighbors of  $v$ . The degree of  $v$  is given by  $\deg(v) = |N(v)|$ .

For a graph  $G = (V, E)$ , a *cover* of  $G$  is a partition of  $V(G)$ . An *independent set*  $S$  is a subset of  $V(G)$  such that no two vertices in  $S$  are adjacent. The *independence number* of  $G$ , denoted  $\alpha(G)$ , is defined by

$$\alpha(G) = \max\{|S| : S \text{ is an independent set in } G\}.$$

Given two simple graphs  $G$  and  $H$ , the *cartesian product* of  $G$  and  $H$ , denoted  $G \times H$ , is the graph whose vertex set is the cartesian product  $V(G) \times V(H)$ , and any two vertices  $(u, u')$  and  $(v, v')$  are adjacent in  $G \times H$  if and only if either  $u = v$  and  $u'$  is adjacent to  $v'$  in  $H$ , or  $u' = v'$  and  $u$  is adjacent to  $v$  in  $G$ . The union of  $G$  and  $H$ , denoted  $G \cup H$ , is the graph with vertex set  $V(G) \cup V(H)$  and edge set  $E(G) \cup E(H)$ .

Throughout this paper, we often denote an edge  $\{u, v\}$  by  $uv$ . An edge  $uv$  is called a *bridge* of  $G$  if  $C - uv$  is disconnected, where  $C$  is the component of  $G$  with  $uv \in E(C)$  and  $C - uv$  denotes the subgraph obtained from  $C$  by deleting the edge  $uv$ . Note that if  $e = uv$  is a bridge, then  $e = uv$  is a simple edge.

## 2. Some bounds for the tree cover number

In this section, we give an upper bound on the tree cover number of a graph using the size of an independent set in the graph. We also provide upper and lower bounds on the tree cover number of a subgraph of  $G$  obtained by deleting an edge from  $G$ . In addition, we observe that subdividing an edge of a graph does not change the tree cover number.

The following proposition shows that, for a connected simple graph, we are able to bound the tree cover number by the difference between the order of the graph and the size of an independent set of vertices of the graph.

**Proposition 2.** *Let  $G = (V, E)$  be a connected simple graph, and let  $S \subseteq V(G)$  be an independent set. Then,  $T(G) \leq |G| - |S|$ . In particular,  $T(G) \leq |G| - \alpha(G)$ , where  $\alpha(G)$  is the independence number of  $G$ . Furthermore, this bound is tight.*

*Proof.* Let  $V(G) = \{v_1, v_2, \dots, v_n\}$  and suppose that  $S = \{v_1, \dots, v_k\}$  is an independent set. We construct a tree cover of size  $n - k$  by the following iterative process: for  $i = k + 1$ , let  $T_{v_i}$  be the tree induced by the set of vertices  $\{v_{k+1}\} \cup \{N(v_{k+1}) \cap S\}$ . For  $i = k + 2$  to  $n$ , let  $T_{v_i}$  be the tree induced by the set of vertices in  $\{v_i\} \cup \{N(v_i) \cap S\}$  that do not belong to  $V(T_{v_j})$  for  $k + 1 \leq j < i$ . Since  $G$  is connected, each  $s \in S$  has at least one neighbor in  $\{v_{k+1}, \dots, v_n\}$ , so this process produces a tree cover of  $G$  (where all components are stars) of size  $n - k$ . Thus,  $T(G) \leq n - k$ . In particular,  $T(G) \leq n - \alpha(G)$ . The star  $K_{1,n}$  shows that the bound  $T(G) \leq |G| - \alpha(G)$  is tight.  $\square$

In connection with the conjecture that  $T(G) \leq M_+(G)$ , we show that for some bounds on  $M_+(G)$ , analogous bounds hold for  $T(G)$ .

For a graph  $G = (V, E)$  and  $e \in E(G)$ , let  $G - e$  denote the graph obtained from  $G$  by deleting the edge  $e$ . In [Booth et al. 2011], it was shown that

$$M_+(G) - 1 \leq M_+(G - e) \leq M_+(G) + 1,$$

when  $G$  is a simple graph. We show that an analogous bound holds for the tree cover number of a multigraph  $G$ .

**Theorem 3.** *For a graph  $G = (V, E)$  and  $e \in E(G)$ ,*

$$T(G) - 1 \leq T(G - e) \leq T(G) + 1.$$

*Proof.* Let  $u, v \in V(G)$  such that  $e = uv$ . Consider the graph  $G - e$  obtained from  $G$  by deleting  $e$  (note that  $e$  could be a multiedge). Let  $\mathcal{T}$  be a minimum tree cover of  $G - e$ . If  $u$  and  $v$  are in disjoint trees in  $\mathcal{T}$ , then  $\mathcal{T}$  is a tree cover of  $G$ . So,  $T(G) \leq T(G - e)$ . If  $u$  and  $v$  are in the same tree in  $\mathcal{T}$ , denoted by  $T_{uv}$ , then the graph induced by the vertices of  $T_{uv}$  contains a cycle in  $G$ , so  $\mathcal{T}$  is not a tree cover of  $G$ . However, we may partition the vertices of  $T_{uv}$  into two sets

$A$  and  $B$ , such that the tree induced by the vertices in  $A$  contains  $u$  and the tree induced by the vertices in  $B$  contains  $v$ . Denote these trees by  $T_A$  and  $T_B$ . Then,  $(\mathcal{T} \setminus T_{uv}) \cup T_A \cup T_B$  is a tree cover of  $G$  of size  $T(G - e) + 1$ . This shows that  $T(G) - 1 \leq T(G - e)$ .

We now show that  $T(G - e) \leq T(G) + 1$ . Suppose there is a minimum tree cover  $\mathcal{T}$  of  $G$  such that  $u$  and  $v$  are in separate trees. Then  $\mathcal{T}$  is a tree cover of  $G - e$ , so  $T(G - e) \leq T(G)$ . Otherwise, let  $\mathcal{T}$  be a minimum tree cover of  $G$  that uses the edge  $e$  (so  $e$  is a simple edge by the definition of a tree cover), and let  $T_e$  be the tree in  $\mathcal{T}$  that contains  $e$ . By deleting  $e$  from  $T_e$ , we produce a tree cover of  $G - e$  of size  $T(G) + 1$ . This shows that  $T(G - e) \leq T(G) + 1$ , which completes the proof.  $\square$

The next theorem gives a bound that holds for the positive semidefinite maximum nullity of a graph, but the example that follows demonstrates that the analogous bound for the tree cover number fails.

A 2-separation of a graph  $G = (V, E)$  is a pair of subgraphs  $(G_1, G_2)$  such that  $V(G_1) \cup V(G_2) = V$ ,  $|V(G_1) \cap V(G_2)| = 2$ ,  $E(G_1) \cup E(G_2) = E$ , and  $E(G_1) \cap E(G_2) = \emptyset$ .

**Theorem 4** [van der Holst 2009, Theorem 2.8]. *Let  $(G_1, G_2)$  be a 2-separation of a graph  $G = (V, E)$ , and let  $H_1$  and  $H_2$  be obtained from  $G_1 = (V_1, E_1)$  and  $G_2 = (V_2, E_2)$ , respectively, by adding an edge between the vertices of  $R = \{r_1, r_2\} = V_1 \cap V_2$ . Then*

$$M_+(G) = \max\{M_+(G_1) + M_+(G_2) - 2, M_+(H_1) + M_+(H_2) - 2\}.$$

The analogous bound does not hold for the tree cover number. The next example provides a counterexample.

**Example 5.** For the graphs  $G, G_1, G_2, H_1, H_2$  given in Figure 1, we have that  $M_+(G_i) = 2$ ,  $M_+(H_i) = 3$ , and  $T(G_i) = T(H_i) = 2$  for  $i \in \{1, 2\}$ . So by Theorem 4,  $M_+(G) = 4$ . However,

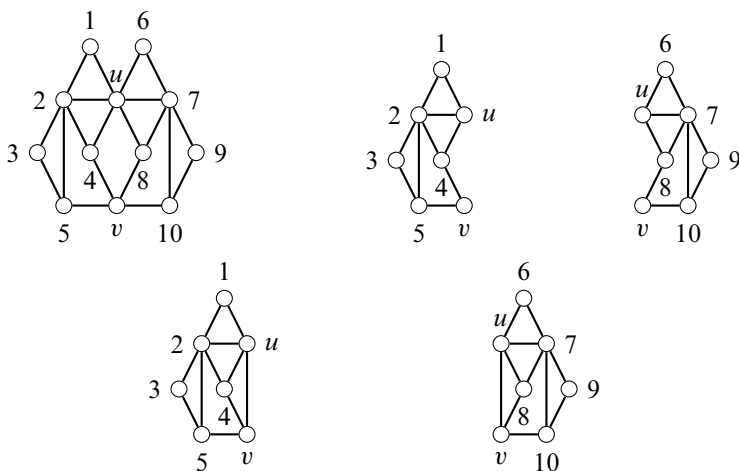
$$3 = T(G) > \max\{T(G_1) + T(G_2) - 2, T(H_1) + T(H_2) - 2\} = 2.$$

### 3. Characterizing edges required in a minimum tree cover

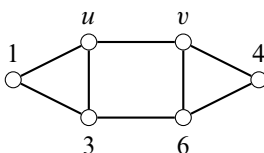
**Proposition 6.** *Let  $G = (V, E)$  be a graph such that  $uv \in E(G)$  is a bridge. Then  $uv$  is in a tree in every minimum tree cover of  $G$ .*

*Proof.* Note that there is no path from  $u$  to  $v$  that does not include  $uv$ . Therefore, for any tree cover that does not include  $uv$ , it must be the case that  $u$  and  $v$  are in separate trees. These two trees can be consolidated into one tree by adding the edge  $uv$ .  $\square$

We then ask the question: if an edge is required in every minimum tree cover, must it be a bridge? Figure 2 shows that such an edge is not necessarily a bridge.



**Figure 1.** Graphs of  $G$  (top left),  $G_1$  (top middle),  $G_2$  (top right),  $H_1$  (bottom left), and  $H_2$  (bottom right).



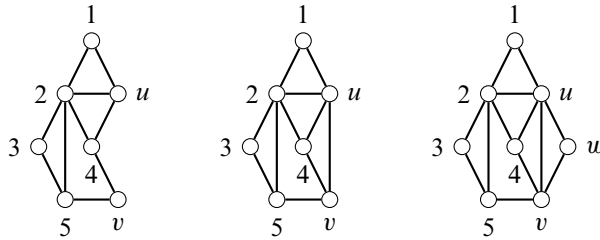
**Figure 2.** Graph for Example 7.

**Example 7.** Figure 2 gives a graph whose tree cover number is 2. However, although  $uv$  is not a bridge, any tree cover that does not include  $uv$  is of size at least 3.

The next lemma gives us a way to determine if an edge is required in every minimum tree cover, given that we are able to compute the necessary tree cover numbers.

**Lemma 8.** *Let  $G$  be a graph,  $u, v \in V(G)$ , and  $uv$  is a simple edge in  $E(G)$ . Let  $H$  be the graph obtained from  $G$  by adding a vertex such that  $V(H) = V(G) \cup \{w\}$  and  $E(H) = E(G) \cup \{uw, vw\}$ , where  $uw$  and  $vw$  are simple edges. Then,  $uv$  is required in every minimum tree cover of  $G$  if and only if  $T(H) = T(G) + 1$ .*

*Proof.* First observe that  $T(H) \leq T(G) + 1$  since any tree cover of  $G$  together with  $\{w\}$  is a tree cover for  $H$ . Let  $\mathcal{T} = \{T_1, T_2, \dots, T_k\}$  be a minimum tree cover of  $H$  such that  $w \in T_i$  for some  $i$ . Since  $w, u,$  and  $v$  cannot all be in the same tree, then either  $w$  is a leaf in  $T_i$  or  $T_i = \{w\}$ . If  $w$  is a leaf in  $T_i$ , then  $T_1, T_2, \dots, T_i - w, T_{i+1}, \dots, T_k$  is a tree cover of  $G$ , so  $T(G) \leq T(H)$ . If  $T_i = \{w\}$ , then  $\mathcal{T} \setminus T_i$  is a tree cover for  $G$ , so  $T(G) \leq T(H) - 1$ . This shows that  $T(H) = T(G)$  or  $T(H) = T(G) + 1$ .



**Figure 3.** Graphs of  $G$ ,  $H$ , and  $\hat{H}$  for Example 9.

Suppose that  $uv$  is required in every minimum tree cover of  $G$ . If  $w$  is a leaf in  $T_i$ , then  $T_1, T_2, \dots, T_i - w, T_{i+1}, \dots, T_k$  is a tree cover of  $G$  with  $u$  and  $v$  in separate trees, so it follows that  $T(H) = T(G) + 1$ . If  $T_i = \{w\}$ , then we also have that  $T(H) = T(G) + 1$ .

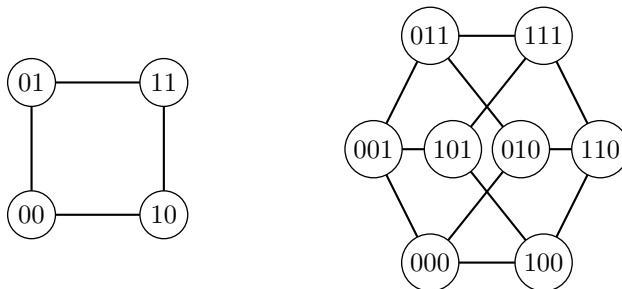
Suppose that there exists a minimum tree cover  $\mathcal{T} = \{T_1, T_2, \dots, T_k\}$  of  $G$  such that  $u$  and  $v$  are in different trees. If  $u \in T_i$ , we can create a tree cover of  $H$  of size  $k$  by adding the edge  $uw$  to  $E(T_i)$ . In this case,  $T(G) = T(H)$ .  $\square$

One might think that if  $H$  is a graph obtained from  $G$  by adding the edge  $uv$ , and  $uv$  is required in every minimum tree cover of  $H$ , then  $T(G) = T(H) + 1$ . However, this is not true. Example 9 provides a counterexample.

**Example 9.** It is easy to see that  $T(G) = T(H) = 2$  (for  $H$ , take the set  $\{1, u, v, 5\}$  and  $\{2, 3, 4\}$  for example). It can also be verified that  $T(\hat{H}) = 3$ . By Lemma 8, it follows that the edge  $uv$  is required in every minimum tree cover of  $H$ .

#### 4. Tree cover number of the hypercube

The  $d$ -dimensional hypercube, denoted  $Q_d$ , is the simple graph with vertex set  $\{0, 1\}^d$  where two vertices are adjacent if and only if they differ in exactly one position. For example, the 2-dimensional hypercube is a square (see left figure below) and the 3-dimensional hypercube is a cube (right figure). Equivalently,



hypercubes can be inductively defined as the cartesian product of  $d$  copies of the complete graph  $K_2$ . Hypercubes are a particular case of a larger family of graphs

called Hamming graphs. The  $d$ -dimensional Hamming graph, denoted  $H(d, q)$ , is the graph with vertex set  $\{0, \dots, n - 1\}^d$  where two vertices are adjacent if and only if they differ in exactly one position. Hamming graphs are of use in many areas including error-correcting codes, modeling heat diffusion, and association schemes in statistics. In this section, we show that the tree cover number of the  $d$ -dimensional hypercube is 2 for all  $d \geq 2$ .

**Theorem 10.** *Let  $Q_d$  be the  $d$ -dimensional hypercube graph. For all  $d \geq 2$ ,  $T(Q_d) = 2$ .*

*Proof.* We first list explicit sets which induce a tree cover of size 2 for  $Q_d$ , for  $d \in \{2, 3, 4, 5\}$ :

$$\begin{aligned}
 T_{1_2} &= \{(00), (01)\}, & T_{2_2} &= \{(11), (10)\}. \\
 T_{1_3} &= \{(010), (000), (001), (110)\}, & T_{2_3} &= \{(111), (011), (100), (101)\}. \\
 T_{1_4} &= \{(0011), (0010), (0000), (0110), (1111), (1011), (1100), (1101)\}, \\
 T_{2_4} &= \{(0001), (0111), (0100), (0101), (1010), (1000), (1001), (1110)\}. \\
 T_{1_5} &= \{(00100), (00011), (00000), (00110), (01111), (01011), (01100), (01101), \\
 &\quad (10001), (10111), (10100), (10101), (11010), (11000), (11001), (11110)\}, \\
 T_{2_5} &= \{(00010), (00001), (00111), (00101), (01010), (01000), (01001), (01110), \\
 &\quad (10011), (10010), (10000), (10110), (11111), (11011), (11100), (11101)\}.
 \end{aligned}$$

Other values of  $d$  are handled by induction. Throughout the proof, the sets  $\widehat{T}_{1_j}$  and  $\widehat{T}_{2_j}$  are covers that will be used as preliminary steps to obtain the sets  $T_{1_j}$  and  $T_{2_j}$  that will induce a tree cover of size two for  $Q_j$ . The proof proceeds as follows: first we give a cover and a tree cover of size two for  $Q_6$ ; due to the volume of data this information is presented in an online-only supplement. Then, using this tree cover, we construct a cover and a tree cover of size two for  $Q_7$ , which again appears in the supplement. We then inductively show that for  $d \geq 8$  we can systematically construct a tree cover of size two using the covers and tree covers constructed for  $Q_{d-1}$  and  $Q_{d-2}$ .

Consider the sets  $\widehat{T}_{1_6}$  and  $\widehat{T}_{2_6}$  given in the supplement. Note that  $\{\widehat{T}_{1_6}, \widehat{T}_{2_6}\}$  is a cover for  $Q_6$ , and that  $Q_6[\widehat{T}_{1_6}]$  and  $Q_6[\widehat{T}_{2_6}]$  are both forests, each consisting of two disjoint trees. Let  $x_{1_6} = (001101)$ ,  $x_{2_6} = (110010)$ ,  $y_{1_6} = (001001)$ ,  $y_{2_6} = (110100)$ . Then  $x_{1_6}$  and  $x_{2_6}$  are in  $\widehat{T}_{1_6}$ , and they are not in the same tree in  $Q_6[\widehat{T}_{1_6}]$ . Similarly,  $y_{1_6}$  and  $y_{2_6}$  are in  $\widehat{T}_{2_6}$ , and they are not in the same tree in  $Q_6[\widehat{T}_{2_6}]$ . By swapping  $x_{1_6}$  and  $y_{1_6}$ , the resulting sets  $T_{1_6}$  and  $T_{2_6}$  (listed in the supplement) induce a tree cover for  $Q_6$  of size two.

To obtain a tree cover of size two for  $Q_7$ , we begin by adding a 0 to the beginning of each element in  $T_{1_6}$ , and a 1 to the beginning of each element in  $T_{2_6}$ . Denote



these sets by  $T_{1_6,0}$  and  $T_{2_6,1}$ , respectively, and let  $\widehat{T}_{1_7} := T_{1_6,0} \cup T_{2_6,1}$ . Similarly, we construct the sets  $T_{1_6,1}$  and  $T_{2_6,0}$ , and let  $\widehat{T}_{2_7} := T_{1_6,1} \cup T_{2_6,0}$  (see supplement). Then, both  $Q_7[\widehat{T}_{1_7}]$  and  $Q_7[\widehat{T}_{2_7}]$  are forests consisting of two disjoint trees. By swapping  $0x_{2_6}$  and  $0y_{2_6}$ , the resulting sets  $T_{1_7}$  and  $T_{2_7}$  (given in supplement) induce a tree cover of size two for  $Q_7$ .

We proceed by induction to prove the claim for  $Q_d$  with  $d \geq 8$ . Suppose that we have constructed the sets  $\widehat{T}_{1_{d-2}} = \{x_1, x_2, \dots, x_n\}$  and  $\widehat{T}_{2_{d-2}} = \{y_1, y_2, \dots, y_n\}$  such that  $\{\widehat{T}_{1_{d-2}}, \widehat{T}_{2_{d-2}}\}$  gives a cover for  $Q_{d-2}$  satisfying the following conditions:

- (1)  $Q_{d-2}[\widehat{T}_{1_{d-2}}]$  and  $Q_{d-2}[\widehat{T}_{2_{d-2}}]$  are forests composed of two disjoint trees.
- (2) Swapping  $x_1$  and  $y_1$  results in sets

$$T_{1_{d-2}} = \{y_1, x_2, \dots, x_n\}, \quad T_{2_{d-2}} = \{x_1, y_2, \dots, y_n\},$$

that induce a tree cover of  $Q_{d-2}$  of size two.

- (3) For the cover

$$\begin{aligned} \widehat{T}_{1_{d-1}} &= T_{1_{d-2,0}} \cup T_{2_{d-2,1}} = \{0y_1, 0x_2, 0x_3, \dots, 0x_n, 1x_1, 1y_2, \dots, 1y_n\}, \\ \widehat{T}_{2_{d-1}} &= T_{2_{d-2,0}} \cup T_{1_{d-2,1}} = \{0x_1, 0y_2, 0y_3, \dots, 0y_n, 1y_1, 1x_2, \dots, 1x_n\}, \end{aligned}$$

of  $Q_{d-1}$ , swapping  $0x_2 \in \widehat{T}_{1_{d-1}}$  and  $0y_2 \in \widehat{T}_{2_{d-1}}$  results in sets

$$\begin{aligned} T_{1_{d-1}} &= \{0y_1, 0y_2, 0x_3, \dots, 0x_n, 1x_1, 1y_2, \dots, 1y_n\}, \\ T_{2_{d-1}} &= \{0x_1, 0x_2, 0y_3, \dots, 0y_n, 1y_1, 1x_2, \dots, 1x_n\}, \end{aligned}$$

for  $Q_{d-1}$  that induced a tree cover of  $Q_{d-1}$  of size two.

- (4)  $x_1$  and  $x_2$  are not in the same induced tree in  $\widehat{T}_{1_{d-2}}$ .
- (5)  $y_1$  and  $y_2$  are not in the same induced tree in  $\widehat{T}_{2_{d-2}}$ .

We are also assuming that  $x_i \neq x_j$ ,  $y_i \neq y_j$  for  $i \neq j$ , and  $x_i \neq y_j$  for all  $i, j$ .

Then we can construct a cover for  $Q_d$  such that swapping two of the elements in the cover will result in a tree cover of size two for  $Q_d$ . Furthermore, we show that the constructed cover and tree cover for  $Q_d$ , together with the constructed cover and tree cover for  $Q_{d-1}$ , still satisfy the above hypotheses, which proves the claim for all  $d \geq 8$ .

We first construct a cover  $\{\widehat{T}_{1_d}, \widehat{T}_{2_d}\}$  for  $Q_d$  in the following way:

$$\begin{aligned} \widehat{T}_{1_d} &= T_{1_{d-1,0}} \cup T_{2_{d-1,1}} \\ &= \{00y_1, 00y_2, 00x_3, \dots, 00x_n, 01x_1, 01y_2, 01y_3, \dots, 01y_n, \\ &\quad 10x_1, 10x_2, 10y_3, \dots, 10y_n, 11y_1, 11x_2, 11x_3, \dots, 11x_n\}. \end{aligned}$$

$$\begin{aligned}\widehat{T}_{2_d} &= T_{2_{d-1},0} \cup T_{1_{d-1},1} \\ &= \{00x_1, 00x_2, 00y_3, \dots, 00y_n, 01y_1, 01x_2, 01x_3, \dots, 01x_n, \\ &\quad 10y_1, 10y_2, 10x_3, \dots, 10x_n, 11x_1, 11y_2, 11y_3, \dots, 11y_n\}.\end{aligned}$$

Note that since  $Q_{d-1}[T_{1_{d-1}}]$  and  $Q_{d-1}[T_{2_{d-1}}]$  are two disjoint trees, it follows that  $Q_d[\widehat{T}_{1_d}]$  is a forest consisting of two disjoint trees. Similarly,  $Q_d[\widehat{T}_{2_d}]$  is a forest consisting of two disjoint trees. By swapping  $01x_1$  and  $01y_1$ , we obtain the sets

$$\begin{aligned}T_{1_d} &= \{00y_1, 00y_2, 00x_3, \dots, 00x_n, 01y_1, 01y_2, 01y_3, \dots, 01y_n, \\ &\quad 10x_1, 10x_2, 10y_3, \dots, 10y_n, 11y_1, 11x_2, 11x_3, \dots, 11x_n\}, \\ T_{2_d} &= \{00x_1, 00x_2, 00y_3, \dots, 00y_n, 01x_1, 01x_2, 01x_3, \dots, 01x_n, \\ &\quad 10y_1, 10y_2, 10x_3, \dots, 10x_n, 11x_1, 11y_2, 11y_3, \dots, 11y_n\}.\end{aligned}$$

We now show that  $\{Q_d[T_{1_d}], Q_d[T_{2_d}]\}$  is a tree cover for  $Q_d$  of size two by showing:

- (1)  $Q_d[T_{1_d}]$  and  $Q_d[T_{2_d}]$  are forests (i.e., there are no cycles in each of  $Q_d[T_{1_d}]$  and  $Q_d[T_{2_d}]$ ).
- (2) Both  $Q_d[T_{1_d}]$  and  $Q_d[T_{2_d}]$  are connected graphs.

We show that  $Q_d[T_{1_d}]$  is a forest (a similar argument shows that  $Q_d[T_{2_d}]$  is a forest). From our construction  $Q_d[\widehat{T}_{1_d}]$  is a forest composed of 2 trees, denoted  $Q_d[\widehat{A}]$  and  $Q_d[B]$ , where

$$\begin{aligned}\widehat{A} &:= \{00y_1, 00y_2, 00x_3, \dots, 00x_n, 01x_1, 01y_2, 01y_3, \dots, 01y_n\}, \\ B &:= \{10x_1, 10x_2, 10y_3, \dots, 10y_n, 11y_1, 11x_2, 11x_3, \dots, 11x_n\}.\end{aligned}$$

By definition  $T_{1_d} = (\widehat{T}_{1_d} \setminus \{01x_1\}) \cup \{01y_1\}$ . By removing  $01x_1$  from  $\widehat{T}_{1_d}$ ,  $B$  is not affected, and  $Q_d[\widehat{A} \setminus \{01x_1\}]$  is now the union of  $\deg(01x_1)$  disjoint trees. We now show that by adding  $01y_1$  to  $\widehat{T}_{1_d} \setminus \{01x_1\}$ , no cycles are created in  $Q_d[T_{1_d}]$ . Define  $A = \{00y_1, 00y_2, 00x_3, \dots, 00x_n, 01y_1, 01y_2, 01y_3, \dots, 01y_n\}$  (note that  $T_{1_d} = A \cup B$ ). Between  $A$  and  $B$ , the only vertices that are adjacent are  $01y_1$  and  $11y_1$  (everything else differs in more than one position). Hence, if there is a cycle in  $Q_d[T_{1_d}]$ , it must be in  $Q_d[A]$ . Since  $Q_d[A \setminus \{01y_1\}]$  (which equals  $Q_d[\widehat{A} \setminus \{01x_1\}]$ ) is a forest composed of  $\deg(01x_1)$  trees, if there is a cycle in  $Q_d[A]$  it must involve  $01y_1$ . We will now show that it is not possible to have a cycle involving  $01y_1$ , hence no cycle is possible in  $Q_d[T_{1_d}]$ .

Note that there is an edge between  $00y_1$  and  $01y_1$ , and that there are no edges between  $01y_1$  and any of  $00y_2, 00x_3, \dots, 00x_n$ . Thus, the neighbors of  $01y_1$  in  $Q_d[A]$  are  $00y_1$  and a subset of  $\{01y_3, 01y_4, \dots, 01y_n\}$  (since  $y_1$  is not adjacent to  $y_2$  by condition (5) above, then  $01y_1$  is not adjacent to  $01y_2$ ). Let  $01y_i$  and  $01y_j$ ,  $i \neq j$ , be arbitrary neighbors of  $01y_1$ . We show that:

- (a) There is no path from  $01y_i$  to  $01y_j$  in  $Q_d[A]$  for  $i, j \in \{3, 4, \dots, n\}$  that does not include  $01y_1$ .
- (b) There is no path from  $00y_1$  to  $01y_i$  in  $Q_d[A]$  that does not include  $01y_1$ .

To see (a), note that from condition (1),  $Q_{d-2}[\{y_1, \dots, y_n\}]$  is a forest of two disjoint trees. This implies that  $Q_d[\{01y_1, \dots, 01y_n\}]$  is a forest of two disjoint trees. Then, within  $Q_d[\{01y_1, \dots, 01y_n\}]$  there is no path from  $01y_i$  to  $01y_j$  that does not include  $01y_1$ . Note that vertices of  $\{01y_3, 01y_4, \dots, 01y_n\}$  are not adjacent to any vertices in  $A$  except for possibly each other and  $01y_1$  and  $01y_2$ . Thus, any path from  $01y_i$  to  $01y_j$  not including  $01y_1$  must include  $01y_2$ . By condition (1),  $y_1$  and  $y_2$  are not in the same induced tree of  $Q_{d-2}[\{y_1, \dots, y_n\}]$ , so  $01y_1$  and  $01y_2$  are not in the same induced tree of  $Q_d[\{01y_1, \dots, 01y_n\}]$ . Since  $01y_i$  and  $01y_j$  are neighbors of  $01y_1$ , and  $01y_1$  is not in the same induced tree as  $01y_2$  in  $Q_d[\{01y_1, \dots, 01y_n\}]$ , then  $01y_i$  and  $01y_j$  are not in the same induced tree as  $01y_2$ . Thus, the only path from  $01y_i$  to  $01y_j$  is  $(01y_i, 01y_1, 01y_j)$ .

For (b), we have that the vertices in the set  $\{01y_3, 01y_4, \dots, 01y_n\}$  are not connected in  $Q_d[A]$  to any vertices in  $A$  except for possibly each other and  $01y_1$ . We also have that  $01y_i$  is not adjacent to  $00y_1$  in  $Q_d[A]$ . So any path from  $01y_i$  to  $00y_1$  must include  $01y_1$ .

Next we show that  $Q_d[T_{1_d}]$  is connected (a similar argument shows that  $Q_d[T_{2_d}]$  is connected). Recall from the hypotheses that

$$Q_{d-2}[\widehat{T}_{2_{d-2}}] = Q_{d-2}[\{y_1, y_2, \dots, y_n\}]$$

is a forest consisting of two disjoint trees, and

$$Q_{d-2}[T_{2_{d-2}}] = Q_{d-2}[\{x_1, y_2, \dots, y_n\}]$$

is a tree. This implies that  $y_1$  has exactly one fewer neighbor among  $y_2, \dots, y_n$  than  $x_1$ . To see this, note that  $Q_{d-2}[\widehat{T}_{2_{d-2}} \setminus \{y_1\}]$  is composed of  $1 + \deg(y_1)$  trees. Since

$$Q_{d-2}[T_{2_{d-2}}] = Q_{d-2}[(\widehat{T}_{2_{d-2}} \setminus \{y_1\}) \cup \{x_1\}]$$

is a tree, we must have  $\deg(x_1) = 1 + \deg(y_1)$ . Therefore,  $01y_1$  must have one less neighbor than  $01x_1$  among  $01y_2, \dots, 01y_n$ . Hence,  $01y_1$  and  $01x_1$  have the same number of neighbors in  $A$ , and thus  $01y_1$  has one more neighbor than  $01x_1$  in  $T_{1_d}$ . We will now show that this last statement implies that  $Q_d[T_{1_d}]$  is connected.

Since the graphs induced by  $T_{1_{d-1}}$  and  $T_{2_{d-1}}$  are trees, then

$$Q_d[T_{1_d}] = Q_d[T_{1_{d-1,0}} \cup T_{2_{d-1,1}}]$$

is a forest consisting of two disjoint trees. Hence,  $Q_d[\widehat{T}_{1_d} \setminus \{01x_1\}]$  is a forest consisting of  $1 + \deg(01x_1)$  trees. Since  $\deg(01y_1) = 1 + \deg(01x_1)$ , and since  $Q_d[T_{1_d}]$  has no cycles, we have that each of the edges of  $01y_1$  must be connected

to a different component of the forest. Therefore,  $Q_d[T_{1_d}]$  is a tree. An analogous argument shows that  $Q_d[T_{2_d}]$  is a tree. Thus,  $\{Q_d[T_{1_d}], Q_d[T_{2_d}]\}$  is a tree cover of size two of  $Q_d$ .

We now show that the covers and tree covers constructed for  $Q_{d-1}$  and  $Q_d$  satisfy the induction hypotheses. Note that since  $Q_{d-2}[T_{1_{d-2}}]$  and  $Q_{d-2}[T_{2_{d-2}}]$  are two disjoint trees, it follows from construction that  $Q_d[\widehat{T}_{1_{d-1}}]$  is a forest consisting of two disjoint trees. Similarly,  $Q_d[\widehat{T}_{2_{d-1}}]$  is a forest consisting of two disjoint trees, satisfying condition (1). For clarity, we relabel the vertices of  $\widehat{T}_{1_{d-1}}$  and  $\widehat{T}_{2_{d-1}}$  such that  $\widehat{T}_{1_{d-1}} = \{w_1, \dots, w_m\}$  and  $\widehat{T}_{2_{d-1}} = \{z_1, \dots, z_m\}$  where  $w_1 = 0x_2$ ,  $w_2 = 1x_1$ ,  $z_1 = 0y_2$ , and  $z_2 = 1y_1$ . Then by condition (3), swapping  $w_1$  and  $z_1$  results in sets  $T_{1_{d-1}} = \{z_1, w_2, \dots, w_m\}$  and  $T_{2_{d-1}} = \{w_1, z_2, \dots, z_m\}$  that induce a tree cover of  $Q_{d-1}$  of size two, which shows that condition (2) is satisfied. Note that with this relabeling, the sets  $\widehat{T}_{1_d}$  and  $\widehat{T}_{2_d}$  become

$$\begin{aligned} \widehat{T}_{1_d} &= T_{1_{d-1,0}} \cup T_{2_{d-1,1}} = \{0z_1, 0w_2, 0w_3, \dots, 0w_m, 1w_1, 1z, \dots, 1z_m\} \\ \widehat{T}_{2_d} &= T_{2_{d-1,0}} \cup T_{1_{d-1,1}} = \{0w_1, 0z_2, 0y_3, \dots, 0z_m, 1z_1, 1w_2, \dots, 1w_m\}, \end{aligned}$$

and we have shown above that swapping  $0w_2 = 01x_1$  and  $0z_2 = 01y_1$  results in the sets  $T_{1_d}$  and  $T_{2_d}$  which induce a tree cover of size two for  $Q_d$ , satisfying condition (3). Furthermore, since  $w_1 = 0x_2 \in T_{1_{d-2,0}}$  and  $w_2 = 1x_1 \in T_{2_{d-2,1}}$ , we have that  $w_1$  and  $w_2$  are not in the same induced tree in  $Q_{d-1}[\widehat{T}_{1_{d-1}}]$ . Similarly,  $z_1 = 0y_2 \in T_{2_{d-2,0}}$  and  $z_2 = 1y_1 \in T_{1_{d-2,1}}$ , so  $z_1$  and  $z_2$  are not in the same induced tree in  $Q_{d-1}[\widehat{T}_{2_{d-1}}]$ , showing that conditions (4) and (5) are satisfied.

Since the hypotheses still hold with the constructed covers and tree covers of  $Q_{d-1}$  and  $Q_d$ , then it follows, by inductively applying the above argument, that  $T(Q_d) = 2$  for all  $d$ . □

One may wonder why the base case of the proof starts with  $Q_6$  and  $Q_7$ . We would like to note that starting as early as  $d = 2$ , we were able to use a tree cover of  $Q_d$  to produce a cover for  $Q_{d+1}$  such that there exists two vertices that could be swapped in order to produce a tree cover for  $Q_{d+1}$ . In fact, this is how we constructed the tree covers for  $Q_3, Q_4, Q_5$  given at the start of the proof. However, there is a choice to be made when switching vertices, and the point at which the above constructive pattern holds is dependent upon the initial choice of vertices that are swapped. For example, we experimented with using a different initial swap and found that the pattern did not hold until  $d = 11$  or later. It may also be the case that there is an initial swap that allows the pattern to begin sooner than  $d = 8$ . This is a very interesting phenomenon that is worth further exploration.

We also investigated the idea of generalizing the above proof to all Hamming graphs. For  $H(2, 3)$ , we found that  $T(H(2, 3)) = 3$ , and evidence suggests that  $T(H(d, q)) = q$ .

**Conjecture 11.**  $T(H(d, q)) = q$ , for  $H(d, q)$  the Hamming graph of dimension  $d$ .

### Acknowledgements

This research was conducted at Iowa State University's 2015 Mathematics REU program. We would like to thank the Department of Mathematics at Iowa State University, as well as the NSF for supporting the research. In addition we would like to thank Dr. Leslie Hogben for her contributed insight to this project and for her suggestions that improved the presentation of this paper.

### References

- [Barioli et al. 2011] F. Barioli, S. M. Fallat, L. H. Mitchell, and S. K. Narayan, "Minimum semidefinite rank of outerplanar graphs and the tree cover number", *Electron. J. Linear Algebra* **22** (2011), 10–21. MR Zbl
- [Booth et al. 2011] M. Booth, P. Hackney, B. Harris, C. R. Johnson, M. Lay, T. D. Lenker, L. H. Mitchell, S. K. Narayan, A. Pascoe, and B. D. Sutton, "On the minimum semidefinite rank of a simple graph", *Linear Multilinear Algebra* **59**:5 (2011), 483–506. MR Zbl
- [Ekstrand et al. 2012] J. Ekstrand, C. Erickson, D. Hay, L. Hogben, and J. Roat, "Note on positive semidefinite maximum nullity and positive semidefinite zero forcing number of partial 2-trees", *Electron. J. Linear Algebra* **23** (2012), 79–87. MR Zbl
- [Fallat and Hogben 2007] S. M. Fallat and L. Hogben, "The minimum rank of symmetric matrices described by a graph: a survey", *Linear Algebra Appl.* **426**:2-3 (2007), 558–582. MR Zbl
- [Fallat and Hogben 2014] S. M. Fallat and L. Hogben, "Minimum rank, maximum nullity, and zero forcing number of graphs", Chapter 46 in *Handbook of linear algebra*, 2nd ed., edited by L. Hogben, CRC Press, Boca Raton, FL, 2014. Zbl
- [van der Holst 2009] H. van der Holst, "On the maximum positive semi-definite nullity and the cycle matroid of graphs", *Electron. J. Linear Algebra* **18** (2009), 192–201. MR Zbl

Received: 2015-11-13      Revised: 2016-09-07      Accepted: 2016-09-07

cbozeman@iastate.edu	<i>Department of Mathematics, Iowa State University, Ames, IA 50011, United States</i>
catralm@xavier.edu	<i>Department of Mathematics and Computer Science, Xavier University, 3000 Victory Parkway, Cincinnati, OH 45207, United States</i>
cookb@carleton.edu	<i>Department of Mathematics, Carleton College, Northfield, MN 55067, United States</i>
oscar.gonzalez3@upr.edu	<i>Department of Mathematics, University of Puerto Rico, San Juan 00931, Puerto Rico</i>
reinh196@iastate.edu	<i>Department of Mathematics, University of Minnesota, Minneapolis, MN 55455, United States</i>



# Matrix completions for linear matrix equations

Geoffrey Buhl, Elijah Cronk, Rosa Moreno,  
Kirsten Morris, Dianne Pedroza and Jack Ryan

(Communicated by Chi-Kwong Li)

A matrix completion problem asks whether a partial matrix composed of specified and unspecified entries can be completed to satisfy a given property. This work focuses on determining which patterns of specified and unspecified entries correspond to partial matrices that can be completed to solve three different matrix equations. We approach this problem with two techniques: converting the matrix equations into linear equations and examining bases for the solution spaces of the matrix equations. We determine whether a particular pattern can be written as a linear combination of the basis elements. This work classifies patterns as admissible or inadmissible based on the ability of their corresponding partial matrices to be completed to satisfy the matrix equation. Our results present a partial or complete characterization of the admissibility of patterns for three homogeneous linear matrix equations.

## 1. Introduction

A matrix completion problem asks whether a partial matrix, one with some entries given and others freely chosen, can be completed to satisfy a desired property. In this work, we classify patterns for entries in a partial matrix so that the partial matrix can almost always be completed to satisfy certain linear matrix equations. We establish limits on the number of specified entries in patterns and on the locations of specified and unspecified entries.

Examples of matrix completion problems include determining completions for  $M$ -matrices and inverse  $M$ -matrices where the desired property is that a nonnegative partial matrix pattern of any order has an inverse  $M$ -matrix [Johnson and Smith 1996], where  $M$ -matrices are  $Z$ -matrices such that each eigenvalue of the matrix has positive real parts. A  $Z$ -matrix is one whose off-diagonal entries are less than or equal to zero. The inverse  $M$ -matrix completion problem can also be evaluated using a graph theoretic approach [Hogben 1998; 2000]. Other classical matrix

---

*MSC2010:* primary 15A83; secondary 15A27.

*Keywords:* matrix completion problems, partial matrices, matrix commutativity, matrix equations.

completion problems involve completing partial Hermitian matrices and positive definite matrices to determine which partial positive definite matrices have a positive definite completion [Grone et al. 1984], while others look at completing TP or TN matrices with the goal of preserving low-rank [Johnson and Wei 2013]. A TP, or totally positive, matrix is a square matrix such that the determinant of each square submatrix (including minors) is positive. Equivalently, each of the eigenvalues of such a matrix is nonnegative. TN matrices are totally nonnegative matrices.

Another matrix completion problem is the titled completion problem, which asks if, given a conventional partial matrix, there exist values for the unspecified entries resulting in a conventional matrix that is either doubly nonnegative (DN) or completely positive (CP) [Drew et al. 2000]. Additionally, for partial matrices that are symmetric and have specified entries along the diagonal, it is known there is a  $P$ -matrix completion if and only if every given principal submatrix has a positive determinant [Johnson and Kroschel 1996]. Any  $4 \times 4$  pattern also has a  $P$ -completion if it contains eight or fewer off-diagonal positions [DeAlba and Hogben 2000]. A graph theoretic approach can also be used to evaluate the  $P$ -completion problem [Hogben 2001]. There are also results for matrix completions involving the Euclidean distance. For example, for every partial distance matrix in  $\mathbb{R}^k$  such that the graph of specified entries is chordal, there exists a completion to a distance matrix in  $\mathbb{R}^k$  [Bakonyi and Johnson 1995]. These classic matrix completion problems determine the condition under which a partial matrix can be completed, so that the resulting matrix has a certain property. Only one matrix is involved in these problems, the partial matrix itself.

In this work, we determine if a partial matrix can be completed to satisfy certain matrix equations. In this case the admissibility of a pattern is relative to other matrices in the matrix equation. We focus on determining which patterns of specified and unspecified entries for partial matrices can almost always be completed to satisfy the following matrix equations: the skew-symmetric equation  $AX - A^T X = 0$ , the commutativity equation  $AX - XA = 0$ , and the skew-Lyapunov equation  $AX - XA^T = 0$ . It is not possible to, in general, solve these matrix completion problems for all matrices  $A$ . So, we look to solve the completion for almost all matrices  $A$ . That is, we assume  $A$  has a certain property that almost all matrices satisfy, and we show that any partial matrix can be completed for almost all of these “generic”  $A$ . In this work, we assume either  $A$  has distinct eigenvalues or is nonderogatory.

We use two approaches to classify patterns. The column space approach converts the matrix equations to linear equations and uses linearly independent columns to determine unspecified entry locations. The nullspace approach uses a basis of the solution space of a homogeneous matrix equation to determine specified entry locations. We classify patterns as admissible or inadmissible based on the ability or inability of corresponding partial matrices to be completed to satisfy the matrix equation for a “generic” matrix  $A$ .



We discuss the important ideas and definitions relevant to completions of matrix equations in Section 2. Sections 3 and 4 explain the two principle methods used for classifying partial matrix patterns: the column space and nullspace approaches. We apply the column space and nullspace approaches to the skew-symmetric, commutativity, and skew-Lyapunov equations in Section 5 to classify patterns for these equations.

### 2. Preliminaries

In this section, we define a partial matrix pattern, a partial matrix, a partial matrix completion, and the admissibility or inadmissibility of matrix patterns. We include relevant definitions and theorems from linear algebra, including the Kronecker product and the vec function.

**Definition 2.1.** An  $n \times n$  partial matrix pattern

$$\alpha = \{(i_t, j_t) \mid 1 \leq i_t, j_t \leq n, t = 1, \dots, n\}$$

is a set of specified entry locations in an  $n \times n$  matrix. For a partial matrix pattern  $\alpha$ , the  $n \times n$  rectangular array  $\mathcal{X} = [x_{ij}]$  is an  $\alpha$ -partial matrix if the only specified entries correspond to the locations in  $\alpha$ .

A pattern describes locations in a matrix as specified or unspecified. A pattern becomes a partial matrix when the specified entry locations have values assigned.

**Definition 2.2.** A completion of an  $\alpha$ -partial matrix  $\mathcal{X} = [x_{ij}]$  is a matrix  $\widehat{\mathcal{X}} = [\widehat{x}_{ij}] \in M_n(\mathbb{R})$  in which  $\widehat{x}_{ij} = x_{ij}$  whenever  $(i, j) \in \alpha$ .

Throughout this paper,  $\mathcal{X}$  will represent a partial matrix, and  $\widehat{\mathcal{X}}$  will represent a completion of  $\mathcal{X}$ . For example, consider a  $3 \times 3$  pattern  $\alpha = \{(1, 1), (1, 3), (2, 2), (3, 2), (3, 3)\}$ . The following are the pattern  $\alpha$ , an  $\alpha$ -partial matrix  $\mathcal{X}$ , and a completion  $\widehat{\mathcal{X}}$ :

$$\alpha = \begin{bmatrix} \# & \square & \# \\ \square & \# & \square \\ \square & \# & \# \end{bmatrix}, \quad \mathcal{X} = \begin{bmatrix} 1 & x_{12} & 4 \\ x_{21} & 5 & x_{23} \\ x_{31} & 9 & 11 \end{bmatrix}, \quad \widehat{\mathcal{X}} = \begin{bmatrix} 1 & 15 & 4 \\ 13 & 5 & 19 \\ 2 & 9 & 11 \end{bmatrix}.$$

**Definition 2.3.** An  $n \times n$  partial matrix pattern  $\alpha$  is *admissible* for the matrix equation

$$A_1 X B_1 + A_2 X B_2 + \dots + A_k X B_k = C$$

if for all  $\alpha$ -partial matrices  $\mathcal{X}$  there exists a completion  $\widehat{\mathcal{X}}$  such that

$$A_1 \widehat{\mathcal{X}} B_1 + A_2 \widehat{\mathcal{X}} B_2 + \dots + A_k \widehat{\mathcal{X}} B_k = C,$$

where  $A_1, A_2, \dots, A_k, B_1, B_2, \dots, B_k, C \in M_n(\mathbb{R})$ .

Because the admissibility of a pattern, in this work, depends on the fully specified matrices in the matrix equation, the problem of classifying admissible patterns becomes unwieldy without some restrictions on these matrices. In this paper, we restrict our attention to two large categories of matrices: nonderogatory matrices and matrices with distinct eigenvalues. These restrictions are necessary in order to calculate the maximum number of specified entry locations for the matrix equations we examine. Both nonderogatory and distinct eigenvalues are “generic” matrix properties in the sense that almost all matrices satisfy these properties.

There may be some versions of matrix equations for which a given partial matrix may not be completed to satisfy the particular instance of the matrix equation. For example with the  $2 \times 2$  pattern  $\alpha = \{(1, 2), (2, 2)\}$ , not all  $\alpha$ -partial matrices can be completed to commute with a diagonal matrix with distinct eigenvalues. However, the only matrices  $A$  for which not all of these  $\alpha$ -partial matrices can be completed to commute with  $A$  are those matrices  $A$  with a 0 in the  $(1, 2)$  position. The set of such matrices is a set of measure zero. So we say that  $\alpha$  is admissible for the commutativity equation in general, which is to say that  $\alpha$  is admissible for the matrix equation  $AX - XA = 0$  for almost all “generic”  $A$ , which we show in Section 5.

In Sections 3 and 4, we construct conditions for the admissibility of patterns given matrices  $A_1, A_2, \dots, A_k, B_1, B_2, \dots, B_k, C$ . For the matrix equations in Section 5, there is only one matrix  $A$  that is fully specified, so admissibility of a pattern for the general form of a matrix equation means any partial matrix can be completed for almost all “generic”  $A$ . Admissibility depends on the matrix equation as well; a pattern may be admissible for  $AX - A^T X = 0$  but not admissible for  $AX - X A^T = 0$ . The matrix equation for which a pattern is admissible or inadmissible should be clear from context.

**Definition 2.4.** An admissible pattern  $\alpha$  is *maximally admissible* if and only if  $|\beta| \leq |\alpha|$  for every admissible pattern  $\beta$ .

In Section 4 we show the dimension of the solution space of the matrix equations gives the size of the maximally admissible patterns

**Definition 2.5.** The *Kronecker product* of  $A = [a_{ij}] \in M_{m,n}(\mathbb{R})$  and  $B = [b_{ij}] \in M_{p,q}(\mathbb{R})$  is denoted by  $A \otimes B$  and is defined to be the block matrix

$$A \otimes B \equiv \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix} \in M_{mp,nq}(\mathbb{R}).$$

**Definition 2.6.** Given  $A = [a_{ij}] \in M_{m,n}(\mathbb{R})$ , the function  $\text{vec} : M_{m,n}(\mathbb{R}) \rightarrow \mathbb{R}^{mn}$  is defined as

$$\text{vec}(A) = [a_{11} \cdots a_{m1} \ a_{12} \cdots a_{m2} \ \cdots \ a_{1n} \cdots a_{mn}]^T.$$

The following theorem describes how to use the  $\text{vec}$  function and Kronecker product to transform linear matrix equations into linear equations.

**Theorem 2.7** [Neudecker 1969]. *If  $A, B, I \in M_n(\mathbb{R})$ , where  $I$  is the identity matrix, then*

$$\text{vec}(AB) = (I \otimes A) \text{vec}(B) = (B^T \otimes I) \text{vec}(A).$$

The following notation describes the submatrices corresponding to certain rows or columns.

**Definition 2.8.** If  $A \in M_{m,n}(\mathbb{R})$  and  $\varepsilon \subseteq \{1, \dots, m\}$ , then  $A[\varepsilon]$  is defined as the submatrix of  $A$  lying in the rows  $\varepsilon$ . The notation  $A[s]$  may also be used to indicate the  $s$ -th row in  $A$ .

**Definition 2.9.** If  $A \in M_{m,n}(\mathbb{R})$  and  $\varepsilon \subseteq \{1, \dots, n\}$ , then  $A(\varepsilon)$  is defined as the submatrix of  $A$  lying in the columns  $\varepsilon$ . The notation  $A(s)$  may also be used to indicate the  $s$ -th column in  $A$ .

For example, let  $A \in M_3(\mathbb{R})$  and let  $\varepsilon = \{1, 3\}$ . If we have

$$A = \begin{bmatrix} 35 & 24 & 19 \\ 39 & 76 & 14 \\ 12 & 7 & 20 \end{bmatrix}, \text{ then } A[\varepsilon] = \begin{bmatrix} 35 & 24 & 19 \\ 12 & 7 & 20 \end{bmatrix} \text{ and } A(\varepsilon) = \begin{bmatrix} 35 & 19 \\ 39 & 14 \\ 12 & 20 \end{bmatrix}.$$

### 3. The column space approach

The  $\text{vec}$  function is a vector space isomorphism which is used to convert linear matrix equations into linear equations. In this section, we show that unspecified entry locations in maximally admissible patterns correspond to full rank submatrices of a certain matrix.

Let  $A_1, \dots, A_k, B_1, \dots, B_k, C$  be  $n \times n$  real matrices. Applying Theorem 2.7 to the matrix equation  $A_1XB_1 + \dots + A_kXB_k = C$  yields the linear equation

$$(B_1^T \otimes A_1 + \dots + B_k^T \otimes A_k) \text{vec}(X) = \text{vec}(C).$$

The solution space of  $A_1XB_1 + A_2XB_2 + \dots + A_kXB_k = \mathbb{0}$  is isomorphic to the nullspace of  $B_1^T \otimes A_1 + B_2^T \otimes A_2 + \dots + B_k^T \otimes A_k$ . Throughout this section, we denote this  $n^2 \times n^2$  matrix  $B_1^T \otimes A_1 + \dots + B_k^T \otimes A_k$  as  $K$ .

**Lemma 3.1.** *Let  $A_1, \dots, A_k, B_1, \dots, B_k, C \in M_n(\mathbb{R})$  and  $\alpha$  be an  $n \times n$  partial matrix pattern. There exists a completion  $\widehat{\mathcal{X}}$  of the  $\alpha$ -partial matrix  $\mathcal{X}$  satisfying  $A_1\widehat{\mathcal{X}}B_1 + \dots + A_k\widehat{\mathcal{X}}B_k = C$  if and only if*

$$\text{vec}(C) - \sum_{(i,j) \in \alpha} x_{ij} K(i + (j - 1)n) \in \text{span}\{K(i + (j - 1)n) \mid (i, j) \notin \alpha\}.$$

*Proof.* The matrix equation  $A_1\mathcal{X}B_1 + \dots + A_k\mathcal{X}B_k = C$  is equivalent to the equation  $(B_1^T \otimes A_1 + \dots + B_k^T \otimes A_k) \text{vec}(\mathcal{X}) = \text{vec}(C)$  where  $\mathcal{X}$  has specified and unspecified entries. As above, let  $K = B_1^T \otimes A_1 + \dots + B_k^T \otimes A_k$ . Separating the specified and unspecified entries of  $\mathcal{X}$  we rewrite this equation as

$$\sum_{(i,j) \notin \alpha} x_{ij}K(i+(j-1)n) + \sum_{(i,j) \in \alpha} x_{ij}K(i+(j-1)n) = \text{vec}(C),$$

where  $x_{ij}$  are the entries in the partial matrix  $\mathcal{X}$ . In the first sum, the entries are unspecified while in the second sum, the entries  $x_{ij}$  are specified. Moving the specified entries to the right-hand side yields the linear equation

$$\sum_{(i,j) \notin \alpha} x_{ij}K(i+(j-1)n) = \text{vec}(C) - \sum_{(i,j) \in \alpha} x_{ij}K(i+(j-1)n).$$

This is solvable if and only if the vector on the right-hand side lies in  $\text{span}\{K(i+(j-1)n \mid (i,j) \notin \alpha)\}$ . □

This lemma tells us precisely when a partial matrix can be completed to satisfy a linear matrix equation and describes the linear system that must be solvable in order to complete a partial matrix. If  $C$  is the zero matrix, then the condition for the existence of a completion simplifies to

$$\sum_{(i,j) \in \alpha} x_{ij}K(i+(j-1)n) \in \text{span}\{K(i+(j-1)n \mid (i,j) \notin \alpha)\},$$

which can be answered by determining which sets of columns of  $K$  have rank equal to the rank of  $K$ . With some abuse of notation, let  $K(\alpha)$  denote the submatrix of columns of  $K$  corresponding to specified entries and  $K(\bar{\alpha})$  denote the submatrix of columns of  $K$  corresponding to unspecified entries.

**Theorem 3.2.** *Let  $A_1, \dots, A_k, B_1, \dots, B_k, C \in M_n(\mathbb{R})$ ,  $\alpha$  be an  $n \times n$  partial matrix pattern, and  $K = B_1^T \otimes A_1 + \dots + B_k^T \otimes A_k$ . Then, the following statements are equivalent:*

- (1) *For a given  $\alpha$ -partial matrix  $\mathcal{X}$  and any  $C \in M_n(\mathbb{R})$  such that  $\text{vec}(C) \in \text{span}\{K(1), \dots, K(n^2)\}$ , there exists a completion  $\widehat{\mathcal{X}}$  of  $\mathcal{X}$  such that  $A_1\widehat{\mathcal{X}}B_1 + \dots + A_k\widehat{\mathcal{X}}B_k = C$ .*
- (2)  $\text{rank}(K) = \text{rank}(K(\bar{\alpha}))$ .

*Proof.* Assuming (1), by Lemma 3.1  $\text{vec}(C) - \sum_{(i,j) \in \alpha} x_{ij}K(i+(j-1)n)$  is in the span of  $\{K(i+(j-1)n \mid (i,j) \notin \alpha)\}$  for all  $\text{vec}(C)$  in the span of the columns of  $K$ . Since it is possible to choose  $C$  so that it is any vector in the column space of  $K$ , it follows that the column space of  $K$  is contained in the column space of  $K(\bar{\alpha})$ , and  $\text{rank}(K) = \text{rank}(K(\bar{\alpha}))$ , proving the second statement.

Assuming (2), for any  $\text{vec}(C)$  in the column space of  $K(\bar{\alpha})$  and any  $\alpha$ -partial matrix  $\mathcal{X}$ , the column space  $K(\bar{\alpha})$  is the column space of  $K$ , since  $K(\bar{\alpha})$  is contained in the column space of  $K$  and both matrices have the same rank. In particular, the column space of  $K(\alpha)$  is contained in the column space of  $K(\bar{\alpha})$ . Then for any  $\text{vec}(C)$  in the column space of  $K$ ,  $\text{vec}(C)$  lies in the column space of  $K(\bar{\alpha})$ , and by Lemma 3.1 there exists a completion  $\widehat{\mathcal{X}}$  of the  $\alpha$ -partial matrix  $\mathcal{X}$  such that  $A_1\widehat{\mathcal{X}}B_1 + \cdots + A_k\widehat{\mathcal{X}}B_k = C$ , establishing the first statement.  $\square$

In this paper, the specific matrix equations of interest are homogeneous. The following corollary gives the condition that we use to classify patterns for this column space approach: the rank of the columns of  $K$  corresponding to unspecified entries must equal the rank of  $K$ . That is, the sets of columns of  $K$  with full rank correspond to unspecified entry locations in admissible patterns.

**Corollary 3.3.** *Let  $A_1, \dots, A_k, B_1, \dots, B_k \in M_n(\mathbb{R})$ ,  $\alpha$  be an  $n \times n$  matrix pattern, and  $K = B_1^T \otimes A_1 + \cdots + B_k^T \otimes A_k$ . Then, the following statements are equivalent:*

- (1) *The matrix pattern  $\alpha$  is admissible for the matrix equation  $A_1XB_1 + \cdots + A_kXB_k = 0$ .*
- (2)  $\text{rank}(K) = \text{rank}(K(\bar{\alpha}))$ .

*Proof.* This follows from the definition of admissibility, Theorem 3.2, and the fact that  $0$  is in the span of the columns of  $K$ .  $\square$

Corollary 3.3 gives the size of a maximally admissible pattern, namely  $n^2 - \text{rank}(K)$ .

**Corollary 3.4.** *Let  $A_1, \dots, A_k, B_1, \dots, B_k \in M_n(\mathbb{R})$  and  $K = B_1^T \otimes A_1 + \cdots + B_k^T \otimes A_k$ . If  $\alpha$  is an admissible  $n \times n$  partial matrix pattern for the matrix equation  $A_1XB_1 + \cdots + A_kXB_k = 0$ ,*

$$|\alpha| \leq n^2 - \text{rank}(K).$$

*Proof.* If  $|\alpha| > n^2 - \text{rank}(K)$ , then the number of columns corresponding to unspecified entries is strictly less than the  $\text{rank}(K)$  and condition (2) of Corollary 3.3 can never be satisfied.  $\square$

Given a linear matrix equation, the patterns  $\alpha$  that are admissible are exactly the patterns that set unspecified entries against a set of columns of  $K$  whose span is equal to the span of all the columns of  $K$ . With the column space approach we think of the specified entries of  $X$  as removing certain columns from  $K$ . We then look at the submatrix formed by the remaining columns of  $K$  and determine its rank. An  $\alpha$ -partial pattern is admissible if the rank of the columns of  $K$  corresponding to unspecified entries is equal to the rank of  $K$ .

The following lemmas establish two basic properties of matrix patterns: subpatterns of admissible patterns are admissible and patterns that contain inadmissible patterns are inadmissible.

**Lemma 3.5.** *Let  $\alpha$  and  $\beta$  be partial matrix patterns such that  $\alpha$  is admissible for the matrix equation  $A_1XB_1 + \dots + A_kXB_k = 0$ , where  $A_1, \dots, A_k, B_1, \dots, B_k \in M_n(\mathbb{R})$ , and let  $K = B_1^T \otimes A_1 + \dots + B_k^T \otimes A_k$ . If  $\beta \subseteq \alpha$ , then  $\beta$  is admissible for the matrix equation  $A_1XB_1 + \dots + A_kXB_k = 0$ .*

*Proof.* By Corollary 3.3,  $\alpha$  is admissible if and only if  $\text{rank}(K(\bar{\alpha})) = \text{rank}(K)$ . Since  $\beta \subseteq \alpha$ ,  $\text{rank}(K(\bar{\alpha})) \leq \text{rank}(K(\bar{\beta}))$ . This forces  $\text{rank}(K(\bar{\beta})) = \text{rank}(K)$ , and  $\beta$  is admissible for the matrix equation  $A_1XB_1 + \dots + A_kXB_k = 0$ .  $\square$

**Lemma 3.6.** *Let  $\alpha$  and  $\beta$  be partial matrix patterns such that  $\alpha$  is inadmissible for the matrix equation  $A_1XB_1 + \dots + A_kXB_k = 0$ , where  $A_1, \dots, A_k, B_1, \dots, B_k \in M_n(\mathbb{R})$ , and let  $K = B_1^T \otimes A_1 + \dots + B_k^T \otimes A_k$ . If  $\alpha \subseteq \beta$ , then  $\beta$  is also inadmissible for the matrix equation  $A_1XB_1 + \dots + A_kXB_k = 0$ .*

*Proof.* By Corollary 3.3,  $\alpha$  is admissible if and only if  $\text{rank}(K(\bar{\alpha})) = \text{rank}(K)$ . Since  $\alpha$  is inadmissible,  $\text{rank}(K(\bar{\alpha})) < \text{rank}(K)$ . Since  $\alpha \subseteq \beta$ ,  $\text{rank}(K(\bar{\beta})) \leq \text{rank}(K(\bar{\alpha}))$ . This forces  $\text{rank}(K(\bar{\beta})) < \text{rank}(K)$ , and  $\beta$  is also inadmissible for the matrix equation  $A_1XB_1 + \dots + A_kXB_k = 0$ .  $\square$

#### 4. The nullspace approach

In this section we develop a second criterion for admissible patterns for the homogeneous matrix equation  $A_1XB_1 + \dots + A_kXB_k = 0$ . We show that if the specified entry locations of a pattern correspond to full rank submatrices of a matrix constructed from a basis of the solution space of the homogeneous matrix equation, the pattern is admissible. We also construct a basis for the solution space of two special cases of this matrix equation

**Nullspace criterion.** Given a partial matrix, we need to determine if the specified entries of the partial matrix can be written as a linear combination of basis elements for the solution space of  $A_1XB_1 + \dots + A_kXB_k = 0$ . Let  $\{V_1, V_2, \dots, V_n\}$  be a basis for the solution space, then  $\{\text{vec}(V_1), \text{vec}(V_2), \dots, \text{vec}(V_n)\}$  is a basis for the nullspace of  $B_1^T \otimes A_1 + \dots + B_k^T \otimes A_k$ . Throughout this paper we denote this matrix  $[\text{vec}(V_1) \text{vec}(V_2) \dots \text{vec}(V_n)]$  as  $N$ .

The partial matrix has a completion if there exist scalars  $c_1, \dots, c_n$  such that the specified entries of  $\mathcal{X}$  satisfy

$$\mathcal{X} = c_1V_1 + c_2V_2 + \dots + c_nV_n.$$

Applying the  $\text{vec}$  function to this equation yields

$$\text{vec}(\mathcal{X}) = [\text{vec}(V_1) \text{vec}(V_2) \dots \text{vec}(V_n)]\mathbf{c} = N\mathbf{c},$$

where  $\mathbf{c} = [c_1 \ c_2 \ \cdots \ c_n]^T$ . Only the rows in  $\text{vec}(\mathcal{X})$  which are specified are of interest because the unspecified entries can be freely chosen. Let  $\varepsilon = \{i + (j - 1)n \mid (i, j) \in \alpha\}$ , the set of integer values corresponding to the rows of  $\text{vec}(\mathcal{X})$  which contain specified entries. Solving the equation

$$\text{vec}(\mathcal{X})[\varepsilon] = [\text{vec}(V_1)[\varepsilon] \ \text{vec}(V_2)[\varepsilon] \ \cdots \ \text{vec}(V_n)[\varepsilon]] = N[\varepsilon]\mathbf{c}$$

is equivalent to determining if the specified entries of  $\mathcal{X}$  can be written as a linear combination of basis elements to the solution space of our linear equation.

The following theorem describes the nullspace condition for admissibility: the submatrix of rows of  $N$  corresponding to specified entries must have rank at least equal to the number of specified entries in  $\mathcal{X}$

**Theorem 4.1.** *Let  $\alpha$  be an  $n \times n$  partial matrix pattern and  $\{V_1, V_2, \dots, V_\ell\}$  be a basis for the solution space of the matrix equation  $A_1 X B_1 + \cdots + A_k X B_k = 0$ . The matrix pattern  $\alpha$  is admissible for this matrix equation if and only if  $\text{rank}(N[\varepsilon]) \geq |\alpha|$ , where  $\varepsilon = \{i + (j - 1)n \mid (i, j) \in \alpha\}$  and*

$$N[\varepsilon] = [\text{vec}(V_1)[\varepsilon] \ \text{vec}(V_2)[\varepsilon] \ \cdots \ \text{vec}(V_\ell)[\varepsilon]].$$

*Proof.* The matrix completion problem is equivalent to determining if there exists a solution to the linear equation  $\text{vec}(X)[\varepsilon] = N[\varepsilon]\mathbf{c}$ .  $N[\varepsilon]$  is an  $n \times |\alpha|$  matrix, so this equation is solvable for all  $\text{vec}(X)[\varepsilon]$  if and only if  $\text{rank}(N[\varepsilon]) \geq |\alpha|$ . If so, there exists a completion  $\widehat{\mathcal{X}}$  for any  $\mathcal{X}$  satisfying  $A_1 \widehat{\mathcal{X}} B_1 + \cdots + A_k \widehat{\mathcal{X}} B_k = 0$ .

If  $\text{rank}(N[\varepsilon]) < |\alpha|$ , then  $\text{vec}(X)[\varepsilon] = N[\varepsilon]\mathbf{c}$  has a solution if  $\text{vec}(X)[\varepsilon]$  lies in the span of the columns of  $N[\varepsilon]$ . Since  $\text{rank}(N[\varepsilon]) < |\alpha|$  and  $\text{vec}(X)[\varepsilon]$  is an  $|\alpha|$ -dimensional vector, there exists an  $\alpha$ -partial matrix  $\mathcal{X}$  such that  $\text{vec}(X)[\varepsilon]$  does not lie in the span of the columns of  $N[\varepsilon]$ . Hence for this  $\alpha$ -partial matrix  $\mathcal{X}$  there does not exist a completion of  $A_1 X B_1 + \cdots + A_k X B_k = 0$ . Since this  $\alpha$  does not have a completion for all  $\alpha$ -partial matrices,  $\alpha$  is inadmissible.  $\square$

For maximal patterns, the condition for admissibility is that the number of specified entries in  $\mathcal{X}$  must equal the rank of  $N[\varepsilon]$ .

**Corollary 4.2.** *Let  $\alpha$  be an  $n \times n$  partial matrix pattern for the matrix equation  $A_1 X B_1 + \cdots + A_k X B_k = 0$  and let  $\{V_1, V_2, \dots, V_\ell\}$  be a basis for the solution space of the given matrix equation. An admissible pattern  $\alpha$  is maximal if and only if  $|\alpha| = \ell$ .*

*Proof.* First assume that the admissible pattern is maximally admissible to show that the number of specified entries equals the dimension of the solution space. For the pattern to be admissible, the rank of  $N[\varepsilon]$  must be greater than  $|\alpha|$ , but also must not exceed the number of columns in  $N[\varepsilon]$ . Then, the greatest possible value for the rank of  $N[\varepsilon]$  is  $\ell$ , namely the dimension of the solution space.

We next assume that the number of specified entries equals the dimension of the solution space to show that the admissible pattern is maximal. Then, since the dimension of the solution space is  $\ell$ ,  $|\alpha| = \ell$ . Since  $N[\varepsilon]$  has  $\ell$  columns and by Theorem 4.1,  $|\alpha| \leq \text{rank}(N[\varepsilon]) \leq \ell$ , the rank of  $N[\varepsilon]$  must equal  $\ell$ . Therefore,  $\alpha$  is maximally admissible because the dimension of  $\alpha$  is as large as possible while maintaining admissibility.  $\square$

**Construction of bases for the nullspace.** We construct a basis for the solution space of the matrix equation  $AX + XB = 0$  using eigenvectors of the matrices  $A$  and  $B$ . This basis is used to classify patterns for the commutativity equation  $AX - XA = 0$  and the skew-Lyapunov equation  $AX - XA^T = 0$  in Section 5.

**Theorem 4.3** [Horn and Johnson 1991]. *Let  $A \in M_n(\mathbb{R})$  and  $B \in M_m(\mathbb{R})$  be given. If  $\lambda$  is an eigenvalue of  $A$  and  $\mathbf{x} \in \mathbb{C}^n$  is a corresponding eigenvector of  $A$ , and if  $\mu$  is an eigenvalue of  $B$  and  $\mathbf{y} \in \mathbb{C}^m$  is a corresponding eigenvector of  $B$ , then  $\lambda + \mu$  is an eigenvalue of  $(I_m \otimes A) + (B \otimes I_n)$ , and  $\mathbf{y} \otimes \mathbf{x} \in \mathbb{C}^{nm}$  is a corresponding eigenvector. Every eigenvalue of  $(I_m \otimes A) + (B \otimes I_n)$  arises as such a sum of eigenvalues of  $A$  and  $B$ , and  $I_m \otimes A$  commutes with  $B \otimes I_n$ . If the set of eigenvalues of  $A$  equals  $\{\lambda_1, \lambda_2, \dots, \lambda_n\}$  and the set of eigenvalues of  $B$  equals  $\{\mu_1, \mu_2, \dots, \mu_m\}$ , then the set of eigenvalues of  $(I_m \otimes A) + (B \otimes I_n)$  equals  $\{\lambda_i + \mu_j \mid i = 1, \dots, n, j = 1, \dots, m\}$  (including algebraic multiplicities in all three cases).*

We use the lemma below to construct bases for  $I \otimes A - A^T \otimes I$  and  $I \otimes A - A \otimes I$ .

**Lemma 4.4.** *If  $\{\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^n\}$  and  $\{\mathbf{y}^1, \mathbf{y}^2, \dots, \mathbf{y}^n\}$  are each linearly independent sets of nonzero vectors, then  $\{\mathbf{y}^1 \otimes \mathbf{x}^1, \mathbf{y}^2 \otimes \mathbf{x}^2, \dots, \mathbf{y}^n \otimes \mathbf{x}^n\}$  is linearly independent.*

*Proof.* Let

$$\mathbf{x}^i = [x_1^i \ x_2^i \ \dots \ x_n^i]^T \quad \text{and} \quad \mathbf{y}^i = [y_1^i \ y_2^i \ \dots \ y_n^i]^T.$$

By the definition of the Kronecker product,

$$\mathbf{y}^i \otimes \mathbf{x}^i = [y_1^i \mathbf{x}^i \ y_2^i \mathbf{x}^i \ \dots \ y_n^i \mathbf{x}^i]^T.$$

We want to show that

$$a_1(\mathbf{y}^1 \otimes \mathbf{x}^1) + a_2(\mathbf{y}^2 \otimes \mathbf{x}^2) + \dots + a_n(\mathbf{y}^n \otimes \mathbf{x}^n) = \mathbf{0} \quad \text{only when } a_1 = a_2 = \dots = a_n = 0.$$

Using the Kronecker product definition, this can be rewritten as

$$\begin{aligned} (a_1 y_1^1) \mathbf{x}^1 + (a_2 y_1^2) \mathbf{x}^2 + \dots + (a_n y_1^n) \mathbf{x}^n &= \mathbf{0}, \\ (a_1 y_2^1) \mathbf{x}^1 + (a_2 y_2^2) \mathbf{x}^2 + \dots + (a_n y_2^n) \mathbf{x}^n &= \mathbf{0}, \\ &\vdots \\ (a_1 y_n^1) \mathbf{x}^1 + (a_2 y_n^2) \mathbf{x}^2 + \dots + (a_n y_n^n) \mathbf{x}^n &= \mathbf{0}. \end{aligned}$$



Since  $\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^n$  are linearly independent,

$$a_1 \mathbf{y}^1 = 0, a_2 \mathbf{y}^2 = 0, \dots, a_n \mathbf{y}^n = 0.$$

Since  $\mathbf{y}^1, \mathbf{y}^2, \dots, \mathbf{y}^n$  are nonzero vectors, there exists at least one nonzero entry in each vector. This implies that  $a_1 = a_2 = \dots = a_n = 0$ . Therefore  $\{\mathbf{y}^1 \otimes \mathbf{x}^1, \mathbf{y}^2 \otimes \mathbf{x}^2, \dots, \mathbf{y}^n \otimes \mathbf{x}^n\}$  is linearly independent.  $\square$

**Remark 4.5.** If we further assume that  $A$  has distinct eigenvalues, then the nullities of  $(I \otimes A) - (A^T \otimes I)$  and  $(I \otimes A) - (A \otimes I)$  are both  $n$  (see Section 5). This and Lemma 4.4 imply that  $\{\mathbf{y}^1 \otimes \mathbf{x}^1, \mathbf{y}^2 \otimes \mathbf{x}^2, \dots, \mathbf{y}^n \otimes \mathbf{x}^n\}$  is a basis for the nullspace of  $(I \otimes A) - (A^T \otimes I)$ , where  $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  is a basis of eigenvectors for  $A$  corresponding to eigenvalues  $\lambda_1, \dots, \lambda_n$  and  $\{\mathbf{y}_1, \dots, \mathbf{y}_n\}$  is a basis of eigenvectors for  $-A^T$  corresponding to eigenvalues  $-\lambda_1, \dots, -\lambda_n$ . Similarly  $\{\mathbf{x}^1 \otimes \mathbf{x}^1, \mathbf{x}^2 \otimes \mathbf{x}^2, \dots, \mathbf{x}^n \otimes \mathbf{x}^n\}$  is a basis for the nullspace of  $(I \otimes A) - (A \otimes I)$ .

### 5. Admissible patterns for certain matrix equations

In this section, we apply the column space and nullspace approaches to three matrix equations: the skew-symmetric equation, the commutativity equation, and the skew-Lyapunov equation. For the skew-symmetric equation, we completely characterize admissible patterns. For the other two matrix equations we classify certain patterns as admissible or inadmissible.

For the skew-symmetric equation,  $AX - A^T X = 0$ , Theorem 5.2 states that a maximal pattern is admissible if and only if it contains one specified entry in each column of an  $\alpha$ -partial matrix  $\mathcal{X}$ . We also show all admissible patterns are subpatterns of maximal patterns.

For the commutativity equation,  $AX - XA = 0$ , Theorem 5.8 states that maximal patterns with no diagonal entries specified are inadmissible. Theorem 5.9 states that patterns in which all of the specified entries are in the same row or in the same column are admissible.

For the skew-Lyapunov equation,  $AX - XA^T = 0$ , Theorem 5.12 states that a pattern is admissible if all of the specified entries reside in the  $i$ -th row or column without  $(i, j)$  and  $(j, i)$  both being in the pattern for any  $j$ . Corollary 5.15 states that if any pattern contains two specified entries which are located across the main diagonal from each other, then the pattern is inadmissible.

**Patterns for the skew-symmetric equation.** Applying the vec function to  $AX - A^T X = 0$  yields the linear equation

$$(I \otimes (A - A^T)) \text{vec}(X) = 0.$$

The matrix  $A - A^T$  is skew-symmetric, so  $(I \otimes (A - A^T))$  is a block diagonal matrix and is skew-symmetric. We denote  $I \otimes (A - A^T)$  as  $S_A$ .

Since  $A - A^T$  is skew-symmetric, it is also diagonalizable and its eigenvalues are purely imaginary or zero [Rukmangadachari 2010]. The rank of  $A - A^T$  is dependent upon whether  $n$  is odd or even.

In this section, we assume that  $A - A^T$  has maximum rank. So  $\text{rank}(A - A^T) = n$  if  $n$  is even, and  $\text{rank}(A - A^T) = n - 1$  if  $n$  is odd. The set of matrices  $A$  with which  $\text{rank}(A - A^T)$  is strictly less than the maximum possible rank is a set of measure zero. So in this section our “generic” property of  $A$  is that  $\text{rank}(A - A^T)$  is maximal.

Since  $S_A$  is a block-diagonal matrix consisting of the matrix  $A - A^T$  down the main diagonal,  $\text{rank}(S_A) = n \cdot \text{rank}(A - A^T)$ . By Corollary 4.2 maximally admissible patterns for  $S_A$  contain  $n$  specified entries for  $n$  odd. Since the nullity of  $S_A$  is zero when  $n$  is even, only the empty pattern, the pattern with no specified entries is admissible.

From this point forward, we only consider the case when  $n$  is odd. We first construct a basis for the nullspace of  $S_A$  in order to apply the nullspace approach.

**Lemma 5.1.** *Let  $A \in M_n(\mathbb{R})$  with  $n$  odd and  $\text{rank}(A - A^T) = n - 1$ , and let  $\{v\}$  be a basis for the nullspace of  $A - A^T$ . If  $n$  is odd, then*

$$\mathcal{B} = \{[v \ 0 \ \dots \ 0], [0 \ v \ 0 \ \dots \ 0], \dots, [0 \ \dots \ 0 \ v]\}$$

*is a basis for the solution space of  $AX - A^T X = 0$ .*

*Proof.* Each element of  $\mathcal{B}$  is a solution to  $AX - A^T X = 0$ . The matrices in  $\mathcal{B}$  are clearly linearly independent. The dimension of the solution space of  $AX - A^T X = 0$  is  $n$ , and  $\mathcal{B}$  contains  $n$  elements. So  $\mathcal{B}$  is a basis for the solution space of  $AX - A^T X = 0$ . □

We now consider maximally admissible patterns for the skew-symmetric equation, and determine whether they are admissible or inadmissible.

**Theorem 5.2.** *Let  $\alpha$  be an  $n \times n$  partial matrix pattern with  $|\alpha| = n$ , and let  $n$  be odd. The matrix pattern  $\alpha$  is maximally admissible for the matrix equation  $AX - A^T X = 0$  for almost all  $A$  with  $\text{rank}(A - A^T) = n - 1$  if and only if  $\alpha = \{(i_1, 1), (i_2, 2), \dots, (i_n, n)\}$ , where  $1 \leq i_k \leq n$ .*

*Proof.* We first show that if  $\alpha$  is admissible, then  $\alpha = \{(i_1, 1), (i_2, 2), \dots, (i_n, n)\}$ , where  $1 \leq i_k \leq n$ . We proceed by contraposition, assuming that

$$\alpha \neq \{(i_1, 1), (i_2, 2), \dots, (i_n, n)\}$$

to show that  $\alpha$  is inadmissible. By Lemma 5.1, a basis for the solution space of  $(A - A^T)X = 0$  is  $\{V_1, \dots, V_n\}$  where the  $i$ -th column of  $V_i$  is  $v$  and all other columns only contain zeros. Following the nullspace approach, the matrix completion

problem is equivalent to solving

$$\text{vec}(\mathcal{X})[\varepsilon] = [\text{vec}(V_1)[\varepsilon] \text{vec}(V_2)[\varepsilon] \cdots \text{vec}(V_n)[\varepsilon]]\mathbf{c},$$

where  $\mathbf{c} = [c_1 \ c_2 \ \cdots \ c_n]^T$  and  $\varepsilon = \{i + (j - 1)n \mid (i, j) \in \alpha\}$ . Let  $N$  be the matrix containing the column vectors of the basis elements, so

$$N[\varepsilon] = [\text{vec}(V_1)[\varepsilon] \text{vec}(V_2)[\varepsilon] \cdots \text{vec}(V_n)[\varepsilon]].$$

From our assumption, there exists at least one column in  $\mathcal{X}$  that does not have a specified entry. Without loss of generality, assume that the  $k$ -th column in  $\mathcal{X}$  does not have a specified entry. Any row in  $\text{vec}(V_k)$  that contains an element of  $\mathbf{v}$  will be excluded when  $\text{vec}(V_k)$  is restricted to  $\text{vec}(V_i)[\varepsilon]$ . We have, then, that  $\text{vec}(V_k)[\varepsilon] = \mathbf{0}$ . The rank of  $N[\varepsilon]$  is therefore strictly less than  $|\alpha|$ , and therefore  $\alpha$  is inadmissible.

We next show that if  $\alpha = \{(i_1, 1), (i_2, 2), \dots, (i_n, n)\}$ , where  $1 \leq i_k \leq n$ , then  $\alpha$  is admissible. Following the nullspace approach as above, this completion problem is equivalent to

$$\begin{aligned} \text{vec}(X)[\varepsilon] &= [\text{vec}(V_1)[\varepsilon] \text{vec}(V_2)[\varepsilon] \cdots \text{vec}(V_n)[\varepsilon]]\mathbf{c} \\ &= \begin{bmatrix} v_{i_1} & 0 & \cdots & 0 \\ 0 & v_{i_2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & v_{i_n} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix}, \end{aligned}$$

where  $v_{i_\ell}$  are entries in  $\mathbf{v}$ . For almost all  $A$ ,  $v_{i_\ell} \neq 0$  for all  $1 \leq \ell \leq n$ , and the rank of  $N[\varepsilon]$  is  $n$ . This means that the columns of  $N[\varepsilon]$  spans  $\mathbb{R}^n$ , and therefore any values that can be specified for  $\mathcal{X}$  are in the span of the columns of  $N[\varepsilon]$ . So any  $\alpha$ -partial matrix for the  $\alpha$  pattern can be completed to satisfy the skew-symmetric equation, and  $\alpha = \{(i_1, 1), (i_2, 2), \dots, (i_n, n)\}$  is admissible.  $\square$

This tells us that  $\alpha$  is maximally admissible if and only if  $\alpha$  contains exactly one specified entry in each column. Again “almost all” is used to say that these patterns are admissible for the given matrix equation, with  $A$  satisfying the given conditions, except for a set of matrices  $A$  of measure zero. In this case, we can be more specific. The set of matrices that these patterns are not admissible for are those matrices  $A$  for which the vector  $\mathbf{v}$  has zero entries, where  $\mathbf{v}$  is the basis for the nullspace of  $A - A^T$ . The following theorem shows that admissible patterns appear as subpatterns of maximal patterns.

**Theorem 5.3.** *Let  $A \in M_n(\mathbb{R})$  be nonderogatory with  $n$  odd. A pattern  $\beta$  is admissible for the matrix equation  $AX - A^T X = 0$  for almost all  $A$  with  $\text{rank}(A - A^T) = n - 1$  if and only if  $\beta \subseteq \{(i_1, 1), (i_2, 2), \dots, (i_n, n)\}$  with  $1 \leq i_k \leq n$ .*

*Proof.* By Theorem 5.2,  $\alpha = \{(i_1, 1), (i_2, 2), \dots, (i_n, n)\}$  is admissible. If  $\beta \subseteq \alpha$  then by Lemma 3.5  $\beta$  is also admissible.

If  $\beta \not\subseteq \{(i_1, 1), (i_2, 2), \dots, (i_n, n)\}$  then  $\{(i, k), (j, k)\} \subseteq \beta$  for some  $i \neq j$ . Then  $\varepsilon = \{i + (k - 1)n, j + (k - 1)n\}$  and

$$[\text{vec}(V_1)[\varepsilon] \text{vec}(V_2)[\varepsilon] \cdots \text{vec}(V_n)[\varepsilon]] = \begin{bmatrix} 0 & \cdots & 0 & v_i & 0 & \cdots & 0 \\ 0 & \cdots & 0 & v_j & 0 & \cdots & 0 \end{bmatrix}.$$

This matrix does not have full rank, so the pattern  $\{(i, k), (j, k)\}$  is inadmissible by the nullspace criterion. Since  $\{(i, k), (j, k)\} \subseteq \beta$ ,  $\beta$  is inadmissible by Lemma 3.6. □

Finally we give formulas for the number of maximally admissible and admissible patterns.

**Corollary 5.4.** *For  $A \in M_n(\mathbb{R})$  where  $n$  is odd and  $\text{rank}(A - A^T) = n - 1$ , the number of maximally admissible patterns for the skew-symmetric equation is  $n^n$ .*

*Proof.* From Theorem 5.2, if  $\alpha$  is admissible for the skew-symmetric equation, each column in  $\mathcal{X}$  has one specified entry. Each of the  $n$  columns has  $n$  possible locations where an entry can be specified, so the total number of admissible patterns is  $n^n$ . □

**Corollary 5.5.** *For  $A \in M_n(\mathbb{R})$  where  $n$  is odd and  $\text{rank}(A - A^T) = n - 1$ , the number of admissible patterns for the skew-symmetric equation is  $(1 + n)^n$ .*

*Proof.* We have by Theorem 5.3 that if  $\beta \subseteq \alpha$ , where  $\alpha = \{(i_1, 1), (i_2, 2), \dots, (i_n, n)\}$  and  $1 \leq i_k \leq n$ , then  $\beta$  is admissible for the skew-symmetric equation.

Suppose  $\beta$  has  $i$  specified entries, there are  $\binom{n}{i}$  choices for columns and  $n$  choices within each column. Summing over  $i$  and using the binomial theorem, the total number of admissible patterns is

$$\sum_{i=0}^n \binom{n}{i} n^i = (1 + n)^n. \quad \square$$

**Patterns for the commutativity equation.** We next classify patterns for the commutativity equation,  $AX - XA = 0$ . The conditions under which two matrices commute are well known, but there still are interesting questions that can be asked about matrix commutativity with regard to partial matrix completions [Horn and Johnson 1991]. We are interested in finding answers to the following: if given a partial matrix pattern  $\alpha$  and a matrix  $A$ , what are the conditions on the specific entries in an  $\alpha$ -partial matrix  $\mathcal{X}$  so that  $\mathcal{X}$  has a completion that commutes with  $A$ ? Which patterns  $\alpha$  allow any  $\alpha$ -partial matrix  $\mathcal{X}$  to be completed to commute with almost all  $A \in M_n(\mathbb{R})$ ?

We use the column space approach to convert the matrix equation into a linear equation. The  $\text{vec}$  function applied to the commutativity equation yields  $[(I \otimes A) - (A^T \otimes I)] \text{vec}(X) = 0$ . We denote  $(I \otimes A) - (A^T \otimes I)$  as  $\Omega_A$ .

**Lemma 5.6** [Horn and Johnson 1991]. *If  $A \in M_n(\mathbb{R})$  has  $k$  eigenvalues  $\{\lambda_1, \lambda_2, \dots, \lambda_k\}$ , then the dimension of the nullspace of  $\Omega_A$  is*

$$\sum_{i=1}^k m_a(\lambda_i)m_g(\lambda_i),$$

where  $m_a(\lambda), m_g(\lambda)$  are the algebraic and geometric multiplicities of  $\lambda$  respectively.

**Lemma 5.7** [Horn and Johnson 1991]. *For  $A \in M_n(\mathbb{R})$ , the dimension of the commutant of  $A$  is at least  $n$ , and the dimension of the commutant is equal to  $n$  if and only if  $A$  is nonderogatory.*

Because the solutions to the commutativity equation are exactly the elements of the commutant, the rank of  $\Omega_A$  is  $n^2 - n$  if and only if  $A$  is nonderogatory. Maximal patterns for the commutativity equation contain at most  $n$  specified entries for  $A$  nonderogatory.

We use two different bases for the nullspace of  $\Omega_A$  to classify admissible and inadmissible patterns. If  $A$  is nonderogatory, then only polynomials in  $A$  commute with  $A$  [Horn and Johnson 1985]. So one basis for the null space of  $\Omega_A$  is

$$\{\text{vec}(I), \text{vec}(A), \text{vec}(A^2), \dots, \text{vec}(A^{n-1})\}.$$

By Remark 4.5 if  $A$  has distinct eigenvalues then  $\{\mathbf{y}^1 \otimes \mathbf{x}^1, \mathbf{y}^2 \otimes \mathbf{x}^2, \dots, \mathbf{y}^n \otimes \mathbf{x}^n\}$  is also a second basis for the nullspace where  $\{\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^n\}$  is a set of eigenvectors for  $A$  and  $\{\mathbf{y}^1, \mathbf{y}^2, \dots, \mathbf{y}^n\}$  is a set of eigenvectors for  $-A^T$  corresponding to eigenvalues  $\{\lambda_1, \lambda_2, \dots, \lambda_n\}$  and  $\{-\lambda_1, -\lambda_2, \dots, -\lambda_n\}$  respectively.

We first show maximally admissible patterns must have a diagonal entry specified.

**Theorem 5.8.** *Let  $\alpha$  be an  $n \times n$  partial matrix pattern with  $|\alpha| = n$  and  $A \in M_n(\mathbb{R})$  be nonderogatory. If  $(i, i) \notin \alpha$  for all  $1 \leq i \leq n$ , then any  $\alpha$ -partial matrix  $\mathcal{X}$  is inadmissible for the matrix equation  $AX - XA = 0$ .*

*Proof.* Using the nullspace approach and the basis  $\{\text{vec}(I), \text{vec}(A), \dots, \text{vec}(A^{n-1})\}$ , the partial matrix completion problem for the commutativity equation is equivalent to solving

$$\text{vec}(\mathcal{X})[\varepsilon] = [\text{vec}(I)[\varepsilon] \text{vec}(A)[\varepsilon] \text{vec}(A^2)[\varepsilon] \dots \text{vec}(A^{n-1})[\varepsilon]]\mathbf{c},$$

where  $\varepsilon = \{i + (j - 1)n \mid (i, j) \in \alpha\}$ .

From our assumption, we have that  $(i, i) \notin \alpha$  for all  $1 \leq i \leq n$ . That is, no entries along the main diagonal are specified. Then, any row in  $\text{vec}(I)$  that contains a 1 will be excluded in  $\text{vec}(I)[\varepsilon]$ , so  $\text{vec}(I)[\varepsilon] = \mathbf{0}$ .

This means that  $\text{rank}(N[\varepsilon]) < n = |\alpha|$ . By Theorem 4.1,  $\alpha$  is inadmissible.  $\square$

We now partially classify maximally admissible patterns for the commutativity equation.

**Theorem 5.9.** *Let  $\alpha$  be an  $n \times n$  partial matrix pattern with  $|\alpha| = n$ . If  $\alpha = \{(i, 1), (i, 2), \dots, (i, n)\}$  or  $\alpha = \{(1, j), (2, j), \dots, (n, j)\}$  where  $1 \leq i, j \leq n$ , then  $\alpha$  is maximally admissible for the commutativity equation  $AX - XA = 0$  for almost all  $A$ , where all  $A$  have distinct eigenvalues.*

*Proof.* By Remark 4.5,  $\{\mathbf{y}^1 \otimes \mathbf{x}^1, \mathbf{y}^2 \otimes \mathbf{x}^2, \dots, \mathbf{y}^n \otimes \mathbf{x}^n\}$  is a basis for the nullspace of  $\Omega_A$  where  $\{\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^n\}$  is a set of eigenvectors for  $A$  and  $\{\mathbf{y}^1, \mathbf{y}^2, \dots, \mathbf{y}^n\}$  is a set of eigenvectors for  $-A^T$ .

Following the nullspace approach, the commutativity matrix completion problem is equivalent to solving

$$\begin{aligned} \text{vec}(\mathcal{X})[\varepsilon] &= [\text{vec}(\mathbf{y}^1 \otimes \mathbf{x}^1)[\varepsilon] \text{vec}(\mathbf{y}^2 \otimes \mathbf{x}^2)[\varepsilon] \dots \text{vec}(\mathbf{y}^n \otimes \mathbf{x}^n)[\varepsilon]]\mathbf{c} \\ &= [x_i^1 \mathbf{y}^1 \ x_i^2 \mathbf{y}^2 \ \dots \ x_i^n \mathbf{y}^n]\mathbf{c}, \end{aligned}$$

where  $\mathbf{c} = [c_1 \ c_2 \ \dots \ c_n]^T$  and  $\varepsilon = \{i + (j - 1)n \mid (i, j) \in \alpha\}$ .

Since  $\{\mathbf{y}^1, \mathbf{y}^2, \dots, \mathbf{y}^n\}$  is linearly independent,  $\{x_i^1 \mathbf{y}^1, x_i^2 \mathbf{y}^2, \dots, x_i^n \mathbf{y}^n\}$  is linearly independent because its elements are scalar multiples of the elements in the linearly independent set  $\{\mathbf{y}^1, \mathbf{y}^2, \dots, \mathbf{y}^n\}$  and for almost all  $A$ , we have  $x_j^i \neq 0$ , because for almost all  $A$ , it follows that  $x_j^i \neq 0$ . As a result, the columns of  $N[\varepsilon]$  span  $\mathbb{R}^n$ . As such, any  $\text{vec}(\mathcal{X})[\varepsilon]$  lies in the span of the columns of  $N[\varepsilon]$ . Therefore  $\alpha$  is admissible.

The proof that  $\alpha = \{(1, j), (2, j), \dots, (n, j)\}$ , where  $1 \leq j \leq n$ , is admissible is similar. □

This shows that patterns including an entire row or entire column of specified entries is maximally admissible. For specific  $n$ , we can show that there exist other admissible patterns, and we conjecture that a pattern with  $n$  specified entries is admissible if and only if it has at least one diagonal entry specified. The following corollary describes a subset of admissible patterns.

**Corollary 5.10.** *If*

$$\beta \subseteq \{(i, 1), (i, 2), \dots, (i, n)\} \quad \text{or} \quad \beta \subseteq \{(1, j), (2, j), \dots, (n, j)\},$$

where  $1 \leq i, j \leq n$ , then  $\beta$  is admissible.

*Proof.* This follows by Theorem 5.9 and Lemma 3.5. □

**Patterns for the skew-Lyapunov equation.** Lastly we classify patterns for the skew-Lyapunov equation,  $AX - XA^T = 0$ . Applying the  $\text{vec}$  function to  $AX - XA^T = 0$  yields the linear equation  $[(I \otimes A) - (A \otimes I)] \text{vec}(X) = 0$ . We denote  $(I \otimes A) - (A \otimes I)$  as  $\Psi_A$ . The rank of  $\Psi_A$  determines the maximum number of specified entries in an admissible pattern. In this section, we assume  $A$  has distinct eigenvalues, and consider the rank of  $\Psi_A$  under this condition. The following result gives us an upper bound for the nullity of  $\Psi_A$ .

**Lemma 5.11** [Morris 2015]. *Let  $A \in M_n(\mathbb{R})$  and  $B \in M_n(\mathbb{R})$  be similar matrices with eigenvalues  $\{\lambda_1, \lambda_2, \dots, \lambda_k\}$ , then*

$$\text{nullity}(I_n \otimes A + (-B^T) \otimes I_n) \leq \sum_{i=1}^k a_i^2$$

and

$$n^2 - \sum_{i=1}^k m_a(\lambda_i)^2 \leq \text{rank}(I_n \otimes A + (-B^T) \otimes I_n) \leq n^2.$$

For  $A \in M_n(\mathbb{R})$  with distinct eigenvalues, the maximum nullity of  $\Psi_A$  is  $n$ , and we can construct  $n$  linearly independent vectors in the nullspace.

Since the nullity of  $\Psi_A$  is  $n$ , maximally admissible patterns for  $AX - XA^T = 0$  will have  $n$  specified entries. We proceed by determining a basis for the solution space of the skew-Lyapunov equation. This is equivalent to finding a basis for the nullspace of  $\Psi_A$ .

The following theorem partially classifies maximally admissible patterns for the skew-Lyapunov equation. Maximally admissible patterns contain  $n$  specified entries by Corollary 3.4. We first show that if the same numbered column and row have a total of  $n$  specified entries, then the pattern is admissible.

**Theorem 5.12.** *Let  $A \in M_n(\mathbb{R})$  with distinct eigenvalues and  $\alpha$  be an  $n \times n$  partial matrix pattern. Given  $k \in \{1, \dots, n\}$ , if exactly one of  $(k, i)$  or  $(i, k)$  is in  $\alpha$  for all  $1 \leq i \leq n$ , then  $\alpha$  is maximally admissible for the matrix equation  $AX - XA^T = 0$  for almost all  $A$ , where all  $A$  have distinct eigenvalues.*

*Proof.* Noting that the rows of  $N$  corresponding to the  $(i, j)$  and  $(j, i)$  entries are equal, this theorem is a special case of Theorem 5.9 with  $\{\mathbf{x}^1 \otimes \mathbf{x}^1, \mathbf{x}^2 \otimes \mathbf{x}^2, \dots, \mathbf{x}^n \otimes \mathbf{x}^n\}$  as a basis for the solution space. □

**Corollary 5.13.** *For  $A \in M_n(\mathbb{R})$  with distinct eigenvalues, if  $\beta \subseteq \{(1, k), \dots, (n, k)\}$  or  $\beta \subseteq \{(k, 1), \dots, (k, n)\}$  then  $\beta$  is admissible for the matrix equation  $AX - XA^T = 0$  for almost all  $A$ , where all  $A$  have distinct eigenvalues.*

*Proof.* This follows by Theorem 5.12 and Lemma 3.5. □

We next classify patterns as inadmissible. If  $\alpha$  is admissible, then there are no pairs of specified entries which reside opposite the main diagonal from each other. Equivalently, if there exists a pair of specified entries such that they are across the main diagonal from each other, then the pattern will be inadmissible.

**Theorem 5.14.** *For  $A \in M_n(\mathbb{R})$ , if  $\alpha = \{(i, j), (j, i)\}$  such that  $i \neq j$  and  $1 \leq i, j \leq n$ , then  $\alpha$  is inadmissible for the skew-Lyapunov equation  $AX - XA^T = 0$ .*

*Proof.* By Remark 4.5,  $\{\mathbf{x}^1 \otimes \mathbf{x}^1, \mathbf{x}^2 \otimes \mathbf{x}^2, \dots, \mathbf{x}^n \otimes \mathbf{x}^n\}$  is a basis for the nullspace of  $\Psi_A$  where  $\{\mathbf{x}^1, \dots, \mathbf{x}^n\}$  is a basis of eigenvectors for  $A$ . Following the nullspace

approach, we form  $N[\varepsilon]$  where  $\varepsilon = \{i + j(n - 1) \mid (i, j) \in \alpha\}$ . So,

$$N[\varepsilon] = \begin{bmatrix} x_{1j}x_{1i} & x_{2j}x_{2i} & \dots & x_{nj}x_{ni} \\ x_{1j}x_{1i} & x_{2j}x_{2i} & \dots & x_{nj}x_{ni} \end{bmatrix}$$

and we have that  $\text{rank}(N[\varepsilon]) = 1$  which is strictly less than the size of this pattern, 2. So by Theorem 4.1 the pattern  $(i, j), (j, i)$  with  $i \neq j$  is inadmissible for the matrix equation  $AX - XA^T = 0$ .  $\square$

**Corollary 5.15.** *If  $\alpha = \{(i, j), (j, i)\} \subseteq \beta$  where  $i \neq j$  then  $\beta$  is inadmissible for the matrix equation  $AX - XA^T = 0$ .*

*Proof.* This follows by Theorem 5.14 and Lemma 3.6.  $\square$

### Acknowledgments

The California State University Channel Islands Mathematics Research Experience for Undergraduates (REU) is funded through NSF grant DMS-1359165, HDR-0802628, and CI-LSAMP. Special thanks to all of the research mentors affiliated with REU for their support, guidance, and wisdom. Finally, the undergraduate authors would like to thank the individuals at their home institutions for their mentoring, support, and letters of recommendation which enabled them to participate in the REU.

### References

- [Bakonyi and Johnson 1995] M. Bakonyi and C. R. Johnson, “The Euclidean distance matrix completion problem”, *SIAM J. Matrix Anal. Appl.* **16**:2 (1995), 646–654. MR Zbl
- [DeAlba and Hogben 2000] L. M. DeAlba and L. Hogben, “Completions of P-matrix patterns”, *Linear Algebra Appl.* **319**:1–3 (2000), 83–102. MR Zbl
- [Drew et al. 2000] J. H. Drew, C. R. Johnson, S. J. Kilner, and A. M. McKay, “The cycle completable graphs for the completely positive and doubly nonnegative completion problems”, *Linear Algebra Appl.* **313**:1–3 (2000), 141–154. MR Zbl
- [Grone et al. 1984] R. Grone, C. R. Johnson, E. M. Sá, and H. Wolkowicz, “Positive definite completions of partial Hermitian matrices”, *Linear Algebra Appl.* **58** (1984), 109–124. MR Zbl
- [Hogben 1998] L. Hogben, “Completions of inverse M-matrix patterns”, *Linear Algebra Appl.* **282**:1–3 (1998), 145–160. MR Zbl
- [Hogben 2000] L. Hogben, “Inverse M-matrix completions of patterns omitting some diagonal positions”, *Linear Algebra Appl.* **313**:1–3 (2000), 173–192. MR Zbl
- [Hogben 2001] L. Hogben, “Graph theoretic methods for matrix completion problems”, *Linear Algebra Appl.* **328**:1–3 (2001), 161–202. MR Zbl
- [Horn and Johnson 1985] R. A. Horn and C. R. Johnson, *Matrix analysis*, Cambridge University Press, 1985. Reprinted in 1994. MR Zbl
- [Horn and Johnson 1991] R. A. Horn and C. R. Johnson, *Topics in matrix analysis*, Cambridge University Press, 1991. Reprinted in 1994. MR Zbl
- [Johnson and Kroschel 1996] C. R. Johnson and B. K. Kroschel, “The combinatorially symmetric P-matrix completion problem”, *Electron. J. Linear Algebra* **1** (1996), 59–63. MR Zbl



- [Johnson and Smith 1996] C. R. Johnson and R. L. Smith, “The completion problem for  $M$ -matrices and inverse  $M$ -matrices”, *Linear Algebra Appl.* **241–243** (1996), 655–667. MR Zbl
- [Johnson and Wei 2013] C. R. Johnson and Z. Wei, “Asymmetric TP and TN completion problems”, *Linear Algebra Appl.* **438**:5 (2013), 2127–2135. MR Zbl
- [Morris 2015] K. Morris, “On the rank of a Kronecker sum of similar matrices”, capstone project, Georgia College & State University, 2015.
- [Neudecker 1969] H. Neudecker, “A note on Kronecker matrix products and matrix equation systems”, *SIAM J. Appl. Math.* **17**:3 (1969), 603–606. MR Zbl
- [Rukmangadachari 2010] E. Rukmangadachari, *Mathematical methods*, Dorling Kindersley, New Delhi, 2010.

Received: 2015-11-22    Revised: 2016-06-14    Accepted: 2016-10-06

geoffrey.buhl@csuci.edu      *Department of Mathematics, CA State Univ Channel Islands,  
1 University Dr., Camarillo, CA 93012, United States*

ecronk1@ithaca.edu      *Department of Mathematics, Ithaca College, 953 Danby Rd.,  
Ithaca, NY 14850, United States*

rosa.moreno544@myci.csuci.edu      *Department of Mathematics,  
California State University Channel Islands, 1 University Dr.,  
Camarillo, CA 93012, United States*

kirsten.morris25@uga.edu      *Department of Mathematics, The University of Georgia,  
University of Georgia, Athens, GA 30602, United States*

pedrozad@ripon.edu      *Department of Mathematics, Ripon College, 300 Seward St.,  
Ripon, WI 54971, United States*

jryan23@vols.utk.edu      *Department of Mathematics, The University of Tennessee  
Knoxville, 1403 Circle Dr., Knoxville, TN 37996, United States*



# The Hamiltonian problem and $t$ -path traceable graphs

Kashif Bari and Michael E. O’Sullivan

(Communicated by Ronald Gould)

The problem of characterizing maximal non-Hamiltonian graphs may be naturally extended to characterizing graphs that are maximal with respect to nontraceability and beyond that to  $t$ -path traceability. We show how  $t$ -path traceability behaves with respect to disjoint union of graphs and the join with a complete graph. Our main result is a decomposition theorem that reduces the problem of characterizing maximal  $t$ -path traceable graphs to characterizing those that have no universal vertex. We generalize a construction of maximal nontraceable graphs by Zelinka to  $t$ -path traceable graphs.

## 1. Introduction

The motivating problem for this article is the characterization of maximal non-Hamiltonian (MNH) graphs. The first broad family of MNH graphs was given in [Skupień 1979], and all MNH graphs with ten or fewer vertices were described in [Jamrozik et al. 1982], a paper where Skupień and his coauthors gave three constructions, called types  $A1$ ,  $A2$ ,  $A3$ , with a similar structure. Zelinka [1998] gave two constructions of graphs that are maximal nontraceable; that is, they have no Hamiltonian path, but the addition of any edge gives a Hamiltonian path. The join of such a graph with a single vertex gives an MNH graph. Zelinka’s first family produces, under the join with  $K_1$ , the original MNH graphs of Skupień. Zelinka’s second family is a broad generalization of the type  $A1$ ,  $A2$ , and  $A3$  graphs of [Jamrozik et al. 1982]. Further examples of infinite families of maximal nontraceable graphs appeared in [Bullock et al. 2008].

In this article, we work with two closely related invariants of a graph  $G$ ,  $\check{\mu}(G)$  and  $\mu(G)$ . The  $\mu$ -invariant, introduced by Ore [1961] and also used by Noorvash [1975], is the minimal number of paths in  $G$  required to cover the vertex set of  $G$ . We define  $\check{\mu}(G)$  to be the smallest integer  $\ell$  such that the join of  $K_\ell$  with  $G$  is Hamiltonian. We show that  $\check{\mu}(G) = \mu(G)$  unless  $G$  is Hamiltonian, when  $\check{\mu}(G) = 0$ . Maximal

---

*MSC2010:* 05C45.

*Keywords:* maximal non-hamiltonian, hamiltonian, graph theory,  $t$ -path traceable.

non-Hamiltonian graphs are maximal with respect to  $\check{\mu}(G) = 1$ , and maximal nontraceable graphs are maximal with respect to  $\check{\mu}(G) = 2$ . It is useful to broaden the perspective to study, for arbitrary  $t$ , graphs that are maximal with respect to  $\check{\mu}(G) = t$ , which we call  $t$ -path traceable graphs.

In Section 2 we show how the  $\check{\mu}$  and  $\mu$  invariants behave with respect to disjoint union of graphs and the join with a complete graph. Section 3 derives the main result, a decomposition theorem that reduces the problem of characterizing maximal  $t$ -path traceable graphs to characterizing those that have no universal vertex, which we call *trim*. Section 4 presents a generalization of the Zelinka construction to  $t$ -path traceable graphs.

## 2. Traceability and Hamiltonicity

It will be notationally convenient to say that the complete graphs  $K_1$  and  $K_2$  are Hamiltonian. As justification for this view, consider an undirected graph as a directed graph with each edge having a conjugate edge in the reverse direction. This perspective does not affect the Hamiltonicity of a graph with more than three vertices, but it does give  $K_2$  a Hamiltonian cycle. Similarly, adding loops to any graph with more than two vertices does not alter the Hamiltonicity of the graph, but  $K_1$ , with an added loop, has a Hamiltonian cycle.

Let  $G$  be a graph. A vertex,  $v \in V(G)$ , is called a *universal vertex* if  $\deg(v) = |V(G)| - 1$ . A universal vertex is also known as a dominating vertex. Let  $\bar{G}$  denote the *graph complement* of  $G$ , having vertex set  $V(G)$  and edge set  $E(K_n) \setminus E(G)$ . We will use the disjoint union of two graphs,  $G \sqcup H$  and the join of two graphs  $G * H$ . The latter is  $G \sqcup H$  together with the edges  $\{vw \mid v \in V(G) \text{ and } w \in V(H)\}$ .

**Definition 1.** A set of  $s$  disjoint paths in a graph  $G$  that includes every vertex in  $G$  is an  $s$ -*path covering* of  $G$ . We define the following invariants:

$$\begin{aligned}\mu(G) &:= \min\{s \in \mathbb{N} \mid \text{there exists an } s\text{-path covering of } G\}, \\ \check{\mu}(G) &:= \min\{l \in \mathbb{N}_0 \mid K_l * G \text{ is Hamiltonian}\}, \\ i_H(G) &:= \begin{cases} 1 & \text{if } G \text{ is Hamiltonian,} \\ 0 & \text{otherwise.} \end{cases}\end{aligned}$$

We will say  $G$  is  $t$ -*path traceable* when  $\mu(G) = t$ . A set of  $t$  disjoint paths that covers a  $t$ -path traceable graph  $G$  is a *minimal path covering*.

Note that  $K_r * (K_s * G) = K_{r+s} * G$ . If  $G$  is Hamiltonian then so is  $K_r * G$  for  $r \geq 0$  (in particular, this is true for  $G = K_1$  and  $G = K_2$ ).

We now present a series of lemmas that leads to the main result of this section, which is a formula showing how the  $\mu$ -invariant and  $\check{\mu}$ -invariant behave with respect to the disjoint union and the join with a complete graph.

**Lemma 2.**  $\check{\mu}(G) = \min\{l \in \mathbb{N}_0 \mid \bar{K}_l * G \text{ is Hamiltonian}\}.$

*Proof.* Since  $\bar{K}_l * G$  is a subgraph of  $K_l * G$ , a Hamiltonian cycle in  $\bar{K}_l * G$  would also be one in  $K_l * G$ .

Let  $\check{\mu}(G) = a$ . Suppose  $C$  is a Hamiltonian cycle in  $K_a * G$  and write  $C$  as  $v \sim P_1 \sim Q_1 \sim \dots \sim P_s \sim Q_s \sim v$ , where  $v$  is a vertex in  $G$  and the paths  $P_i$  in  $G$  and  $Q_i$  in  $K_a$ . If any  $Q_i$  contains two vertices or more, say  $u$  and  $w_1, \dots, w_k$  with  $k \geq 1$ , then we may simply remove all the vertices, except  $u$ , and end up with a Hamiltonian graph on  $K_{a-k}$ . This contradicts the minimality of  $a = \check{\mu}(G)$ . Therefore,  $C$  must not contain any paths of length greater than two in the subgraph  $K_a$ , and any Hamiltonian cycle on  $K_a * G$  is also a Hamiltonian cycle on  $\bar{K}_a * G$ .  $\square$

**Lemma 3.**  $\check{\mu}(G) = \mu(G) - i_H(G).$

*Proof.* If  $G$  is Hamiltonian (including  $K_1$  and  $K_2$ ) then  $\check{\mu}(G) = 0, \mu(G) = 1$  so the equality holds. Suppose  $G$  is non-Hamiltonian with  $\mu(G) = t$  and  $t$ -path covering  $P_1, \dots, P_t$ . Let  $K_t$  have vertices  $u_1, \dots, u_t$ . In the graph  $K_t * G$ , there is a Hamiltonian cycle:  $v_1 \sim P_1 \sim v_2 \sim P_2 \sim \dots \sim v_t \sim P_t \sim v_1$ . Thus  $\check{\mu}(G) \leq t = \mu(G)$ .

Let  $\check{\mu}(G) = a$ , so there is a Hamiltonian cycle in  $K_a * G$ . Removing the vertices of  $K_a$  breaks the cycle into at most  $a$  disjoint paths covering  $G$ . Thus  $\mu(G) \leq \check{\mu}(G)$ .  $\square$

**Lemma 4.**  $\mu(G \sqcup H) = \mu(G) + \mu(H)$  and  $\check{\mu}(G \sqcup H) = \check{\mu}(G) + \check{\mu}(H) + i_H(G) + i_H(H).$

*Proof.* A path covering of  $G$  may be combined with a path covering of  $H$  to create one for  $G \sqcup H$  so  $\mu(G \sqcup H) \leq \mu(G) + \mu(H)$ . Conversely, paths in a  $t$ -path covering of  $G \sqcup H$  can be partitioned into those contained in  $G$  and those contained in  $H$ , giving a path covering of  $G$  and one of  $H$ . Consequently,  $\mu(G \sqcup H) \geq \mu(G) + \mu(H)$ .

Since  $G \sqcup H$  is not Hamiltonian we have

$$\begin{aligned} \check{\mu}(G \sqcup H) &= \mu(G \sqcup H) + i_H(G \sqcup H) \\ &= \mu(G) + \mu(H) \\ &= \check{\mu}(G) + i_H(G) + \check{\mu}(H) + i_H(H). \end{aligned} \quad \square$$

**Lemma 5.** For any graph  $G$ ,

$$\begin{aligned} \mu(K_s * G) &= \max\{1, \mu(G) - s\}, \\ \check{\mu}(K_s * G) &= \max\{0, \check{\mu}(G) - s\}. \end{aligned}$$

In particular, if  $K_s * G$  is Hamiltonian then  $\mu(K_s * G) = 1$  and  $\check{\mu}(K_s * G) = 0$ ; otherwise,  $\mu(K_s * G) = \mu(G) - s$  and  $\check{\mu}(K_s * G) = \check{\mu}(G) - s$ .

*Proof.* The formula for  $\check{\mu}$  is immediate when  $G$  is Hamiltonian since we have observed that this forces  $K_s * G$  to be Hamiltonian. Otherwise, it follows from

$K_r * (K_s * G) = K_{r+s} * G$ : if  $\check{\mu}(G) = a$ , then  $K_r * (K_s * G)$  is Hamiltonian if and only if  $r + s \geq a$ .

The formula for  $\mu$  may be derived from the result for  $\check{\mu}$  using Lemma 3. □

The main result of this section is the following two formulas for the  $\mu$  and  $\check{\mu}$  invariants of the disjoint union of graphs, and the join with a complete graph.

**Proposition 6.** *Let  $\{G_j\}_{j=1}^m$  be graphs. Then*

$$\begin{aligned} \mu\left(\bigsqcup_{j=1}^m G_j\right) &= \sum_{j=1}^m \mu(G_j), \\ \check{\mu}\left(\bigsqcup_{j=1}^m G_j\right) &= \sum_{j=1}^m \check{\mu}(G_j) + \sum_{j=1}^m i_H(G_j). \end{aligned}$$

Furthermore,

$$\check{\mu}\left(\left(\bigsqcup_{j=1}^m G_j\right) * K_r\right) = \max\left\{0, \sum_{j=1}^m \check{\mu}(G_j) + \sum_{j=1}^m i_H(G_j) - r\right\}.$$

*Proof.* We proceed by induction. The base case  $k = 2$  is exactly Lemma 4. Assume the formula holds for  $k$  graphs; we will prove it for  $k + 1$  graphs.

$$\begin{aligned} \mu\left(\bigsqcup_{j=1}^{k+1} G_j\right) &= \mu\left(\left(\bigsqcup_{j=1}^k G_j\right) \sqcup G_{k+1}\right) = \mu\left(\bigsqcup_{j=1}^k G_j\right) + \mu(G_{k+1}) \\ &= \sum_{j=1}^k \mu(G_j) + \mu(G_{k+1}) = \sum_{j=1}^{k+1} \mu(G_j). \end{aligned}$$

By Lemma 3 and the fact that disjoint graphs are not Hamiltonian, we have

$$\begin{aligned} \check{\mu}\left(\bigsqcup_{j=1}^m G_j\right) &= \mu\left(\bigsqcup_{j=1}^m G_j\right) + i_H\left(\bigsqcup_{j=1}^m G_j\right) \\ &= \sum_{j=1}^m (\check{\mu}(G_j) + i_H(G_j)) = \sum_{j=1}^m \check{\mu}(G_j) + \sum_{j=1}^m i_H(G_j). \end{aligned}$$

Therefore, we have by Lemma 5,

$$\begin{aligned} \check{\mu}\left(\left(\bigsqcup_{j=1}^m G_j\right) * K_r\right) &= \max\left\{0, \check{\mu}\left(\bigsqcup_{j=1}^m G_j\right) - r\right\} \\ &= \max\left\{0, \sum_{j=1}^m \check{\mu}(G_j) + \sum_{j=1}^m i_H(G_j) - r\right\}. \end{aligned} \quad \square$$

The following lemma will be useful in the next section. To express it succinctly, we introduce the following Boolean condition. For a graph  $G$  and vertex  $v \in V(G)$ ,  $T(v, G)$  is true if and only if  $v$  is a terminal vertex in some minimal path covering of  $G$ .

**Lemma 7.** *Let  $v \in V(G)$  and  $w \in V(H)$ . Then we have*

$$\mu((G \sqcup H) + vw) = \begin{cases} \mu(G \sqcup H) - 1 & \text{if } T(v, G) \text{ and } T(w, H), \\ \mu(G \sqcup H) & \text{otherwise.} \end{cases}$$

*Proof.* Let  $\mu(G) = c$ ,  $\mu(H) = d$  and  $\mu((G \sqcup H) + vw) = t$ . Clearly,  $t \leq c + d$ .

Let  $R_1, \dots, R_t$  be a minimal path cover of  $(G \sqcup H) + vw$ . If no  $R_i$  contains  $vw$  then this is also a minimal path cover of  $(G \sqcup H)$  so  $t = c + d$ . Suppose  $R_1$  contains  $vw$  and note that  $R_1$  is the only path with vertices in both  $G$  and  $H$ . Removing  $vw$  gives two paths  $P \subseteq G$  and  $Q \subseteq H$ . Paths  $P$  and  $Q$  along with  $R_2, \dots, R_t$  cover  $G \sqcup H$ , so  $t + 1 \geq c + d$ . Thus,  $t$  can either be  $c + d$  or  $c + d - 1$ .

If  $t = c + d - 1$ , then we have the minimal  $(t + 1)$ -path covering  $P, Q, R_2, \dots, R_t$  of  $G \sqcup H$ , as above. We note that  $v$  must be a terminal point of  $P$  and  $w$  must be a terminal point of  $Q$ , by construction. This path covering may be partitioned into a  $c$ -path covering of  $G$  containing  $P$  and a  $d$ -path covering of  $H$  containing  $Q$ . Thus,  $T(v, G)$  and  $T(w, G)$  hold.

Conversely, suppose  $T(u, G)$  and  $T(w, H)$  both hold. Let  $P_1, \dots, P_c$  be a minimal path of  $G$  with  $v$  a terminal vertex of  $P_1$  and let  $Q_1, \dots, Q_d$  be a minimal path cover of  $H$  with  $w$  a terminal vertex of  $Q_1$ . The edge  $vw$  knits  $P_1$  and  $Q_1$  into a single path and  $P_1 \sim Q_1, P_1, \dots, P_c, Q_1, \dots, Q_d$  is a  $c + d - 1$  cover of  $(G \sqcup H) + vw$ . Consequently,  $t \leq c + d - 1$ .

Thus,  $T(u, G)$  and  $T(w, H)$  both hold if and only if  $t = c + d - 1$ . Otherwise,  $t = c + d$ . □

**Corollary 8.** *Let  $v \in V(G)$  and  $w \in V(H)$ . Then we have*

$$\check{\mu}((G \sqcup H) + vw) = \begin{cases} \check{\mu}(G \sqcup H) - 2 & \text{if } G = H = K_1, \\ \check{\mu}(G \sqcup H) - 1 & \text{if } T(v, G) \text{ and } T(w, H), \\ \check{\mu}(G \sqcup H) & \text{otherwise.} \end{cases}$$

*Proof.* Let  $\delta = 1$  if  $T(v, G)$  and  $T(w, H)$  are both true and  $\delta = 0$  otherwise. Then

$$\begin{aligned} \check{\mu}((G \sqcup H) + vw) &= \mu((G \sqcup H) + vw) - i_H((G \sqcup H) + vw) \\ &= \mu((G \sqcup H) - \delta - i_H((G \sqcup H) + vw)). \end{aligned}$$

The final term is  $-1$  if and only if  $G = H = K_1$ . □

### 3. Decomposing maximal $t$ -path traceable graphs

In this section we prove our main result, a maximal  $t$ -path traceable graph may be uniquely written as the join of a complete graph and a disjoint union of graphs that are also maximal with respect to traceability, but which are also either complete or have no universal vertex. We work with the families of graphs  $\mathcal{M}_t$  for  $t \geq 0$  and  $\mathcal{N}_t$  for  $t \geq 1$ :

$$\begin{aligned} \mathcal{M}_t &:= \{G \mid \check{\mu}(G) = t \text{ and } \check{\mu}(G + e) < t, \forall e \in E(\overline{G})\}, \\ \mathcal{N}_t &:= \{G \in \mathcal{M}_t \mid G \text{ is connected and has no universal vertex}\}. \end{aligned}$$

The set  $\mathcal{M}_0$  is the set of complete graphs. The set  $\mathcal{M}_1$  is the set of graphs with a Hamiltonian path but no Hamiltonian cycle, that is, maximal non-Hamiltonian graphs. For  $t > 1$ ,  $\mathcal{M}_t$  is also the set of graphs  $G$  such  $\mu(G) = t$  and  $\mu(G + e) = t - 1$  for any  $e \in E(\overline{G})$ . We will call these *maximal  $t$ -path traceable graphs*. A graph in  $\mathcal{N}_t$  will be called *trim*.

**Proposition 9.** *For  $0 \leq r < t$ ,  $G \in \mathcal{M}_t$  if and only if  $K_r * G \in \mathcal{M}_{t-r}$ .*

*Proof.* We have  $\check{\mu}(K_r * G) = \check{\mu}(G) - r$ , by Lemma 5, so we just need to show that  $K_r * G$  is maximal if and only if  $G$  is maximal. The only edges that can be added to  $K_r * G$  are those between vertices of  $G$ , that is,  $E(\overline{K_r * G}) = E(\overline{G})$ . For such an edge  $e$ ,

$$\check{\mu}((K_r * G) + e) = \check{\mu}(K_r * (G + e)) = \check{\mu}(G + e) - r. \tag{1}$$

Thus,  $\check{\mu}(G + e) = \check{\mu}(G) - 1$  if and only if  $\check{\mu}((K_r * G) + e) = \check{\mu}(K_r * G) - 1$ .  $\square$

Note that the proposition is false for  $r = t > 0$  since  $K_r * G$  will not be a complete graph and  $\mathcal{M}_0$  is the set of complete graphs. The proof breaks down in (1).

As a key step before the main theorem, the next lemma shows that in a maximal graph, each vertex is either universal or it is a terminal vertex in a minimal path covering (but not both).

**Lemma 10.** *Let  $c \geq 1$  and  $G \in \mathcal{M}_c$ . For any two nonadjacent vertices  $v, w$  in  $G$ , there is a  $c$ -path covering of  $G$  in which both  $v$  and  $w$  are terminal points of paths. Moreover, a vertex  $v \in V(G)$  is a terminal point in some  $c$ -path covering if and only if  $v$  is not universal.*

*Proof.* Suppose  $c > 1$  and let  $v, w$  be nonadjacent in  $G$ . Since  $G$  is maximal,  $G + vw$  has a  $(c - 1)$ -path covering,  $P_1, \dots, P_{c-1}$ . The edge  $vw$  must be contained in some  $P_i$  because  $G$  has no  $(c - 1)$ -path covering. Removing that edge gives a  $c$ -path covering of  $G$  with  $v$  and  $w$  as terminal vertices. The special case  $c = 1$  is well known, adding the edge  $vw$  gives a Hamiltonian cycle, and removing it leaves a path with endpoints  $v$  and  $w$ . A consequence is that any nonuniversal vertex is the terminal point of some path in a  $c$ -path covering.



Suppose  $P_1, \dots, P_c$  is a  $c$ -path covering of  $G \in \mathcal{M}_c$  with  $v$  a terminal point of  $P_i$ . Then  $v$  is not adjacent to any of the terminal points of  $P_j$  for  $j \neq i$ , for otherwise two paths could be combined into a single one. In the case  $c = 1$ ,  $v$  cannot be adjacent to the other terminal point of  $P_1$ , otherwise  $G$  would have a Hamiltonian cycle. Consequently, a universal vertex is not a terminal point in a  $c$ -path covering of  $G$ .  $\square$

**Proposition 11.** *Let  $G \in \mathcal{M}_c$  and  $H \in \mathcal{M}_d$ . The following are equivalent:*

- (1)  $G \sqcup H \in \mathcal{M}_{c+d+i_H(G)+i_H(H)}$ .
- (2) *Each of  $G$  and  $H$  is either complete or has no universal vertex.*

*Proof.* We have already shown that  $\check{\mu}(G \sqcup H) = c + d + i_H(G) + i_H(H)$ . We have to consider whether adding an edge to  $G \sqcup H$  reduces the  $\check{\mu}$ -invariant. There are three cases to consider: the extra edge may be in  $E(\bar{G})$  or  $E(\bar{H})$  or it may join a vertex in  $G$  to one in  $H$ . Since  $G$  is maximal, adding an edge to  $G$  is either impossible, when  $G$  is complete, or it reduces the  $\check{\mu}$ -invariant of  $G$ . This edge would also reduce the  $\check{\mu}$ -invariant of  $G \sqcup H$  by Lemma 4. The case for adding an edge of  $H$  is the same. Consider the edge  $vw$  for  $v \in V(G)$  and  $w \in V(H)$ . By Corollary 8 the  $\check{\mu}$ -invariant will drop if and only if  $v$  is the terminal point of a path in a minimal path covering of  $G$  and similarly for  $w$  in  $H$ , that is,  $T(v, G)$  and  $T(w, H)$ . Clearly this holds for all vertices in a complete graph. Lemma 10 shows that  $T(v, G)$  holds for  $G \in \mathcal{M}_c$  with  $c > 0$  if and only if  $v$  is not a universal vertex in  $G$ . Thus, in order for  $G \sqcup H$  to be maximal,  $G$  must either be complete or be maximal itself and have no universal vertex, and similarly for  $H$ .  $\square$

**Theorem 12.** *For any  $G \in \mathcal{M}_t$ ,  $t > 0$ ,  $G$  may be uniquely decomposed as*

$$K_r * (G_1 \sqcup \dots \sqcup G_m),$$

where  $r$  is the number of universal vertices of  $G$ , and each  $G_j$  is either complete or  $G_j \in \mathcal{N}_{t_j}$  for some  $t_j > 0$ . Furthermore  $t = \sum_{j=1}^m t_j + \sum_{j=1}^m i_H(G_j) - r$ .

*Proof.* Suppose  $G \in \mathcal{M}_t$  and let  $r$  be the number of universal vertices of  $G$ . Let  $m$  be the number of components in the graph obtained by removing the universal vertices from  $G$ , let  $G_1, \dots, G_m$  be the components and let  $\check{\mu}(G_j) = t_j$ . Then  $G = K_r * (G_1 \sqcup \dots \sqcup G_m)$ .

Proposition 6 shows that  $t = \sum_{j=1}^m t_j + \sum_{j=1}^m i_H(G_j) - r$ . By Proposition 9, we have that  $G \in \mathcal{M}_t$  if and only if  $G_1 \sqcup \dots \sqcup G_m \in \mathcal{M}_{t+r}$ . Each  $G_i$  must be maximal, otherwise the disjoint union would not be maximal (add an appropriate edge to a  $G_i$  in Proposition 6). Inductively applying Proposition 11 to  $G_1 \sqcup \dots \sqcup G_m \in \mathcal{M}_{t+r}$ , where  $t + r = \sum_{j=1}^m t_j + \sum_{j=1}^m i_H(G_j)$ , we have that each  $G_j$  is complete or is trim ( $G_j \in \mathcal{N}_{t_j}$  for  $t_j > 0$ ).  $\square$

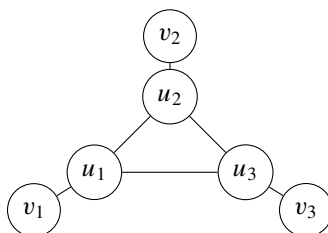
#### 4. Trim maximal $t$ -path traceable graphs

Skupień [1979] discovered the first family of maximal non-Hamiltonian graphs, that is, graphs in  $\mathcal{M}_1$ . These graphs are formed by taking the join of  $K_r$  with the disjoint union of  $r + 1$  complete graphs [Marczyk and Skupień 1991]. The smallest graph in  $\mathcal{N}_2$  is shown in Figure 1. Chvátal [1973] identified its join with  $K_1$  as the smallest maximal non-Hamiltonian graph that is not 1-tough, that is, not one of the Skupień family. Jamrozik, Kalinowski and Skupień [1982] generalized this example to three different families. Family A1 replaces each edge  $u_i v_i$  in Figure 1 with an arbitrary complete graph containing  $u_i$  and replaces the  $K_3$  formed by the  $u_i$  with an arbitrary complete graph. The result — a type A1 graph — has four cliques, the first three disjoint from each other but each intersecting the fourth clique in a single vertex. An A1 graph is in  $\mathcal{N}_2$  and its join with  $K_1$  gives a maximal non-Hamiltonian graph. Family A2 is formed by taking the join with  $K_2$  of the disjoint union of a complete graph and an A1 graph. Theorem 12 shows that the resulting graph is in  $\mathcal{M}_1$ . Family A3 is a modification of the A1 family based on the graph in Figure 2, which is in  $\mathcal{N}_2$ .

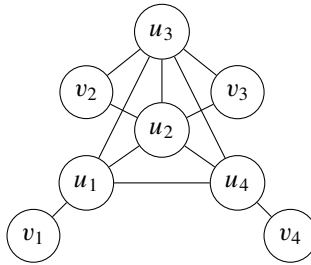
More than two decades later, Bullock, Frick, Singleton and van Aardt [2008] recognized that two constructions of Zelinka [1998] give maximal nontraceable graphs, that is, elements of  $\mathcal{M}_2$ . Zelinka's first construction is like the Skupień family: formed from  $r + 1$  complete graphs followed by the join with  $K_{r-1}$ . The Zelinka type II family contains graphs in  $\mathcal{N}_2$  that are a significant generalization of the graphs in Figures 1 and 2. In this section we generalize this family further to get graphs in  $\mathcal{N}_t$  for arbitrary  $t$ . Our starting point is the graph in Figure 3, which is in  $\mathcal{N}_3$ .

**Example 13.** Consider  $K_m$  with  $m = 2t - 1$  and vertices  $u_1, \dots, u_m$ . Let  $G$  be the graph containing  $K_m$  along with vertices  $v_1, \dots, v_{2t-1}$  and edges  $u_i v_i$ . The case with  $t = 3$  and  $m = 5 = 2t - 1$  is Figure 3. We claim  $G \in \mathcal{N}_t$ .

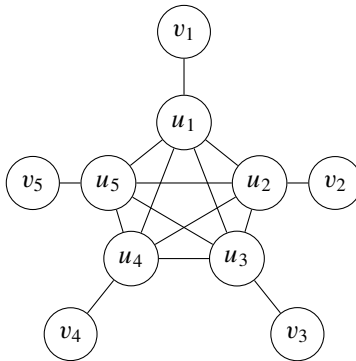
One can readily check that this graph is  $t$ -path covered using  $v_{2i-1} \sim u_{2i-1} \sim u_{2i} \sim v_{2i}$  for  $i = 1, \dots, t - 1$  and  $v_{2t-1} \sim u_{2t-1} \sim u_{2t} \sim \dots \sim u_m$ . We check that  $G$  is maximal. By the symmetry of the graph, we need only consider the addition



**Figure 1.** Smallest graph in  $\mathcal{N}_2$ .



**Figure 2.** The join of this graph with  $K_1$  is the smallest graph in the  $A_3$  family.



**Figure 3.** Whirligig in  $\mathcal{N}_3$ .

of the edge  $v_1u_m$  or  $v_1u_2$  or  $v_1v_2$ . In each case, the last and the first paths listed above may be combined into one, either

$$\begin{aligned}
 &v_{2t-1} \sim u_{2t-1} \sim \dots \sim u_m \sim v_1 \sim u_1 \sim u_2 \sim v_2, \text{ or} \\
 &v_{2t-1} \sim u_{2t-1} \sim \dots \sim u_m \sim u_1 \sim v_1 \sim u_2 \sim v_2, \text{ or} \\
 &v_{2t-1} \sim u_{2t-1} \sim \dots \sim u_m \sim u_1 \sim v_1 \sim v_2 \sim u_2.
 \end{aligned}$$

Thus, adding an edge creates a  $(t - 1)$ -path covered graph, proving maximality.

The next proposition shows that the previous example is the only way to have a trim maximal  $t$ -path covered graph with  $2t - 1$  degree-one vertices. We start with a technical lemma.

**Lemma 14.** *Let  $G$  be a connected graph and  $u_1, v_1, v_2, v_3 \in V(G)$  with  $\deg(v_i) = 1$ , and  $u$  adjacent to  $v_1$  and  $v_2$  but not  $v_3$ . Then  $\mu(G) = \mu(G + uv_3)$ .*

*Proof.* Let  $P_1, \dots, P_r$  be a minimal path covering of  $G + uv_3$ ; it is enough to show that there are  $r$ -paths covering  $G$ . If the covering doesn't include  $uv_3$ , then  $P_1, \dots, P_r$  also give a minimal path covering of  $G$ , establishing the claim of the lemma. Otherwise, suppose  $uv_3$  is an edge of  $P_1$ . We consider two cases.

Suppose  $P_1$  contains the edge  $uv_1$  (or similarly  $uv_2$ ). Then  $P_1$  has  $v_1$  as a terminal point and one of the other paths, say  $P_2$ , must be a length-0 path containing simply  $v_2$ . Let  $Q$  be obtained by removing  $uv_1$  and  $uv_3$  from  $P_1$ . Then  $v_1 \sim u \sim v_2, Q, P_3, \dots, P_r$ , gives an  $r$ -path covering of  $G$ .

Suppose  $P_1$  contains neither  $uv_1$  nor  $uv_2$ . Then each of  $v_1$  and  $v_2$  must be on a length-0 path in the covering, say  $P_2$  and  $P_3$  are these paths. Furthermore  $u$  must not be a terminal point of  $P_1$ ; if it were, the path could be extended to include  $v_1$  or  $v_2$ , reducing the number of paths required to cover  $G$ . Removing  $u$  from  $P_1$  yields two paths,  $Q_1, Q_2$ . Then  $v_1 \sim u \sim v_2, Q_1, Q_2, P_4, \dots, P_r$  gives an  $r$ -path cover of  $G$ . This proves the lemma.  $\square$

**Proposition 15.** *Let  $G \in \mathcal{N}_t$ . The number of degree-one vertices in  $G$  is at most  $2t - 1$ . This occurs if and only if the  $2t - 1$  vertices of degree-one have distinct neighbors and removing the degree-one vertices leaves a complete graph.*

*Proof.* Each degree-one vertex must be a terminal point in a path covering. So any graph  $G$  covered by  $t$  paths can have at most  $2t$  degree-one vertices. Aside from the case  $t = 1$  and  $G = K_2$ , we can see that a graph with  $2t$  degree-one vertices cannot be maximal  $t$ -path traceable as follows. It is easy to check that a  $2t$  star is not  $t$ -path traceable (it is also not trim). A  $t$ -path traceable graph with  $2t$  degree-one vertices must therefore have an interior vertex  $w$  that is not connected to at least one of the degree-one vertices  $v$ . Such a graph is not maximal because the edge  $vw$  can be added leaving  $2t - 1$  degree-one vertices. This resulting graph cannot be  $(t - 1)$ -path covered.

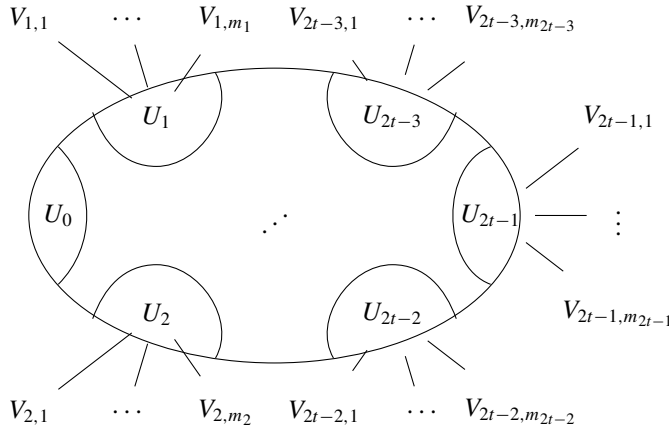
Suppose that  $G \in \mathcal{N}_t$  with  $2t - 1$  degree-one vertices,  $v_1, \dots, v_{2t-1}$ . Lemma 14 shows that no two of the  $v_i$  can be adjacent to the same vertex, for that would violate maximality of  $G$ . So, the  $v_i$  have distinct neighbors. Furthermore, all the vertices except the  $v_i$  can be connected to each other and a path covering will still require at least  $t$  paths since there remain  $2t - 1$  degree-one vertices. This proves the necessity of the structure claimed in the proposition. The previous example showed that the graph is indeed in  $\mathcal{N}_t$ .  $\square$

We can now generalize the Zelinka family.

**Construction 16.** Let  $U_0, U_1, \dots, U_{2t-1}$  be disjoint sets of vertices and

$$U = \bigsqcup_{i=0}^{2t-1} U_i.$$

Let  $m_i = |U_i|$  and assume that for  $i > 0$  the  $U_i$  are nonempty, so  $m_i > 0$ . For  $i = 1, \dots, 2t - 1$  (but not  $i = 0$ ) and  $j = 1, \dots, m_i$ , let  $V_{ij}$  be nonempty sets of vertices disjoint from each other and from  $U$ . Form the graph  $W$  with vertex set  $U \sqcup \left( \bigsqcup_{i=1}^{2t-1} \left( \bigsqcup_{j=1}^{m_i} V_{ij} \right) \right)$  and edges  $uu'$  for  $u, u' \in U$  and  $uv$  for any  $u \in U_i$



**Figure 4.** Generalization of the whirligig,  $W$ .

and  $v \in V_{ij}$  with  $i = 1, \dots, 2t - 1$  and  $j = 1, \dots, m_i$  and all edges within each set  $V_{ij}$ . The cliques of this graph are  $K_U$  and  $K_{U_i \sqcup V_{ij}}$  for each  $i = 1, \dots, 2t - 1$  and  $j = 1, \dots, m_i$ .

The graph in Figure 2 has  $m_0 = 0$ ,  $m_1 = m_2 = 1$  and  $m_3 = 2$ , and the graph in Figure 4 indicates the general construction.

**Theorem 17.** *The graph  $W$  in Construction 16 is a trim, maximal  $t$ -path traceable graph.*

*Proof.* We must show that  $W$  is  $t$ -path covered and not  $(t - 1)$ -path covered, and that the addition of any edge yields a  $(t - 1)$ -path covered graph. The argument is analogous to the one in Example 13.

Let  $R$  be a Hamiltonian path in  $U_0$ . For each  $i = 1, \dots, 2t - 1$  and  $j = 1, \dots, m_i$ , let  $Q_{ij}$  be a Hamiltonian path in  $K_{V_{ij}}$ . Let  $P_i$  be the path

$$P_i : Q_{i1} \sim u_{i1} \sim Q_{i2} \sim u_{i2} \sim \dots \sim Q_{im_i} \sim u_{im_i},$$

and let  $\overleftarrow{P}_i$  be the reversal of  $P_i$ .

Since there is an edge  $u_{im_i}u_{jm_j}$  there is a path  $P_i \sim \overleftarrow{P}_j$  for any  $i \neq j \in \{1, \dots, 2t - 1\}$ . Therefore the graph  $W$  has a  $t$ -path covering  $P_{2i-1} \sim \overleftarrow{P}_{2i}$  for  $i = 1, \dots, (t - 1)$ , along with  $P_{2t-1} \sim R$ . We leave to the reader the argument that there is no  $(t - 1)$ -path cover.

To show  $W$  is maximal we show that after adding an edge  $e$ , we can join two paths in the  $t$ -path cover above, with a bit of rearrangement. There are three types of edges to consider, the edge  $e$  might join  $V_{ij}$  to  $U_{i'}$  for  $i \neq i'$ ; or  $V_{ij}$  to  $V_{i'j'}$  for  $j \neq j'$ ; or  $V_{ij}$  to  $V_{i'j'}$  for  $i \neq i'$ . Because of the symmetry of  $W$ , we may assume

$i = 1$  and  $j = 1$  and that the vertex chosen from  $V_{ij} = V_{1,1}$  is the initial vertex of  $Q_{1,1}$ . Other simplifications due to symmetry will be evident in what follows.

In the first case there are two subcases — determined by  $i' \geq 2t$  or not — and after permutation, we may consider the edge  $e$  from the initial vertex of  $Q_{1,1}$  to the terminal vertex of  $R$ , or to the terminal vertex of  $P_{2t-1}$ . We can then join two paths in the  $t$ -path cover: either  $P_{2t-1} \sim R \stackrel{e}{\sim} P_1 \sim \overleftarrow{P}_2$  or  $P_{2t-1} \stackrel{e}{\sim} P_1 \sim R \sim \overleftarrow{P}_2$ .

Suppose next that we join the initial vertex of  $Q_{11}$  with the terminal vertex of  $Q_{12}$ . We then rearrange  $P_1$  and join two paths in the  $t$ -path cover to get

$$P_{2t-1} \sim R \sim u_{1,1} \sim Q_{1,1} \stackrel{e}{\sim} Q_{1,2} \sim u_{1,2} \sim \cdots \sim Q_{1m_1} \sim u_{1m_1} \sim \overleftarrow{P}_2.$$

Finally, suppose that we join the initial vertex of  $Q_{1,1}$  with the initial vertex of  $Q_{2t-1,1}$ . Then we rearrange to

$$\overleftarrow{R} \sim \overleftarrow{P}_{2t-1} \stackrel{e}{\sim} P_1 \sim \overleftarrow{P}_2. \quad \square$$

## References

- [Bullock et al. 2008] F. Bullock, M. Frick, J. Singleton, S. van Aardt, and K. Mynhardt, “Maximal nontraceable graphs with toughness less than one”, *Electron. J. Combin.* **15**:1 (2008), #R18, 19 pp. MR Zbl
- [Chvátal 1973] V. Chvátal, “Tough graphs and Hamiltonian circuits”, *Discrete Math.* **5** (1973), 215–228. MR Zbl
- [Jamrozik et al. 1982] J. Jamrozik, R. Kalinowski, and Z. Skupień, “A catalogue of small maximal non-Hamiltonian graphs”, *Discrete Math.* **39**:2 (1982), 229–234. MR Zbl
- [Marczyk and Skupień 1991] A. Marczyk and Z. Skupień, “Maximum non-Hamiltonian tough graphs”, *Discrete Math.* **96**:3 (1991), 213–220. MR Zbl
- [Noorvash 1975] S. Noorvash, “Covering the vertices of a graph by vertex-disjoint paths”, *Pacific J. Math.* **58**:1 (1975), 159–168. MR Zbl
- [Ore 1961] O. Ore, “Arc coverings of graphs”, *Ann. Mat. Pura Appl.* (4) **55** (1961), 315–321. MR Zbl
- [Skupień 1979] Z. Skupień, “On maximal non-Hamiltonian graphs”, *Rostock. Math. Kolloq.* **11** (1979), 97–106. MR Zbl
- [Zelinka 1998] B. Zelinka, “Graphs maximal with respect to absence of Hamiltonian paths”, *Discuss. Math. Graph Theory* **18**:2 (1998), 205–208. MR Zbl

Received: 2016-02-07    Revised: 2016-06-23    Accepted: 2016-07-24

kashbari@math.tamu.edu    Department of Mathematics and Statistics, San Diego State University, San Diego, CA 92182, United States

mosullivan@mail.sdsu.edu    Department of Mathematics and Statistics, San Diego State University, San Diego, CA 92182, United States

# Relations between the conditions of admitting cycles in Boolean and ODE network systems

Yunjiao Wang, Bamidele Omidiran,  
Franklin Kigwe and Kiran Chilakamarri

(Communicated by Richard Rebarber)

*This paper is dedicated to our dear friend Professor Kiran Chilakamarri who passed away due to a sudden illness in 2015.*

Boolean (BL) systems and coupled ordinary differential equations (ODEs) are popular models for studying biological networks. BL systems can be set up without detailed reaction mechanisms and rate constants and provide qualitatively useful information, but they cannot capture the continuous dynamics of biological systems. On the other hand, ODEs are able to capture the continuous dynamic features of biological networks and provide more information on how the activities of components depend on other components and parameter values. However, a useful coupled ODE model requires details about interactions and parameter values. The introduction of the relationships between the two types of models will enable us to leverage their advantages and better understand the target network systems. In this paper, we investigate the relations between the conditions of the existence of limit cycles in ODE networks and their homologous discrete systems. We prove that for a single feedback loop, as long as the corresponding governing functions of the homologous continuous and discrete systems have the same upper and lower asymptotes, the limit cycle borne via Hopf bifurcation corresponds to the cycle of the discrete system. However, for some coupled feedback loops, besides having the same upper and lower asymptotes, parameters such as the decay rates also play crucial roles.

## 1. Introduction

Since the end of twentieth century, due to dramatic advances in technology, biological networks such as gene regulatory networks, protein interaction networks, biochemical reaction networks and neuronal networks have attracted attention from

---

*MSC2010:* 37G99.

*Keywords:* feedback loops, limit cycles, Boolean networks, coupled differential equations.

many different research fields. Mathematical models have shown to be indispensable tools for investigating mechanisms behind biological phenomena. Network systems are often represented by directed graphs, wherein components are represented by nodes and interactions by arrows. Among various modeling frameworks, coupled differential equations (ODEs) and Boolean (BL) networks are popular for modeling regulatory networks.

An  $n$ -node BL network is a discrete dynamical system with the form

$$x_i(t + 1) = f_i(x_1(t), x_2(t), \dots, x_n(t)), \quad (1.1)$$

where  $x_i$  is the state variable of the  $i$ -th node and  $f_i$  is a BL function with the value being either 0 or 1. Since the seminal work of Kauffman [1969], BL networks have been widely used to model biological regulatory networks [Campbell et al. 2011; Thakar et al. 2012; Li et al. 2006; Saez-Rodriguez et al. 2007; Sánchez and Thieffry 2001; Albert and Othmer 2003; Espinosa-Soto et al. 2004; Albert and Wang 2009; Abou-Jaoudé et al. 2009; Glass and Kauffman 1973]. They can be set up in situations where the detailed kinetic characterization of interaction is not available and provide valuable insights [Saadatpour et al. 2013; Glass and Kauffman 1973; Snoussi 1989; Thomas and D'Ari 1990; Edwards and Glass 2000; Edwards et al. 2001; Veliz-Cuba et al. 2014]. However, BL systems cannot faithfully represent the dynamics of biological networks that evolve continuously in time [Tyson and Novák 2010].

An  $n$ -node ODE network has the form

$$\dot{x}_i = F_i(x_1, x_2, \dots, x_n), \quad i = 1, 2, \dots, n,$$

where  $x_i$  is the state variable of the  $i$ -th node and  $F_i$  describes how  $x_i$  depends on other variables. Many researchers have used the ODE framework to study biological network systems [Tyson et al. 2001; 2003; Mogilner et al. 2006; Aldridge et al. 2006; Turner et al. 2010]. Compared with BL models, ODE systems are able to capture the continuous dynamic feature of biological networks and provide more information on how the activities of components depend on other components and parameter values. However, it requires detailed information on interactions and parameter values to set up a useful model.

Often, main dynamical features can be captured by both ODE and BL models [Davidich and Bornholdt 2008; Wittmann et al. 2009; Veliz-Cuba et al. 2014; Abou-Jaoudé et al. 2009; Ouattara et al. 2010]. Given a network, the two different types of models are subject to the same set of constraints resulting from the network structure. This leads to the expectation that their dynamics are closely related as shown in many instances [Abou-Jaoudé et al. 2009; Ouattara et al. 2010; Glass and Kauffman 1973; Veliz-Cuba et al. 2012; 2014; Wittmann et al. 2009; Mendoza and Xenarios 2006; Snoussi 1989]. It was proved under certain conditions that if a continuous network model is monotonic, has distinct upper and lower asymptotes



and has appropriate parameter values corresponding to its discrete homologue, then they may have the same set of stable steady-states, or at least a stable steady-state in the BL network implies a stable steady-state in the homologous continuous one [Glass and Kauffman 1973; Veliz-Cuba et al. 2012; Wittmann et al. 2009; Mendoza and Xenarios 2006; Snoussi 1989]. Glass and Kauffman [1973] also showed that when each node received only one input from other nodes, then a stable limit cycle gives a stable cycle in the BL system. However, the relations between the cycles of ODEs and BL models are still not clear.

To address this issue, we study the relations between the conditions needed to have a cycle in BL networks and those in their homologous ODE networks. Instead of depending on specific reaction mechanisms and rate constants, the ODE systems we consider here are rather qualitative. In this way, we can focus on the differences of the dynamics due to the contrast between discreteness and continuity. More specifically, the ODE network systems we are interested in are in the form

$$\dot{x}_i = \gamma_i \left( \frac{1}{1 + e^{-\sigma_i(a_i + \sum_j v_{ij} x_j)}} - x_i \right), \quad (1.2)$$

where  $i \in \{1, \dots, n\}$ ,  $\gamma_i$ ,  $\sigma_i$  and  $a_i$  are constants, as well as  $v_{ij} = \alpha_{ij} - \beta_{ij}$  ( $\alpha_{ij} \geq 0, \beta_{ij} \geq 0$ ). Here  $\alpha_{ij}$  is the activating coupling from node  $j$  to node  $i$ , and  $\beta_{ij}$  is the inhibitory coupling from node  $j$  to node  $i$ . *If the coupling from node  $j$  to node  $i$  is positive then  $v_{ij} = \alpha_{ij}$  and if the coupling is negative, then  $v_{ij} = -\beta_{ij}$ .* Throughout this paper, we assume that there is no self-regulation, i.e., we assume that  $v_{ii} = 0$ . This type of ODE system was first used by Reinitz et al. [1991] to model gene regulatory networks and then employed by Tyson et al. [2010] to study functional motifs in biochemical reaction networks. So we assume that using the ODE systems in (1.2) to represent biological networks are acceptable. We analytically compare the conditions for supporting a stable limit cycle and find that for a single feedback loop, as long as the corresponding governing functions of the homologous continuous and discrete systems have the same upper and lower asymptotes, a branch of limit cycle borne via Hopf bifurcation corresponds to the cycle of its discrete homologue. However, for coupled feedback loops, besides having the same upper and lower asymptotes, parameters such as the decay rates also play a crucial role.

This paper is constructed as follows. In Section 2, we express the Jacobian matrix as the function of equilibrium, which will facilitate the computation in the later sections. In Section 3, we prove that a negative feedback loop with more than 2 nodes can have stable oscillations borne from Hopf bifurcation. We show in Section 4 that a negative feedback loop of a BL network supports a cycle if and only if each node that has an inhibitor has a background activation (i.e., high basal production rate). Comparing the results from Sections 3 and 4, we conclude that with the same upper and lower asymptotes, a cycle in a BL feedback loop gives a

stable limit cycle in the homologous continuous system. In Section 5, we show that the conditions of Hopf bifurcation occurring in an ODE network, which consists of coupled positive and negative feedback loops, include a restriction on the relations between the decay rates which cannot be implied from the BL network.

## 2. Preliminary: Jacobian matrix at equilibrium

In this section, we give a form of the Jacobian matrix at equilibrium, which will be needed for the computation in the following sections.

**Lemma 2.1.** *Let  $X_0 = (x_1, x_2, \dots, x_n)$  (where  $n \geq 2$ ) be an equilibrium to the system (1.2). Then the Jacobian matrix at the equilibrium is*

$$\begin{pmatrix} -\gamma_1 & f_{12} & \cdots & f_{1,n-1} & f_{1n} \\ f_{21} & -\gamma_2 & \cdots & f_{2,n-2} & f_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ f_{n1} & f_{n2} & \cdots & f_{n,n-1} & -\gamma_n \end{pmatrix},$$

where

$$f_{ij} = \gamma_i x_i (1 - x_i) \sigma_i v_{ij}.$$

*Proof.* Denote the right-hand side of (1.2) by  $f_i$ . Then

$$\frac{df_i}{dx_i} = -\gamma_i,$$

and when  $j \neq i$ ,

$$\begin{aligned} \frac{df_i}{dx_j}(X_0) &= \gamma_i \frac{1}{(1 + e^{-\sigma_i(a_i + \sum_j v_{ij} x_j)})^2} e^{-\sigma_i(a_i + \sum_j v_{ij} x_j)} \sigma_i v_{ij} \\ &= \gamma_i x_i^2 \frac{1 - x_i}{x_i} \sigma_i v_{ij} = \gamma_i x_i (1 - x_i) \sigma_i v_{ij}, \end{aligned}$$

where the second equality is due to the fact that at the equilibrium,

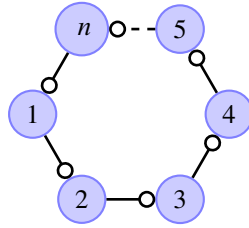
$$\frac{1}{1 + e^{-\sigma_i(a_i + \sum_j v_{ij} x_j)}} = x_i.$$

Hence the Jacobian matrix at the equilibrium  $X_0$  is

$$\begin{pmatrix} -\gamma_1 & f_{12} & \cdots & f_{1,n-1} & f_{1n} \\ f_{21} & -\gamma_2 & \cdots & f_{2,n-2} & f_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ f_{n1} & f_{n2} & \cdots & f_{n,n-1} & -\gamma_n \end{pmatrix},$$

where

$$f_{ij} = \gamma_i x_i (1 - x_i) \sigma_i v_{ij}. \quad \square$$



**Figure 1.** A feedback loop with  $n$  nodes.

### 3. Dynamics of negative feedback loop with ODE equations

In this section, we focus on the dynamics of feedback loops with  $n$  nodes, where the arrows can be either inhibiting or activating. If there are an odd number of inhibitory arrows, then the network is a *negative* feedback loop; otherwise, it is a *positive* feedback loop.

The equations associated to the loop in Figure 1 are

$$\begin{cases} \dot{x}_1 = \gamma_1 \left( \frac{1}{1 + e^{-\sigma_1(a_1 + v_{1,n}x_n)}} - x_1 \right), \\ \dot{x}_i = \gamma_i \left( \frac{1}{1 + e^{-\sigma_i(a_i + v_{i,i-1}x_{i-1})}} - x_i \right), \end{cases} \quad (3.3)$$

where  $i \in \{2, \dots, n\}$  and  $v_{i,j} = \alpha_{ij} - \beta_{ij}$  ( $\alpha_{ij} > 0, \beta_{ij} > 0$ ).

Next we show a result that has been proved in a couple of papers including [Leite and Wang 2010]. Since it is a simple proof, we reproduce it for our system as follows.

**Lemma 3.1.** *Suppose the network associated to system (3.3) is a negative feedback loop. Then the system has a unique equilibrium.*

*Proof.* An equilibrium  $X_0 = (x_1, x_2, \dots, x_n)$  of system (3.3) satisfies

$$\begin{cases} \frac{1}{1 + e^{-\sigma_1(a_1 + v_{1,n}x_n)}} = x_1, \\ \frac{1}{1 + e^{-\sigma_i(a_i + v_{i,i-1}x_{i-1})}} = x_i, \end{cases} \quad (3.4)$$

where  $i \in \{2, \dots, n\}$ .

Let

$$h_i(x) = \frac{1}{1 + e^{-\sigma_i(a_i + v_{i,i-1}x)}}$$

for  $i \in \{1, \dots, n\}$ . Then  $h_i$  are obviously strictly monotonic functions.

Since the coordinates of the equilibrium satisfy (3.4), we have

$$\begin{aligned} x_1 &= h_1(x_n) = h_1 \circ h_n(x_{n-1}) \cdots \\ &= h_1 \circ h_n \circ h_{n-1}(x_{n-2}) \cdots = h_1 \circ h_n \circ h_{n-1} \circ \cdots \circ h_2(x_1). \end{aligned} \quad (3.5)$$

Note that the composition of two monotonic functions is monotonic. Since it is a negative feedback loop,  $h_1 \circ h_n \circ h_{n-1} \circ \dots \circ h_2$  is a strictly monotonically decreasing. Hence there is at most one solution to (3.5).

Now consider the existence of the equilibrium. Note that  $0 \leq x_1 \leq 1$ . When  $x_1 = 0$ , we know  $h_1 \circ h_n \circ h_{n-1} \circ \dots \circ h_2(x_1) > 0$  since  $h_i(x) > 0$  for any value of  $x$ . That is,

$$h_1 \circ h_n \circ h_{n-1} \circ \dots \circ h_2(0) > 0.$$

On the other hand,  $h_i(x) < 1$  for any value of  $x$ . It follows that

$$h_1 \circ h_n \circ h_{n-1} \circ \dots \circ h_2(1) < 1.$$

Now if we let

$$p(x) = h_1 \circ h_n \circ h_{n-1} \circ \dots \circ h_2(x) - x,$$

then  $p(0) > 0$  and  $p(1) < 0$ . By the intermediate value theorem, there is a value of  $x$ , say  $x^*$ , such that  $p(x^*) = 0$ . That is, there exists a  $x^*$  such that

$$h_1 \circ h_n \circ h_{n-1} \circ \dots \circ h_2(x^*) = x^*.$$

So we prove the existence. Therefore, there is a unique equilibrium for any negative feedback loop whose equations have the form of (3.3). □

**Theorem 3.2.** *Let  $X_0 = (x_1, x_2, \dots, x_n)$  be an equilibrium of an  $n$ -node negative feedback loop with associated equations in the form of (3.3). Suppose  $\gamma_i = \gamma > 0$ . Then:*

(1) *The eigenvalues of the Jacobian matrix at the equilibrium are*

$$\lambda_k = -\gamma + \left| \prod_{i=1}^n \gamma x_i (1 - x_i) \sigma_i v_{i,i-1} \right|^{1/n} e^{i(\pi/n + 2k\pi/n)}, \tag{3.6}$$

where  $k = 0, 1, \dots, n - 1$ .

(2) *When  $n = 2$ , the unique equilibrium is always stable.*

(3) *When  $n \geq 3$  and  $a_i = -\frac{1}{2}v_{i,i-1}$ , a branch of periodic solutions can bifurcate from equilibrium with  $x_i = 0.5$  by varying one of the parameters  $\sigma_i$  and fixing remaining other parameter values.*

*Proof.* Without loss of generality, we assume  $\gamma = 1$  (otherwise, we can always rescale the time so that  $\gamma = 1$ ). By Lemma 2.1, the Jacobian matrix of the system (3.3) has the form

$$\begin{pmatrix} -1 & 0 & \dots & 0 & f_{1n} \\ f_{21} & -1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & f_{n,n-1} & -1 \end{pmatrix}.$$

So the characteristic equation of the Jacobian matrix at the equilibrium is

$$0 = (\lambda + 1)^n - f_{1n} f_{21} \dots f_{n,n-1} = (\lambda + 1)^n - \prod_{i=1}^n x_i (1 - x_i) \sigma_i v_{i,i-1}.$$

Hence,

$$(\lambda + 1)^n = \prod_{i=1}^n x_i (1 - x_i) \sigma_i v_{i,i-1}. \tag{3.7}$$

Note that when the feedback loop is negative,  $\prod_{i=1}^n v_{i,i-1}$  is negative. So is the right-hand side of (3.7). Let  $\Delta = \prod_{i=1}^n x_i (1 - x_i) \sigma_i v_{i,i-1}$ , then  $\Delta = |\Delta| e^{i\pi}$ . It follows that

$$\lambda = -1 + |\Delta|^{1/n} e^{i(\pi/n + 2k\pi/n)}$$

for  $0 \leq k \leq n - 1$ .

Note that when  $n = 2$ ,

$$\lambda_1 = -1 + |\Delta|^{1/2} e^{i(\pi/2)} = -1 + i|\Delta|^{1/2}$$

and

$$\lambda_2 = -1 + |\Delta|^{1/2} e^{i(\pi/2 + \pi)} = -1 - i|\Delta|^{1/2}.$$

It follows that  $\text{Re}(\lambda_k) = -1$  for  $k \in \{1, 2\}$ . Hence, a negative feedback loop with only two nodes must only have a stable equilibrium.

When  $n \geq 3$ , the pair of conjugate roots  $-1 + |\Delta|(\cos \pi/n \pm i \sin \pi/n)$  have the largest real part:  $-1 + |\Delta| \cos \pi/n = -1 + |\prod_{i=1}^n x_i (1 - x_i) \sigma_i v_{i,i-1}| \cos \pi/n$ . Note that when  $a_i = -\frac{1}{2} v_{i,i-1}$ , it is straightforward to show that  $\{x_i = \frac{1}{2}\}$  is an equilibrium. Since the expression of eigenvalues is independent of  $a_i$ , we can vary  $\sigma_i$  so that the real part changes from negative to positive and leaves the other eigenvalues with negative real parts. Therefore, a branch of limit cycles can be borne through Hopf bifurcation.  $\square$

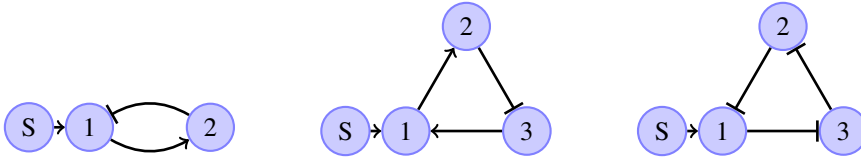
**Remark 3.3.** The theorem states that when  $a_i = -\frac{1}{2} v_{i,i-1}$ , by varying the parameter of steepness  $\sigma_i$  a limit cycle can be obtained via Hopf bifurcation. Note that  $v_{i,i-1}$  is either equal to the activating coupling parameter  $\alpha_{i,i-1}$  or equal to the inhibitory coupling parameter  $-\beta_{i,i-1}$  in the feedback loop. Note that  $1/(1 + e^{-\sigma_i a_i})$  is the basal production rate. That is, when node  $i$  has inhibitory input, its basal production rate has to be relatively high since  $a_i > 0$ .

Also with the parameter setting  $a_i = -\frac{1}{2} v_{i,i-1}$ , we have

$$\left. \frac{1}{1 + e^{-\sigma_i (a_i + v_{i,i-1} x_{i-1})}} \right|_{x_{i-1}=1} = \frac{1}{1 + e^{-1/2 \sigma_i v_{i,i-1}}}$$

and

$$\left. \frac{1}{1 + e^{-\sigma_i (a_i + v_{i,i-1} x_{i-1})}} \right|_{x_{i-1}=0} = \frac{1}{1 + e^{1/2 \sigma_i v_{i,i-1}}}.$$



**Figure 2.** Two- and three-node negative feedback loops with a constant signal to node 1, as in [Tyson and Novák 2010]. The dynamics of the corresponding ODEs are equivalent to those without the signal.

We consider the values of  $x_i \in [0, 1]$ . So if  $v_{i,i-1} = \alpha_{i,i-1} > 0$  (i.e., node  $i$  has an activating input from node  $i - 1$ ), the sigmoidal has maximum value  $(1 + e^{-1/2\sigma_i\alpha_{i,i-1}})^{-1}$  at  $x_{i,i-1} = 1$  that goes towards 1 as  $\sigma_i \rightarrow \infty$  and has minimum value  $(1 + e^{1/2\sigma_i\alpha_{i,i-1}})^{-1}$  at  $x_{i,i} = 0$  that goes towards 0 as  $\sigma_i \rightarrow \infty$ . On the other hand, if  $v_{i,i-1} = -\beta_{i,i-1} < 0$  (i.e., node  $i$  has an inhibitory input from node  $i - 1$ ), the sigmoidal has maximum value  $(1 + e^{1/2\sigma_i\beta_{i,i-1}})^{-1}$  at  $x_{i,i-1} = 0$  that goes towards 1 as  $\sigma_i \rightarrow \infty$  and has minimum value  $(1 + e^{-1/2\sigma_i\beta_{i,i-1}})^{-1}$  at  $x_{i,i-1} = 1$  that goes towards 0 as  $\sigma_i \rightarrow \infty$ .

**Remark 3.4.** We recall dynamics of some networks studied by Tyson et al. [2010] (reproduced in Figure 2). It was assumed that node 1 has a constant basal production that is indicated by  $S$ . The equations to the network in Figure 2 are in the form

$$\begin{cases} \dot{x}_1 = \gamma_1 \left( \frac{1}{1 + e^{-\sigma_1(S+a_1+v_{1,n}x_n)}} - x_1 \right), \\ \dot{x}_i = \gamma_i \left( \frac{1}{1 + e^{-\sigma_i(a_i+v_{i,i-1}x_{i-1})}} - x_i \right), \end{cases} \tag{3.8}$$

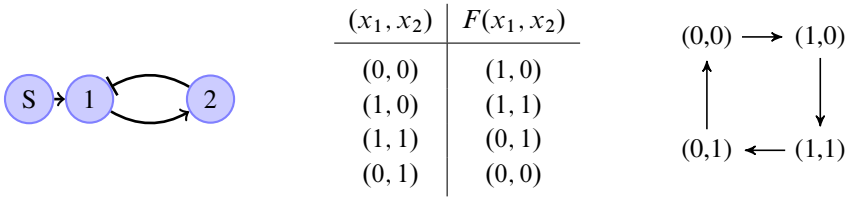
where  $i = 2$  or  $i \in \{2, 3\}$  and  $v_{i,j} = \alpha_{i,j} - \beta_{i,j}$ .

Note that both  $S$  and  $a_1$  are constants. We can relabel  $S + a_1$  by  $a_1^*$ . Then (3.8) is again in the form of the feedback loop without signal as (3.3). So the dynamics are the same as we discussed in Section 3.

#### 4. Dynamics of negative feedback loop with Boolean functions

With a given interaction network, there are many ways to choose BL functions for the nodes. Here we adopt the well-cited assumptions for the associated BL functions proposed by Albert and Othmer [2003]. We make the following assumptions, which we will refer to as *axioms*:

- (1) The effects of activators and inhibitors are never additive, but rather, inhibitors are dominant.
- (2) The activity of a node will be “on” in the next time step if at least one of its activators is “on” and all inhibitors are “off”.



**Figure 3.** Left: two-node negative feedback loop that admits a cycle; middle: BL map; right: transition graph.

- (3) The activity of a node will be “off” in the next time step if none of its activators are “on”.
- (4) If a node has a background activation, then we assume that the node has an activator that is permanently “on”.

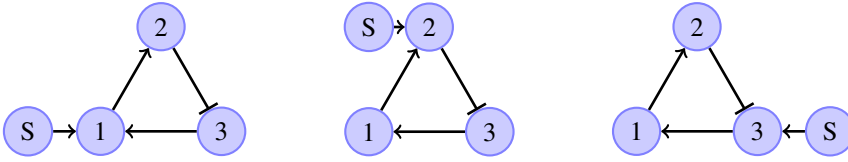
Let  $I(i)$  be the set of inhibitors and  $A(i)$  be the set of activators of the  $i$ -th node. Then we can express the *axioms* by the following logic function:

$$x_i(t+1) = \begin{cases} (\neg \bigvee_{j \in I(i)} x_j(t)) \wedge \bigvee_{k \in A(i)} x_k(t) & \text{when node } i \text{ has no background activation,} \\ \neg \bigvee_{j \in I(i)} x_j(t) & \text{when node } i \text{ has a background activation.} \end{cases}$$

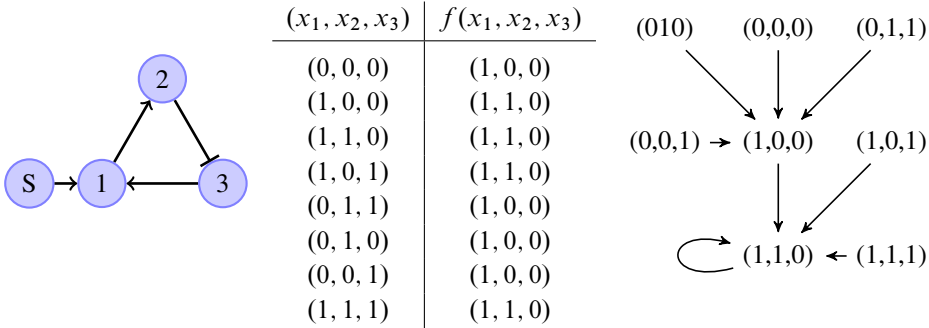
For example, for the network in Figure 3, node 1 receives two inputs: one inhibitor and another one is a background activator, and node 2 receives one activator from node 1. So if  $x_1 = 0$  and  $x_2 = 1$ , then in the next time step,  $x_1$  remains 0 since its inhibitor node 2 is on and  $x_2 = 0$  since its only activator is off. It is straightforward to check that the BL function associated to the network must be the one listed in the table in Figure 3. The dynamics of the two-node network in Figure 2 can be described by the transition graph in Figure 3. Note that this network admits a cycle.

**Three-node negative feedback loop.** Next we consider a network of three-node negative feedback loops. We assume one of the nodes has a background activation. Then there are three cases as shown in Figure 4.

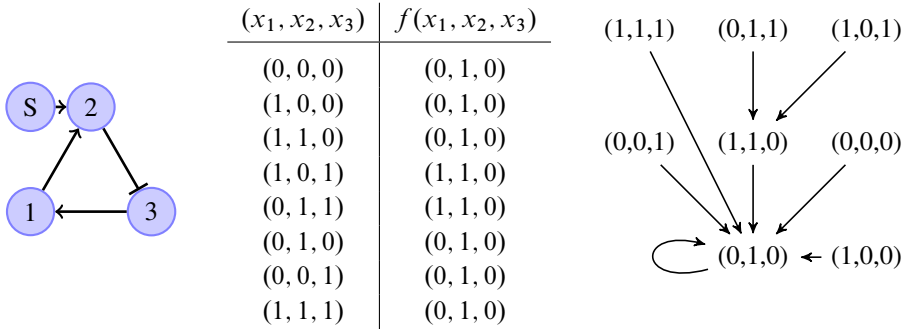
*Case I.* Assume node 1 has the background activator shown in Figure 5, left. Then following the *axioms*, the BL functions associated to the network is the one in Figure 5, middle, and the transition diagram is as in Figure 5, right. We can see that  $(1, 1, 0)$  is a fixed point and all other points will converge to the fixed point over the time. As a result, no cycle exists.



**Figure 4.** Three different background activation locations.



**Figure 5.** Case I, left: background signal is on node 1; middle: BL map; right: transition graph.

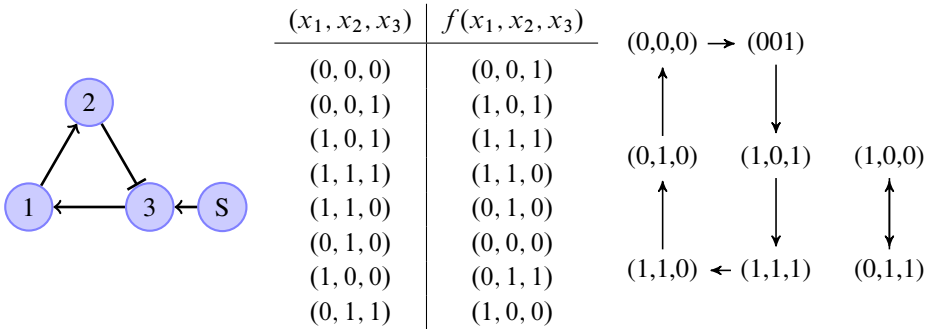


**Figure 6.** Case II, left: background signal is on node 2; middle: BL map; right: transition graph.

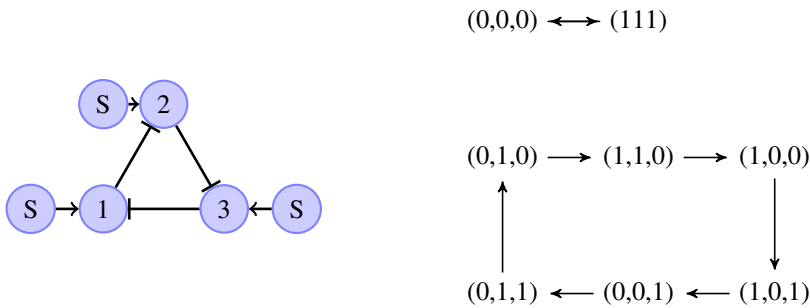
*Case II.* Assume node 2 has the background activation as Figure 6, left. Then the corresponding BL function and transition graph are Figure 6, middle, and Figure 6, right, respectively. Again, we can see that the system has only a stable fixed point  $(1, 1, 0)$ . As a result, no cycle exists.

*Case III.* Assume node 3 has the background activation. Similarly, we can determine its associated BL function and transition graph as in Figure 7. Different from the other two cases, there exists a cycle with length 6.





**Figure 7.** Case III, left: background signal is on node 3; middle: BL map; right: transition graph.



**Figure 8.** BL network that admits cycles.

Similarly, we can show for the three-node network in Figure 8, the network admits cycle only if each node receives a background activation.

**Dynamics of  $n$ -node negative feedback loop in Figure 2, right.** The analysis of BL three-node negative feedback networks discussed in the previous section shows that a network admits cycles only if each node that receives inhibitory input has background activation. This observation can be generalized to any  $n$ -node negative feedback loop.

Let  $x_i^m$  be the value of the state variable of node  $i$  at the  $m$ -th time step. Then the BL system of the feedback loop in Figure 1 has the form

$$\begin{cases} x_1^{m+1} = f_1(x_n^m), \\ x_i^{m+1} = f_i(x_{i-1}^m). \end{cases} \tag{4.9}$$

**Lemma 4.1.** *Let  $\mathcal{G}$  be an  $n$ -node feedback loop with associated BL system having the form of (4.9). Then for any  $m > n$ ,*

$$x_1^{m+1} = f_1 \circ f_n \circ f_{n-1} \circ \dots \circ f_2(x_1^{m+1-n}).$$

*Proof.* It follows straightforwardly from (4.9). □

**Lemma 4.2.** *Let  $\mathcal{G}$  be a feedback loop with the associated BL system satisfying the axioms. Suppose each node with an inhibitory input from some other node has a background activation. Then*

- (1) *if node  $i$  receives a negative input, then the associated BL function is  $f_i(0) = 1$  and  $f_i(1) = 0$ ;*
- (2) *if node  $i$  receives a positive input, then the associated BL function is  $f_i(0) = 0$  and  $f_i(1) = 1$ ;*

*and the compositions of  $f_i$  are bijections.*

*Proof.* Items (1) and (2) follow straightforwardly from the axioms. So all  $f_i$  are bijections. It then follows that the compositions of  $f_i$  are bijections.  $\square$

**Lemma 4.3.** *Let  $\mathcal{C}$  be a node of an  $n$ -node feedback loop  $\mathcal{G}$ . Suppose the value of the state variable of  $\mathcal{C}$  stays constant after a finite number of time steps. Then the associated BL system does not have nontrivial cycles.*

*Proof.* Without loss of generality, we relabel the nodes of  $\mathcal{G}$  so that  $\mathcal{C}$  is node 1 and the rest of the nodes are relabeled as in Figure 1. Let  $x_i^m$  be the value of the state variable of node  $i$  at the  $m$ -th time step. Then the BL system has the form of (4.9).

Since this is a deterministic system, when the value  $x_1$  is fixed after a finite series of steps, say  $M$ , then by Lemma 4.1, the values of all other  $x_i$  will be fixed after  $M + n$  time steps. So the system only has fixed points and does not admit nontrivial cycles.  $\square$

**Theorem 4.4.** *Let  $\mathcal{G}$  be a negative feedback loop with the associated BL system satisfying the axioms. Then  $\mathcal{G}$  admits cycles if and only if each node with a negative input from some other node has a background activation.*

*Proof.* We first prove by contradiction that if one of the nodes with negative inputs from other nodes has no background activation, then the system does not admit cycle. Suppose node  $\mathcal{C}$  of the negative feedback loop  $\mathcal{G}$  has a negative input from the other node and has no background excitation. Note that if the initial state value of  $\mathcal{C}$  is zero, then the value of  $\mathcal{C}$  stays zero forever; if the initial state value of  $\mathcal{C}$  is 1, then because  $\mathcal{C}$  has no excitation input, the value of  $\mathcal{C}$  becomes zero in the next time step and remains zero forever. By Lemma 4.3, the negative feedback does not have a cycle.

Next we prove that if all suppressed nodes have background activation, then a cycle exists. It is sufficient to show that the value of each state variable changes over time. By Lemma 4.1, for any  $m > n$ ,

$$x_1^{m+1} = f_1 \circ f_n \circ f_{n-1} \circ \cdots \circ f_2(x_1^{m+1-n}).$$

Since  $\mathcal{G}$  is a negative feedback loop,

$$f_1 \circ f_n \circ f_{n-1} \circ \cdots \circ f_2(0) = 1 \quad \text{and} \quad f_1 \circ f_n \circ f_{n-1} \circ \cdots \circ f_2(1) = 0.$$

So the value of the state variable of node 1 changes every  $n$  steps. Because  $f_i$  are bijections, the values of other node states also change over time.  $\square$

*Comparison.* Theorem 4.4 states that a BL feedback loop admits a cycle only if every node with an inhibitory input has a background activation. In the other words, the governing BL function of a node  $i$  with an inhibitory (negative) input must be  $f_i(x) = 1 - x$ .

Compared with the results for the discrete homologue, the conditions for the continuous system are essentially the same. As discussed in Remark 3.3, each node with an inhibitory input must have a relatively high basal production rate, and as the steepness parameter  $\sigma_i$  goes to  $\infty$ , the governing function is the same as the BL system.

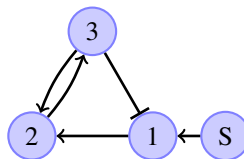
## 5. Networks with two or more feedback loops

*Dynamics of the network in Figure 9.* The first network we examine is the one in Figure 9, which was studied by Tyson et al. [2010]. The authors showed that without the positive input from node 3 to node 2 (i.e.,  $\alpha_{23} = 0$ ), the network of ODEs demonstrates oscillations in a certain range of the parameter value  $S$  (with other parameter values fixed). The oscillating range of  $S$  shrinks as the coupling parameter  $\alpha_{23}$  increases and it disappears when  $\alpha_{23}$  increases to a certain value. We show next that the effect of the parameter  $\alpha_{23}$  can be captured by two BL systems:

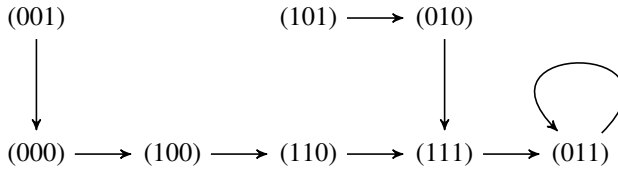
- (1) Besides functions based on the *axioms*, the governing function of node 2 which has two inputs,  $f_2(x_1, x_3)$ , satisfies

$$f_2(1, *) = 1 \quad \text{and} \quad f_2(0, *) = 0,$$

i.e., the activity of node 2 is dominated by the activity of node 1 and the effect of node 3 is negligible. For this setting, the dynamics of the network is the same as the three-node network feedback loop in Figure 7 and it has a stable cycle. This BL system can capture the dynamics of the corresponding ODE system with  $\alpha_{23} = 0$  or relatively small.



**Figure 9.** A network consists of two feedback loops.



**Figure 10.** Transitions of the network in Figure 9 with the governing function for node 2 is node 2 is on if either node 1 or node 2 is on.

- (2) Besides functions based on *axioms*, the governing function of node 2,  $f_2(x_1, x_3)$ , satisfies

$$f_2(1, *) = 1, \quad f_2(*, 1) = 1, \quad \text{and} \quad f_2(0, 0) = 0,$$

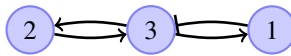
i.e., node 2 is on if either node 1 or node 3 is on.

This transition diagram of the system is shown in Figure 10. It is clear that the system only has a fixed point which captures the case when  $\alpha_{23}$  is sufficiently large.

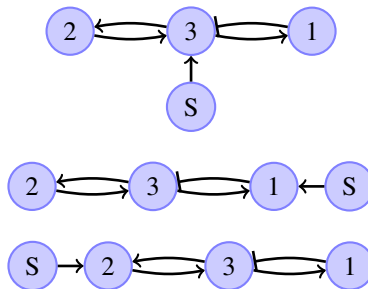
**Remark 5.1.** From this example, we can observe that ODE systems can be viewed as “organizing centers” of BL systems.

**Boolean system of the network in Figure 11.** Suppose the network in Figure 11 receives a signal through one of the nodes. The possible networks are as in Figure 12.

It is rather straightforward to check that when the signal goes to node 1 or 2, the corresponding BL network can only have a stable steady-state, and when the signal goes through node 3, then it has a stable cycle:  $(000) \rightarrow (001) \rightarrow (111) \rightarrow (110) \rightarrow (000)$ .



**Figure 11.** Another network consisting of two feedback loops.



**Figure 12.** Three possible signal input places.

**Dynamics of the ODE systems of the network in Figure 11.** By Lemma 2.1, the Jacobian matrix of an ODE system associated to the network in Figure 11 at an equilibrium has the form

$$\begin{pmatrix} -\gamma_1 & 0 & f_{13} \\ 0 & -\gamma_2 & f_{23} \\ f_{31} & f_{32} & -\gamma_3 \end{pmatrix},$$

where  $f_{ij} = \gamma_i x_i (1 - x_i) \sigma_i v_{ij}$ .

Therefore, the characteristic polynomial equation of the matrix is

$$\begin{aligned} |\lambda I - J| &= (\lambda + \gamma_1)(\lambda + \gamma_2)(\lambda + \gamma_3) - (\lambda + \gamma_2)f_{31}f_{13} - (\lambda + \gamma_1)f_{32}f_{23} \\ &= \lambda^3 + (\gamma_1 + \gamma_2 + \gamma_3)\lambda^2 + (\gamma_1\gamma_2 + \gamma_2\gamma_3 + \gamma_1\gamma_3 - f_{13}f_{31} - f_{23}f_{32})\lambda \\ &\quad + \gamma_1\gamma_2\gamma_3 - \gamma_2f_{13}f_{31} - \gamma_1f_{23}f_{32}. \end{aligned} \tag{5.10}$$

**Theorem 5.2.** *The condition  $\gamma_2 > \gamma_1$  is necessary for Hopf bifurcation to occur.*

*Proof.* Let us label the coefficient of  $\lambda^2$  as  $c_1$ , the coefficient of  $\lambda$  as  $c_2$  and the constant term as  $c_3$ . Then the conditions for having a pair of pure imaginary eigenvalues are:

- $c_1 = \gamma_1 + \gamma_2 + \gamma_3 > 0.$  (5.11)

- $c_2 = \gamma_1\gamma_2 + \gamma_2\gamma_3 + \gamma_1\gamma_3 - f_{13}f_{31} - f_{23}f_{32} > 0.$  It follows that

$$\gamma_1\gamma_2 + \gamma_2\gamma_3 + \gamma_1\gamma_3 - f_{13}f_{31} > f_{23}f_{32}. \tag{5.12}$$

- $c_3 - c_1c_2 = 0,$  i.e.,

$$\begin{aligned} &\gamma_1\gamma_2\gamma_3 - \gamma_2f_{13}f_{31} - \gamma_1f_{23}f_{32} \\ &\quad - (\gamma_1 + \gamma_2 + \gamma_3)(\gamma_1\gamma_2 + \gamma_2\gamma_3 + \gamma_1\gamma_3 - f_{13}f_{31} - f_{23}f_{32}) = 0. \end{aligned} \tag{5.13}$$

It follows that

$$\begin{aligned} &(\gamma_2 + \gamma_3)f_{23}f_{32} - \gamma_1^2(\gamma_2 + \gamma_3) - \gamma_2^2(\gamma_1 + \gamma_3) \\ &\quad - \gamma_3^2(\gamma_1 + \gamma_2) - 2\gamma_1\gamma_2\gamma_3 + (\gamma_1 + \gamma_3)f_{13}f_{31} = 0. \end{aligned} \tag{5.14}$$

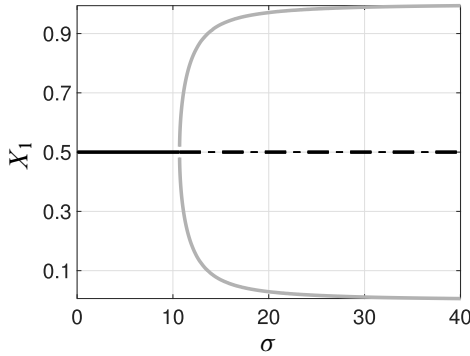
Inequality (5.12) and equation (5.14) imply that

$$\begin{aligned} &(\gamma_2 + \gamma_3)(\gamma_1\gamma_2 + \gamma_2\gamma_3 + \gamma_1\gamma_3 - f_{13}f_{31}) \\ &\quad > \gamma_1^2(\gamma_2 + \gamma_3) + \gamma_2^2(\gamma_1 + \gamma_3) + \gamma_3^2(\gamma_1 + \gamma_2) \\ &\quad \quad \quad + 2\gamma_1\gamma_2\gamma_3 - (\gamma_1 + \gamma_3)f_{13}f_{31}. \end{aligned} \tag{5.15}$$

Simplifying (5.15), we have

$$-(\gamma_2 - \gamma_1)f_{13}f_{31} > \gamma_1^2(\gamma_2 + \gamma_3). \tag{5.16}$$

By the condition of the network,  $f_{13}f_{31} < 0$ , so inequality (5.16) implies  $\gamma_2 > \gamma_1$ .  $\square$



**Figure 13.** Bifurcation diagram at the parameter values  $\sigma_1 = \sigma_3 = \sigma_2 = \sigma$ ,  $\gamma_1 = \gamma_3 = 1$  and  $\gamma_2 = 1.5$ ,  $\alpha_{13} = \beta_{31} = \alpha_{23} = \alpha_{32} = 1$ ,  $a_1 = a_2 = -0.5$  and  $a_3 = 0$ , and with the rest of the parameters being zero. Here the gray curve represents a branch of stable limit cycle, the solid black line represents a branch of stable equilibria and the black dashed line a branch of unstable equilibria.

How can we choose parameter values so that we will observe sustained oscillations that close to the Hopf bifurcation point? Suppose the bifurcation is supercritical; then near the bifurcation point,  $c_3 - c_1c_2 \geq 0$  while  $c_1$  and  $c_2$  remain positive. Now by substituting  $f_{ij}$  by  $\gamma_i x_i(1 - x_i)\sigma_i v_{ij}$  in inequalities (5.16), (5.12) and  $c_3 - c_1c_2 \geq 0$ , we obtain

$$(\gamma_2 - \gamma_1)\gamma_1\gamma_3x_1x_3(1 - x_1)(1 - x_3)\sigma_1\sigma_3\beta_{31}\alpha_{13} > \gamma_1^2(\gamma_2 + \gamma_3), \tag{5.17}$$

$$\begin{aligned} \gamma_1\gamma_2 + \gamma_2\gamma_3 + \gamma_1\gamma_3 + \gamma_1\gamma_3x_1x_3(1 - x_1)(1 - x_3)\sigma_1\sigma_3\beta_{31}\alpha_{13} \\ > \gamma_2\gamma_3x_2x_3(1 - x_2)(1 - x_3)\sigma_2\sigma_3\alpha_{32}\alpha_{23} \end{aligned} \tag{5.18}$$

and

$$\begin{aligned} \gamma_1^2(\gamma_2 + \gamma_3) + \gamma_2^2(\gamma_1 + \gamma_3) + \gamma_3^2(\gamma_1 + \gamma_2) + 2\gamma_1\gamma_2\gamma_3 \\ + (\gamma_1 + \gamma_3)\gamma_1\gamma_3x_1x_3(1 - x_1)(1 - x_3)\sigma_1\sigma_3\beta_{31}\alpha_{13} \\ \leq (\gamma_2 + \gamma_3)\gamma_2\gamma_3x_2x_3(1 - x_2)(1 - x_3)\sigma_2\sigma_3\alpha_{32}\alpha_{23}. \end{aligned} \tag{5.19}$$

Focusing on the equilibria with  $x_i = 0.5$ , we can find a range of parameter values that satisfy conditions (5.16), (5.18) and (5.19). For example,  $\sigma_1 = \sigma_3 = \sigma_2 = \sigma$ ,  $\gamma_1 = \gamma_3 = 1$  and  $\gamma_2 = 1.5$ ,  $\alpha_{13} = \beta_{31} = \alpha_{23} = \alpha_{32} = 1$ ,  $a_1 = a_2 = -0.5$  and  $a_3 = 0$ . By setting the rest of the parameters to zero and varying the value  $\sigma$ , we can find a branch of limit cycle occurring through Hopf bifurcation; see Figure 13.

*Comparison of discrete and continuous homologues.* Now we compare the conditions for the homologous systems of the network in Figure 11. The BL system

requires that node 3 has a background activation, which is reflected in the choices of parameter values associated to basal production rates in the ODE system:  $a_1 = a_2 = -0.5$  and  $a_3 = 0$ , where  $a_3$  is actually the summation of the two parameters  $a_3$  and signal  $S$  with  $a_3 = -0.5$  and  $S = 0.5$ . In order to realize oscillations in the continuous system, we need to find suitable values for other parameters as well. For example, we need to impose a restriction on the relation of decay rates  $\gamma_2 > \gamma_1$  in order to observe stable oscillations. Such requirements in the parameter values of ODE systems do not have correspondence in the BL systems.

## 6. Discussion

Glass and Kauffman [1973] showed that a stable limit cycle of a continuous network gives a cycle in its discrete homologue under the condition that each node has only one input from other nodes. In this work, we compared the conditions for each type possessing a stable cycle for the case where each node has one input and also examined two cases when some nodes have two inputs. Our strategy of focusing on the type of ODE systems in a rather abstract form enables us to perform analytical examinations and to possibly extract essential dynamical differences between the two types of network models. The strategy has the potential to be used for more extensive study of the relations as to provide more efficient algorithms for converting between continuous and discrete network systems.

## 7. Acknowledgements

We would like to thank the reviewer for the thoughtful comments. Yunjiao Wang was supported by DHS-14-ST-062-001 and seeds grant from Texas Southern University. Kigwe and Omidiran are undergraduates and were supported by the Summer Undergraduate Research Program from COSET at Texas Southern University.

## References

- [Abou-Jaoudé et al. 2009] W. Abou-Jaoudé, D. A. Ouattara, and M. Kaufman, “From structure to dynamics: frequency tuning in the p53-Mdm2 network, I: Logical approach”, *J. Theoret. Biol.* **258**:4 (2009), 561–577.
- [Albert and Othmer 2003] R. Albert and H. G. Othmer, “The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in *Drosophila melanogaster*”, *J. Theoret. Biol.* **223**:1 (2003), 1–18. MR
- [Albert and Wang 2009] R. Albert and R.-S. Wang, “Discrete dynamic modeling of cellular signaling networks”, pp. 281–306 in *Computer methods, B*, edited by M. L. Johnson and L. Brand, Methods in Enzymology **467**, Academic Press, 2009.
- [Aldridge et al. 2006] B. B. Aldridge, J. M. Burke, D. A. Lauffenburger, and P. K. Sorger, “Physico-chemical modelling of cell signalling pathways”, *Nat. Cell Biol.* **8**:11 (2006), 1195–1203.

- [Campbell et al. 2011] C. Campbell, J. Thakar, and R. Albert, “Network analysis reveals cross-links of the immune pathways activated by bacteria and allergen”, *Phys. Rev. E* **84** (2011), art. id. 031929, 12 pp.
- [Davidich and Bornholdt 2008] M. Davidich and S. Bornholdt, “The transition from differential equations to Boolean networks: a case study in simplifying a regulatory network model”, *J. Theoret. Biol.* **255**:3 (2008), 269 – 277.
- [Edwards and Glass 2000] R. Edwards and L. Glass, “Combinatorial explosion in model gene networks”, *Chaos* **10**:3 (2000), 691–704. MR Zbl
- [Edwards et al. 2001] R. Edwards, H. T. Siegelmann, K. Aziza, and L. Glass, “Symbolic dynamics and computation in model gene networks”, *Chaos* **11**:1 (2001), 160–169.
- [Espinosa-Soto et al. 2004] C. Espinosa-Soto, P. Padilla-Longoria, and E. R. Alvarez-Buylla, “A gene regulatory network model for cell-fate determination during *Arabidopsis thaliana* flower development that is robust and recovers experimental gene expression profiles”, *Plant Cell* **16**:11 (2004), 2923–2939.
- [Glass and Kauffman 1973] L. Glass and S. A. Kauffman, “The logical analysis of continuous, non-linear biochemical control networks”, *J. Theoret. Biol.* **39**:1 (1973), 103–129.
- [Kauffman 1969] S. Kauffman, “Metabolic stability and epigenesis in randomly constructed genetic nets”, *J. Theoret. Biol.* **22**:3 (1969), 437–467.
- [Leite and Wang 2010] M. C. A. Leite and Y. Wang, “Multistability, oscillations and bifurcations in feedback loops”, *Math. Biosci. Eng.* **7**:1 (2010), 83–97. MR Zbl
- [Li et al. 2006] S. Li, S. M. Assmann, and R. Albert, “Predicting essential components of signal transduction networks: a dynamic model of guard cell abscisic acid signaling”, *PLoS Biol* **4**:10 (2006), art. id. e312, 17 pp.
- [Mendoza and Xenarios 2006] L. Mendoza and I. Xenarios, “A method for the generation of standardized qualitative dynamical systems of regulatory networks”, *Theor. Biol. Med. Model.* **3** (2006), art. id. 13, 18 pp.
- [Mjolsness et al. 1991] E. Mjolsness, D. H. Sharp, and J. Reinitz, “A connectionist model of development”, *J. Theoret. Biol.* **152**:4 (1991), 429–453.
- [Mogilner et al. 2006] A. Mogilner, R. Wollman, and W. F. Marshall, “Quantitative modeling in cell biology: what is it good for?”, *Dev. Cell* **11**:3 (2006), 279–287.
- [Ouattara et al. 2010] D. A. Ouattara, W. Abou-Jaoudé, and M. Kaufman, “From structure to dynamics: frequency tuning in the p53-Mdm2 network, II: Differential and stochastic approaches”, *J. Theoret. Biol.* **264**:4 (2010), 1177–1189.
- [Saadatpour et al. 2013] A. Saadatpour, R. Albert, and T. C. Reluga, “A reduction method for Boolean network models proven to conserve attractors”, *SIAM J. Appl. Dyn. Syst.* **12**:4 (2013), 1997–2011. MR Zbl
- [Saez-Rodriguez et al. 2007] J. Saez-Rodriguez, L. Simeoni, J. A. Lindquist, R. Hemenway, U. Bommhardt, B. Arndt, U.-U. Haus, R. Weismantel, E. D. Gilles, S. Klamt, and B. Schraven, “A logical model provides insights into T cell receptor signaling”, *PLoS Comput. Biol.* **3**:8 (2007), art. id. e163, 11 pp.
- [Sánchez and Thieffry 2001] L. Sánchez and D. Thieffry, “A logical analysis of the *Drosophila* gap-gene system”, *J. Theoret. Biol.* **211**:2 (2001), 115–141.
- [Snoussi 1989] E. H. Snoussi, “Qualitative dynamics of piecewise-linear differential equations: a discrete mapping approach”, *Dynam. Stability Systems* **4**:3-4 (1989), 189–207. MR Zbl



- [Thakar et al. 2012] J. Thakar, A. K. Pathak, L. Murphy, R. Albert, and I. M. Cattadori, “Network model of immune responses reveals key effectors to single and co-infection dynamics by a respiratory bacterium and a gastrointestinal helminth”, *PLoS Comput. Biol.* **8**:1 (2012), art. id. e1002345, 19 pp.
- [Thomas and D’Ari 1990] R. Thomas and R. D’Ari, *Biological feedback*, CRC Press, Boca Raton, 1990. Zbl
- [Turner et al. 2010] D. A. Turner, P. Paszek, D. J. Woodcock, D. E. Nelson, C. A. Horton, Y. Wang, D. G. Spiller, D. A. Rand, M. R. H. White, and C. V. Harper, “Physiological levels of TNF $\alpha$  stimulation induce stochastic dynamics of NF- $\kappa$ b responses in single living cells”, *J. Cell Sci.* **123**:16 (2010), 2834–2843.
- [Tyson and Novák 2010] J. J. Tyson and B. Novák, “Functional motifs in biochemical reaction networks”, *Annu. Rev. Phys. Chem.* **61** (2010), 219–240.
- [Tyson et al. 2001] J. J. Tyson, K. Chen, and B. Novak, “Network dynamics and cell physiology”, *Nat. Rev. Mol. Cell Biol.* **2**:12 (2001), 908–916.
- [Tyson et al. 2003] J. J. Tyson, K. C. Chen, and B. Novak, “Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell”, *Curr. Opin. Cell Biol.* **15**:2 (2003), 221–231.
- [Veliz-Cuba et al. 2012] A. Veliz-Cuba, J. Arthur, L. Hochstetler, V. Klomps, and E. Korpi, “On the relationship of steady states of continuous and discrete models arising from biology”, *Bull. Math. Biol.* **74**:12 (2012), 2779–2792. MR Zbl
- [Veliz-Cuba et al. 2014] A. Veliz-Cuba, A. Kumar, and K. Josić, “Piecewise linear and Boolean models of chemical reaction networks”, *Bull. Math. Biol.* **76**:12 (2014), 2945–2984. MR Zbl
- [Wittmann et al. 2009] D. M. Wittmann, J. Krumsiek, J. Saez-Rodriguez, D. A. Lauffenburger, S. Klamt, and F. J. Theis, “Transforming Boolean models to continuous models: methodology and application to T-cell receptor signaling”, *BMC Syst. Biol.* **3**:1 (2009), 98.

Received: 2016-03-18

Revised: 2016-08-11

Accepted: 2016-08-17

wangyx@tsu.edu

*Department of Mathematics, Texas Southern University,  
3100 Cleburne Street, Houston, TX 77004, United States*

bomidiran@gmail.com

*Department of Economics, University of Pennsylvania,  
Philadelphia, PA 19104, United States*

fkigwe@gmail.com

*Department of Engineering, Texas Southern University,  
Houston, TX 77004, United States*

varu\_chilakamarri@yahoo.com

*Department of Mathematics, Texas Southern University,  
3100 Cleburne Street, Houston, TX 77004, United States*



# Weak and strong solutions to the inverse-square brachistochrone problem on circular and annular domains

Christopher Grimm and John A. Gemmer

(Communicated by John Baxley)

In this paper we study the brachistochrone problem in an inverse-square gravitational field on the unit disk. We show that the time-optimal solutions consist of either smooth strong solutions to the Euler–Lagrange equation or weak solutions formed by appropriately patched together strong solutions. This combination of weak and strong solutions completely foliates the unit disk. We also consider the problem on annular domains and show that the time-optimal paths foliate the annulus. These foliations on the annular domains converge to the foliation on the unit disk in the limit of vanishing inner radius.

## 1. Introduction

In 1696 Johann Bernoulli posed the following problem: given two points  $A$ ,  $B$ , find a curve connecting  $A$  and  $B$  such that a particle traveling from  $A$  to  $B$  under the influence of a uniform gravitational field takes the minimum time. This is called the *brachistochrone problem*, from the Greek terms *brachistos* for shortest and *chronos* for time. It was solved the following year by Leibniz, L'Hospital, Newton, and others [Dunham 1990]. While the solution to the brachistochrone problem has limited applications, the techniques from calculus used to solve it were novel at the time. Namely, rudimentary techniques from what would later be called the calculus of variations were developed. Euler and Lagrange later formalized these initial approaches into their celebrated necessary conditions for optimality, what we now call the Euler–Lagrange equations. For certain types of functionals, this approach reduces the optimization problem to solving a differential equation, i.e., the Euler–Lagrange equation, corresponding to the functional. Indeed, the reduction of an optimization problem to that of solving a differential equation can be

---

*MSC2010:* 49K05, 49K30, 49S05.

*Keywords:* brachistochrone problem, calculus of variations of one independent variable, eikonal equation, geometric optics.

directly applied to many classical optimization problems such as the isoperimetric problem [Blåsjö 2005], determining the shape of a minimal surface [Sagan 1969; Oprea 2000] and calculating the path of a geodesic on a surface [McCleary 2013]. Moreover, in classical mechanics the dynamics of a system can be derived using the Euler–Lagrange equations to extremize the so-called “action” of the system [Goldstein et al. 2014]. This approach to classical mechanics is equivalent to Newtonian mechanics but leads to deeper insights which are critical to our current mathematical understanding of quantum mechanics, general relativity, and other branches of physics.

While the Euler–Lagrange equations have been very successfully applied to many problems in engineering and physics, they do not provide the complete picture. In particular, as necessary conditions for optimality, their derivation implicitly assumes existence and smoothness of a minimum. In modern mathematics and applications these assumptions are naive. For instance, many problems in continuum mechanics have minimizers which lack enough regularity to be classical solutions to the Euler–Lagrange equations [Müller 1999]. The existence of these nonstandard solutions is not simply a mathematical curiosity but can be realized in practice as the blister and herringbone patterns in compressed thin sheets [Ortiz and Gioia 1994; Song et al. 2008], branched domain structures in ferromagnets [DeSimone et al. 2000], self-similar patterns in shape memory alloys [Bhattacharya 2003], the network of ridges in crumpled paper [Witten 2007], and even the fractal-like patterns in leaves and torn elastic sheets [Audoly and Boudaoud 2003; Sharon et al. 2007; Gemmer et al. 2016]. To understand such systems, local solutions of the Euler–Lagrange equations must be “patched together” along singularities in a manner that is consistent with the overall variational structure of the problem; see [Kohn 2007] for an introduction to this approach.

In this paper our focus is more modest. Namely, we study the problem of determining brachistochrone solutions for particles falling in an inverse-square gravitational field. This problem has been studied in [Parnovsky 1998; Tee 1999; Gemmer et al. 2011] using standard techniques from the calculus of variations. However, in these works they only considered “strong” solutions to the Euler–Lagrange equations, which limits the scope of the optimal paths considered. In particular, in [Parnovsky 1998; Tee 1999; Gemmer et al. 2011] it was shown that there is a “forbidden” region through which strong solutions to the Euler–Lagrange equations do not penetrate. In this paper, we show that by considering appropriate “weak” solutions constructed from strong solutions patched together at the singular origin of the gravitational field, the full space of optimal paths is more robust. In particular, these solutions enter the forbidden region and are characteristics for the Hamilton–Jacobi equation. This lends credence to the notion that our solutions are the natural extensions of the strong solutions that penetrate the forbidden region

and are optimal. Moreover, we also consider the inverse-square problem on an annular domain. Using variational inequalities, we show that our weak solutions are obtained in the limit as the inner radius of the annulus vanishes.

The paper is organized as follows. In Section 2 we outline the general framework of brachistochrone problems, present the strong and weak versions of the Euler–Lagrange equations, and draw a connection to geometric optics using results from optimal control theory. In Section 3 we restrict our focus to the case of the inverse-square gravitational field. We first briefly reproduce the results in [Parnovsky 1998; Tee 1999], namely that there exists a forbidden region through which strong solutions cannot penetrate. Next we present our construction of weak solutions that penetrate into this region. In Section 4 we take a pragmatic approach and consider the problem on an annular domain that excises the singularity at the origin. In doing so, we prove that under the assumption that the strong solutions are global minimizers outside of the forbidden region, our weak solutions are time-optimal. We conclude with a discussion section.

## 2. Mathematical framework and governing equations

**2.1. Mathematical framework.** In this section we summarize the essential definitions and equations which we use to study brachistochrone problems in generic settings. First, let  $V : \mathbb{R}^n \rightarrow \mathbb{R}$  be the potential for a gravitational field; i.e.,  $V$  is a smooth function except possibly at isolated singularities. Suppose  $A, B \in \mathbb{R}^n$  satisfy  $V(A) > V(B)$  and there exists a smooth curve  $\alpha : [0, 1] \rightarrow \mathbb{R}^n$  satisfying  $\alpha(0) = A$ ,  $\alpha(1) = B$  and  $V(\alpha(s)) \leq V(A)$  for all  $0 \leq s \leq 1$ . Now, for a particle released in the gravitational field and constrained to fall along  $\alpha$ , it follows that if friction is neglected, mechanical energy is conserved along the path

$$|\alpha'(s)|^2 \left( \frac{ds}{dt} \right)^2 + V(\alpha(s)) = V(A), \tag{1}$$

where  $t$  denotes time traveled on  $\alpha$  and we have absorbed the standard factor of  $1/2$  in the kinetic energy into the potential. The total time of flight to  $B$  can then be directly computed:

$$T[\alpha] = \int_0^1 \frac{|\alpha'(s)|}{\sqrt{V(A) - V(\alpha(s))}} ds. \tag{2}$$

This time of flight is still well defined if instead of smooth functions we consider *absolutely continuous functions* for which  $V(\alpha) \leq V(A)$ .<sup>1</sup> That is, we define the

---

<sup>1</sup>The space of absolutely continuous functions from  $[0, 1]$  into  $\mathbb{R}^n$  consists of all functions for which there exists a Lebesgue measurable function  $\beta : [0, 1] \rightarrow \mathbb{R}^n$  satisfying  $\alpha(s) = \alpha(0) + \int_0^s \beta(\bar{s}) d\bar{s}$  and is denoted by  $AC([0, 1]; \mathbb{R}^n)$  [Leoni 2009]. For  $\alpha \in AC([0, 1]; \mathbb{R}^n)$  the (weak) notion of the first derivative is defined by  $\alpha'(s) = \beta(s)$ .

admissible set  $\mathcal{A}$  by

$$\mathcal{A} = \left\{ \alpha \in AC([0, 1]; \mathbb{R}^n) : \alpha(0) = A, \alpha(1) = B \text{ and } V(\alpha(s)) \leq V(A) \right\} \quad (3)$$

and define the functional  $T : \mathcal{A} \rightarrow \mathbb{R}$  by (2). The generalized brachistochrone problem for the potential  $V$  is to find a curve  $\alpha^* \in \mathcal{A}$  that minimizes the time of flight to  $B$ . We call such curves *brachistochrone solutions* for the potential  $V$ .

The contours, i.e., the equipotential curves, of  $V$  naturally partition  $\mathbb{R}^n$  into domains

$$U(A) = \{x \in \mathbb{R}^n : V(A) - V(x) \geq 0\}$$

that contain points that (possibly) can be reached by brachistochrone solutions. For example, for the uniform gravitational potential  $V : \mathbb{R}^2 \rightarrow \mathbb{R}$ , defined by  $V(x, y) = -y$ , a particle released at the point  $A = (0, 0)$  can only reach points in the set  $U(A) = \{(x, y) \in \mathbb{R}^2 : y \leq 0\}$ . To completely solve the brachistochrone problem for this potential, one is naturally led to the question of finding all brachistochrone solutions that foliate  $U(A)$ .

**2.2. Euler–Lagrange equations for brachistochrone problems.** We now follow classical techniques presented in [Sagan 1969] to derive the Euler–Lagrange equations for (2). First, suppose  $\inf_{\alpha \in \mathcal{A}} T[\alpha] < \infty$  and  $\alpha^* \in \mathcal{A}$  satisfies  $T[\alpha^*] = \inf_{\alpha \in \mathcal{A}} T[\alpha]$ , i.e.,  $\alpha^*$  is a minimizer. Since  $\alpha^* \in AC([0, 1]; \mathbb{R}^n)$  and  $T[\alpha^*] < \infty$ , the set of points in  $[0, 1]$  for which  $V(\alpha^*(s)) = V(A)$  has Lebesgue measure zero. Consequently, if we further assume that  $V(\alpha^*(s)) = V(A)$  only at  $s = 0$  and possibly at  $s = 1$  if the terminal point satisfies  $V(B) = V(A)$ , then for all  $\eta \in C_0^\infty([0, 1]; \mathbb{R}^n)^2$  there exists  $\bar{h} > 0$  such that  $|h| < \bar{h}$  implies  $\alpha^* + h\eta \in \mathcal{A}$ . Define the function  $f : [-\bar{h}, \bar{h}] \rightarrow \mathbb{R}$  by  $f(x) = T[\alpha^* + x\eta]$ . From the regularity assumptions on  $V$  and  $\alpha^*$ , it follows that  $f$  is differentiable in  $h$  and consequently, since  $\alpha^*$  minimizes  $T$ , it follows that  $f'(0) = 0$ . Therefore, we have the following necessary condition for optimality:

$$f'(0) = \int_0^1 \frac{\alpha^{*'}(s)}{|\alpha^{*'}(s)|\sqrt{V(A) - V(\alpha^*(s))}} \cdot \eta'(s) ds + \frac{1}{2} \int_0^1 \frac{|\alpha^{*'}(s)|}{(V(A) - V(\alpha^*(s)))^{3/2}} \nabla V(\alpha^*(s)) \cdot \eta ds, \quad (4)$$

which must be satisfied for all  $\eta \in C_0^\infty([0, 1]; \mathbb{R}^n)$  [Evans 1998]. Equation (4) is known as the *weak formulation of the Euler–Lagrange equations for the brachistochrone problem*. If we further assume that the minimizing curve  $\alpha^*$  is twice

---

<sup>2</sup>  $C_0^\infty([0, 1]; \mathbb{R}^n)$  denotes the space of smooth functions from  $[0, 1]$  into  $\mathbb{R}^n$  with compact support [Royden and Fitzpatrick 2010].

differentiable then (4) can be integrated by parts to yield

$$0 = \int_0^1 \left( \frac{1}{2} \frac{|\alpha^{*'}(s)|}{(V(A) - V(\alpha^*(s)))^{3/2}} \nabla V(\alpha^*(s)) - \frac{d}{ds} \left( \frac{\alpha^{*'}(s)}{|\alpha^{*'}(s)| \sqrt{V(A) - V(\alpha^*(s))}} \right) \right) \cdot \eta \, ds. \quad (5)$$

By the so-called “fundamental theorem of the calculus of variations”, since  $\eta$  was arbitrary the necessary condition satisfied by a twice differentiable curve  $\alpha^*$  is the following differential equation [Sagan 1969]:

$$0 = \frac{1}{2} \frac{|\alpha^{*'}(s)|}{(V(A) - V(\alpha^*(s)))^{3/2}} \nabla V(\alpha^*(s)) - \frac{d}{ds} \left( \frac{\alpha^{*'}(s)}{|\alpha^{*'}(s)| \sqrt{V(A) - V(\alpha^*(s))}} \right). \quad (6)$$

Equation (6) is known as the *strong formulation of the Euler–Lagrange equations for the brachistochrone problem*.

Note, however, that in deriving the strong formulation of the Euler–Lagrange equations we made the additional assumption that  $\alpha^*$  is twice differentiable. In many applications this assumption is too restrictive. For example, the functional  $J : AC([-1, 1]; \mathbb{R}) \rightarrow \mathbb{R}$  defined by  $J[y] = \int_{-1}^1 (1 - y'(s)^2)^2 \, ds$  with boundary conditions  $y(-1) = y(1) = 1$  is minimized by  $y(x) = |x|$ . In this example the two strong solutions  $y = x$  and  $y = -x$  are joined together at  $x = 0$ . However simply gluing together two strong solutions does not guarantee that the resulting combination is a weak solution. If  $\alpha^*$  is twice differentiable everywhere except at a point  $c \in (0, 1)$  and satisfies (6) away from  $c$ , we can integrate (4) by parts to obtain the necessary condition

$$\lim_{s \rightarrow c^-} \left( \frac{\alpha^{*'}(s)}{|\alpha^{*'}(s)| \sqrt{V(A) - V(\alpha(s))}} \right) = \lim_{s \rightarrow c^+} \left( \frac{\alpha^{*'}(s)}{|\alpha^{*'}(s)| \sqrt{V(A) - V(\alpha(s))}} \right). \quad (7)$$

Equation (7) is commonly called the *Weierstrass–Erdmann corner condition* [Sagan 1969] and must be satisfied by piecewise smooth solutions of (6).

We now make some additional comments about the Weierstrass–Erdmann corner conditions which will be relevant to the discussion in later sections. First, away from a singularity in the potential  $V$ , i.e., if we assume that  $(V(A) - V(c))^{-1/2} \neq 0$ , (7) corresponds to continuity of the tangent vector  $\alpha'(s)$  at  $c$ . Moreover, away from singularities, this condition physically corresponds to conservation of classical momentum at  $c$ . However, if  $(V(A) - V(c))^{-1/2} = 0$ , this necessary condition is trivially satisfied. That is, at a singularity in the gravitational field, a minimizer could violate conservation of momentum. This result should not be surprising since at a singularity  $V(c) = \infty$ , implying that the instantaneous speed, as well as the acceleration of the particle, is infinite. This fact will be critical in our later construction of weak brachistochrone solutions in an inverse-square gravitational field.

**2.3. Connection with geometrical optics through control theory.** While directly solving the Euler–Lagrange equations given by (6) will solve the brachistochrone problem, there is another approach, originally taken by Johann Bernoulli. Namely, Bernoulli realized that the brachistochrone problem is equivalent to finding the path traced out by a ray of light in a medium with index of refraction  $n(\mathbf{x}) = (V(A) - V(\mathbf{x}))^{-1/2}$ . His solution method was prescient in that it applied Snell’s law of refraction to what would now be called a finite element approximation to the problem with a piecewise linear basis [Erlichson 1999; Sussmann and Willems 1997]. The connection to geometrical optics was later exploited by Hamilton and finalized by Jacobi to derive what we now call the Hamilton–Jacobi equations for a variational problem [Broer 2014; Sussmann and Willems 1997; Nakane and Fraser 2002]. Specifically, the Hamilton–Jacobi equation is a quasilinear partial differential equation whose characteristic equations are precisely the Euler–Lagrange equations for the system [Evans 1998]. In particular, the Hamilton–Jacobi equation governs the dynamics of wave-fronts propagating in a medium with index of refraction  $n(\mathbf{x})$  and the Euler–Lagrange equations are the evolution equations for the normals to the wave-fronts.

We will now show how the geometric optics interpretation of the brachistochrone problem can be directly derived using modern optimal control theory. To reinterpret the brachistochrone problem as a control problem we follow [Sussmann and Willems 1997] and first define the set of admissible controls by

$$\mathcal{U} = \{ \mathbf{u} : [0, \mathbb{R}) \rightarrow \mathbb{R}^n : \mathbf{u} \text{ is piecewise smooth and } |\mathbf{u}| = 1 \}, \quad (8)$$

and to satisfy (1) we constrain the dynamics of the system by

$$\dot{\alpha} = \sqrt{V(A) - V(\mathbf{x})} \mathbf{u}. \quad (9)$$

We define the *trajectory of a control* to be the curve  $\alpha$  defined by (9) and also define  $\mathcal{T}_B : \mathcal{U} \rightarrow \mathbb{R} \cup \{+\infty\}$  to be the first time a trajectory corresponding to a control  $\mathbf{u}$  reaches the point  $B$ . The optimal control problem corresponding to the brachistochrone problem is to find  $\mathbf{u}^* \in \mathcal{U}$  that steers a trajectory  $\alpha(t)$  to a point  $B \in U(A)$  in the minimal amount of time. That is, find  $\mathbf{u}^* \in \mathcal{U}$  such that  $T_B[\mathbf{u}^*] = \inf_{\mathbf{u} \in \mathcal{U}} T_B[\mathbf{u}]$ . Clearly, this optimal control problem is equivalent to our previous formulation of the brachistochrone problem and  $\inf_{\mathbf{u} \in \mathcal{U}} T_B[\mathbf{u}] = \inf_{\alpha \in \mathcal{A}} T[\alpha]$ .

One technique for solving such an optimal control problem is to apply Bellman’s technique of dynamic programming [Bertsekas 1995]. Namely, if we define the *value function*  $\mathcal{V} : U(A) \rightarrow \mathbb{R}$  by

$$\mathcal{V}(\mathbf{x}) = \inf_{\mathbf{u} \in \mathcal{U}} \mathcal{T}_{\mathbf{x}}[\mathbf{u}] \quad (10)$$

then the dynamic programming principle states that for  $\Delta t > 0$  sufficiently small

$$\mathcal{V}(\mathbf{x}) = \min_{\substack{\mathbf{u} \in \mathcal{U} \\ 0 < s < \Delta t}} \{ \mathcal{V}(\bar{\alpha}(\Delta t)) + \Delta t \}, \quad (11)$$



where  $\bar{\alpha}$  corresponds to the *time-reversed* solution of (9) with initial condition  $\bar{\alpha}(0) = \mathbf{x}$  and control  $\mathbf{u}$ . If we assume that  $\mathcal{V}$  is smooth, we can formally Taylor expand:

$$\mathcal{V}(\mathbf{x}) = \min_{|\mathbf{u}(0)|=1} \left\{ \mathcal{V}(\mathbf{x}) + \nabla \mathcal{V}(\mathbf{x}) \sqrt{V(A) - V(\mathbf{x})} \mathbf{u}(0) \Delta t + \Delta t + \mathcal{O}(\Delta t^2) \right\}.$$

Consequently, taking the limit as  $\Delta t \rightarrow 0$ , we obtain

$$-1 = \min_{|\mathbf{u}(0)|=1} \nabla \mathcal{V}(\mathbf{x}) \sqrt{V(A) - V(\mathbf{x})} \mathbf{u}(0).$$

Finally, this minimum is obtained by  $\mathbf{u}(0) = -\nabla \mathcal{V}(\mathbf{x}) / |\nabla \mathcal{V}(\mathbf{x})|$  and hence we can conclude that the value function  $\mathcal{V}$  satisfies the partial differential equation

$$|\nabla \mathcal{V}|^2 = \frac{1}{V(A) - V(\mathbf{x})} = n^2(\mathbf{x}). \tag{12}$$

In geometrical optics, (12) is an eikonal equation for a medium with index of refraction  $n(\mathbf{x})$ . That is, the level sets of solutions to (12) correspond to wave fronts for light traveling through the medium and the light rays correspond to curves that are everywhere tangent to the normals of the level sets of  $\mathcal{V}$ . Consequently, if we let  $\beta(s)$  be an arc-length parametrization of such a light ray, it follows that

$$\nabla \mathcal{V}(\beta(s)) = n(\beta(s)) \frac{d\beta}{ds}. \tag{13}$$

Differentiating with respect to  $ds$  and switching the order of differentiation,

$$\nabla \left( \nabla \mathcal{V}(\beta(s)) \cdot \frac{d\beta}{ds} \right) = \frac{d}{ds} \left( n(\beta(s)) \frac{d\beta}{ds} \right).$$

Therefore, by (12) and (13), the governing equation for the rays is

$$\nabla n(\beta) = \frac{d}{ds} \left( n(\beta(s)) \frac{d\beta}{ds} \right). \tag{14}$$

Finally, it follows immediately that (14) is simply a version of (6) that is parametrized by arc-length. That is, to solve the brachistochrone problem we could, in principle, solve the eikonal equation (12) and use (13) to reconstruct the brachistochrone solution. More importantly, we could solve the brachistochrone problem directly by solving the Euler–Lagrange equations (6) and compute the time of flight along these solution curves to compute the solution to the eikonal equation (12).

### 3. Brachistochrone problem in an inverse-square gravitational field

**3.1. Framework.** Consider two points  $A, B \in \mathbb{R}^2$  with  $|B| \leq |A|$ . For an inverse-square field, the potential  $V : \mathbb{R}^2 \rightarrow \mathbb{R}$  is defined by  $V(x) = -|x|^{-2}$ . The inverse-square brachistochrone problem is to construct a curve connecting  $A$  and  $B$  such

that a particle traversing the curve from  $A$  to  $B$  under the influence a gravitational field centered at the origin with potential  $V$  has the least time of flight. In this case the admissible set  $\mathcal{A}_0$  is defined by

$$\mathcal{A}_0 = \{ \alpha \in AC([0, 1]; \mathbb{R}^2) : \alpha(0) = A, \alpha(1) = B \text{ and } \forall t \in [0, 1], |\alpha(t)| \leq |A| \} \quad (15)$$

and the time of flight  $T : \mathcal{A}_0 \rightarrow \mathbb{R}^+$  is given by

$$T[\alpha] = \int_0^1 \frac{|\alpha'(s)|}{\sqrt{|\alpha(s)|^{-2} - |A|^{-2}}} ds. \quad (16)$$

Again, this functional arises from classical conservation of mechanical energy and the constraint  $|\alpha(s)| \leq |A|$  — a necessary condition for this functional to be well defined — is equivalent to the condition that the particle cannot gain mechanical energy.

To study minimizers of (16) it is natural to work in a polar-coordinate representation of the form

$$\alpha(s) = (r(s) \cos(\theta(s)), r(s) \sin(\theta(s))), \quad (17)$$

where  $r : [0, 1] \rightarrow [0, |A|]$  and  $\theta : [0, 1] \rightarrow [-\pi, \pi]$  are (weakly) differentiable functions satisfying  $r(0) = |A|$ ,  $r(1) = |B|$ ,  $\theta(0) = \theta_0$ ,  $\theta(1) = \theta_f$ , with  $\theta_0, \theta_f$  the angular coordinates of  $A, B$  respectively; see Figure 1 (left). By rotational symmetry and radial invariance we can assume without loss of generality that  $A = (1, 0)$ ; see Figure 1 (right). In this representation, (16) becomes

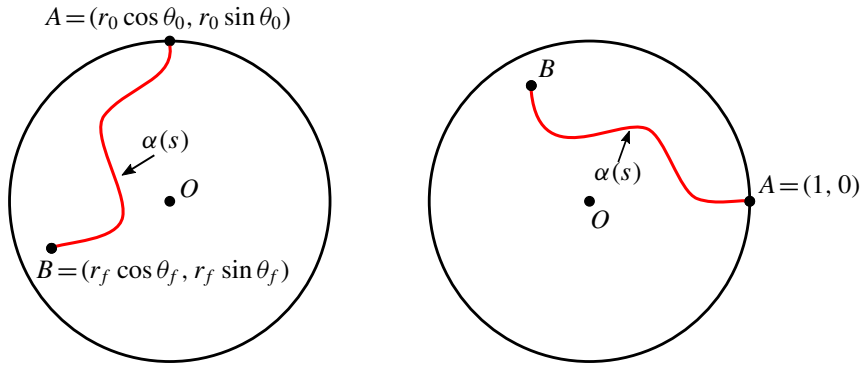
$$T[r, \theta] = \int_0^1 \sqrt{\frac{r'(s)^2 + r(s)^2 \theta'(s)^2}{r(s)^{-1} - 1}} ds = \int_0^1 L_2(r(s), r'(s), \theta'(s)) ds, \quad (18)$$

where  $L_2 : \mathbb{R}^3 \rightarrow \mathbb{R}$  denotes the Lagrangian for this functional. To reduce encumbering notation we write  $(r(s), \theta(s)) \in \mathcal{A}_0$  as a proxy for the statement that there exists  $\alpha \in \mathcal{A}_0$  with corresponding radial and angular components  $r(s)$  and  $\theta(s)$  respectively.

We now deduce geometric properties of minimizers using the structure of the Lagrangian. We first show that if  $(r^*, \theta^*) \in \mathcal{A}_0$  minimizes  $T$  then  $\theta^*$  must be a monotone function. This property prevents a minimizer from “turning back” to its starting point. The idea of the proof is to construct for all  $(r, \theta) \in \mathcal{A}_0$  a modified curve  $(r, \bar{\theta}) \in \mathcal{A}$  with  $\bar{\theta}$  monotone in  $s$  and show  $T[r, \bar{\theta}] \leq T[r, \theta]$ .

**Proposition 1.** *If  $(r^*(s), \theta^*(s)) \in \mathcal{A}_0$  minimizes  $T$  then  $\theta^*$  is monotone in  $s$ .*

*Proof.* Let  $(r(s), \theta(s)) \in \mathcal{A}_0$  terminate at the point  $(r_f \cos \theta_f, r_f \sin \theta_f)$  and assume  $\theta_f \geq 0$ . Define  $\bar{\theta} : [0, 1] \rightarrow [0, 2\pi]$  by  $\bar{\theta}(s) = \min\{\theta_f, \sup\{\theta(t) : 0 \leq t \leq s\}\}$ . From the absolute continuity of  $\theta$ , it follows that  $\bar{\theta}$  is absolutely continuous, monotone



**Figure 1.** Left: Plot of a curve  $\alpha : [0, 1] \rightarrow \mathbb{R}^2$  connecting  $A = (r_0 \cos \theta_0, r_0 \sin \theta_0)$  to  $B = (r_f \cos \theta_f, r_f \sin \theta_f)$  in an inverse-square gravitational field centered at the origin  $O$ . The circle of radius  $r_0$  is an equipotential for the inverse-square gravitational field. For a particle falling along this curve, conservation of mechanical energy requires that  $|\alpha(s)| \leq r_0$ . Right: By rotational and radial scale invariance of this problem, we can assume without loss of generality that  $A = (1, 0)$ .

increasing and satisfies  $\bar{\theta}(1) = \theta_f$ . Moreover, there exists a closed set  $I$  on which  $\bar{\theta} = \theta$  and an open set  $\bar{I} = [0, 1] \setminus I$  on which  $d\bar{\theta}/ds = 0$ . Therefore,

$$\begin{aligned} T[r, \bar{\theta}] &= \int_I \sqrt{\frac{r'(s)^2 + r(s)^2 \theta'(s)^2}{r(s)^{-1} - 1}} ds + \int_{\bar{I}} \sqrt{\frac{r'(s)^2}{r(s)^{-1} - 1}} ds \\ &\leq \int_I \sqrt{\frac{r'(s)^2 + r(s)^2 \theta'(s)^2}{r(s)^{-1} - 1}} ds + \int_{\bar{I}} \sqrt{\frac{r'(s)^2 + r(s)^2 \theta'(s)^2}{r(s)^{-1} - 1}} ds = T[r, \theta], \end{aligned}$$

with equality if and only if  $\theta$  is monotone increasing. Thus, if  $(r^*(s), \theta^*(s)) \in \mathcal{A}_0$  minimizes  $T$  then  $\theta^*$  is monotone increasing in  $s$ . A similar argument proves that  $\theta^*$  must be monotone decreasing if  $\theta_f < 0$ . □

We now prove that without loss of generality we can assume minimizers are symmetric about the angle  $\theta_f/2$ . Specifically, if in polar coordinates  $(r^*(s), \theta^*(s)) \in \mathcal{A}_0$  minimizes  $T$  and terminates at the final point  $(r_f = 1, \theta_f)$  then the image of  $(r(s), \theta(s))$  is symmetric about the line  $\theta = \theta_f/2$ . Similar to the previous proof, the idea is to modify a curve  $\alpha \in \mathcal{A}_0$  by constructing symmetric versions and comparing the times of flight.

**Proposition 2.** *If  $(r^*(s), \theta^*(s)) \in \mathcal{A}_0$  minimizes  $T$  with terminal point  $(r(1), \theta(1)) = (1, \theta_f)$  then there exists a version of  $(r^*(s), \theta^*(s))$  in  $\mathbb{R}^2$  that is symmetric about the line  $\theta = \theta_f/2$  that also minimizes  $T$ .*

*Proof.* Let  $(r(s), \theta(s)) \in \mathcal{A}_0$  and  $t(s)$  be a reparameterization in which  $\theta(1/2) = \theta_f/2$  and if  $t > 1/2$  then  $\theta(t) > \theta_f/2$ . Define the two possible reflections of  $(r(t), \theta(t))$  about the line  $\theta = \theta_f/2$  by

$$\begin{aligned}
 r_1(t) &= \begin{cases} r(t), & 0 \leq t \leq \frac{1}{2}, \\ r(1-t), & \frac{1}{2} < t \leq 1, \end{cases} & \text{and} & \theta_1(t) = \begin{cases} \theta(t), & 0 \leq t \leq \frac{1}{2}, \\ \theta_f - \theta(1-t), & \frac{1}{2} < t \leq 1, \end{cases} \\
 r_2(t) &= \begin{cases} r(1-t), & 0 \leq t \leq \frac{1}{2}, \\ r(t), & \frac{1}{2} < t \leq 1, \end{cases} & \text{and} & \theta_2(t) = \begin{cases} \theta(1-t) - \theta_f, & 0 \leq t \leq \frac{1}{2}, \\ \theta(t), & \frac{1}{2} < t \leq 1. \end{cases}
 \end{aligned}$$

By construction, the images of the curves  $(r_1(t), \theta_1(t))$  and  $(r_2(t), \theta_2(t))$  in  $\mathbb{R}^2$  are symmetric about the line  $\theta = \theta_f/2$ . It follows from symmetry that

$$\begin{aligned}
 T[r_1, \theta_1] &= 2 \int_0^{1/2} \sqrt{\frac{r'(s)^2 + r(s)^2 \theta'(s)^2}{r(s)^{-1} - 1}} ds, \\
 T[r_2, \theta_2] &= 2 \int_{1/2}^1 \sqrt{\frac{r'(s)^2 + r(s)^2 \theta'(s)^2}{r(s)^{-1} - 1}} ds.
 \end{aligned}$$

Therefore,  $T[r_1, \theta_1] + T[r_2, \theta_2] = 2T[r, \theta]$  from which it follows that

$$\min\{T[r_1, \theta_1], T[r_2, \theta_2]\} \leq T[r, \theta].$$

Thus if  $(r^*(s), \theta^*(s)) \in \mathcal{A}_0$  minimizes  $T$  then either  $(r_1(s), \theta_1(s))$  or  $(r_2(s), \theta_2(s))$  must also minimize  $T$ . □

**3.2. Strong solutions to Euler–Lagrange equations.** In this subsection we review the construction of smooth minimizers to  $T$  that was originally presented in [Parnovsky 1998; Tee 1999]. The classic method for finding time-minimizing curves is to derive the Euler–Lagrange equations for  $T$  and solve the resulting boundary value problem. Specifically, if we assume there exists a twice differentiable curve  $\alpha^*(s) \in \mathcal{A}_0$  with angular component  $\theta^*(s)$  and radial component  $r^*(s)$  that (locally) minimizes  $T$  then the resulting boundary value problem is

$$\begin{cases} \left(\frac{\partial L}{\partial r} - \frac{d}{ds} \frac{\partial L}{\partial r'}\right) \Big|_{(r^*(s), \theta^*(s))} = 0, \\ \frac{d}{ds} \frac{\partial L}{\partial \theta'} \Big|_{(r^*(s), \theta^*(s))} = 0, \\ r^*(0) = 1, r^*(1) = |B|, \theta^*(0) = 0, \theta(1) = \theta_f. \end{cases} \tag{19}$$

If we make the assumption that  $\theta^*$  is a function of  $r^*$ , i.e., we assume the ansatz  $r^*(s) = (|B| - 1)s + 1$ , then (19) reduces to the differential equation

$$\frac{d}{dt} \frac{\partial L}{\partial \theta'} = 0. \tag{20}$$

Formally, (20) can be integrated to yield a separable differential equation with solution

$$\theta^*(r^*) = \pm \int_1^{r^*} \sqrt{\frac{2(1/u - 1)D}{u^4 - u^2(2(1/u - 1))D}} du, \tag{21}$$

where  $D > 0$  is a constant of integration which can be determined from the boundary conditions. The assumption that  $\theta^*$  is globally a function of  $r^*$  is valid if  $(d\theta^*/dr^*)^{-1} \neq 0$  for all  $r^* \in (0, 1)$ . It follows from (21) that for fixed  $D > 0$  this condition is equivalent to the nonexistence of solutions to the equation  $r^3 + 2Dr - 2D = 0$  for  $r \in (0, 1]$ . The following proposition makes this statement precise and identifies the critical radius  $r_c$  in terms of the integration constant  $D$ .

**Proposition 3.** *For fixed  $D > 0$ , there exists a unique  $r_c(D) \in [0, 1]$  such that  $\theta^*(r^*)$  defined by (21) satisfies*

$$\lim_{r \rightarrow r_c(D)^+} \left. \frac{d\theta^*}{dr^*} \right|_r = \infty.$$

Moreover, the mapping  $D \rightarrow r_c(D)$  is a bijection from  $(0, \infty)$  into  $(0, 1)$ .

*Proof.* Fix  $D > 0$  and define  $g : (0, 1) \rightarrow \mathbb{R}$  by  $g(r) = r^3 + 2Dr - 2D$ . Since  $g(0) = -2D$ ,  $g(1) = 1$ , and  $g'(r) > 0$ , by the intermediate value theorem there exists unique  $r_c(D) \in (0, 1)$  such that  $g(r_c(D)) = 0$ . Consequently, it follows from (21) that

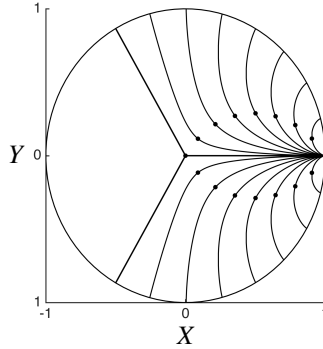
$$\lim_{r \rightarrow r_c(D)^+} \left. \frac{\partial \theta^*}{\partial r^*} \right|_r = \infty.$$

The bijection is proved by noting that the inverse mapping from  $r_c(D)$  to  $D$  given by  $D(r_c) = (r_c^3/2)(1 - r_c)$  satisfies  $\lim_{r_c \rightarrow 0^+} D(r_c) = 0$ ,  $\lim_{r_c \rightarrow 1^-} D(r_c) = \infty$  and is monotone increasing in  $r_c$ . □

To extend smooth solutions beyond the point where  $\theta^*$  is no longer a function of  $r^*$  it follows from Proposition 2 that it is necessary to reflect the solutions across the line  $\theta = \theta_f/2$ . Specifically, for  $D \in (0, \infty)$  if we define  $r_c(D)$  as in Proposition 3 then we obtain the following family of solutions expressed in parametric form  $\alpha(s) = (r_D^S(s) \cos(\theta_D^S(s)), r_D^S(s) \sin(\theta_D^S(s)))$  with

$$r_D^S(s) = \begin{cases} 2(r_c(D) - 1)s + 1, & 0 \leq s \leq \frac{1}{2}, \\ 2(1 - r_c(D))(s - 1) + 1, & \frac{1}{2} < s \leq 1, \end{cases} \tag{22}$$

$$\theta_D^S(s) = \begin{cases} \pm \int_1^{r_D^S(s)} \sqrt{\frac{2(1-u)D}{u^5 - 2u^2D(1-u)}} du, & 0 \leq s \leq \frac{1}{2}, \\ \mp \int_{r_c(D)}^{r_D^S(s)} \sqrt{\frac{2(1-u)D}{u^5 - 2u^2D(1-u)}} du \pm \theta_D^S\left(\frac{1}{2}\right), & \frac{1}{2} \leq s \leq 1, \end{cases} \tag{23}$$



**Figure 2.** Plot of 16 smooth strong solution curves  $\alpha(s)$  defined by (22) and (23) with the final angular coordinate  $\theta_D^S(1)$  uniformly spaced from  $-\pi/3$  to  $\pi/3$ . The value of  $D$  was found by the bisection method, i.e., the shooting method, applied to (23). The points indicate the critical radius  $r_c(D)$  where  $d\theta_D^S/dr_D^S = \pm\infty$  and the curve begins receding away from the origin.

where we are using the superscript “S” to denote that these are strong solutions to the Euler–Lagrange equations. Note, that while the individual functions  $r_D^S(s)$  and  $\theta_D^S(s)$  are not smooth, the curve  $\alpha$  itself is a smooth function from  $[0, 1]$  into  $\mathbb{R}^2$ .

Interestingly, as  $D$  ranges over values in  $(0, \infty)$  the curves defined by (22) and (23) do not foliate the unit disk  $x^2 + y^2 \leq 1$ ; see Figure 2. Indeed, if we define the sector  $S$  by

$$S = \{\theta : -\pi \leq \theta < -2\pi/3\} \cup \{\theta : 2\pi/3 < \theta < \pi\}, \tag{24}$$

it was shown in [Tee 1999; Gemmer et al. 2011] that these curves do not enter  $S$ . This is made precise by the following proposition whose proof we adapt from [Gemmer et al. 2011].

**Proposition 4.** *For all  $s \in [0, 1]$  and  $D \in (0, \infty)$  the curves  $\alpha(s)$  with radial and angular components  $r_D^S(s)$  and  $\theta_D^S(s)$  defined by (22) and (23) satisfy  $\theta_D^S(s) \notin S$ .*

*Proof.* For  $D > 0$  let  $\alpha(s) = (r_D^S(s) \cos(\theta_D^S(s)), r_D^S(s) \sin(\theta_D^S(s)))$  be defined by (22) and (23) with the “–” branch. Differentiating,

$$\frac{d\theta_D^S}{ds} = \frac{d\theta_D^S}{dr_D^S} \frac{dr_D^S}{ds} = 2r_c(D) \sqrt{\frac{2(1-r_D^S)D}{(r_D^S)^5 - 2(r_D^S)^2 D(1-r_D^S)}} > 0$$

with equality only at  $r_D^S = 1$ . Hence,  $d\theta_D^S/ds$  is monotone increasing in  $s$  with the maximum angular coordinate  $\bar{\theta}(D)$  satisfying

$$\bar{\theta}(D) = \max_{0 \leq s \leq 1} \theta_D^S(s) = 2 \int_{r_c(D)}^1 \sqrt{\frac{2(1-u)D}{u^5 - 2u^2 D(1-u)}} du.$$

Since  $\lim_{D \rightarrow \infty} r_c(D) = 1$ , it follows that  $\lim_{D \rightarrow \infty} \bar{\theta}(D) = 0$ . Now, from uniqueness of solutions to (20) we can deduce that  $\bar{\theta}(D)$  must be monotone decreasing in  $D$  and hence has a limit as  $D \rightarrow 0$ . By making the change of variables  $x = r_c(D)/u^{3/2}$ , we obtain

$$\begin{aligned} \frac{1}{2} \lim_{D \rightarrow 0} \bar{\theta}(D) &= \lim_{D \rightarrow 0} \frac{2}{3} \int_{r_c(D)}^1 \sqrt{\frac{1-r_c(D)x^{-2/3}}{(1-r_c(D))-(1-r_c(D)x^{-2/3})x^2}} dx \\ &= \lim_{D \rightarrow 0} \frac{2}{3} \int_0^1 \sqrt{\frac{1}{(1-r_c(D))/(1-r_c(D)x^{-2/3})-x^2}} \mathbb{I}\{x > r_c(D)^{3/2}\} dx, \end{aligned}$$

where  $\mathbb{I}$  denotes the standard indicator function. Now, observing that the integrand of the above equation forms a sequence of functions bounded by  $(1-x^2)^{-1/2}$ , it follows from Lebesgue’s dominated convergence theorem that

$$\lim_{D \rightarrow 0} \bar{\theta}(D) = \int_0^1 \frac{4}{3} \sqrt{\frac{1}{1-x^2}} dx = \frac{4}{3}(\arcsin(1) - \arcsin(0)) = \frac{2\pi}{3}.$$

The exact same arguments hold if we consider the “+” branch in (22) and (23) except the limiting angle is  $-2\pi/3$ . Consequently we can conclude for all  $s \in [0, 1]$  and  $D \in (0, \infty)$  that  $\theta_D^S(s)$  given by (23) satisfies

$$-\frac{2\pi}{3} \leq \theta_D^S(s) \leq \frac{2\pi}{3}. \quad \square$$

The following proposition immediately follows from Propositions 3 and 4.

**Proposition 5.** *For all  $\theta_f \in (0, 2\pi/3)$  there exists  $D \in (0, \infty)$  such that the solution curve with angular and radial coordinates  $(\theta_D^S(s), r_D^S(s))$  as defined by (22) and (23) satisfies  $\theta_D^S(0) = 0$ , and  $\theta_D^S(1) = \theta_f$ . Moreover, the mapping  $\theta_f \rightarrow D$  is a bijection.*

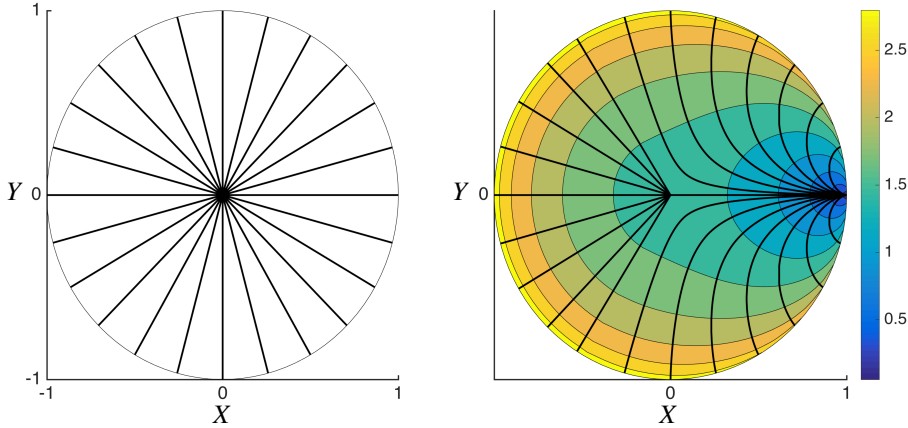
**Remark.** It follows from Propositions 3 and 4 that  $\theta_f, r_c$  and  $D$  characterize a unique solution curve of the form defined by (22) and (23).

**3.3. Weak solutions to the Euler–Lagrange equations.** One family of weak solutions to (16) is given by the parametrization

$$r^W(s) = \begin{cases} 1 - 2s, & 0 \leq s \leq \frac{1}{2}, \\ 2s - 1, & \frac{1}{2} < s \leq 1, \end{cases} \tag{25}$$

$$\theta_{\theta_f}^W(s) = \begin{cases} 0, & 0 \leq s \leq \frac{1}{2}, \\ \theta_f, & \frac{1}{2} \leq s \leq 1, \end{cases} \tag{26}$$

where the superscript “W” is used to denote that these are weak solutions to the Euler–Lagrange equations. This parametrization indeed satisfies the corner



**Figure 3.** Left: Plot of weak solution curves defined by (25) and (26) with final angular coordinate  $\theta_f$  spaced evenly from  $-\pi$  to  $\pi$ . Right: Foliation of the unit disk by weak solution curves in  $S$  and classic solutions outside of  $S$ . Beneath the solution curves is a contour plot of the time of flight calculated along the solution curves. The contours correspond to level sets of the value  $\mathcal{V}$  satisfying the eikonal equation defined by (12).

condition given by (7):

$$\begin{aligned} & \lim_{s \rightarrow c^-} \left( \frac{\alpha^{*'}(s)}{|\alpha^{*'}(s)| \sqrt{V(A) - V(\alpha(s))}} \right) - \lim_{s \rightarrow c^+} \left( \frac{\alpha^{*'}(s)}{|\alpha^{*'}(s)| \sqrt{V(A) - V(\alpha(s))}} \right) \\ &= (1, 0) \lim_{s \rightarrow 1/2^+} \left( \sqrt{\frac{r^W(s)}{r^W(s) - 1}} \right) + (\cos \theta_f, \sin \theta_f) \lim_{s \rightarrow 1/2^-} \left( \sqrt{\frac{r^W(s)}{r^W(s) - 1}} \right) = (0, 0). \end{aligned}$$

It is important to note that as it is defined,  $\theta_{\theta_f}^W(s)$  is not weakly differentiable. Specifically,  $\theta_{\theta_f}^W(s)$  is only differentiable in the distributional sense with a derivative given by a delta mass centered at  $s = 1/2$ . However, this is only an artifact of the  $r = 0$  coordinate singularity for polar coordinates and the curve  $\alpha^W(s)$  itself is weakly differentiable. Moreover, for  $s < 1/2$  this curve is simply the solution curve given by (22) and (23) with  $D = 0$  and the weak solution is constructed by joining appropriately rotated copies of this strong solution at the origin.

The family of solutions to the Euler–Lagrange equations defined by (25) and (26) completely foliate the unit disk; see Figure 3 (left). Hence these solution curves are natural candidates for time minimizers that enter the sector  $S$ . In Figure 3 (right) we plot the unit disk foliated by a combination of strong and weak solutions to the Euler–Lagrange equations. More specifically, for a given  $\theta_f$  we use (22) and (23) or (25) and (26) depending on whether  $|\theta_f| > 2\pi/3$ . The contour beneath the curves in corresponds to the time of flight computed along the solution curves and



confirms our intuition that the classic solutions have shorter time of flight outside of  $S$ . Notice that the contours in Figure 3 (right) are smooth and intersect the strong and weak solution curves orthogonally as expected from (13). Moreover, the value function  $\mathcal{V}$  defined in Section 2 is a solution to the eikonal equation defined by (12).

#### 4. Constrained inverse-square brachistochrone problem

**4.1. Variational inequality.** In the previous section we solved the inverse-square brachistochrone problem using a combination of weak and strong extremizers. However, the solutions are impractical in that, as a consequence of the singular gravitational field, a particle following along an extremizer will experience infinite acceleration at the origin. To alleviate this problem we now consider a modified version of the inverse-square brachistochrone problem that restricts the radial coordinate to remain bounded away from the origin. Specifically, for  $\epsilon > 0$  we define the annulus  $\mathcal{O}_\epsilon = \{(x, y) \in \mathbb{R}^2 : \sqrt{x^2 + y^2} \geq \epsilon\}$  and consider the problem of minimizing  $T$  over the admissible set  $\mathcal{A}_\epsilon \subset \mathcal{A}_0$  defined by

$$\mathcal{A}_\epsilon = \{(r(s), \theta(s)) \in \mathcal{A}_0 : (r(s) \cos(\theta(s)), r(s) \sin(\theta(s))) \in \mathcal{O}_\epsilon \text{ for } s \in [0, 1]\}. \tag{27}$$

This formulation of the problem is equivalent to an ‘‘obstacle problem’’ with the obstacle being the circle of radius  $\epsilon$  centered at the origin.

To derive necessary conditions satisfied by minimizers of  $T$  over  $\mathcal{A}_\epsilon$  we follow [Evans 1998, Chapter 8, Section 4] and derive a variational inequality that plays the role of the Euler–Lagrange equations. First, suppose  $\alpha^*(s) \in \mathcal{A}_\epsilon$  is the global minimizer of  $T$  over  $\mathcal{A}_\epsilon$  with radial and angular components  $r^*(s)$  and  $\theta^*(s)$  respectively. Letting  $\beta(s) \in \mathcal{A}_\epsilon$  with radial and angular components  $q(s)$  and  $\theta^*(s)$  respectively, it follows from the convexity of  $\mathcal{A}_\epsilon$  that for all  $\lambda \in [0, 1]$  the curve  $\gamma(s)$  with radial component  $r^*(s) + \lambda(q(s) - r^*(s))$  and angular component  $\theta^*(s)$  also satisfies  $\gamma(s) \in \mathcal{A}_\epsilon$ . Consequently

$$T[r^*(s) + \lambda(q(s) - r^*(s)), \theta^*(s)] - T[r^*(s), \theta^*(s)] \geq 0 \tag{28}$$

and thus taking the limit  $\lambda \rightarrow 0$  we obtain the following necessary condition satisfied by a minimizer:

$$\int_0^1 \left( (q(s) - r^*(s)) \frac{\partial L}{\partial r} \Big|_{r^*(s), \theta^*(s)} + (q'(s) - r^{*'}(s)) \frac{\partial L}{\partial r'} \Big|_{r^*(s), \theta^*(s)} \right) ds \geq 0. \tag{29}$$

Since we can perturb  $\theta^*(s)$  by any smooth function  $\xi$  compactly supported on  $[0, 1]$ , we again obtain the weak Euler–Lagrange equation

$$\int_0^1 \xi'(s) \frac{\partial L}{\partial \theta'} \Big|_{r^*(s), \theta^*(s)} ds = 0. \tag{30}$$

We now illustrate how (29) and (30) can be used to derive further necessary conditions satisfied by a minimizer  $(r^*(s), \theta^*(s)) \in \mathcal{A}_\epsilon$ . Suppose  $(r^*(s), \theta^*(s)) \in \mathcal{A}_\epsilon$  minimizes  $T$  over  $\mathcal{A}_\epsilon$  and define the sets

$$U = (r^*)^{-1}\{\epsilon\}, \tag{31}$$

$$U^c = [0, 1] \setminus U. \tag{32}$$

Since  $r^*(s)$  is continuous,  $U$  is a closed subset of  $[0, 1]$ . On  $U$  it follows that (29) is automatically satisfied since

$$\begin{aligned} \int_U \left( (q(s) - r^*(s)) \frac{\partial L}{\partial r} \Big|_{r^*(s), \theta^*(s)} + (q'(s) - r^{*\prime}(s)) \frac{\partial L}{\partial r'} \Big|_{r^*(s), \theta^*(s)} \right) ds \\ = \int_U (q(s) - \epsilon) \frac{|\theta^{*\prime}(s)|}{1 - \epsilon} \frac{3 - 2\epsilon}{2} ds \geq 0. \end{aligned}$$

On  $U^c$  consider the perturbation  $q(s) = \tau v(s) + r^*(s)$ , where  $v$  is any smooth function compactly supported on  $V$  and  $\tau \in \mathbb{R}$  is small enough in magnitude that  $q(s) \in \mathcal{A}_\epsilon$ . Substituting into (29) yields

$$\tau \int_{U^c} \left( v(s) \frac{\partial L}{\partial r} \Big|_{r^*(s), \theta^*(s)} + v'(s) \frac{\partial L}{\partial r'} \Big|_{r^*(s), \theta^*(s)} \right) ds \geq 0. \tag{33}$$

Since  $\tau$  is of arbitrary sign, the above inequality is actually an equality. That is, for  $s \in U^c$  we recover the weak Euler–Lagrange equations for  $r^*(s)$ .

**Remark.** Taken together, the above necessary conditions imply that potential minimizers of  $T$  over  $\mathcal{A}_\epsilon$  consist of the family of curves satisfying the Euler–Lagrange equations away from the constraint. That is, potential minimizers consist of piecewise smooth curves satisfying (22) and (23), joined with circular arcs of radius  $\epsilon$ .

**4.2. Piecewise smooth minimum.** As in the case with no constraint, i.e.,  $\epsilon = 0$ , we now foliate  $\mathcal{O}_\epsilon$  by curves that minimize the time of flight. By symmetry we only foliate the upper half-annulus  $\mathcal{O}_\epsilon^+ = \{(x, y) \in \mathcal{O}_\epsilon : y \geq 0\}$ . To construct the foliation we examine the behavior of potential minimizers with terminal coordinates  $(r_t, \theta_f)$  satisfying  $(r_f, \theta_f) \in \partial\mathcal{O}_\epsilon^+$  (the boundary of  $\mathcal{O}_\epsilon^+$ ) which can be naturally divided into four regions:

$$\begin{aligned} R_1 &= \{(r, \theta) \in \partial\mathcal{O}_\epsilon^+ : \theta = 0\}, & R_2 &= \{(r, \theta) \in \partial\mathcal{O}_\epsilon^+ : r = 1\}, \\ R_3 &= \{(r, \theta) \in \partial\mathcal{O}_\epsilon^+ : \theta = \pi\}, & R_4 &= \{(r, \theta) \in \partial\mathcal{O}_\epsilon^+ : r = \epsilon\}. \end{aligned}$$

Each of these regions is considered as separate cases below.

**4.2.1. Minimizers terminating on  $R_1$ .** It follows from (22) and (23) that if  $D = 0$ , the strong solution to the Euler–Lagrange equation is a straight line connecting  $(1, 0)$  to the origin. In particular this implies that if  $(r_f, \theta_f) \in R_1$  then straight lines are the natural candidate minimizers.

**4.3. Minimizers terminating on  $R_2$ .** Suppose  $(r_f, \theta_f) \in R_2$ . For  $\theta_f$  sufficiently small we expect the minimizers to consist of the smooth strong solution curves defined by (22) and (23). However, if  $\theta_f > \pi/3$ , the strong solutions to the Euler–Lagrange equations will necessarily intersect the obstacle. Note that from the convexity of the strong solutions there exists a unique critical angle  $\theta_c^\epsilon \in (0, \pi/3)$  in which the strong solutions intersect the obstacle tangentially. Specifically,  $\theta_c^\epsilon$  is defined by

$$\left(r_{D(\epsilon)}^S(1/2), \theta_{D(\epsilon)}^S(1/2)\right) = (\epsilon, \theta_c^\epsilon) \quad \text{and} \quad \left.\frac{dr_{D(\epsilon)}^S}{ds}\right|_{1/2} = 0. \tag{34}$$

The critical angle  $\theta_c^\epsilon$  serves as a boundary in the sense that if the final angular coordinate  $\theta_f$  satisfies  $\theta_f/2 > \theta_c^\epsilon$  then it is necessary to consider piecewise-defined curves as candidate minimizers. This is made precise by the following proposition.

**Proposition 6.** *Suppose that the smooth solution curves given by (22) and (23) are global minimizers of  $T$  over  $\mathcal{A}_0$ . For  $\epsilon > 0$ , if  $\theta_f/2 < \theta_c^\epsilon$  then there exists a  $D \geq D(\epsilon) > 0$  such that  $(r_D^S(s), \theta_D^S(s)) \in \mathcal{A}_\epsilon$  minimizes  $T$  over curves in  $\mathcal{A}_\epsilon$  terminating at the angular coordinate  $\theta_f$ .*

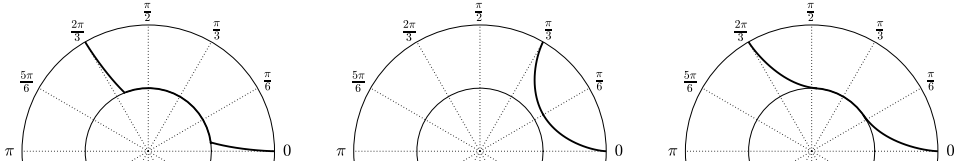
*Proof.* Let  $\theta_f \in (0, 2\pi/3)$  satisfy  $\theta_f/2 < \theta_c^\epsilon$  and  $(r_D^S(s), \theta_D^S(s))$  parametrize the smooth solutions given by (22) and (23) terminating at  $(\cos \theta_f, \sin \theta_f)$ . By Propositions 4 and 5,  $\theta_f$  and  $r_c$  are monotonically decreasing and increasing in  $D$  respectively, and thus  $r_c(D(\theta_f)) \geq \epsilon$ . Furthermore, since  $r_D^S(s)$  is convex in  $s$ , it follows that  $r_D^S(s) \geq \epsilon$  and thus  $(r_D^S(s), \theta_D^S(s)) \in \mathcal{A}_\epsilon$ . Finally, since  $\mathcal{A}_\epsilon \subset \mathcal{A}_0$  and  $(r_D^S(s), \theta_D^S(s))$  is assumed to minimize  $T$  over curves in  $\mathcal{A}_0$  which terminate at  $(\cos \theta_f, \sin \theta_f)$ , it follows that  $(r_D^S(s), \theta_D^S(s))$  also minimizes  $T$  over curves in  $\mathcal{A}_\epsilon$  which terminate at  $(\cos \theta_f, \sin \theta_f)$ .  $\square$

For a strong solution given by (22) and (23) satisfying  $\theta_f/2 > \theta_c^\epsilon$ , let

$$s_D^\epsilon = \min\{(r_D^S)^{-1}\{\epsilon\}\},$$

i.e., the first point of intersection with the obstacle. The natural generalizations of Propositions 1 and 2 can be shown to hold for  $\mathcal{A}_\epsilon$  and consequently we know, without loss of generality, that minimizers consist of curves symmetric about the angle  $\theta_f/2$  that are smooth solutions given by (22) and (23) away from the constraint, ride along it for a finite amount of time, and then rejoin a rotated and reflected version of the latter half of the same smooth solution; see Figure 4. This family of minimizers is

$$\mathcal{F}_{D, \theta_f}^\epsilon(s) = \begin{cases} \left(r_D^S(h(s)) \cos(h(s)), r_D^S(h(s)) \sin(\theta_D^S(h(s)))\right), & s \in \left[0, \frac{1}{3}\right), \\ \left(\epsilon \cos(t(s)), \epsilon \sin(t(s))\right) & s \in \left[\frac{1}{3}, \frac{2}{3}\right], \\ \left(r_D^S(s) \cos(l(s)), r_D^S(s) \sin(l(s))\right) & s \in \left(\frac{2}{3}, 1\right], \end{cases} \tag{35}$$



**Figure 4.** Plots of curves  $\mathcal{F}_{D,\theta_f}^\epsilon(s)$  defined by (35) with  $\epsilon = 0.5$ . Left: a curve  $\mathcal{F}_{D,\theta_f}^\epsilon(s)$  with  $\theta_f = 2\pi/3$ ,  $D = 0.0204$  and  $\theta_c = \pi/4$  which reaches the obstacle and rides along it; (middle) a curve  $\mathcal{F}_{D,\theta_f}^\epsilon(s)$  with  $\theta_f = \pi/3$ ,  $D = 0.2300$  and  $\theta_c = \pi/6$  which does not reach the obstacle ( $s_\epsilon = 0.5$ ); (right) a curve  $\mathcal{F}_{D,\theta_f}^\epsilon(s)$  with  $\theta_f = 2\pi/3$ ,  $D = 0.1250$  and  $\theta_c = 0.5981$  which approaches the obstacle at a tangent and rides along it.

where  $r_D^S(s), \theta_D^S(s)$  are given by (22) and (23) with  $\theta_f(D)$  satisfying  $\theta_f(D)/2 \geq \theta_c^\epsilon$ ,

$$\begin{aligned} h(s) &= s/(3s_D^\epsilon), & l(s) &= \theta_D^S(j_{D,\epsilon}(s)) + \theta_f - \theta_D^S(1), \\ j(s) &= 3s_D^\epsilon s + 1 - 3s_D^\epsilon, & t(s) &= 3(\theta_f - 2\theta^S(s_D^\epsilon))s + \theta_f - \theta^S(s_D^\epsilon). \end{aligned}$$

The following proposition characterizes the minimum of  $T$  over the family of curves given by (35); namely they consist of the curves defined by (35) that meet the constraint at a tangent.

**Proposition 7.** *Suppose that the smooth solution curves given by (22) and (23) are global minimizers in  $\mathcal{A}_0$ . For  $\epsilon > 0$ , if  $\theta_f/2 \geq \theta_c^\epsilon$  then the unique minimizer of  $T$  over the family of curves defined by (35) intersects the constraint tangentially.*

*Proof.* Let  $\theta_f \in (0, \pi)$  satisfy  $\theta_f/2 \geq \theta_c^\epsilon$ . Let  $\mathcal{F}_{D(\epsilon),\theta_f}^\epsilon(s)$  be the unique curve which intersects the constraint at a tangent and terminates at angular coordinate  $\theta_f$ . Let  $\mathcal{F}_{D,\theta_f}^\epsilon(s)$  be another curve with  $D < D(\epsilon)$  that terminates at angular coordinate  $\theta_f$ . By Propositions 4 and 5,  $\theta_f$  and  $r_c$  are monotonically decreasing and increasing in  $D$  respectively; hence  $r_c(D) \leq \epsilon$  and  $\theta_c(D) \geq \theta_c^\epsilon$ . Moreover, it follows from the monotonicity of  $\theta_D^S(s)$  in  $s$  that  $\mathcal{F}_{D,\theta_f}^\epsilon(s)$  intersects the constraint at some angular coordinate  $\theta_0 < \theta_c^\epsilon$  and intersects  $\mathcal{F}_{D(\epsilon),\theta_f}^\epsilon(s)$  at angular coordinate  $\theta_c^\epsilon$ . Since  $\mathcal{F}_{D(\epsilon),\theta_f}^\epsilon(s)$  is of the form  $(r_{D(\epsilon)}^S(s), \theta_{D(\epsilon)}^S(s))$  for  $s < s_{D(\epsilon)}^\epsilon$ , it follows from our assumption that smooth solutions given by (22) and (23) minimize  $T$  that  $\mathcal{F}_{D(\epsilon),\theta_f}^\epsilon(s)$  minimizes the time of flight to angular coordinate  $\theta_c^\epsilon$ . Moreover, both curves have the same time of flight along the constraint from angular coordinates  $\theta_c^\epsilon$  to  $\theta_f - \theta_c^\epsilon$ , and consequently  $T[\mathcal{F}_{D(\epsilon),\theta_f}^\epsilon(s)] < T[\mathcal{F}_{D,\theta_f}^\epsilon(s)]$ .  $\square$

**4.3.1. Minimizers terminating on  $R_3$ .** Let  $(r, \pi) \in R_3$ . We know from our prior analysis in  $R_2$  that all minimizers must ride along the obstacle until at least angular coordinate  $\pi - \theta_c^\epsilon$ . Hence, to minimize over curves terminating at  $(r, \pi)$ , we

need only minimize  $T$  over curves from  $(\epsilon, \pi - \theta_c^\epsilon)$  to  $(r, \pi)$ . Note that from the convexity of strong solutions given by (22) and (23), there exists a unique angle  $\phi_c^\epsilon \geq \pi - \theta_c^\epsilon$  such that the smooth solution comes off the obstacle tangentially at  $\phi_c^\epsilon$  and intersects  $(r, \pi)$ . It can be further shown that this curve minimizes the time of flight  $T$  between coordinates  $(\epsilon, \pi - \theta_c^\epsilon)$  and  $(r, \pi)$ . This is made precise by the following proposition.

**Proposition 8.** *Suppose that the smooth solution curves given by (22) and (23) are global minimizers in  $\mathcal{A}_0$ . For  $\epsilon > 0$  and  $(r, \pi) \in R_3$ , a minimizer of  $T$  over  $\mathcal{A}_\epsilon$  that terminates at  $(r, \pi)$  leaves the obstacle at a tangent.*

*Proof.* Let  $\alpha^* \in \mathcal{A}_\epsilon$  denote the candidate minimizer that leaves the obstacle at a tangent from the angular coordinate  $\phi_c^\epsilon$  and terminates at the polar coordinate  $(r, \pi)$ . From the convexity of smooth solutions, we know that any other candidate minimizer  $\alpha \in \mathcal{A}_\epsilon$  terminating at the polar coordinate  $(r, \pi)$  must come off the obstacle at some angular coordinate  $\phi_0 > \phi_c^\epsilon$ . Since the piece of the curve  $\alpha^*$  that connects the polar coordinates  $(\epsilon, \phi_c^\epsilon)$  to  $(r, \pi)$  is a smooth solution curve given by (22) and (23), it follows from our assumption that it minimizes the time of flight from polar coordinates  $(\epsilon, \phi_c^\epsilon)$  to  $(r, \pi)$ . Consequently, the time of flight from polar coordinates  $(\epsilon, \phi_0)$  to  $(r, \pi)$  along  $\alpha$  is larger.

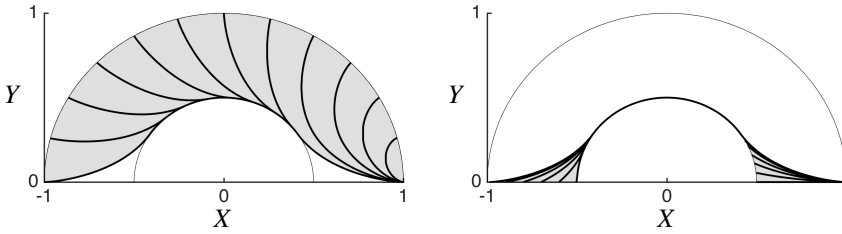
Using the same method as in the proof of Proposition 7, it can be shown that it follows from our assumption that smooth solutions are given by (22) and (23), that a candidate minimizer coming off the obstacle at angular coordinate  $\phi_0$  has greater time of flight than one coming off at angular coordinate  $\phi_c^\epsilon$ .  $\square$

**4.3.2. Minimizers terminating on  $R_4$ .** Let  $(\epsilon, \theta) \in R_4$ . If  $\theta < \theta_c^\epsilon$ , there is a smooth solution given by (22) and (23) that, by assumption, minimizes the time of flight to  $(\epsilon, \theta)$ . If  $\theta \in (\theta_c^\epsilon, \pi - \theta_c^\epsilon)$ , there exists a curve in the family (35) that minimizes the time of flight to  $(\epsilon, \theta)$ . If  $\theta > \theta_c^\epsilon$ , it follows from Proposition 8 that there is a curve minimizing the time of flight that approaches the obstacle at a tangent and rides along until angular coordinate  $\theta$ .

**Remark.** The solution curves connected to the boundary foliate the domain. Specifically, there are three distinct areas  $A_1, A_2$  and  $A_3$  satisfying  $\mathcal{O}_\epsilon = A_1 \cup A_2 \cup A_3$  that are foliated by curves connected to the boundary in  $R_2, R_3$  and  $R_4$  respectively. This is illustrated in Figure 5.

**4.4. Convergence to weak solutions.** In the previous section, Propositions 6 and 7 describe the behavior of a family of curves that minimizes  $T$  to terminal polar coordinates  $(1, \theta_f) \in R_2$ . For a given value of  $\epsilon \in (0, 1)$  and  $\theta_f \in (0, \pi)$ , we denote this family as

$$\alpha_{\theta_f}^\epsilon(s) = \begin{cases} (r_{D(\theta_f)}^S(s) \cos(\theta_{D(\theta_f)}^S(s)), r_{D(\theta_f)}^S(s) \sin(\theta_{D(\theta_f)}^S(s))) & \text{if } \theta_f/2 \leq \theta_c^\epsilon, \\ \mathcal{F}_{D(\epsilon), \theta_f}^\epsilon(s) & \text{if } \theta_f/2 > \theta_c^\epsilon, \end{cases} \quad (36)$$



**Figure 5.** Left: The annulus  $\mathcal{O}_\epsilon$  with  $A_1$  shaded in.  $A_1$  consists of the set of points which lie on solution curves terminating on  $R_2$ . Overlaid on  $A_1$  are evenly spaced solution curves terminating on  $R_2$  given by (35). Right: The annulus  $\mathcal{O}_\epsilon$  with  $A_2, A_3$  shaded in.  $A_2$  and  $A_3$  consist of the sets of points which lie on solution curves terminating on  $R_3$  and  $R_4$  respectively. Overlaid on  $A_2$  and  $A_3$  are evenly spaced solution curves terminating in  $R_3$  and  $R_4$ .

where  $(r_{D(\theta_f)}^S(s), \theta_{D(\theta_f)}^S(s))$  are the radial and angular coordinates of the unique smooth solution given by (22) and (23), and  $\mathcal{F}_{D(\epsilon), \theta_f}^\epsilon(s)$  is a member of the family described by (35). Moreover, as  $\epsilon$  approaches 0, this family converges to the natural foliation of the unit disk described in Figure 3 (right) and given by

$$\alpha_{\theta_f}(s) = \begin{cases} (r_{D(\theta_f)}^S(s) \cos(\theta_{D(\theta_f)}^S(s)), r_{D(\theta_f)}^S(s) \sin(\theta_{D(\theta_f)}^S(s))) & \text{if } \theta_f < 2\pi/3, \\ (r^W(s) \cos(\theta^W(s)), r^W(s) \sin(\theta^W(s))) & \text{if } \theta_f \geq 2\pi/3, \end{cases} \quad (37)$$

where  $r^W(s), \theta^W(s)$  are the radial and angular coordinates of the unique smooth solution given by (25) and (26) with  $|B| = 1$  and  $\theta_f$ . This is made precise in the following proposition.

**Proposition 9.** For  $\theta_f \in (0, \pi)$ ,

$$\lim_{\epsilon \rightarrow 0} d(\alpha_{\theta_f}^\epsilon, \alpha_{\theta_f}) = 0,$$

where

$$d(\alpha_{\theta_f}, \alpha_{\theta_f}) = \sup_{0 \leq s \leq 1} \inf_{0 \leq t \leq 1} |\alpha_{\theta_f}^\epsilon(s) - \alpha_{\theta_f}(t)|$$

is the natural distance between the images of curves in the uniform norm.

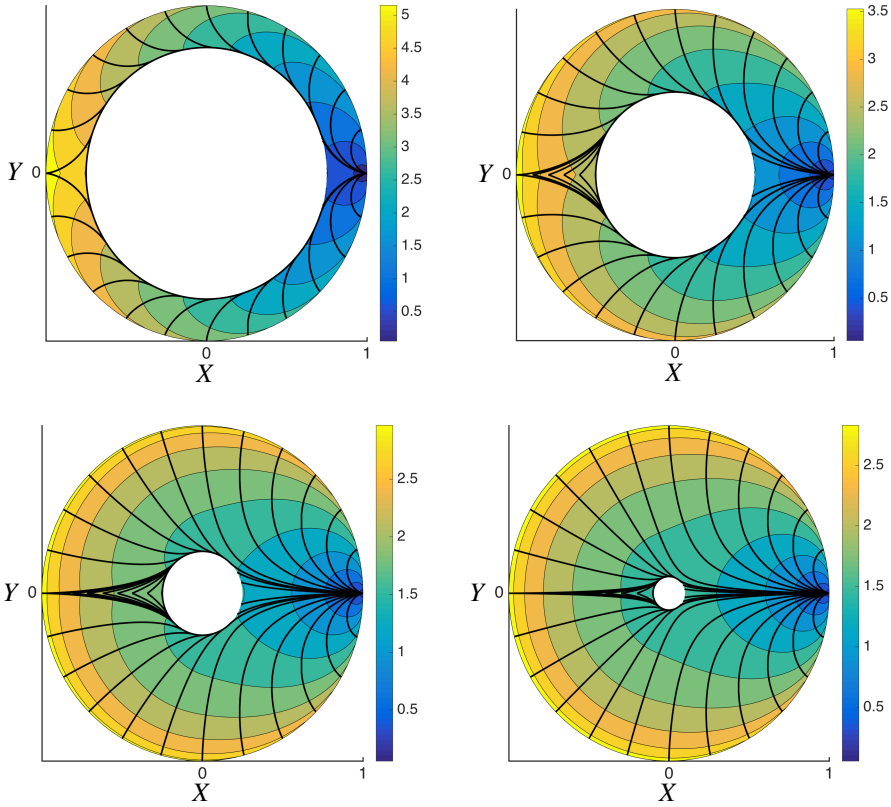
*Proof.* Let  $\theta_f \in (0, \pi)$  and define the sequence of functions  $\alpha_{\theta_f}^\epsilon(s)$  by (36).

(1) If  $\theta_f < 2\pi/3$ , there exists some  $D(\theta_f) > 0$  such that the smooth solution  $(r_{D(\theta_f)}^S(s), \theta_{D(\theta_f)}^S(s))$  given by (22) and (23) terminates at  $(1, \theta_f)$ . By Proposition 5,

$$\lim_{\epsilon \rightarrow 0} \theta_c^\epsilon = \pi/3 \geq \theta_f/2.$$

Therefore, there exists  $\epsilon^*$  such that  $\epsilon < \epsilon^* \implies \alpha_{\theta_f}^\epsilon = \alpha_{\theta_f}$ . Thus, for  $\theta_f < 2\pi/3$ ,

$$\lim_{\epsilon \rightarrow 0} d(\alpha_{\theta_f}^\epsilon, \alpha_{\theta_f}) = 0.$$



**Figure 6.** A foliation of the annuli  $\mathcal{O}_{0.75}$ ,  $\mathcal{O}_{0.5}$ ,  $\mathcal{O}_{0.25}$ ,  $\mathcal{O}_{0.1}$  by evenly spaced solution curves terminating on the boundary of the annulus. Beneath the solution curves of each subfigure is a contour plot of the time of flight from  $(1, 0)$  to each point on the annulus by solution curves of the form given by (36).

(2) If  $\theta_f \geq 2\pi/3$ , then  $\alpha_{\theta_f}^\epsilon$  approaches the obstacle tangentially at  $(r, \epsilon_c^\epsilon)$  along the path of a smooth solution for  $s \in [0, 1/3]$ . It follows from the convexity of smooth solutions that  $\alpha_{\theta_f}^\epsilon([0, 1/3])$  is contained in the rectangular region

$$\mathcal{R}_\epsilon = \{(x, y) \in \mathbb{R}^2 : x \in [\epsilon \cos(\theta_c^\epsilon), 1], y \in [0, \epsilon \sin(\theta_c^\epsilon)]\}.$$

As  $\epsilon \rightarrow 0$ , the region  $\mathcal{R}_\epsilon$  limits to the line  $\{(x, y) \in \mathbb{R}^2 : x \in [0, 1], y = 0\}$ . Moreover,  $\alpha_{\theta_f}^\epsilon([1/3, 2/3])$  is on the obstacle and consequently limits to the origin as  $\epsilon \rightarrow 0$ . It follows immediately from radial symmetry that a rotated rectangular region can be constructed around  $\alpha_{\theta_f}^\epsilon([2/3, 1])$  that limits to the line  $\theta = \theta_f$ . Hence each point on  $\alpha_{\theta_f}^\epsilon$  limits to a point along the weak solution  $(r^W(s) \cos(\theta^W(s)), r^W(s) \sin(\theta^W(s)))$  and thus for  $\theta_f > 2\pi/3$ ,

$$\lim_{\epsilon \rightarrow 0} d(\alpha_{\theta_f}^\epsilon, \alpha_{\theta_f}) = 0. \quad \square$$

The solution curves depicted in Figure 6 again intersect the level sets of the value function orthogonally. This is consistent with (13); i.e.,  $\mathcal{V}$  satisfies the eikonal equation defined by (12) on an annular domain.

## 5. Discussion and conclusion

In this paper we solved the brachistochrone problem in the inverse-square gravitational field. Namely, we constructed solutions that enter the so-called forbidden region first mentioned in [Parnovsky 1998; Tee 1999; Gemmer et al. 2011]. Furthermore we considered the constrained problem where solutions are restricted to lie outside of a ball around the origin. This restricted problem is more physically relevant since it avoids the particle experiencing infinite acceleration at the origin. Moreover, the solutions in the annular domain recover our prior solutions on the disk in the limit of vanishing inner radius. Consequently these solutions on the annular domain correspond to “regularized” brachistochrone solutions that avoid the singularity.

In the future, this work could be extended to problems with multiple singularities. That is, a natural extension of this work is to consider brachistochrone problems with multiple point sources of gravity. Natural questions to consider would be what role if any does the existence of a forbidden region play in the selection of strong or weak solutions. If weak solutions do exist, we conjecture that they would form a network of strong solutions patched together at singularities of the gravitational field. We expect that many of our results would hold locally near a singularity. However, by adding multiple singularities we break the radial invariance which we exploited to explicitly construct global solutions.

We also should mention that we have only considered necessary conditions for optimality. Specifically, this problem is not completely solved in the modern sense without a proof of the existence of a minimizer. This is not a trivial task since the functional is not coercive and is not convex at the singular origin and hence the direct method of the calculus of variations cannot be applied. We conjecture that the general results for noncoercive integrals presented in [Botteron and Marcellini 1991] or the technique of convex rearrangement presented in [Greco 2012] can be adapted to prove existence on the annular domain. Consequently, we expect that we could prove an existence result on the entire disk by considering the limit of vanishing inner radius.

## Acknowledgments

We wish to acknowledge Zachary Nado for many fruitful conversations. Grimm would also like to thank the MAA for providing travel support to attend MathFest 2015, where this work was first presented. Gemmer is currently supported by NSF-RTG grant DMS-1148284.



## References

- [Audoly and Boudaoud 2003] B. Audoly and A. Boudaoud, “Self-similar structures near boundaries in strained systems”, *Phys. Rev. Lett.* **91** (Aug 2003), art. id. 086105, 4 pp.
- [Bertsekas 1995] D. P. Bertsekas, *Dynamic programming and optimal control*, vol. 1, Athena Scientific, Belmont, MA, 1995. MR Zbl
- [Bhattacharya 2003] K. Bhattacharya, *Microstructure of martensite: why it forms and how it gives rise to the shape-memory effect*, Oxford Series on Materials Modelling **2**, Oxford University Press, 2003. MR Zbl
- [Blåsjö 2005] V. Blåsjö, “The isoperimetric problem”, *Amer. Math. Monthly* **112**:6 (2005), 526–566. MR Zbl
- [Botteron and Marcellini 1991] B. Botteron and P. Marcellini, “A general approach to the existence of minimizers of one-dimensional non-coercive integrals of the calculus of variations”, *Ann. Inst. H. Poincaré Anal. Non Linéaire* **8**:2 (1991), 197–223. MR Zbl
- [Broer 2014] H. W. Broer, “Bernoulli’s light ray solution of the brachistochrone problem through Hamilton’s eyes”, *Internat. J. Bifur. Chaos Appl. Sci. Engrg.* **24**:8 (2014), art. id. 1440009, 15 pp. MR Zbl
- [DeSimone et al. 2000] A. DeSimone, R. V. Kohn, S. Müller, and F. Otto, “Magnetic microstructures: a paradigm of multiscale problems”, pp. 175–190 in *ICIAM 99: proceedings of the Fourth International Congress on Industrial and Applied Mathematics* (Edinburgh, 1999), edited by J. M. Ball and J. C. R. Hunt, Oxford Univ. Press, 2000. MR Zbl
- [Dunham 1990] W. Dunham, *Journey through genius: the great theorems of mathematics*, Wiley, New York, 1990. MR Zbl
- [Erlichson 1999] H. Erlichson, “Johann Bernoulli’s brachistochrone solution using Fermat’s principle of least time”, *European J. Phys.* **20**:5 (1999), 299–304. MR Zbl
- [Evans 1998] L. C. Evans, *Partial differential equations*, Graduate Studies in Mathematics **19**, American Mathematical Society, Providence, RI, 1998. MR Zbl
- [Gemmer et al. 2011] J. A. Gemmer, M. Nolan, and R. Umble, “Generalizations of the brachistochrone problem”, *Pi Mu Epsilon Journal* **13**:4 (2011), 207–218.
- [Gemmer et al. 2016] J. Gemmer, E. Sharon, T. Shearman, and S. C. Venkataramani, “Isometric immersions, energy minimization and self-similar buckling in non-Euclidean elastic sheets”, *Europhys. Lett.* **114**:2 (2016), art. id. 24003, 6 pp.
- [Goldstein et al. 2014] H. Goldstein, C. P. Poole, and J. L. Safko, *Classical mechanics*, 3rd ed., Pearson, Harlow, 2014.
- [Greco 2012] A. Greco, “Minimization of non-coercive integrals by means of convex rearrangement”, *Adv. Calc. Var.* **5**:2 (2012), 231–249. MR Zbl
- [Kohn 2007] R. V. Kohn, “Energy-driven pattern formation”, pp. 359–383 in *International Congress of Mathematicians* (Madrid, 2006), vol. 1, edited by M. Sanz-Solé et al., Eur. Math. Soc., Zürich, 2007. MR Zbl
- [Leoni 2009] G. Leoni, *A first course in Sobolev spaces*, Graduate Studies in Mathematics **105**, American Mathematical Society, Providence, RI, 2009. MR Zbl
- [McCleary 2013] J. McCleary, *Geometry from a differentiable viewpoint*, 2nd ed., Cambridge University Press, 2013. MR Zbl
- [Müller 1999] S. Müller, “Variational models for microstructure and phase transitions”, pp. 85–210 in *Calculus of variations and geometric evolution problems* (Cetraro, 1996), edited by S. Hildebrandt and M. Struwe, Lecture Notes in Math. **1713**, Springer, 1999. MR Zbl

- [Nakane and Fraser 2002] M. Nakane and C. G. Fraser, “The early history of Hamilton–Jacobi dynamics 1834–1837”, *Centaurus* **44**:3–4 (2002), 161–227. MR Zbl
- [Oprea 2000] J. Oprea, *The mathematics of soap films: explorations with Maple*, Student Mathematical Library **10**, American Mathematical Society, Providence, RI, 2000. MR Zbl
- [Ortiz and Gioia 1994] M. Ortiz and G. Gioia, “The morphology and folding patterns of buckling-driven thin-film blisters”, *J. Mech. Phys. Solids* **42**:3 (1994), 531–559. MR Zbl
- [Parnovsky 1998] A. Parnovsky, “Some generalisations of brachistochrone problem”, *Acta Physica Polonica A* **93**:Supplement (1998), S–55.
- [Royden and Fitzpatrick 2010] H. Royden and P. Fitzpatrick, *Real analysis*, 4th ed., Pearson, Boston, 2010. Zbl
- [Sagan 1969] H. Sagan, *Introduction to the calculus of variations*, McGraw-Hill, New York, 1969. Reprinted Dover, New York, 1992. MR
- [Sharon et al. 2007] E. Sharon, B. Roman, and H. L. Swinney, “Geometrically driven wrinkling observed in free plastic sheets and leaves”, *Phys. Rev. E* **75**:4 (2007), art. id. 046211, 7 pp.
- [Song et al. 2008] J. Song, H. Jiang, W. M. Choi, D. Y. Khang, Y. Huang, and J. A. Rogers, “An analytical study of two-dimensional buckling of thin films on compliant substrates”, *J. Appl. Phys.* **103**:1 (2008), art. id. 014303, 10 pp.
- [Sussmann and Willems 1997] H. J. Sussmann and J. C. Willems, “300 years of optimal control: from the brachistochrone to the maximum principle”, *IEEE Control Systems* **17**:3 (1997), 32–44.
- [Tee 1999] G. J. Tee, “Isochrones and brachistochrones”, *Neural Parallel Sci. Comput.* **7**:3 (1999), 311–341. MR Zbl
- [Witten 2007] T. A. Witten, “Stress focusing in elastic sheets”, *Rev. Modern Phys.* **79**:2 (2007), 643–675. MR Zbl

Received: 2016-05-05      Accepted: 2016-07-24

christopher\_grimm@brown.edu      *Department of Computer Science, Brown University,  
Providence, RI 02912, United States*

gemmerj@wfu.edu      *Department of Mathematics, Wake Forest University,  
Winston Salem, NC 27109, United States*

# Numerical existence and stability of solutions to the distributed spruce budworm model

Hala Al-Khalil, Catherine Brennan, Robert Decker,  
Aslihan Demirkaya and Jamie Nagode

(Communicated by John Baxley)

This paper presents the steady-state solutions and traveling wave solutions for a spatially distributed PDE version of the spruce budworm model. The ODE (undistributed) model has been used in practical scenarios to model the outbreaks of the spruce budworm in forest environments, alongside the study of concepts involving fixed points and bifurcations in introductory differential equations courses. This study represents the spread of an outbreak from one end of a forest to the other. Numerical simulations are conducted using spectral methods.

## 1. Introduction

In the early 1900s, regions of eastern Canada began to see periodic outbreaks in the spruce budworm population, occurring approximately forty years apart [Williams and Birdsey 2003]. These outbreaks caused severe forest devastation, particularly in conifer tree species that are preferred by the budworms. In response to these population explosions, researchers at the University of British Columbia sought to explain and predict the outbreaks using mathematical models. The spruce budworm model, introduced in [Ludwig et al. 1978], is a modified logistic growth equation with an additional term,  $p(N)$ , to account for budworm mortality due to predation. Specifically,

$$\frac{dN}{d\tau} = r_B N \left(1 - \frac{N}{K_B}\right) - p(N) \quad \text{with} \quad p(N) = \frac{BN^2}{A^2 + N^2}, \quad (1)$$

where  $N$  represents the spruce budworm population,  $r_B$  represents the intrinsic growth rate and  $K_B$  is the carrying capacity of the budworm population. The predation term  $p(N)$  is determined by the *switching value*  $A$  and the predation efficiency  $B$ . The switching value for predation refers to the minimum budworm

---

*MSC2010:* 34B15, 35B32, 35B35, 35C07.

*Keywords:* spruce budworm, steady states, traveling waves, stability.

population required to cause birds to take interest in them as a source of food. Predation efficiency refers to the degree of accuracy exhibited by predatory birds in the capture of budworms.

Equation (1) contains variables of varying dimensions, making numerical analysis a challenge. To simplify (1), we seek to remove physical dimension from the variables. Substituting

$$u = N/A, \quad r = Ar_B/B, \quad q = K_B/A, \quad \text{and} \quad t = B\tau/A$$

into (1), we find the nondimensionalized spruce budworm model

$$\frac{du}{dt} = ru\left(1 - \frac{u}{q}\right) - h(u) \quad \text{with} \quad h(u) = \frac{u^2}{1 + u^2}, \tag{2}$$

where  $u$  represents the budworm population density and  $t$  represents time. As with the logistic growth model,  $r$  and  $q$  correspond with the natural growth rate and the carrying capacity of the population respectively.

The traditional spruce budworm model simulates a stationary population over time. It does not account for the spatial layout of the budworm habitat or the diffusion of the population across this habitat. In order to make the spruce budworm model mimic a diffusive insect population, the addition of a diffusion term is necessary. The fundamental differential equation of diffusion in one spatial dimension  $x$  is given by

$$C_t = aC_{xx},$$

where  $C$  is the concentration of the diffusing substance,  $t$  is the time variable,  $x$  is the spatial variable and  $a$  is the diffusion coefficient. The term  $C_t$  represents the change in the concentration of the diffusing substance with respect to time, and the term  $C_{xx}$  accounts for the diffusing substance changing over space, or along the  $x$  axis. Making use of Fick's second law of diffusion, we can deduce that the diffusion of the spruce bud worm population  $u$  across a linear habitat defined by  $x$  can be modeled by the second derivative of  $u$  in respect to  $x$ . The addition of the diffusion term  $au_{xx}$  to (2) leaves us with the distributed spruce budworm model

$$u_t = au_{xx} + ru\left(1 - \frac{u}{q}\right) - \frac{u^2}{1 + u^2}, \tag{3}$$

which simulates a migratory population that is both time and space-dependent.

In this paper, we study the numerical existence of the steady-state and the traveling wave solutions of (3). First we use the shooting method to determine the steady-state solutions at various diffusion rates ( $a$ ) and identify bifurcation values that produce additional steady-state solutions. Then we vary the carrying capacity values ( $q$ ) and determine the growth rate ( $r$ ) where the traveling solutions travel to the right, to the left or stay there without a movement, and numerically estimate the velocities

for various combinations of  $r$  and  $q$ . Finally, we study the relation between the carrying capacity and the growth rate for various values of traveling velocities.

**2. Numerical methods and the region of exploration**

**2.1. Numerical methods.** We use numerical methods to compute and simulate the steady-state and traveling wave solutions of (3). We discretize in the spatial ( $x$ ) direction, and use a spectral differentiation matrix  $D_{xx}$  as in [Trefethen 2000] to approximate  $u_{xx}$  as  $D_{xx}u$ . This turns the PDE into a system of ODEs. We then use the shooting method along with the Matlab fsolve command to identify the steady states. Spectral differentiation matrices were paired with Matlab’s built in ODE solver ode45 to form a PDE solver that we used to verify steady-state solutions found from the shooting method and fsolve, and to simulate traveling wave solutions. The spatial range is chosen to be  $-1 \leq x \leq 1$ .

**2.2. Parameter ranges of exploration.** As the carrying capacity value  $q$ , growth rate  $r$  and diffusion constant  $a$  vary, the number of steady states and traveling waves of (3) changes. First we find the steady-state solutions (fixed point solutions) of the undistributed system (2) which satisfy the equation

$$ru\left(1 - \frac{u}{q}\right) = \frac{u^2}{1 + u^2}. \tag{4}$$

Since (4) can be written as a quartic equation, we expect a maximum of four solutions. Our interest is the case where four fixed solutions exist. In order to find these solutions, we look for the intersection points of the two functions

$$y_1 = ru\left(1 - \frac{u}{q}\right) \quad \text{and} \quad y_2 = \frac{u^2}{1 + u^2}.$$

In Figure 1, upper left, we present the intersection points of these two function curves when  $r = 0.5$  and  $q = 10$ . Clearly  $u = 0$  is a solution, so we have divided both  $y_1$  and  $y_2$  by  $u$  and graphed both functions to visualize the other three intersection points. For these values, the corresponding fixed point solutions are  $u = 0$ ,  $u = 0.6834$ ,  $u = 2.0000$  and  $u = 7.3166$ . In Figure 1, upper right, for these  $q$  and  $r$  values, we show the corresponding direction field of (2). As shown in this direction field,  $u = 0.6834$  and  $u = 7.3166$  are stable solutions, while  $u = 0$  and  $u = 2.0000$  are unstable solutions. In Figure 1, lower left and lower right, we present  $q$  and  $r$  values (on different scales) that gives us four intersection points of the two curves  $y_1$  and  $y_2$ , i.e., the four fixed solutions of (2).

When there are four solutions to (4), the solution  $u = 0$  will always be one of them, and it will be unstable. The smallest nonzero solution we will refer to as the “refuge level” and the largest nonzero solution as the “outbreak level”, which are

both stable. Between these two stable equilibria is an unstable one that we will refer to as “intermediate”.

As the carrying capacity  $q$  gets larger, the range of  $r$  values that provide four intersection points approaches  $0 < r < 0.5$ . To show this, consider (4) and then let  $q \rightarrow \infty$ . This results in a cubic equation in  $u$ , with  $u = 0$  being one of the roots. The discriminant of the resulting quadratic equation (after  $u = 0$  is factored out) is  $-4r^2 + 1$ , and hence to get three solutions (the fourth has gone to infinity) we require  $0 < r < 0.5$ , assuming positive  $r$ . The smallest value of  $q$  for which there are four intersection points is about  $q = 5$ .

Until now we only considered the fixed point solutions of the nondistributed model (2) and found  $q$  and  $r$  values that give us the maximum number of fixed point solutions (and hence two stable equilibria). We might expect that these  $q$  and  $r$  values would also give us two stable steady-state solutions (refuge and outbreak) to the distributed model and perhaps the maximum number of steady states of the distributed model (3). In fact we will see that as  $a$ , the diffusion constant, gets smaller, the number of steady states gets bigger. Also, the existence of the refuge and outbreak steady states are  $a$  dependent.

Finally, we use  $a$  values in the range  $0.0005 < a < 0.1$ . This range includes  $a$  values appropriate to both steady-state and traveling wave solutions that illustrate our findings.

### 3. Steady state solutions

In this section we will present the numerical steady-state solutions of (3) with the boundary conditions

$$u(-1, t) = 0, \quad u(1, t) = 0. \tag{5}$$

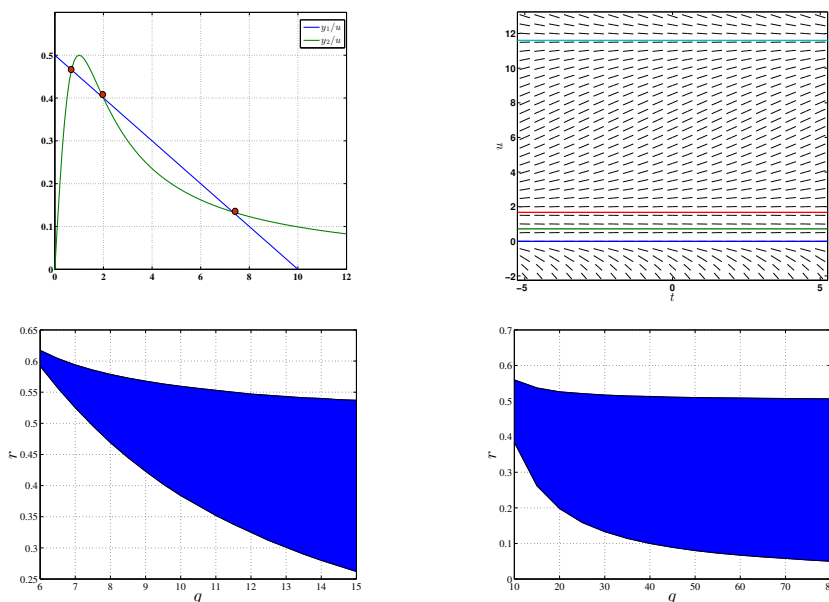
As we discussed in Section 2.2, we are interested in  $q$  and  $r$  values that will provide us the maximum number of steady-state solutions for the nondistributed case. For this purpose we now illustrate our results for  $r = 0.5$  and  $q = 10$ .

Steady state solutions  $u(x, t) = \phi(x)$  to (3) do not change over time, i.e.,  $\phi_t = 0$ . Thus  $\phi$  satisfies the following ordinary differential equation:

$$0 = a\phi'' + r\phi\left(1 - \frac{\phi}{q}\right) - \frac{\phi^2}{1 + \phi^2}, \tag{6}$$

with  $\phi(-1) = 0$  and  $\phi(1) = 0$ . We can change this second-order differential equation into a first-order system by defining  $y_1 = \phi$  and  $y_2 = \phi'$ . Then we get the system

$$y_1' = y_2, \quad y_2' = \frac{-ry_1}{a}\left(1 - \frac{y_1}{q}\right) + \frac{y_1^2}{a(1 + y_1^2)}. \tag{7}$$

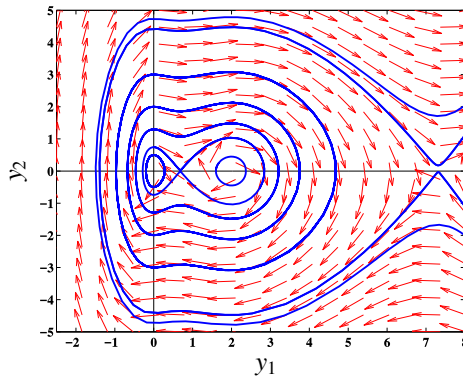


**Figure 1.** Upper left: the nonzero intersections of  $y_1$  and  $y_2$  occurs at  $u = 0$ ,  $u = 0.6834$ ,  $u = 2.0000$  and  $u = 7.3166$  when  $r = 0.5$  and  $q = 10$ . Upper right: the direction field of the nondistributed model (2) when  $r = 0.5$  and  $q = 10$ . Lower left and lower right:  $r$  vs.  $q$  values that gives exactly four fixed point solutions to (2).

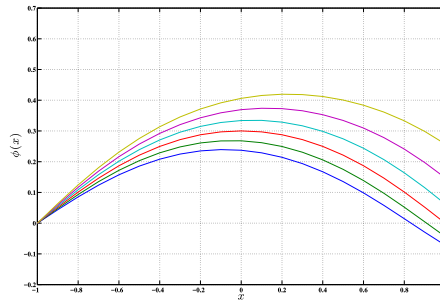
A phase portrait of the system of equations (7) is shown in Figure 2, for  $r = 0.5$ ,  $q = 10$  and  $a = 0.1$ . Other  $a$  values give a similar phase portrait (with a different scale on the  $y$  axis). In the phase portrait, one can see the four fixed points for the undistributed model, now as centers and saddles. The first (a center) is at the origin, the second (a saddle) is at  $\phi = 0.6834$ , the third (a center) is at  $\phi = 2$  and the fourth (a saddle) is at  $\phi = 7.3166$ .

**3.1. The shooting method.** The shooting method is a numerical technique for solving two-point boundary value problems (BVP's) by reformulating them as initial value problems (IVP's). The objective of this method is to determine initial conditions for the corresponding IVP that produce solutions that satisfy the original BVP. Solutions are found by fixing the left boundary point of the solution and guessing the initial slope until the right-hand boundary condition is satisfied.

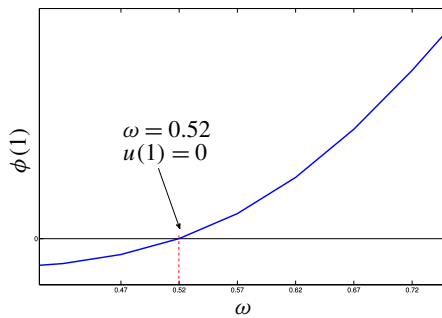
Several sample solutions to (6) on  $[-1, 1]$  with initial conditions  $\phi(-1) = 0$  and  $\phi'(-1) = \omega$  are plotted in Figure 3. The value of  $\omega$ , or the initial slope of the solution, is varied until the right endpoint of the solution,  $\phi(1)$ , meets the desired boundary value at zero.



**Figure 2.** Phase portrait of (7) when  $a = 0.1$ ,  $r = 0.5$  and  $q = 10$ .



**Figure 3.** Solutions to the IVP (bottom to top)  $\omega = 0.44, 0.48, 0.52, 0.56, 0.60$  and  $0.64$  and for  $r = 0.5$ ,  $q = 10$  and  $a = 0.1$ .



**Figure 4.** The solution  $\phi(1)$  as a function of  $\omega$  for  $r = 0.5$ ,  $q = 10$  and  $a = 0.1$ .

A plot of the right endpoints  $\phi(1)$  versus the initial slope values  $\omega$  can be used to determine the appropriate initial conditions to produce a solution to (6); see Figure 4. When the  $\phi(1)$  vs.  $\omega$  curve intersects the  $\omega$  axis,  $\phi(1) = 0$  and the right

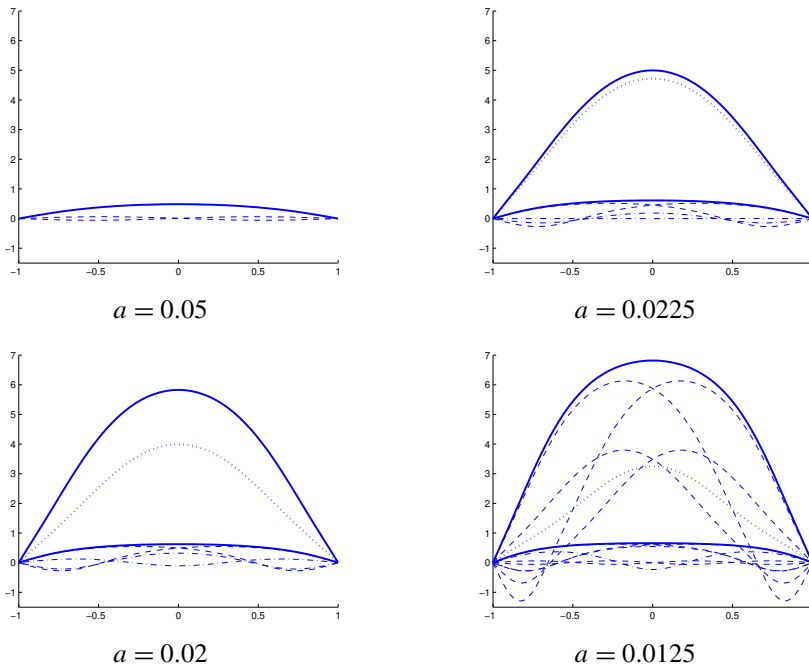


boundary condition is met. Each value of  $\omega$  that causes  $\phi(1) = 0$  represents a steady-state solution. Similar results apply to other  $a$  values.

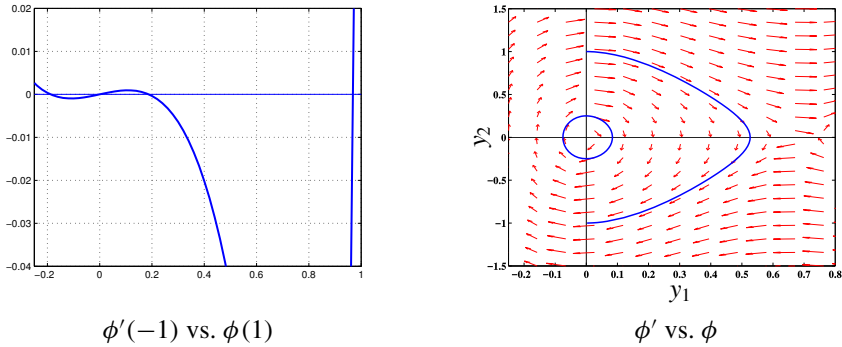
**3.2. Steady states for various  $a$  values.** In Figure 5 we show all nonzero steady-state solutions superimposed for a few  $a$  values. Solid lines represent stable steady-states, and dashed, dotted or dash-dot lines represent unstable ones.

The steady states for each diffusion rate, or  $a$  value, were determined using both the shooting method and the phase portrait of (7), which is shown in Figure 2. Within the shooting method plots, we expect a new steady-state solution to emerge each time the  $\omega$  axis is intersected. Furthermore, the  $\omega$  value at the point of intersection corresponds with the initial slope of the equilibrium solution. The phase portrait helps to make sure that no steady-state solutions are missed; each steady-state solution must start on the  $\phi'$  axis and end on the  $\phi'$  axis ensuring that  $\phi = 0$  at  $x = -1$  and  $x = 1$ , as required by the boundary value problem.

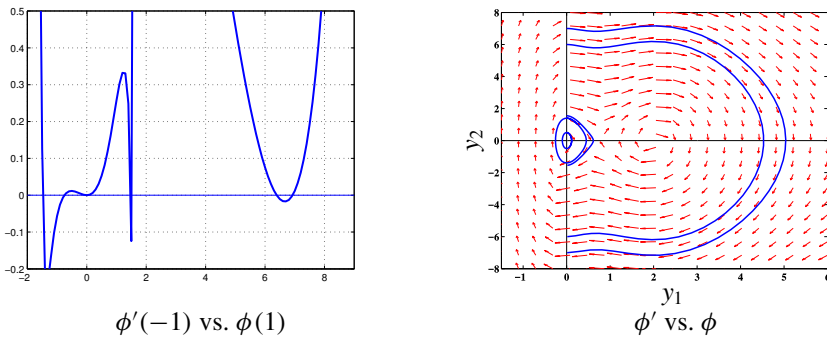
At  $a = 0.05$ , there are two positive initial conditions that force the boundary condition at  $\phi(1)$  to meet zero:  $\omega \approx 0.19$  and  $\omega \approx 0.97$ . In Figure 6, left, we see these values as points where the shooting plot crosses the  $\omega$  axis, and in Figure 6, right, we see these values as the starting values of the phase plots of the steady-state



**Figure 5.** Nonzero steady-state solutions for several  $a$  values and for  $r = 0.5$  and  $q = 10$ .



**Figure 6.** Steady-state solutions for  $a = 0.05$ ,  $r = 0.5$  and  $q = 10$ .

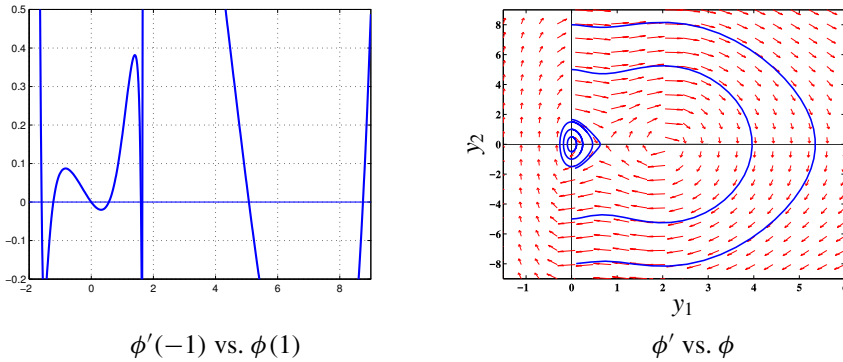


**Figure 7.** Solutions for  $a = 0.0225$ ,  $r = 0.5$  and  $q = 10$ .

solutions. For the smaller  $\omega$ -value, the phase plots overlap themselves and thus appear to form a closed loop. Finally, if the initial slope is  $\omega \approx -0.19$ , we get another steady-state solution, and the phase plot is indistinguishable from the one for which  $\omega \approx 0.19$ .

Looking back to Figure 5, upper left, we again see the three (nontrivial) steady-state solutions. The one with the larger positive initial slope ( $\omega \approx 0.97$ ) is stable and corresponds to the refuge level of the undistributed model. The unstable solution with the smaller positive initial slope corresponds to the unstable (dashed) solution that goes from slightly positive (on the left) to slightly negative (on the right). The third unstable solution has initial slope  $\omega \approx -0.19$  and is a mirror image (left-right) of the other unstable solution.

We now look at what happens when  $a$  is lowered to  $a = 0.0225$ . In Figure 7, right, we find two new solutions whose phase plots wrap around the saddle at  $\phi = 0.6834$  and the center at  $\phi = 2.000$ . This indicates the appearance of a stable outbreak equilibrium solution, and a slightly smaller unstable intermediate steady-state solution, as shown in Figure 5, upper right. At this point we have steady-state



**Figure 8.** Solutions for  $a = 0.02$ ,  $r = 0.5$  and  $q = 10$ .

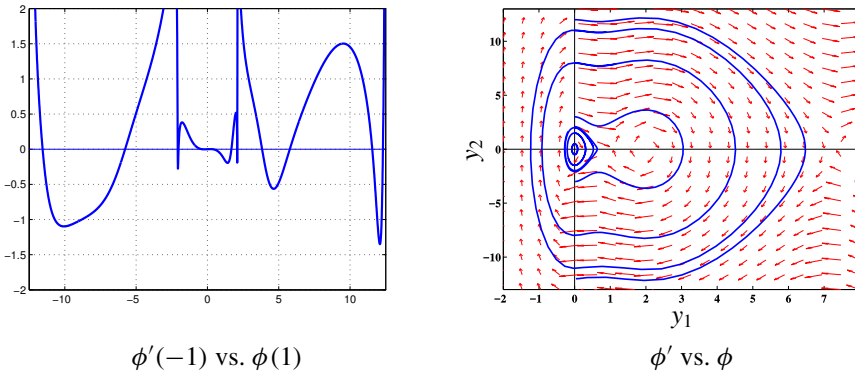
solutions corresponding to each of the four fixed points of the nondistributed model, and with similar stability types.

If you look closely at Figure 7, left, it appears that there will be six nonzero equilibrium solutions. There are a pair of positive solutions near  $\omega = 1.5$  and another pair of positive solutions near  $\omega = 6.65$  (outbreak and intermediate solutions). In addition there are negative solutions near  $\omega = -1.5$  and  $\omega = -0.75$ . Close to zero the situation is not so clear, but upon closer inspection we find another positive solution near  $\omega = 0.0055$  (as well as the trivial zero solution).

*Three solution types.* The solution types can be broken into three groups using phase plots. We define group I as steady-state solutions that start on the positive  $\phi'$  axis and end on the negative  $\phi'$  axis, and form exactly one-half of a loop. These are the solutions that correspond directly to the fixed-points of the nondistributed model, and represent physically realistic solutions. We define group II as solutions that loop around both centers and the smaller saddle one or more times (including half loops such as 1.5 or 2.5 loops). We will also refer to these as “big loops”, and they appear as “big waves” in the  $\phi$  vs.  $x$  plots of Figure 2. Because these solutions have negative  $\phi$  values, they are not physically realistic. Group III then consists of solutions that loop around only the origin one or more times (“small loops” in the phase plane or “small waves” in the  $\phi$  vs.  $x$  plots). These solutions are not physically realistic.

Thus as  $a$  changes from 0.05 to 0.0225 there are two bifurcations; there are two new group I steady-state solutions, and two new group III solutions. The group III solutions are 1.5 loop solutions around the origin (one with positive initial slope and one with equal and opposite negative initial slope).

As  $a$  is further reduced to 0.02 (Figure 8), the two larger steady-state solutions (stable outbreak and unstable intermediate half loops) grow more distinct and easily perceivable. Also, the shooting plot now shows that the small positive solution,



**Figure 9.** Solutions for  $a = 0.0125$ ,  $r = 0.5$  and  $q = 10$ .

which could not be distinguished for  $a = 0.0225$ , is now clearly visible, and shows up as a loop in the phase plot and a small wave in Figure 2. Thus no bifurcations occur between the  $a$  values 0.0225 and 0.02.

Note finally that the two 1-loop inner solutions (one for positive initial slope and one for equal and opposite initial slope) coincide in the phase plane and so cannot be distinguished from each other there. On the other hand, the two 1.5-loop inner solutions (initial slopes positive and negative but not equal and opposite), which are closer to the origin than the 1-loop solutions, are distinguishable in the phase plane.

Finally, when  $a$  is lowered again from 0.02 to 0.0125 (see Figure 9), we see two new group II solutions (“big loops” that wrap around once) as well as two group III solutions (“small loops” that wrap around two times). The two small loop solutions have equal and opposite initial slopes, and hence are indistinguishable in the phase plane. Thus two more bifurcations have occurred.

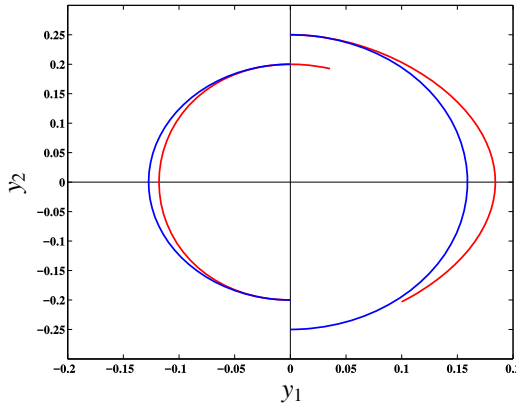
**3.3. Determination of bifurcation values.** [Aron et al. 2014, Theorem 3.4] states that the eigenvalues of the linear boundary value problem

$$\phi'' + \lambda^2 \phi = 0, \quad \phi(-1) = 0, \quad \phi(1) = 0, \tag{8}$$

correspond to the bifurcation values of the nonlinear boundary value problem

$$\phi'' + \lambda^2(\phi - \phi^3) = 0, \quad \phi(-1) = 0, \quad \phi(1) = 0. \tag{9}$$

The proof is based on the property that close to the origin, the solution curves of the nonlinear problem approach those of the linear one, and that as the solution curves move clockwise around the origin, the ones corresponding to the nonlinear problem move slower (in the sense of the angle in polar coordinates), and are farther from the origin (in the sense of the radius in polar coordinates), than those of the linear problem.



**Figure 10.** Linear (blue) vs. nonlinear (red) system near the origin.

The spruce budworm BVP (see (6)) can be written as

$$\phi'' + \frac{r}{a}\phi(1 - \phi) - \frac{\phi^2}{a(1 + \phi^2)} = 0, \quad \phi(-1) = 0, \quad \phi(1) = 0. \quad (10)$$

This equation can also be linearized, giving

$$\phi'' + \frac{r}{a}\phi = 0, \quad \phi(-1) = 0, \quad \phi(1) = 0, \quad (11)$$

which with the identification  $\lambda^2 = r/a$  is again (8). We hypothesize that this will give us some of the bifurcation values for the spruce budworm BVP of (10). Solving for  $a$ , we have  $a = r/\lambda^2$ . When  $\lambda_n = \frac{1}{2}n\pi$  (eigenvalues from (8)) is substituted into this equation, we find some of the expected bifurcation values of (3) in terms of  $a$ :

$$a_n = \frac{r}{\left(\frac{1}{2}n\pi\right)^2} \quad \text{for } n = 1, 2, 3, 4 \dots \quad (12)$$

The bifurcations calculated from (12) correspond to the emergence of a new small half loop (for  $n = 1$ ) and new small loops (for  $n > 1$ ) in the terminology of the previous section. That these can be calculated analytically is a result of the linearization of the problem for solution curves near the origin. For bifurcation values corresponding to the emergence of new big loop solutions (group II) we have estimated the bifurcation values using numerical exploration.

Finally, there are bifurcations that lead to new small loop solutions that are not given by (12). This is a result of the lack of left-right symmetry in the vector field for the nonlinear spruce budworm BVP, which leads to the property that solution curves in the phase plane travel faster around the origin (in terms of angle in polar coordinates), and closer to it (in terms of the radius in polar coordinates) than those of the linearized equation for  $x < 0$ , but slower around the origin (and farther

from it) than those of the linearized equation for  $x > 0$ , as long as the solution curves are sufficiently close to the origin (see Figure 10).

This allows new solutions to the BVP that start on the negative  $y$  axis and end on the positive  $y$  axis (1.5 loop, 2.5 loop, etc.) to occur for  $a$  values slightly larger than the predicted bifurcations values of (12). Only solutions that start on the negative  $y$  axis and end on the positive  $y$  axis can “outrun” the corresponding linear solution (since they spend more time in the fast region  $x < 0$ ). Thus  $a$  values corresponding to the appearance of these types of solutions are the only inner loop bifurcations that must be calculated using numerical experimentation.

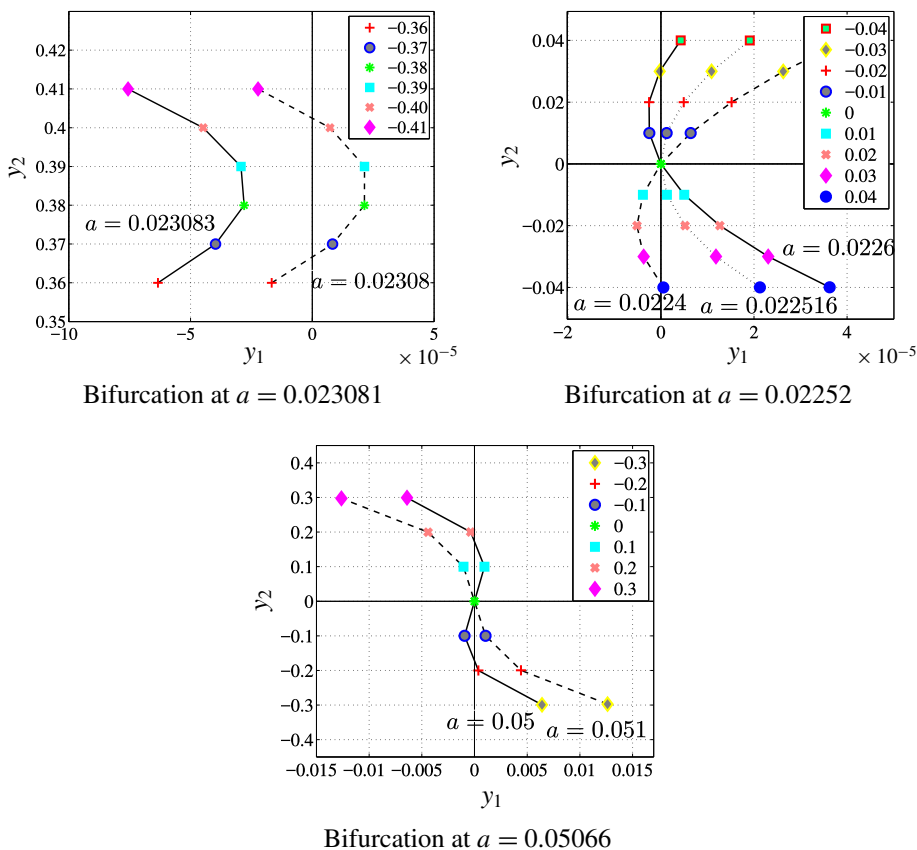
In Figure 11, left, we demonstrate a numerically calculated bifurcation of this type at approximately  $a = 0.0230814$ . To do this we show the endpoints only (with connecting lines for readability) of solution curves for  $a = 0.0230830$  and  $a = 0.0230800$ , corresponding to several initial conditions along the negative  $y = \phi'$  axis. The initial conditions used are labeled in the figure. These endpoints correspond to solutions that wrap around the origin about 1.5 times.

One sees that for  $a = 0.0230830$  the solution curves do not reach the  $y$  axis, and hence they are not solutions to the BVP. For  $a = 0.0230800$  the longer curves pass the  $y$  axis and the shorter ones fall short of it, showing that there are exactly two new solutions to the BVP. At some point in between these two cases there must be an  $a$  value for which the longest solution curve just touches the  $y$  axis (this value is about  $a = 0.0230814$ ).

This type of bifurcation is similar to a saddle-node bifurcation for a first-order ordinary differential equation, where at the bifurcation point a single fixed point appears where there was previously none, then this single fixed point splits into two fixed points which grow farther apart. This is also how new big-loop steady-state solutions are created; they must also be estimated using numerical exploration.

Figure 11, right and bottom, shows the two other types of bifurcation that occur to create new steady-state solutions. Figure 11, right, illustrates the type of bifurcation that occurs at a bifurcation point calculated by (12) when new solutions with a fractional number of loops are created (which corresponds to  $n$  odd in (12)). For  $a$  just larger than the bifurcation value, one observes a solution with negative initial condition and the zero solution. At the bifurcation value there is just the zero solution, and for  $a$  just smaller than the bifurcation value there is the zero solution and solution with positive initial condition. This is somewhat similar to a transcritical bifurcation for first-order ODE's.

Bifurcations of this type occur for  $a$  slightly smaller than the type shown in Figure 11, left. Thus as  $a$  gets smaller, first two new small loop solutions are created which have negative initial conditions (Figure 11, left), and then shortly after that the negative solution that is closest to zero switches over to become positive (Figure 11, right). The net result is one new solution with negative initial



**Figure 11.** Bifurcation types for  $r = 0.5$  and  $q = 10$ .

condition and one with positive initial condition (in addition to the zero solution) after both bifurcations.

Finally, Figure 11, bottom, shows the type of bifurcation that occurs when new inner loop solutions are created for  $n$  even in (12) (nonfractional number of loops). This type of bifurcation can be compared to the pitchfork bifurcation of first-order ODEs; as  $a$  is reduced, the zero solution gives rise to two new solutions, one with positive initial condition and one with negative initial condition (the zero solution continues). Note that the end of result of the two bifurcations in Figures 11, left, and 11, right, is similar to the bifurcation in Figure 11, bottom, in that including the zero solution, the number of solutions goes from one to three, corresponding to one new solution with positive initial condition and one negative. The difference is that for the case of even  $n$  the initial conditions that correspond to steady-state solutions have equal and opposite sign, but not for the case of odd  $n$ .

bifurcation type	bifurcation values in terms of $a$
small loop from $a_n = r/(\frac{1}{2}n\pi)^2 (n > 1)$	0.05066, 0.02252, 0.01267, 0.00811
small loop estimated numerically	0.023081
small half loop from $a_n = r/(\frac{1}{2}n\pi)^2 (n = 1)$	0.20264
big loop estimated numerically	0.0225578, 0.015, 0.0106, 0.0045

**Table 1.** Bifurcations values for  $r = 0.5$  and  $q = 10$ .

In Table 1 we show all bifurcations that occur for  $0.00811 \leq a \leq 0.20264$ . From that table we see that the bifurcation just described at  $a = 0.023814$  occurs just before (as  $a$  gets smaller) the one calculated by (12) at  $a = 0.02252$ .

**3.4. Stability analysis.** Our numerical simulations have shown that some of the equilibrium solutions found in Section 3.2 are stable and some are unstable. This has motivated us to check the eigenvalues of the linearized operator of (3) about the steady-state solution  $\phi$ .

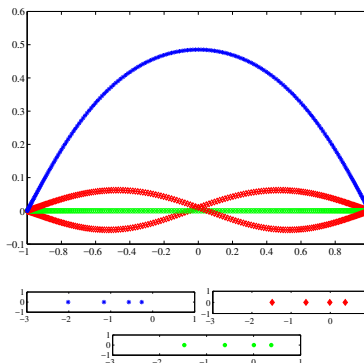
Let  $v$  be a small perturbation and  $u = \phi + v$  the solution to (3), then if we substitute it into (3), we get

$$(\phi + v)_t = a(\phi + v)_{xx} + r(\phi + v)\left(1 - \frac{\phi + v}{q}\right) - \frac{(\phi + v)^2}{1 + (\phi + v)^2}. \tag{13}$$

Since  $\phi$  is a steady-state solution, we have  $a\phi_{xx} + r\phi(1 - \phi/q) - \phi^2/(1 + \phi^2) = 0$ .

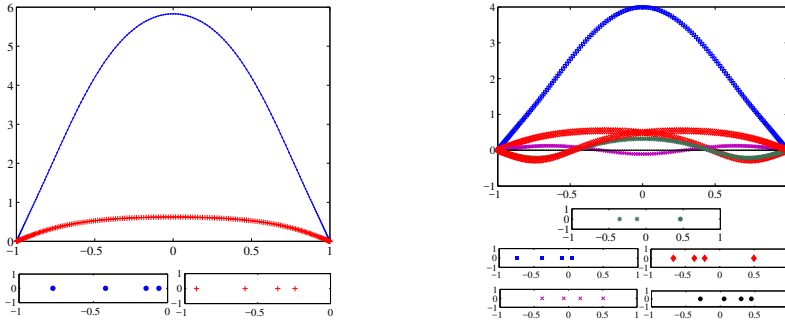
If we linearize the nonlinear terms about the steady-state  $\phi$ , we get

$$v_t = av_{xx} + f(\phi)v,$$



**Figure 12.** Stable and unstable solutions with color-coded eigenvalue spectrum for  $a = 0.05$ ,  $r = 0.05$  and  $q = 10$ .





**Figure 13.** Left: stable solutions with spectrum for  $a = 0.002$ ,  $r = 0.5$  and  $q = 10$ . Right: unstable solutions with spectrum for  $a = 0.002$ ,  $r = 0.5$  and  $q = 10$ .

where

$$f(\phi) = r - \frac{2r\phi}{q} - \frac{2\phi}{(1 + \phi^2)^2}.$$

Then the corresponding linear system is

$$v_t = \mathcal{H}v, \quad \text{where} \quad \mathcal{H} = a \frac{d^2}{dx^2} + f(\phi). \tag{14}$$

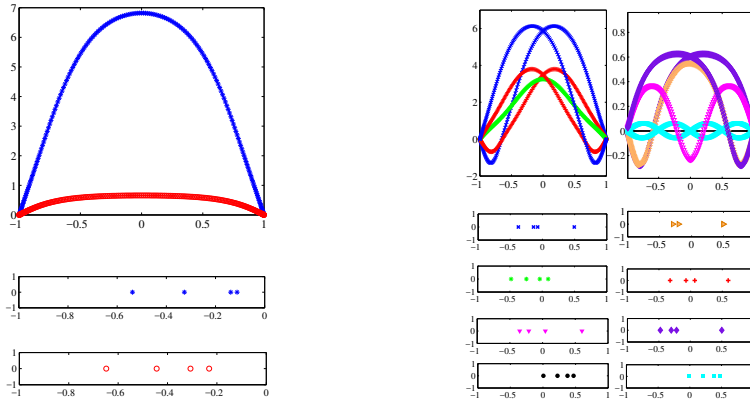
Our interest is the sign of the real part of the largest eigenvalue of  $\mathcal{H}$  for each steady state  $\phi$ . If that value is negative, we expect the perturbation from the steady state to shrink until the perturbed solution conforms to the steady state.

In Figures 12, 13 and 14 we show graphs of the steady-state solutions and the corresponding eigenvalues of  $\mathcal{H}$  for the  $\alpha$  values 0.05, 0.002 and 0.00125.

In Figure 15 we show snapshots of an animation of a perturbed initial condition and how it converges to a stable steady state. Notice that variations in the initial condition and large deviations from the original steady state do not affect the long term behavior of the solutions. This is typical of the outbreak and refuge solutions, for which all eigenvalues are negative.

Conversely, if the largest eigenvalue has a positive real part, we will expect the perturbation to grow, distancing the perturbed solution from the original steady state; see Figures 16 and 17. Equilibria with positive eigenvalues are unstable and achieved only under specific initial conditions [Seydel 2010]. Subtle changes to an initial condition in the neighborhood of an unstable equilibrium will alter the long term behavior of the solution. Figure 16 shows a small loop solution and Figure 17 shows an intermediate half-loop solution (between outbreak and refuge levels)

In some cases, the perturbed solution will rest near the steady state for a period of time, then slowly gravitate to a new, distinct resting place. This sort of behavior



**Figure 14.** Left: stable solutions with spectrum for  $a = 0.00125$ ,  $r = 0.5$  and  $q = 10$ . Right: unstable solutions with spectrum for  $a = 0.00125$ ,  $r = 0.5$  and  $q = 10$ .

is typical of equilibria that are “almost stable” in the sense that there is only one positive eigenvalue and it is very small.

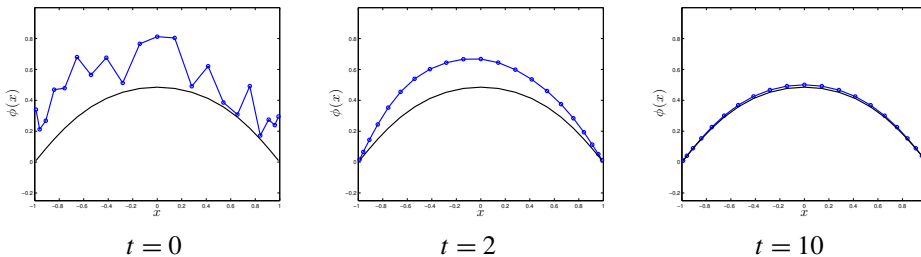
#### 4. Traveling wave solutions

We study the traveling wave solutions of (3) that are in the form  $u(x, t) = \phi(x - vt)$ , where  $v$  is the speed of the wave with the boundary conditions

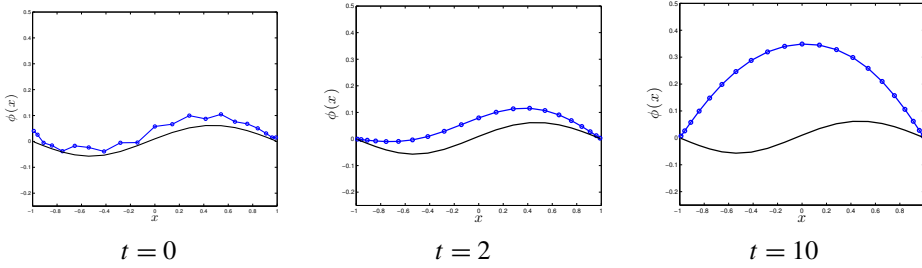
$$u(-1, t) = h, \quad \text{and} \quad u(1, t) = k$$

and the initial condition

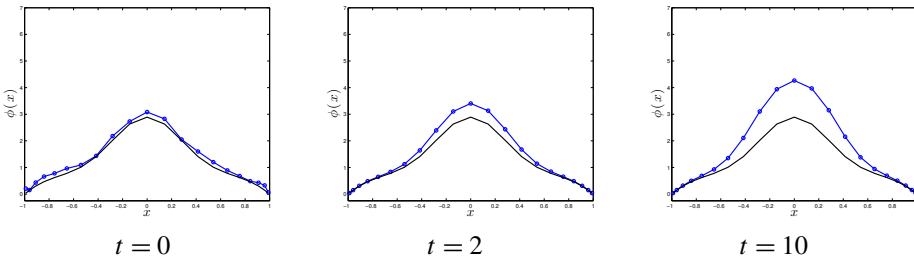
$$u(x, 0) = h + \frac{2(k-h)}{\pi} \tan^{-1} e^{Cx}. \tag{15}$$



**Figure 15.** A stable equilibrium solution (printed in black) and a perturbed solution (printed in blue) are plotted each at  $t = 0$ ,  $t = 2$  and  $t = 10$ . The perturbation from the steady state is amplified to highlight insensitivity to changes in the initial condition.



**Figure 16.** An unstable equilibrium solution (black) and a perturbed solution (blue).



**Figure 17.** An unstable equilibrium solution (black) and a perturbed solution (blue).

With these boundary and initial conditions (and appropriately chosen  $C$ ), the solutions are close in shape to traveling wavefronts, and thus quickly converge to traveling wavefronts and end in steady-state solutions. These waves represent growth/decay of the population as a function of the spatial dimension.

**4.1. Choosing boundary conditions.** We pick the boundary conditions as

$$u(-1, t) = h \quad \text{and} \quad u(1, t) = k,$$

where  $h$  and  $k$  are the fixed point solutions to the nondistributed model (2). As explained in Section 2.2, we consider  $q$  and  $r$  values that give us the four fixed point solutions to (2). Two of these solutions are stable and the other two are unstable. One of the unstable fixed solutions is the zero solution and if  $u_{s_1}$  and  $u_{s_2}$  are stable and  $u_u$  is the unstable solution, we have the following inequality:

$$0 < u_{s_1} < u_u < u_{s_2}. \tag{16}$$

We are interested in the traveling waves that converge to stable fixed point solutions at  $\pm 1$ , i.e.,  $h = u_{s_1}$  and  $k = u_{s_2}$ .

**4.2. Results.**

**4.2.1. The movement of the wavefront.** Traveling wavefronts move according to the boundary conditions, growth constant  $r$  and carrying constant  $q$ . Fixing  $q$  and varying  $r$ , a critical  $r = r^*$  was found at different  $q$  values such that:

- for  $r < r^*$ , the wave travels to the right, and the population dies out;
- for  $r > r^*$ , the wave travels to the left, and infestation occurs;
- for  $r = r^*$ , the wave does not travel, and no population growth or decay occurs.

The behavior of the wave movement was observed after incrementally selecting the values of  $q$  from [9, 15] with the increment 1 while  $r$  values were varied continuously within an interval for which both the refuge and outbreak levels existed. The effect of changing  $r$  and  $q$  over a selected range, with  $a = 0.001$  fixed, is recorded in Table 2. When the wave moves to the right, it means that the wave favors moving to the refuge solution and the population decreases. On the other hand, the wave favors outbreak and an increase in population when it moves to the left. This behavior is presented in Figure 18 for  $a = 0.001$  and  $q = 14.5$ , with  $r$  changing in value from its lowest to highest value within the range of  $r$  specified within the four solutions case for (2) as shown Figure 1, lower left. In Figure 18 the wave is plotted at the critical  $r$  value where the wave does not move and the thus the population does not change in time.

Note that for a wavefront that starts at the outbreak level on the left and ends at the refuge level on the right, the movement of the wavefront would be in the opposite direction of that just described.

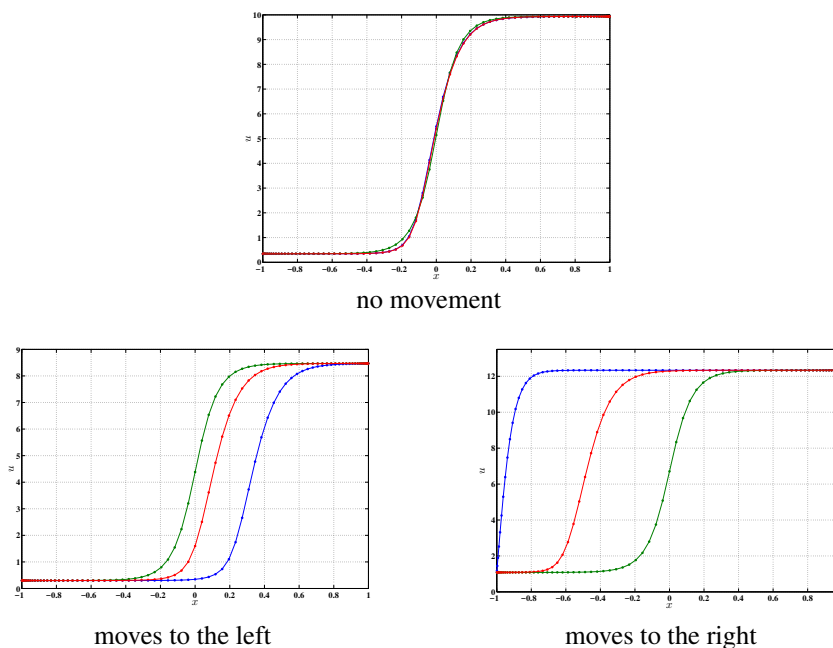
It can be shown that for  $a = 1$  that there is an integral condition [Murray 2005] that determines the value of  $r^*$  for fixed  $q$  for which the velocity is zero. The condition is

$$\int_{u_{s_1}}^{u_{s_2}} ru \left(1 - \frac{u}{q}\right) - \frac{u^2}{u^2 + 1} du = 0. \tag{17}$$

In fact, by inspecting the proof in the reference just given, it is clear that this condition works for all positive  $a$  values. This condition was checked against the numerically calculated values in Table 2, and the results were consistent. This

$q$	9	10	11	12	13	14	15
$r^*$	0.4605	0.4258	0.3956	0.3692	0.3459	0.3252	0.3067

**Table 2.** Critical  $r^*$  that represents zero velocity wavefronts for different  $q$  values, calculated numerically for  $a = 0.001$ , but valid for other  $a$  values. Larger  $r$  means wavefront moves left (outbreak level increasing) and for smaller  $r$  wavefront moves right (outbreak level decreasing).



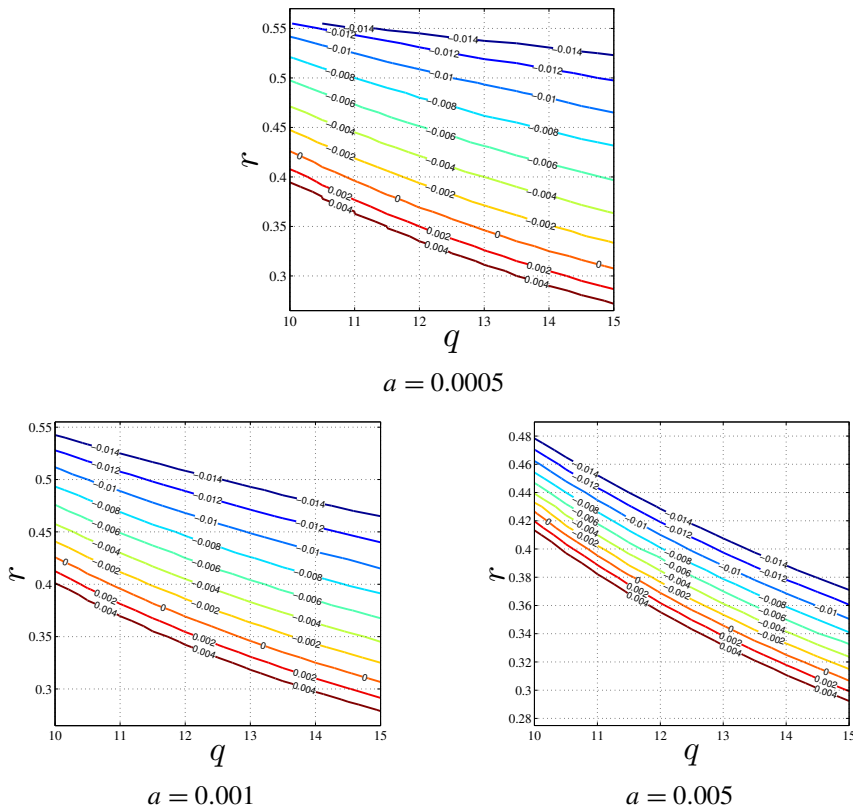
**Figure 18.** The plots captured at  $t = 0$  (blue curve),  $t = 5$  (red curve) and  $t = 10$  (green curve) with  $a = 0.001$  and  $q = 14.5$ . Top:  $r = 0.3165$ ,  $h = 0.345918$  and  $k = 9.93485$ . Left:  $r = 0.28$ ,  $h = 0.2987$  and  $k = 8.47012$ . Right:  $r = 0.538$ ,  $h = 1.03008$  and  $k = 12.3281$ .

means that even though the values given in Table 2 were calculated with  $a = 0.001$ , they are valid for other  $a$  values.

**4.2.2. Velocity as a function of  $r$  and  $q$ , with  $a$  fixed.** We have studied how the velocity of a traveling wave depends on  $r$  and  $q$ . Figure 19 shows that relation for a few different values of  $a$ . For these charts, the speed was calculated via simulation of the PDE for various  $r$  and  $q$  values, and then a contour plot of the data was created using Matlab.

These charts can be used to estimate the speed at which an outbreak spreads within the parameter ranges shown.

**4.2.3. The approximately linear relation between  $v$  and  $r$ , with  $q$  and  $a$  fixed.** For a fixed  $q$  value, by utilizing the ranges of velocities for each  $q$  and range of  $r$ , we observed an approximately linear relation between  $r$  and  $v$ . Thus, for a fixed carrying capacity, we can estimate the velocity of the budworm wave as a function of the growth rate of the insect. Figure 20 shows this relation for different values



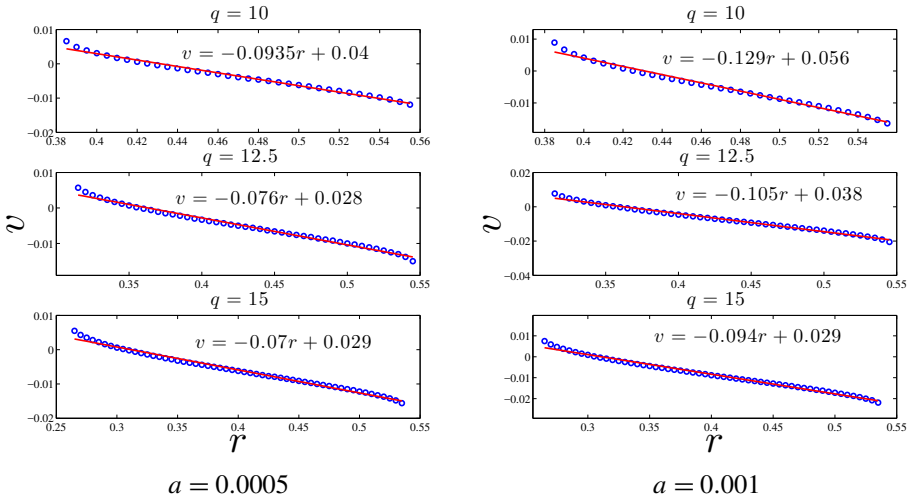
**Figure 19.** The speed contour plots.

of  $q$  when  $a = 0.0005$  and when  $a = 0.001$ . These equations are consistent with Figure 19.

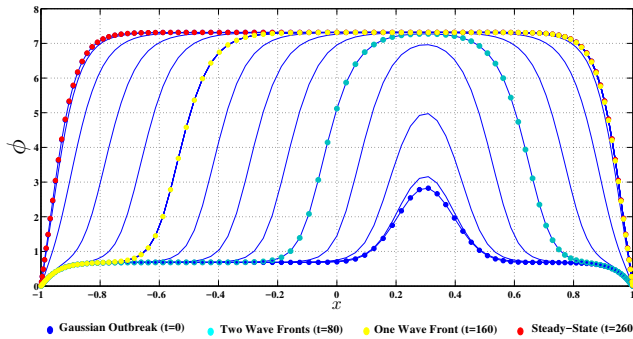
### 5. How outbreaks spread

In this section we demonstrate one possible way for an outbreak of budworms localized in space can spread to the entire forest (the region  $-1 \leq x \leq 1$ ). We choose  $r = 0.5$  and  $q = 10$  again, so that we are in the  $r - q$  region where there are four fixed points. We also choose  $a = 0.0005$  which makes sure that the outbreak and subsistence levels occur in the distributed model as well.

We show how a large enough perturbation of the steady-state solution that represents the subsistence level of budworms can create a traveling wavefront of the type studied in the last section, and end in the steady-state solution that represents the outbreak level. The speed of the wave can be estimated using the charts and equations from that last section as well.



**Figure 20.** Graphs of  $r$  vs.  $v$  for fixed  $q$  and  $a$ .



**Figure 21.** Initial Gaussian increase in budworm population from subsistence level spreads to outbreak.

In Figure 21 we show the effect of imposing a Gaussian bump, representing a small normally distributed increase in the budworm population, on top of a steady-state subsistence solution. In that figure we see snapshots of an animation every 20 time units, starting at  $t = 0$  (blue dots). Also highlighted are the times  $t = 80$  (cyan dots),  $t = 160$  (dark green dots) and  $t = 260$  (red dots). The snapshot at  $t = 80$  represents the point at which the initial disturbance to the subsistence level has grown so that the top has reached the outbreak level. At this point there are two wavefronts of the type described in Section 4 (one moving left and one moving right) as well as two regions that conform to the steady-state subsistence level ( $-1 \leq x \leq -0.2$  and  $0.8 \leq x \leq 1$ ) as described in Section 3.

At  $t = 160$  there are three regions; for  $-0.25 \leq x \leq 1$  we observe the population conforming to the steady-state outbreak level of Section 3, for  $-0.75 \leq x \leq -0.25$  we have a traveling wavefront as in Section 4, and for  $-1 \leq x \leq -0.75$  the population conforms to the steady-state subsistence level of Section 3. Finally at  $t = 260$  the population has completely reached the steady-state outbreak level.

Finally, from Figure 19 we can estimate the speed of the wavefront to be slightly larger than  $-0.006$  (for the left-moving front); the equation from Figure 20, right, gives  $-0.0067$ . In the 80 time units that separate the snapshots at  $t = 80$  and  $t = 160$  we would expect the wavefront to move about  $-0.5$  units to the left, which is what is seen in Figure 21.

## 6. Conclusion

The original spruce budworm model is an ordinary differential equation and it models the outbreaks of the spruce budworm in forest environments. By adding the diffusion term  $au_{xx}$  to the original equation, we got the distributed model, which is a partial differential equation. By using spectral numerical methods in the spatial direction, and Matlab's ode45 solver in the time direction, we studied the numerical existence of steady-state and traveling wave solutions of the equation.

In particular, we found bifurcations values in terms of the diffusion parameter  $a$  for which new steady-state solutions emerge, and we determined the stability of each steady-state solution found. We were able to numerically estimate the speed of a traveling wave solution given the values of the growth rate and carrying capacity parameters. Finally we showed how a small Gaussian perturbation of the refuge level can lead to the steady-state outbreak level, and estimate how quickly that can happen.

## References

- [Aron et al. 2014] M. Aron, P. Bowers, N. Byer, R. Decker, A. Demirkaya, and J. H. Ryu, "Numerical results on existence and stability of steady state solutions for the reaction-diffusion and Klein-Gordon equations", *Involve* 7:6 (2014), 723–742. MR Zbl
- [Ludwig et al. 1978] D. Ludwig, D. D. Jones, and C. S. Holling, "Qualitative analysis of insect outbreak systems: the spruce budworm and forest", *J. Anim. Ecol.* 47:1 (1978), 315–332.
- [Murray 2005] J. Murray, *Mathematical biology, I: An introduction*, 3rd ed., Interdisciplinary applied mathematics 17, Springer, New York, 2005. Zbl
- [Seydel 2010] R. Seydel, *Practical bifurcation and stability analysis*, 3rd ed., Interdisciplinary Applied Mathematics 5, Springer, 2010. MR Zbl
- [Trefethen 2000] L. N. Trefethen, *Spectral methods in MATLAB*, Software, Environments, and Tools 10, Society for Industrial and Applied Mathematics, Philadelphia, 2000. MR Zbl
- [Williams and Birdsey 2003] D. W. Williams and R. A. Birdsey, "Historical patterns of spruce budworm defoliation and bark beetle outbreaks in North American conifer forests: an atlas and description of digital maps", general technical report NE-308, U.S. Department of Agriculture,



Forest Service, Northeastern Research Station, Newtown Square, PA, 2003, available at <http://www.treesearch.fs.fed.us/pubs/5521>.

Received: 2016-06-06

Accepted: 2016-08-21

alkhalil@stanford.edu

*Stanford University, Stanford, CA 94305, United States*

cbrenna1@ucsc.edu

*UC Santa Cruz, 1156 High Street, Santa Cruz, CA 95064, United States*

rdecker@hartford.edu

*Department of Mathematics, University of Hartford, 200 Bloomfield Ave, West Hartford, CT 06117, United States*

demirkaya@hartford.edu

*Department of Mathematics, University of Hartford, Dano Hall 210, 200 Bloomfield Ave, West Hartford, CT 06117, United States*

nagode@colostate.edu

*Colorado State University, Fort Collins, CO 80523, United States*



# Integer solutions to $x^2 + y^2 = z^2 - k$ for a fixed integer value $k$

Wanda Boyer, Gary MacGillivray, Laura Morrison,  
C. M. (Kieka) Mynhardt and Shahla Nasserar

(Communicated by Chi-Kwong Li)

For a given integer  $k$ , general necessary and sufficient conditions for the existence of integer solutions to an equation of the form  $x^2 + y^2 = z^2 - k$  are given. It is shown that when there is a solution, there are infinitely many solutions. An elementary method for finding the solutions, when they exist, is described.

## 1. Introduction

Finding solutions to quadratic Diophantine equations in three or more variables has been of interest since ancient times. One example is *Pythagoras' equation*  $x^2 + y^2 = z^2$ , which was studied at least 3500 years ago by the Babylonians. Another example is its generalization  $x^2 + y^2 + w^2 = z^2$ , which was completely solved by Catalan [1885] (also see [Ayoub 1984]). A further generalization is the equation  $x^2 + y^2 = z^2 - k$  for a given integer  $k \neq 0$ . Frink [1987] gave a complete solution to the equations of the form  $x^2 + y^2 = z^2 + 1$ . Moreover, solutions to the equation  $x^2 + y^2 = z^2 - k$  with  $k = 1, 2$  were crucial in finding the minimum number of arcs in primitive digraphs with smallest large exponent; see [MacGillivray et al. 2008]. When  $k$  is a perfect square, the solution set can be found using Catalan's method. In the previous reference, the solution set is described when  $k = 1, 2$ .

We study the equation  $x^2 + y^2 = z^2 - k$  for any fixed integer value of  $k$ . It is advantageous to write  $z = x + t$  for some integer  $t$ . Hence we seek solutions  $x, y, t$  to the Diophantine equation

$$x^2 + y^2 = (x + t)^2 - k. \quad (1)$$

We give conditions on  $k$  and  $t$  for which the equation has no solution, and describe an elementary method for finding all solutions to the equation in the cases when

---

*MSC2010*: primary 11D09; secondary 11A07, 11A15.

*Keywords*: Diophantine equations, congruences, residue systems, Pythagorean triples.

they exist. If  $t = 0$  then (1) becomes  $y^2 = -k$ , which has a solution if and only if  $-k$  is a perfect square. Thus in the sequel we consider only nonzero integers  $t$ .

## 2. Background

In an attempt to make this article self-contained, we review some relevant background from elementary number theory. The results and proofs in this section can be found in standard number theory books; for example, see [Apostol 1976; Kumanduri and Romero 1998].

We shall make use of quadratic congruences, that is, congruences of the form  $x^2 \equiv a \pmod{m}$ , for integers  $a$  and  $m$ . The integer  $a$  is a *quadratic residue* modulo  $m$  if the congruence  $x^2 \equiv a \pmod{m}$  has a solution, and a *quadratic nonresidue* modulo  $m$  otherwise.

Suppose  $p$  is an odd prime and  $p$  does not divide  $a$ . The *Legendre symbol*, denoted by  $(a/p)$ , is defined by

$$\left(\frac{a}{p}\right) = \begin{cases} 1 & \text{if } a \text{ is a quadratic residue modulo } p, \\ -1 & \text{if } a \text{ is a quadratic nonresidue modulo } p. \end{cases}$$

**Theorem 1** [Kumanduri and Romero 1998, p. 216]. *Suppose  $p$  is an odd prime which divides neither  $a$  nor  $b$ . Then:*

$$(1) \left(\frac{a^2}{p}\right) = 1.$$

$$(2) \left(\frac{ab}{p}\right) = \left(\frac{a}{p}\right)\left(\frac{b}{p}\right).$$

$$(3) \text{ Euler's criterion: } a^{(p-1)/2} \equiv \left(\frac{a}{p}\right) \pmod{p}.$$

**Proposition 2** [Apostol 1976, p. 181; Kumanduri and Romero 1998, p. 414]. *Suppose  $p$  is an odd prime with  $p \neq 3$ . Then*

$$\left(\frac{-1}{p}\right) = \begin{cases} 1 & \text{if } p \equiv 1 \pmod{4}, \\ -1 & \text{if } p \equiv 3 \pmod{4}, \end{cases} \quad (2)$$

$$\left(\frac{2}{p}\right) = \begin{cases} 1 & \text{if } p \equiv 1, 7 \pmod{8}, \\ -1 & \text{if } p \equiv 3, 5 \pmod{8}, \end{cases} \quad (3)$$

$$\left(\frac{3}{p}\right) = \begin{cases} 1 & \text{if } p \equiv 1, 11 \pmod{12}, \\ -1 & \text{if } p \equiv 5, 7 \pmod{12}. \end{cases} \quad (4)$$

**Proposition 3** [Kumanduri and Romero 1998, p. 428]. *For every odd prime  $p \neq 5$ ,*

$$\left(\frac{5}{p}\right) = \begin{cases} 1 & \text{if } p \equiv 1, 4 \pmod{5}, \\ -1 & \text{if } p \equiv 2, 3 \pmod{5}. \end{cases}$$

We use the notation  $(a, b)$  to denote the *greatest common divisor* of integers  $a$  and  $b$ .

**Proposition 4.** *Suppose  $a, b \in \mathbb{N}$  and  $p$  is a prime, and assume  $k = k_1 p^b$  with  $(k_1, p) = 1$ . Consider the congruence*

$$y^2 \equiv -k \pmod{p^a}. \tag{5}$$

- (1) *If  $b < a$ , then the congruence (5) has an integer solution if and only if  $b$  is even and  $-k_1$  is a quadratic residue modulo  $p^{a-b}$ .*
- (2) *If  $b \geq a$ , then the congruence (5) always has a solution.*

*Proof.* (1) The congruence  $y^2 \equiv -k \pmod{p^a}$  has a solution if and only if there exists an integer  $m$  such that  $m^2 = -k + p^a q = p^b(-k_1 + p^{a-b}q)$ . Since  $p$  does not divide  $-k_1 + p^{a-b}q$ , we have  $p^b \mid m^2$  but  $p^{b+1}$  does not divide  $m^2$ , thus  $b$  is even. Now, divide both sides of  $m^2 = p^b(-k_1 + p^{a-b}q)$  by  $p^b$ . Then  $m_1^2 = -k_1 + p^{a-b}q$ , for some integer  $m_1$ , which implies  $x^2 \equiv -k_1 \pmod{p^{a-b}}$  has a solution. The converse is trivial.

(2) If  $b \geq a$ , then  $y^2 \equiv -k \pmod{p^a}$  has a solution if and only if there exists an integer  $m$  such that  $m^2 = -k + p^a q = p^a(-k_1 p^{b-a} + q)$ . If  $a$  is even, say  $a = 2\beta$  for some integer  $\beta$ , then for any integer  $u$ , any number of the form  $m = \pm u p^\beta$  satisfies  $m^2 = (\pm p^\beta)^2(-k_1 p^{b-a} + u^2 + k_1 p^{b-a})$ . So any such  $m$  with  $0 \leq m \leq p^a - 1$  is a solution to the congruence  $y^2 \equiv -k \pmod{p^a}$ . If  $a = 2\beta + 1$  is odd, then by a similar argument  $m = \pm u p^{\beta+1}$ , with  $0 \leq m \leq p^a - 1$ , is a solution to the congruence  $y^2 \equiv -k \pmod{p^a}$ . □

For any integer  $n > 1$ , and given congruence  $f(x) \equiv 0 \pmod{n}$ , let  $N(n)$  denote the number of solutions to the congruence  $f(x) \equiv 0 \pmod{n}$ .

**Lemma 5** [Apostol 1976, p. 118]. *Suppose  $f(x)$  is a polynomial with integer coefficients. Let  $t = p_1^{e_1} p_2^{e_2} \dots p_r^{e_r}$  be the prime factorization of  $t$ .*

- (1) *The congruence  $f(x) \equiv 0 \pmod{t}$  has a solution if and only if each of the congruences  $f(x) \equiv 0 \pmod{p_i^{e_i}}$ ,  $i = 1, 2, \dots, r$ , has a solution.*
- (2)  $N(t) = \prod_i^r N(p_i^{e_i})$ .

The following results will also be used in solving (1).

**Lemma 6** [Apostol 1976, p. 178]. *If  $p$  is an odd prime and  $p$  does not divide  $k$ , then  $y^2 \equiv -k \pmod{p}$  has either exactly two distinct solutions or no solution.*

**Lemma 7** [Nasserar 2007, p. 38]. *If  $p$  is an odd prime and  $(k, p) = 1$ , then every solution to the congruence  $y^2 \equiv -k \pmod{p^e}$ ,  $e \geq 2$ , generates a solution to the congruence  $y^2 \equiv -k \pmod{p}$  and conversely.*

If the modulus in Lemma 6 is a composite number, we have the following result.

**Lemma 8.** *If  $t = p_1^{e_1} p_2^{e_2} \dots p_r^{e_r}$ , where  $p_1, p_2, \dots, p_r$  are distinct odd primes,  $r, e_i \in \mathbb{N}$ , and  $(k, t) = 1$ , then  $y^2 \equiv -k \pmod{t}$  has  $2^r$  distinct solutions  $y$  if  $-k$*

is a quadratic residue modulo  $p_i$  for each  $p_i$ ,  $i = 1, 2, \dots, r$ , and no solution otherwise.

*Proof.* Suppose for each  $p_i$ ,  $i = 1, 2, \dots, r$  there is a solution to  $y^2 \equiv -k \pmod{p_i}$ . Using Lemma 6, there are exactly two solutions for each congruence. Lemma 5 implies that  $s^2 \equiv -k \pmod{t}$  has exactly  $2^r$  distinct solutions. If one of the congruences  $s^2 \equiv -k \pmod{p_i}$ ,  $i = 1, 2, \dots, r$ , has no solution, then by Lemma 5, the congruence  $y^2 \equiv -k \pmod{t}$  has no solution.  $\square$

The following is a special case of  $y^2 \equiv -k \pmod{p}$  when  $k$  is a perfect square.

**Lemma 9.** *Let  $p$  be an odd prime, and  $a$  be an integer such that  $p$  does not divide  $a$ . Then the congruence  $y^2 \equiv -a^2 \pmod{p}$  has exactly two distinct solutions if  $p \equiv 1 \pmod{4}$  and no solution otherwise.*

*Proof.* The congruence  $y^2 \equiv -a^2 \pmod{p}$  has exactly two distinct solutions if and only if

$$\left(\frac{-a^2}{p}\right) = \left(\frac{-1}{p}\right)\left(\frac{a^2}{p}\right) = 1. \tag{6}$$

Since  $(a^2/p) = 1$ , (6) holds if and only if  $(-1/p) = 1 = (-1)^{(p-1)/2}$  (using Euler’s criterion). The last equation holds if and only if  $p \equiv 1 \pmod{4}$ .  $\square$

### 3. General results

We give solutions to the equation

$$x^2 + y^2 = (x + t)^2 - k.$$

First, we show that it is possible to remove common divisors of  $k$  and  $t$ .

**Proposition 10.** *Suppose  $t$  has prime factorization of the form  $t = \prod_{i=1}^r p_i^{e_i}$  and let  $k = k_1 p_{i_0}^{f_{i_0}}$ , where  $1 \leq i_0 \leq r$  and  $p_{i_0} \nmid k_1$ . Then the equation  $x^2 + y^2 = (x + t)^2 - k$  is equivalent to  $x_1^2 + y_1^2 = (x_1 + t_1)^2 - k_1$ , where  $p_{i_0}^2 \nmid (k_1, t_1)$ .*

*Proof.* We prove the statement for the case  $f_{i_0} \leq e_{i_0}$ . The case  $f_{i_0} > e_{i_0}$  is similar. Depending on whether  $f_{i_0}$  is even or odd we have  $f_{i_0} = 2\alpha + \beta$  with  $\beta = 0, 1$ . Since  $p_{i_0}^{2\alpha} \mid (k, t)$ , if the equation has a solution, then  $p_{i_0}^{2\alpha} \mid y^2$ . Thus, dividing both sides of the equation  $y^2 = 2xt + t^2 - k$  by  $p_{i_0}^{2\alpha}$  implies

$$\left(\frac{y}{p_{i_0}^\alpha}\right)^2 = 2\left(\frac{x}{p_{i_0}^\alpha}\right)\left(\frac{t}{p_{i_0}^\alpha}\right) + \left(\frac{t}{p_{i_0}^\alpha}\right)^2 - \left(\frac{k}{p_{i_0}^{2\alpha}}\right).$$

This is equivalent to  $x_1^2 + y_1^2 = (x_1 + t_1)^2 - k_1 p_{i_0}^\beta$ , and the result follows.  $\square$

In Proposition 10, if  $\alpha = 1$ , in solving the equation we can consider  $k/p_{i_0}^2$  and  $t/p_{i_0}$  instead of  $k$  and  $t$ , respectively. By repeating this process on each common

prime factor  $p$  of  $k$  and  $t$  such that  $p^2 \mid k$  and  $p \mid t$ , we arrive to an equation of the form  $x^2 + y^2 = x^2 + 2xt + t^2 - k$  with a few possibilities for common divisors of  $k$  and  $t$  listed below.

**Lemma 11.** *For every common prime factor  $p$  of  $k$  and  $t$ , (1) can be reduced to an equation of a similar form where  $k$  and  $t$  satisfy one of the following conditions:*

- (1)  $(k, t) = 1$ .
- (2)  $(k, t) = sp$  where  $p$  does not divide  $s$ ,  $p^2$  does not divide  $k$  and  $p^2 \mid t$ .
- (3)  $(k, t) = sp$  where  $p$  does not divide  $s$ ,  $p^2$  does not divide  $k$ , and  $p^2$  does not divide  $t$ .

Therefore, without loss of generality, in solving (1) we may assume that  $k$  and  $t$  satisfy one of the conditions in Lemma 11.

We consider the cases for  $t$  odd and  $t$  even separately.

**3.1. Solutions to  $x^2 + y^2 = (x + t)^2 - k$  when  $t$  is odd.** If  $t$  is odd and  $y$  is a variable, the solutions to  $y^2 \equiv t^2 - k \pmod{2t}$  and  $y^2 \equiv -k \pmod{t}$  are related.

**Lemma 12.** *Suppose  $t$  is odd and  $k$  is an even integer. Then  $m$  is a solution to  $y^2 \equiv t^2 - k \pmod{2t}$  if and only if it is an odd solution to  $y^2 \equiv -k \pmod{t}$ .*

*Proof.* If  $m$  is a solution to  $y^2 \equiv t^2 - k \pmod{2t}$ , then there exists  $q \in \mathbb{Z}$  such that  $m^2 = -k + t(2q + t)$ . Since  $t$  is odd and  $k$  is even,  $m$  is an odd solution to  $y^2 \equiv -k \pmod{t}$ . For the converse, note that if  $m$  is an odd solution to  $y^2 \equiv -k \pmod{t}$ , then  $m$  is a solution to  $y^2 \equiv t^2 - k \pmod{t}$ . Since  $m^2 - t^2 + k$  is even and  $t$  is odd,  $m$  is a solution to  $y^2 \equiv t^2 - k \pmod{2t}$ .  $\square$

In this case, if all solutions to  $y^2 \equiv -k \pmod{t}$  are odd, then they all generate distinct solutions to  $y^2 \equiv t^2 - k \pmod{2t}$ . However, if  $y^2 \equiv -k \pmod{t}$  has an even solution  $v$ , then  $v + t$  is an odd solution to  $y^2 \equiv -k \pmod{t}$  and thus it is a solution to  $y^2 \equiv t^2 - k \pmod{t}$ . That is, for  $t = \prod_{i=1}^r p_i^{e_i}$ , we can choose  $2^r$  distinct solutions to the congruence  $y^2 \equiv -k \pmod{t}$  to be odd, and they will generate  $2^r$  distinct solutions to the congruence  $y^2 \equiv t^2 - k \pmod{2t}$ .

**Lemma 13.** *Suppose  $t$  and  $k$  are odd integers. Then  $m$  is a solution to  $y^2 \equiv t^2 - k \pmod{2t}$  if and only if it is an even solution to  $y^2 \equiv -k \pmod{t}$ .*

*Proof.* If  $m$  is a solution to  $y^2 \equiv t^2 - k \pmod{2t}$ , then  $m$  is even and there exists  $q \in \mathbb{Z}$  such that  $m^2 = -k + t(2q + t)$ . Since  $t$  and  $k$  are odd,  $m$  is an even solution to  $y^2 \equiv -k \pmod{t}$ . If  $m$  is an even solution to  $y^2 \equiv -k \pmod{t}$ , then  $m$  is a solution to  $y^2 \equiv t^2 - k \pmod{t}$ . Now,  $m^2 - t^2 + k$  is even and  $t$  is odd, so  $m$  is a solution to  $y^2 \equiv t^2 - k \pmod{2t}$ .  $\square$

Similarly, in this case, if all solutions to  $y^2 \equiv -k \pmod{t}$  are even, then they all generate distinct solutions to  $y^2 \equiv t^2 - k \pmod{2t}$ . However, if  $y^2 \equiv t^2 - k \pmod{t}$

has an odd solution  $v$ , then  $v + t$  is an even solution to  $y^2 \equiv -k \pmod{t}$  and thus is a solution to  $y^2 \equiv t^2 - k \pmod{2t}$ . Similar to the previous case, we can generate  $2^r$  distinct solutions to the congruence  $y^2 \equiv t^2 - k \pmod{2t}$  by choosing enough even solutions to  $y^2 \equiv -k \pmod{t}$ .

Therefore, when  $t$  is odd, solving the congruence  $y^2 \equiv -k \pmod{t}$  is critical in solving (1). We study the cases of  $(k, t) \neq 1$  and  $(k, t) = 1$  separately.

**Lemma 14.** *Let  $t = \prod_{i=1}^r p_i^{e_i}$  be the prime factorization of  $t$ . Consider the equation  $x^2 + y^2 = (x + t)^2 - k$ :*

- (1) *If  $(k, t) = sp$  and  $p$  does not divide  $s$ ,  $p^2$  does not divide  $k$ , and  $p^2 \mid t$ , for some common prime factor  $p$  of  $k$  and  $t$ , then the equation has no solution.*
- (2) *If the above case does not hold for any common prime factor of  $k$  and  $t$ , and there exists a prime  $p$  such that  $(k, t) = sp$ ,  $p$  does not divide  $s$ ,  $p^2$  does not divide  $k$ , and  $p^2$  does not divide  $t$ , then the equation has a solution if and only if every congruence  $y^2 \equiv -k \pmod{p_i^{e_i}}$  with  $p_i \neq p$  has a solution of the form  $y \equiv 0 \pmod{p}$ .*

*Proof.* (1) In this case, one of the congruences obtained from the congruence  $y^2 \equiv -k \pmod{t}$  is equivalent to  $y^2 \equiv -k_1 p \pmod{p^2}$ , where  $(k_1, p) = 1$ . Using Proposition 4, this congruence has no solution, which implies that (1) has no solution.

(2) In this case, one of the congruences obtained from the congruence  $y^2 \equiv -k \pmod{t}$  is equivalent to  $y^2 \equiv -k_1 p \pmod{p}$ . Using Proposition 4, this congruence always has a solution, namely  $y \equiv 0 \pmod{p}$ . Since  $y^2 \equiv -k \pmod{t}$  has a solution if and only if each of the congruences  $y^2 \equiv -k \pmod{p_i^{e_i}}$  with  $p_i \neq p$  for all other prime divisors of  $t$  has a solution, the result follows. □

Now consider the case where  $(k, t) = 1$  and  $t$  is odd.

Using Lemma 5, if  $t = \prod_{i=1}^r p_i^{e_i}$  is an odd integer, then the congruence  $y^2 \equiv -k \pmod{t}$  is equivalent to the system of congruences

$$\begin{aligned} y^2 &\equiv -k \pmod{p_1^{e_1}}, \\ y^2 &\equiv -k \pmod{p_2^{e_2}}, \\ &\vdots \\ y^2 &\equiv -k \pmod{p_r^{e_r}}. \end{aligned}$$

That is, if one of the above congruences does not have a solution, then the congruence  $y^2 \equiv -k \pmod{t}$  has no solution. Now, if all of the above congruences have solutions, then each congruence can be replaced by a linear congruence, and the resulting system of congruences can be solved using the Chinese remainder theorem.

The following is a consequence of Lemmas 8, 12, and 13.



**Corollary 15.** *If  $t = p_1^{e_1} p_2^{e_2} \dots p_r^{e_r}$ , where  $p_1, p_2, \dots, p_r$  are distinct odd primes and  $r, e_i \in \mathbb{N}$ , then  $y^2 \equiv t^2 - k \pmod{2t}$  has  $2^r$  distinct solutions for  $y$  if  $-k$  is a quadratic residue modulo  $p_i$  for each  $p_i, i = 1, 2, \dots, r$ , and no solution otherwise.*

**Theorem 16.** *Suppose  $t$  is odd and  $k$  is an integer with  $(k, t) = 1$ . The equation  $x^2 + y^2 = (x + t)^2 - k$  has integer solutions  $x, y, t$  if and only if  $-k$  is a quadratic residue modulo  $p_i$  for every prime divisor  $p_i$  of  $t$ . For any such  $t$ , there are infinitely many solutions.*

*Proof.* Suppose  $x^2 + y^2 = (x + t)^2 - k$  has integer solutions  $x, y, t$ . Then,  $y^2 \equiv t^2 - k \pmod{2t}$ , so by Corollary 15,  $-k$  is a quadratic residue modulo every prime divisor of  $t$ .

Now, suppose  $t = p_1^{e_1} p_2^{e_2} \dots p_r^{e_r}$  where  $-k$  is a quadratic residue of every  $p_i$ . By Corollary 15,  $y^2 \equiv t^2 - k \pmod{2t}$  has  $2^r$  distinct solutions. Let  $m$  be such a solution that is also a least residue of  $y$  modulo  $2t$ . Now,  $x, y, t$  with  $y = m + 2tq, x = (y^2 - t^2 + k)/(2t)$ , is a solution to the equation  $x^2 + y^2 = (x + t)^2 - k$  for all  $q \in \mathbb{Z}$ . Therefore, for any such  $t$ , there are infinitely many solutions.  $\square$

The above results give an algorithm for computing the solutions to the equation  $x^2 + y^2 = (x + t)^2 - k$  when  $t$  is odd. To illustrate this algorithm, we present an example for each of the cases  $k \equiv 0, 1, 2, 3 \pmod{4}$ . For this we consider  $k = 12, 5, 6, 15$ , respectively.

**3.1.1. Examples for  $k \equiv 0, 1, 2, 3 \pmod{4}$ .** For the case  $k \equiv 0 \pmod{4}$ , consider the example  $k = 12$ . That is, we want to solve  $x^2 + y^2 = (x + t)^2 - 12$  when  $t$  is odd and has a prime factorization  $t = \prod_{i=1}^r p_i^{e_i}$ . Since  $t$  is odd, the only possibilities for  $(12, t)$  are 3 and 1. First we consider  $(12, t) = 3$ . Using Lemma 14, if  $9 \mid t$ , then there is no solution to the equation; if 9 does not divide  $t$ , then there is a solution to the equation if and only if  $y \equiv 0 \pmod{3}$  and  $y^2 \equiv -12 \pmod{p_i^{e_i}}$  has a solution for each  $p_i \neq 3, i = 1, 2, \dots, r$ . Since  $(12, p_i) = 1$  for  $p_i \neq 3$ , the latter congruence is equivalent to finding whether or not  $-12$  is a quadratic residue modulo each  $p_i^{e_i}$ ; this can be done using Euler's criterion or quadratic reciprocity. The result for each congruence will be a linear congruence and then the Chinese remainder theorem can be used. Now, consider the case  $(12, t) = 1$ .

The parity of  $t$  depends on the parity of  $x$  and the parity of  $y$  as follows:

- If both  $x$  and  $y$  are even, then  $x^2 + y^2 \equiv 0 \pmod{4}$ . This leads to  $(x + t)^2 \equiv 0 \pmod{4}$ , which implies that  $t$  is even.
- If  $x$  and  $y$  are both odd, then  $x^2 + y^2 \equiv 2 \pmod{4}$ . Then  $(x + t)^2 \equiv 2 \pmod{4}$ , which is a contradiction since no square is congruent to 2 modulo 4.
- If  $x$  and  $y$  are of opposite parity, then  $x^2 + y^2 \equiv 1 \pmod{4}$ . This implies that  $(x + t)^2 \equiv 1 \pmod{4}$ , meaning that  $x$  and  $t$  are of opposite parity.

**Proposition 17.** (i) *If  $p \neq 3$  is an odd prime, then  $s^2 \equiv -12 \pmod{p}$  has exactly two distinct solutions if  $p \equiv 1 \pmod{6}$  and no solution otherwise.*

(ii) *If  $t = p_1^{e_1} p_2^{e_2} \dots p_r^{e_r}$ , where  $p_1, p_2, \dots, p_r$  are distinct odd primes, all greater than 3, and  $r, e_i \in \mathbb{N}$ , then  $s^2 \equiv -12 \pmod{t}$  has  $2^r$  distinct solutions if  $p_i \equiv 1 \pmod{6}$  for each  $i = 1, 2, \dots, r$ , and no solution otherwise.*

*Proof.* (i) First suppose  $s^2 \equiv -12 \pmod{p}$  has exactly two distinct solutions. Since  $(4/p) = (2^2/p) = 1$ ,

$$\left(\frac{-12}{p}\right) = \left(\frac{-1}{p}\right) \left(\frac{3}{p}\right) \left(\frac{4}{p}\right) = 1 \implies \left(\frac{-1}{p}\right) \left(\frac{3}{p}\right) = 1.$$

We consider the two cases  $(-1/p) = (3/p) = 1$  and  $(-1/p) = (3/p) = -1$ :

- (1)  $(-1/p) = (3/p) = 1$ . Then, using (2) and (4), we get one of the following:
  - (i)  $p \equiv 1 \pmod{4}$  and  $p \equiv 1 \pmod{12}$ . These congruences imply  $p \equiv 1 \pmod{2}$  and  $p \equiv 1 \pmod{3}$ , respectively. By the Chinese remainder theorem  $p \equiv 1 \pmod{6}$ .
  - (ii)  $p \equiv 1 \pmod{4}$  and  $p \equiv 11 \pmod{12}$ , which is impossible.
- (2)  $(-1/p) = (3/p) = -1$ . Then, using (2) and (4), we get one of the following:
  - (i)  $p \equiv 3 \pmod{4}$  and  $p \equiv 5 \pmod{12}$ , which is impossible.
  - (ii)  $p \equiv 3 \pmod{4}$  and  $p \equiv 7 \pmod{12}$ . These congruences imply  $p \equiv 1 \pmod{2}$  and  $p \equiv 1 \pmod{3}$ , respectively. By the Chinese remainder theorem,  $p \equiv 1 \pmod{6}$ .

For the converse, suppose  $p \equiv 1 \pmod{6}$ . Then either  $p \equiv 1 \pmod{12}$ , which implies  $(-12/p) = (-1/p)(3/p)(4/p) = (1)(1)(1) = 1$ , or  $p \equiv 7 \pmod{12}$ , which implies  $(-12/p) = (-1/p)(3/p)(4/p) = (-1)(-1)(1) = 1$ . In either case,  $s^2 \equiv -12 \pmod{p}$  has exactly two distinct solutions.

(ii) If  $r = 1$ , then the result follows from the Case (1). For  $r > 1$ , suppose that for  $i = 1, 2, \dots, r$ , the prime  $p_i$  is congruent to 1 modulo 6. Then the result follows from the Case (1) and Lemma 8. For the converse, suppose  $s^2 \equiv -12 \pmod{t}$  has exactly  $2^r$  distinct solutions. Then each congruence  $s^2 \equiv -12 \pmod{p_i}$ ,  $i = 1, 2, \dots, r$ , has a solution and by the Case (1),  $p_i$  is congruent to 1 modulo 6 for all  $i = 1, 2, \dots, r$ . □

Also, if  $t = p_1^{e_1} p_2^{e_2} \dots p_r^{e_r}$  where  $p_1, p_2, \dots, p_r$  are distinct odd primes and  $r, e_i \in \mathbb{N}$ , then  $s^2 \equiv t^2 - 12 \pmod{2t}$  has  $2^r$  distinct solutions if each  $p_i \equiv 1 \pmod{6}$  and no solution otherwise.

**Proposition 18.** *Let  $t$  be an odd number with  $(12, t) = 1$ . The equation  $x^2 + y^2 = (x + t)^2 - 12$  has integer solutions for  $x, y, t$  if and only if every prime divisor of  $t$  is congruent to 1 modulo 6. For any such  $t$ , there are infinitely many solutions.*

*Proof.* Note that  $x^2 + y^2 = (x + t)^2 - 12$  implies that  $y^2 \equiv t^2 - 12 \pmod{2t}$ . Then by Lemma 12 and Proposition 17, every prime divisor of  $t$  is congruent to 1 modulo 6. For the converse, suppose  $t = p_1^{e_1} p_2^{e_2} \dots p_r^{e_r}$  with  $p_i \equiv 1 \pmod{6}$  for all  $i = 1, 2, \dots, r$ . By Lemma 12 and Proposition 17,  $y^2 \equiv t^2 - k \pmod{2t}$  has  $2^r$  distinct solutions. Let  $m$  be such a solution that is also a least residue of  $y$  modulo  $2t$ . Then,  $x, y, t$  with  $y = m + 2tq$ ,  $x = (y^2 - t^2 + 12)/(2t)$ , is a solution to the equation  $x^2 + y^2 = (x + t)^2 - 12$  for  $q \in \mathbb{Z}$ . Therefore, for any such  $t$ , there are infinitely many solutions.  $\square$

For the case  $k \equiv 1 \pmod{4}$ , we consider  $k = 5$ . In this case,  $(5, t)$  equals 1 or 5. If  $(5, t) = 5$ , and 25 does not divide  $t$ , then the equation has no solution. If  $(5, t) = 5$ , and  $25 \mid t$ , then the equation has a solution if and only if the following system of equations has a solution:

$$y \equiv 0 \pmod{5} \quad \text{and} \quad y^2 \equiv -5 \pmod{p_i^{e_i}} \quad \text{for all } p_i \neq 5.$$

Similarly to the previous example, this system can be reduced to linear equations. We now consider the case  $(5, t) = 1$ .

The next lemma can be obtained from Proposition 3.

**Lemma 19.** (i) *If  $p \neq 5$  is an odd prime, then  $s^2 \equiv -5 \pmod{p}$  has exactly two distinct solutions if  $p \equiv 1, 3, 7, 9 \pmod{20}$  and no solution otherwise.*

(ii) *If  $t = p_1^{e_1} p_2^{e_2} \dots p_r^{e_r}$  where  $p_1, p_2, \dots, p_r$  are distinct odd primes,  $p_i \neq 5$  for all  $i = 1, 2, \dots, r$ , and  $r, e_i \in \mathbb{N}$ , then  $s^2 \equiv -5 \pmod{t}$  has  $2^r$  distinct solutions modulo  $t$  if each  $p_i \equiv 1, 3, 7, 9 \pmod{20}$  and no solution otherwise.*

We now have the following.

**Proposition 20.** *Suppose  $t$  is odd with  $(5, t) = 1$ . The equation*

$$x^2 + y^2 = (x + t)^2 - 5$$

*has integer solutions  $x, y, t$  if and only if every prime divisor of  $t$  is congruent to 1, 3, 7, 9 modulo 20. For any such  $t$  there are infinitely many solutions.*

For the case  $k \equiv 2 \pmod{4}$  we consider  $k = 6$ . In this case, since  $t$  is odd, we have either  $(6, t) = 3$  or  $(6, t) = 1$ . If  $9 \mid t$ , there is no solution; if 9 does not divide  $t$ , then the equation has a solution if and only if there is a solution to

$$y \equiv 0 \pmod{3} \quad \text{and} \quad y^2 \equiv -6 \pmod{p_i^{e_i}} \quad \text{for all } p_i \neq 3.$$

Hence we consider the case when  $(6, t) = 1$ . We shall use a lemma which follows from (3).

**Lemma 21.** (i) *If  $p \neq 3$  is an odd prime, then  $s^2 \equiv -6 \pmod{p}$  has exactly two distinct solutions if  $p \equiv 1, 5, 7, 11 \pmod{24}$  and no solution otherwise.*

(ii) If  $t = p_1^{e_1} p_2^{e_2} \dots p_r^{e_r}$ , where  $p_1, p_2, \dots, p_r$  are distinct odd primes,  $(6, t) = 1$  and  $r, e_i \in \mathbb{N}$ , then  $s^2 \equiv -6 \pmod{t}$  has  $2^r$  distinct solutions modulo  $t$  if each  $p_i \equiv 1, 5, 7, 11 \pmod{24}$  and no solution otherwise.

**Proposition 22.** *The equation  $x^2 + y^2 = (x + t)^2 - 6$  with  $(t, 6) = 1$  has integer solutions  $x, y, t$  if and only if every prime divisor of  $t$  is congruent to  $1, 5, 7, 11$  modulo  $24$ . For any such  $t$  there are infinitely many solutions.*

Finally,  $k = 15$  is considered as an example for the case  $k \equiv 3 \pmod{4}$ . The cases when  $(15, t) = 3, 5, 15$  are similar to the previous examples. We only consider the case when  $(15, t) = 1$ .

**Lemma 23.** (i) *If  $p \geq 7$  is an odd prime, then  $s^2 \equiv -15 \pmod{p}$  has exactly two distinct solutions if  $p \equiv 1, 7, 17, 19, 23, 31, 43, 47, 49, 53 \pmod{60}$  and no solution otherwise.*

(ii) *If  $t = p_1^{e_1} p_2^{e_2} \dots p_r^{e_r}$  where  $p_1, p_2, \dots, p_r$  are distinct odd primes,  $(15, t) = 1$ , and  $r, e_i \in \mathbb{N}$ , then  $s^2 \equiv -15 \pmod{t}$  has  $2^r$  distinct solutions modulo  $t$  if each  $p_i$  is congruent to  $1, 7, 17, 19, 23, 31, 43, 47, 49, 53$  modulo  $60$ , and no solution otherwise.*

**Proposition 24.** *The equation  $x^2 + y^2 = (x + t)^2 - 15$  with  $(15, t) = 1$  has integer solutions  $x, y, t$  if and only if every prime divisor of  $t$  is congruent to  $1, 7, 17, 19, 23, 31, 43, 47, 49, 53$  modulo  $60$ . For any such  $t$  there are infinitely many solutions.*

**3.2. Solutions to  $x^2 + y^2 = (x + t)^2 - k$  when  $t$  is even.** Now we consider the equation  $x^2 + y^2 = (x + t)^2 - k$  when  $t$  is even.

**Proposition 25.** *Let  $k, t$  be integers and suppose  $t$  is even. Then  $m$  is a solution to the congruence  $y^2 \equiv t^2 - k \pmod{2t}$  if and only if it is a solution to the congruence  $y^2 \equiv -k \pmod{2t}$ .*

*Proof.* Note that since  $t$  is even,  $2t \mid t^2$ . Now,  $m$  is a solution for  $y^2 \equiv t^2 - k \pmod{2t}$  if and only if  $2t \mid (m^2 - t^2 + k)$  if and only if  $2t \mid (m^2 + k)$ . □

Thus, in this section our focus is on congruences of the form  $y^2 \equiv t^2 - k \pmod{2t}$ . We first show that when  $p = 2$ , there is no solution to (1) in Case (2) or Case (3) of Lemma 11.

**Lemma 26.** *Consider integers  $k, t$ :*

- (1) *If  $2 \mid (k, t)$  but  $4$  does not divide  $k$ , and  $4 \mid t$ , then the congruence  $y^2 \equiv -k \pmod{2t}$  has no solution.*
- (2) *If  $2 \mid (k, t)$  but  $4$  divides neither  $k$  nor  $t$ , then the congruence  $y^2 \equiv -k \pmod{2t}$  has no solution.*

*Proof.* (1) Suppose  $k = 2\alpha$  and  $t = 4\beta$  for some integers  $\alpha$  and  $\beta$ , where  $\alpha$  is odd. The congruence  $y^2 \equiv -k \pmod{2t}$  has a solution if and only if there exist integers  $m, q$  such that  $m^2 = -2(\alpha + 4\beta q)$ . This is not possible since  $\alpha + 4\beta q$  is odd.

(2) Suppose  $k = 2\alpha$  and  $t = 2\beta$  for some odd integers  $\alpha$  and  $\beta$ . As above, the congruence  $y^2 \equiv -k \pmod{2t}$  has a solution if and only if there exist integers  $m, q$  such that  $m^2 = -2(\alpha + 2\beta q)$ . This is not possible since  $\alpha + 2\beta q$  is odd.  $\square$

If 2 does not divide  $(k, t)$  but  $(k, t) \neq 1$ , the same argument as the case of  $t$  odd can be used. Thus, without loss of generality, we can assume that  $(k, t) = 1$ . This implies that  $k$  is odd. Let  $t = 2^r s$  where  $r \in \mathbb{N}$  and  $s = \prod_{i=1}^u p_i^{e_i}$  is an odd integer. Since  $(2^{r+1}, s) = 1$ , using Lemma 5, the congruence  $y^2 \equiv -k \pmod{2t}$  can be reduced to two congruences:

$$y^2 \equiv -k \pmod{2^{r+1}}, \quad \text{and} \quad y^2 \equiv -k \pmod{s}.$$

The congruence  $y^2 \equiv -k \pmod{s}$  can be solved using the results from the previous section. We now consider different cases for  $r$  for the remaining congruence  $y^2 \equiv -k \pmod{2^{r+1}}$ .

The following result can be found in most number theory books; see [Kumanduri and Romero 1998, p. 231] for example. We restate it using the notation used in this work.

**Lemma 27** [Kumanduri and Romero 1998, p. 231]. *Suppose  $k$  is odd and  $r \geq 1$ . Consider the congruence*

$$y^2 \equiv -k \pmod{2^{r+1}}. \tag{7}$$

- (1) *If  $r = 1$ , the congruence (7) has exactly two distinct solutions if  $-k \equiv 1 \pmod{4}$  and no solution otherwise.*
- (2) *If  $r \geq 2$ , the congruence (7) has exactly four distinct solutions if  $-k \equiv 1 \pmod{8}$  and no solution otherwise. If  $y_0$  is a solution, then  $-y_0$  and  $\pm y_0 + 2^r$  are also solutions.*

An application of the above results can solve (1) when  $t$  is even, as follows.

**Theorem 28.** *Assume  $k$  is odd and consider (1) with  $t = 2^r s$ , where  $s$  is an odd integer and  $r > 0$ :*

- (1) *If  $r = 1$ , then (1) has a solution if and only if  $-k \equiv 1 \pmod{4}$  and  $y^2 \equiv -k \pmod{s}$  has a solution.*
- (2) *If  $r \geq 2$ , then (1) has a solution if and only if  $-k \equiv 1 \pmod{8}$  and  $y^2 \equiv -k \pmod{s}$  has a solution.*

*In each case, if there is one solution, there are infinitely many solutions.*

*Proof.* Using Proposition 25, we know that  $x^2 + y^2 = (x + t)^2 - k$  has a solution if and only if  $y^2 \equiv -k \pmod{2t}$  has a solution. The conditions for the existence of a solution in each case follow from Lemma 27 and the discussion preceding it. Now, suppose  $m$  is a solution to  $y^2 \equiv -k \pmod{2t}$ . Using Proposition 25, we see that it is also a solution to  $y^2 \equiv t^2 - k \pmod{2t}$ . Thus, the triple  $(x, y, t)$  with  $y = m + 2tq$ ,  $x = (y^2 - t^2 + k)/(2t)$ , is a solution to the equation  $x^2 + y^2 = (x + t)^2 - k$  for all  $q \in \mathbb{Z}$ . Since  $q$  can be chosen arbitrarily, there are infinitely many solutions.  $\square$

## References

- [Apostol 1976] T. M. Apostol, *Introduction to analytic number theory*, Springer, 1976. MR Zbl
- [Ayoub 1984] A. B. Ayoub, “Integral solutions to the equation  $x^2 + y^2 + z^2 = u^2$ : a geometrical approach”, *Math. Mag.* **57**:4 (1984), 222–223. MR Zbl
- [Catalan 1885] E. Catalan, “Questions d’analyse indéterminée”, *Bull. Acad. Roy. Sci. Belgique* (3) **9** (1885), 531–534. JFM
- [Frink 1987] O. Frink, “Almost Pythagorean triples”, *Math. Mag.* **60**:4 (1987), 234–236. MR Zbl
- [Kumanduri and Romero 1998] R. Kumanduri and C. Romero, *Number theory with computer applications*, Prentice Hall, Upper Saddle River, NJ, 1998. Zbl
- [MacGillivray et al. 2008] G. MacGillivray, S. Nasserar, D. D. Olesky, and P. van den Driessche, “Primitive digraphs with smallest large exponent”, *Linear Algebra Appl.* **428**:7 (2008), 1740–1752. MR Zbl
- [Nasserar 2007] S. Nasserar, *Primitive digraphs with smallest large exponent*, master’s thesis, University of Victoria, 2007, available at <http://hdl.handle.net/1828/184>.

Received: 2016-07-27      Accepted: 2016-09-25

wbkboyer@uvic.ca	<i>Department of Mathematics and Statistics, University of Victoria, P.O. Box 1700 STN CSC, Victoria BC V8W 2Y2, Canada</i>
gmacgill@uvic.ca	<i>Department of Mathematics and Statistics, University of Victoria, P.O. Box 1700 STN CSC, Victoria BC V8W 2Y2, Canada</i>
laura.may.morrison@gmail.com	<i>Department of Mathematics and Statistics, University of Victoria, P.O. Box 1700 STN CSC, Victoria BC V8W 2Y2, Canada</i>
kieka@uvic.ca	<i>Department of Mathematics and Statistics, University of Victoria, P.O. Box 3060 STN CSC, Victoria BC V8W 3R4, Canada</i>
snasserar@nova.edu	<i>Department of Mathematics, Nova Southeastern University, Fort Lauderdale, FL 33324, United States</i>

# A solution to a problem of Frechette and Locus

Chenthuran Abeyakaran

(Communicated by Ken Ono)

In a recent paper, Frechette and Locus examined and found expressions for the infinite product  $D_m(q) := \prod_{t=1}^{\infty} (1 - q^{mt}) / (1 - q^t)$  in terms of products of  $q$ -series of the Rogers–Ramanujan type coming from Hall–Littlewood polynomials, when  $m \equiv 0, 1, 2 \pmod{4}$ . These  $q$ -series were originally discovered in 2014 by Griffin, Ono, and Warnaar in their work on the framework of the Rogers–Ramanujan identities. Extending this framework, Rains and Warnaar also recently discovered more  $q$ -series and their corresponding infinite products. Frechette and Locus left open the case where  $m \equiv 3 \pmod{4}$ . Here we find such an expression for the infinite products for  $m \equiv 3 \pmod{4}$  by making use of the new  $q$ -series obtained by Rains and Warnaar.

## 1. Introduction

The Rogers–Ramanujan identities [Andrews 1971]

$$G(q) := \sum_{n=0}^{\infty} \frac{q^{n^2}}{(1-q)(1-q^2)\dots(1-q^n)} = \prod_{n=0}^{\infty} \frac{1}{(1-q^{5n+1})(1-q^{5n+4})}, \quad (1-1)$$

$$H(q) := \sum_{n=0}^{\infty} \frac{q^{n^2+n}}{(1-q)(1-q^2)\dots(1-q^n)} = \prod_{n=0}^{\infty} \frac{1}{(1-q^{5n+2})(1-q^{5n+3})}, \quad (1-2)$$

have inspired research and discoveries in many areas of mathematics and physics, such as modular forms and elliptic curves, conformal field theory, knot theory, probability, and statistical mechanics. (See next citation for some discussion.) Given the importance of these identities, it had been an open problem for nearly a century to build a theory suggested by these two Rogers–Ramanujan identities. In 2014 Griffin, Ono, and Warnaar [Griffin et al. 2016] discovered<sup>1</sup> a more general framework for identities similar to that of Rogers–Ramanujan, where an infinite sum, defined using Hall–Littlewood polynomials  $P_{\lambda}(x; q)$ , is equal to an infinite product with periodic exponents.

*MSC2010:* 11P84.

*Keywords:* Rogers–Ramanujan Identities.

<sup>1</sup>Their work was named the 15th top story in science in 2014 by *Discover* magazine.

In order to define the Hall–Littlewood polynomials, we recall the definition of an integer partition and the following notation. A *partition* is a nonincreasing sequence of nonnegative integers with finitely many nonzero terms. For a partition  $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_n)$ , we define the *weight* of the partition to be  $|\lambda| := \lambda_1 + \lambda_2 + \dots + \lambda_n$  and the *length* of the partition  $\lambda$  to be  $n$ . In addition, we let  $2\lambda := (2\lambda_1, 2\lambda_2, \dots, 2\lambda_n)$ . Let  $m_i$  denote the multiplicity of size  $i$  parts. Also, let  $(q)_k = (q; q)_k$  denote the  $q$ -Pochhammer symbol, which is defined as follows:

$$(a)_k := (a; q)_k = \begin{cases} (1 - a)(1 - aq)(1 - aq^2) \cdots (1 - aq^{k-1}) & \text{if } k \geq 0, \\ \prod_{n=0}^{\infty} (1 - aq^n) & \text{if } k = \infty. \end{cases}$$

If  $\lambda$  has length  $n$ , the Hall–Littlewood polynomial is a symmetric function in  $n$  variables, namely  $x_1, x_2, \dots, x_n$ , defined as

$$P_\lambda(x; q) = \frac{1}{v_\lambda(q)} \sum_{w \in S_n} w \left( x^\lambda \prod_{i < j} \frac{x_i - qx_j}{x_i - x_j} \right), \tag{1-3}$$

where  $x^\lambda := x_1^{\lambda_1} x_2^{\lambda_2} x_3^{\lambda_3} \dots x_n^{\lambda_n}$ ,  $v_\lambda(q) := \prod_{i=0}^n (q)_{m_i} / (1 - q)^{m_i}$ , and the symmetric group  $S_n$  acts on  $x$  by permuting all the  $x_i$ .

For ordered pairs  $v = (c, d) \in \{(1, -1), (2, -1), (1, 0), (2, -2)\}$  and arbitrary  $a, b \geq 1$ , Griffin, Ono, and Warnaar [Griffin et al. 2016], and more recently Rains and Warnaar [2015], defined the  $q$ -series

$$R_v(a, b; q) = \sum_{\lambda, \lambda_1 \leq a} q^{c|\lambda|} P_{2\lambda}(1, q, q^2, \dots; q^{2b+d}), \tag{1-4}$$

$$S(a, b; q) = \sum_{\lambda, \lambda_1 \leq a} q^{|\lambda|/2} P_\lambda(1, q, q^2, \dots; q^{2b}), \tag{1-5}$$

and

$$T(a, b; q) = \sum_{\lambda, \lambda_1 \leq a} q^{|\lambda|} \left( \prod_{i=1}^{a-1} (-q^{b-1/2}; q^{b-1/2})_{m_i}(\lambda) \right) P_\lambda(1, q, q^2, \dots; q^{2b-1}). \tag{1-6}$$

Here the  $P_\lambda(1, q, q^2, \dots; q^T)$  are Hall–Littlewood  $q$ -series in infinitely many parameters. To define these Hall–Littlewood  $q$ -series, we must first express the Hall–Littlewood polynomial  $P_\lambda(x_1, x_2, \dots, x_n; q^T)$  in terms of the  $r$ -th power sum symmetric function,  $x_1^r + x_2^r + \dots + x_n^r$ . This is possible due to a well-known fact in abstract algebra which states that every symmetric polynomial can be written as a sum of products of  $r$ -th power sum symmetric functions with rational coefficients [Macdonald 1995]. Now we obtain the Hall–Littlewood polynomial  $P_\lambda(1, q, q^2, \dots; q^T)$  by replacing  $(x_1^r + x_2^r + \dots + x_n^r)$  with  $1^r + q^r + q^{2r} + \dots = 1/(1 - q^r)$ .

To motivate our work, consider another interesting property of the Roger–Ramanujan identities. When we take the product of both Rogers–Ramanujan



identities, we can see that

$$\begin{aligned} G(q) \cdot H(q) &= \prod_{n=0}^{\infty} \frac{1}{(1-q^{5n+1})(1-q^{5n+4})} \cdot \prod_{n=0}^{\infty} \frac{1}{(1-q^{5n+2})(1-q^{5n+3})} \\ &= \prod_{n=0}^{\infty} \frac{1}{(1-q^{5n+1})(1-q^{5n+4})(1-q^{5n+2})(1-q^{5n+3})} \\ &= \prod_{n=1}^{\infty} \frac{(1-q^{5n})}{(1-q^n)}. \end{aligned}$$

Inspired by this emergence of this infinite product, Frechette and Locus [2016] explored the natural question of more generalized products of  $q$ -series of the Rogers–Ramanujan type that result in an infinite product of the form

$$D_m(q) := \prod_{t=1}^{\infty} \frac{(1-q^{mt})}{(1-q^t)}. \tag{1-7}$$

Making use of the Rogers–Ramanujan framework of [Griffin et al. 2016], Frechette and Locus obtained explicit formulas for  $D_m(q)$  when  $m \equiv 0, 1, 2 \pmod{4}$ . When  $m \geq 8$  and is even, they found that

$$D_m(q) = \frac{R_{(1,0)}\left(2, \frac{m}{2} - 3; q\right) \cdot R_{(2,-2)}\left(\frac{m}{2}, 2; q\right)}{R_{(2,-2)}\left(\frac{m}{2} - 3, 3; q\right)}. \tag{1-8}$$

They also found for  $m \equiv 1 \pmod{4}$  and  $m > 1$ ,

$$D_m(q) = \frac{R_{(1,-1)}\left(\frac{m-1}{2} - 1, 1; q\right) \cdot R_{(2,-1)}\left(\frac{m-1}{4}, \frac{m-1}{4}; q\right)}{R_{(2,-1)}\left(\frac{m-1}{4} + 1, \frac{m-1}{4} - 1; q\right)}. \tag{1-9}$$

However, they were unable to construct  $D_m(q)$  for positive integers  $m \equiv 3 \pmod{4}$ . Their difficulty arose from the fact that the Rogers–Ramanujan framework of [Griffin et al. 2016]. Rains and Warnaar [2015] recently found this extension. In this paper, we address this case and provide such a formula.

**Theorem 1.1.** *If  $m \equiv 3 \pmod{4}$  and  $m \geq 7$ , we have*

$$D_m(q) = \frac{T\left(\frac{m+1}{2}, \frac{m+1}{4}; q\right) S\left(\frac{m+1}{2}, \frac{m+1}{4} - 1; q\right) R_{(1,-1)}\left(\frac{m-1}{2} - 1, 1; q\right)}{T\left(\frac{m+1}{2} + 2, \frac{m+1}{4} - 1; q\right) S\left(\frac{m+1}{2} - 2, \frac{m+1}{4}; q\right)}.$$

In Section 2, we cover preliminaries on  $q$ -series and state the results of [Griffin et al. 2016; Rains and Warnaar 2015; Frechette and Locus 2016]. In Section 3, we use these results to prove Theorem 1.1.

### 2. Preliminaries

In order to simplify the products involved in writing  $R_\nu(a, b; q)$ ,  $S(a, b; q)$ , and  $T(a, b; q)$ , we use a modified theta function

$$\theta(a; q) := (a; q)_\infty (q/a; q)_\infty, \tag{2-1}$$

where  $(a; q)_\infty$  denotes the  $q$ -Pochhammer symbol we previously defined. We then define

$$\theta(a_1, a_2, \dots, a_n; q) = \theta(a_1; q) \cdot \theta(a_2; q) \cdots \theta(a_n; q). \tag{2-2}$$

**Theorem 2.1** [Griffin et al. 2016, Theorem 1.1]. *If  $a$  and  $b$  are positive integers and  $\kappa := 2a + 2b + 1$ , we have*

$$\begin{aligned} R_{(1,-1)}(a, b; q) &:= \sum_{\lambda, \lambda_1 \leq a} q^{|\lambda|} P_{2\lambda}(1, q, q^2, \dots; q^{2b-1}) \\ &= \frac{(q^\kappa; q^\kappa)_\infty^b}{(q)_\infty^b} \cdot \prod_{i=1}^b \theta(q^{i+a}; q^\kappa) \prod_{1 \leq i < j \leq b} \theta(q^{j-i}, q^{i+j-1}; q^\kappa) \\ &= \frac{(q^\kappa; q^\kappa)_\infty^a}{(q)_\infty^a} \cdot \prod_{i=1}^a \theta(q^{i+1}; q^\kappa) \prod_{1 \leq i < j \leq a} \theta(q^{j-i}, q^{i+j+1}; q^\kappa). \end{aligned}$$

This result has generalizations to the functions  $S$  and  $T$ :

**Theorem 2.2** [Rains and Warnaar 2015, Theorem 5.10]. *If  $a$  and  $b$  are positive integers and  $\kappa := a + 2b + 1$ , we have*

$$\begin{aligned} S(a, b; q) &:= \sum_{\lambda, \lambda_1 \leq a} q^{|\lambda|/2} P_\lambda(1, q, q^2, \dots; q^{2b}) \\ &= \frac{(q^\kappa; q^\kappa)_\infty^{b-1} (q^{\kappa/2}; q^{\kappa/2})_\infty}{(q)_\infty^{b-1} (q^{1/2}; q^{1/2})_\infty} \prod_{i=1}^b \theta(q^i; q^{\kappa/2}) \prod_{1 \leq i < j \leq b} \theta(q^{j-i}, q^{i+j}; q^\kappa). \end{aligned}$$

**Theorem 2.3** [Rains and Warnaar 2015, Remark 5.13]. *If  $a$  and  $b$  are positive integers and  $\kappa := a + 2b - 1$ , we have*

$$\begin{aligned} T(a, b; q) &:= \sum_{\lambda, \lambda_1 \leq a} q^{|\lambda|} \left( \prod_{i=1}^{a-1} (-q^{b-1/2}; q^{b-1/2})_{m_i(\lambda)} \right) P_\lambda(1, q, q^2, \dots; q^{2b-1}) \\ &= \frac{(q^\kappa; q^\kappa)_\infty^b}{(q)_\infty^{b-1} (q^{1/2}; q^{1/2})_\infty} \prod_{i=1}^b \theta(q^{i-1/2}; q^\kappa) \prod_{1 \leq i < j \leq b} \theta(q^{j-i}, q^{i+j-1}; q^\kappa). \end{aligned}$$

To prove our result, we combine Theorem 2.3 with the following proposition.

**Proposition 2.4.** *If  $i \in \mathbb{Z}^+$ , we have*

$$\theta(q^i; q^{m/2}) = \theta(q^i; q^m)\theta(q^{m/2-i}; q^m).$$

*Proof.* Using the definition of the modified theta function, we have

$$\begin{aligned} \theta(q^i; q^{m/2}) &= (q^i; q^{m/2})_\infty (q^{m/2-i}; q^{m/2})_\infty \\ &= \prod_{n=0}^\infty (1 - q^i \cdot q^{mn/2}) \prod_{n=0}^\infty (1 - q^{m/2-i} \cdot q^{mn/2}). \end{aligned}$$

Separating terms in the infinite product on the right-hand side based on the parity of  $n$ , we have

$$\begin{aligned} \theta(q^i; q^{m/2}) &= \prod_{n=0}^\infty (1 - q^i \cdot q^{mn}) (1 - q^{i+m/2} \cdot q^{mn}) \prod_{n=0}^\infty (1 - q^{m/2-i} \cdot q^{mn}) (1 - q^{m-i} \cdot q^{mn}) \\ &= \prod_{n=0}^\infty (1 - q^i \cdot q^{mn}) (1 - q^{m-i} \cdot q^{mn}) (1 - q^{m/2-i} \cdot q^{mn}) (1 - q^{i+m/2} \cdot q^{mn}) \\ &= (q^i; q^m)_\infty (q^{m-i}; q^m)_\infty (q^{m/2-i}; q^m)_\infty (q^{m/2+i}; q^m)_\infty \\ &= \theta(q^i; q^m)\theta(q^{m/2-i}; q^m). \quad \square \end{aligned}$$

### 3. Proof of Theorem 1.1

We shall now prove Theorem 1.1. By Theorem 2.3, when  $m = a + 2b - 1$ , we have

$$\begin{aligned} T\left(\frac{m+1}{2}, \frac{m+1}{4}; q\right) &= \frac{(q^m; q^m)_{\infty}^{(m+1)/4}}{(q)_{\infty}^{(m+1)/4-1} (q^{1/2}; q^{1/2})_{\infty}} \prod_{i=1}^{(m+1)/4} \theta(q^{i-1/2}; q^m) \\ &\quad \prod_{1 \leq i < j \leq (m+1)/4} \theta(q^{j-i}, q^{i+j-1}; q^m) \quad (3-1) \end{aligned}$$

and

$$\begin{aligned} T\left(\frac{m+1}{2} + 2, \frac{m+1}{4} - 1; q\right) &= \frac{(q^m; q^m)_{\infty}^{(m-3)/4}}{(q)_{\infty}^{(m-7)/4} (q^{1/2}; q^{1/2})_{\infty}} \prod_{i=1}^{(m-3)/4} \theta(q^{i-1/2}; q^m) \\ &\quad \prod_{1 \leq i < j \leq (m-3)/4} \theta(q^{j-i}, q^{i+j-1}; q^m), \quad (3-2) \end{aligned}$$

which gives

$$\begin{aligned}
 & \frac{T\left(\frac{m+1}{2}, \frac{m+1}{4}; q\right)}{T\left(\frac{m+1}{2} + 2, \frac{m+1}{4} - 1; q\right)} \\
 &= \frac{(q^m; q^m)_\infty}{(q)_\infty} \theta(q^{(m-1)/4}; q^m) \prod_{i=1}^{(m+1)/4-1} \theta(q^{(m+1)/4-i}, q^{(m+1)/4-1+i}; q^m) \\
 &= \frac{(q^m; q^m)_\infty}{(q)_\infty} \theta(q^{(m-1)/4}; q^m) \prod_{i=1}^{(m+1)/4-1} \theta(q^i; q^m) \prod_{i=(m+1)/4}^{(m+1)/2-2} \theta(q^i; q^m) \\
 &= \frac{(q^m; q^m)_\infty}{(q)_\infty} \theta(q^{(m-1)/4}; q^m) \prod_{i=1}^{(m-1)/2} \theta(q^i; q^m). \tag{3-3}
 \end{aligned}$$

Using Theorem 2.2 and Proposition 2.4, we have

$$\begin{aligned}
 & S\left(\frac{m+1}{2} - 2, \frac{m+1}{4}; q\right) \\
 &= \frac{(q^m; q^m)_\infty^{(m+1)/4-1} (q^{m/2}; q^{m/2})_\infty}{(q)_\infty^{(m+1)/4-1} (q^{1/2}; q^{1/2})_\infty} \prod_{i=1}^{(m+1)/4} \theta(q^i; q^m) \theta(q^{m/2-i}; q^m) \\
 & \quad \prod_{1 \leq i < j \leq (m+1)/4} \theta(q^{j-i}, q^{i+j}; q^m), \tag{3-4}
 \end{aligned}$$

and

$$\begin{aligned}
 & S\left(\frac{m+1}{2}, \frac{m+1}{4} - 1; q\right) \\
 &= \frac{(q^m; q^m)_\infty^{(m+1)/4-2} (q^{m/2}; q^{m/2})_\infty}{(q)_\infty^{(m+1)/4-2} (q^{1/2}; q^{1/2})_\infty} \prod_{i=1}^{(m+1)/4-1} \theta(q^i; q^m) \theta(q^{m/2-i}; q^{m/2}) \\
 & \quad \prod_{1 \leq i < j \leq (m+1)/4-1} \theta(q^{j-i}, q^{i+j}; q^m). \tag{3-5}
 \end{aligned}$$

Now, we evaluate the following quotient:

$$\begin{aligned}
 & \frac{S\left(\frac{m+1}{2} - 2, \frac{(m+1)}{4}; q\right)}{S\left(\frac{m+1}{2}, \frac{m+1}{4} - 1; q\right)} \\
 &= \frac{(q^m; q^m)_\infty}{(q)_\infty} \theta(q^{(m+1)/4}; q^m) \theta(q^{(m-1)/4}; q^m) \\
 & \quad \prod_{i=1}^{(m+1)/4-1} \theta(q^{(m+1)/4-i}, q^{(m+1)/4+i}; q^m)
 \end{aligned}$$

$$\begin{aligned}
 &= \frac{(q^m; q^m)_\infty}{(q)_\infty} \theta(q^{(m+1)/4}; q^m) \theta(q^{(m-1)/4}; q^m) \\
 &\quad \prod_{i=1}^{(m+1)/4-1} \theta(q^i; q^m) \prod_{i=(m+1)/4+1}^{(m+1)/2-1} \theta(q^i; q^m) \\
 &= \frac{(q^m; q^m)_\infty}{(q)_\infty} \theta(q^{(m-1)/4}; q^m) \prod_{i=1}^{(m-1)/2} \theta(q^i; q^m). \tag{3-6}
 \end{aligned}$$

Dividing (3-3) by (3-6), we obtain

$$\begin{aligned}
 &\frac{T\left(\frac{m+1}{2}, \frac{m+1}{4}; q\right) S\left(\frac{m+1}{2}, \frac{m+1}{4} - 1; q\right)}{T\left(\frac{m+1}{2} + 2, \frac{m+1}{4} - 1; q\right) S\left(\frac{m+1}{2} - 2, \frac{m+1}{4}; q\right)} \\
 &= \frac{(q^m; q^m)_\infty / (q)_\infty \theta(q^{(m-1)/4}; q^m) \prod_{i=1}^{(m-3)/2} \theta(q^i; q^m)}{(q^m; q^m)_\infty / (q)_\infty \theta(q^{(m-1)/4}; q^m) \prod_{i=1}^{(m-1)/2} \theta(q^i; q^m)} \tag{3-7} \\
 &= \frac{1}{\theta(q^{(m-1)/2}; q^m)}.
 \end{aligned}$$

By Theorem 2.1, we have the identity

$$R_{(1,-1)}\left(\frac{m-1}{2} - 1, 1; q\right) = \frac{(q^m; q^m)_\infty}{(q)_\infty} \theta(q^{(m-1)/2}; q^m). \tag{3-8}$$

Multiplying (3-8) and (3-7) gives us our desired result:

$$\begin{aligned}
 &\frac{T\left(\frac{m+1}{2} + 1, \frac{m+1}{4}; q\right) S\left(\frac{m+1}{2}, \frac{m+1}{4} - 1; q\right) R_{(1,-1)}\left(\frac{m-1}{2} - 1, 1; q\right)}{T\left(\frac{m+1}{2} + 3, \frac{m+1}{4} - 1; q\right) S\left(\frac{m+1}{2} - 2, \frac{m+1}{4}; q\right)} \\
 &= \frac{(q^m; q^m)_\infty}{(q)_\infty} \theta(q^{(m-1)/2}; q^m) \cdot \frac{1}{\theta(q^{(m-1)/2}; q^m)} \\
 &= \frac{(q^m; q^m)_\infty}{(q)_\infty} = \prod_{t=1}^\infty \frac{(1-q^{mt})}{(1-q^t)} = D_m(q). \quad \square
 \end{aligned}$$

**Acknowledgements**

I would like to thank Ken Ono for introducing me to this problem and for his guidance throughout this project. I would also like to thank Tessa Cotron, Sarah Fleming, and Robert Dicks for proofreading this paper and checking my work.

**References**

[Andrews 1971] G. E. Andrews, *Number theory*, Saunders, Philadelphia, 1971. Corrected reprint by Dover, New York, 1994. MR Zbl

- [Frechette and Locus 2016] C. Frechette and M. Locus, “Combinatorial properties of Rogers–Ramanujan type identities arising from Hall–Littlewood polynomials”, *Ann. Comb.* **20**:2 (2016), 345–360. MR Zbl
- [Griffin et al. 2016] M. J. Griffin, K. Ono, and S. O. Warnaar, “A framework of Rogers–Ramanujan identities and their arithmetic properties”, *Duke Math. J.* **165**:8 (2016), 1475–1527. MR Zbl
- [Macdonald 1995] I. G. Macdonald, *Symmetric functions and Hall polynomials*, 2nd ed., Oxford University Press, New York, 1995. MR Zbl
- [Rains and Warnaar 2015] E. M. Rains and S. O. Warnaar, “Bounded Littlewood identities”, preprint, 2015. arXiv

Received: 2016-07-29      Revised: 2016-08-18      Accepted: 2016-08-21

ChenthuranJA@gmail.com      *Chamblee Charter High School, 3688 Chamblee Dunwoody Rd,  
Chamblee, GA 30341, 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

2017 vol. 10 no. 5

Algorithms for finding knight's tours on Aztec diamonds	721
SAMANTHA DAVIES, CHENXIAO XUE AND CARL R. YERGER	
Optimal aggression in kleptoparasitic interactions	735
DAVID G. SYKES AND JAN RYCHTÁŘ	
Domination with decay in triangular matchstick arrangement graphs	749
JILL COCHRAN, TERRY HENDERSON, AARON OSTRANDER AND RON TAYLOR	
On the tree cover number of a graph	767
CHASSIDY BOZEMAN, MINERVA CATRAL, BRENDAN COOK, OSCAR E. GONZÁLEZ AND CAROLYN REINHART	
Matrix completions for linear matrix equations	781
GEOFFREY BUHL, ELIJAH CRONK, ROSA MORENO, KIRSTEN MORRIS, DIANNE PEDROZA AND JACK RYAN	
The Hamiltonian problem and $t$ -path traceable graphs	801
KASHIF BARI AND MICHAEL E. O'SULLIVAN	
Relations between the conditions of admitting cycles in Boolean and ODE network systems	813
YUNJIAO WANG, BAMIDELE OMIDIRAN, FRANKLIN KIGWE AND KIRAN CHILAKAMARRI	
Weak and strong solutions to the inverse-square brachistochrone problem on circular and annular domains	833
CHRISTOPHER GRIMM AND JOHN A. GEMMER	
Numerical existence and stability of steady state solutions to the distributed spruce budworm model	857
HALA AL-KHALIL, CATHERINE BRENNAN, ROBERT DECKER, ASLIHAN DEMIRKAYA AND JAMIE NAGODE	
Integer solutions to $x^2 + y^2 = z^2 - k$ for a fixed integer value $k$	881
WANDA BOYER, GARY MACGILLIVRAY, LAURA MORRISON, C. M. (KIEKA) MYNHARDT AND SHAHLA NASSERASR	
A solution to a problem of Frechette and Locus	893
CHENTHURAN ABEYAKARAN	



1944-4176(2017)10:5;1-8