

# involve

a journal of mathematics

## Editorial Board

Kenneth S. Berenhaut, *Managing Editor*

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



# involve

msp.org/involve

## EDITORS

### MANAGING EDITOR

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

### BOARD OF EDITORS

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

## PRODUCTION


Silvio Levy, Scientific Editor

See inside back cover or msp.org/involve for submission instructions. The subscription price for 2014 is US \$120/year for the electronic version, and \$165/year (+\$35, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues from the last three years and changes of subscribers address should be sent to MSP.

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

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

PUBLISHED BY

 **mathematical sciences publishers**  
nonprofit scientific publishing

<http://msp.org/>

© 2014 Mathematical Sciences Publishers

# An interesting proof of the nonexistence of a continuous bijection between $\mathbb{R}^n$ and $\mathbb{R}^2$ for $n \neq 2$

Hamid Reza Daneshpajouh, Hamed Daneshpajouh and Fereshte Malek

(Communicated by Joel Foisy)

We show that there is no continuous bijection from  $\mathbb{R}^n$  onto  $\mathbb{R}^2$  for  $n \neq 2$  by an elementary method. This proof is based on showing that for any cardinal number  $\beta \leq 2^{\aleph_0}$ , there is a partition of  $\mathbb{R}^n$  ( $n \geq 3$ ) into  $\beta$  arcwise connected dense subsets.

## 1. Introduction

In 1877 Cantor discovered a bijection of  $\mathbb{R}$  onto  $\mathbb{R}^n$  for any  $n \in \mathbb{N}$ . Cantor's map was discontinuous, but the discovery of the Peano curve in 1890 showed that there existed continuous (although not injective) maps of  $\mathbb{R}$  onto  $\mathbb{R}^n$ . Between then and 1910, several mathematicians showed that there does not exist a bicontinuous bijection (homeomorphism) from  $\mathbb{R}^m$  onto  $\mathbb{R}^n$  for the cases  $m = 2$  and  $m = 3$  and  $n > m$ . Finally in 1911, Brouwer showed that there does not exist a homeomorphism between  $\mathbb{R}^m$  and  $\mathbb{R}^n$  for  $n \neq m$  (for a modern treatment, see [Munkres 1984, p. 109]). The present paper proves the nonexistence of a continuous bijection from  $\mathbb{R}^n$  onto  $\mathbb{R}^2$  for  $n \neq 2$  by an elementary method.

Rudin [1963] showed that for any countable cardinal  $\alpha > 2$ , we cannot partition the plane into  $\alpha$  arcwise connected dense subsets. In this paper we show that for any cardinal number  $\beta \leq 2^{\aleph_0}$ , there is a partition of  $\mathbb{R}^n$  ( $n \geq 3$ ) into  $\beta$  arcwise connected dense subsets; then by using this we show that there is no continuous bijection from  $\mathbb{R}^n$  onto  $\mathbb{R}^2$  for  $n \neq 2$ .

**Lemma 1.** *There is a partition of  $\mathbb{R}^+$  into  $2^{\aleph_0}$  dense subsets.*

*Proof.* Consider the additive group  $(\mathbb{R}, +)$ . The quotient group  $\mathbb{R}/\mathbb{Q}$  has  $2^{\aleph_0}$  elements which are dense subsets of  $\mathbb{R}$ . Intersect them with  $\mathbb{R}^+$ .  $\square$

**Theorem 1.** *There is a partition of  $\mathbb{R}^3$  into  $2^{\aleph_0}$  arcwise connected dense subsets.*

*MSC2010:* primary 54-XX; secondary 54CXX.

*Keywords:* arcwise connected, dense subset, homeomorphism.

*Proof.* Let  $\{S_i \mid i \in I\}$  be a partition of  $\mathbb{R}^+$  into  $2^{\aleph_0}$  dense subsets. The set  $I$  is just an index set, so we may suppose that  $I = (0, 1)$ . Define  $L_i = \{(t, it, 0) \mid t > 0\}$  and  $M = \bigcup_{i \in I} L_i$  and let  $A_i$  be the union of all spheres with center at the origin and radius from  $S_i$ , that is,  $A_i = \{x \in \mathbb{R}^3 \mid \|x\| \in S_i\}$ . Let  $B_i = (A_i \setminus M) \cup L_i$ . If  $S$  is a sphere centered at the origin, then  $S \setminus M$  is a sphere with a small arc removed. Therefore  $A_i \setminus M$  is the union of some arcwise connected punctured spheres. Open half-line  $L_i$  pastes these punctured spheres together, so  $B_i$  is arcwise connected. It is obvious that  $\{B_i \mid i \in I\}$  is a partition of  $\mathbb{R}^3$  with size  $2^{\aleph_0}$ . Since  $S_i$  is dense in  $\mathbb{R}^+$ ,  $A_i$  and consequently  $B_i$  are dense in  $\mathbb{R}^3$ .  $\square$

**Corollary 1.** *There is a partition of  $\mathbb{R}^n$  into  $2^{\aleph_0}$  arcwise connected dense subsets for  $n \geq 3$ .*

*Proof.* It is enough to set  $B_i^{(n)} = B_i \times \mathbb{R}^{n-3}$ , in which  $B_i$  is as above. The collection  $\{B_i^{(n)} \mid i \in I\}$  is a partition of  $\mathbb{R}^n$  satisfying the claim.  $\square$

Note that the union of any number of the sets  $B_i^{(n)}$  is an arcwise connected dense subset of  $\mathbb{R}^n$ , hence:

**Corollary 2.** *For any cardinal number  $\beta \leq 2^{\aleph_0}$ , there is a partition of  $\mathbb{R}^n$  ( $n \geq 3$ ) into  $\beta$  arcwise connected dense subsets.*

**Theorem 2.** *For any countable cardinal  $\alpha > 2$ , we cannot partition the plane into  $\alpha$  arcwise connected dense subsets.*

*Proof.* This statement is proved in [Rudin 1963].  $\square$

**Lemma 2.** *Let  $X, Y$  be metric spaces and  $T : X \rightarrow Y$  be a continuous map.*

- (a) *If  $A$  is dense in  $X$  and  $T$  is surjective, then  $T(A)$  is dense in  $Y$ .*
- (b) *If  $B \subset X$  is arcwise connected, then  $T(B)$  is also arcwise connected.*

**Theorem 3.** *There is no continuous bijection from  $\mathbb{R}$  onto  $\mathbb{R}^m$  for  $m \neq 1$ .*

*Proof.* Suppose the contrary: Let  $g : \mathbb{R} \rightarrow \mathbb{R}^m$  be a continuous bijective map. We put  $B_n = [-n, n]$ , and so we have  $\mathbb{R}^m = g(\bigcup_{n=1}^{\infty} B_n) = \bigcup_{n=1}^{\infty} g(B_n)$ . Since  $\mathbb{R}^m$  is not in the first category, at least one of the  $g(B_n)$ , for example  $g(B_k)$ , has nonempty interior in  $\mathbb{R}^m$ . Suppose  $B(x, r) \subset g(B_k)$ . Since  $B_k$  is compact,  $f : B_k \rightarrow g(B_k)$  is a homeomorphism. It follows that  $B(x, r)$  is homeomorphic with an interval in  $\mathbb{R}$ . This is a contradiction, because if we remove 3 points from  $B(x, r)$  it remains connected, but this is not the case for the intervals in  $\mathbb{R}$ .  $\square$

**Theorem 4.** *There is no continuous bijection from  $\mathbb{R}^n$  onto  $\mathbb{R}^2$  for  $n \neq 2$ .*

*Proof.* Suppose the contrary:

- (a) If  $n > 2$ , then according to Corollary 2 and Lemma 2 we can partition  $\mathbb{R}^2$  into 3 arcwise connected dense subsets, and this contradicts Theorem 2.
- (b) If  $n = 1$ , then this contradicts Theorem 3.  $\square$

### Acknowledgments

The authors are grateful to Professor Nicolas Hadjisavvas for his valuable advice and comments. The authors are also grateful to the referee for an extensive critical report including helpful hints and corrections, and extend our special thanks to Johannes Hahn for the useful point leading to the solution of the problem for case  $n = 1$ .

### References

- [Munkres 1984] J. R. Munkres, *Elements of algebraic topology*, Addison-Wesley, Menlo Park, CA, 1984. MR 85m:55001 Zbl 0673.55001
- [Rudin 1963] M. E. Rudin, "Arcwise connected sets in the plane", *Duke Math. J.* **30** (1963), 363–366. MR 27 #1923 Zbl 0131.38004

Received: 2012-06-03    Revised: 2012-11-27    Accepted: 2012-12-01

hr.daneshpajouh@ipm.ir    *School of Mathematics, Institute for Research in Fundamental Sciences, P.O. Box 19395-5746, Tehran, Iran*

hdp@mehr.sharif.ir    *Department of Mathematical Sciences, Sharif University of Technology, P.O. Box 11155-9415 Tehran, Iran*

malek@kntu.ac.ir    *Department of Mathematics, Faculty of Science, K. N. Toosi University of Technology, P.O. Box 16315-1618, Tehran, Iran*



# Analysing territorial models on graphs

Marie Bruni, Mark Broom and Jan Rychtář

(Communicated by Kenneth S. Berenhaut)

Evolutionary graph theory combines evolutionary games with population structure, induced by the graph. The games used are limited to pairwise games occurring on the edges of the graph. Multiplayer games can be important in biological modelling, however, and so recently a new framework for modelling games in structured populations allowing games with arbitrary numbers of players was introduced. In this paper we develop the model to investigate the effect of population structure on the level of aggression, as opposed to a well-mixed population for two specific types of graph, using a multiplayer hawk-dove game. We find that the graph structure can have a significant effect on the level of aggression, and that a key factor is the variability of the group sizes formed to play the games; the more variable the group size, the lower the level of aggression, in general.

## 1. Introduction

Evolutionary graph theory has been developed to more realistically model evolution in populations [Lieberman et al. 2005; Antal and Scheuring 2006; Nowak 2006; Broom and Rychtář 2008]. These models use standard games like the Prisoner's Dilemma and the hawk-dove game, and embed them within a graph structure [Ohtsuki et al. 2006; Santos and Pacheco 2006; Hadjichrysanthou et al. 2011] representing a finite inhomogeneous population, as opposed to traditional evolutionary game theory models which generally consider infinite well-mixed populations. Earlier work also considered similar models which depart from the infinite well-mixed case, in particular [Schaffer 1988] considered a hawk-dove game in a finite population, and [Killingback and Doebeli 1996] considered a hawk-dove game on a lattice. A limitation of the evolutionary graph theory approach is that games can only involve two players, which interact through the graph edges. However, many real animal interactions can involve many players, e.g., in African wild dogs [Ginsberg and Macdonald 1990] or roadrunners [Kelley et al. 2011]. In addition useful theoretical

---

*MSC2010:* primary 91A22; secondary 05C57, 91A43, 92B05.

*Keywords:* structured populations, evolution, game theory, territory.

models which we might want to utilise also describe such multiplayer interactions. In some cases many groups can interact at significant food sources, and often food loss to neighbours can be considerable [Jetz et al. 2004].

In [Broom and Rychtář 2012] we developed a new framework of territorial behaviour modelling how a structured population involved the interaction of different sized groups of individuals at different times. As well as developing the general framework, they also introduced some specific models of interaction. One such was the territorial raider model, where individuals each owned a territory and could either stay in their own territory or move to a neighbouring territory at each time point. Whenever a group of individuals met on a territory, they interacted by means of playing a (potentially) multiplayer game. In the same paper we considered an example of a multiplayer hawk-dove game on a star. One important conclusion was that the level of aggressiveness was less on the star graph than on the equivalent well-mixed graph, i.e., an unstructured population. Thus it is possible that the population structure can have a significant effect on how the population behaves, and it may be that in real structured populations aggression is lower than that predicted by models which do not take the structure into account.

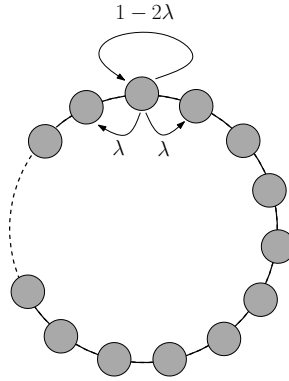
In this paper we follow [Broom and Rychtář 2012] and model the same interaction using different example graphs, again comparing these structured populations with their equivalent well-mixed population model. We show that in different circumstances the level of aggression can be noticeably higher or lower than would be expected in the equivalent well-mixed population, and that even graphs with superficially similar structures can lead to either a significant increase or decrease in the level of aggression. Indeed a particular graph structure can lead to either more or less aggression than the well-mixed population, depending upon other parameter values. Thus to model group interactions properly, it may be important to develop a strong understanding of the nature of interactions and the population structure.

We consider  $N$  individuals  $I_1, \dots, I_N$  living in their own respective territories  $P_1, \dots, P_N$ . The individuals can also move to one of the territories neighbouring theirs. This situation is modelled by a graph  $(V, E)$  where the vertices represent the individuals and the territories that they occupy, and an edge between two vertices means that they are neighbours, and so one individual can raid the territory of the other.

## 2. A territorial raider model on the circle

A circle is a connected graph with every vertex having degree 2. In this model, each individual can go from its territory to one next to its own with a probability  $\lambda$  and stays on its territory with a probability  $1 - 2\lambda$ . The circle model is shown in Figure 1.





**Figure 1.** The circle representation. In this model, each individual starts on one of the vertices which represent their territories. From this vertex they can either “stay at home” with a probability  $1 - 2\lambda$  or explore a neighbouring territory connected to it through an edge with a probability  $\lambda$  to go to each neighbour (every individual has two neighbours in this model).

**2.1. Group sizes.** For any population of size  $N \geq 3$  in the circle model, we have the following probabilities that a given individual is in a group of size  $i$ , denoted by  $P(|G| = i)$ :

$$\begin{aligned} P(|G| = 1) &= (1 - 2\lambda)(1 - \lambda)^2 + 2\lambda(2\lambda)(1 - \lambda) \\ &= 1 - 4\lambda + 9\lambda^2 - 6\lambda^3, \end{aligned}$$

$$\begin{aligned} P(|G| = 2) &= 2(1 - 2\lambda)(1 - \lambda)\lambda + 2\lambda((1 - 2\lambda)(1 - \lambda) + 2\lambda^2) \\ &= 4\lambda(1 - 3\lambda + 3\lambda^2), \end{aligned}$$

$$P(|G| = 3) = 3\lambda^2(1 - 2\lambda),$$

$$P(|G| = k) = 0, \quad k > 3.$$

Note that these probabilities do not depend on  $N$ . From here, we find that the mean group size is

$$E[|G|] = 1 + 4\lambda - 6\lambda^2. \quad (1)$$

**2.2. A multiplayer hawk-dove game.** We suppose that the individuals on the circle structure play a multiplayer hawk-dove game as in [Broom and Rychtář 2012]; i.e., if several individuals are on the same territory then they compete for a reward of value  $V$ . If all individuals are doves, they split the reward equitably and if there are hawks all the doves give up and get nothing, while the hawks fight for the reward so that one hawk receives the reward  $V$  and all the other hawks get a cost  $C$  (see [Broom and Rychtář 2012] for more details on the calculations). If all individuals

play hawk with a probability  $\alpha$ , except our focal individual, we find that the average payoff for a dove player will be

$$E_d(\alpha) = V\{1 - 2\lambda + 4\lambda^2 - 2\lambda^3 + \alpha(4\lambda^2 - 2\lambda - 2\lambda^3) + \alpha^2(\lambda^2 - 2\lambda^3)\}$$

and the average payoff for a hawk player will be

$$E_h(\alpha) = V\{1 + \alpha(3\lambda^2 - 2\lambda) + \alpha^2(\lambda^2 - 2\lambda^3)\} + C\{-2\alpha\lambda + 3\alpha\lambda^2 + \alpha^2\lambda^2 - 2\alpha^2\lambda^3\}.$$

Then the difference of payoff between a hawk player and a dove player, will be given by the incentive function

$$\begin{aligned} h_C(\alpha) &= E_h(\alpha) - E_d(\alpha) \\ &= V\{2\lambda - 4\lambda^2 - \alpha\lambda^2 + 2\lambda^3 + 2\alpha\lambda^3\} - C\{2\alpha\lambda + 3\alpha\lambda^2 + \alpha^2\lambda^2 - 2\alpha^2\lambda^3\}. \end{aligned} \quad (2)$$

In [Broom and Rychtář 2012] we considered examples involving  $V = 1$  and  $C = 2$ , and here we shall also use these values. In this case

$$h_C(\alpha) = 2\lambda - 4\lambda^2 + 2\lambda^3 + \alpha(-4\lambda + 5\lambda^2 + 2\lambda^3) + \alpha^2(2\lambda^2 - 4\lambda^3).$$

To find mixed evolutionarily stable strategies, i.e., with  $0 < \alpha < 1$ , we need to set  $h_C(\alpha) = 0$ . We then have the discriminant for  $\alpha$  given by

$$\Delta = 36\lambda^6 - 60\lambda^5 + 73\lambda^4 - 56\lambda^3 + 16\lambda^2$$

and obtain the roots

$$\alpha_1 = \frac{-5\lambda^2 + 4\lambda - 2\lambda^3 + \sqrt{\Delta}}{2(2\lambda^2 - 4\lambda^3)} \quad \text{and} \quad \alpha_2 = \frac{-5\lambda^2 + 4\lambda - 2\lambda^3 - \sqrt{\Delta}}{2(2\lambda^2 - 4\lambda^3)}.$$

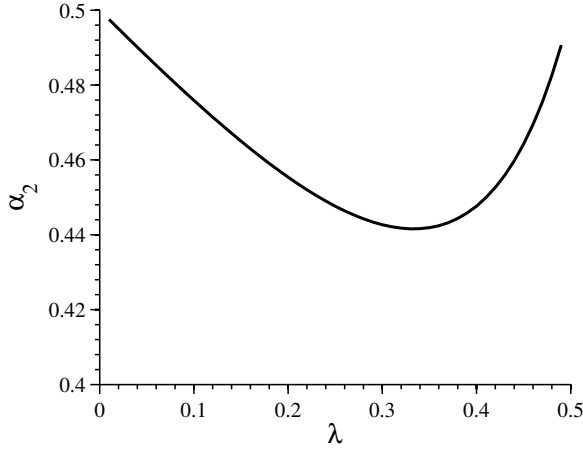
We now have to see if those values are in  $(0, 1)$  or not to find possible ESSs.

**Example 1.** For any population size  $N$ , if we take  $\lambda = \frac{1}{3}$ , we find that the roots of  $h_C(\alpha) = 0$  are  $\alpha_1 = 9.0584$  and  $\alpha_2 = 0.4416$ . The first root is outside of  $[0, 1]$  and it is easy to show that the other is stable, whilst the two pure strategies 0 and 1 are unstable, so that 0.4416 is the unique ESS of this case.

More generally the value of  $\alpha_2$  is shown for the full range of values of  $\lambda$  in Figure 2.  $\alpha_2$  is low for intermediate values of  $\lambda$  when the variability of group size is the largest, and high for the extreme values when group size variability is lower.

**2.3. The equivalent well-mixed population model.** As described in [Broom and Rychtář 2012], to find an equivalent well-mixed population, we want to identify a  $p$  so that we have a Binomial( $N - 1, p$ ) distribution with equal mean group size to that of the circle. Here we have the equation  $E[|G|] = 1 + (N - 1)p$ , which is to say,

$$p = \frac{\lambda(4 - 6\lambda)}{N - 1}. \quad (3)$$



**Figure 2.** The values of the biological meaningful root  $\alpha_2$  for the multiplayer hawk-dove game with  $V = 1$  and  $C = 2$  on the circle, representing the probability of playing hawk in the mixed ESS. The values of  $\alpha_2$  for all allowable  $\lambda$  are shown.

Following [Broom and Rychtář 2012], the same hawk-dove game as played above leads to

$$E_h(\alpha) = V \frac{1 - (1 - p\alpha)^N}{Np\alpha} + C \left( -1 + \frac{1 - (1 - p\alpha)^N}{Np\alpha} \right)$$

and

$$E_d(\alpha) = V \left( \frac{(1 - p\alpha)^N - (1 - p)^N}{Np(1 - \alpha)} \right);$$

i.e., the incentive function is

$$h_W(\alpha)$$

$$= V \frac{1 - (1 - p\alpha)^N}{Np\alpha} + C \left( -1 + \frac{1 - (1 - p\alpha)^N}{Np\alpha} \right) - V \left( \frac{(1 - p\alpha)^N - (1 - p)^N}{Np(1 - \alpha)} \right),$$

or again

$$h_W(\alpha) = \frac{1}{Np\alpha(1 - \alpha)} \left\{ (1 - \alpha)(V + C) - (V + C)(1 - p\alpha)^N + C\alpha(1 - p\alpha)^N - CNp\alpha(1 - \alpha) + \alpha V(1 - p)^N \right\}. \quad (4)$$

In [Broom and Rychtář 2012] it was stated that there is at most one root of (4) in the interval  $[0, 1]$ . A proof of this statement is given in Appendix A, where it is shown that there is a root between 0 and 1 for  $p \neq 0$  and  $C > 0$  if and only if

$$\frac{V}{C} < \frac{Np + (1 - p)^N - 1}{1 - (1 - p)^N - Np(1 - p)^{N-1}}. \quad (5)$$

**Example 2.** We again take  $\lambda = \frac{1}{3}$ , and find the ESS value  $\alpha_N$  for the well-mixed population of size  $N$ , for various  $N$ .

- a) We consider  $N = 3$ , corresponding to the smallest possible circle. We find the same result as in the circle case:  $\alpha_{\text{circle}} = \alpha_3 = 0.4416$ . This is as we would expect as for  $N = 3$  the circle and the well-mixed population are identical for  $\lambda = \frac{1}{3}$ .
- b) For  $N = 5$ , we find  $\alpha_5 = 0.4208 < \alpha_{\text{circle}}$ .
- c) For  $N = 50$  we find  $\alpha_{50} = 0.4046 < \alpha_5 < \alpha_{\text{circle}}$ .

Thus for the well-mixed population the ESS hawk probability declines with the population size. In particular, except for  $N = 3$ , the ESS hawk proportion is higher for the circle than for the well-mixed population. This is in contrast to the star form from [Broom and Rychtář 2012]. These results are consistent because in the circle case, the hawk cannot be in a territory with more than two other hawks whereas the equivalent well-mixed population allows bigger groups which disfavour hawk players. The star in turn allowed such bigger groups to form with greater probability.

### 3. A territorial raider model on a complete bipartite graph

A bipartite graph is a graph whose vertices can be divided into two disjoint sets  $U$  and  $V$ , with  $n$  and  $m$  elements respectively, such that every edge connects a vertex in  $U$  to one in  $V$ . A complete bipartite graph is a special kind of bipartite graph where every vertex of the first set is connected to every vertex of the second set. We shall assume that each individual in  $U$  has a probability  $\lambda$  of going to each territory in  $V$  and a probability  $1 - m\lambda$  of staying in its own territory, and similarly each individual in  $V$  has probability  $\mu$  of going to each territory in  $U$  and a probability  $1 - n\mu$  of staying in its own territory. The general bipartite model is illustrated in Figure 3.

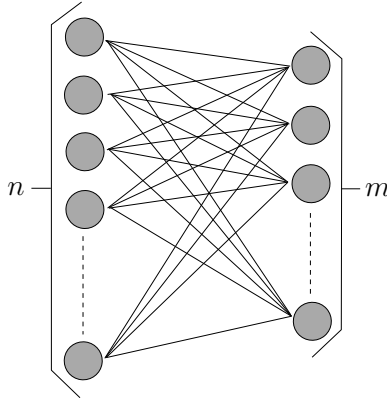
We shall again find the distribution of group sizes, and compare these to the equivalent well-mixed population.

**3.1. Group size.** Without loss of generality we assume that  $m \leq n$ . For an individual in a territory on the right side (in the smaller side, with  $m$  individuals), we find:

For any  $1 \leq k \leq \min(n + 1, m + 1) = m + 1$ :

$$P(|G| = k) =$$

$$\binom{n}{k-1} (1 - n\mu) \lambda^{k-1} (1 - \lambda)^{n-k+1} + n\mu (1 - m\lambda) \binom{m-1}{k-2} \mu^{k-2} (1 - \mu)^{m-k+1} \\ + nm\lambda\mu \binom{m-1}{k-1} \mu^{k-1} (1 - \mu)^{m-k}.$$



**Figure 3.** The  $n$ - $m$  complete bipartite graph representation. The vertices represent territories. The edges represent the possible moves from one territory to another. Here we assume that each individual on the left has a probability  $\lambda$  to go to each of the right territories and a probability  $1 - m\lambda$  to stay in its own territory, and similarly each individual on the right has a probability  $\mu$  to go to each territory on the left and a probability  $1 - n\mu$  to stay in its own territory.

For any  $m + 2 \leq k \leq \max(n + 1, m + 1) = n + 1$  we have

$$P(|G| = k) = \binom{n}{k-1} (1 - n\mu) \lambda^{k-1} (1 - \lambda)^{n-k+1}.$$

Similarly for an individual in a territory on the left, we find:

For any  $1 \leq k \leq m + 1$ :

$$\begin{aligned} P(|G| = k) = & \binom{m}{k-1} (1 - m\lambda) \mu^{k-1} (1 - \mu)^{m-k+1} \\ & + m\lambda(1 - n\mu) \binom{n-1}{k-2} \lambda^{k-2} (1 - \lambda)^{n-k+1} + nm\lambda\mu \binom{n-1}{k-1} \lambda^{k-1} (1 - \lambda)^{n-k}. \end{aligned}$$

For any  $m + 2 \leq k \leq n + 1$ , we find:

$$P(|G| = k) = m\lambda(1 - n\mu) \binom{n-1}{k-2} \lambda^{k-2} (1 - \lambda)^{n-k+1} + nm\lambda\mu \binom{n-1}{k-1} \lambda^{k-1} (1 - \lambda)^{n-k}.$$

Finally we find the average probability for an individual on this structure:

For any  $1 \leq k \leq m + 1$ :

$$\begin{aligned}
 P(|G| = k) &= \frac{n}{n+m} \binom{m}{k-1} (1-m\lambda)\mu^{k-1}(1-\mu)^{m-k+1} \\
 &\quad + \frac{n}{n+m} \binom{n-1}{k-2} m\lambda(1-n\mu)\lambda^{k-2}(1-\lambda)^{n-k+1} \\
 &\quad + \frac{n}{n+m} \binom{n-1}{k-1} nm\lambda\mu\lambda^{k-1}(1-\lambda)^{n-k} \\
 &\quad + \frac{m}{n+m} \binom{n}{k-1} (1-n\mu)\lambda^{k-1}(1-\lambda)^{n-k+1} \\
 &\quad + \frac{m}{n+m} n\mu(1-m\lambda) \binom{m-1}{k-2} \mu^{k-2}(1-\mu)^{m-k+1} \\
 &\quad + \frac{m}{n+m} nm\lambda\mu \binom{m-1}{k-1} \mu^{k-1}(1-\mu)^{m-k}.
 \end{aligned}$$

For any  $m + 2 \leq k \leq n + 1$ :

$$\begin{aligned}
 P(|G| = k) &= \\
 &\frac{n}{n+m} \left( m\lambda(1-n\mu) \binom{n-1}{k-2} \lambda^{k-2}(1-\lambda)^{n-k+1} + nm\lambda\mu \binom{n-1}{k-1} \lambda^{k-1}(1-\lambda)^{n-k} \right) \\
 &\quad + \frac{m}{n+m} \left( (1-n\mu) \binom{n}{k-1} \lambda^{k-1}(1-\lambda)^{n-k+1} \right).
 \end{aligned}$$

Now we can use these results (see Appendix B) to show that the mean group size is given by

$$\begin{aligned}
 \mathbf{E}(|G|) &= \\
 &1 + \frac{2nm\mu - 2nm^2\lambda\mu + 2nm\lambda - 2n^2m\lambda\mu + n^2m\lambda^2 - nm\lambda^2 + n\mu^2m^2 - nm\mu^2}{n+m}. \quad (6)
 \end{aligned}$$

**3.2. The equivalent well-mixed population.** In the equivalent well-mixed population with  $N = n + m$  individuals, with the number of individuals in the same patch as a focal individual following a Binomial( $N - 1, p$ ) distribution, we want the same mean group size as before. For a well-mixed population equivalent to the  $n$ - $m$  structure, we will have

$$1 + p(n + m - 1) = E(|G|).$$

This leads directly from the previous calculation to

$$p = \frac{2nm\mu - 2nm^2\lambda\mu + 2nm\lambda - 2n^2m\lambda\mu + n^2m\lambda^2 - nm\lambda^2 + n\mu^2m^2 - nm\mu^2}{(n+m)(n+m-1)}. \quad (7)$$

**Example 3.** If we take  $n = m$  and  $\lambda = \mu = 1/(n + 1)$ , clearly the probability distribution of the group size of an individual from the left is identical to that of an individual from the right. For any  $1 \leq k \leq n + 1$ ,

$$\begin{aligned} P(|G| = k) &= \binom{n}{k-1} \left(1 - \frac{n}{n+1}\right) \left(\frac{1}{n+1}\right)^{k-1} \left(1 - \frac{1}{n+1}\right)^{n-k+1} \\ &\quad + \frac{n}{(n+1)^2} \binom{n-1}{k-2} \left(\frac{1}{n+1}\right)^{k-2} \left(1 - \frac{1}{n+1}\right)^{n-k+1} \\ &\quad + \frac{n^2}{(n+1)^2} \binom{n-1}{k-1} \left(\frac{1}{n+1}\right)^{k-1} \left(1 - \frac{1}{n+1}\right)^{n-k} \\ &= \frac{n^{n-k+1}}{(n+1)^{n+1}} \left( \frac{n!}{(k-1)!(n-k+1)!} + \frac{n!}{(k-2)!(n-k+1)!} + \frac{n!}{(k-1)!(n-k)!} \right) \\ &= \left(\frac{1}{n+1}\right)^{k-1} \left(1 - \frac{1}{n+1}\right)^{n-(k-1)} \binom{n}{k-1}. \end{aligned}$$

Thus this bipartite graph with equally sized parts has a binomially distributed group size, and this is equivalent to a well-mixed population with  $n + 1$  individuals and mean group size  $(2n + 1)/(n + 1)$ . For large  $n$  this is approximately a Poisson distribution which is also a good approximation for the well-mixed population with  $2n$  individuals and mean group size  $(2n + 1)/(n + 1)$ . Thus for large  $n$  this graph is approximately the same as its equivalent well-mixed population.

**3.3. A complete bipartite graph with  $n = 3$  and  $m = 2$ .** We now consider a complete bipartite graph with  $n = 3$  and  $m = 2$ . There is a representation of this model in Figure 4.

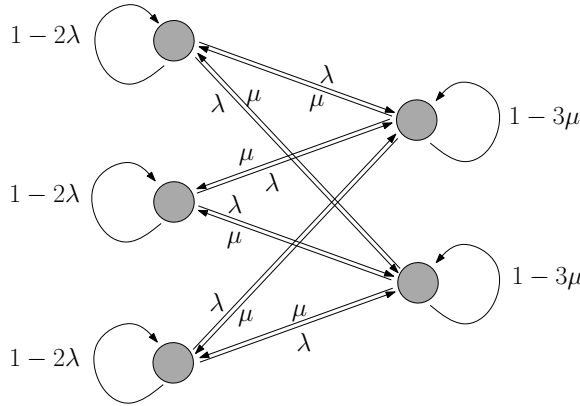
The group size probabilities are as follows. For an individual on the right,

$$\begin{aligned} P(|G| = 1) &= (1 - 3\mu)(1 - \lambda)^3 + 3\mu(2\lambda)(1 - \mu), \\ P(|G| = 2) &= 3(1 - 3\mu)\lambda(1 - \lambda)^2 + 3\mu(1 - 2\lambda)(1 - \mu) + 3\mu^2(2\lambda), \\ P(|G| = 3) &= 3(1 - 3\mu)\lambda^2(1 - \lambda) + 3\mu^2(1 - 2\lambda), \\ P(|G| = 4) &= (1 - 3\mu)\lambda^3, \end{aligned}$$

and for an individual on the left,

$$\begin{aligned} P(|G| = 1) &= (1 - 2\lambda)(1 - \mu)^2 + 2\lambda(3\mu)(1 - \lambda)^2, \\ P(|G| = 2) &= 2(1 - 2\lambda)\mu(1 - \mu) + 2\lambda((1 - 3\mu)(1 - \lambda)^2 + 6\lambda\mu(1 - \lambda)), \\ P(|G| = 3) &= (1 - 2\lambda)\mu^2 + 2\lambda(2(1 - 3\mu)\lambda(1 - \lambda) + 3\lambda^2\mu), \\ P(|G| = 4) &= 2(1 - 3\mu)\lambda^3. \end{aligned}$$

Thus we find the mean group size as



**Figure 4.** The complete bipartite graph with  $n = 3$  and  $m = 2$  representation. The vertices represent territories and the edges represent the possible moves from one territory to another. Here  $\lambda$  is the probability for an individual on the left to move to each of its neighbours on the right; it stays in its own territory with probability  $1 - 2\lambda$ .  $\mu$  is the equivalent probability for an individual on the right.

$$E[|G|] = 1 + \frac{12}{5}\lambda + \frac{12}{5}\mu - 12\mu\lambda + \frac{6}{5}\mu^2 + \frac{12}{5}\lambda^2. \tag{8}$$

As for the circle we consider the multiplayer hawk-dove game. We find then for a hawk-dove game, with probability  $\alpha$  of playing hawk, the following payoffs: For a hawk player the payoff is

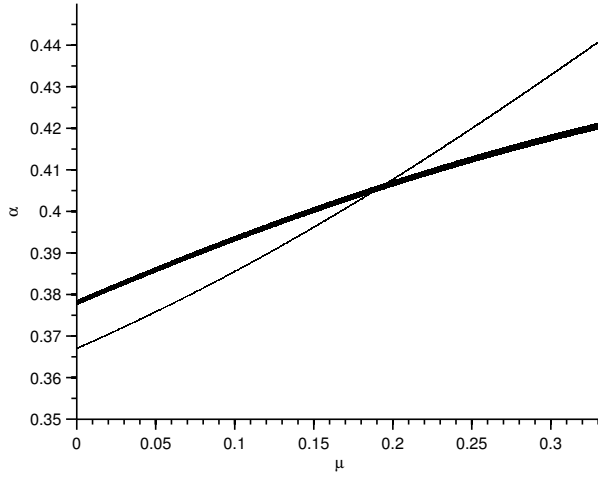
$$\begin{aligned} E_h(\alpha) = & C \left\{ \left( -\frac{6}{5}\mu - \frac{3}{5}\mu^2 + 6\mu\lambda - \frac{6}{5}\lambda - \frac{6}{5}\lambda^2 \right) \right. \\ & + \alpha^2 \left( \frac{3}{5}\mu^2 - \frac{6}{5}\mu^2\lambda - \frac{18}{5}\mu\lambda^2 + \frac{6}{5}\lambda^2 + \frac{2}{5}\lambda^3 \right) + \alpha^3 \left( -\frac{2}{5}\lambda^3 + \frac{6}{5}\mu\lambda^3 \right) \left. \right\} \\ & + V \left\{ 1 + \alpha \left( -\frac{6}{5}\mu - \frac{3}{5}\mu^2 - \frac{6}{5}\lambda + 6\lambda\mu - \frac{6}{5}\lambda^2 \right) \right. \\ & \left. + \alpha^2 \left( \frac{3}{5}\mu^2 - \frac{6}{5}\mu^2\lambda + \frac{6}{5}\lambda^2 - \frac{18}{5}\mu\lambda^2 + \frac{2}{5}\lambda^3 \right) + \alpha^3 \left( -\frac{2}{5}\lambda^3 + \frac{6}{5}\mu\lambda^3 \right) \right\}. \end{aligned}$$

For a dove player the payoff is

$$\begin{aligned} E_d(\alpha) = & V \left\{ 1 - \frac{6}{5}\mu - \frac{6}{5}\lambda + 6\lambda\mu - \frac{6}{5}\mu^2\lambda - \frac{18}{5}\mu\lambda^2 + \frac{6}{5}\mu\lambda^3 \right. \\ & + \alpha \left( -\frac{6}{5}\mu - \frac{6}{5}\lambda + 6\lambda\mu - \frac{6}{5}\mu^2\lambda - \frac{18}{5}\mu\lambda^2 + \frac{6}{5}\mu\lambda^3 \right) \\ & \left. + \alpha^2 \left( \frac{3}{5}\mu^2 - \frac{6}{5}\mu^2\lambda + \frac{6}{5}\lambda^2 - \frac{18}{5}\mu\lambda^2 + \frac{6}{5}\mu\lambda^3 \right) + \alpha^3 \left( -\frac{2}{5}\lambda^3 + \frac{6}{5}\mu\lambda^3 \right) \right\}. \end{aligned}$$

**Example 4.** We consider the case when  $C = 2$ ,  $V = 1$ ,  $\mu = 0.2$  and  $\lambda = \frac{1}{3}$ . Here there is an ESS on the graph with hawk probability  $\alpha = 0.4086$ .





**Figure 5.** The evolutionarily stable proportion of hawks on a complete bipartite graph with  $n = 3$  and  $m = 2$  (thin line) and a well-mixed population (thick line), where  $\lambda = \frac{1}{3}$  and  $\mu$  varies from 0 to  $\frac{1}{3}$ .

For the well-mixed population we obtain

$$p = \frac{3}{5}\lambda + \frac{3}{5}\mu - \frac{15}{5}\mu\lambda + \frac{3}{10}\mu^2 + \frac{3}{5}\lambda^2 = \frac{79}{375}.$$

We find for  $\mu = 0.2$ , that in the well-mixed population there is an ESS with hawk probability  $\alpha = 0.4066 < 0.4077$ . Thus it appears that the hawk probability is somewhat bigger in this 3-2 model than the corresponding well-mixed population.

However, if we vary the parameter  $\mu$  as in Figure 5, we see that for small  $\mu$  the level of aggression is higher in the well-mixed population, and for large  $\mu$  it is bigger on the graph.

**3.4. A complete bipartite graph with  $n = 5$  and  $m = 2$ .** Let us now study another concrete example of this  $n$ - $m$  bipartite graph model. Taking  $n = 5$  and  $m = 2$ , we find that the corresponding probabilities are as follows, where we denote  $P(|G| = k)$  by  $P_k$ :

$$P_1 = \frac{1}{7}(7 - 20\mu + 5\mu^2 - 20\lambda + 20\lambda^2 - 20\lambda^3 + 10\lambda^4 - 2\lambda^5 + 140\lambda\mu - 30\mu^2\lambda - 300\mu\lambda^2 + 400\mu\lambda^3 - 250\mu\lambda^4 + 60\mu\lambda^5),$$

$$P_2 = \frac{1}{7}(20\mu - 20\mu^2 - 140\mu\lambda + 20\lambda - 80\lambda^2 + 120\lambda^3 - 80\lambda^4 + 20\lambda^5 + 60\mu^2\lambda + 600\mu\lambda^2 - 1200\mu\lambda^3 + 1000\mu\lambda^4 - 300\mu\lambda^5),$$

$$P_3 = \frac{1}{7}(15\mu^2 + 60\lambda^2 - 180\lambda^3 + 180\lambda^4 - 60\lambda^5 - 30\mu^2\lambda - 300\mu\lambda^2 + 1200\mu\lambda^3 - 1500\mu\lambda^4 + 600\mu\lambda^5),$$

$$\begin{aligned}
P_4 &= \frac{1}{7}(80\lambda^3 - 160\lambda^4 + 80\lambda^5 - 400\mu\lambda^3 + 1000\mu\lambda^4 - 600\mu\lambda^5), \\
P_5 &= \frac{1}{7}(50\lambda^4 - 50\lambda^5 - 250\mu\lambda^4 + 300\mu\lambda^5), \\
P_6 &= \frac{1}{7}(12\lambda^5 - 60\mu\lambda^5).
\end{aligned}$$

We can calculate the payoff for a dove player as

$$\begin{aligned}
E_d(\alpha) &= P_1 + \frac{P_2}{2} + \frac{P_3}{3} + \frac{P_4}{4} + \frac{P_5}{5} + \frac{P_6}{6} - \alpha\left(\frac{P_2}{2} + 2\frac{P_3}{3} + 3\frac{P_4}{4} + 4\frac{P_5}{5} + \frac{P_6}{6}\right) \\
&\quad + \alpha^2\left(\frac{P_3}{3} + 3\frac{P_4}{4} + 6\frac{P_5}{5} + 10\frac{P_6}{6}\right) - \alpha^3\left(\frac{P_4}{4} + 4\frac{P_5}{5} + 10\frac{P_6}{6}\right) \\
&\quad + \alpha^4\left(\frac{P_5}{5} + 5\frac{P_6}{6}\right) - \frac{P_6}{6}\alpha^5.
\end{aligned}$$

The payoff for a hawk player is similarly

$$\begin{aligned}
E_h(\alpha) &= V - \alpha(V + C)\left(\frac{P_2}{2} + P_3 + \frac{3P_4}{2} + 2P_5 + \frac{5P_6}{2}\right) \\
&\quad + \alpha^2(V + C)\left(\frac{P_3}{3} + P_4 + 2P_5 + \frac{10P_6}{3}\right) - \alpha^3(V + C)\left(\frac{P_4}{4} + P_5 + \frac{5P_6}{2}\right) \\
&\quad + \alpha^4(V + C)\left(\frac{P_5}{5} + P_6\right) + \alpha^5\left(\frac{41VP_6}{6} - \frac{CP_6}{6}\right).
\end{aligned}$$

**Example 5.** For  $\lambda = \frac{1}{3}$  and  $\mu = \frac{1}{6}$  we obtain

$$P_1 = \frac{289}{756}, \quad P_2 = \frac{590}{1701}, \quad P_3 = \frac{1255}{6804}, \quad P_4 = \frac{40}{567}, \quad P_5 = \frac{25}{1701} \quad \text{and} \quad P_6 = \frac{2}{1701}.$$

Using the payoffs  $V = 1$  and  $C = 2$  we find that the ESS occurs when  $\alpha = 0.3603$ .

For the equivalent well-mixed population, we find that for the same values of  $\lambda$  and  $\mu$  as above, we have

$$E[|G|] = P_1 + 2P_2 + 3P_3 + 4P_4 + 5P_5 + 6P_6 = 1.9921.$$

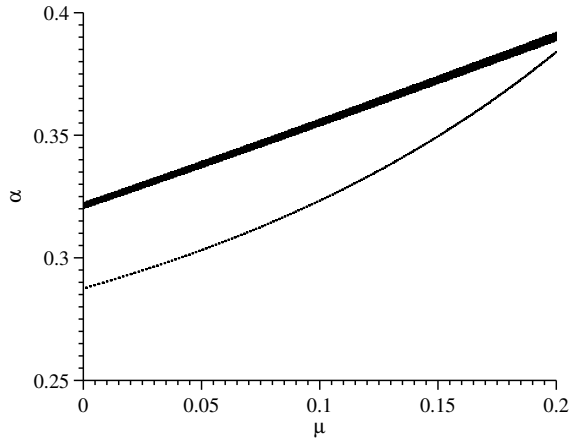
So according to [Broom and Rychtář 2012] we have  $1 + 6p = 1.9921$ , or

$$p = 0.1653.$$

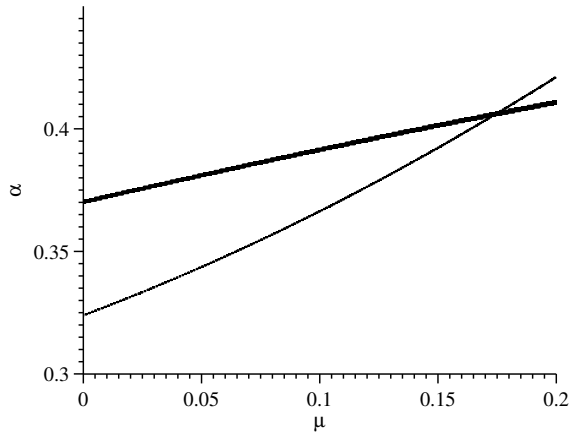
We then find  $\alpha = 0.3785$  as the unique ESS in this equivalent well-mixed population. We notice that  $\alpha_{5-2} = 0.3603 < 0.3785 = \alpha_7$ . Thus in this case hawks prefer the equivalent well-mixed population to the corresponding 5-2 model.

However, again, if we vary the parameter  $\mu$  as we did in Figure 5, we see that for small  $\mu$  the level of aggression is much higher in the well-mixed population, and for large  $\mu$  this advantage is reduced (see Figure 6). There is nevertheless a higher level of aggression for the well-mixed population in all cases. As we see in Figure 7, this is not the case for the alternative value of  $\lambda = \frac{1}{4}$ .

Thus different bipartite graphs can inhibit or encourage aggression, in comparison to the baseline well-mixed populations. In the multiplayer hawk-dove game,



**Figure 6.** The evolutionarily stable proportion of hawks on a complete bipartite graph with  $n = 5$  and  $m = 2$  (thin line) and a well-mixed population (thick line), where  $\lambda = \frac{1}{3}$  and  $\mu$  varies from 0 to 0.2.



**Figure 7.** The evolutionarily stable proportion of hawks on a complete bipartite graph with  $n = 5$  and  $m = 2$  (thin line) and a well-mixed population (thick line), where  $\lambda = \frac{1}{4}$  and  $\mu$  varies from 0 to 0.2.

hawks do particularly badly in large groups. Thus when there is a significant risk of a large group forming, selection will favour lower aggression. This is the case in more asymmetric bipartite graphs like the 5-2 model when the parameter  $\lambda$  is large (and the star which is an  $(N-1)$ -1 bipartite graph), where vertices on the smaller side can play host to such large groups.

#### 4. Discussion

Evolutionary graph theory has made a valuable contribution to the understanding of evolution in structured populations [Lieberman et al. 2005; Nowak 2006; Broom and Rychtář 2008]. However it has certain limitations; in particular the interactions between individuals, usually modelled by evolutionary games, are limited to pairwise ones. Hence a new framework was introduced in [Broom and Rychtář 2012] for modelling structured populations which allows interactions between an arbitrarily large number of individuals. The main purpose of the paper was to introduce the framework, and a secondary purpose was to give examples of different models of interaction, one of which was the territorial raider model. However, no single model was considered in any great detail. In this paper we applied results from [Broom and Rychtář 2012] to several different examples of graphs for the territorial raider model and compared the multiplayer hawk-dove game played on these graphs to equivalent well-mixed populations.

We studied two main graphs: the circle and the  $n$ - $m$  complete bipartite graph. The observation of the different cases leads to interesting results. First we notice that for the same mean group size, hawks favour the model in which it is less likely to meet many other individuals, i.e., be a member of a large group; comparing different populations with identical means, it seems that small variance is preferred by hawks. In the circle case, since the maximal number of individuals in one territory is three no matter the number of individual considered, and since the equivalent well-mixed population will allow  $N$  individuals in one territory, hawks prefer the circle model for any  $N$  larger than three. In the  $n$ - $m$  bipartite model the results observed are different. For the 3-2 bipartite graph hawks prefer the 3-2 graph to the equivalent well-mixed population except for small values of  $\mu$  but for the 5-2 graph, hawks generally prefer the equivalent well-mixed population. In [Broom and Rychtář 2012] we considered the star, the  $n$ -1 bipartite graph, and in particular the 4-1 model, where hawks also prefer the equivalent well-mixed population. Here for large numbers of individuals, hawks favour the well-mixed population. From these observations, we understand that the structure of a model has a major influence on the strategy of the individuals.

One of the key components of our population model is the evolutionary game used. We considered a multiplayer hawk-dove game, but there are a number of alternatives that could have been applied. Multiplayer matrix games [Broom et al. 1997] provide a more general class of games, and it is possible to have games which involve coalitionary behaviour, so that perhaps forming large groups can be beneficial, in contrast to the hawk-dove game example. The results will be game-specific, in general. For instance we demonstrated the fact that there is at most one mixed ESS in the well-mixed population model, however for arbitrary multiplayer games there can be many ESSs.

Potential future work will thus include investigating different example games and structures as mentioned above. An important future direction of this research is the incorporation of evolutionary dynamics in the new structure, as at present the theory has concentrated on the development of the framework and key static properties. The type of dynamics used in evolutionary graph theory, such as the invasion process where a random individual gives birth with probability proportional to its fitness and then replaces one of its neighbours at random, will be applicable to our system with suitable redefinition of the term neighbour to more properly interpret the interactions between individuals, though there may be other potential dynamics as well. The combination of dynamics, game and structure will provide a flexible framework for analysing population interactions.

**Appendix A: A proof that Equation (4) has at most one root, and conditions for such a root to occur**

Set  $v = V/C$ . Then

$$h_W(\alpha) = \frac{C}{Np\alpha(1-\alpha)} \left( (1-\alpha)(v+1) - (v+1)(1-p\alpha)^N + \alpha(1-p\alpha)^N - Np\alpha(1-\alpha) + \alpha v(1-p)^N \right).$$

Denoting  $1-\alpha$  by  $\beta$ , we then have

$$h_W(\alpha) = \frac{C}{Np\alpha(1-\alpha)} \left( \beta(v+1) - (v+\beta)(1-p+p\beta)^N - Np\beta(1-\beta) + (1-\beta)v(1-p)^N \right).$$

We now define

$$f(\beta) = \beta(v+1) - (v+\beta)(1-p+p\beta)^N - Np\beta(1-\beta) + (1-\beta)v(1-p)^N.$$

This function is differentiable as many times as we want and its third derivative is

$$f'''(\beta) = -3p^2N(N-1)(1-p+p\beta)^{N-2} - (v+\beta)N(N-1)(N-2)p^3(1-p+p\beta)^{N-3}.$$

It is clear that  $f'''(\beta) < 0$ . Moreover, we have  $f(0) = -v(1-p)^N + v(1-p)^N = 0$ , and  $f(1) = (v+1) - (v+1) = 0$ .

Thus  $f'$  is concave, increasing and then decreasing, and therefore  $f'$  can't have more than two roots. From there we can say that  $f$  has at most three roots. Since we know that  $f(0) = f(1) = 0$ , there is at most one other root in  $\mathbb{R}$  so at most one root in  $(0, 1)$ .

Since we have

$$\frac{C}{Np\beta(1-\beta)} > 0 \quad \text{for all } \beta \in (0, 1),$$

(i.e., there is no root in  $(0, 1)$ ), we can say that  $f(\beta)$  has at most one root in this interval. From here, we can also conclude that  $h_W(\alpha)$  has at most one root in this interval. That concludes the proof.

We now investigate what are the conditions on  $V$  and  $C$  to give a root in  $(0, 1)$ . First, let us calculate  $h_W(\alpha)$  when  $\alpha = 0$  and  $\alpha = 1$ . We have, for  $p \neq 0$

$$h_W(0) = V \left( 1 - \frac{1}{Np} + \frac{(1-p)^N}{Np} \right) > 0$$

and

$$h_W(1) = (V + C) \frac{1 - (1-p)^N}{Np} - C - V(1-p)^{N-1}.$$

So  $h_W$  is positive if  $p \neq 0$  when  $\alpha = 0$  and  $h_W$  has at most one root in  $(0, 1)$ . Thus we can say that either  $h_W$  is nonnegative for any  $\alpha$  if  $h_W(1) \geq 0$  or there is one  $\alpha$  in  $(0, 1)$  such as  $h_W(\alpha) = 0$  (and then we have  $h_W(1) \leq 0$ ).

Let us now study the sign of  $h_W(1)$  for  $p \neq 0$ . We have

$$\begin{aligned} (V + C) \frac{1 - (1-p)^N}{Np} - C - V(1-p)^{N-1} &\geq 0 \\ \iff (1 - (1-p)^N - Np(1-p)^{N-1})V &\geq (Np + (1-p)^N - 1)C \\ \iff \frac{V}{C} &\geq \frac{Np + (1-p)^N - 1}{1 - (1-p)^N - Np(1-p)^{N-1}}. \end{aligned}$$

So there is a root between 0 and 1 for  $p \neq 0$  and  $C > 0$  if and only if

$$\frac{V}{C} < \frac{Np + (1-p)^N - 1}{1 - (1-p)^N - Np(1-p)^{N-1}}. \quad (9)$$

### Appendix B: Mean group size for the complete bipartite graph

The mean group size can be expressed as the sum

$$\mathbf{E}(|G|) = \sum_{k=1}^{n+1} P(|G| = k)k.$$

This is divided into nine distinct terms from the calculations from Section 3.1, six for group sizes less than or equal to  $m + 1$  and three for larger groups. These nine terms are simplified below, and the final expression for the mean group size from Equation (6) is found by summing them.

$$\sum_{k=1}^{m+1} \frac{nk}{n+m} \binom{m}{k-1} (1-m\lambda)\mu^{k-1}(1-\mu)^{m-k+1}$$

$$\begin{aligned}
&= \sum_{k=0}^m \frac{n(k+1)}{n+m} \binom{m}{k} (1-m\lambda)\mu^k (1-\mu)^{m-k} \\
&= \frac{n(1-m\lambda)}{n+m} \left( 1 + \sum_{k=0}^m k \binom{m}{k} \mu^k (1-\mu)^{m-k} \right) \\
&= \frac{n(1-m\lambda)}{n+m} \left( 1 + \sum_{k=1}^m \frac{m!}{(k-1)!(m-k)!} \mu^k (1-\mu)^{m-k} \right) \\
&= \frac{n(1-m\lambda)}{n+m} \left( 1 + m\mu \sum_{k=0}^{m-1} \frac{(m-1)!}{(k)!(m-1-k)!} \mu^k (1-\mu)^{m-k-1} \right) \\
&= \frac{n(1-m\lambda)}{n+m} + \frac{nm\mu(1-m\lambda)}{n+m}.
\end{aligned}$$

$$\begin{aligned}
&\sum_{k=1}^{m+1} \frac{nk}{n+m} \binom{n-1}{k-2} m\lambda(1-n\mu)\lambda^{k-2}(1-\lambda)^{n-k+1} \\
&\quad + \sum_{k=m+2}^{n+1} \frac{nmk}{n+m} (1-n\mu) \binom{n-1}{k-2} \lambda^{k-1}(1-\lambda)^{n-k+1} \\
&= \sum_{k=0}^m \frac{n(k+1)}{n+m} \binom{n-1}{k-1} m\lambda(1-n\mu)\lambda^{k-1}(1-\lambda)^{n-k} \\
&\quad + \sum_{k=m+1}^n \frac{nm(k+1)}{n+m} (1-n\mu) \binom{n-1}{k-1} \lambda^k (1-\lambda)^{n-k} \\
&= \frac{nm\lambda(1-n\mu)}{n+m} \sum_{k=0}^n \binom{n-1}{k-1} (1+k)\lambda^{k-1}(1-\lambda)^{n-k} \\
&= \frac{nm\lambda(1-n\mu)}{n+m} \sum_{k=1}^n \binom{n-1}{k-1} (1+k)\lambda^{k-1}(1-\lambda)^{n-k} \\
&= \frac{nm\lambda(1-n\mu)}{n+m} \sum_{k=0}^{n-1} \binom{n-1}{k} (2+k)\lambda^k (1-\lambda)^{n-k-1} \\
&= 2\frac{nm\lambda(1-n\mu)}{n+m} + \frac{nm\lambda(1-n\mu)}{n+m} \sum_{k=0}^{n-1} \binom{n-1}{k} k\lambda^k (1-\lambda)^{n-k-1} \\
&= 2\frac{nm\lambda(1-n\mu)}{n+m} + \frac{nm\lambda(1-n\mu)}{n+m} \sum_{k=1}^{n-1} \frac{(n-1)!}{(k-1)!(n-k-1)!} \lambda^k (1-\lambda)^{n-k-1}
\end{aligned}$$

$$\begin{aligned}
&= 2 \frac{nm\lambda(1-n\mu)}{n+m} + \frac{nm\lambda^2(n-1)(1-n\mu)}{n+m} \sum_{k=0}^{n-2} \binom{n-2}{k} \lambda^k (1-\lambda)^{n-k-2} \\
&= 2 \frac{nm\lambda(1-n\mu)}{n+m} + \frac{nm\lambda^2(n-1)(1-n\mu)}{n+m}.
\end{aligned}$$

$$\begin{aligned}
&\sum_{k=1}^{m+1} \frac{nk}{n+m} \binom{n-1}{k-1} nm\mu \lambda^k (1-\lambda)^{n-k} + \sum_{k=m+2}^{n+1} \frac{nmk\lambda\mu}{n+m} \binom{n-1}{k-1} n\lambda^{k-1} (1-\lambda)^{n-k} \\
&= \sum_{k=0}^m \frac{n(k+1)}{n+m} \binom{n-1}{k} nm\mu \lambda^{k+1} (1-\lambda)^{n-k-1} \\
&\quad + \sum_{k=m+1}^n \frac{nm(k+1)\lambda\mu}{n+m} \binom{n-1}{k} n\lambda^k (1-\lambda)^{n-k-1} \\
&= \sum_{k=0}^n \frac{n(k+1)}{n+m} \binom{n-1}{k} nm\mu \lambda^{k+1} (1-\lambda)^{n-k-1} \\
&= \frac{n^2 m \lambda \mu}{n+m} \sum_{k=0}^n (k+1) \binom{n-1}{k} \lambda^k (1-\lambda)^{n-k-1} \\
&= \frac{n^2 m \lambda \mu}{n+m} \sum_{k=0}^{n-1} (k+1) \binom{n-1}{k} \lambda^k (1-\lambda)^{n-k-1} \\
&= \frac{n^2 m \lambda \mu}{n+m} \left( 1 + \sum_{k=1}^{n-1} k \binom{n-1}{k} \lambda^k (1-\lambda)^{n-k-1} \right) \\
&= \frac{n^2 m \lambda \mu}{n+m} \left( 1 + \sum_{k=0}^{n-2} \frac{(n-1)!}{(k)! (n-2-k)!} \lambda^{k+1} (1-\lambda)^{n-k-2} \right) \\
&= \frac{n^2 m \lambda \mu}{n+m} \left( 1 + (n-1)\lambda \sum_{k=0}^{n-2} \binom{n-2}{k} \lambda^k (1-\lambda)^{n-k-2} \right) \\
&= \frac{n^2 m \lambda \mu}{n+m} + \frac{n^2 m (n-1) \lambda^2 \mu}{n+m}.
\end{aligned}$$

$$\begin{aligned}
&\sum_{k=1}^{m+1} \frac{mk}{n+m} \binom{n}{k-1} (1-n\mu) \lambda^{k-1} (1-\lambda)^{n-k+1} \\
&\quad + \sum_{k=m+2}^{n+1} \frac{mk}{n+m} (1-n\mu) \binom{n}{k-1} \lambda^{k-1} (1-\lambda)^{n-k+1}
\end{aligned}$$



$$\begin{aligned}
&= \sum_{k=0}^m \frac{m(k+1)}{n+m} \binom{n}{k} (1-n\mu) \lambda^k (1-\lambda)^{n-k} \\
&\quad + \sum_{k=m+1}^n \frac{m(k+1)}{n+m} (1-n\mu) \binom{n}{k} \lambda^k (1-\lambda)^{n-k} \\
&= \frac{m(1-n\mu)}{n+m} \sum_{k=0}^n (k+1) \binom{n}{k} \lambda^k (1-\lambda)^{n-k} \\
&= \frac{m(1-n\mu)}{n+m} \left( 1 + \sum_{k=1}^n k \binom{n}{k} \lambda^k (1-\lambda)^{n-k} \right) \\
&= \frac{m(1-n\mu)}{n+m} + \frac{m(1-n\mu)}{n+m} n \lambda \sum_{k=0}^{n-1} \frac{(n-1)!}{k! (n-k-1)!} \lambda^k (1-\lambda)^{n-k-1} \\
&= \frac{m(1-n\mu)}{n+m} + \frac{nm\lambda(1-n\mu)}{n+m}.
\end{aligned}$$

$$\begin{aligned}
&\sum_{k=1}^{m+1} \frac{nmk\mu}{n+m} (1-m\lambda) \binom{m-1}{k-2} \mu^{k-2} (1-\mu)^{m-k+1} \\
&= \sum_{k=0}^m \frac{nm(k+1)\mu}{n+m} (1-m\lambda) \binom{m-1}{k-1} \mu^{k-1} (1-\mu)^{m-k} \\
&= \frac{nm(1-m\lambda)\mu}{n+m} \sum_{k=1}^m (k+1) \binom{m-1}{k-1} \mu^{k-1} (1-\mu)^{m-k} \\
&= \frac{nm(1-m\lambda)\mu}{n+m} \\
&\quad \cdot \left( \sum_{k=0}^{m-1} \binom{m-1}{k} \mu^k (1-\mu)^{m-1-k} + \binom{m-1}{k} (k+1) \mu^k (1-\mu)^{m-1-k} \right) \\
&= \frac{2nm(1-m\lambda)\mu}{n+m} \\
&\quad + \frac{nm(m-1)(1-m\lambda)\mu}{n+m} \sum_{k=1}^{m-1} \frac{(m-2)!}{(k-1)! (m-1-k)!} \mu^k (1-\mu)^{m-1-k} \\
&= \frac{2nm(1-m\lambda)\mu}{n+m} \\
&\quad + \frac{nm(m-1)(1-m\lambda)\mu^2}{n+m} \sum_{k=0}^{m-2} \frac{(m-2)!}{(k)! (m-2-k)!} \mu^k (1-\mu)^{m-2-k} \\
&= \frac{2nm(1-m\lambda)\mu}{n+m} + \frac{nm(m-1)(1-m\lambda)\mu^2}{n+m}.
\end{aligned}$$

$$\begin{aligned}
& \sum_{k=1}^{m+1} \frac{nm^2 k \lambda \mu}{n+m} \binom{m-1}{k-1} \mu^{k-1} (1-\mu)^{m-k} \\
&= \sum_{k=0}^m \frac{nm^2 (k+1) \lambda \mu}{n+m} \binom{m-1}{k} \mu^k (1-\mu)^{m-1-k} \\
&= \frac{nm^2 \lambda \mu}{n+m} \sum_{k=0}^m \binom{m-1}{k} (k+1) \mu^k (1-\mu)^{m-1-k} \\
&= \frac{nm^2 \lambda \mu}{n+m} \left( \sum_{k=0}^{m-1} \binom{m-1}{k} \mu^k (1-\mu)^{m-1-k} + \sum_{k=1}^{m-1} \binom{m-1}{k} k \mu^k (1-\mu)^{m-1-k} \right) \\
&= \frac{nm^2 \lambda \mu}{n+m} (1 + \mu(m-1)) \sum_{k=0}^{m-2} \frac{(m-2)!}{k! (m-2-k)!} \mu^k (1-\mu)^{m-2-k} \\
&= \frac{nm^2 \lambda \mu}{n+m} + \frac{nm^2 (m-1) \mu^2 \lambda}{n+m}.
\end{aligned}$$

So we have

$\mathbf{E}(|G|)$

$$\begin{aligned}
&= \frac{n(1-m\lambda)}{n+m} + \frac{nm\mu(1-m\lambda)}{n+m} + 2\frac{nm\lambda(1-n\mu)}{n+m} + \frac{nm\lambda^2(n-1)(1-n\mu)}{n+m} \\
&\quad + \frac{n^2 m \lambda \mu}{n+m} + \frac{n^2 m (n-1) \lambda^2 \mu}{n+m} + \frac{m(1-n\mu)}{n+m} + \frac{nm\lambda(1-n\mu)}{n+m} \\
&\quad + \frac{2nm(1-m\lambda)\mu}{n+m} + \frac{nm(m-1)(1-m\lambda)\mu^2}{n+m} + \frac{nm^2 \lambda \mu}{n+m} + \frac{nm^2(m-1)\mu^2 \lambda}{n+m} \\
&= 1 + \frac{2nm\mu - 2nm^2 \lambda \mu + 2nm\lambda - 2n^2 m \lambda \mu + n^2 m \lambda^2 - nm\lambda^2 + n\mu^2 m^2 - nm\mu^2}{n+m}.
\end{aligned}$$

## References

- [Antal and Scheuring 2006] T. Antal and I. Scheuring, “Fixation of strategies for an evolutionary game in finite populations”, *Bull. Math. Biol.* **68**:8 (2006), 1923–1944. MR 2293829
- [Broom and Rychtář 2008] M. Broom and J. Rychtář, “An analysis of the fixation probability of a mutant on special classes of non-directed graphs”, *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **464**:2098 (2008), 2609–2627. MR 2439285 Zbl 1152.92341
- [Broom and Rychtář 2012] M. Broom and J. Rychtář, “A general framework for analysing multiplayer games in networks using territorial interactions as a case study”, *J. Theoret. Biol.* **302** (2012), 70–80. MR 2909517
- [Broom et al. 1997] M. Broom, C. Cannings, and G. T. Vickers, “Multi-player matrix games”, *Bull. Math. Biol.* **59**:5 (1997), 931–952. Zbl 0923.92034
- [Ginsberg and Macdonald 1990] J. R. Ginsberg and D. D. W. Macdonald, *Foxes, wolves, jackals, and dogs: An action plan for the conservation of canids*, IUCN, Gland, Switzerland, 1990.

- [Hadjichrysanthou et al. 2011] C. Hadjichrysanthou, M. Broom, and J. Rychtář, “Evolutionary games on star graphs under various updating rules”, *Dyn. Games Appl.* **1**:3 (2011), 386–407. MR 2012i:91047 Zbl 1252.91016
- [Jetz et al. 2004] W. Jetz, C. Carbone, J. Fulford, and J. H. Brown, “The scaling of animal space use”, *Science* **306**:5694 (2004), 266–268.
- [Kelley et al. 2011] S. W. Kelley, D. Ransom, Jr., J. A. Butcher, G. G. Schulz, B. W. Surber, W. E. Pinchak, C. A. Santamaria, and L. A. Hurtado, “Home range dynamics, habitat selection, and survival of Greater Roadrunners”, *J. Field Ornithol.* **82**:2 (2011), 165–174.
- [Killingback and Doebeli 1996] T. Killingback and M. Doebeli, “Spatial evolutionary game theory: hawks and doves revisited”, *Proc. R. Soc. Lond. B* **263**:1374 (1996), 1135–1144.
- [Lieberman et al. 2005] E. Lieberman, C. Hauert, and M. A. Nowak, “Evolutionary dynamics on graphs”, *Nature* **433**:7023 (2005), 312–316.
- [Nowak 2006] M. A. Nowak, *Evolutionary dynamics: exploring the equations of life*, Belknap, Cambridge, MA, 2006. MR 2007g:92001 Zbl 1115.92047
- [Ohtsuki et al. 2006] H. Ohtsuki, C. Hauert, E. Lieberman, and M. A. Nowak, “A simple rule for the evolution of cooperation on graphs and social networks”, *Nature* **441**:7092 (2006), 502–505.
- [Santos and Pacheco 2006] F. C. Santos and J. M. Pacheco, “A new route to the evolution of cooperation”, *J. Evol. Biol.* **19**:3 (2006), 726–733.
- [Schaffer 1988] M. E. Schaffer, “Evolutionary stable strategies for a finite population and a variable contest size”, *J. Theor. Biol.* **132**:4 (1988), 469–478.

Received: 2012-06-08

Revised: 2012-10-23

Accepted: 2012-11-17

azur7777@aol.com

*UFR de Mathématique et d'Informatique, University of Strasbourg, 7 rue René Descartes, 67084 Strasbourg Cedex, France*

mark.broom@city.ac.uk

*Department of Mathematical Science, City University London, Northampton Square, London, EC1V 0HB, United Kingdom*

rychtar@uncg.edu

*Department of Mathematics and Statistics, The University of North Carolina at Greensboro, Greensboro, NC 27402, United States*



# Binary frames, graphs and erasures

Bernhard G. Bodmann, Bijan Camp and Dax Mahoney

(Communicated by Stephan Garcia)

This paper examines binary codes from a frame-theoretic viewpoint. Binary Parseval frames have convenient encoding and decoding maps. We characterize binary Parseval frames that are robust to one or two erasures. These characterizations are given in terms of the associated Gram matrix and with graph-theoretic conditions. We illustrate these results with frames in lowest dimensions that are robust to one or two erasures. In addition, we present necessary conditions for correcting a larger number of erasures. As in a previous paper, we emphasize in which ways the binary theory differs from the theory of frames for real and complex Hilbert spaces.

## 1. Introduction

In the last decades, frame theory has matured into a field with relevance in pure and applied mathematics as well as in engineering [Christensen 2003; Kovačević and Chebira 2007a; 2007b]. The simplest examples of frames are finite frames, finite spanning sequences in finite-dimensional real or complex Hilbert spaces. The possibility of having linear dependencies among the frame vectors can be used for error correction when a vector is encoded in terms of its frame coefficients, the inner products with the frame vectors [Goyal et al. 1998]. A common type of error considered in this context is an erasure, when part of the frame coefficients becomes corrupted or inaccessible and one has to recover the encoded vector from partial data [Marshall 1984; 1989]. The performance of frames for decoding erasures was studied, and in certain cases optimal frames could be characterized in a geometric fashion [Casazza and Kovačević 2003; Strohmer and Heath 2003; Holmes and Paulsen 2004; Püschel and Kovačević 2005], which was further extended with graph-theoretic or algebraic methods [Bodmann and Paulsen 2005; Xia et al. 2005; Kalra 2006; Bodmann et al. 2009b; Bodmann and Elwood 2010].

---

*MSC2010:* primary 42C15; secondary 94B05, 05C50.

*Keywords:* frames, Parseval frames, finite-dimensional vector spaces, binary numbers, codes, switching equivalence, Gram matrices, adjacency matrix, graphs.

This research was supported by NSF grant DMS-1109545.

Apart from the presence of the inner product, one could say that these applications in frame theory are similar to earlier work on error correcting linear codes over finite fields [MacWilliams and Sloane 1977; Betten et al. 2006]. Motivated by the literature in frame theory, a previous paper studied an analogue of Parseval frames in the setting of binary vector spaces [Bodmann et al. 2009a]; see also [Hotovy et al. 2012]. In this paper, we continue this direction of research and ask whether concepts from frame theory yield new insights for binary linear codes. We study how the Gram matrix of a binary frame relates to its robustness, its resilience to erasures. The space spanned by the columns of the Gram matrix is the set of all codewords, so the main question is in which way the robustness of a frame manifests itself. Interpreting the Gram matrix as the adjacency matrix of a graph gives a natural reformulation of conditions for robustness in terms of the connectivity properties of the graph. Note that this graph is different from the so-called Tanner graph of a binary code, which is a bipartite graph associated with the parity check matrix [Betten et al. 2006]. The space of code words is annihilated by the parity check matrix, so one can expect complementary insights from properties of the Gram and Tanner matrices with their associated graphs. While the structure of Tanner graphs has been studied with sophisticated methods in coding theory [Forney 2001; 2003; 2011], the Gram matrix and its role for erasures seems to appear mostly in the literature on frames; see, for example, [Holmes and Paulsen 2004; Bodmann and Paulsen 2005].

The remainder of this paper is structured as follows. In Section 2, we fix notation and define frames and Parseval frames for finite-dimensional binary vector spaces. Section 3 gives a motivation for the use of such frames as binary codes. In Section 4, we study robustness to erasures. Section 5 presents the results on robustness in graph-theoretic terms and gives the smallest frames with robustness to one or two erasures.

## 2. Preliminaries

We define binary frames and Parseval frames without appealing to the concept of an inner product, as in [Bodmann et al. 2009a]. The vector space that these families of vectors span is of the form  $\mathbb{Z}_2^n = \mathbb{Z}_2 \oplus \cdots \oplus \mathbb{Z}_2$  for some  $n \in \mathbb{N}$ , with the binary numbers  $\mathbb{Z}_2$  as the ground field.

**Definition 2.1.** A *frame* for  $\mathbb{Z}_2^n$  is a family of vectors  $\mathcal{F} = \{f_1, \dots, f_k\}$  such that

$$\text{span } \mathcal{F} = \mathbb{Z}_2^n.$$

To define a Parseval frame over  $\mathbb{Z}_2^n$ , we use a bilinear form that resembles the usual dot product on  $\mathbb{R}^n$ . For other choices of bilinear forms and a more general theory of binary frames, see [Hotovy et al. 2012].

**Definition 2.2.** The *dot product* on  $\mathbb{Z}_2^n$  is the bilinear map  $(\cdot, \cdot) : \mathbb{Z}_2^n \times \mathbb{Z}_2^n \rightarrow \mathbb{Z}_2$  given by

$$\left( \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}, \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \right) := \sum_{i=1}^n a_i b_i.$$

The dot product provides a natural map between vectors and linear functionals on  $\mathbb{Z}_2^n$ . With the help of this dot product, we define a Parseval frame for  $\mathbb{Z}_2^n$ .

**Definition 2.3.** A *Parseval frame* for  $\mathbb{Z}_2^n$  is a family of vectors  $\mathcal{F} = \{f_1, \dots, f_k\}$  in  $\mathbb{Z}_2^n$  such that

$$x = \sum_{j=1}^k (x, f_j) f_j \quad \text{for all } x \in \mathbb{Z}_2^n.$$

According to this definition, a Parseval frame provides a simple, redundant expansion for any vector  $x$  in  $\mathbb{Z}_2^n$ . Unless otherwise noted, when we speak of a frame or of a Parseval frame in this paper, we always mean families of vectors in  $\mathbb{Z}_2^n$  with the properties specified in Definitions 2.1 and 2.3, respectively. In the next section, we present a motivating example that explains the design problem of such frames as codes for erasures.

### 3. Binary frames as codes for erasures

Suppose Alice wants to send Bob a message that consists of a sequence of 0's and 1's. We can represent this message as the column vector

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{Z}_2^n,$$

where the entries  $x_1, x_2, \dots, x_n$  are the 1st, 2nd,  $\dots$ ,  $n$ -th digits of the message. Alice is aware that the message is sent through a somewhat unreliable medium, so she decides to *encode* it, that is, convert it into a new message which is generated from a codebook known to both Alice and Bob. The encoded message should have a reasonable chance of withstanding *erasures*, that is, removals of entries in the message that might occur. If the codebook is properly chosen, Bob will be able to recover the original message  $x$  from the fragments of the encoded message that remain.

The encoding is a linear map associated with a binary frame. Let the family of vectors  $\mathcal{F} = \{f_1, f_2, \dots, f_k\}$  be a frame for the vector space  $\mathbb{Z}_2^n$ , and let

$$\Theta_{\mathcal{F}} = \begin{pmatrix} \leftarrow f_1 \rightarrow \\ \leftarrow f_2 \rightarrow \\ \vdots \\ \leftarrow f_k \rightarrow \end{pmatrix} = \begin{pmatrix} f_{1,1} & f_{1,2} & \cdots & f_{1,n} \\ f_{2,1} & f_{2,2} & \cdots & f_{2,n} \\ \vdots & \vdots & & \vdots \\ f_{k,1} & f_{k,2} & \cdots & f_{k,n} \end{pmatrix},$$

where the entry  $f_{i,j}$  is the  $j$ -th entry of the  $i$ -th vector  $f_i \in \mathcal{F}$ . Alice encodes her message  $x$  by left-multiplying it with the matrix  $\Theta_{\mathcal{F}}$ . Consequently, Alice's encoded message will be a  $k \times 1$  matrix, where the  $i$ -th entry is the dot product  $(x, f_i)$ :

$$\Theta_{\mathcal{F}} x = \begin{pmatrix} f_{1,1} & f_{1,2} & \cdots & f_{1,n} \\ f_{2,1} & f_{2,2} & \cdots & f_{2,n} \\ \vdots & \vdots & & \vdots \\ f_{k,1} & f_{k,2} & \cdots & f_{k,n} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} (x, f_1) \\ (x, f_2) \\ \vdots \\ (x, f_k) \end{pmatrix}.$$

For convenience, let us abbreviate Alice's encoded message  $\Theta_{\mathcal{F}} x$  as

$$\Theta_{\mathcal{F}} x = y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{pmatrix}.$$

A first requirement for the choice of  $\mathcal{F}$  is that, if the encoded message arrives unaltered, then Bob can easily extract  $x$  from it. If  $\mathcal{F}$  is a Parseval frame, then this is indeed the case. In terms of  $\Theta_{\mathcal{F}}$ , the reconstruction identity in Definition 2.3 is

$$\Theta_{\mathcal{F}}^* \Theta_{\mathcal{F}} = I_n,$$

where  $\Theta_{\mathcal{F}}^*$  denotes the transpose of  $\Theta_{\mathcal{F}}$ .

Imagine at least one entry in the message  $y$  gets “erased”; that is, suppose Bob only receives the  $r \times 1$  matrix

$$\tilde{y} = \begin{pmatrix} y_{j_1} \\ y_{j_2} \\ \vdots \\ y_{j_r} \end{pmatrix},$$

where  $\{j_1, j_2, \dots, j_r\} \subset \{1, 2, \dots, k\}$ . For example, if there had been two erasures, then Bob would have received a  $(k-2) \times 1$  matrix with two of the original entries in  $y$  missing.



The goal is to reconstruct the original message  $x$  from the received matrix  $\tilde{y}$ . This can be achieved by finding an  $n \times r$  matrix  $\tilde{L}$  such that

$$x = \tilde{L} \begin{pmatrix} y_{j_1} \\ y_{j_2} \\ \vdots \\ y_{j_r} \end{pmatrix}.$$

A notationally more convenient way to formulate this problem is to use the full message without erasures but require reconstruction with a matrix  $L$  that has columns of zeros for the erased entries. To see this, let the columns of  $\tilde{L}$  be denoted by

$$\tilde{L} = \begin{pmatrix} \uparrow & \uparrow & \cdots & \uparrow \\ l_{j_1} & l_{j_2} & \cdots & l_{j_r} \\ \downarrow & \downarrow & \cdots & \downarrow \end{pmatrix},$$

and let the entries  $y_1, y_2,$  and  $y_4$  in  $y$  be erased. Then the matrix  $L$  is

$$L = \begin{pmatrix} \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \cdots & \uparrow \\ 0 & 0 & l_{j_1} & 0 & l_{j_2} & \cdots & l_{j_r} \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \cdots & \downarrow \end{pmatrix},$$

and there exists  $L$  of the above form such that  $x = Ly$  if and only if there exists  $\tilde{L}$  with  $x = \tilde{L}\tilde{y}$ . To characterize the requirement on  $L$  having columns of zeros, we write  $L = LE$ , where  $E$  is a diagonal 0-1-matrix with a 1 on the diagonal for any digit which gets transmitted and a 0 for every erased digit. With this terminology, we can reformulate the problem of correcting erasures as that of finding *any*  $L$  such that  $x = LE \Theta_{\mathcal{F}} x$  for each  $x \in \mathbb{Z}_2^n$ , that is, whether  $E \Theta_{\mathcal{F}}$  has a left inverse.

**Definition 3.1.** Let  $\mathcal{F} = \{f_1, f_2, \dots, f_k\}$  be a frame for  $\mathbb{Z}_2^n$ , and let  $E_J$  be a diagonal  $k \times k$  matrix associated with an erasure of digits indexed by  $J \subset \{1, 2, \dots, k\}$ , where  $(E_J)_{j,j} = 0$  if  $j \in J$  and  $(E_J)_{j,j} = 1$  otherwise. We say that the frame  $\mathcal{F}$  can *correct* the erasure if  $E_J \Theta_{\mathcal{F}}$  has a left inverse. We also say that the erasure of digits indexed by  $J$  is *correctable*.

The existence of a left inverse is equivalent to a rank condition and to the spanning property of the family of vectors corresponding to unaffected digits.

**Proposition 3.2.** Let  $\mathcal{F} = \{f_1, f_2, \dots, f_k\}$  be a frame for  $\mathbb{Z}_2^n$  and let  $J \subset \{1, 2, \dots, k\}$ . The following are equivalent:

- (1) The erasure of digits indexed by  $J$  is correctable.
- (2) The map  $E_J \Theta_{\mathcal{F}}$  is one-to-one.
- (3) The subfamily  $\tilde{\mathcal{F}} = \{f_j : j \notin J\}$  spans  $\mathbb{Z}_2^n$ ; that is, it is a frame.

(4) *The matrix  $E_J \Theta_{\mathcal{F}}$  has rank  $n$ .*

*Proof.* The equivalence of (1) and (2) is a standard exercise in linear algebra. We next prove the equivalence of (1) and (4). Let  $E_J \Theta_{\mathcal{F}}$  have rank  $n$ . Since  $\mathcal{F}$  is a frame,  $k \geq n$ . By elementary row operations,  $E_J \Theta_{\mathcal{F}}$  can be transformed into reduced row echelon form. However, this sequence of row operations can be obtained by multiplying with a suitable invertible matrix on the left. Thus, there is a  $k \times k$  matrix  $R$  such that

$$RE_J \Theta_{\mathcal{F}} = \begin{pmatrix} I_n \\ 0_{k-n,n} \end{pmatrix}.$$

Henceforth, we adopt block matrix notation and let  $I_n$  denote the  $n \times n$  identity matrix and  $0_{m,n}$  the  $m \times n$  zero matrix with  $m, n \in \mathbb{N}$ . Next, left multiplying this matrix by  $(I_n \ 0_{n,k-n})$  gives

$$(I_n \ 0_{n,k-n})RE_J \Theta_{\mathcal{F}} = I_n.$$

Thus, the required left inverse is  $L = (I_n \ 0_{n,k-n})R$ . On the other hand, if there is a left inverse for  $E_J \Theta_{\mathcal{F}}$  then this matrix must have the maximal possible rank,  $n$ .

To see the equivalence of (3) and (4), we observe that  $\tilde{\mathcal{F}}$  is spanning if and only if  $\Theta_{\tilde{\mathcal{F}}}$  has rank  $n$ , and the same is true for the matrix  $E_J \Theta_{\tilde{\mathcal{F}}}$ , where the frame vectors belonging to erased digits have been replaced by zero vectors.  $\square$

#### 4. Robustness to erasures

Next, we consider sets of erasures. A natural ordering is to consider erasures of at most one coefficient, then erasures of up to two, etc. A measure for robustness of a frame is how many erasures it can correct.

**Definition 4.1.** A frame  $\mathcal{F} = \{f_1, f_2, \dots, f_k\}$  for  $\mathbb{Z}_2^n$  is robust to  $m$  erasures if  $E_J \Theta_{\mathcal{F}}$  has a left inverse for any  $J \subset \{1, 2, \dots, k\}$  of size  $|J| \leq m$ .

Dimension counting gives a simple necessary condition for the size of a frame robust to  $m$  erasures.

**Proposition 4.2.** *If  $\mathcal{F} = \{f_1, f_2, \dots, f_k\}$  is a frame for  $\mathbb{Z}_2^n$  which is robust to  $m$  erasures, then  $k \geq n + m$ .*

*Proof.* If  $J \subset \{1, 2, \dots, k\}$  has size  $|J| = m$  then by assumption  $E_J \Theta_{\mathcal{F}}$  has a left inverse, and the subfamily  $\tilde{\mathcal{F}} = \{f_j : j \notin J\}$  spans  $\mathbb{Z}_2^n$ . Thus, the cardinality of  $\tilde{\mathcal{F}}$  is bounded by  $|\tilde{\mathcal{F}}| = k - m \geq n$ .  $\square$

Next, we wish to establish sufficient conditions which ensure robustness. If an erasure indexed by  $J$  is not correctable, then  $E_J \Theta_{\mathcal{F}}$  is not one-to-one and there exists a nonzero  $x \in \mathbb{Z}_2^n$  such that  $E_J \Theta_{\mathcal{F}} x = 0$ . For Parseval frames, there appears to be a simple condition in terms of an eigenvalue problem for submatrices of the Gramian. We prepare this with a lemma.

**Lemma 4.3.** *Let  $A$  be an  $n \times k$  matrix. The matrix  $AA^*$  has eigenvalue 1 if and only if  $A^*A$  has eigenvalue 1.*

*Proof.* Suppose that  $A^*A$  does have an eigenvalue equal to 1. That is, suppose that  $A^*Ax = x$ . Then  $y = Ax$  is nonzero and  $AA^*y = AA^*Ax = Ax = y$ . Hence,  $AA^*$  has an eigenvalue equal to 1. Switching the roles of  $A$  and  $A^*$  gives the converse.  $\square$

**Proposition 4.4.** *Let  $\mathcal{F} = \{f_1, f_2, \dots, f_k\}$  be a Parseval frame for  $\mathbb{Z}_2^n$  and let  $J \subset \{1, 2, \dots, k\}$ . If  $E_{J^c} \Theta_{\mathcal{F}} \Theta_{\mathcal{F}}^* E_{J^c}$  does not have eigenvalue one, where  $J^c$  is the complement of  $J$  in  $\{1, 2, \dots, k\}$ , then the erasure is correctable.*

*Proof.* We use the fact that  $AA^*$  has eigenvalue one if and only if  $A^*A$  does. Here,  $A = E_{J^c} \Theta_{\mathcal{F}} = (I - E_J) \Theta_{\mathcal{F}}$ . Assuming there is no eigenvector of eigenvalue one for  $A^*A$  means there exists no nonzero  $x$  such that

$$\Theta_{\mathcal{F}}^* (I - E_J) (I - E_J) \Theta_{\mathcal{F}} x = \Theta_{\mathcal{F}}^* (I - E_J) \Theta_{\mathcal{F}} x = x.$$

By assumption,  $\Theta_{\mathcal{F}}^* \Theta_{\mathcal{F}} = I$ , so this implies that there is no  $x \neq 0$  with

$$\Theta_{\mathcal{F}}^* E_J \Theta_{\mathcal{F}} x = 0.$$

Consequently,  $(\Theta_{\mathcal{F}}^* E_J \Theta_{\mathcal{F}})^{-1} \Theta_{\mathcal{F}}^* E_J \Theta_{\mathcal{F}} = I$  and the required left inverse of  $E_J \Theta_{\mathcal{F}}$  is

$$L = (\Theta_{\mathcal{F}}^* E_J \Theta_{\mathcal{F}})^{-1} \Theta_{\mathcal{F}}^*. \quad \square$$

At first glance, robustness against one erasure would motivate the search for frames whose vectors contain only an even number of ones, because then the diagonal of the Gram matrix  $\Theta_{\mathcal{F}} \Theta_{\mathcal{F}}^*$  would be zero, avoiding the eigenvalue one condition. However, such frames do not exist because any linear combination of vectors with an even number of ones still has an even number of ones. Thus, a family of such vectors cannot be spanning for all of  $\mathbb{Z}_2^n$ .

In addition, the above eigenvalue condition is sufficient for recovery, but not necessary. We present an example for this:

**Example 4.5.** Let  $n = 1$ ,  $\mathcal{F} = \{1, 1, 1\}$ , and  $J = \{2, 3\}$ . The encoded “vector”  $x \in \{0, 1\}$  is  $\Theta_{\mathcal{F}} x = (x \ x \ x)^*$ , and thus  $E_J \Theta_{\mathcal{F}}$  has the left inverse  $(1 \ 0 \ 0)$ . However,  $E_{J^c} \Theta_{\mathcal{F}} \Theta_{\mathcal{F}}^* E_{J^c} = E_{J^c}$  has eigenvalue one.

This motivates the search for a more general condition which ensures robustness. To this end, we introduce a function counting the number of 1’s in a vector, the (Hamming) weight.

**Definition 4.6.** A vector  $x \in \mathbb{Z}_2^n$  has *weight*  $w(x) = |\{j : x_j = 1\}|$ . We also speak of the *parity* of a vector, which is even or odd, depending on whether the weight is an even or an odd number.

**Theorem 4.7.** *Let  $\mathcal{F}$  be a Parseval frame with Gram matrix  $G = \Theta_{\mathcal{F}}\Theta_{\mathcal{F}}^*$ . The frame  $\mathcal{F}$  is robust to  $m$  erasures if and only if all the eigenvectors of  $G$  corresponding to eigenvalue one have a weight of at least  $m + 1$ .*

*Proof.* If  $\mathcal{F}$  is a Parseval frame, then any eigenvector of eigenvalue one of the Gram matrix is a possible message and vice versa. This is true because if  $y = \Theta_{\mathcal{F}}x$  then  $\Theta_{\mathcal{F}}\Theta_{\mathcal{F}}^*y = y$  and conversely if  $y$  is an eigenvector of eigenvalue one then  $y = \Theta_{\mathcal{F}}x$  for  $x = \Theta_{\mathcal{F}}^*y$ .

Assume that each such eigenvector has weight at least  $m + 1$ . If  $|J| \leq m$ , then applying  $E_J$  to  $y$  can only change at most  $m$  ones to zero, so  $E_Jy \neq 0$  and thus  $E_J\Theta_{\mathcal{F}}x \neq 0$  unless  $x = 0$ . This proves that  $E_J\Theta_{\mathcal{F}}$  is one-to-one.

Conversely, given a nonzero message  $y$ , if for each  $J \subset \{1, 2, \dots, k\}$  with  $|J| \leq m$  we have  $E_Jy \neq 0$ , then  $y$  must have weight at least  $m + 1$ .  $\square$

It is implicit in this characterization that the robustness of a frame against erasures is determined by the Gram matrix. If two frames have the same Gram matrix, then the two frames have identical robustness. Since the weight of a vector is invariant under permutations of its entries, the same holds if the Gram matrices differ only by a permutation of rows and columns. This means that the search for robust frames can be restricted to representatives of equivalence classes introduced in [Bodmann et al. 2009a].

**Definition 4.8.** Two frames  $\mathcal{F} = \{f_1, f_2, \dots, f_k\}$  and  $\mathcal{G} = \{g_1, g_2, \dots, g_k\}$  for  $\mathbb{Z}_2^n$  are called *switching equivalent* if there is a binary  $n \times n$  matrix  $U$  such that  $U^*U = UU^* = I$  and a permutation  $\sigma$  on the set  $\{1, 2, \dots, k\}$  such that  $f_j = Ug_{\sigma(j)}$  for all  $j \in \{1, 2, \dots, k\}$ .

**Theorem 4.9.** *If two frames  $\mathcal{F}$  and  $\mathcal{G}$  for  $\mathbb{Z}_2^n$  are switching equivalent, then  $\mathcal{F}$  is robust to  $m$  erasures if and only if  $\mathcal{G}$  is.*

*Proof.* If  $\mathcal{F}$  and  $\mathcal{G}$  are switching equivalent, then the Gram matrices of  $\mathcal{F}$  and  $\mathcal{G}$  differ by a permutation of rows and columns. The same is true for the eigenvectors corresponding to eigenvalue one. However, the weight of the eigenvectors is invariant under permutation of coordinates. This means, according to the preceding theorem, if  $\mathcal{F}$  is robust to  $m$  erasures, so is  $\mathcal{G}$ , and vice versa.  $\square$

In the context of real or complex Hilbert spaces, equal-norm frames characterize optimality for one erasure among Parseval frames [Casazza and Kovačević 2003]. In the binary setting, the equal-norm condition would correspond to a frame in which the vectors all have the same parity. Linear combinations of even vectors remain even, so there cannot be a frame consisting only of vectors having even parity, which leaves only the possibility of Parseval frames having only odd vectors. However, we show below that such frames have severe limitations for their robustness. We

prepare this with a lemma which is essentially a result in [Haemers et al. 1999, Lemma 2.2].

**Lemma 4.10.** *Let  $\mathcal{F} = \{f_1, f_2, \dots, f_k\}$  be a Parseval frame and  $G$  the associated Gram matrix; then the vector  $y$  with entries  $y_j = G_{j,j}$  for  $j \in \{1, 2, \dots, k\}$  is an eigenvector of  $G$  corresponding to eigenvalue one.*

*Proof.* Since  $G$  is idempotent, is enough to show that  $y_i = G_{i,i}$  defines a vector in the range of  $G$ . To see this, let  $(\text{ran } G)^\perp = \{x \in \mathbb{Z}_2^k, (x, z) = 0 \text{ for all } z \in \text{ran } G\}$  and recall  $((\text{ran } G)^\perp)^\perp = \text{ran } G$  because  $\text{ran } G \subset ((\text{ran } G)^\perp)^\perp$  by definition and  $\dim(\text{ran } G)^\perp + \dim \text{ran } G = k$ . If  $x \in (\text{ran } G)^\perp$ , then, by setting  $z = Gx$  and binary arithmetic,  $0 = (z, x) = \sum_{i,j=1}^k G_{i,j}x_i x_j = \sum_{j=1}^k G_{j,j}x_j$ . Thus,  $(x, y) = 0$  for each such  $x$ , and  $y$  is necessarily in  $\text{ran } G$ .  $\square$

Next, we examine how many erasures a binary Parseval frame can possibly correct. It turns out that, in some cases, the inequality necessary for correcting all  $m$ -erasures,  $k \geq n + m$ , can be strengthened considerably.

**Theorem 4.11.** *If  $\mathcal{F} = \{f_1, f_2, \dots, f_k\}$  is a Parseval frame of which  $p$  vectors are odd, then the frame cannot be robust to more than  $\min\{p - 1, k - p/2 - 1\}$  erasures.*

*Proof.* We recall that the vector  $y$  defined by  $y_j = G_{j,j}$ , the diagonal of the Gram matrix  $G$ , is an eigenvector of  $G$  corresponding to eigenvalue one, and that it has weight  $p$ . It is clear that the frame cannot correct more than  $p - 1$  erasures. On the other hand, assume that the minimal weight  $q$  among the vectors in the range of  $G$  is assumed by  $x$ , so  $p \geq q$ . The vector  $z = x + y$  is then also in the range of  $G$ . Define  $\Delta = q + p - k$ ; then the two vectors have at least  $\Delta$  indices in common for which the entries of both vectors are one. Thus, the weight of  $z$  is bounded by  $q \leq w(z) \leq q - \Delta + p - \Delta = 2k - q - p$ . This inequality gives  $q \leq k - p/2$ .  $\square$

This result shows that binary Parseval frames containing only odd vectors, the binary analogue of real or complex equal-norm Parseval frames, have a severe limitation for robustness.

**Corollary 4.12.** *If  $\mathcal{F} = \{f_1, f_2, \dots, f_k\}$  is a Parseval frame which consists only of odd vectors, then it cannot correct more than  $k/2 - 1$  erasures.*

Moreover, maximizing the upper bound for robustness yields that a binary Parseval frame achieves the best possible robustness when  $p - 1 = k - p/2 - 1$ , so  $p = 2k/3$ .

**Corollary 4.13.** *If  $\mathcal{F} = \{f_1, f_2, \dots, f_k\}$  is a Parseval frame for  $\mathbb{Z}_2^n$ , then it cannot correct more than  $2k/3 - 1$  erasures.*

### 5. Binary Parseval frames, graphs and erasures

With a binary symmetric  $k \times k$  matrix  $A$ , we associate a graph  $\gamma$  on  $k$  vertices. An entry  $A_{i,j} = 1$  means there is an edge connecting vertices  $i$  and  $j$ ; otherwise there is no edge between them. If  $A_{i,i} = 1$ , then vertex  $i$  has a loop, and we say that  $i$  is adjacent to itself; otherwise,  $i$  has no loop. The graph  $\gamma$  determines the matrix  $A$ , often called its adjacency matrix. We characterize binary Parseval frames in terms of the adjacency structure of the graph associated with the Gram matrix.

**Theorem 5.1.** *If  $\mathcal{F}$  is a binary frame and  $G = \Theta_{\mathcal{F}}\Theta_{\mathcal{F}}^*$  is its Gram matrix, then  $\mathcal{F}$  is a Parseval frame if and only if all of the following conditions hold for the graph  $\gamma$  associated with  $G$ :*

- (1) *Every vertex  $i$  has an even number of neighbors in the set  $\{1, 2, \dots, k\} \setminus \{i\}$ .*
- (2) *If two vertices of  $\gamma$  are not adjacent, then the two vertices have an even number of common neighbors.*
- (3) *If two vertices of  $\gamma$  are adjacent, then the two vertices have an odd number of common neighbors.*

*Proof.* First, suppose  $\mathcal{F}$  is Parseval. Then  $G^2 = \Theta_{\mathcal{F}}\Theta_{\mathcal{F}}^*\Theta_{\mathcal{F}}\Theta_{\mathcal{F}}^* = \Theta_{\mathcal{F}}\Theta_{\mathcal{F}}^* = G$ . From this, we conclude that the three properties are true.

(1) Let  $G_{i,i} = 1$ . Then  $\sum_j G_{i,j}G_{j,i} = \sum_j G_{i,j} = 1$ . Hence,  $\sum_{j,j \neq i} G_{i,j} = 0$ . On the other hand, let  $G_{i,i} = 0$ . Then  $\sum_j G_{i,j}G_{j,i} = 0$ , and consequently  $\sum_{j,j \neq i} G_{i,j} = 0$ . Thus, any vertex  $i$  has an even number of neighbors in the set of vertices not including  $i$ .

(2) If two vertices  $j$  and  $k$ ,  $j \neq k$ , are nonadjacent then  $0 = G_{j,k} = \sum_l G_{j,l}G_{l,k}$ . The nodes  $j$  and  $k$  then necessarily have an even number of common neighbors.

(3) If vertices  $j$  and  $k$  are adjacent nodes then  $1 = G_{j,k} = \sum_l G_{j,l}G_{l,k}$  and they have an odd number of common neighbors.

On the other hand, if these three properties hold then  $G^2 = G$  can be verified by a similar discussion of entries on the diagonal and on the off-diagonal: The property (1) implies that  $G_{i,i} = (G^2)_{i,i}$ , while (2) and (3) imply  $G_{j,k} = (G^2)_{j,k}$ . If  $\mathcal{F}$  is a frame, then the matrix  $\Theta_{\mathcal{F}}$  has rank  $n$ . Thus by appropriate elementary row operations it can be transformed into the row-reduced echelon form. These row operations amount to left multiplication with an invertible matrix  $R$ ,  $R\Theta_{\mathcal{F}} = \begin{pmatrix} I_n \\ 0_{n-k,n} \end{pmatrix}$ , and consequently  $\Theta_{\mathcal{F}}^*R^* = (I_n \ 0_{n,n-k})$ . If  $G^2 = G$ , then

$$RG^2R^* = \begin{pmatrix} I_n \\ 0_{n-k,n} \end{pmatrix} \Theta_{\mathcal{F}}^* \Theta_{\mathcal{F}} (I_n \ 0_{n,n-k}) = \begin{pmatrix} I_n & 0_{n,n-k} \\ 0_{n-k,n} & 0_{k,k} \end{pmatrix} = RGR^*,$$

and the middle equality shows that  $\Theta_{\mathcal{F}}^* \Theta_{\mathcal{F}} = I_n$ , so  $\mathcal{F}$  is Parseval.  $\square$

A graph that satisfies conditions (2) and (3) of Theorem 5.1 is not a strongly regular graph since the exact number of common neighbors may fluctuate between pairs of adjacent vertices and between pairs of nonadjacent vertices. However, since the number of common neighbors remains even or odd between pairs of adjacent or nonadjacent vertices, respectively, we propose the term *strongly parity regular graph* to refer to graphs that satisfy (2) and (3) of Theorem 5.1.

Next, we discuss graph-theoretic criteria for robustness to erasures. With Theorem 4.7, we have a characterization of robustness to  $m$  erasures in terms of the weights of the eigenvectors of the Gram matrix  $G$  corresponding to eigenvalue one. Because of the relation  $G^2 = G$ , these are precisely the vectors in the range of  $G$ . We can deduce a simple necessary and sufficient condition for the graph associated with a Parseval frame that is robust to one or two erasures.

**Theorem 5.2.** *Let  $\mathcal{F}$  be a Parseval frame for  $\mathbb{Z}_2^n$ ,  $G$  its Gram matrix and  $\gamma$  the associated graph. The frame  $\mathcal{F}$  is robust to one erasure if and only if every vertex of  $\gamma$  has at least two neighbors other than itself and is part of a cycle of length at most 4.*

*Proof.* First, we prove that robustness against one erasure implies the graph-theoretic properties. From the Parseval property, we know that each vertex has an even number of neighbors other than itself. If we pick a vertex  $i$  then the neighbors of it are encoded in the  $i$ -th column of the Gram matrix  $G$ . On the other hand, this column vector is in the range of  $G$ . If the frame corrects one erasure, then this vector must have at least weight two. Consequently, each vertex has to have at least two neighbors other than itself in order to correct one erasure.

Given a vertex  $i$  and two of its neighbors  $j$  and  $l$ ,  $i \neq j \neq l \neq i$ , then either the vertices  $j$  and  $l$  are adjacent and  $i$  is part of a 3-cycle, or they are not adjacent. In this case,  $j$  and  $l$  have an even number of common neighbors, so there is another vertex  $i'$  adjacent to  $j$  and  $l$ . Thus  $i, j, i'$  and  $l$  form a 4-cycle.

Next, we prove that the graph-theoretic properties ensure robustness against one erasure. For this, we only need to make the weaker assumption that each vertex has a neighbor other than itself. We note that a one-erasure not being correctable requires that there is a vector  $e_l$  from the standard basis, with some  $l \in \{1, 2, \dots, k\}$ , such that  $Ge_l = e_l$ . This implies that  $G_{j,l} = \delta_{j,l}$  for all  $j$ , so the  $l$ -th vertex is only a neighbor to itself. This is excluded by the assumption.  $\square$

Additional conditions characterize robustness against two erasures.

**Definition 5.3.** We say that a vertex  $i$  *discriminates* between two other vertices  $j$  and  $l$  if it is a neighbor to only one of them. We also say that the pair  $\{j, l\}$  has a *discriminating vertex*  $i$ .

**Theorem 5.4.** *Let  $\mathcal{F}$  be a Parseval frame for  $\mathbb{Z}_2^n$ ,  $G$  its Gram matrix and  $\gamma$  the associated graph. The frame  $\mathcal{F}$  is robust to two erasures if and only if the conditions for correcting one erasure hold and if every nonadjacent pair of vertices that are both adjacent to themselves and every adjacent pair of vertices that are both nonadjacent to themselves have a discriminating vertex.*

*Proof.* We first note that the graph-theoretic conditions in the preceding theorem are implied by robustness against two erasures which is a stronger requirement than correcting all one-erasures.

Next, we recall that Theorem 4.7 characterizes robustness in terms of the existence of certain eigenvectors. Assuming robustness against 1 erasure, an erasure of  $m = 2$  digits is not correctable if and only if there is a pair  $\{l, l'\}$  and  $h = e_l + e_{l'}$  satisfying  $Gh = h$ . Then,  $G_{l,l} = G_{l',l'} = 1$  and  $G_{l,l'} = 0$  or  $G_{l,l} = G_{l',l'} = 0$  and  $G_{l,l'} = 1$ . The first case corresponds to two nonadjacent vertices that are neighbors to themselves and the second one is a pair of adjacent vertices that are not neighbors to themselves. In both cases, the eigenvalue equation requires that  $G_{j,l} = G_{j,l'}$  for all  $j \notin \{l, l'\}$ . This means if a vertex  $j$  is adjacent to  $l$  then it is adjacent to  $l'$  and vice versa. We conclude that the eigenvalue equation is satisfied by  $h$  if and only if there is no vertex which discriminates between  $l$  and  $l'$ . Hence, all erasures of  $m = 2$  indices are correctable if and only if all one-erasures are and if there is a discriminating vertex for any nonadjacent pair of vertices that are both adjacent to themselves and any adjacent pair of vertices that are both nonadjacent to themselves.  $\square$

To illustrate these results, we use them to identify binary Parseval frames in 3 and 4 dimensions that achieve robustness to one or two erasures. We briefly mention that the canonical basis vectors form a Parseval frame that cannot correct any erasure, because they are minimal spanning sets. This means our search starts with 4 vectors in  $\mathbb{Z}_2^3$  and 5 vectors in  $\mathbb{Z}_2^4$ . Removing zero vectors from a frame does not affect the robustness as well as the Parseval property, so we can restrict ourselves to binary Parseval frames which do not contain zero vectors. Apart from zero vectors, identical pairs of vectors do not contribute to the reconstruction identity in Definition 2.3, which can be interpreted as a trivial form of incorporating redundancy in the encoding.

**Definition 5.5** [Bodmann et al. 2009a]. A binary Parseval frame  $\{f_1, f_2, \dots, f_k\}$  for  $\mathbb{Z}_2^n$  is called *trivially redundant* if there is  $j \in \{1, 2, \dots, k\}$  with  $f_j = 0$ , or if there are two indices  $i \neq j$  with  $f_i = f_j$ .

We restrict our study of robustness to binary Parseval frames that are not trivially redundant. This implies an upper bound on the number of frame vectors:

**Theorem 5.6** [Bodmann et al. 2009a]. *Let  $n \geq 3$ . Let  $\mathcal{F} = \{f_i\}_{i=1}^k$  be a family without repeated vectors in  $\mathbb{Z}_2^n$  and  $\mathcal{G} = \mathbb{Z}_2^n \setminus \mathcal{F}$ . If  $\mathcal{F}$  is a Parseval Frame for  $\mathbb{Z}_2^n$ , then  $\mathcal{G}$  is also a Parseval frame.*



**Corollary 5.7.** *If  $n \geq 3$  and  $\mathcal{F} = \{f_i\}_{i=1}^k$  is not trivially redundant, then  $k \leq 2^n - n - 1$ .*

*Proof.* If  $\mathcal{F}$  is Parseval, then so is  $\mathcal{G}$ . Removing the zero vector from  $\mathcal{G}$  gives a spanning set  $\mathcal{G} \setminus \{0\}$ , so it has at least  $n$  vectors. The union of  $\mathcal{F}$  and  $\mathcal{G} \setminus \{0\}$  has a total of  $2^n - 1$  vectors, so comparing sizes gives  $k + n \leq 2^n - 1$ .  $\square$

Switching equivalence allows a further simplification of the search. Since the robustness is the same for all representatives of a switching equivalence class, we can extract frames which are robust to one or two erasures from the classification of binary Parseval frames for  $\mathbb{Z}_2^3$  and  $\mathbb{Z}_2^4$  that are not trivially redundant [Bodmann et al. 2009a].

In  $n = 3$  dimensions, the above corollary limits the number of vectors in a binary Parseval frame that is not trivially redundant by  $k \leq 2^3 - 3 - 1 = 4$ . Up to switching equivalence, there are only two such binary Parseval frames for  $\mathbb{Z}_2^3$ : the canonical basis with 3 vectors and a binary Parseval frame with 4 vectors [ibid.]. Robustness to one erasure rules out the canonical basis, which leaves the case of 4 vectors. We examine the graph belonging to this Parseval frame, for readability purposes labeling vertices by the corresponding rows in  $\Theta_{\mathcal{F}}$ .

**Example 5.8.** The Parseval frame  $\mathcal{F}$  for  $\mathbb{Z}_2^3$  with encoding matrix

$$\Theta_{\mathcal{F}} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

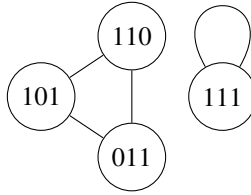
cannot correct one erasure because the graph associated with  $\Theta_{\mathcal{F}}\Theta_{\mathcal{F}}^*$  has an isolated vertex, as shown in Figure 1.

By the limit on the number of vectors, a Parseval frame for  $\mathbb{Z}_2^3$  which is robust to one erasure contains at least one repeated vector. We do not pursue this any further because it is a case of trivial redundancy.

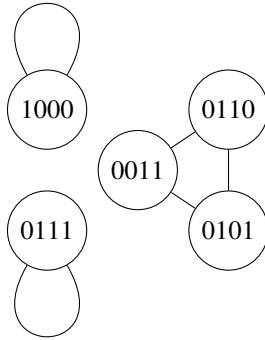
We proceed to  $n = 4$ . Here, the corollary limits the size of the frames we consider to  $k \leq 2^4 - 4 - 1 = 11$  vectors. As above, any graph with an isolated vertex prevents robustness to one erasure. This happens for the switching equivalence class of binary Parseval frames of 5 vectors for  $\mathbb{Z}_2^4$ .

**Example 5.9** [Bodmann et al. 2009a]. A Parseval frame  $\mathcal{F}$  for  $\mathbb{Z}_2^4$  with 5 vectors is, up to switching equivalence, given by

$$\Theta_{\mathcal{F}} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix}.$$



**Figure 1.** The graph associated with the Parseval frame of 4 vectors in  $\mathbb{Z}^3$  given in Example 5.8. Vertices are labeled by the corresponding rows of the encoding matrix. The presence of the isolated vertex (1 1 1) implies that this frame cannot correct one erasure.



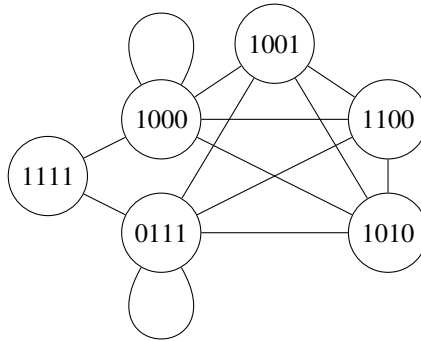
**Figure 2.** The graph belonging to the Parseval frame of 5 vectors in  $\mathbb{Z}^4$  given in Example 5.9 has the isolated vertices (1 0 0 0) and (0 1 1 1), so an erasure of the first frame coefficient or of the last one cannot be corrected.

The graph associated with the Gram matrix has two isolated vertices as shown in Figure 2, so the frame cannot correct one erasure.

Next, we identify a smallest binary Parseval frame for  $\mathbb{Z}_2^4$  which is not trivially redundant and can correct one erasure. There is only one switching equivalence class of Parseval frames for  $\mathbb{Z}_2^4$  containing 6 vectors [Bodmann et al. 2009a], so it is enough to investigate one representative.

**Example 5.10.** Let  $\mathcal{F}$  be the Parseval frame for  $\mathbb{Z}_2^4$  with

$$\Theta_{\mathcal{F}} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}.$$



**Figure 3.** The graph belonging to the Parseval frame of 6 vectors in  $\mathbb{Z}^4$  given in Example 5.10. Its adjacency structure satisfies the conditions in Theorem 5.2; thus it can correct one erasure. However,  $\mathcal{F}$  is not robust to two erasures since no vertices discriminate between the nonadjacent vertices (0 1 1 1) and (1 0 0 0) which are both adjacent to themselves.

The graph of  $\Theta_{\mathcal{F}}\Theta_{\mathcal{F}}^*$  satisfies the conditions for correcting one erasure stated in Theorem 5.2, which can be confirmed by inspecting Figure 3. However, it cannot correct more than one because it fails the requirement of discriminating vertices stated in Theorem 5.4.

The next larger Parseval frames form again a unique switching equivalence class [Bodmann et al. 2009a]. They fail to be robust to two erasures as well.

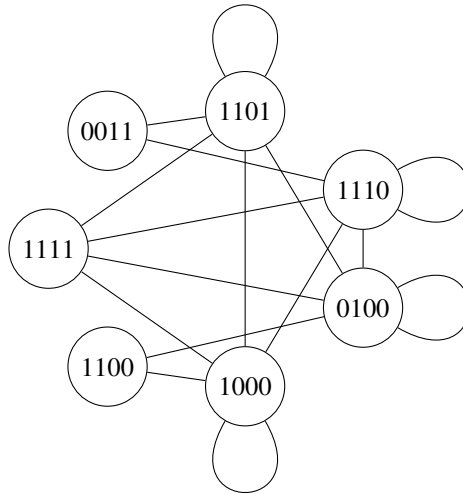
**Example 5.11.** Let  $\mathcal{F}$  be the binary Parseval frame for  $\mathbb{Z}_2^4$  containing seven vectors with

$$\Theta_{\mathcal{F}} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}.$$

The associated graph shown in Figure 4 satisfies the conditions of Theorem 5.2, but fails the conditions for correcting more than one, as described in Theorem 5.4.

Up to switching equivalence, the next example is the smallest binary Parseval frame for  $\mathbb{Z}_2^4$  which is not trivially redundant and can correct 2 erasures.

**Example 5.12.** Consider the binary Parseval frame  $\mathcal{F}$  for  $\mathbb{Z}_2^4$  with 8 vectors given



**Figure 4.** The graph associated with the Parseval frame of 7 vectors in  $\mathbb{Z}_2^4$  given in Example 5.11. It satisfies the connectivity conditions for correcting one erasure, but fails to be robust to two erasures because the nonadjacent vertices  $(1\ 1\ 0\ 1)$  and  $(1\ 1\ 1\ 0)$  are both adjacent to themselves and do not have any discriminating vertex.

by the matrix

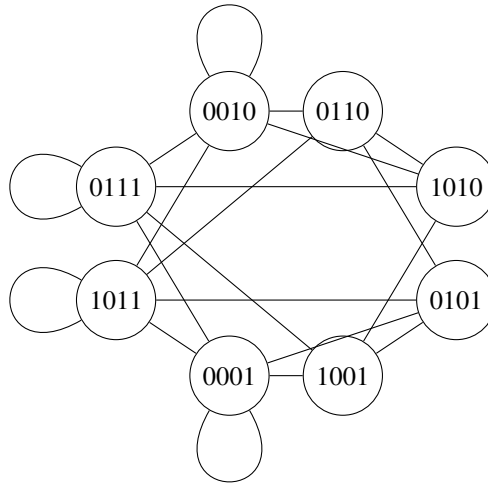
$$\Theta_{\mathcal{F}} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix}.$$

The associated graph shown in Figure 5 satisfies the conditions of Theorem 5.4, so it can correct up to two erasures.

Finally, we provide necessary conditions for correcting  $m$ -erasures, which require increased connectivity.

**Theorem 5.13.** *Let  $\mathcal{F}$  be a Parseval frame for  $\mathbb{Z}_2^n$ ,  $G$  its Gram matrix and  $\gamma$  the associated graph. If  $\mathcal{F}$  is robust to  $m \geq 1$  erasures, then every vertex has at least  $m + 1$  neighbors, possibly including itself, and it is part of at least  $m(m - 1)/2$  cycles of length at most 4.*

*Proof.* This condition follows again from the weights of the columns of  $G$ . If a vertex  $i$  is adjacent to itself then it needs at least  $m$  edges to other vertices. If it is



**Figure 5.** The graph associated with the Parseval frame of 8 vectors in  $\mathbb{Z}_2^4$  given in Example 5.12. It can correct up to two erasures because it satisfies the conditions of Theorem 5.4. For example, the vertex (0 1 1 0) discriminates between the nonadjacent vertices (1 0 1 1) and (0 1 1 1), and the vertex (0 0 1 0) discriminates between the adjacent vertices (1 0 1 0) and (1 0 0 1).

not adjacent to itself, it requires  $m + 1$  edges. Thus, there are at least  $m(m - 1)/2$  pairs of edges to other vertices. Any pair of such edges leads to either an adjacent pair or to a nonadjacent pair of vertices. If the pair is adjacent, then it forms a 3-cycle with the vertex  $i$ . Otherwise, the nonadjacent pair has a common neighbor other than the vertex  $i$ , forming a 4-cycle as in Theorem 5.2.  $\square$

Such necessary conditions are useful when searching for binary Parseval frames that are maximally robust. This could, in principle, be done by enumerating all Parseval frames and by testing their robustness against erasures exhaustively. The properties of the examples we have examined in  $\mathbb{Z}_2^3$  and  $\mathbb{Z}_2^4$  would, for example, be accessible by studying linear dependencies among the frame vectors. However, because of the combinatorial nature of robustness, it is advantageous for the search in higher dimensions if testing can be restricted to the subset of Parseval frames satisfying the necessary conditions.

### References

[Betten et al. 2006] A. Betten, M. Braun, H. Friepertinger, A. Kerber, A. Kohnert, and A. Wassermann, *Error-correcting linear codes*, Algorithms and Computation in Mathematics **18**, Springer, Berlin, 2006. MR 2008h:94001 Zbl 1102.94001

- [Bodmann and Elwood 2010] B. G. Bodmann and H. J. Elwood, “Complex equiangular Parseval frames and Seidel matrices containing  $p$ th roots of unity”, *Proc. Amer. Math. Soc.* **138**:12 (2010), 4387–4404. MR 2011h:42041 Zbl 1209.42020
- [Bodmann and Paulsen 2005] B. G. Bodmann and V. I. Paulsen, “Frames, graphs and erasures”, *Linear Algebra Appl.* **404** (2005), 118–146. MR 2006a:42047 Zbl 1088.46009
- [Bodmann et al. 2009a] B. G. Bodmann, M. Le, L. Reza, M. Tobin, and M. Tomforde, “Frame theory for binary vector spaces”, *Involve* **2**:5 (2009), 589–602. MR 2011b:42102 Zbl 1195.15001
- [Bodmann et al. 2009b] B. G. Bodmann, V. I. Paulsen, and M. Tomforde, “Equiangular tight frames from complex Seidel matrices containing cube roots of unity”, *Linear Algebra Appl.* **430**:1 (2009), 396–417. MR 2010b:42040 Zbl 1165.42007
- [Casazza and Kovačević 2003] P. G. Casazza and J. Kovačević, “Equal-norm tight frames with erasures”, *Adv. Comput. Math.* **18**:2-4 (2003), 387–430. MR 2004e:42046 Zbl 1035.42029
- [Christensen 2003] O. Christensen, *An introduction to frames and Riesz bases*, Birkhäuser, Boston, MA, 2003. MR 2003k:42001 Zbl 1017.42022
- [Forney 2001] G. D. Forney, Jr., “Codes on graphs: normal realizations”, *IEEE Trans. Inform. Theory* **47**:2 (2001), 520–548. MR 2002f:94055 Zbl 0998.94021
- [Forney 2003] G. D. Forney, Jr., “Codes on graphs: constraint complexity of cycle-free realizations of linear codes”, *IEEE Trans. Inform. Theory* **49**:7 (2003), 1597–1610. MR 2004f:94095
- [Forney 2011] G. D. Forney, Jr., “Codes on graphs: duality and MacWilliams identities”, *IEEE Trans. Inform. Theory* **57**:3 (2011), 1382–1397. MR 2012d:94106
- [Goyal et al. 1998] V. K. Goyal, M. Vetterli, and N. T. Thao, “Quantized overcomplete expansions in  $\mathbb{R}^N$ : Analysis, synthesis, and algorithms”, *IEEE Trans. Inform. Theory* **44**:1 (1998), 16–31. MR 99a:94004 Zbl 0905.94007
- [Haemers et al. 1999] W. H. Haemers, R. Peeters, and J. M. van Rijkevorsel, “Binary codes of strongly regular graphs”, *Des. Codes Cryptogr.* **17**:1-3 (1999), 187–209. MR 2000m:94035 Zbl 0938.05062
- [Holmes and Paulsen 2004] R. B. Holmes and V. I. Paulsen, “Optimal frames for erasures”, *Linear Algebra Appl.* **377** (2004), 31–51. MR 2004j:42028 Zbl 1042.46009
- [Hotovy et al. 2012] R. Hotovy, D. Larson, and S. Scholze, “Binary frames”, preprint, 2012, available at <http://www.math.tamu.edu/~larson/binary.pdf>.
- [Kalra 2006] D. Kalra, “Complex equiangular cyclic frames and erasures”, *Linear Algebra Appl.* **419**:2-3 (2006), 373–399. MR 2007j:42024 Zbl 1119.42013
- [Kovačević and Chebira 2007a] J. Kovačević and A. Chebira, “Life beyond bases: the advent of frames, I”, *IEEE Signal Proc. Mag.* **24**:4 (2007), 86–104.
- [Kovačević and Chebira 2007b] J. Kovačević and A. Chebira, “Life beyond bases: the advent of frames, II”, *IEEE Signal Proc. Mag.* **24**:5 (2007), 15–125.
- [MacWilliams and Sloane 1977] F. J. MacWilliams and N. J. A. Sloane, *The theory of error-correcting codes*, North-Holland Mathematical Library **16**, North-Holland, Amsterdam, 1977. MR 57 #5408a Zbl 0369.94008
- [Marshall 1984] T. Marshall, Jr., “Coding of real-number sequences for error correction: a digital signal processing problem”, *IEEE J. Sel. Areas Commun.* **2**:2 (1984), 381–392.
- [Marshall 1989] T. Marshall, “Fourier transform convolutional error-correcting codes”, pp. 658–662 in *Signals, Systems and Computers* (Asilomar, 1989), vol. 2, Maple, San Jose, CA, 1989.
- [Püschel and Kovačević 2005] M. Püschel and J. Kovačević, “Real, tight frames with maximal robustness to erasures”, pp. 63–72 in *Data Compression Conference* (Snowbird, UT, 2005), vol. 15, IEEE Computer Society, 2005.

[Strohmer and Heath 2003] T. Strohmer and R. W. Heath, Jr., “Grassmannian frames with applications to coding and communication”, *Appl. Comput. Harmon. Anal.* **14**:3 (2003), 257–275. MR 2004d:42053 Zbl 1028.42020

[Xia et al. 2005] P. Xia, S. Zhou, and G. B. Giannakis, “Achieving the Welch bound with difference sets”, *IEEE Trans. Inform. Theory* **51**:5 (2005), 1900–1907. MR 2007b:94148a Zbl 1237.94007

Received: 2012-06-25

Revised: 2012-09-14

Accepted: 2012-09-14

bgb@math.uh.edu

*Department of Mathematics, University of Houston,  
Houston, TX 77204, United States*

campx051@umn.edu

*Department of Psychology, University of Minnesota,  
Minneapolis, MN 55414, United States*

daxmahoney@gmail.com

*Department of Mathematics, University of Houston,  
Houston, TX 77204, United States*





# On groups with a class-preserving outer automorphism

Peter A. Brooksbank and Matthew S. Mizuhara

(Communicated by Nigel Boston)

Four infinite families of 2-groups are presented, all of whose members possess an outer automorphism that preserves conjugacy classes. The groups in these families are central extensions of their predecessors by a cyclic group of order 2. For each integer  $r > 1$ , there is precisely one 2-group of nilpotency class  $r$  in each of the four families. All other known families of 2-groups possessing a class-preserving outer automorphism consist entirely of groups of nilpotency class 2.

## 1. Introduction

Let  $G$  be a group,  $\text{Aut}(G)$  the automorphism group of  $G$ , and  $\text{Inn}(G)$  the subgroup of inner automorphisms. Then  $\text{Aut}(G)$  acts naturally on the set of conjugacy classes of  $G$ , and we denote the kernel of this action by  $\text{Aut}_c(G)$ . We refer to the elements of  $\text{Aut}_c(G)$  as *class-preserving automorphisms*. Evidently  $\text{Inn}(G) \trianglelefteq \text{Aut}_c(G)$ , and the elements of  $\text{Out}_c(G) = \text{Aut}_c(G)/\text{Inn}(G)$  will be referred to as *class-preserving outer automorphisms*.

Over a century ago, William Burnside [1911, Note B, p. 463] asked the question: *Are there groups  $G$  such that  $\text{Out}_c(G) \neq 1$ ?* He himself settled the question soon thereafter [Burnside 1913]: *for each prime  $p \equiv \pm 3 \pmod{8}$ , there is a group  $G_p$  of order  $p^6$  and nilpotency class 2 with  $\text{Out}_c(G_p) \neq 1$ .*

Since Burnside's initial discovery, the problem has been revisited on many occasions, and new families of groups  $G$  with  $\text{Out}_c(G) \neq 1$  have been found. Until fairly recently, however, most of those families consisted of  $p$ -groups of nilpotency class 2. The object of this paper is to prove the following result.

**Theorem 1.1.** *There are four distinct infinite families  $\mathcal{H} = \{H_j\}_{j=1}^\infty$ , where  $H_j$  is a 4-generator 2-group of order  $2^{5+j}$  and nilpotency class  $j+1$  such that  $\text{Out}_c(H_j) \neq 1$ .*

MSC2010: 20D15, 20D45, 20E45.

Keywords:  $p$ -groups, class-preserving automorphisms, polycyclic groups.

Project sponsored by the National Security Agency under Grant Number H98230-11-1-0146. The United States Government is authorized to reproduce and distribute reprints notwithstanding any copyright notation herein.

It is evident from the statement of Theorem 1.1 that the nilpotency class of the groups  $H_j$  in each family grows in an elementary way as a function of the group orders. This is because  $H_{j+1}$  is built as a central extension of  $H_j$  by  $\mathbb{Z}/2$ . Indeed, each  $\mathcal{H}$  may be constructed algorithmically using the  $p$ -group generation algorithm [O'Brien 1990]; this is precisely how the families were discovered and studied. Furthermore, the groups in all four families have coclass 4, so we have shown that they are all “mainline groups” in the coclass graph  $\mathcal{G}(2, 4)$  (see [Eick and Leedham-Green 2008]).

Readers interested in the history and applications of Burnside’s problem are referred to the recent comprehensive survey of Yadav [2011]; we restrict ourselves here to a brief summary of those results pertaining directly to Theorem 1.1.

Wall [1947] showed that, for each integer  $m$  divisible by 8, the general linear group  $\mathrm{GL}(1, \mathbb{Z}/m)$  (i.e., the group of linear permutations  $x \mapsto \sigma x + \tau$  on integers modulo  $m$  with  $\sigma, \tau$  integral) has a class-preserving automorphism that is not inner. This family includes the smallest group  $G$  such that  $\mathrm{Out}_c(G) \neq 1$ , namely  $\mathrm{GL}(1, \mathbb{Z}/8)$  of order 32 (there, in fact, are two nonisomorphic groups of order 32 having this property). The 2-groups in Wall’s family, namely  $\mathrm{GL}(1, \mathbb{Z}/2^k)$ , have nilpotency class 2.

Heineken [1979] constructed, for each odd prime  $p$ , an infinite family of  $p$ -groups of nilpotency class 2, *all* of whose automorphisms are class-preserving. As far as we are aware, these are the only known infinite families of groups  $G$  for which  $\mathrm{Aut}_c(G) = \mathrm{Aut}(G)$ .

Hertweck [2001] constructed a family of Frobenius groups as subgroups of affine semilinear groups  $A\Gamma(F)$ , where  $F$  is a finite field, which possess class-preserving automorphisms that are not inner.

Malinowska [1992] exhibited, for each prime  $p > 5$  and each  $r > 2$ , a  $p$ -group  $G$  of nilpotency class  $r$  such that  $\mathrm{Out}_c(G) \neq 1$ . Unlike the groups in our families, however, it is not clear how the order of  $G$  relates to  $r$ .

We remark that the absence of simple groups in the above summary is explained by Feit and Seitz [1989, Section C]: *if  $G$  is a finite simple group then  $\mathrm{Out}_c(G) = 1$ .*

Briefly, the paper is organized as follows. In Section 2 we summarize the necessary background on  $p$ -groups. The families  $\mathcal{H}$  in Theorem 1.1 are introduced in Section 3; they are naturally parametrized by vectors  $\epsilon \in \{0, 1\}^4$ , but there only four distinct families. The proof of Theorem 1.1 is given in Section 4.

## 2. Preliminaries

Our notation and terminology is standard. For elements  $x, y$  of a group, we write  $x^y = y^{-1}xy$  and  $[x, y] = x^{-1}x^y$ . For subsets  $X$  and  $Y$  of a group, we denote by  $[X, Y]$  the subgroup generated by all commutators  $[x, y]$ , where  $x \in X$  and  $y \in Y$ .

The *lower central series* of a group  $G$  is the series

$$G = \gamma_1(G) \geq \gamma_2(G) \geq \dots, \tag{1}$$

where  $\gamma_{i+1}(G) = [G, \gamma_i(G)]$ . A group  $G$  is *nilpotent* if  $\gamma_i(G) = 1$  for some  $i \geq 1$ , in which case the smallest  $r$  such that  $\gamma_{r+1}(G) = 1$  is called the *nilpotency class* (or simply *class*) of  $G$ . A finite group  $G$  is a *p-group* if  $|G| = p^n$  for some prime  $p$ . All  $p$ -groups are nilpotent, and if  $G$  has class  $r$ , then  $G$  has *coclass*  $n - r$ . A  $p$ -group minimally generated by  $d$  elements is called a *d-generator group*.

Each nilpotent group (more generally, each soluble group) possesses a *polycyclic generating sequence* [Holt et al. 2005, Chapter 8]. This in turn gives rise to a *power-conjugate presentation* (or simply *pc-presentation*), an extremely efficient model for computing with soluble groups. We describe these presentations specifically for  $p$ -groups.

Fix a  $p$ -group  $G$ . Let  $X = [x_1, \dots, x_n] \subset G$  be such that if  $P_i = \langle x_i, \dots, x_n \rangle$  ( $i = 1, \dots, n$ ), then  $P_i/P_{i+1}$  has order  $p$ , and  $G = P_1 > P_2 > \dots > P_n > 1$  refines the lower central series in (1). If  $G$  has nilpotency class  $r$ , we define a *weighting*,  $w: X \rightarrow \{1, \dots, r\}$ , where  $w(x_i) = k$  if  $x_i \in \gamma_{k-1}(G) \setminus \gamma_k(G)$ . Evidently,  $w(x_i) \geq w(x_j)$  whenever  $i \geq j$ . Any such sequence  $X$  satisfies the conditions needed to serve as the generating sequence of a *weighted pc-presentation* of  $G$ . The relations,  $R$ , in such a presentation all have the form

$$x_i^p = \prod_{k=i+1}^n x_k^{b(i,k)}, \quad \text{where } 0 \leq b(i,k) < p, \quad 1 \leq i \leq n,$$

or (2)

$$x_j^{x_i} = x_j \prod_{k=j+1}^n x_k^{b(i,j,k)}, \quad \text{where } 0 \leq b(i,j,k) < p, \quad 1 \leq i < j \leq n.$$

We write  $\langle X \mid R \rangle$  to denote the  $p$ -group defined by such a presentation. We adopt the usual convention that an omitted relation  $x_i^p$  implies that  $x_i^p = 1$ , and an omitted relation  $x_j^{x_i}$  implies that  $x_i$  and  $x_j$  commute. We will often find it convenient to write a conjugate relation  $x_j^{x_i} = x_j w$  as a commutator relation  $[x_j, x_i] = w$ .

**Remark 2.1.** In general, one requires that  $G = P_1 > \dots > P_n > 1$  refines a related series called the *exponent p-central series* [Holt et al. 2005, p. 355]. For the families of  $p$ -groups we consider here, however, the two series coincide.

A critical feature of a  $pc$ -presentation for a  $p$ -group is that elements of the group inherit a *normal form*  $x_1^{a_1} x_2^{a_2} \dots x_n^{a_n}$ , where  $0 \leq a_i < p$ . Given  $g \in G$  as a word in  $x_1, \dots, x_n$ , a normal form may be obtained by repeatedly applying the relations in (2) in a process known as *collection*. If each element of  $G$  has a unique normal form, the  $pc$ -presentation is said to be *consistent*. Clearly if  $G$  has a consistent  $pc$ -presentation on  $X = [x_1, \dots, x_n]$ , then  $|G| = p^n$ .

We conclude this section with a useful test for consistency. We state it just for 2-groups — since this is all we need — and refer the reader to [Holt et al. 2005, Theorem 9.22] for the more general version.

**Proposition 2.2.** *A weighted pc-presentation of a  $d$ -generator 2-group of class  $r$  on  $[x_1, \dots, x_n]$  is consistent if the following pairs of words in the generators have the same normal form (the products in parentheses are collected first):*

$$\begin{aligned} (x_k x_j) x_i \text{ and } x_k (x_j x_i), & \quad 1 \leq i < j < k \leq n \text{ and } i \leq d, \quad w(x_i) + w(x_j) + w(x_k) \leq r; \\ (x_j x_i) x_i \text{ and } x_j (x_j x_i), & \quad 1 \leq i < j \leq n \text{ and } i \leq d, \quad w(x_i) + w(x_j) < r; \\ (x_j x_i) x_i \text{ and } x_j (x_i x_i), & \quad 1 \leq i < j \leq n, \quad w(x_i) + w(x_j) < r; \\ (x_i x_i) x_i \text{ and } x_i (x_i x_i), & \quad 1 \leq i \leq n, \quad 2w(x_i) < r. \end{aligned}$$

### 3. The families $\mathcal{H}^\epsilon$

In this section we introduce four infinite families of 4-generator 2-groups of fixed coclass 4. In the next section we will show that each family consists of groups that have a class-preserving outer automorphism, thus proving Theorem 1.1.

We will define the groups in each family by giving consistent pc-presentations. It is convenient to denote the ordered list of pc-generators of the  $n$ -th group in each family by  $X_n = \{x_1, x_2, x_3, x_4, z, y_1, \dots, y_n\}$ , with the group minimally generated by  $\{x_1, x_2, x_3, x_4\}$ . The commutator relations for each family are identical, namely

$$\begin{aligned} C_n = \{[x_2, x_1] = [x_3, x_2] = [x_4, x_1] = z, \quad [x_3, x_1] = y_1, \\ [x_1, y_i] = [x_3, y_i] = y_{i+1} \quad (i = 1, \dots, n-1)\}. \end{aligned} \quad (3)$$

For each  $\epsilon = (\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4) \in \{0, 1\}^4$ , define

$$\begin{aligned} P_n^\epsilon = \{x_j^2 = z^{\epsilon_j} \quad (j = 1, \dots, 4), \quad z^2 = 1, \\ y_n^2 = 1, \quad y_i^2 = y_{i+1} y_{i+2} \quad (i = 1, \dots, n-2), \quad y_{n-1}^2 = y_n\}. \end{aligned} \quad (4)$$

Let  $R_n^\epsilon = C_n \cup P_n^\epsilon$ , define  $H_n^\epsilon = \langle X_n \mid R_n^\epsilon \rangle$ , and put  $\mathcal{H}^\epsilon = \{H_n^\epsilon\}_{n=1}^\infty$ . Note that the pc-presentations for the  $n$ -th group in each family differ only in the power relations of the generators  $x_j$ .

**Proposition 3.1.** *Let  $n$  be a positive integer, and  $\epsilon \in \{0, 1\}^4$ . Then  $H_n^\epsilon = \langle X_n \mid R_n^\epsilon \rangle$  has order  $2^{n+5}$  and class  $n+1$  (hence coclass 4).*

*Proof.* To confirm the order of  $H_n^\epsilon$ , it suffices to check that their defining pc-presentations are consistent, for which we use Proposition 2.2. Although there are  $O(n^3)$  computations involved in that test, the lion's share of these may be treated uniformly for the groups  $H_n^\epsilon$ . The following table lists all of the triples that must be checked, together with their normal forms. Triples involving  $z$  are

omitted (since  $z$  is central), as are triples involving two or more  $y_s$  generators (since  $\langle y_s : s = 1, \dots, n \rangle$  is abelian).

Triple $(a, b, c)$	Conditions	Normal form of $a(bc)$ and $(ab)c$
$(x_3, x_2, x_1)$ $(x_4, x_2, x_1)$ $(x_4, x_3, x_1)$ $(x_4, x_3, x_2)$		$x_1x_2x_3y_1$ $x_1x_2x_4$ $x_1x_3x_4zy_1$ $x_2x_3x_4z$
$(y_s, x_2, x_1)$ $(y_s, x_3, x_1)$ $(y_s, x_4, x_1)$ $(y_s, x_3, x_2)$ $(y_s, x_4, x_2)$ $(y_s, x_4, x_3)$	$s \leq n - 2$ $s \leq n - 2$ $s \leq n - 2$ $s \leq n - 2$ $s \leq n - 2$ $s \leq n - 2$	$x_1x_2zy_sy_{s+1}$ $x_1x_3y_1y_s$ $x_1x_4zy_sy_{s+1}$ $x_2x_3zy_sy_{s+1}$ $x_2x_4y_s$ $x_3x_4y_sy_{s+1}$
$(x_j, x_j, x_i)$ $(y_s, y_s, x_i)$ $(x_j, x_i, x_i)$ $(y_s, x_i, x_i)$	$1 \leq i < j \leq 4$ $s \leq n - 2, i = 1, 3$ $1 \leq i < j \leq 4$ $s \leq n - 2, i \leq 4$	$x_i z^{e_j}$ $x_j y_{s+1}$ $x_j z^{e_i}$ $z^{e_i} y_s$
$(x_i, x_i, x_i)$	$i \leq 4$	$x_i z^{e_i}$

Routine calculations using the pc-relations are all that is needed to verify the normal forms listed in the table. It remains to compute the lower central series of  $H_n^\epsilon$ :

$$\begin{aligned} \gamma_1(H_n^\epsilon) &= H_n^\epsilon, \\ \gamma_2(H_n^\epsilon) &= \langle z, y_i : 1 \leq i \leq n \rangle, \\ \gamma_j(H_n^\epsilon) &= \langle y_i : j - 1 \leq i \leq n \rangle \quad \text{for } j = 3, \dots, n + 1, \\ \gamma_{n+2}(H_n^\epsilon) &= 1. \end{aligned}$$

This shows that  $H_n^\epsilon$  has class  $n + 1$ , as stated. □

Proposition 3.1 suggests that there are 16 families  $\mathcal{H}^\epsilon$ , but the following result shows that there is some duplication.

**Proposition 3.2.** *For each positive integer  $n$ , there are four isomorphism classes among the groups  $\{H_n^\epsilon : \epsilon \in \{0, 1\}^4\}$ .*

*Proof.* Each group  $H = H_n^\epsilon$  determines a quadratic map  $\mathbf{q} = \mathbf{q}^\epsilon$  (independent of  $n$ ) as follows. Let  $V$  denote the largest elementary abelian quotient of  $H$ , namely  $V = H/A \cong (\mathbb{Z}/2)^4$ , where  $A = \langle z, y_1, \dots, y_n \rangle$ . Let  $W$  denote the largest elementary abelian quotient of  $A$ , namely  $W = A/B \cong (\mathbb{Z}/2)^2$ , where  $B = \langle y_2, \dots, y_n \rangle$ . Define maps  $\mathbf{q} : V \rightarrow W$  and  $\mathbf{b} : V \times V \rightarrow W$ , where  $\mathbf{q}(xA) = x^2B$  and  $\mathbf{b}(xA, yA) =$

$[x, y]B$  for all  $x, y \in H$ . Using additive notation in  $V$  and  $W$ , one easily checks that

$$\mathbf{b}(u, v) = \mathbf{q}(u + v) + \mathbf{q}(u) + \mathbf{q}(v) \text{ for all } u, v \in V, \tag{5}$$

so  $\mathbf{b}$  is the symmetric bilinear map associated to  $\mathbf{q}$  in the familiar sense.

If  $H_n^\epsilon$  and  $H_n^\delta$  are isomorphic groups, and  $\alpha : H_n^\epsilon \rightarrow H_n^\delta$  is any isomorphism, then  $\alpha$  induces isomorphisms  $\beta : V^\epsilon \rightarrow V^\delta$  and  $\gamma : W^\epsilon \rightarrow W^\delta$  such that  $\mathbf{q}^\delta(v\beta) = \mathbf{q}^\epsilon(v)\gamma$  for all  $v \in V^\epsilon$ . Thus  $\alpha$  induces a *pseudo-isometry* between  $\mathbf{q}^\epsilon$  and  $\mathbf{q}^\delta$ .

Fixing a basis  $\{v_i\}$  for  $V$ , one can represent a quadratic map  $\mathbf{q}$  as a  $4 \times 4$  matrix  $\mathbf{Q} = [[q_{ij}]]$  with entries in  $W$ , where  $q_{ii} = \mathbf{q}(v_i)$ ,  $q_{ij} = \mathbf{b}(v_i, v_j)$  if  $i < j$ , and  $q_{ij} = 0$  if  $i > j$ . Given  $v \in V$ , write  $v = \sum \lambda_i v_i$  with  $\lambda_i \in \mathbb{Z}/2$ . Using (5) and a finite induction, we see that  $\mathbf{q}(v) = \sum_i \sum_{j \geq i} \lambda_i \lambda_j q_{ij}$ . An easy matrix calculation then shows that  $\mathbf{q}(v) = v\mathbf{Q}v^{\text{tr}}$  for all  $v \in V$ .

Using the basis  $\{x_i A\}$  for  $V$ , and identifying  $A/B$  on basis  $\{zB, y_1 B\}$  with the additive group of the ring  $(\mathbb{Z}/2)[t]/(t^2)$  on the usual basis  $\{1, t\}$ , the matrix representing  $\mathbf{q} = \mathbf{q}^\epsilon$ , where  $\epsilon = (\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4)$ , is

$$\mathbf{Q} = \begin{bmatrix} \epsilon_1 & 1 & t & 1 \\ 0 & \epsilon_2 & 1 & 0 \\ 0 & 0 & \epsilon_3 & 0 \\ 0 & 0 & 0 & \epsilon_4 \end{bmatrix},$$

and the matrix representing the associated bilinear map  $\mathbf{b}$  is  $\mathbf{B} = \mathbf{Q} + \mathbf{Q}^{\text{tr}}$ .

Given maps  $\mathbf{q}^\epsilon$  and  $\mathbf{q}^\delta$  representing groups  $H^\epsilon$  and  $H^\delta$  ( $\epsilon, \delta \in \{0, 1\}^4$ ), one can easily test for pseudo-isometry as follows. Let  $\mathbf{Q}^\epsilon$  and  $\mathbf{Q}^\delta$  be matrices representing  $\mathbf{q}^\epsilon$  and  $\mathbf{q}^\delta$ . If  $g \in \text{GL}(4, 2)$  represents an isomorphism  $H^\epsilon/A^\epsilon \rightarrow H^\delta/A^\delta$  induced by an isomorphism  $H^\epsilon \rightarrow H^\delta$ , then the induced isomorphism  $A^\epsilon/B^\epsilon \rightarrow A^\delta/B^\delta$  is uniquely determined by  $g$ , and its matrix  $h \in \text{GL}(2, 2)$  is easily computed. Extend  $h$  entry-wise to a map  $\mathbb{M}_4(W^\epsilon) \rightarrow \mathbb{M}_4(W^\delta)$ , and denote the image of  $X \in \mathbb{M}_4(W^\epsilon)$  by  $X^h$ . Then  $\mathbf{q}^\epsilon$  and  $\mathbf{q}^\delta$  are pseudo-isometric if and only if there exists  $g \in \text{GL}(4, 2)$  such that

$$g\mathbf{B}^\delta g^{\text{tr}} = (\mathbf{B}^\epsilon)^h \quad \text{and} \quad v_i(g\mathbf{Q}^\delta g^{\text{tr}})v_i^{\text{tr}} = v_i(\mathbf{Q}^\epsilon)^h v_i^{\text{tr}},$$

as  $v_i$  runs over a basis for  $(\mathbb{Z}/2)^4$ .

Thus, the determination of the pseudo-isometry classes of the quadratic maps associated to the families  $\mathcal{H}^\epsilon$  is an elementary matrix calculation in  $\text{GL}(4, 2)$ , which is easily carried out using a computer algebra system such as MAGMA [Bosma et al. 1997]. Those classes are represented by

$$\mathbf{Q}^\epsilon \quad \text{for } \epsilon \in \{(0, 0, 0, 0), (0, 0, 1, 1), (1, 1, 0, 0), (1, 1, 1, 1)\}.$$

Finally, it is not difficult to verify that any pseudo-isometry  $\mathbf{Q}^\epsilon \rightarrow \mathbf{Q}^\delta$  lifts to an

isomorphism  $H^\epsilon \rightarrow H^\delta$ . Thus, for each  $n$ , there are precisely four isomorphism classes of group  $H_n^\epsilon$ , as claimed.  $\square$

#### 4. Proof of Theorem 1.1

In this section we complete the proof of Theorem 1.1 by exhibiting a class-preserving automorphism of each group  $H_n^\epsilon$  that is not inner.

*Proof of Theorem 1.1.* Fix  $n \geq 1$ ,  $\epsilon \in \{0, 1\}^4$ , and put  $H = H_n^\epsilon$ . Define  $\theta: H \rightarrow H$  on generators, sending

$$x \mapsto \begin{cases} x_4 z & \text{if } x = x_4, \\ x & \text{if } x \in X_n \setminus \{x_4\}. \end{cases} \quad (6)$$

One easily verifies (by replacing  $x_4$  with  $x_4 z$  in each pc-relation involving  $x_4$  and evaluating) that  $\theta \in \text{Aut}(H)$ .

First, suppose that  $\theta$  is an inner automorphism. Then there exists  $h \in H$  commuting with  $x_1$  and  $x_3$ , but not with  $x_4$ . Writing

$$h = \prod_{i=1}^4 x_i^{a_i} \cdot z^b \cdot \prod_{j=1}^n y_j^{c_j} \quad (a_i, b, c_j \in \{0, 1\}) \quad (7)$$

and using the defining commutator relations of  $H$ , we see that

$$hx_1 = x_1 h \cdot \left( z^{a_2+a_4} y_1^{a_3} \prod_{j=2}^n y_j^{c_{j-1}} \right).$$

Hence  $h \in C_H(x_1)$  if and only if  $a_2 = a_4$  and  $0 = a_3 = c_1 = \dots = c_{n-1}$ . Also,

$$x_3 h = x_1^{a_1} x_2^{a_2} x_3^{1+a_3} x_4^{a_4} z^{a_2+b} y_1^{a_1+c_1} \prod_{j=2}^n y_j^{c_j},$$

while

$$hx_3 = x_1^{a_1} x_2^{a_2} x_3^{1+a_3} x_4^{a_4} z^b y_1^{c_1} \prod_{j=2}^n y_j^{c_j} \prod_{j=2}^n y_j^{c_{j-1}},$$

so that  $h \in C_H(x_3)$  if and only if  $0 = a_1 = a_2 = c_1 = \dots = c_{n-1}$ . It follows that  $C_H(x_1) \cap C_H(x_3) = \langle z, y_n \rangle = Z(H)$ . Hence  $\theta$  is not inner.

We next show that  $\theta$  is class-preserving. To that end, we must show that, for each  $h \in H$ , there exists  $t = t(h) \in H$  with  $h^t = h\theta$ . Fix  $h \in H$ , and write

$$h = \prod_{i=1}^4 x_i^{a_i} \cdot z^b \cdot \prod_{j=1}^n y_j^{c_j},$$

as in (7). If  $a_4 = 0$ , then  $h\theta = h$  and  $t(h) = 1$  works. Thus, we may assume that  $a_4 = 1$ , and hence that  $h\theta = hz$ .

**Claim.** *If  $h\theta = hz$ , then either  $h^{x_2} = hz$  or  $h^{x_1 x_3} = hz$ .*

It is clear from the pc-relations that  $x_2$  commutes with every  $y_j$ . This is true also of  $x_1x_3$ . For, if  $j < n - 1$ , then  $y_j^{x_1x_3} = (y_jy_{j+1})^{x_3} = y_jy_{j+1}^2y_{j+2}$ . Using the relations (and a finite induction) one sees that  $y_{j+1}^2y_{j+2} = y_{n-1}^2y_n = y_n^2 = 1$ . It is easy to see that  $y_{n-1}^{x_1x_3} = y_{n-1}$  and that  $y_n^{x_1x_3} = y_n$ .

Next, observe that  $x_2$  commutes with  $x_4$ , while  $x_4^{x_1x_3} = (x_4z)^{x_3} = x_4z$ . Thus, it suffices to show that, if  $h = x_1^{a_1}x_2^{a_2}x_3^{a_3}$  with  $(a_1, a_2, a_3) \in \{0, 1\}^3$ , then either  $h^{x_2} = hz$ , or  $h^{x_1x_3} = h$ . First,

$$h^{x_2} = (x_1^{a_1}x_2^{a_2}x_3^{a_3})^{x_2} = x_1^{a_1}x_2^{a_2}x_3^{a_3}z^{a_1+a_3} = hz^{a_1+a_3}.$$

Hence, if  $a_1 \neq a_3$ , then  $h^{x_2} = hz$ , as required. It remains to show that  $x_1x_3$  commutes with  $h$  whenever  $a_1 = a_3$ . If  $a_1 = a_3 = 0$ , then either  $h = 1$  or  $h = x_2$ ; clearly  $x_1x_3$  commutes with 1, and  $x_2^{x_1x_3} = x_2z^2 = x_2$ . Finally, if  $a_1 = a_3 = 1$ , then either  $h = x_1x_3$  or  $h = x_1x_2x_3$ ; clearly  $x_1x_3$  commutes with itself, and

$$\begin{aligned} (x_1x_2x_3)^{x_1x_3} &= (x_1(x_2z)(x_3y_1))^{x_3} \\ &= (x_1y_1^{-1})(x_2z)zx_3(y_1y_2) \\ &= x_1x_2y_1^{-1}x_3y_1y_2 \\ &= x_1x_2x_3y_2^{-1}y_1^{-1}y_1y_2 = x_1x_2x_3. \end{aligned}$$

This establishes our claim, and completes the proof of Theorem 1.1.  $\square$

### Acknowledgments

The authors would like to thank R. Quinlan for bringing this problem to their attention, and the anonymous referee for some helpful suggestions.

### References

- [Bosma et al. 1997] W. Bosma, J. Cannon, and C. Playoust, “The Magma algebra system, I: The user language”, *J. Symbolic Comput.* **24**:3–4 (1997), 235–265. MR 1484478 Zbl 0898.68039
- [Burnside 1911] W. Burnside, *Theory of groups of finite order*, 2nd ed., Cambridge Univ. Press, New York, 1911. Reprinted Dover, New York, 1955. MR 16,1086c Zbl 0064.25105
- [Burnside 1913] W. Burnside, “On the outer isomorphisms of a group”, *Proc. London Math. Soc.* **S2-11**:1 (1913), 40–42. MR 1577234 JFM 43.0198.03
- [Eick and Leedham-Green 2008] B. Eick and C. Leedham-Green, “On the classification of prime-power groups by coclass”, *Bull. Lond. Math. Soc.* **40**:2 (2008), 274–288. MR 2009b:20030 Zbl 1168.20007
- [Feit and Seitz 1989] W. Feit and G. M. Seitz, “On finite rational groups and related topics”, *Illinois J. Math.* **33**:1 (1989), 103–131. MR 90a:20016 Zbl 0701.20005
- [Heineken 1979] H. Heineken, “Nilpotente Gruppen, deren sämtliche Normalteiler charakteristisch sind”, *Arch. Math. (Basel)* **33**:6 (1979), 497–503. MR 81h:20023 Zbl 0413.20017
- [Hertweck 2001] M. Hertweck, “Class-preserving automorphisms of finite groups”, *J. Algebra* **241**:1 (2001), 1–26. MR 2002e:20047 Zbl 0993.20017



- [Holt et al. 2005] D. F. Holt, B. Eick, and E. A. O'Brien, *Handbook of computational group theory*, Chapman & Hall, Boca Raton, FL, 2005. MR 2006f:20001 Zbl 1091.20001
- [Malinowska 1992] I. Malinowska, "On quasi-inner automorphisms of a finite  $p$ -group", *Publ. Math. Debrecen* **41**:1–2 (1992), 73–77. MR 93g:20069 Zbl 0792.20019
- [O'Brien 1990] E. A. O'Brien, "The  $p$ -group generation algorithm", *J. Symbolic Comput.* **9**:5–6 (1990), 677–698. MR 91j:20050 Zbl 0736.20001
- [Wall 1947] G. E. Wall, "Finite groups with class-preserving outer automorphisms", *J. London Math. Soc.* **22** (1947), 315–320. MR 10,8g Zbl 0030.00901
- [Yadav 2011] M. K. Yadav, "Class preserving automorphisms of finite  $p$ -groups: a survey", pp. 569–579 in *Groups St Andrews 2009* (Bath, 2009), vol. II, edited by C. M. Campbell et al., London Math. Soc. Lecture Note Ser. **388**, Cambridge Univ. Press, 2011. MR 2012j:20061 Zbl 1231.20024

Received: 2012-08-04

Revised: 2012-11-07

Accepted: 2012-11-17

pbrooks@bucknell.edu

*Department of Mathematics, Bucknell University,  
380 Olin Science Building, Lewisburg, PA 17837, United States*

msm030@bucknell.edu

*Department of Mathematics, Bucknell University,  
380 Olin Science Building, Lewisburg, PA 17837, United States*



# The sharp log-Sobolev inequality on a compact interval

Whan Ghang, Zane Martin and Steven Waruhiu

(Communicated by Kenneth S. Berenhaut)

We provide a proof of the sharp log-Sobolev inequality on a compact interval.

## 1. Introduction

The Gaussian log-Sobolev inequality, due to A. J. Stam [1959, Equation 2.3] or Paul Federbush [1969, Equation (14)], although often attributed to L. Gross [1975, Corollary 4.2], played a crucial role in Perelman's proof [2002] of the Poincaré conjecture. We consider log-Sobolev inequalities for finite Lebesgue measure. F. Maggi [Morgan 2009] observed that the sharp log-Sobolev inequality on the interval follows from an isoperimetric conjecture of Díaz et al. [2012], which remains open, but provided no proof. We found it in [Wang 1999], which cited Deuschel and Stroock [1990], who gave a proof of the sharp log-Sobolev inequality on the circle. We then traced this result back to [Émery and Yukich 1987, page 1; Rothaus 1980, Theorem 4.3; Weissler 1980, Theorem 1]. Our Theorem 2.2 shows that the interval case follows quickly from the circle case.

## 2. Log-Sobolev inequality on a compact interval

In considering the isoperimetric problem in sectors of the plane with density  $r^p$ , Díaz et al. [2012, Corollary 4.24, Conjecture 4.18] conjectured the inequality

$$\left[ \int_0^1 r^q d\alpha \right]^{1/q} \leq \int_0^1 \sqrt{r^2 + (q-1) \frac{r'^2}{\pi^2}} d\alpha, \quad (1)$$

where  $1 < q \leq 2$ . F. Maggi [Morgan 2009] observed that (1) implies the log-Sobolev inequality of Theorem 2.2. Here we observe that Theorem 2.2 follows from a proposition of Weissler.

---

*MSC2010:* primary 46; secondary 53.

*Keywords:* isoperimetric inequality, log-Sobolev inequality.

**Proposition 2.1** [Weissler 1980, Theorem 1]. *Let  $f$  be a nonnegative  $C^1$  function on the circle  $S^1$  of length 1. Suppose  $\int_{S^1} f^2 = 1$ . Then we have the sharp inequality*

$$4\pi^2 \int_{S^1} f^2 \log f \leq \int_{S^1} f'^2.$$

Various proofs are discussed in Section 3.

**Theorem 2.2.** *Let  $f$  be a nonnegative  $C^1$  function on the interval  $[0, 1]$ . Suppose  $\int_0^1 f^2 = 1$ . Then we have the inequality*

$$\pi^2 \int_0^1 f^2 \log f \leq \int_0^1 f'^2. \quad (2)$$

*Proof.* Let  $f$  be any nonnegative  $C^1$  function on  $[0, 1]$  such that  $\int_0^1 f^2 = 1$ . Define a nonnegative piecewise  $C^1$  function  $g$  on  $S^1$  such that

$$g(x) = \begin{cases} f(2x) & \text{if } 0 \leq x \leq \frac{1}{2}, \\ f(2-2x) & \text{if } \frac{1}{2} < x \leq 1. \end{cases}$$

Then  $\int_{S^1} g^2 = 1$ . By smoothing, Proposition 2.1 applies to  $g$ . By simple computation, we have that

$$\int_{S^1} g^2 \log g = \int_0^1 f^2 \log f \quad \text{and} \quad \int_{S^1} g'^2 = 4 \int_0^1 f'^2.$$

The conclusion follows. □

**Remark 2.3.** Feng-Yu Wang [1999, Example 1.2] suggested an alternative proof of (2), but we don't understand his proof. He considered densities  $C_\epsilon \exp(\epsilon \cos \pi x)$  and functions  $f_\epsilon = \exp(-\epsilon \cos \pi x)$ , with  $C_\epsilon$  chosen to make the integral of  $f_\epsilon^2$  equal to 1. Then  $f_\epsilon$  satisfies the differential equation

$$f_\epsilon'' - \pi \epsilon \sin \pi x f_\epsilon' = -\pi^2 f_\epsilon \log f_\epsilon. \quad (3)$$

He said that it follows that (2) holds for those functions and densities with sharp constant  $\pi^2$ . This might follow if it were known that functions realizing equality exist, but Wang himself [1999, page 655] admits that “the author is not sure yet whether there always exists [such a function].” Indeed, in the case of the circle with unit density, there apparently is no such function. Of course, the sharp inequality for density 1 would follow as  $\epsilon$  approaches 0.

A similar result holds on the interval  $[a, b]$  for a function with root mean square  $m$ .

**Corollary 2.4.** *Let  $f$  be a nonnegative  $C^1$  function on the interval  $[a, b]$ . Suppose*

$$\frac{1}{b-a} \int_a^b f^2 = m^2, \quad m > 0.$$

Then we have the inequality

$$\frac{\pi^2}{(b-a)^2} \left( \int_a^b f^2 \log f - (b-a)m^2 \log m \right) \leq \int_a^b f'^2. \quad (4)$$

*Proof.* Let  $f$  be a nonnegative  $C^1$  function on the interval  $[a, b]$  such that

$$\frac{1}{b-a} \int_a^b f^2 = m^2 > 0, \quad m > 0.$$

Define a function  $g$  on the interval  $[0, 1]$  as

$$g(x) = \frac{1}{m} f((b-a)x + a).$$

Then  $g$  is nonnegative and  $C^1$ . Moreover, we have

$$\int_0^1 g(x)^2 dx = \int_0^1 \frac{1}{m^2} f((b-a)x + a)^2 dx = \frac{1}{(b-a)m^2} \int_a^b f(y)^2 dy = 1.$$

Therefore, we can apply Theorem 2.2 to the function  $g$ . We have

$$\frac{\pi^2}{b-a} \int_0^1 g^2 \log g \leq (b-a) \int_0^1 g'^2. \quad (5)$$

Note that

$$g'(x) = \frac{b-a}{m} f'((b-a)x + a).$$

By direct calculation, we have

$$\int_0^1 g'(x)^2 dx = \frac{(b-a)^2}{m^2} \int_0^1 f'((b-a)x + a)^2 dx = \frac{(b-a)^2}{m^2} \int_a^b f'(x)^2 dx.$$

We also have

$$\begin{aligned} \int_0^1 g(x)^2 \log g(x) dx &= \frac{1}{m^2} \int_0^1 f((b-a)x + a)^2 \log \frac{f((b-a)x + a)}{m} dx \\ &= \frac{1}{(b-a)m^2} \int_a^b f(x)^2 \log \frac{f(x)}{m} dx \\ &= \frac{1}{(b-a)m^2} \int_a^b f^2 (\log f - \log m) \\ &= \frac{1}{(b-a)m^2} \left( \int_a^b f^2 \log f - (b-a)m^2 \log m \right). \end{aligned}$$

Therefore, by plugging these identities into (5), we have

$$\frac{\pi^2}{(b-a)m^2} \left( \int_a^b f^2 \log f - (b-a)m^2 \log m \right) \leq \frac{b-a}{m^2} \int_a^b f'^2.$$

This is equivalent to the desired inequality (4). □

Corollary 2.4 can be written in the following form.

**Corollary 2.5.** *Let  $f$  be a nonnegative  $C^1$  function on the interval  $[a, b]$ . Suppose*

$$\frac{1}{b-a} \int_a^b f = m > 0.$$

*Then we have the inequality*

$$\frac{2\pi^2}{(b-a)^2} \left( \int_a^b f \log f - m \log m \right) \leq \int_a^b \frac{f'^2}{f}.$$

*Proof.* Define a nonnegative piecewise  $C^1$  function  $g$  on the interval  $[a, b]$  as  $g = \sqrt{f}$ . Plugging  $g$  into Corollary 2.4 yields the desired result. □

**Proposition 2.6.** *In Theorem 2.2,  $\pi^2$  is the best possible constant.*

*Proof.* For any  $0 < \epsilon < 1$ , define

$$f_\epsilon(x) = \sqrt{1 - \epsilon^2} + \sqrt{2}\epsilon \cos \pi x.$$

Then by direct computation, we have

$$\lim_{\epsilon \rightarrow 0^+} \frac{\int_0^1 f_\epsilon'^2}{\int_0^1 f_\epsilon^2 \log f_\epsilon} = \pi^2.$$

Therefore, the constant  $\pi^2$  cannot be replaced by a larger constant. □

**Remark 2.7.** The function  $\cos \pi x$  comes from the equality case of a Wirtinger inequality which follows from the log-Sobolev inequality [Morgan 2009].

### 3. Proofs of the sharp log-Sobolev inequality on the circle

We summarize three proofs of Proposition 2.1 given by Rothaus [1980, Theorem 4.3], Weisler [1980, Theorem 1], Émery and Yukich [1987, page 1], and Deuschel and Stroock [1990, Remark 1.14 (i)].

**3.1. Weisler’s proof.** Weisler proved a stronger result than Proposition 2.1 by Fourier expansion of functions of period  $2\pi$ .

**Proposition 3.1** [Weisler 1980, Theorem 1]. *Let  $f(\theta) = \sum_{n=-\infty}^{\infty} a_n e^{in\theta}$  be in  $L^2$  and suppose  $f(\theta) \geq 0$  almost everywhere. Then*

$$\int f^2 \log f \leq \sum_{n=-\infty}^{\infty} |n| |a_n|^2 + \|f\|_2^2 \log \|f\|_2$$

*in the sense that if the right-hand side is finite, then so is the left-hand side and the inequality holds. ( $0^2 \log 0$  is taken to be 0.)*

Obviously the above inequality is stronger than the inequality

$$\int f^2 \log f \leq \sum_{n=-\infty}^{\infty} |n|^2 |a_n|^2 + \|f\|_2^2 \log \|f\|_2,$$

which is equivalent to Proposition 2.1 by change of variables as in Corollary 2.4.

Weissler [1980] cited [Rothaus 1978] but did not have [Rothaus 1980], where Rothaus gave his proof of Proposition 2.1.

**3.2. Rothaus's proof.** Rothaus proved Proposition 2.1 by a variational method. (References in this section are relative to [Rothaus 1980].) He considered an equivalent problem with a positive parameter  $\rho$  in Section 4. If a related constant  $b_\rho$  is zero, then the log-Sobolev inequality on the circle with the constant  $2/\rho$  holds. For each  $b_\rho$ , he showed in Theorem 4.2 that a minimizing function exists, is positive and satisfies a related differential equation. Moreover, for  $\rho > 1/2\pi^2$ , the only positive solution to the differential equation is the constant function 1 (Theorem 4.3) and hence  $b_\rho$  is zero. Therefore in the limit  $b_{1/2\pi^2}$  is zero, and our Proposition 2.1 follows.

Rothaus cited [Weissler 1980], saying that “a result related to Theorem 6.3 appears in” that paper.

**3.3. Émery and Yukich's proof.** Proposition 2.1 was proved by Émery and Yukich [1987, page 1] by using estimates deploying the Brownian motion semigroup.

Émery and Yukich [1987] cited both Weissler [1980] and Rothaus [1980].

**3.4. Deuschel and Stroock's proof.** Deuschel and Stroock considered the log-Sobolev inequality in general spaces with densities. As a special case, they proved [Deuschel and Stroock 1990, Remark 1.14 (i)] that the log-Sobolev constant for the circle of length 1 with Lebesgue measure is the first eigenvalue of the Laplacian, namely  $4\pi^2$  (corresponding to the first eigenfunction  $\sin 2\pi x$ ).

Deuschel and Stroock [1990] cited [Émery and Yukich 1987].

### Acknowledgements

We thank our advisor Frank Morgan for his patience and invaluable input. We thank our friend Andrew Kelly for substantial contributions to this paper. We thank the National Science Foundation for grants to Professor Morgan and the Williams College SMALL Research Experience for Undergraduates, and Williams College for additional funding. We thank the Mathematical Association of America, MIT, the University of Chicago, and Williams College for supporting our trip to talk at MathFest 2012.

The authors were undergraduate students at MIT (Ghang), Williams College (Martin) and University of Chicago (Waruhiu) at the time this work was carried out.

## References

- [Deuschel and Stroock 1990] J.-D. Deuschel and D. W. Stroock, “Hypercontractivity and spectral gap of symmetric diffusions with applications to the stochastic Ising models”, *J. Funct. Anal.* **92**:1 (1990), 30–48. MR 91j:58174 Zbl 0705.60066
- [Díaz et al. 2012] A. Díaz, N. Harman, S. Howe, and D. Thompson, “Isoperimetric problems in sectors with density”, *Adv. Geom.* **14**:4 (2012), 589–619.
- [Émery and Yukich 1987] M. Émery and J. E. Yukich, “A simple proof of the logarithmic Sobolev inequality on the circle”, pp. 173–175 in *Séminaire de Probabilités, XXI*, edited by J. Azéma et al., Lecture Notes in Math. **1247**, Springer, Berlin, 1987. MR 89f:26020 Zbl 0616.46023
- [Federbush 1969] P. Federbush, “Partially alternate derivation of a result of Nelson”, *J. Math. Phys.* **10**:1 (1969), 50–52.
- [Gross 1975] L. Gross, “Logarithmic Sobolev inequalities”, *Amer. J. Math.* **97**:4 (1975), 1061–1083. MR 54 #8263 Zbl 0318.46049
- [Morgan 2009] F. Morgan, “Log-Sobolev inequality”, blog entry, 2009, <http://sites.williams.edu/Morgan/2009/06/11/sobolev-type-inequality/>.
- [Perelman 2002] G. Perelman, “The entropy formula for the Ricci flow and its geometric applications”, preprint, 2002. Zbl 1130.53001 arXiv math/0211159
- [Rothaus 1978] O. S. Rothaus, “Lower bounds for eigenvalues of regular Sturm–Liouville operators and the logarithmic Sobolev inequality”, *Duke Math. J.* **45**:2 (1978), 351–362. MR 58 #1368 Zbl 0435.47049
- [Rothaus 1980] O. S. Rothaus, “Logarithmic Sobolev inequalities and the spectrum of Sturm–Liouville operators”, *J. Funct. Anal.* **39**:1 (1980), 42–56. MR 81m:34025 Zbl 0472.47024
- [Stam 1959] A. J. Stam, “Some inequalities satisfied by the quantities of information of Fisher and Shannon”, *Information and Control* **2** (1959), 101–112. MR 21 #7813 Zbl 0085.34701
- [Wang 1999] F.-Y. Wang, “Harnack inequalities for log-Sobolev functions and estimates of log-Sobolev constants”, *Ann. Probab.* **27**:2 (1999), 653–663. MR 2000i:58067 Zbl 0948.58023
- [Weissler 1980] F. B. Weissler, “Logarithmic Sobolev inequalities and hypercontractive estimates on the circle”, *J. Funct. Anal.* **37**:2 (1980), 218–234. MR 81k:42007 Zbl 0463.46024

Received: 2012-09-15      Accepted: 2013-01-06

ghangh@math.harvard.edu      *Department of Mathematics, Harvard University,  
26 Everett Street #314, Cambridge, MA 02138, United States*

zkm1@williams.edu      *Department of Mathematics and Statistics, Williams College,  
Bronfman Science Center, 18 Hoxsey Street,  
Williamstown, MA 01267, United States*

waruhius@uchicago.edu      *Department of Mathematics, University of Chicago, 5734  
South University Avenue, Chicago, IL 60637, United States*



# Analysis of a Sudoku variation using partially ordered sets and equivalence relations

Ana Burgers, Shelly Smith and Katherine Varga

(Communicated by Ann Trenk)

Sudoku is a popular game of logic, and there are many variations of the standard puzzle. We investigate a variation of Sudoku that uses inequalities between cells rather than numerical clues. We begin with an overview of the rules and strategies of the game. We then examine the solvability of an individual  $m \times n$  block with the use of partially ordered sets, and combine  $2 \times 2$  blocks to form  $4 \times 4$  puzzles.

## 1. Introduction

The basic concepts behind the popular Sudoku number puzzles may be familiar from the newspaper, the internet, or any variety of puzzle books. A *Sudoku board* is a  $9 \times 9$  grid in which the entries 1 through 9 appear exactly once in each row, column, and  $3 \times 3$  block. A *Sudoku puzzle* is created from a board by strategically removing some of the entries, leaving only select clues from which the player must try to reconstruct the original board. In order to be a valid puzzle, the clues must lead to a unique solution. In this paper, we will refer to this game as *standard Sudoku* (see Figure 1, left).

One variation on the basic puzzle is *Greater Than Sudoku*. A *Greater Than Sudoku board* (Figure 1, right) meets the same criteria as the standard board, but has an additional condition: within each block, every pair of adjacent entries, both horizontal and vertical, must satisfy the inequality which separates them. While the standard puzzle begins with some entries filled in, providing the player with numerical clues which will lead to a unique solution, a Greater Than Sudoku puzzle gives the player only the inequalities on an empty grid. Furthermore, the inequalities must be arranged in such a way that a unique solution exists.

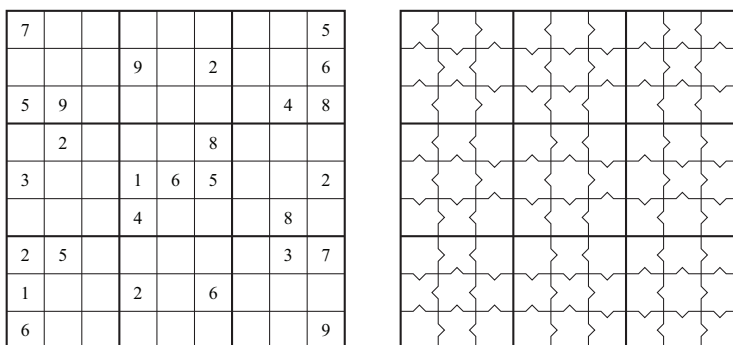
The primary focus of this paper is a smaller version called *Greater Than Shidoku* (Figure 2, left) consisting of a  $4 \times 4$  grid partitioned into four  $2 \times 2$  blocks, which is played with the entries 1 through 4. Many results are also extended to blocks of

---

MSC2010: 06A06, 20B30, 91A46.

Keywords: Sudoku, partial order, total order, equivalence relation.

This research was supported by the National Science Foundation under grant DMS-0451254.



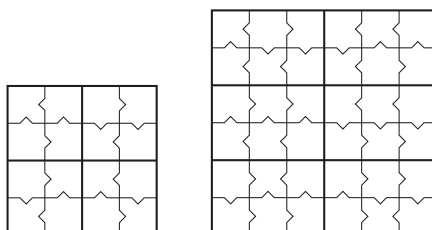
**Figure 1.** Left: standard Sudoku puzzle [Mepham 2011]; right: Greater Than Sudoku puzzle [Sudoku 2006].

larger variations, including *Greater Than Rokudoku* (Figure 2, right), which has six  $2 \times 3$  blocks, and Greater Than Sudoku.

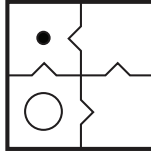
## 2. Playing the game

Solving a Greater Than puzzle of any size requires a slightly different approach than that used to play standard Sudoku, and this approach will also prove to be instrumental in the analysis of Greater Than puzzles of any size. In particular, the player identifies the minimal and maximal cells as well as using the conditions placed on rows, columns, and blocks. A *minimal* cell of a block is any unfilled cell whose inequalities all point inward from adjacent unfilled cells. Similarly, a *maximal* cell is any unfilled cell with all inequalities pointing outward into adjacent unfilled cells. Since these properties depend upon the cells that have not yet been filled, the maximal and minimal cells will change as the game is played. In Figure 3, the unfilled Greater Than Shidoku block contains one minimal cell, identified by ●, and one maximal cell, identified by ○.

Our first step in solving the Greater Than Shidoku puzzle in Figure 2 is to identify where to place the 1 entries. (The solutions to the other puzzles are at the end of



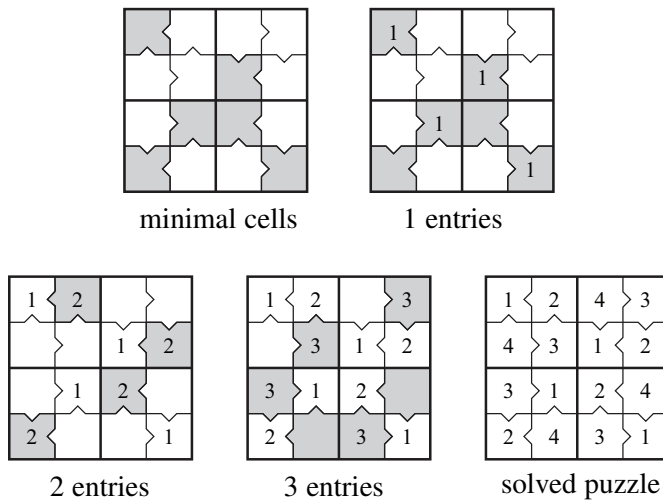
**Figure 2.** Smaller puzzles. Left: Greater Than Shidoku; right: Greater Than Rokudoku.



**Figure 3.** Minimal and maximal cells.

this article.) Since 1 is the smallest element we use, each 1 must be placed in a minimal cell in an unfilled block. For example, in Figure 4, the top two blocks each contain only one (shaded) minimal cell, so we know those cells must contain 1. The bottom two blocks, however, each contain two cells that are minimal. In such cases where the inequalities do not determine unique placement of each 1 entry, the next step is to consider any information provided by the rows and columns. Thus, while there are two possible placements of 1 entries in each block of the lower half of the board, by using the columns it is possible to uniquely determine their proper placement.

Next we will determine where to place the 2 entries by considering the minimal cells among those that remain unfilled. If necessary, the rows and columns may again be used to determine the correct placements. Similarly, a 3 entry must have inequalities pointing inward from each adjacent cell not containing a 1 or a 2, and so on. The player may also begin with the largest entry and work backwards by looking for maximal cells. A 4 must be placed in a maximal cell, where the inequalities all point outward. A 3 would have inequalities pointing out into any cell not containing a 4, and so on.



**Figure 4.** Playing Greater Than Shidoku.

### 3. Inequality blocks and cycles

A Greater Than puzzle contains inequalities that compare adjacent entries; however, only entries within the same block are considered. The player must begin by examining the ways in which individual blocks can be filled before moving on to the puzzle as a whole. Similarly, we begin our investigation of Greater Than puzzles by considering individual blocks.

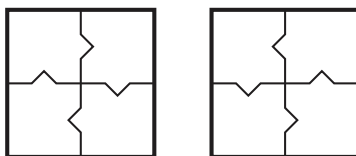
**Definition 1.** An *inequality block* is an  $m \times n$  grid, with  $m, n \in \mathbb{N}$ , in which an inequality separates each pair of horizontally or vertically adjacent cells.

In one block of Greater Than Shidoku there are four inequalities, and each can be oriented in one of two directions. Thus there are  $2^4 = 16$  possible  $2 \times 2$  inequality blocks. There are four cells in each block, so without considering the inequalities there are  $4! = 24$  ways of permuting the entries. Similarly, we can count the number of inequality blocks and permutations of any size block. Greater Than Rokudoku blocks have  $2^7 = 128$  ways of arranging the inequalities and  $6! = 720$  permutations of entries. For Greater Than Sudoku, we have  $2^{12} = 4096$  inequality blocks and  $9! = 362,880$  ways of permuting the entries.

**Definition 2.** An inequality block is *solvable* if there exists at least one permutation of entries satisfying all inequalities in that block. A block is *unsolvable* if no such permutation exists.

Note that for each size block, there are many more ways to permute the entries than there are ways to arrange the inequalities. Each permutation of entries corresponds to one arrangement of the inequalities because, given any filled block, we can insert the inequalities accordingly. However, since there are significantly fewer inequality arrangements than permutations, some inequality arrangements must correspond to more than one permutation. In other words, without considering the other blocks in a puzzle, many inequality blocks have more than one solution. This leads to two natural questions: are all inequality blocks solvable, and for those that are, how many solutions exist? We can only use solvable blocks to create Greater Than puzzles, so our first goal is to determine criteria for deciding which blocks are solvable.

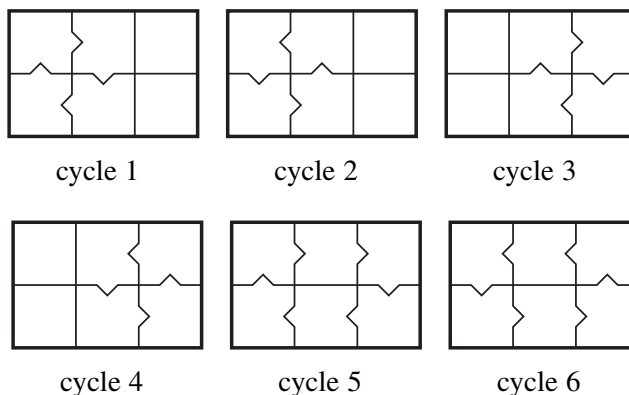
A *path* in an inequality block is a sequence of adjacent cells where the inequalities are always increasing or always decreasing. If a path includes any cell more than once, that path contains a *cycle*. A cycle of cells is impossible to fill with entries without contradicting at least one of the inequalities; thus any inequality block containing a cycle is unsolvable. In a  $2 \times 2$  inequality block, there are two inequality arrangements that produce a cycle of the four cells, shown in Figure 5. These two cycles correspond to two unsolvable inequality blocks, leaving us with 14 that are acyclic.



**Figure 5.** Unsolvable  $2 \times 2$  blocks.

In Greater Than Rokudoku, the  $2 \times 3$  inequality blocks may also contain cycles, but more than two unsolvable blocks result from such arrangements. There are six different cycles that may appear in a  $2 \times 3$  block, shown in Figure 6.

A block that contains a cycle is unsolvable, but some inequality arrangements may contain more than one cycle, so to count the number of blocks with cycles, we use the principle of inclusion-exclusion. Let  $C_i$  be the set of all blocks containing cycle  $i$  for  $1 \leq i \leq 6$ . Blocks from sets  $C_1$  through  $C_4$  each have three inequalities that are not involved in the given cycle, so  $|C_i| = 2^3$  for  $1 \leq i \leq 4$ . Sets  $C_5$  and  $C_6$  consist of blocks with only one inequality not involved in the cycle, thus  $|C_5| = |C_6| = 2$ , and consequently  $\sum_{i=1}^6 |C_i| = 36$ . However, some blocks will be counted in two sets; for example, if the remaining inequality in a block from set  $C_5$  is pointing down, that block also contains cycle 1, thus that block is included in set  $C_1$ . If the inequality is pointing up, that block is included in set  $C_3$ . Similarly, one block in  $C_6$  is also contained in set  $C_2$ , while the other is contained in set  $C_4$ . There is one block containing both cycles 1 and 4, and another containing cycles 2 and 3. There are no  $2 \times 3$  blocks that contain 3 different cycles. Thus we have double counted 6 blocks that are in two sets, and so we subtract this from our previous tally, resulting in a total of 30  $2 \times 3$  inequality blocks with at least one cycle. These 30 blocks are unsolvable, so we eliminate them from the number of inequality blocks we need to consider. This leaves us with 98 acyclic  $2 \times 3$  inequality blocks.



**Figure 6.** Cycles in  $2 \times 3$  blocks.

We employ the same strategy with  $3 \times 3$  Sudoku blocks, but now we have twenty-six possible cycles that can be formed among the inequalities (see if you can find them all!). Many inequality arrangements contain multiple cycles. To count the number of blocks that contain at least one cycle, we again use inclusion-exclusion. There are 1698 such puzzle blocks that are impossible to fill in, so of the 4096  $3 \times 3$  inequality blocks, 2398 are acyclic. The results of this section are summarized in the following theorem.

**Theorem 1.** *There are 14 acyclic  $2 \times 2$  inequality blocks, 98 acyclic  $2 \times 3$  inequality blocks, and 2398 acyclic  $3 \times 3$  inequality blocks.*

#### 4. Posets and solvable blocks

We have shown that every inequality block containing a cycle is unsolvable; however, it remains to be seen that every acyclic inequality block is solvable. While playing the game, we compared cells using the inequalities and identified minimal cells, but we found that minimal cells were not always unique. This suggests considering an acyclic block as a *partially ordered set* and leads us to another way of describing solutions of the block.

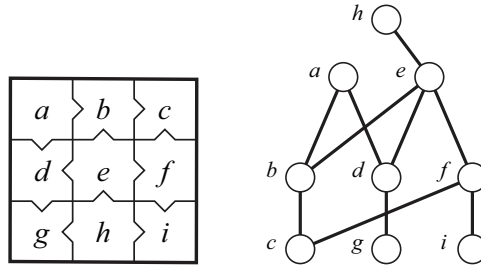
**Definition 3.** A *partial order*  $\preceq$  on a set  $A$  is a binary relation that is reflexive, antisymmetric, and transitive. A *partially ordered set*, or *poset*, is a pair  $(A, \preceq)$ , where  $\preceq$  is a partial order on the set  $A$ .

We now define a relation on inequality blocks of arbitrary size and show that it satisfies the above definition:

**Definition 4.** Let  $A = \{a_1, a_2, \dots, a_{mn}\}$  be the set of cells of an  $m \times n$  acyclic inequality block. For all  $a_i, a_j \in A$ , we define a relation  $\preceq$  on  $A$  such that  $a_i \preceq a_j$  if  $a_i = a_j$  or if  $a_i$  precedes  $a_j$  in an increasing path.

**Theorem 2.** *With  $A$  and  $\preceq$  as defined above,  $(A, \preceq)$  is a partially ordered set.*

*Proof.* Let  $a_i \in A$ . Since  $a_i = a_i$ , then  $a_i \preceq a_i$  and consequently  $\preceq$  is reflexive. Now let  $a_i, a_j \in A$ , where  $a_i \preceq a_j$  and  $a_j \preceq a_i$ , and assume that  $a_i \neq a_j$ . Then  $a_i$  precedes  $a_j$  in an increasing path, and  $a_j$  precedes  $a_i$  in an increasing path. The concatenation of these two increasing paths will contain a cycle, which contradicts our assumption that the block is acyclic. Thus  $a_i = a_j$ , and  $\preceq$  is antisymmetric. Finally, let  $a_i, a_j, a_k \in A$ , where  $a_i \preceq a_j$  and  $a_j \preceq a_k$ . If  $a_i = a_j$  or  $a_j = a_k$ , it is clear that  $a_i \preceq a_k$ , so let us consider the case where  $a_i \neq a_j$  and  $a_j \neq a_k$ . This means that  $a_i$  precedes  $a_j$  in an increasing path, and  $a_j$  precedes  $a_k$  in an increasing path. The concatenation of these paths forms an increasing path in which  $a_i$  precedes  $a_k$ . Thus  $a_i \preceq a_k$ , and  $\preceq$  is transitive. Therefore,  $\preceq$  is a partial order on  $A$ , and  $(A, \preceq)$  is a poset.  $\square$



**Figure 7.** Inequality block (left) and Hasse diagram (right) of poset of cells.

We may visualize a poset by creating a *Hasse diagram* of the set. In a Hasse diagram, the vertices represent the elements of the set. If  $x \leq y$ , then the vertex for  $x$  is placed below the vertex for  $y$ . If  $x \neq y$ ,  $x \leq y$ , and there is no intermediate element  $z \neq x, y$  such that  $x \leq z \leq y$ , then we say that  $y$  *covers*  $x$ , and an edge is drawn connecting the two elements. However, if there is such an element  $z$ , an edge from  $x$  to  $z$ , and one from  $z$  to  $y$ , then  $x \leq y$  by transitivity. Figure 7 shows an acyclic  $3 \times 3$  inequality block as well as the corresponding Hasse diagram.

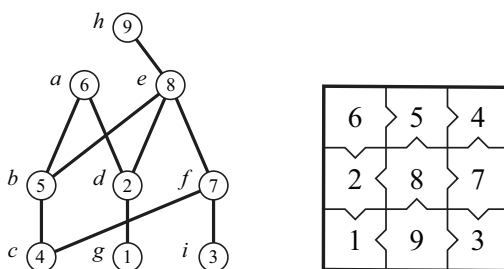
This type of relation is called a *partial* order because it may not be possible to use the relation to compare all of the elements in the set.

**Definition 5.** Let  $\leq$  be a partial order on a set  $A$ . Elements  $a$  and  $b$  are called *comparable* if and only if either  $a \leq b$  or  $b \leq a$ . Otherwise,  $a$  and  $b$  are *incomparable*.

Even though cells  $c$  and  $h$  are not adjacent in the above inequality block, there is an increasing path  $c, b, e, h$ ; therefore  $c \leq h$  and we know any solution for this block must have a smaller element in cell  $c$  than in cell  $h$ . On the other hand, there is no such increasing path between  $b$  and  $g$ , so those two cells are incomparable and we cannot predict which cell will contain the larger entry. A useful fact about posets is that any finite, nonempty poset has a minimal element, and furthermore, any subset of a poset is also a poset [Epp 2004]. This means that if we remove a minimal element from a poset, we will always have at least one minimal element among the remaining cells.

This brings us back to our technique for solving the Greater Than Shidoku puzzle by identifying minimal cells in each block. When we placed the 1 entries, we effectively removed those cells from the posets for each block, then we identified the minimal cells in the resulting posets in order to place the 2 entries. Previously proven results about posets give us another way to view our solution to the puzzle.

**Definition 6.** If  $\leq$  is a relation on a set  $A$ , and for any two elements  $a$  and  $b$  in  $A$  either  $a \leq b$  or  $b \leq a$ , then  $\leq$  is a *total order* on  $A$ . A *linear extension* is obtained by putting a total order on a poset  $(A, \leq)$  which preserves the partial order  $\leq$ .



**Figure 8.** Creating a linear extension to solve a block.

**Theorem 3.** *Every partial order may be extended to a total order.*

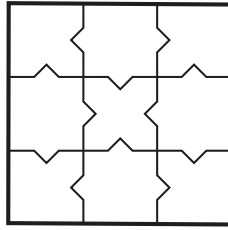
This theorem was first proven by Szpilrajn in [1930], and from it we may conclude that we can put a total order on the poset of cells of any acyclic inequality block. Furthermore, creating a linear extension of the poset of cells is the same as finding a solution for the block, which leads us directly to the following corollary.

**Corollary 4.** *Every acyclic  $m \times n$  inequality block is solvable.*

To demonstrate, we will create a linear extension of our poset in Figure 8. First, we pick a minimal cell  $g$  and label it with 1. Once  $g$  has been labeled and thus removed from future comparisons, our new set of minimal cells consists of  $c$ ,  $d$ , and  $i$ . We arbitrarily choose  $d$  and label it 2. The minimal cells are now  $c$  and  $i$ ; we choose cell  $i$  and label it 3. We continue to choose and label minimal elements until all are labeled. Figure 8 shows one example of how the remaining entries can be labeled, and the corresponding solution of the inequality block. However, at each step in the process, there were often multiple minimal cells to choose from, so the solution in the figure is only one of the many solutions we could have chosen.

Now that we can find a solution of any acyclic inequality block, the next step is to find a method of counting the number of solutions of any such block. This is essential because some blocks have a large number of solutions, so it is often tedious to attack this task by hand. Many researchers have studied the question of creating and counting linear extensions of posets; we used *A Maple Package for Posets*, created by John R. Stembridge [2009] of the University of Michigan. This package includes a command to count the number of linear extensions of any given poset, and when we applied it to the  $3 \times 3$  block in Figure 7, we found that there are actually 261 solutions to the block. Surprising as this might seem, it is far from the highest number of solutions of a  $3 \times 3$  block. After testing all possible inequality combinations on a block, we find that there are 34  $3 \times 3$  blocks that have over 1000 solutions. In fact, the block in Figure 9 has 4800 solutions!





**Figure 9.** Block with 4800 solutions.

### 5. Equivalent Shidoku blocks

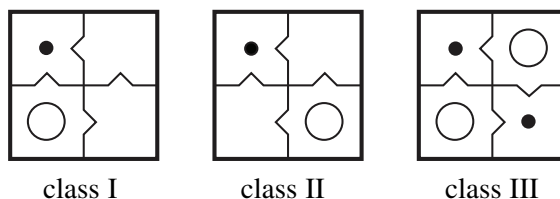
There are 14 acyclic, and therefore solvable, inequality blocks that we can use to create a Greater Than Shidoku puzzle. Each  $4 \times 4$  grid is comprised of four blocks, which means that there are  $14^4 = 38,416$  possible combinations of inequality blocks to choose from, although we shall see that the number of Greater Than Shidoku boards and puzzles is considerably smaller. For Greater Than Rokudoku and Sudoku, the number of possible combinations are  $98^6 \approx 8.9 \times 10^{11}$  and  $2398^9 \approx 2.6 \times 10^{30}$ , respectively. For this reason, we are interested in grouping similar blocks together, thereby reducing the number of possible combinations to a more manageable size; thus we will define a method of grouping inequality blocks based on the positions of minimal and maximal cells in an unfilled block.

In Greater Than Shidoku, each block has four entries, and we want to find all combinations of maximal and minimal cells. Recall that every inequality block must have at least one maximal and at least one minimal cell. Further note that, given any two adjacent cells, the entry in one must be larger than that of the other, thus it is not possible to have two adjacent minimal cells nor two adjacent maximal cells. Finally, we recognize that if we have two maximal cells diagonal from each other, the inequalities of the remaining two cells are determined, and those cells are forced to be minimal. Consequently there are two cases for the number of minimal and maximal cells: we may either have one minimal and one maximal cell, or we may have two of each. To take all the different arrangements into account, we define the following relation.

**Definition 7.** Let  $S_{2,2}$  be the set of all solvable  $2 \times 2$  inequality blocks. We define a relation  $\sim$  as follows. Let  $A, B \in S_{2,2}$ . Then  $A \sim B$  if and only if  $A$  can be transformed into  $B$  using some sequence of reflections across the vertical, horizontal, and diagonal axes of the block.

**Theorem 5.** *The relation  $\sim$  is an equivalence relation on  $S_{2,2}$ .*

*Proof.* To prove  $\sim$  is an equivalence relation, we must show that it is reflexive, symmetric, and transitive. Let  $A \in S_{2,2}$ . Clearly  $A \sim A$  because no transformation



**Figure 10.** Representatives of the three equivalence classes.

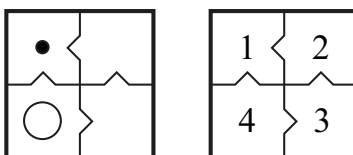
is necessary, and thus  $\sim$  is reflexive. Now let  $A, B \in S_{2,2}$  such that  $A \sim B$ . Then  $A$  can be transformed into  $B$  by some sequence of reflections. Applying these reflections to  $B$  in reverse order transforms  $B$  into  $A$ . Thus  $B \sim A$ , and  $\sim$  is symmetric. Finally, let  $A, B, C \in S_{2,2}$ , with  $A \sim B$  and  $B \sim C$ . Then there is a sequence of reflections that will transform  $A$  into  $B$ , and another sequence which will transform  $B$  into  $C$ . The concatenation of these sequences yields a sequence of reflections that will transform  $A$  into  $C$ . Thus the relation  $\sim$  is transitive, and  $\sim$  is an equivalence relation.  $\square$

Using equivalence relation  $\sim$ , the set of solvable blocks can be partitioned into equivalence classes. Blocks within a class are *equivalent*, and those from different classes are said to be *distinct*.

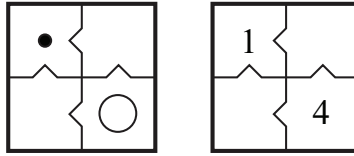
**Theorem 6.** *There are three equivalence classes of  $2 \times 2$  inequality blocks.*

This is easily verified by checking the 14 blocks in  $S_{2,2}$ . An example from each class is shown in Figure 10. We next consider the number of solutions of each block. For the following section, we will also find it helpful to observe that within an equivalence class, the same entry is always placed diagonally from 1 in the block.

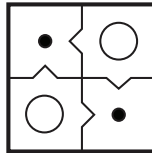
Consider a block from class I in Figure 11. When solving the block, we see there is only one possible position for the 1 entry, and similarly, only one way to place the 4 entry. The 2 and 3 entries are adjacent to one another, and so the 3 must go in the greater of these two cells. There is only one way to fill in this block, and reflections do not change the number of solutions. In fact, blocks of class I correspond to all blocks with unique solutions. We further note that in each block, entries 1 and 3 will be placed diagonally in the block.



**Figure 11.** Class I blocks have 1 solution.



**Figure 12.** Class II blocks have 2 solutions.



**Figure 13.** Class III blocks have 4 solutions.

Following the same procedure with blocks from class II, we can see in Figure 12 that the 1 and 4 entries are uniquely placed. However, examining the remaining two cells, we see that the entries in each must be greater than the 1 entry and less than the 4 entry. It is not possible from this arrangement to uniquely determine placement of the remaining two entries. Thus, blocks from class II correspond to blocks with two solutions, and in each solution 1 and 4 will be placed diagonally.

In class III blocks, however, we have two possible positions for the 1 entry. Similarly, there are two possible placements of the 4 entry. Once 1 and 4 are placed in the cells, there is only one way to place 2 and 3. Each puzzle block from this class has four solutions as shown in Figure 13; entries 1 and 2 will be in the minimal cells, which are placed diagonally.

## 6. Greater Than Shidoku puzzles

Now that we have a better understanding of the different types of inequality blocks, we are able to examine ways in which they can be combined to form puzzles. Recall that a Greater Than board is an  $mn \times mn$  grid, where  $m, n \in \mathbb{N}$ , in which the numbers 1 through  $mn$  must satisfy the inequalities between adjacent cells and appear exactly once in each row, column, and  $m \times n$  block. If when the numerical entries are removed there is a unique solution to the board, the unfilled board is a *Greater Than puzzle*.

It is important to note that, by definition, *every* Greater Than board is solvable when the entries are removed. It is not necessarily the case, however, that each board has a unique solution and is therefore a puzzle. In this section, we will first find a way to create Greater Than Shidoku boards, then determine whether the unfilled boards are puzzles. Previously, we saw that each of the three equivalence

$a$	$b$		
$c$	$d$		

**Figure 14.** Initial block.

classes of blocks may be identified by the entry that is diagonal from 1 when a block is solved. This is a useful tool in proving Theorem 7, which states a rule for combining blocks to create boards. Although the proof begins by considering only entries without inequalities, a Greater Than board can be formed from the standard board by inserting the appropriate inequalities between adjacent cells within each block.

**Theorem 7.** *Every block of a Greater Than Shidoku board must be horizontally or vertically adjacent to another block from the same equivalence class.*

*Proof.* Assume, to the contrary, that a block need not be adjacent to another block from the same equivalence class. Without loss of generality, consider the filled block in Figure 14.

To complete the top row, we place  $c$  and  $d$  in one of two ways. Once these are placed, we then position  $a$  and  $b$  in the second row to ensure that the top two blocks are from different classes. Although we don't know which cell will contain the 1 entry, it is sufficient to ensure that the blocks do not contain any common diagonal. We use similar logic on the first two columns to fill in the bottom-left block in one of two ways, again ensuring that it is not equivalent to the first block. This gives us the four cases shown in Figure 15. In each case we attempt to complete the board by filling in the last block. There is only one cell where we can place the  $a$  entry, but then we find that we are unable to place the  $d$  entry without violating the condition that an entry may only appear once in each row and column, leading

$a$	$b$	$c$	$d$
$c$	$d$	$b$	$a$
$b$	$c$	$a$	
$d$	$a$		

Case 1

$a$	$b$	$c$	$d$
$c$	$d$	$b$	$a$
$d$	$a$		
$b$	$c$	$a$	

Case 2

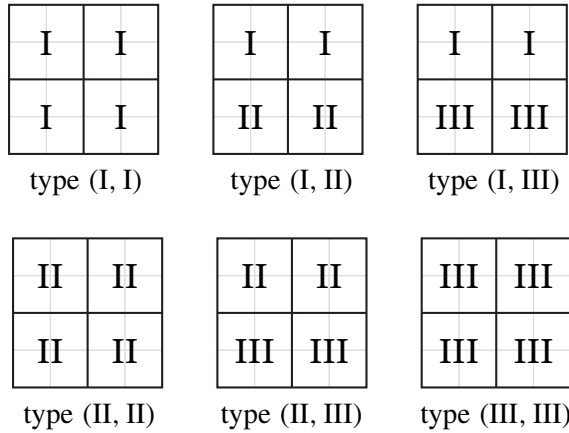
$a$	$b$	$d$	$c$
$c$	$d$	$a$	$b$
$b$	$c$		$a$
$d$	$a$		

Case 3

$a$	$b$	$d$	$c$
$c$	$d$	$a$	$b$
$d$	$a$		
$b$	$c$		$a$

Case 4

**Figure 15.** Each case leads to a contradiction.



**Figure 16.** Six types of Greater Than Shidoku boards.

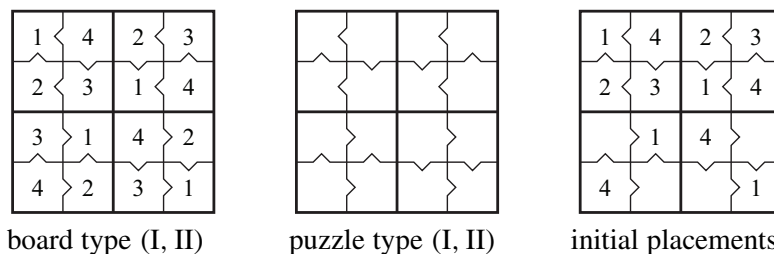
to our desired contradiction. Thus, every block must be adjacent to at least one equivalent block to form a Greater Than Shidoku board.  $\square$

**Corollary 8.** *There are six types of Greater Than Shidoku boards.*

This corollary follows directly from counting the possible combinations of our three equivalence classes, shown in Figure 16. A board of type (I, II), for example, is comprised of two blocks from class I and two from class II. Note that boards comprised of two different block classes may be written in four different ways, taking rotations of  $90^\circ$ ,  $180^\circ$ , and  $270^\circ$  into consideration. Each of these types may be used to form Greater Than Shidoku boards, so our next goal is to determine which of these boards have unique solutions when the entries are removed, and are therefore Greater Than Shidoku puzzles. In the following lemmas, we will see that 4 of these 6 board types correspond to puzzles, and we will count the number of puzzles of each type.

**Lemma 9.** *Every board of type (I, I) has a unique solution when the entries are removed, and therefore corresponds to a Greater Than Shidoku puzzle. There are 32 puzzles of type (I, I).*

*Proof.* Consider any board of type (I, I) and remove all entries, leaving only inequalities. This board consists of Greater Than blocks from class I, and each of these blocks has a unique solution, so there is only one way to fill in entries on the entire board. Thus every board of type (I, I) corresponds to a puzzle. To create a board, there are eight ways to order the first block, since there are four cells in which to place the 1 entry, two ways to place 4 adjacent to 1, and then the cells containing 2 and 3 are uniquely determined. The second and third blocks can each be arranged in two ways, similar to the argument in the proof of Theorem 7, and the fourth block is uniquely determined by the first three. We then place the



**Figure 17.** Type (I, II) example.

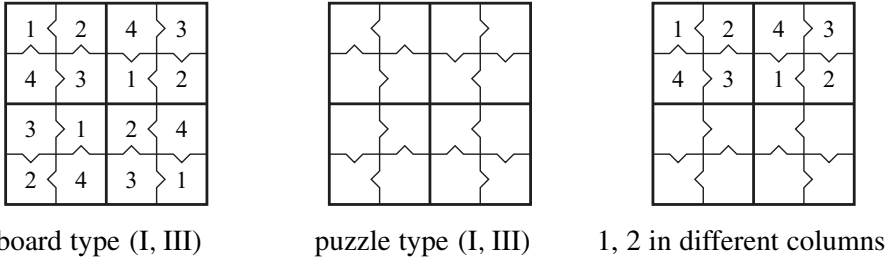
appropriate inequalities to finish the board. Consequently, there are  $(8)(2)(2) = 32$  puzzles of type (I, I).  $\square$

**Lemma 10.** *Each of the 64 type (I, II) boards corresponds to a puzzle.*

*Proof.* Consider a board of type (I, II) such as that in Figure 17, and remove the entries. Without loss of generality, suppose the top two blocks are from class I and the lower two from class II. Since blocks from class I can only be filled in one way and blocks from class II have uniquely determined 1 and 4 entries, the only entries not uniquely determined by inequalities are the 2 and 3 entries on the blocks from class II. However, these entries are placed diagonally in their block, so each column has only one unfilled cell. Thus, by standard Shidoku rules, there is only one possible entry that can be placed in each unfilled cell, leading to a unique solution for the unfilled board. To count these puzzles, we will start by counting boards with blocks placed as in Figure 17. In the top-left block, there are four ways to place the 1 entry, then the 3 entry must be diagonal from 1. There are two ways to place the remaining 2 and 4 entries. In both the top-right block (class I) and the class II block on the bottom-left, we have two choices for placing 1 so that it isn't in the same row or column as the 1 in the first block. Once those choices are made, the placement of the other entries in those blocks is uniquely determined. All of the entries in the last block are uniquely determined, and once again we finish by writing in the inequalities. Thus there are  $(4)(2)(2) = 16$  puzzles of type (I, II) in the form described, however, since each of these puzzles may be rotated  $90^\circ$ ,  $180^\circ$ , or  $270^\circ$  to create new puzzles, there are 64 puzzles of this type.  $\square$

**Lemma 11.** *There are 64 type (I, III) puzzles.*

*Proof.* As in the previous case, we will consider a board of type (I, III) such as that in Figure 18 with class I blocks on top and class III below. Again, the class I blocks have a unique solution. The blocks from class III all have the entry 2 placed diagonally from 1, with 3 and 4 on the other diagonal. Since entries 1 and 2 must be in different columns in the class III blocks, the two blocks from class I must be oriented so entries 1 and 2 are also in different columns to avoid contradiction with

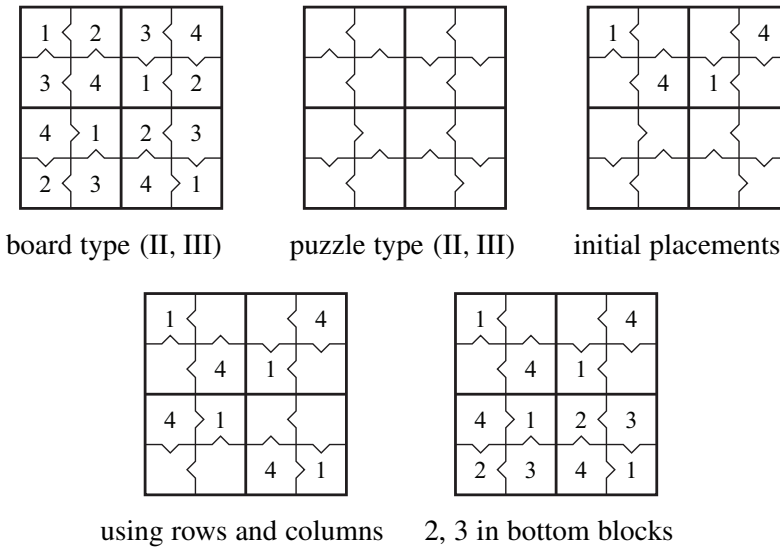


**Figure 18.** Type (I, III) example.

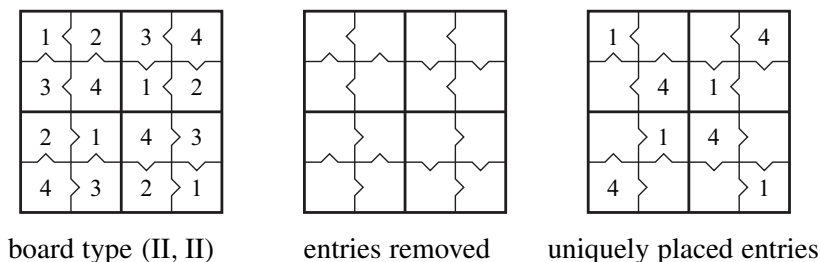
the class III blocks. Furthermore, we recall that each column in a class III block contains both a maximal and minimal cell. Each column of the board contains either a 1 or a 2 in the top two blocks; the other is placed in the minimal cell in that column. Similarly, each column already contains either a 3 or a 4; the other must go in the remaining cell, which is a maximal cell. Thus each cell is filled uniquely, and the board corresponds to a puzzle of type (I, III). We count the puzzles as in the previous lemma: four ways to fill the first block, two ways to fill blocks to right and below, then one way to complete the last block. Including rotations, there are 64 type (I,III) puzzles. □

**Lemma 12.** *There are 64 type (II, III) puzzles.*

*Proof.* Suppose our board has class II blocks on top and class III blocks below, such as in Figure 19. The entries are uniquely placed in the top blocks. As argued in the proof of Lemma 11, there is a maximal and minimal cell in each column of the



**Figure 19.** Type (II, III) example.



**Figure 20.** Type (II, II) example.

bottom blocks. Since the 1 entries have already been placed in two of the columns of the puzzle, the minimal entries in each of the remaining columns on the lower band must contain 1 entries. Similarly, the 4 entries have already been placed in two columns; the maximal entries in each of the remaining columns on the lower band must contain 4 entries as well. There are now two remaining cells in each bottom block. These cells are adjacent, and thus the 2 entry is placed in the lesser of the cells, while the 3 is placed in the greater of the two. So the bottom blocks are uniquely filled in. Now, returning to the top blocks we see that there is only one remaining unfilled cell in each column. Therefore, there is only one possible entry for each cell, which completes the unique solution, so the type (II, III) board corresponds to a puzzle.

There are four choices in placing the 1 in the top-left class II block; after that the 4 must be placed diagonally from 1. The 1 and 2 entries cannot be in the same column, since 1 and 2 must be in different columns in the class III block below it, so the placement of the 2 and 3 entries in the first block is uniquely determined. There are two choices for orienting each of the blocks adjacent to the top-left block, and one way to complete the remaining block. Thus there are  $(4)(2)(2) = 16$  ways to fill in the board, and taking into consideration the 4 possible rotations there are 64 puzzles.  $\square$

**Lemma 13.** *Boards of type (II, II) do not correspond to puzzles.*

*Proof.* Consider a block from class II. The 1 and 4 elements are uniquely determined, but there are two remaining cells which contain precisely the same inequality set. Thus, given any class II block, it is not possible to identify a unique placement of either the 2 or the 3 entries. As we see in Figure 20, even when the 1 and 4 entries are placed in a type (II, II) board we still have two choices for placing 2 and 3, and therefore removing the entries does not create a puzzle.  $\square$

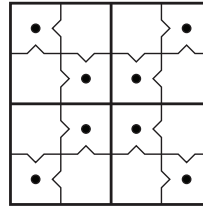
**Lemma 14.** *Boards of type (III, III) do not correspond to puzzles.*

*Proof.* Class III blocks each have two minimal and two maximal cells. Regardless of how we orient the blocks in a type (III, III) board, when we remove the entries



1	3	4	2
4	2	1	3
3	1	2	4
2	4	3	1

board type (III, III)



no unique placements

**Figure 21.** Type (III, III) example.

the unfilled board will have two minimal cells in each row and column. As seen in Figure 21, we have two choices for placing the 1 entry in the first block, which then determines the placement of the remaining 1 entries as well as the 2 entries. Similarly, having two maximal cells in each row and column will result in two different ways to place the set of 3 and 4 entries. Thus the unfilled board has 4 solutions and is not a valid puzzle.  $\square$

Combining these six lemmas, we can now make a statement about Greater Than Shidoku puzzles.

**Theorem 15.** *There are 224 Greater Than Shidoku puzzles.*

## 7. Further explorations

When creating Greater Than puzzles of any size, we must avoid the use of blocks that contain cycles because these blocks are unsolvable. We have shown that every acyclic inequality block is solvable, thus the acyclic Greater Than Shidoku, Rokudoku, and Sudoku blocks that were counted in Section 3 may all potentially be used to create Greater Than puzzles. Nevertheless, we have also seen that not all types of solvable blocks may be used together to form a valid puzzle. The task of combining blocks to form puzzles becomes increasingly complex as the size of the puzzle increases, and research related to standard Sudoku is not always directly applicable. For example, Rosenhouse and Taalman [2011] showed that there are 288 standard Shidoku boards, but those corresponding to types (II, II) and (III, III) do not have a unique solution when the numbers are removed and only the inequalities remain, so there are fewer Greater Than Shidoku puzzles than boards. The equivalence relation defined on  $2 \times 2$  acyclic blocks can also be applied to other sets of  $m \times n$  acyclic blocks (excluding reflection across the diagonals if  $m \neq n$ ). This will allow us to partition the sets of acyclic blocks into equivalence classes that may facilitate our investigation, but Felgenhauer and Jarvis' computer-aided calculation [2006] of 6,670,903,752,021,072,936,960 standard  $9 \times 9$  Sudoku boards hints at the challenge presented by task.

### Puzzle solutions

2	1	6	3	4	5	3
3	4	5	2	1	6	9
1	2	3	6	5	4	9
6	5	4	1	2	3	8
5	6	1	4	3	2	7
4	3	2	5	6	1	8

9	6	7	8	5	1	2	4	3
1	4	5	7	3	2	8	9	6
3	2	8	9	4	6	1	7	5
5	3	4	2	6	9	7	8	1
6	8	1	4	7	5	9	3	2
2	7	9	3	1	8	6	5	4
8	9	6	9	3	1	2	4	5
7	1	6	5	8	3	4	2	9
4	5	2	6	9	7	3	1	8

9	1	2	7	4	5	8	3	6
4	5	8	6	3	2	9	7	1
7	3	6	9	1	8	4	5	2
1	8	1	3	4	2	5	6	9
2	9	2	5	4	6	7	8	3
3	6	3	5	8	9	7	1	4
8	4	8	3	1	7	6	2	5
6	7	1	1	2	5	9	3	4
5	2	9	4	8	3	8	6	7

### References

- [Epp 2004] S. S. Epp, *Discrete mathematics with applications*, 3rd ed., Brooks Cole, Belmont, CA, 2004. Zbl 0714.00001
- [Felgenhauer and Jarvis 2006] B. Felgenhauer and F. Jarvis, “Mathematics of Sudoku, I”, *Mathematical Spectrum* **39**:1 (2006), 15–22.
- [Mepham 2011] M. Mepham, “Sudoku”, *Grand Rapids Press* (April 17, 2011), H5.
- [Rosenhouse and Taalman 2011] J. Rosenhouse and L. Taalman, *Taking Sudoku seriously: the math behind the world’s most popular pencil puzzle*, Oxford University Press, Oxford, 2011. MR 2859240 Zbl 1239.00014
- [Stembridge 2009] J. R. Stembridge, “A Maple package for posets”, 2009, Available at <http://math.lsa.umich.edu/~jrs/maple.html#posets>.
- [Sudoku 2006] Psycho Sudoku, “Greater than Sudoku”, *Boston Phoenix* (June 15, 2006).
- [Szpilrajn 1930] E. Szpilrajn, “Sur l’extension de l’ordre partiel”, *Fundamenta Mathematicae* **16** (1930), 386–389. JFM 56.0843.02

Received: 2012-12-13

Accepted: 2013-03-30

burge180@umn.edu

University of Minnesota, 214 Folwell Hall,  
9 Pleasant Street, SE, Minneapolis, MN 55455, United States

smithshe@gvsu.edu

Department of Mathematics, Grand Valley State University,  
1 Campus Drive, Allendale, MI 49401-9403, United States

kyvarga@ncsu.edu

Department of Mathematics, North Carolina State University,  
2108 SAS Hall, Box 8205, Raleigh, NC 27695, United States

# Spanning tree congestion of planar graphs

Hiu Fai Law, Siu Lam Leung and Mikhail I. Ostrovskii

(Communicated by Joseph A. Gallian)

This paper is devoted to estimates of the spanning tree congestion for some planar graphs. We present three main results: (1) We almost determined (up to  $\pm 1$ ) the maximal possible spanning tree congestion for planar graphs. (2) The value of congestion indicator introduced by Ostrovskii [*Discrete Math.* **310**, 1204–1209] can be very far from the value of the spanning tree congestion. (3) We find some more examples in which the congestion indicator can be used to find the exact value of the spanning tree congestion.

## 1. Introduction

Let  $G$  be a graph and let  $T$  be a spanning tree in  $G$ . We follow the terminology and notation of [Clark and Holton 1991]. For each edge  $e$  of  $T$ , let  $A_e$  and  $B_e$  be the vertex sets of the components of  $T - e$  (see Figure 1). By  $e_G(A_e, B_e)$  we denote the number of edges in  $G$  with one end vertex in  $A_e$  and the other end vertex in  $B_e$ . We define the *edge congestion* of  $G$  in  $T$  by

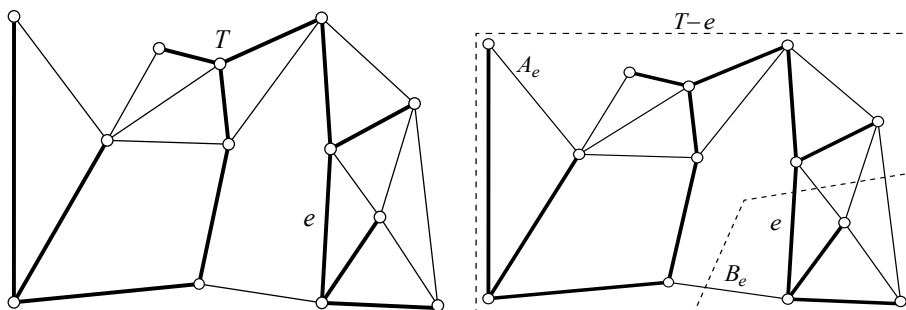
$$ec(G : T) = \max_{e \in E(T)} e_G(A_e, B_e).$$

The number  $e_G(A_e, B_e)$  is called the *congestion* in  $e$ . The name comes from the following analogy. Imagine that edges of  $G$  are roads, and edges of  $T$  are those roads which are cleaned of snow after snowstorms. If we assume that each edge in  $G$  bears the same amount of traffic, and that after a snowstorm each driver takes the corresponding (unique) detour in  $T$ , then  $ec(G : T)$  describes the traffic congestion at the most congested road of  $T$ . Clearly, it is interesting for applications to find a spanning tree which minimizes the congestion.

*MSC2010:* primary 05C05; secondary 05C10, 05C35.

*Keywords:* dual graph, dual spanning tree, minimum congestion spanning tree, planar graph.

Hiu Fai Law was a doctoral student at Oxford University (UK) at the time of work on this paper. The paper contains the main results of Siu Lam Leung's Master's thesis, written under the supervision of Mikhail Ostrovskii. Ostrovskii was supported in part by NSF DMS-1201269. The authors would like to thank the referee for helpful criticism of the first version of this paper.



**Figure 1.** Left: spanning tree  $T$  of a graph  $G$ . Right: subgraph  $T - e$  of  $G$ . In this case,  $e_G(A_e, B_e) = 5$ .

We define the *spanning tree congestion* of  $G$  by

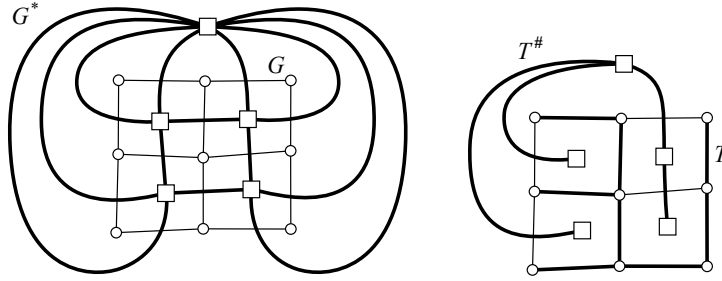
$$s(G) = \min\{ec(G : T) : T \text{ is a spanning tree of } G\}. \quad (1)$$

Each spanning tree  $T$  in  $G$  satisfying  $ec(G : T) = s(G)$  is called a *minimum congestion spanning tree*. The definitions of  $ec(G : T)$  and  $s(G)$  were introduced and their study initiated in [Ostrovskii 2004]. Closely related parameters were introduced earlier in [Simonson 1987, p. 236; Khuller et al. 1993]. After the publication of [Ostrovskii 2004], the spanning tree congestion became the object of active study. As a result the spanning tree congestion was computed and estimated for many families of graphs — see [Law and Ostrovskii 2010; Otachi 2011] for surveys of such results and further references. Algorithmic issues of the problem were studied in [Bodlaender et al. 2012; Löwenstein 2010; Otachi et al. 2010]. In [Löwenstein 2010, Section 5.6] and [Otachi et al. 2010] it was independently discovered that the spanning tree congestion is computationally hard. The contents of the latter were incorporated in [Bodlaender et al. 2012], which contains a systematic analysis and the strongest known results on the algorithmic complexity of problems related to the spanning tree congestion. In a note to Lemma 8, we mention what seems to be the easiest known way to show that the spanning tree congestion problem is NP-hard even for planar graphs.

In this paper we restrict our attention to the study of the spanning tree congestion for planar graphs. In this case some additional tools are available, but computing the spanning tree congestion is still NP-hard and offers some challenging problems.

The main results of this paper:

- (1) We almost determined (up to  $\pm 1$ ) the maximal possible spanning tree congestion for planar graphs; see Section 3.
- (2) The computational hardness of the spanning tree congestion problem makes us interested in parameters which approximate the spanning tree congestion. We



**Figure 2.** Left: The dual graph  $G^*$  of a graph  $G$ . Right: the dual tree  $T^\sharp$  of a spanning tree  $T$ .

find that the value of the congestion indicator introduced in [Ostrovskii 2010] is very far from the value of the spanning tree congestion for some graphs; see Section 4.

- (3) We find more examples in which the congestion indicator introduced in [Ostrovskii 2010] can be used to find the exact value of the spanning tree congestion; see Section 5.

### 2. Dual graphs, indices and center-tail systems

In this section we introduce tools which can be used to estimate the spanning tree congestion and which are available for planar graphs only. By a *plane graph* we mean a planar graph whose planar drawing is fixed.

**Definition 1.** The *dual graph*  $G^*$  of a plane graph  $G$  is defined to be the multigraph whose vertices correspond to the faces of  $G$ , including the exterior face  $O$ . Two faces are joined by an edge if and only if they have a common edge in their boundaries. (If two faces have several common edges in their boundaries, the corresponding edges are multiple edges.) Note that an edge  $e^* \in E(G^*)$  corresponding to  $e \in E(G)$  joins the faces of  $G$  (equal to the vertices of  $V(G^*)$ ) whose boundaries contain  $e$ . If  $T$  is a spanning tree of  $G$ , then the *dual tree*  $T^\sharp$  is defined as a spanning subgraph of  $G^*$  such that  $e^* \in E(T^\sharp)$  if and only if  $e \notin E(T)$  (see Figure 2). As is well known,  $T^\sharp$  is a spanning tree in  $G^*$  (see [Lovász 2007, solution of Problem 5.23] for an explanation).

**Definition 2** [Ostrovskii 2010]. An edge  $e \in E(G)$  is said to be an *outer edge* of  $G$  if it lies on the boundary of the exterior face. The *index*  $i(F, e)$ , where  $F$  is a bounded face and  $e$  is an outer edge, is defined to be the length of the shortest path in  $G^*$  which joins the exterior face  $O$  with  $F$  and satisfies the condition that  $e^*$  is the first edge in the path.

**Definition 3** [Ostrovskii 2010]. A *center-tail system*  $S$  in the dual graph  $G^*$  of a plane graph  $G$  consists of:

- (1) A connected set  $C$  of vertices of  $G^*$ , which is called a *center*.
- (2) A set of paths in  $G^*$  which join some vertices of the center  $C$  with the exterior face. Such a path is called a *tail*. The *tip* of a tail is the last vertex of the corresponding path before it reaches the exterior face.
- (3) An assignment of *opposite tails* for outer edges of  $G$ . This means that for each outer edge  $e$ , a tail is assigned to be the opposite tail, which is denoted by  $N(e)$  and its tip by  $t(e)$ .

**Definition 4** [Ostrovskii 2010]. The *congestion indicator*  $CI(S)$  of a center-tail system  $S$  is defined as the minimum of three numbers:

- (1)  $\min_{F,H,f,h} (i(F, f) + i(H, h) + 1)$ , where the minimum is taken over all pairs  $F, H$  of adjacent vertices in the center  $C$  and over all pairs  $f, h$  of outer edges with  $f \neq h$ . In the case where the center consists of just one vertex, we assume that the minimum is  $\infty$ .
- (2)  $\min_e i(t(e), e) + 1$ , where the minimum is taken over all outer edges of  $G$ .
- (3)  $\min_e \min_{F \in N(e)} \min_{\tilde{e} \neq e} (i(F, e) + i(H, \tilde{e}) + 1)$ , where the first minimum is taken over all outer edges of  $G$ ; the second minimum is over vertices  $F$  from the path  $N(e)$  different from  $t(e)$  and the exterior face, and  $H$  is the vertex in  $N(e)$  which follows immediately after  $F$  if one moves along  $N(e)$  from  $F$  to  $t(e)$ ; and the third minimum is over all outer edges different from  $e$ .

**Theorem 5** [Ostrovskii 2010]. Let  $S$  be any center-tail system in a connected planar graph  $G$ . Then  $s(G) \geq CI(S)$ .

**Definition 6** [Ostrovskii 2010]. The *absolute index*  $i(F)$  of a face  $F$  is defined as  $\min_e i(F, e)$ , where the minimum is over all outer edges.

**Theorem 7** [Ostrovskii 2010]. For each connected planar graph  $G$  with at least two bounded adjacent faces, we have  $s(G) \leq \max(i(F) + i(H)) + 1$ , where the maximum is over all pairs  $F, H$  of bounded faces which have a common edge in their boundaries.

For the study of maximal spanning tree congestion, we make use of results on graph radius. Recall that given a connected graph  $G$ , the *radius* is

$$\text{rad}(G) = \min_{x \in V(G)} \max_{y \in V(G)} d_G(x, y). \quad (2)$$

A vertex  $x$  for which the minimum in (2) is attained is called *central*. (Warning: this notion of centrality is not related to the center-tail systems introduced above.)

For planar graphs the spanning tree congestion is closely related to the widely used notion of *stretch*; see [Peleg 2000, p. 166].

If  $H$  is a connected spanning subgraph in  $G$ , then its *stretch* is defined by

$$\text{Stretch}(H) = \max_{u,v \in V(G)} \frac{d_H(u, v)}{d_G(u, v)}. \tag{3}$$

The following observations can be found in [Ostrovskii 2010; Otachi et al. 2010; Peleg 2000].

**Lemma 8.** *Let  $G$  be a connected planar graph.*

(a) *If  $T$  is a spanning tree in  $G$  and  $T^\sharp$  is its dual tree, then*

$$\text{ec}(G : T) = \text{Stretch}(T^\sharp) + 1.$$

(b)  *$s(G) = \inf_{T^\sharp} \text{Stretch}(T^\sharp) + 1$ , where the infimum is over all spanning trees  $T^\sharp$  in the dual graph  $G^*$ .*

(c)  $\min_{T^\sharp} \text{Stretch}(T^\sharp) \leq 2 \text{rad}(G^*)$ .

*Proof.* It is easy to see that the number of detours using an edge  $e \in T$  is the length of the cycle obtained by adding the edge  $e^*$  to  $T^\sharp$ . On the other hand, the length of this cycle is exactly  $d_{T^\sharp}(u, v) + 1$ , where  $u, v$  are the ends of  $e^*$ . Therefore  $\text{ec}(G : T) = \text{Stretch}(T^\sharp) + 1$ , proving (a).

The statement (b) follows immediately from (a).

To prove (c) it suffices to observe that any breadth-first search (BFS) tree  $T^\sharp$  in  $G^*$  rooted at one of its central vertices  $C$  satisfies  $\text{Stretch}(T^\sharp) \leq 2 \text{rad}(G^*)$ . (See [Rosen et al. 2000, Section 9.2.1] or [Nishizeki and Chiba 1988, p. 31] for information on BFS trees.) To see the inequality  $\text{Stretch}(T^\sharp) \leq 2 \text{rad}(G^*)$  we need only the defining property of a BFS tree in  $G^*$  rooted at  $C$ : it is a spanning tree in  $G^*$  in which the distance between any vertex and  $C$  is the same as in  $G^*$ , and therefore is  $\leq \text{rad}(G^*)$ .  $\square$

**Note.** Fekete and Kremer [2001] proved that the determination of the least  $t$  for which a planar graph has a spanning tree  $T$  with  $\text{Stretch}(T) = t$  is NP-hard. Combining this with Lemma 8 we get that the problem of computation of  $s(G)$  for planar graphs is also NP-hard.

### 3. On the maximal spanning tree congestion of planar graphs

The purpose of this section is to find sharp estimates of the quantity

$$\mu_p(n) = \max\{s(G) : G \text{ is a planar graph with } n \text{ vertices}\}.$$

Graphs  $G$  with  $n$  vertices satisfying  $s(G) = \mu_p(n)$  can be called *the most congested planar graphs with  $n$  vertices*.

**Note.** A consequence of Euler's formula is that a simple planar graph with  $n \geq 3$  vertices has at most  $3n - 6$  edges. As  $n - 1$  of them are in a spanning tree, they are detours for themselves. Therefore the spanning tree congestion cannot exceed  $3n - 6 - (n - 1) + 1$ . Thus  $\mu_p(n) \leq 2n - 4$ . Our purpose is to get more precise estimates for  $\mu_p(n)$ .

**Theorem 9.** *Let  $n \geq 5$ . If  $n$  is even, then  $n \leq \mu_p(n) \leq n + 1$ . If  $n$  is odd, then  $n - 1 \leq \mu_p(n) \leq n$ .*

The proof of this theorem naturally splits into two parts: estimates from above (Section 3.1) and estimates from below (Section 3.2).

**Problem 10.** Fill the gap of size 1 between the upper and lower estimates in Theorem 9.

**3.1. Estimates from above.** We need some terminology and notation of [Diestel 2000]. A plane graph is called a *plane triangulation* if all faces of it are triangles. Adding some edges (but not vertices) to an arbitrary planar graph  $G$  we get a plane triangulation  $G_t$  which we call a *triangulation* of  $G$ . It is easy to construct examples showing that  $G_t$ , in general, is not uniquely determined by  $G$ .

**Lemma 11.**  $\text{rad } G_t^* \geq \text{rad } G^*$ .

*Proof.* To see this, it suffices to observe that  $G^*$  is a minor of  $G_t^*$ , obtained if sets of triangular faces of  $G_t$  that originated from the same face of  $G$  are considered as branch sets (see [Diestel 2000, p. 16] for minor-related definitions). It is clear that such sets are connected in  $G_t^*$  and the corresponding minor is isomorphic to  $G^*$ . Since in creating this minor we did not delete any edges or vertices, the radius of the resulting graph can only be less than the radius of  $G_t^*$ , and we get the desired inequality.  $\square$

The following two facts are well known; see, for example, [Diestel 2000, Section 4.4; Exercise 40 in Chapter 4].

**Lemma 12.** *A triangulation of a planar graph with at least 4 vertices is 3-connected.*

**Lemma 13.** *The dual graph of a 3-connected planar graph is a 3-connected planar graph.*

Finally we need the following tight estimate for a radius of a 3-connected graph obtained in [Iida 2007]. (See [Egawa and Inoue 1999; Harant 1993; Harant and Walther 1981; Iida and Kobayashi 2006; Inoue 1996] for preceding and related estimates.)

**Theorem 14** [Iida 2007]. *Let  $G$  be a 3-connected graph with radius  $r$ . Then*

$$|V(G)| \geq 4r - 4.$$



**Lemma 15.** *Let  $n \geq 4$ . Then*

$$\mu_p(n) \leq \begin{cases} n + 1 & \text{if } n \text{ is even,} \\ n & \text{if } n \text{ is odd.} \end{cases}$$

*Proof.* Let  $G$  be a plane graph with  $n$  vertices satisfying  $s(G) = \mu_p(n)$ . By Lemma 12 the graph  $G_t$  is 3-connected. By Lemma 13 the graph  $G_t^*$  is also 3-connected. An easy computation with Euler’s formula shows that  $G_t^*$  has  $2n - 4$  vertices. By Theorem 14 we get  $\text{rad}(G_t^*) \leq 2n/4 = n/2$ . By Lemma 11 we get  $\text{rad}(G^*) \leq n/2$ . Therefore  $\text{rad}(G^*) \leq n/2$  if  $n$  is even and  $\text{rad}(G^*) \leq (n - 1)/2$  if  $n$  is odd. Combining these inequalities with Lemma 8 we get

$$s(G) \leq \begin{cases} n + 1 & \text{if } n \text{ is even,} \\ n & \text{if } n \text{ is odd.} \end{cases} \quad \square$$

**3.2. Estimates from below.** For  $n \geq 5$ , we let  $B_n$  be graphs of bipyramids whose bases are  $(n - 2)$ -gons. These graphs can be constructed in the following way: we start with  $C_{n-2}$  (cycle of length  $n - 2$ ), then introduce two more vertices and join each of them with each of the vertices in the cycle.

**Lemma 16.** *Let  $n \geq 5$ . Then*

$$s(B_n) = \begin{cases} n & \text{if } n \text{ is even,} \\ n - 1 & \text{if } n \text{ is odd.} \end{cases} \quad (4)$$

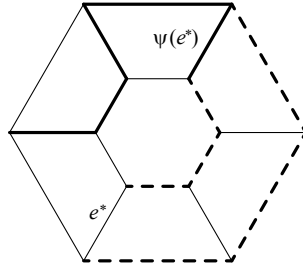
*Proof.* Observe that the dual of  $B_n$  is  $B_n^* = C_{n-2} \times K_2$ .

Denote by  $\ell(n)$  the case-defined function given by the right-hand side of (4). By the proof of Lemma 8, in order to prove  $s(B_n) \geq \ell(n)$  it suffices to show that for an arbitrary spanning tree  $T^\sharp$  in  $B_n^*$  there is an edge  $e^*$  in  $B_n^*$  which is not in  $T^\sharp$  and such that  $T^\sharp \cup \{e^*\}$  contains a cycle of length  $\geq \ell(n)$ . The inequality  $s(B_n) \leq \ell(n)$  will also follow from our argument, but it is clear that the main point of Lemma 16 is the lower estimate.

An edge in  $B_n^*$  is called *vertical* if its end vertices are  $(c, k_1)$  and  $(c, k_2)$ , where  $c$  is a vertex of  $C_{n-2}$  and  $k_1, k_2$  are vertices of  $K_2$ ; otherwise, it is *horizontal*. Two horizontal edges form a *couple* if they correspond to the same edge in  $C_{n-2}$ .

If all vertical edges are in  $T^\sharp$ , then there is a couple  $e^*, f^*$  of horizontal edges which are both not in  $T^\sharp$  (otherwise  $T^\sharp$  would contain a cycle). Clearly, at least one of  $e^*, f^*$  creates together with edges of  $T^\sharp$  a cycle of length at least  $n \geq \ell(n)$ .

Now suppose that there are vertical edges which are not in  $T^\sharp$ . Let  $e^*$  be one of the vertical edges in  $E(B_n^*) \setminus E(T^\sharp)$ . Then  $T^\sharp \cup \{e^*\}$  contains a cycle. If this cycle contains an edge from each couple of the horizontal edges, we say that it *goes around*. It is clear that if the cycle contained in  $T^\sharp \cup \{e^*\}$  goes around, then it has length  $\geq n \geq \ell(n)$ . If it does not go around, then it contains exactly one more vertical edge.



**Figure 3.** Different sides of  $e^*$  and  $\psi(e^*)$  in  $B_8^*$ .

Therefore, if there are no cycles of the described type which go around, then there is a mapping  $\psi$  from the set of vertical edges which are not in  $E(T^\sharp)$  to the set of vertical edges which are in  $E(T^\sharp)$  satisfying this condition: all couples of horizontal edges on one of the “sides” between  $e^*$  and  $\psi(e^*)$  belong to  $T^\sharp$ . To clarify the meaning of the word “sides” in the previous sentence we show different sides in Figure 3 using dashed and continuous lines, respectively, attribution of vertical edges to sides does not matter; the tree  $T^\sharp$  is shown using thick lines, dashed or continuous. In this way, vertical edges split into groups having the common image under  $\psi$ . We include  $f^*$  in the group of edges  $e^*$  for which  $\psi(e^*) = f^*$ . It is clear that all vertical edges between  $e^*$  and  $\psi(e^*)$  which are on the suitable side (see above) belong to the same group as  $e^*$ . Therefore, the groups partition the vertex set of the cycle  $C_{n-2}$  into connected pieces.

If there is just one connected piece, then there is just one vertical edge in  $E(T^\sharp)$ , and all but two horizontal edges are in  $E(T^\sharp)$ . It is clear that the missing horizontal edges should form a couple (otherwise there would be a vertical edge  $e^*$  for which the cycle in  $T^\sharp \cup \{e^*\}$  goes around). It is in this case that we get a weaker estimate for odd  $n$ .

In fact, if the end vertices of the only vertical edge of  $T^\sharp$  divide those pieces of  $C_{n-2} \times \{k_1\}$  and  $C_{n-2} \times \{k_2\}$  which are in  $T^\sharp$  into parts of equal length (this is possible if  $n$  is odd), then the maximal length of the cycle in  $T^\sharp \cup \{e^*\}$  over  $e^* \in E(B_n^*) \setminus E(T^\sharp)$  is  $n - 1$ . (Otherwise, the longest cycle in  $T^\sharp \cup \{e^*\}$  has length at least  $n + 1$ .)

On the other hand, if  $n$  is even, the cycle obtained by adding to  $E(T^\sharp)$  the vertical edge which is most distant from the one contained in  $E(T^\sharp)$  produces a cycle of length at least  $n$ .

Now we suppose that there are at least two connected pieces. We consider horizontal edges between the neighboring intervals. It is easy to check that if there are at least three intervals, there is a pair of neighboring intervals with no edges in

$T^\sharp$  between them. If there are two intervals, then on one side there are no edges in  $T^\sharp$  between them.

Let  $e_1^*$  and  $e_2^*$  be the corresponding missing horizontal edges. Then  $E(T^\sharp) \cup \{e_1^*\}$  or  $E(T^\sharp) \cup \{e_2^*\}$  contains a cycle which contains vertical edges and therefore has length  $\geq n \geq \ell(n)$ .  $\square$

#### 4. Limitations of center-tail systems

In this section we show that for some classes of planar graphs the estimates of the spanning tree congestion given by center-tail systems (see Theorem 5) are far from being sharp. More precisely we prove the following result.

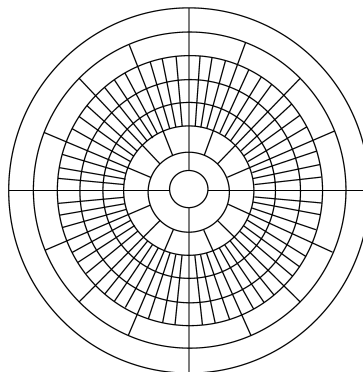
**Theorem 17.** *There exists a sequence  $\{G_n\}_{n=1}^\infty$  of planar graphs such that*

$$\lim_{n \rightarrow \infty} s(G_n) = \infty,$$

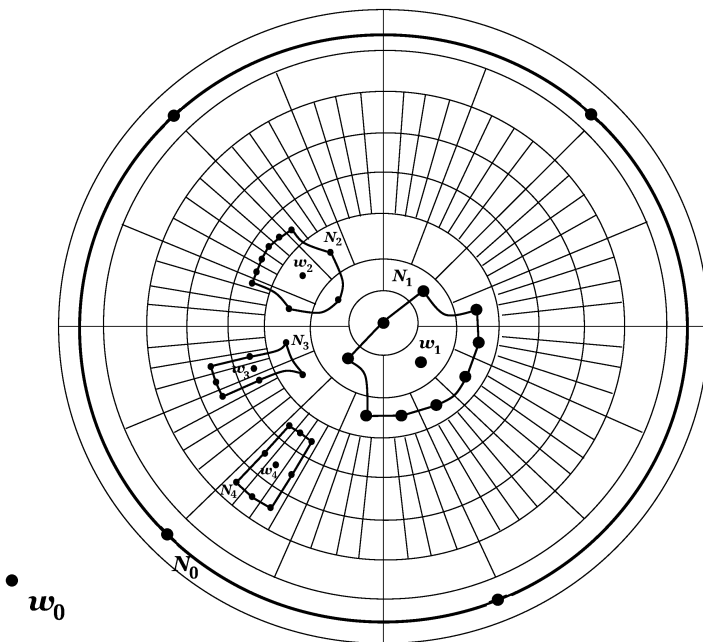
but for any center-tail system  $S_n$  in  $G_n$ , we have  $CI(S_n) \leq 6$ .

**Note.** By a center-tail system for a planar graph we mean a center-tail system of any of its drawings. In particular, any of the faces of the graph can be regarded as its exterior face.

*Proof.* Before defining the graphs  $G_n$ , it is convenient to define a two-parametric family of graphs, which we denote  $\{Q_{n,m}\}_{n,m=1}^\infty$ . To construct the graph  $Q_{n,m}$  we start with a family of  $2n+m$  concentric circles. They cut out of the plane  $2n+m-1$  concentric annuli. We cut both the outer and the inner annuli into 4 pieces each using radial cuts (see Figure 4). We next cut annuli, both from the inner and the outer side, into  $4^2$  equal pieces using radial cuts. We make these radial cuts in such a way that they extend the radial cuts done in the first step (see Figure 4). Continuing, for each  $k \leq n$  we cut the  $k$ -th annuli, both from the inner and the outer



**Figure 4.** A planar graph  $Q_{3,2}$ .



**Figure 5.** A planar graph  $Q_{3,2}$  with pieces of the dual graph needed to estimate the congestion indicator.

side, into  $4^k$  equal pieces using radial cuts. We cut the remaining  $m - 1$  annuli in the same way as the annuli in the last set, that is, using  $4^n$  radial cuts. In Figure 4, we show the resulting graph in the case where  $n = 3$  and  $m = 2$ .

Then  $G_n$  is defined as  $Q_{n,4^n-2}$ . Now we estimate the congestion indicator. Recall that CI is a minimum of three terms, one of which is

$$\min_e i(t(e), e) + 1.$$

Clearly this term, in the case where the face playing the role of the exterior face is denoted by  $w$ , does not exceed

$$\max_{u,v} d(u, v) + 2, \tag{5}$$

where the maximum is over pairs  $u, v$  of vertices in  $G_n^*$ , both of which are adjacent to  $w$ , and  $d$  is the graph distance in  $G_n^* - w$ . To estimate from above the value of (5), we observe that vertices adjacent to  $w$  in  $G_n^*$  belong to a cycle in  $G_n^* - w$  whose length is between 4 and 9. See Figure 5, in which we denote several possible choices of  $w$  by  $w_0, w_1, w_2, w_3$ , and  $w_4$ , and denote the cycles described in the previous sentence by  $N_0, N_1, N_2, N_3$ , and  $N_4$ , respectively. It is clear that the distance between any two vertices of such a cycle of length  $\leq 9$  does not exceed 4, so the

maximum in (5) does not exceed 6. We get the desired estimate: the congestion indicator of any center-tail system in any of the graphs  $Q_{n,m}$ , and therefore in any of the graphs  $G_n$ , does not exceed 6.

Now we turn to spanning tree congestion estimates. Here we use the approach suggested in [Ostrovskii 2004] using centroids and isoperimetric estimates.

**Definition 18** [Jordan 1869]. Let  $u$  be a vertex of a tree  $T$ . Let the *weight of  $T$  at  $u$*  be the maximal number of vertices in components of  $T - u$ . A vertex  $v$  of  $T$  is called a *centroid* vertex if the weight of  $T$  at  $v$  is minimal.

Let  $T$  be an optimal tree in  $G_n$  so that  $ec(G_n : T) = s(G_n)$ . Let  $u$  be a centroid of  $T$ . Since the maximum degree of  $G_n$  is 4, there are at most 4 edges incident with  $u$ . Let

$$O_{G_n} = \left\lceil \frac{|V(G_n)| - 1}{4} \right\rceil.$$

Since  $u$  is a centroid, it is not hard to see that there is a component of  $T - u$  whose vertex set  $A$  satisfies

$$O_{G_n} \leq |A| \leq \frac{|V(G_n)|}{2}.$$

As the edge connecting  $u$  with  $A$  is used in  $e_{G_n}(A, V(G_n) - A)$  detours, any lower bound of this number, where  $A$  runs over sets of size within the above range, is a lower bound of  $s(G_n)$ .

We use the following special case of the isoperimetric result of Bollobás and Leader [1991, Theorem 3]. Let  $R(k)$  be the graph with vertex set

$$[k]^2 = \{0, 1, 2, 3, \dots, k - 1\}^2$$

in which  $x = (x_1, x_2)$  is adjacent to  $y = (y_1, y_2)$  if and only if  $|x_i - y_i| = 1$  for some  $i$  and  $x_j = y_j$  for  $j \neq i$ .

**Theorem 19.** *Let  $B$  be a subset of  $[k]^2$  with  $|B| \leq k^2/2$ . Then*

$$e_{R(k)}(B, \bar{B}) \geq \min\{2\sqrt{|B|}, k\}. \tag{6}$$

Let us introduce the function

$$f_k(t) = \min\{k, 2\sqrt{t}\} \quad \text{for } t \in \left[0, \frac{k^2}{2}\right].$$

Observe that the graph  $G_n$  has a subgraph  $S_n$  isomorphic to  $R(4^n)$ . Indeed, we may take  $S_n$  to contain all vertices of the  $4^n$  central circles and all the corresponding edges except one “radial” set of  $4^n$  edges. The subgraph  $S_n$  has  $4^n \times 4^n = 4^{2n}$  vertices. In addition,  $G_n$  has  $2(4 + 4^2 + \dots + 4^{n-1}) = \frac{8}{3}(4^{n-1} - 1)$  vertices on the  $2(n - 1)$  circles which are not in  $S_n$ . It is clear that the intersection of the set  $A$

with the vertex set of  $S_n$  has at most  $2 \cdot 4^{2n-1} + \frac{4}{3}(4^{n-1} - 1)$  vertices. We need also the inequality

$$|A \cap V(S_n)| \geq 4^{2n-1} - 2(4^{n-1} - 1).$$

To get this inequality we recall that

$$|A| \geq \left\lceil \frac{|V(G_n)| - 1}{4} \right\rceil \geq 4^{2n-1} + \frac{2}{3}(4^{n-1} - 1) - \frac{1}{4},$$

and observe that

$$\begin{aligned} |A \cap V(S_n)| &\geq |A| - (|V(G_n)| - |V(S_n)|) \\ &\geq 4^{2n-1} + \frac{2}{3}(4^{n-1} - 1) - \frac{1}{4} - \frac{8}{3}(4^{n-1} - 1) \\ &= 4^{2n-1} - 2(4^{n-1} - 1) - \frac{1}{4}. \end{aligned}$$

We may drop  $\frac{1}{4}$  since  $|A \cap V(S_n)|$  is an integer.

Applying Theorem 19 to the smaller of  $A \cap V(S_n)$  and  $V(S_n) \setminus A$ , we get that the number of edges joining  $A \cap V(S_n)$  with  $V(S_n) \setminus A$  can be estimated from below by

$$\min_t \{f_{4^n}(t)\},$$

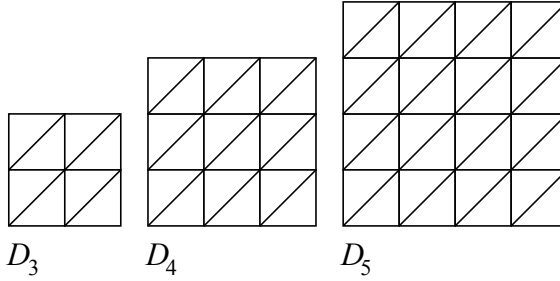
where  $t$  ranges from  $\min\{4^{2n-1} - 2(4^{n-1} - 1), 4^{2n} - (2 \cdot 4^{2n-1} + \frac{4}{3}(4^{n-1} - 1))\}$  to  $2 \cdot 4^{2n-1}$ . It is clear that these minima approach  $\infty$  as  $n \rightarrow \infty$ . Therefore  $\lim_{n \rightarrow \infty} s(G_n) = \infty$ . This completes the proof of the theorem.  $\square$

**Note.** It is known that for some planar graphs, center-tail systems and the corresponding congestion indicators give sharp lower bounds of the spanning tree congestion. However, as the above example shows, in some cases the lower bound given by center-tail systems is very far from the actual value of the spanning tree congestion.

**Problem 20.** Is it possible to define a flexible version of the congestion indicator (FCI) such that for some function  $f: \mathbb{N} \rightarrow \mathbb{N}$  and any planar graph  $G$  we have  $s(G) \leq f(n)$  if the maximal possible value of FCI (on the corresponding analogue of the center-tail system in  $G$ ) has value  $\leq n$ ?

## 5. Computing spanning tree congestion by center-tail systems

Center-tail systems were introduced in [Ostrovskii 2010] as a tool to compute or estimate the spanning tree congestion of some plane graphs. In [Ostrovskii 2010] the computation was performed for the triangular grids. Another grid for which center-tail systems give the exact value of the spanning tree congestion was found in [Bodlaender et al. 2011, Theorem 3.7]. In this section we use the center-tail systems and Theorems 5 and 7 to find the spanning tree congestion of other sets of planar graphs.



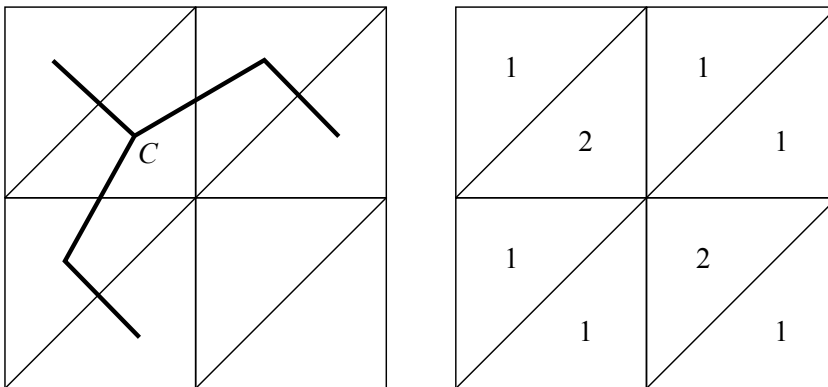
**Figure 6.** A sequence of square-triangular graphs.

**5.1. Square-triangular grids.** Consider the sequence of square-triangular graphs in Figure 6. In this figure, there is a vertex at each intersection of the line segments. The spanning tree congestion of these graphs is computed in the next theorem.

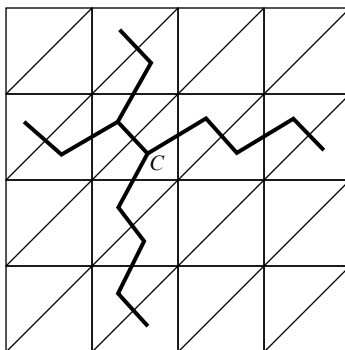
**Theorem 21.** Let  $n \in \mathbb{N}$ . Then

$$s(D_k) = \begin{cases} 4n & \text{if } k = 2n + 1, \\ 4n + 3 & \text{if } k = 2n + 2. \end{cases}$$

*Proof. Case 1.*  $k = 2n + 1$ , where  $n \in \mathbb{N}$ . We start by considering the graph  $D_3$  and its center-tail system  $S_3$  shown in Figure 7. The center for the system  $S_3$  consists of one vertex and is marked with the letter  $C$ . The tail whose tip points to the upper-left corner is assigned as the opposite tail for the outer edges on the right and at the bottom of  $D_3$ . The tail with right-most tip is assigned as the opposite tail for those outer edges on the left. The tail with bottom-most tip is assigned to those outer edges on top. It is easy to see that the congestion indicator  $CI(S_3)$  (see Definition 4) of the center-tail system  $S_3$  is the minimum of three numbers: (i)  $\infty$ ,



**Figure 7.** Left:  $D_3$  with a center-tail system  $S_3$ . Right: absolute indices for  $D_3$ .

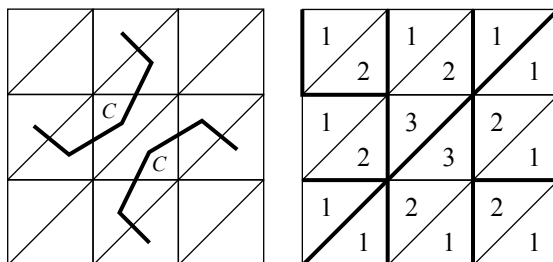


**Figure 8.**  $D_5$  with a center-tail system  $S_5$ .

(ii) 5 and (iii) 4. Hence  $CI(S_3) = 4$  and, by Theorem 5,  $s(D_3) \geq 4$ . According to Theorem 7, we have  $s(D_3) \leq 4$  (see the values of the absolute indices in Figure 7), therefore  $s(D_3) = 4$ .

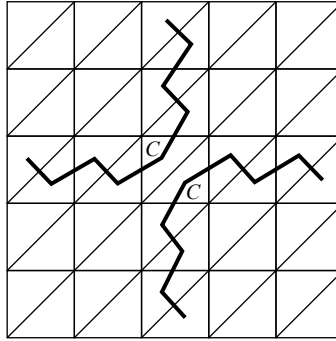
Adding a row on each side of  $D_3$  gives us the graph  $D_5$ . We consider what can be regarded as a natural extension of  $S_3$  to  $S_5$ ; the only feature of this extension which is not completely predictable is that the tail whose tip points to the upper-left corner in  $S_3$  is now splitting into two tails in  $S_5$  (see Figure 8). The tail whose tip points to the left is assigned to the outer edges on the right, and the tail whose tip points upward is assigned to those outer edges at the bottom. Since the indices of the central triangles increase by two as we add a row on each side of the graph, the spanning tree congestion increases by four, so  $s(D_5) = s(D_3) + 4 = 4 + 4 = 8$ . It is clear by induction that  $s(D_{2n+1}) = 4n$  for each  $n \in \mathbb{N}$ .

*Case 2:  $k = 2n + 2$ .* First we consider the graph  $D_4$  and its center-tail system  $S_4$ , described as follows. The center of  $S_4$  consists of two vertices which are labeled  $C$  (see Figure 9). There are four tails, which are drawn in Figure 9 with thick lines. The assignments of opposite tails for outer edges are done in the natural way. For



**Figure 9.** Left: center-tail system  $S_4$ . Right: absolute indices for  $D_4$  and a minimum congestion spanning tree for  $D_4$ .



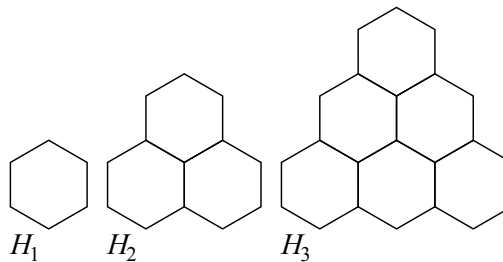


**Figure 10.**  $D_6$  with a center-tail system  $S_6$ .

example, the tail whose tip points to the left is assigned to the outer edges on the right. The tail whose tip points upward is assigned to the outer edges at the bottom of the graph.

It is easy to see that the congestion indicator  $CI(S_4)$  of the center-tail system  $S_4$  is the minimum of the following three numbers: (i)  $3 + 3 + 1 = 7$ , (ii)  $6 + 1 = 7$  and (iii)  $7$ . Hence  $CI(S_4) = \min\{7, 7, 7\} = 7$ . By Theorem 5,  $s(D_4) \geq 7$ . The values of absolute indices  $i(F)$  for  $D_4$  are shown in Figure 9. According to Theorem 7,  $s(D_4) \leq \max(i(F) + i(H)) + 1 = 3 + 3 + 1 = 7$ , where  $F$  and  $H$  are bounded faces with an edge in common, and the maximum is taken over  $F$  and  $H$ . Hence,  $s(D_4) = 7$ . Following the argument of the proof of Theorem 7 in [Ostrovskii 2010], we sketch one of the spanning trees for which the congestion is 7; see Figure 9.

By adding one row on each side of the graph, we obtain the square-triangular grid  $D_6$  (see Figure 10). Addition of a row on each side increases the indices of central triangles by two. Straightforward computation shows that all of the estimates increase by 4, hence  $s(D_6) = s(D_4) + 4 = 11$ . We use induction to show that  $s(D_{2n+2}) = 4n + 3$  for each  $n \in \mathbb{N}$ . □



**Figure 11.** A sequence of hexagonal grids.

**5.2. Hexagonal grids.** A hexagonal grid  $H_k$  is constructed following the pattern shown in Figure 11. Our next purpose is to compute  $s(H_k)$ .

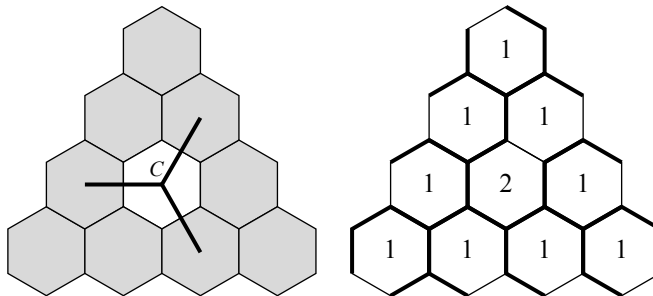
In fact, the following theorem was stated in [Castejon et al. 2007], but its proof was insufficient. The authors of [Castejon et al. 2007] wrote that the proof is the same as their proof for rectangular grids; errors of their proof for rectangular grids were described in [Ostrovskii 2010, p. 1209]. We provide a proof of this theorem using center-tail systems.

**Theorem 22.** *Let  $n \geq 0$  be an integer. Then*

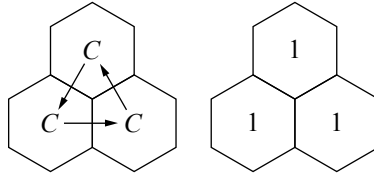
$$s(H_k) = \begin{cases} 2n + 2 & \text{if } k = 3n + 1, \\ 2n + 3 & \text{if } k = 3n + 2, \\ 2n + 3 & \text{if } k = 3n + 3. \end{cases}$$

*Proof. Case 1:  $k = 3n + 1$ .* Since  $H_1$  is isomorphic to  $C_6$ , it is easy to see that  $s(H_1) = 2$ .

By adding one row on each side of  $H_1$ , we obtain the graph  $H_4$  (see Figure 12). Here the center of the center-tail system  $S_4$  consists of one vertex, labeled  $C$ . The tails are drawn with thick lines. The assignments of opposite tails to the outer edges are done in the natural way. The tail whose tip points to the left, downward and upward is assigned to outer edges on the right, the left and at the bottom, respectively (see Figure 12). According to the center-tail system  $S_4$ , we have the three numbers defined in Definition 4: (i)  $\infty$ , since there is only one face in the center, (ii)  $3 + 1 = 4$ , witnessed by an outer edge  $e$  in the middle of any of the three sides since  $i(t(e), e)$  is 3, and (iii)  $2 + 1 + 1 = 4$ . We pick an outer edge  $e$  in the middle of one of the three sides, let  $F = C$ , and  $H$  be the face that contains  $t(e)$ . Then  $i(F, e) = 2$  and  $i(H, \tilde{e}) = 1$ , where  $\tilde{e}$  is an outer edge on the boundary of the face that contains  $H$ . So  $CI(S_4) = \min\{\infty, 4, 4\} = 4$ . By Theorem 5,  $s(H_4) \geq 4$ .



**Figure 12.** Left:  $H_4$  with center-tail system  $S_4$ . The shaded region represents the additional rows added on each side of  $H_1$ . Right: absolute indices for  $H_4$  and minimum spanning congestion tree for  $H_4$ .



**Figure 13.** Left:  $H_2$  with center-tail system  $S_2$ . Right: absolute indices for  $H_2$ .

The absolute indices of faces are shown in Figure 12 (right). The sum of indices of adjacent faces never exceeds 3. Therefore, by Theorem 7, we have  $s(H_4) \leq 4$ . So  $s(H_4) = 4$ .

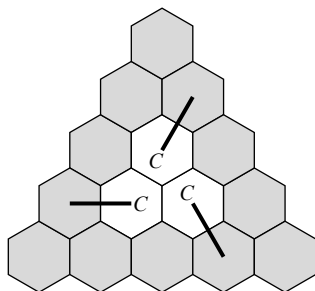
Now we prove that

$$s(H_{3n+1}) = 2n + 2 \tag{7}$$

for each  $n \in \mathbb{N}$ . We use induction. We have shown that (7) holds for  $n = 0, 1$ . It remains to show that  $s(H_{3n+1}) = 2n + 2$  implies  $s(H_{3(n+1)+1}) = 2(n + 1) + 2$ . To see this, we observe that if an additional row is added on each side of  $H_{3n+1}$ , then each side has three more hexagons and the graph becomes  $H_{3n+1+3} = H_{3(n+1)+1}$ . An increment of one row on each side increases the indices of central vertices by one, so each of the three numbers defined in Definition 4, as well as the number  $\max(i(F) + i(H)) + 1$  (see Theorem 7), increase by two. Hence,  $s(H_{3(n+1)+1}) = 2(n + 1) + 2$ .

*Case 2:  $k = 3n + 2$ .* Now consider the graph  $H_2$  with the center-tail system  $S_2$  (see Figure 13). The center of  $S_2$  consists of three vertices labeled  $C$ . The tails are represented by the arrows; their tips correspond to the arrow heads. The tail whose tip points to the right, downward and upward is assigned to the outer edges on the left, the right and the bottom, respectively. According to Definition 4,  $CI(S_2)$  is the minimum of the following three numbers: (i)  $1 + 1 + 1 = 3$ , since the distance from the exterior face  $O$  to any face that contains a vertex of the center is 1. (ii)  $2 + 1 = 3$ , since every tail has length 1, and the distance from  $O$  to any face that contains a vertex of the center is also 1. (iii)  $1 + 1 + 1 = 3$ , based on the same reasoning as in (ii). So  $CI(S_2) = \min\{3, 3, 3\} = 3$ . By Theorem 5,  $s(H_2) \geq 3$ . Since there are only three faces in  $H_2$  and each face is adjacent to one another, by Theorem 7, we have  $s(H_2) \leq 1 + 1 + 1 = 3$ . Therefore,  $s(H_2) = 3$ .

We can obtain the graph  $H_5$  from  $H_2$  by simply adding a row on each side of  $H_2$  (see Figure 14). Notice that the configuration of the center-tail system  $S_5$  for  $H_5$  is different than  $S_2$ . The center of  $S_5$  also consists of three vertices (labeled  $C$ ), and they are located in the middle of the graph (see Figure 14). The tails are drawn with thick lines. The assignment of opposite tails to the outer edge is done in the natural way. The tail whose tip points to the left, downward and upward is assigned to

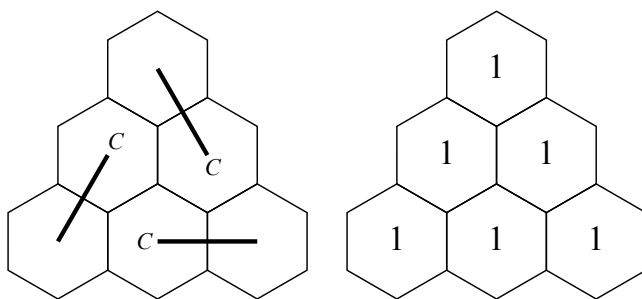


**Figure 14.**  $H_5$  with center-tail system  $S_5$ . The shaded region represents the additional rows added on each side of  $H_2$ .

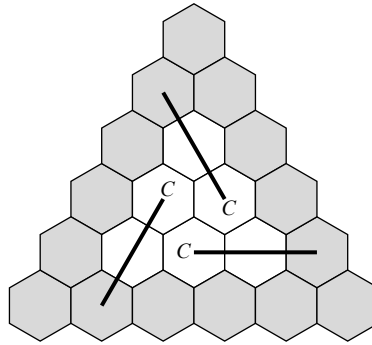
the outer edges on the right, the left and the bottom, respectively. By Definition 4,  $CI(S_5)$  is the minimum of (i)  $2 + 2 + 1 = 5$ , (ii)  $4 + 1 = 5$  and (iii) 5. By Theorem 5,  $s(H_5) \geq CI(S_5) = 5$ , and by Theorem 7,  $s(H_5) \leq 2 + 2 + 1 = 5$ . So  $s(H_5) = 5$ . As in the previous case, we can use the natural extensions of the center-tail system  $S_5$  to prove that  $s(H_{3n+2}) = 2n + 3$  for each  $n \in \{0\} \cup \mathbb{N}$ .

*Case 3:  $n = 3n + 3$ .* The hexagonal grid  $H_3$  and the center-tail system  $S_3$  are shown in Figure 15. The center of  $S_3$  consists of three vertices, labeled  $C$ . The tails for the system are drawn with thick lines. The assignments of opposite tails to the outer edges are natural. The tail whose tip points to the right, upward and downward is assigned to the outer edges on the left, the bottom and the right, respectively (see Figure 15). The congestion indicator  $CI(D_3)$  for  $H_3$  is determined as the minimum of the following three numbers defined in Definition 4: (i)  $1 + 1 + 1 = 3$ , (ii)  $3 + 1 = 4$  and (iii)  $2 + 1 + 1 = 4$ . Hence, by Theorem 5,  $s(H_3) \geq CI(S_3) = 3$ . According to Theorem 7,  $s(H_3) \leq 1 + 1 + 1 = 3$  (see Figure 15). So  $s(H_3) = 3$ .

The graph  $H_6$  can be obtained by adding a row on each side of the graph  $H_3$ . The configuration of the center-tail system  $S_6$  is shown in Figure 16, where the



**Figure 15.** Left:  $H_3$  with center-tail system  $S_3$ . Right: absolute indices for  $H_3$ .



**Figure 16.**  $H_6$  with center-tail system  $S_6$ . The shaded region represents the additional rows added on each side of  $H_3$ .

assignment of opposite tail to the outer edges is done in the obvious and natural way, that is, each tail is assigned to the outer edges in the opposite direction. By Definition 4,  $CI(S_6)$  is the minimum of (i)  $2 + 2 + 1 = 5$ , (ii)  $5 + 1 = 6$  and (iii)  $3 + 2 + 1 = 6$ . By Theorem 5,  $s(H_6) \geq CI(S_6) = 5$ , and by Theorem 7,  $s(H_6) \leq 2 + 2 + 1 = 5$ . So  $s(H_6) = 5$ . Using induction (as in the previous cases) we get  $s(H_{3n+3}) = 2n + 3$  for each  $n \in \{0\} \cup \mathbb{N}$ . This concludes our proof of Theorem 22.  $\square$

**5.3. Rectangular grids.** Let  $R_{m,n}$  denote the rectangular grid consisting of  $m$  horizontal lines and  $n$  vertical lines. The purpose of this section is to show that center-tail systems can be used to prove the following result of Hruska.

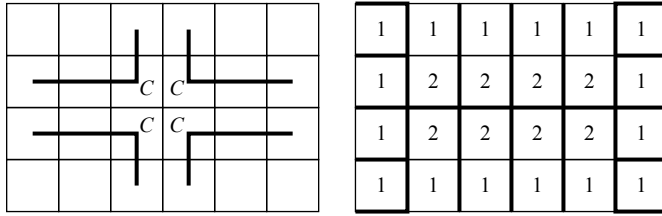
**Theorem 23** [Hruska 2008]. *Suppose  $m < n$ , where  $m$  and  $n$  are natural numbers. Then*

$$s(R_{m,n}) = \begin{cases} m & \text{if } m \text{ is odd,} \\ m + 1 & \text{if } m \text{ is even.} \end{cases}$$

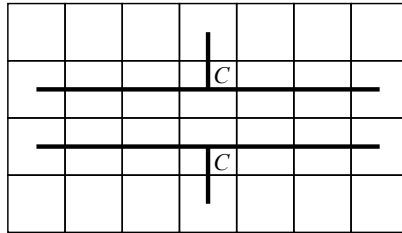
*Proof. Case 1:  $m$  is odd.*

*Subcase 1:  $n$  is also odd.* As an instructive example, we consider  $R_{5,7}$  with the center-tail system  $S_{5,7}$  as shown in Figure 17. The center of  $S_{5,7}$  consists of four vertices, labeled  $C$ . Each tail is assigned to the diagonally opposite outer edges, for example, the tail which is on the left half of the graph and whose tip points upward is assigned to the outer edges at the bottom of the right half of the graph.

The three numbers corresponding (according to Definition 4) to the center-tail system  $S_{m,n}$ ,  $m < n$ ,  $m$  and  $n$  are odd, are (i)  $(m - 1)/2 + (m - 1)/2 + 1 = m$ ; (ii)  $m + 1$ ; and (iii)  $m + 1$ . Hence,  $CI(S_{m,n}) = m$  in the described case. Then by Theorem 5,  $s(R_{m,n}) \geq m$ . On the other hand, by Theorem 7, the values of the absolute indices (see Figure 17 for the absolute indices in the case  $R_{5,7}$ ) imply that  $s(R_{m,n}) \leq m$ . Thus  $s(R_{m,n}) = m$  if both  $m$  and  $n$  are odd and  $m < n$ .



**Figure 17.** Left:  $R_{5,7}$  with center-tail system  $S_{5,7}$ . Right: absolute indices for  $R_{5,7}$  and a minimum spanning congestion tree for  $R_{5,7}$ .

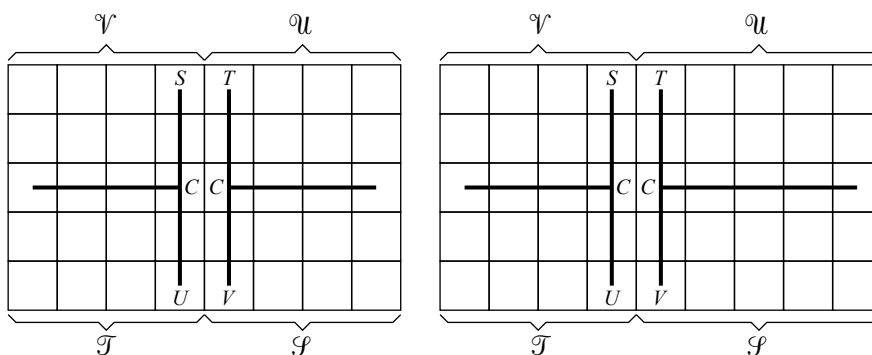


**Figure 18.**  $R_{5,8}$  with center-tail system  $S_{5,8}$ .

*Subcase 2:  $m$  is odd and  $n$  is even.* An instructive example of this type is shown in Figure 18.

We assign the tail pointing upward and downward to the outer edges at the bottom and the top, respectively. Assignments of the tails to the vertical outer edges are the same as before. It is easy to see that the congestion indicator of this center-tail system is equal to 5 in the case of  $R_{5,8}$  and  $m$  in general. Also, it is easy to see that computing the absolute indices (as in Figure 17), we get that  $s(R_{m,n}) = m$  in Subcase 2.

*Case 2:  $m$  is even.* As instructive examples, we consider the cases  $R_{6,9}$  and  $R_{6,10}$  (see Figure 19). The center-tail systems  $S_{6,9}$  and  $S_{6,10}$  are also shown in Figure 19. The centers of  $S_{6,9}$  and  $S_{6,10}$  consist of two vertices, labeled  $C$ . The tails  $S, T, U, V$  are assigned to the outer edges in the regions  $\mathcal{S}, \mathcal{T}, \mathcal{U}, \mathcal{V}$ , respectively. Finally, the tail whose tip points to the left and the right is assigned to those outer edges on the right and the left, respectively. In this case, the three numbers defined in Definition 4 are (i)  $3 + 3 + 1 = 7$ , (ii)  $6 + 1 = 7$  and (iii)  $6 + 1 = 7$ . So  $CI(S_{6,9}) = 7$  and hence, by Theorem 5,  $s(R_{6,9}) \geq 7$  and  $s(R_{6,10}) \geq 7$ . By Theorem 7, we have  $s(R_{6,9}) \leq 7$  and  $s(R_{6,10}) \leq 7$ . Thus  $s(R_{6,9}) = s(R_{6,10}) = 7$ . Observe that the length of the longest side of the rectangular grid does not play an important role in this computation, since we assume  $n > m$ . It is clear that similar center-tail systems can be used to show that  $s(R_{m,n}) = m + 1$  for any even  $m$  and any  $n$  satisfying  $n > m$ . □



**Figure 19.** Left:  $R_{6,9}$  with center-tail system  $S_{6,9}$ . Right:  $R_{6,10}$  with center-tail system  $S_{6,10}$ .

## References

- [Bodlaender et al. 2011] H. L. Bodlaender, K. Kozawa, T. Matsushima, and Y. Otachi, “Spanning tree congestion of  $k$ -outerplanar graphs”, *Discrete Math.* **311**:12 (2011), 1040–1045. MR 2012d:05105 Zbl 1223.05017
- [Bodlaender et al. 2012] H. L. Bodlaender, F. V. Fomin, P. A. Golovach, Y. Otachi, and E. J. van Leeuwen, “Parameterized complexity of the spanning tree congestion problem”, *Algorithmica* **64**:1 (2012), 85–111. MR 2923963 Zbl 1253.68163
- [Bollobás and Leader 1991] B. Bollobás and I. Leader, “Edge-isoperimetric inequalities in the grid”, *Combinatorica* **11**:4 (1991), 299–314. MR 92m:51042 Zbl 0755.05045
- [Castejon et al. 2007] A. Castejon, E. Corbacho, and R. Vidal, “Árboles óptimos en rejillas”, pp. 79–86 in *Avances en matemática discreta en Andalucía* (Cádiz, 2007), edited by J. C. Camacho Moreno et al., Serv. Publ. Univ. Cádiz, Cádiz, 2007. MR 2401733 Zbl 1193.05057
- [Clark and Holton 1991] J. Clark and D. A. Holton, *A first look at graph theory*, World Scientific, Teaneck, NJ, 1991. MR 92i:05077 Zbl 0765.05001
- [Diestel 2000] R. Diestel, *Graph theory*, 2nd ed., Graduate Texts in Mathematics **173**, Springer, New York, 2000. MR 1743598 Zbl 0945.05002
- [Egawa and Inoue 1999] Y. Egawa and K. Inoue, “Radius of  $(2k-1)$ -connected graphs”, *Ars Combin.* **51** (1999), 89–95. MR 99j:05061 Zbl 0977.05040
- [Fekete and Kremer 2001] S. P. Fekete and J. Kremer, “Tree spanners in planar graphs”, *Discrete Appl. Math.* **108**:1-2 (2001), 85–103. MR 2003d:05048 Zbl 0969.68111
- [Harant 1993] J. Harant, “An upper bound for the radius of a 3-connected graph”, *Discrete Math.* **122**:1-3 (1993), 335–341. MR 94i:05046 Zbl 0787.05055
- [Harant and Walther 1981] J. Harant and H. Walther, “On the radius of graphs”, *J. Combin. Theory Ser. B* **30**:1 (1981), 113–117. MR 82e:05081 Zbl 0483.05039
- [Hruska 2008] S. W. Hruska, “On tree congestion of graphs”, *Discrete Math.* **308**:10 (2008), 1801–1809. MR 2009b:05069 Zbl 1151.05014
- [Iida 2007] T. Iida, “Order and radius of 3-connected graphs”, *AKCE Int. J. Graphs Comb.* **4**:1 (2007), 21–49. MR 2008c:05047 Zbl 1138.05018
- [Iida and Kobayashi 2006] T. Iida and K. Kobayashi, “Order and radius of  $(2k-1)$ -connected graphs”, *SUT Journal of Math.* **42**:2 (2006), 335–356. MR 2010a:05073 Zbl 1132.05017

- [Inoue 1996] K. Inoue, “Radius of 3-connected graphs”, *SUT Journal of Math.* **32**:1 (1996), 83–90. MR 97d:05089 Zbl 0890.05025
- [Jordan 1869] C. Jordan, “Sur les assemblages de lignes”, *J. Reine Angew. Math.* **70** (1869), 185–190. JFM 02.0344.01
- [Khuller et al. 1993] S. Khuller, B. Raghavachari, and N. Young, “Designing multi-commodity flow trees”, pp. 433–441 in *Algorithms and data structures* (Montréal, 1993), edited by F. Dehne et al., Lecture Notes in Comput. Sci. **709**, Springer, Berlin, 1993. MR 1257430 Zbl 0803.68087
- [Law and Ostrovskii 2010] H.-F. Law and M. I. Ostrovskii, “Spanning tree congestion: duality and isoperimetry, with an application to multipartite graphs”, *Graph Theory Notes N. Y.* **58** (2010), 18–26. MR 2012j:05238
- [Lovász 2007] L. Lovász, *Combinatorial problems and exercises*, 2nd ed., AMS Chelsea Publishing, Providence, RI, 2007. MR 2321240 Zbl 1120.05001
- [Löwenstein 2010] C. Löwenstein, *In the complement of a dominating set*, PhD thesis, Technische Universität, Ilmenau, 2010, Available at <http://tinyurl.com/LowensteinPhD>.
- [Nishizeki and Chiba 1988] T. Nishizeki and N. Chiba, *Planar graphs: theory and algorithms*, North-Holland Math. Studies **140**, North-Holland, Amsterdam, 1988. MR 89f:05068 Zbl 0647.05001
- [Ostrovskii 2004] M. I. Ostrovskii, “Minimal congestion trees”, *Discrete Math.* **285**:1-3 (2004), 219–226. MR 2005f:05111 Zbl 1051.05032
- [Ostrovskii 2010] M. I. Ostrovskii, “Minimum congestion spanning trees in planar graphs”, *Discrete Math.* **310**:6-7 (2010), 1204–1209. MR 2010m:05082 Zbl 1230.05112
- [Otachi 2011] Y. Otachi, “Designing low-congestion networks with structural graph theory”, *Interdiscip. Inform. Sci.* **17**:3 (2011), 197–216. MR 2012k:05371 Zbl 1247.68199
- [Otachi et al. 2010] Y. Otachi, H. L. Bodlaender, and E. J. van Leeuwen, “Complexity results for the spanning tree congestion problem”, pp. 3–14 in *Graph-theoretic concepts in computer science* (Zarós, 2010), edited by D. M. Thilikos, Lecture Notes in Comput. Sci. **6410**, Springer, Berlin, 2010. MR 2012c:05222 Zbl 05816303
- [Peleg 2000] D. Peleg, *Distributed computing: a locality-sensitive approach*, SIAM Monographs on Discrete Mathematics and Applications **5**, Society for Industrial and Applied Mathematics, Philadelphia, 2000. MR 2001g:68010 Zbl 0959.68042
- [Rosen et al. 2000] K. H. Rosen, J. G. Michaels, J. L. Gross, J. W. Grossman, and D. R. Shier (editors), *Handbook of discrete and combinatorial mathematics*, CRC Press, Boca Raton, FL, 2000. MR 2000g:05001 Zbl 1044.00002
- [Simonson 1987] S. Simonson, “A variation on the min cut linear arrangement problem”, *Math. Systems Theory* **20**:4 (1987), 235–252. MR 89g:05065 Zbl 0643.68094

Received: 2013-03-19      Revised: 2013-07-02      Accepted: 2013-07-03

hiufai.law@gmail.com

*Department of Mathematics, Universität Hamburg,  
55 Bunderstrasse, D-20146 Hamburg, Germany*

siulam.leung10@stjohns.edu

*Department of Mathematics and Computer Science,  
St. John's University, 8000 Utopia Parkway,  
Queens, NY 11439, United States*

ostrovsm@stjohns.edu

*Department of Mathematics and Computer Science,  
St. John's University, 8000 Utopia Parkway,  
Queens, NY 11439, United States*



# Convex and subharmonic functions on graphs

Matthew J. Burke and Tony L. Perkins

(Communicated by Ronald Gould)

We explore the relationship between convex and subharmonic functions on discrete sets. Our principal concern is to determine the setting in which a convex function is necessarily subharmonic. We initially consider the primary notions of convexity on graphs and show that more structure is needed to establish the desired result. To that end, we consider a notion of convexity defined on lattice-like graphs generated by normed abelian groups. For this class of graphs, we are able to prove that all convex functions are subharmonic.

## 1. Introduction

Classical analysis provides several equivalent definitions of a convex function, which have led to several nonequivalent concepts of a convex function on a graph. This is not the case for subharmonic functions, where there appears to be a consensus on how to define subharmonic functions on graphs. In the real variable counterpart, all convex functions are subharmonic. It is the aim of this paper to investigate this relationship in the discrete setting.

We show that in the setting of weighted graphs over a normed abelian group, one can prove analogs of some classical analysis theorems relating convexity to subharmonic functions. In particular: all convex functions are subharmonic (Theorem 13); for a fixed point  $a \in X$ , the distance function  $d(x, a)$  is convex (Lemma 15); and a set  $F$  is convex if and only if the distance function  $d(x, F) = \inf_{y \in F} d(x, y)$  is subharmonic (Propositions 14 and 17).

For a discrete set with metric, there is generally one straightforward way to define convex sets and convex functions on them. For completeness and ease of reference, we present these in Section 2. The definitions we give (or something equivalent to them) can be traced back at least to  $d$ -convexity [German et al. 1973; Soltan 1972] and  $d$ -convex functions [Soltan and Soltan 1979], and possibly much earlier. Graphs admit a natural metric — the length of the shortest path between two vertices — which leads to one notion of convexity on graphs studied in [Soltan

---

*MSC2010:* primary 26A51; secondary 31C20.

*Keywords:* convex, subharmonic, discrete, graphs.

1983; 1991]. The notion of  $d$ -convexity on graphs when  $d$  is the standard graph metric is equivalent to the more common notion of geodesic convexity [Cáceres et al. 2005; Farber and Jamison 1986].

Common to [Cáceres et al. 2005; Farber and Jamison 1986; Soltan 1983; 1991], one starts with a graph and then puts a convexity theory on it by using the graph metric. However, in Section 3 we show that convex sets and functions defined on graphs with respect to the graph metric extend well for some, but not all, properties.

Another approach taken in Section 4 is to allow the vertices themselves to have some underlying structure, for example, a normed abelian group, and force the edges to be compatible with this metric. In the setting of a normed abelian group there are many notions of a convex function (see [Kiselman 2004] and references therein). One introduced in [Kiselman 2004] provides a natural extension of geodesic convexity that makes use of the additional abelian group structure. In this setting, convex and subharmonic functions are of particular interest to image analysis, for example, [Kiselman 2004; 2005]. In this setting, we are able to prove theorems analogous to several standard results from classical analysis.

## 2. Fundamental concepts

We will always assume that a graph is locally finite.

**2.1. Convexity.** Let  $X$  be an at most countable set with a metric  $d$ , that is,

$$d: X \times X \rightarrow \mathbb{R},$$

with these properties:

- (i)  $d(x, y) \geq 0$  for all  $x, y \in X$  with  $d(x, y) = 0$  if and only if  $x = y$ .
- (ii)  $d(x, y) = d(y, x)$ .
- (iii)  $d(x, y) \leq d(x, z) + d(z, y)$ .

Traditionally, a set  $A$  is convex if for all points  $x, y \in A$  every point on the line segment connecting them is also in  $A$ . Notice that a point  $z$  is on the line segment connecting  $x, y \in A$  if and only if  $d(x, y) = d(x, z) + d(z, y)$ . Hence we take the following definitions:

For  $A \subset X$  define

$$c_1(A) = \{z \in X: d(x, y) = d(x, z) + d(z, y) \text{ for some } x, y \in A\}$$

(this gives  $c_1(A) = \emptyset$  when  $A = \emptyset$ ), and inductively set  $c_n(A) = c_1(c_{n-1}(A))$ . Note that  $0 = d(x, x) = d(x, x) + d(x, x)$ , hence  $A \subseteq c_1(A) \subseteq \cdots \subseteq c_n(A)$  for all  $n$ .

**Definition 1.** Let  $A \subset X$ . The *convex hull* of  $A$  is

$$\text{cvx}(A) = \bigcup_{n=1}^{\infty} c_n(A).$$

Naturally, the set  $A$  is said to be *convex* if  $\text{cvx}(A) = A$ . Clearly  $\emptyset$  and  $X$  are convex.

We say that the point  $z$  is *in between*  $x$  and  $y$  whenever  $d(x, y) = d(x, z) + d(z, y)$  is satisfied.

**Lemma 2.** A set  $A \subset X$  is convex if and only if  $A = c_1(A)$ .

*Proof.* If  $A = c_1(A)$  then  $c_2(A) = c_1(c_1(A)) = c_1(A) = A$ . Hence by induction  $c_n(A) = A$  and so  $A = \bigcup c_n(A) = \text{cvx}(A)$ . Thus  $A$  is convex.

Suppose that  $A$  is convex. Then  $A = \text{cvx}(A) = \bigcup c_n(A) \supset c_1(A) \supset A$ . Thus  $A = c_1(A)$ .  $\square$

**Proposition 3.** For all sets  $A, B \subset X$ ,

$$A \subset \text{cvx}(A), \tag{1}$$

$$A \subset B \Rightarrow \text{cvx}(A) \subset \text{cvx}(B), \tag{2}$$

$$\text{cvx}(A) = \text{cvx}(\text{cvx}(A)). \tag{3}$$

*Proof.* (1) We've already shown that  $A \subset c_1(A) \subset \dots \subset c_n(A)$  for all  $n$  and so  $A \subset \bigcup c_n(A) = \text{cvx}(A)$ .

(2) For any sets  $X$  and  $Y$ , if  $X \subset Y$  then  $c_1(X) \subset c_1(Y)$ . Indeed for any  $z \in c_1(X)$  there exists by definition  $x_1, x_2 \in X$  so that  $d(x_1, x_2) = d(x_1, z) + d(z, x_2)$ , but as  $x_1, x_2 \in X \subset Y$  this shows that  $z \in c_1(Y)$ . Then as  $A \subset B$ , we have  $c_1(A) \subset c_1(B)$ . Then by induction,  $c_n(A) \subset c_n(B)$ . Therefore  $\text{cvx}(A) \subset \text{cvx}(B)$ .

(3) The claim  $\text{cvx}(A) = \text{cvx}(\text{cvx}(A))$  amounts to saying that  $\text{cvx}(A)$  is convex. We will use Lemma 2 to show this. Consider any  $z \in c_1(\text{cvx}(A))$ . This means there exists  $x, y \in \text{cvx}(A) = \bigcup c_n(A)$  so that  $d(x, y) = d(x, z) + d(z, y)$ . However, as  $A \subset c_1(A) \subset c_2(A) \subset \dots \subset c_n(A) \subset \dots$  we know  $x, y \in c_n(A)$  for some  $n$ , and so  $z \in c_1(c_n(A)) = c_{n+1}(A) \subset \text{cvx}(A)$ . Hence  $c_1(\text{cvx}(A)) = \text{cvx}(A)$ .  $\square$

The following proposition shows that our definition of convex hull is equivalent to the usual one, that is, the convex hull of  $A$  is the intersection of all convex sets that contain  $A$ .

**Proposition 4.** For any  $A \subset X$ , the set  $\text{cvx}(A)$  is the intersection of all convex sets that contain  $A$ .

*Proof.* Let  $B \subset X$  be a convex set containing  $A$ . As noted previously,  $A \subset B$  implies  $\text{cvx}(A) \subset \text{cvx}(B)$ . However,  $\text{cvx}(B) = B$  by hypothesis. Hence,  $\text{cvx}(A) \subset B$  for all convex  $B$  containing  $A$ . Therefore

$$\text{cvx}(A) \subset \bigcap \{B : A \subset B \text{ and } B \text{ convex}\}.$$

As  $\text{cvx}(A)$  is convex and  $A \subset \text{cvx}(A)$ , it must be included in the intersection above. Thus

$$\bigcap \{B : A \subset B \text{ and } B \text{ convex}\} \subset \text{cvx}(A). \quad \square$$

**Proposition 5.** *If  $A$  and  $B$  are convex, then  $A \cap B$  is convex.*

*Proof.* Let  $A$  and  $B$  be convex. Then by Lemma 2,  $A = c_1(A)$  and  $B = c_1(B)$ . We will show that  $c_1(A \cap B) = c_1(A) \cap c_1(B) = A \cap B$ . We've already noted that  $A \cap B \subset c_1(A \cap B)$ .

Suppose that  $z \in c_1(A \cap B)$ . Then there exists  $x, y \in A \cap B$  such that  $d(x, y) = d(x, z) + d(z, y)$ . Hence  $z \in c_1(A)$  and  $z \in c_1(B)$ , that is,  $z \in c_1(A) \cap c_1(B)$ . As  $A = c_1(A)$  and  $B = c_1(B)$ , we now have  $z \in c_1(A) \cap c_1(B) = A \cap B$ . Therefore  $c_1(A \cap B) \subset A \cap B$ . Thus  $A \cap B = c_1(A \cap B)$ , and so  $A \cap B$  is convex.  $\square$

**Proposition 6.** *Let  $I$  be an ordered set and take  $\{A_\alpha\}, \alpha \in I$  to be a collection of convex sets in  $X$  where  $A_\alpha \subset A_\beta$  whenever  $\alpha < \beta$  and  $\alpha, \beta \in I$ . The set formed by taking the union of  $A_\alpha$  for  $\alpha \in I$  is convex.*

*Proof.* We must show that  $\bigcup A_\alpha$  is convex. Consider the set  $c_1(\bigcup A_\alpha)$ . For any  $z \in c_1(\bigcup A_\alpha)$ , we can find  $x, y \in \bigcup A_\alpha$  so that  $d(x, y) = d(x, z) + d(z, y)$ . However,  $x, y \in \bigcup A_\alpha$  implies that  $x \in A_\alpha$  and  $y \in A_\beta$  for some  $\alpha, \beta \in I$ . Without loss of generality, we assume that  $\alpha < \beta$ . By hypothesis,  $A_\alpha \subset A_\beta$ . Hence  $x, y \in A_\beta$ . Since  $z$  satisfies  $d(x, y) = d(x, z) + d(z, y)$  for  $x, y \in A_\beta$  with  $A_\beta$  convex, we see that  $z \in c_1(A_\beta) = A_\beta$ . As  $z$  was arbitrarily chosen from  $c_1(\bigcup A_\alpha)$ , we have  $c_1(\bigcup A_\alpha) \subset \bigcup A_\alpha$ .

By construction the reverse inclusion  $\bigcup A_\alpha \subset c_1(\bigcup A_\alpha)$  is immediate. Hence  $c_1(\bigcup A_\alpha) = \bigcup A_\alpha$ . Recall from Lemma 2 that a set  $A$  is convex if and only if  $A = c_1(A)$ . Therefore  $\bigcup A_\alpha$  is convex.  $\square$

**Definition 7.** Let  $A$  be a convex set. A function  $f: A \rightarrow \mathbb{R}$  is *convex at the point*  $z \in A$  if

$$f(z) \leq \frac{d(y, z)}{d(x, y)} f(x) + \frac{d(x, z)}{d(x, y)} f(y)$$

whenever  $z$  is in between  $x, y \in A$ , that is,  $d(x, y) = d(x, z) + d(z, y)$ . A function is said to be *convex on*  $A$  if it is convex at every point in  $A$ . Furthermore, a function is simply called *convex* when it is convex on the entire set  $X$ .

The vertices of a graph admit a natural metric defined as the length of the shortest path between them. With this, the notions of convex and convex functions extend naturally to all graphs; see [Cáceres et al. 2005; Farber and Jamison 1986; Soltan 1983; 1991].

**2.2. Subharmonic functions on a graph.** Introductions to various aspects of the theory can be found in [Biyikoğlu et al. 2007; Kiselman 2005; Soardi 1994; Woess 1994].

Consider a graph  $G$ . The vertices of this graph will be denoted  $X$  (to stay consistent with above), which shall be the domain of our (sub)harmonic functions. A function  $f : X \rightarrow \mathbb{R}$  is said to be *harmonic* at  $x \in X$  if

$$f(x) = \frac{1}{\deg(x)} \sum_{y \sim x} f(y),$$

and subharmonic at  $x \in X$  if

$$f(x) \leq \frac{1}{\deg(x)} \sum_{y \sim x} f(y),$$

where  $\deg(x)$  denotes the degree of  $x$  and  $y \sim x$  means that  $y$  is adjacent to  $x$ . A function is (sub)harmonic if it is (sub)harmonic at every point  $x \in X$ . Observe that constant functions are always harmonic (thereby subharmonic too), and so these classes of functions are never empty.

**Lemma 8.** *If the graph  $X$  is connected, regular of degree two and triangle free, then a subharmonicity is the same as convexity.*

*Proof.* Each vertex  $z$  has only two neighbors  $x, y$ . As the graph is triangle free, we have  $d(x, y) = 2$ . Hence

$$\frac{1}{\deg(z)} \sum_{\zeta \sim z} f(\zeta) = \frac{1}{2}(f(x) + f(y)) = \frac{d(y, z)}{d(x, y)} f(x) + \frac{d(x, z)}{d(x, y)} f(y).$$

By definition  $f$  is subharmonic at  $z$  if  $f(z)$  is less than or equal to the left side of the equation above, and  $f$  is convex at  $z$  if  $f(z)$  is less than or equal to the right side of the equation above. Therefore subharmonicity and convexity are equivalent when these conditions are met.  $\square$

We will also use a standard modification of the definition of subharmonic functions on graphs to allow for positive edge weights. Namely, a function  $f : X \rightarrow \mathbb{R}$  is subharmonic at  $x$  if

$$0 \leq \sum_{y \sim x} e(x, y)[f(y) - f(x)],$$

which with some arithmetic becomes

$$f(x) \leq \frac{1}{M_x} \sum_{y \sim x} e(x, y) f(y),$$

where  $e(x, y) = e(y, x) \geq 0$  is the edge weight and  $M_x = \sum_{y \sim x} e(x, y)$ . If the edge weights are all taken to be one, then this definition is identical to the first.

### 3. The distance is given by the graph metric

In this section we provide two simple theorems which show that for a large class of graphs, convex functions are indeed subharmonic.

**Theorem 9.** *Let  $z$  be a point in  $X$ . Suppose that  $\deg(z) > 1$  and that  $z$  is not part of any triangle. If  $f$  is convex at  $z$ , then  $f$  is subharmonic at  $z$ . Consequently, if the graph has no triangles or vertices of degree less than 2, then every convex function is subharmonic.*

*Proof.* Let  $B = \{y \in X : y \sim z\}$  be all the vertices adjacent to  $z$ . By hypothesis, we have  $\deg(z) = |B| > 1$ , and so there are at least two vertices  $y_1, y_2 \in B$ . As  $z$  is adjacent to both  $y_1$  and  $y_2$  and as  $z$  is assumed to not be a part of a triangle,  $y_1$  is not adjacent to  $y_2$ . Hence  $z$  is in between  $y_1$  and  $y_2$ , that is, on a geodesic connecting  $y_1$  and  $y_2$ . In fact  $2 = d(y_1, y_2) = d(y_1, z) + d(z, y_2)$ , with  $d(y_1, z) = d(z, y_2) = 1$ . Hence, for all  $y_1, y_2 \in B$ , we have

$$2f(z) \leq f(y_1) + f(y_2) \quad (4)$$

by convexity.

Now we sum the inequality (4) over all unordered pairs of points  $y_1, y_2 \in B$ . Naturally, there are  $\binom{\deg(z)}{2}$  such pairs and each vertex  $y \in B$  will appear precisely  $\deg(z) - 1$  times. (Recall  $B = \{y : y \sim z\}$  and so  $|B| = \deg(z)$ .) Hence

$$\binom{\deg(z)}{2} 2f(z) \leq (\deg(z) - 1) \sum_{y \sim z} f(y),$$

which simplifies to

$$f(z) \leq \frac{1}{\deg(z)} \sum_{y \sim z} f(y).$$

Thus  $f$  is subharmonic at  $z$ . □

**Theorem 10.** *Let  $z$  be a point in  $X$ . If the neighbors of  $z$  can be partitioned into pairs such that the vertices in each pair are nonadjacent, then a function being convex at  $z$  implies that it is also subharmonic at  $z$ .*

*Proof.* For any vertices  $y_1, y_2$  in a pairing of the partition of the neighbors of  $z$  that are nonadjacent, the vertex  $z$  must be between them, and hence,

$$2f(z) \leq f(y_1) + f(y_2)$$

for any function  $f$  subharmonic at  $z$ . Consequently, if we sum this inequality over all  $\deg(z)/2$  pairings, we have

$$2 \frac{\deg(z)}{2} f(z) \leq \sum_{y \sim z} f(y).$$

Therefore  $f$  is subharmonic at  $z$ . □

Notice that for the standard square lattice, both theorems imply that a convex function is subharmonic. If  $z$  was connected to an odd number of nonadjacent points, then only the first theorem implies that a function convex at  $z$  is subharmonic at  $z$ . Similarly, when the graph is the standard triangular tiling of the plane, only the second theorem would show that every convex function is subharmonic.

**Theorem 11.** *Let  $F$  be any subset of  $X$ . If the distance function*

$$d(\cdot, F) := \inf\{d(\cdot, f) : f \in F\}$$

*is convex, then  $F$  is convex.*

*Proof.* Consider any point  $z \in X$  that lies between  $x, y \in F$ . If the distance function is convex, we have

$$0 \leq d(z, F) \leq \frac{d(y, z)}{d(x, y)} d(x, F) + \frac{d(x, z)}{d(x, y)} d(y, F),$$

but  $d(x, F) = d(y, F) = 0$  as  $x, y \in F$ . Therefore  $d(z, F) = 0$ , and so  $z$  must also be a point in  $F$ . □

**Example 12.** Consider a cycle on four vertices, that is,  $X = \{a, x, y, z\}$  with  $a \sim x$ ,  $x \sim y$ ,  $y \sim z$ ,  $z \sim a$ . One would easily believe that  $F = \{a\}$  is convex. Hence  $d(x, F) = d(z, F) = 1$ , and  $y$  is in between  $x$  and  $z$ . However

$$2 = d(y, a) \not\leq \frac{1}{2}d(x, a) + \frac{1}{2}d(z, a) = 1.$$

Hence  $d(\cdot, a)$  is not convex and certainly not subharmonic.

Observe also the set  $\{x, y, z\}$  is *not* convex. We believe this reveals part of the problem with this definition of convexity. Namely, a geodesic line segment need not be convex. It seems that few graphs have convex geodesics. (However  $X = \mathbb{Z}$ , with  $x \sim y$  when  $|x - y| = 1$ , and the standard triangular tiling of the plane are two such graphs.)

It would seem that more structure is needed to have a workable theory.

#### 4. Graphs over a normed abelian group

For the remainder of this paper, we consider weighted graphs where the vertex set  $X$  is a normed abelian group and the graph is compatible with the norm. We will denote the norm  $\|\cdot\|$ . We say that the graph structure is *compatible with the norm* if there is a constant  $r > 0$  such that  $x \sim y$  if and only if  $\|x - y\| \leq r$  and the edge weights are given by the norm  $e(x, y) = \|x - y\| \leq r$ .

In particular, graphs of this type include all lattice graphs. By rescaling  $X$  by  $r$  we can always assume without loss of generality that  $r = 1$ .

Graphs of this type pick up a number of traits from analysis. One such trait is a local similarity property. When one does analysis in a domain  $D \subset \mathbb{R}^n$  (or on a manifold) every point  $z \in D$  has a neighborhood which is locally like a ball in  $\mathbb{R}^n$ . We see the same property here.

This can also be viewed as a translation invariance property; we could translate any point  $x_0$  to the origin by taking  $X \mapsto X - x_0$  and nothing would change. More explicitly, we denote  $B_r(x_0) := \{y \in X : y \sim x_0\}$ , and for every  $x_0$  in  $X$  there is a simple one-to-one correspondence between  $B_r(x_0)$  and  $B_r(0)$ . If  $y \in B_r(x_0)$ , then  $z = y - x_0 \in B_r(0)$ , and if  $z \in B_r(0)$ , then  $x_0 + z \in B_r(x_0)$ .

Furthermore, if  $\zeta \in B_r(0)$ , then  $-\zeta \in B_r(0)$ . Hence

$$\{y \in X : y \sim x\} := B_r(x) = \{x + \zeta : \zeta \in B_r(0)\} = \{x - \zeta : \zeta \in B_r(0)\}. \quad (5)$$

We maintain the same notion of a convex function, namely

$$\|x - y\|f(z) \leq \|y - z\|f(x) + \|x - z\|f(y),$$

whenever  $\|x - y\| = \|x - z\| + \|z - y\|$ . However in this context we can work with midpoints.

Kiselman [1996] defines a function  $f$  on an abelian group  $X$  to be *midpoint convex* if

$$f(x) \leq \frac{1}{2}f(x + z) + \frac{1}{2}f(x - z)$$

for all  $x$  and  $z$  in  $X$ . (Actually he uses the notion of upper addition for functions defined on the extended real line, that is,  $\mathbb{R} \cup \{\pm\infty\}$ , but we will not be needing such subtleties here.) Trivially a convex function is always midpoint convex.

We will now see that this notion of midpoint convexity allows us to achieve our goals.

**Theorem 13.** *Consider a weighted graph where the vertex set  $X$  is a normed abelian group and the graph is compatible with the norm. Every midpoint convex function is subharmonic.*

*Proof.* Pick any  $x \in X$ . Observe that by (5)

$$\begin{aligned} \sum_{y \sim x} e(x, y)f(y) &= \frac{1}{2} \sum_{z \in B_r(0)} e(x, x + z)f(x + z) + \frac{1}{2} \sum_{z \in B_r(0)} e(x, x - z)f(x - z) \\ &= \sum_{z \in B_r(0)} e(x, x + z)\left(\frac{1}{2}f(x + z) + \frac{1}{2}f(x - z)\right). \end{aligned}$$

Hence, by (midpoint) convexity

$$f(x)M_x = f(x) \sum_{z \in B_r(0)} e(x, x + z) \leq \sum_{y \sim x} e(x, y)f(y),$$

which shows that  $f$  is subharmonic at  $x$ . □



A set  $A \subset X$  is called *convex* if the function

$$\mathcal{J}_A(x) = \begin{cases} 0 & \text{for } x \in A, \\ +\infty & \text{for } x \in X \setminus A \end{cases}$$

is convex, or, equivalently, if  $z \in A$  whenever there exists  $x, y \in A$  such that  $\|x - y\| = \|x - z\| + \|z - y\|$ . This again easily implies midpoint convexity, that is, if  $z \in A$  whenever there is an  $x \in X$  such that both  $z + x$  and  $z - x$  are in  $A$ .

**Proposition 14.** *Let  $F$  be any subset of  $X$ . If the distance function*

$$d(x, F) = \inf\{\|x - y\| : y \in F\}$$

*is convex, then the set  $F$  is convex.*

*Proof.* Let  $x \in X$  so that there is some  $z \in X$  with  $x \pm z \in F$ . Then by midpoint convexity

$$0 \leq d(x, F) \leq \frac{1}{2}d(x + z, F) + \frac{1}{2}d(x - z, F) = 0.$$

Therefore  $d(x, F) = 0$  and so  $x \in F$ . □

Notice that for the simple case  $F = \{a\}$  we get the converse of the previous result.

**Lemma 15.** *For any fixed  $a \in X$ , the function  $f(z) = \|z - a\|$  is midpoint convex.*

*Proof.* This follows immediately from the triangle inequality on the norm. Indeed, for any  $x, y, z \in X$  with  $\|x - y\| = \|x - z\| + \|z - y\|$  we have

$$\begin{aligned} 2f(x) &= 2\|x - a\| = \|2(x - a)\| = \|(x - a) - z + (x - a) + z\| \\ &\leq \|(x - a) - z\| + \|(x - a) + z\| = f(x - z) + f(x + z). \end{aligned} \quad \square$$

The minimum of two convex functions is in general not a convex function, which is one reason why the following result is interesting.

However, in general the classical proofs rely heavily upon the fact that for any point  $x$  and convex set  $F$  there is always a unique nearest neighbor  $y \in F$  to  $x$ .

**Definition 16.** We say that a set  $F$  has the *nearest neighbor* property if for all  $y_1, y_2 \in F$  and  $z \in X$  there exists a  $y \in F$  (possibly  $y_1$  or  $y_2$ ) such that

$$2\|y - z\| \leq \|y_1 + y_2 - 2z\|.$$

**Proposition 17.** *If  $F$  is a convex subset of  $X$  with the nearest neighbor property, then the distance function  $d(\cdot, F)$  is midpoint convex (and hence subharmonic).*

*Proof.* Pick any  $z \in X \setminus F$ . We will show that  $d(\cdot, F)$  is midpoint convex at  $z$ . By replacing  $F$  with  $F - z$  we may assume without loss of generality that  $z = 0$ .

Clearly it is possible for there to be an  $x \in B_r(0)$  such that  $d(x, F) \leq d(0, F)$ . However, by switching to normed abelian groups we've a strong property to use.

Namely, if  $x \in B_r(0)$  then  $-x \in B_r(0)$ . We will show for convex sets with the nearest neighbor property that

$$2d(0, F) \leq d(x, F) + d(-x, F),$$

that is,  $d(\cdot, F)$  is midpoint convex (and hence subharmonic).

We can find  $y_1, y_2 \in F$  such that  $d(x, F) = \|x - y_1\|$  and  $d(-x, F) = \|(-x) - y_2\|$ . Let  $y$  be a point in  $F$  such that  $2\|y\| \leq \|y_1 + y_2\|$ . Then

$$\begin{aligned} 2d(0, F) &\leq 2\|y\| \leq \|y_1 + y_2\| = \|y_1 + y_2 + x - x\| = \|(y_1 - x) + (y_2 + x)\| \\ &\leq \|y_1 - x\| + \|y_2 + x\| = d(x, F) + d(-x, F). \quad \square \end{aligned}$$

## References

- [Biyikoğlu et al. 2007] T. Biyikoğlu, J. Leydold, and P. F. Stadler, *Laplacian eigenvectors of graphs: Perron–Frobenius and Faber–Krahn type theorems*, Lecture Notes in Mathematics **1915**, Springer, Berlin, 2007. MR 2009a:05119 Zbl 1129.05001
- [Cáceres et al. 2005] J. Cáceres, A. Márquez, O. R. Oellermann, and M. L. Puertas, “Rebuilding convex sets in graphs”, *Discrete Math.* **297**:1-3 (2005), 26–37. MR 2006e:05053 Zbl 1070.05035
- [Farber and Jamison 1986] M. Farber and R. E. Jamison, “Convexity in graphs and hypergraphs”, *SIAM J. Algebraic Discrete Methods* **7**:3 (1986), 433–444. MR 87i:05166 Zbl 0591.05056
- [German et al. 1973] L. F. German, V. P. Soltan, and P. S. Soltan, “Certain properties of  $d$ -convex sets”, *Dokl. Akad. Nauk SSSR* **212** (1973), 1276–1279. In Russian; translated in *Sov. Math. Dokl.* **14** (1973), 1566–1570. MR 48 #12296 Zbl 0295.52010
- [Kiselman 1996] C. O. Kiselman, “Regularity of distance transformations in image analysis”, *Computer Vision and Image Understanding* **64**:3 (1996), 390–398.
- [Kiselman 2004] C. O. Kiselman, “Convex functions on discrete sets”, pp. 443–457 in *Combinatorial image analysis* (Auckland, 2004), edited by R. Klette and J. Žunić, Lecture Notes in Comput. Sci. **3322**, Springer, Berlin, 2004. MR 2166029 Zbl 1113.68585
- [Kiselman 2005] C. O. Kiselman, “Subharmonic functions on discrete structures”, pp. 67–80 in *Harmonic analysis, signal processing, and complexity* (Fairfax, VA, 2004), edited by I. Sabadini et al., Progr. Math. **238**, Birkhäuser, Boston, 2005. MR 2007c:31007 Zbl 1089.31004
- [Soardi 1994] P. M. Soardi, *Potential theory on infinite networks*, Lecture Notes in Mathematics **1590**, Springer, Berlin, 1994. MR 96i:31005 Zbl 0818.31001
- [Soltan 1972] P. S. Soltan, “Helly’s theorem for  $d$ -convex sets”, *Dokl. Akad. Nauk SSSR* **205** (1972), 537–539. In Russian; translated in *Sov. Math. Dokl.* **13** (1972), 975–978. MR 46 #8047 Zbl 0262.52006
- [Soltan 1983] V. P. Soltan, “ $d$ -convexity in graphs”, *Dokl. Akad. Nauk SSSR* **272**:3 (1983), 535–537. In Russian; translated in *Sov. Math. Dokl.* **28** (1983), 419–421. MR 85a:05077 Zbl 0553.05060
- [Soltan 1991] V. P. Soltan, “Metric convexity in graphs”, *Studia Univ. Babeş-Bolyai Math.* **36**:4 (1991), 3–43. MR 95k:52002 Zbl 0882.52001
- [Soltan and Soltan 1979] V. P. Soltan and P. S. Soltan, “ $d$ -convex functions”, *Dokl. Akad. Nauk SSSR* **249**:3 (1979), 555–558. In Russian; translated in *Sov. Math. Dokl.* **20** (1979), 1323–1326. MR 80j:52001 Zbl 0479.52002
- [Woess 1994] W. Woess, “Random walks on infinite graphs and groups: a survey on selected topics”, *Bull. London Math. Soc.* **26**:1 (1994), 1–60. MR 94i:60081 Zbl 0830.60061

Received: 2013-04-01    Revised: 2013-06-21    Accepted: 2013-07-05

mjburke@shc.edu

*Spring Hill College, 4000 Dauphin Street,  
Mobile, AL 36608-1791, United States*

tperkins@shc.edu

*Department of Mathematics, Spring Hill College,  
4000 Dauphin Street, Mobile, AL 36608-1791, United States*



# New results on an anti-Waring problem

Chris Fuller, David R. Prier and Karissa A. Vasconi

(Communicated by Nigel Boston)

The number  $N(k, r)$  is defined to be the first integer such that it and every subsequent integer can be written as the sum of the  $k$ -th powers of  $r$  or more distinct positive integers. For example, it is known that  $N(2, 1) = 129$ , and thus the last number that cannot be written as the sum of one or more distinct squares is 128. We give a proof of a theorem that states if certain conditions are met, a number can be verified to be  $N(k, r)$ . We then use that theorem to find  $N(2, r)$  for  $1 \leq r \leq 50$  and  $N(3, r)$  for  $1 \leq r \leq 30$ .

## 1. Introduction

In 1770, Waring conjectured that for each positive integer  $k$  there exists a  $g(k)$  such that every positive integer is a sum of  $g(k)$  or fewer  $k$ -th powers of positive integers. After Hilbert proved this theorem true in 1909, the challenge that became known as Waring's problem was the question that asks, for each  $k$ , what is the smallest  $g(k)$  such that the statement holds. For more information on Waring's problem, see [Weisstein].

Recently, two papers have tackled the following "anti-Waring" conjecture: *If  $k$  and  $r$  are positive integers, then every sufficiently large positive integer is the sum of  $r$  or more  $k$ -th powers of distinct positive integers.*

The fact that there must be  $r$  or more  $k$ -th powers motivated the choice of the designation *anti-Waring* in [Johnson and Laughlin 2011], where the conjecture was put forth. What sets this statement apart from Waring's problem is the word "distinct". The conjecture was later proved in [Looper and Saritzky 2012]. A natural anti-Waring problem arising from this proven conjecture is to find the smallest integer  $N(k, r)$  such that it and every subsequent integer can be written as the sum of  $r$  or more  $k$ -th powers of distinct positive integers. Johnson and Laughlin proved that  $N(2, 1) = N(2, 2) = N(2, 3) = 129$ .

The following results are restricted to the case when  $k = 2$  and  $k = 3$ .  $N(2, r)$  is the smallest integer such that it and every subsequent integer can be written as the

---

MSC2010: 11A67.

Keywords: number theory, Waring, anti-Waring, series.

sum of  $r$  or more distinct squares.  $N(2, r)$  has been found for  $1 \leq r \leq 50$ .  $N(3, r)$  is the smallest integer such that it and every subsequent integer can be written as the sum of  $r$  or more distinct cubes.  $N(3, r)$  has been found for  $1 \leq r \leq 30$ . For the purposes of this paper we use two definitions.

**Definitions.** An integer is  $(k, r)$ -good if it *can* be written as the sum of  $r$  or more  $k$ -th powers of distinct positive integers. An integer is  $(k, r)$ -bad if it *cannot* be written as the sum of  $r$  or more  $k$ -th powers of distinct positive integers.

To see an example of this idea, consider the case when  $k = 2$  and  $r = 4$ . Since 129 can be written as  $2^2 + 3^2 + 4^2 + 10^2$ , 129 is  $(2, 4)$ -good. However, it is a brief exercise to verify that there is no way to write 128 as the sum of four or more distinct squares, and hence 128 is  $(2, 4)$ -bad. The fact that 129 is  $(2, 4)$ -good also directly implies that it is  $(2, r)$ -good for any integer  $1 \leq r \leq 4$ . Using these definitions, the problem of finding  $N(2, r)$  can be reworded to be the problem of finding the first  $(2, r)$ -good integer such that every subsequent integer is also  $(2, r)$ -good. In the case when  $r = 4$ , the fact that 128 is  $(2, 4)$ -bad implies that  $N(2, 4) \geq 129$ .

As will be seen, an inductive argument used in the following theorems requires a consecutive list of  $(k, r)$ -good integers whose size grows as  $r$  does. Computer software was used to attain these large lists of  $(k, r)$ -good integers as well as to verify that certain key integers are in fact  $(k, r)$ -bad.

## 2. Results

Before stating the general result of this paper, it may be helpful to offer a less general theorem and proof that will serve as valuable context for Theorem 2.2.

**Theorem 2.1.**  $N(2, 4) = 129$ .

*Proof.* As shown previously,  $N(2, 4) \geq 129$ . It is also true that the consecutive integers  $\{129, \dots, 18^2\}$  are  $(2, 4)$ -good. Therefore, if  $n \leq 18^2$  and  $n$  is  $(2, 4)$ -bad, then  $n \leq 128$ . The rest of the proof continues by induction on  $m$  with  $m \geq 18$ .

The induction statement: If  $n \leq m^2$  and  $n$  is  $(2, 4)$ -bad, then  $n \leq 128$ . If  $m = 18$ , the statement is clearly true as we know the consecutive integers  $\{129, \dots, 18^2\}$  are  $(2, 4)$ -good.

Now suppose  $n \leq (m + 1)^2$  and  $n$  is  $(2, 4)$ -bad. If  $n \leq m^2$ , then by the induction hypothesis,  $n \leq 128$ . Thus we can say

$$(m + 1)^2 \geq n \geq m^2 + 1. \quad (1)$$

Consider the integer  $n - (m - 4)^2$ . From (1) and the fact that  $m \geq 18$ , we know that

$$m^2 \geq n - (m - 4)^2 \geq m^2 + 1 - (m - 4)^2 \geq 129. \quad (2)$$

To see that  $n - (m - 4)^2$  is  $(2, 4)$ -bad, suppose that it is  $(2, 4)$ -good and hence

$$n - (m - 4)^2 = a_1^2 + a_2^2 + \dots + a_t^2 \quad \text{with } t \geq 4, a_i \neq a_j \text{ for all } i \text{ and } j,$$

or

$$n = a_1^2 + a_2^2 + \dots + a_t^2 + (m - 4)^2.$$

Since  $n$  is  $(2, 4)$ -bad, there is some  $j \in \{1, 2, \dots, t\}$  such that  $a_j = (m - 4)$ . Therefore

$$n - (m - 4)^2 \geq 1^2 + 2^2 + 3^2 + (m - 4)^2,$$

and equivalently,  $n - m^2 \geq m^2 - 16m + 46$ .

Combining this with (1), we get

$$(m + 1)^2 \geq n \geq 2m^2 - 16m + 46$$

or

$$0 \geq m^2 - 18m + 45,$$

which is untrue when  $m \geq 18$ . Therefore  $n - (m - 4)^2$  must be  $(2, 4)$ -bad, and by (2) and the inductive hypothesis,  $n - (m - 4)^2 \leq 128$ . However, this is a contradiction since by (2) it is also true that  $n - (m - 4)^2 \geq 129$ , and thus there are no  $n$  that are  $(2, 4)$ -bad and satisfy (1).  $\square$

In Theorem 2.1, 129 was the expected result for  $N(2, 4)$  after using computer software to generate a long list of consecutive  $(2, 4)$ -good integers that began with 129. The aim of Theorem 2.2 is to offer a theorem such that under given conditions, expected results for  $N(k, r)$  can be proven for any positive integers  $k$  and  $r$ . To simplify the notation  $S_k(z)$  will be used to represent  $\sum_{i=1}^z i^k$ .

**Theorem 2.2.** *If the consecutive integers  $\{\hat{N}(k, r), \dots, b^k\}$  are all  $(k, r)$ -good,  $\hat{N}(k, r) - 1$  is  $(k, r)$ -bad, and if there exists an integer  $x$  such that*

- (i)  $0 < S_k(r - 1) + 2(m - x)^k - (m + 1)^k$  for all  $m \geq b$ ,
- (ii)  $(m + 1)^k - (m - x)^k \leq m^k$  for all  $m \geq b$ ,
- (iii)  $m^k + 1 - (m - x)^k \geq \hat{N}(k, r)$  for all  $m \geq b$ , and
- (iv)  $0 < x < b - r$ ,

then  $\hat{N}(k, r) = N(k, r)$ .

*Proof.* We use induction on  $m \in \mathbb{N}$  with  $m \geq b$ . The induction statement: If  $n \leq m^k$  and  $n$  is  $(k, r)$ -bad, then  $n \leq \hat{N}(k, r) - 1$ .

If  $m = b$ , the statement is clearly true as we know the consecutive integers  $\{\hat{N}(k, r), \dots, b^k\}$  are all  $(k, r)$ -good.

Now suppose  $n \leq (m+1)^k$  and  $n$  is  $(k, r)$ -bad. If  $n \leq m^k$ , then by the induction hypothesis,  $n \leq \hat{N}(k, r) - 1$ . Thus we can say

$$(m+1)^k \geq n \geq m^k + 1. \quad (3)$$

We will show that  $n$  cannot satisfy (3), and hence all cases have been addressed.

Consider the integer  $n - (m-x)^k$ . Using (3) and condition (iii), we know that

$$n - (m-x)^k \geq m^k + 1 - (m-x)^k \geq \hat{N}(k, r)$$

or

$$n - (m-x)^k \geq \hat{N}(k, r). \quad (4)$$

To see that  $n - (m-x)^k$  is  $(k, r)$ -bad, suppose it is  $(k, r)$ -good. Then

$$n - (m-x)^k = a_1^k + a_2^k + \cdots + a_t^k \quad \text{with } t \geq r, a_i \neq a_j \text{ for all } i \neq j,$$

or

$$n = a_1^k + a_2^k + \cdots + a_t^k + (m-x)^k.$$

Since  $n$  is  $(k, r)$ -bad,  $a_j = m-x$  for some  $j \in \{1, 2, \dots, t\}$ . This, along with condition (iv), implies that  $n - (m-x)^k \geq S_k(r-1) + (m-x)^k$ . Combining this with (3), we get

$$(m+1)^k \geq n \geq S_k(r-1) + 2(m-x)^k,$$

or

$$0 \geq S_k(r-1) + 2(m-x)^k - (m+1)^k.$$

This contradiction of condition (i) means  $n - (m-x)^k$  must be  $(k, r)$ -bad.

Now from (3) and condition (ii),

$$n - (m-x)^k \leq (m+1)^k - (m-x)^k \leq m^k.$$

Thus by the induction hypothesis,  $n - (m-x)^k \leq \hat{N}(k, r) - 1$ . This contradicts (4) and means that there are no  $n$  that are  $(k, r)$ -bad and satisfy (3).  $\square$

As a result of Theorem 2.2, in order to find  $N(k, r)$  one must simply find a suitable list of  $(k, r)$ -good consecutive integers  $\{\hat{N}(k, r), \dots, b^k\}$  such that  $\hat{N}(k, r) - 1$  is  $(k, r)$ -bad and an integer  $x$  that satisfies the four conditions of the theorem. It is this strategy that gives way to the tables of values in Theorems 2.3 and 2.4. Again, computer software was a valuable tool in determining whether a given number was  $(k, r)$ -good or  $(k, r)$ -bad for  $k \in \{2, 3\}$ . For each  $r$  in the following two theorems, corresponding values for  $x$  and  $b$  are listed in Tables 1 and 2 rather than in the proof of the theorem.

**Theorem 2.3.** *Table 1 is a list of  $N(2, r)$  for integers  $1 \leq r \leq 50$ .*



$r$	$N(2, r)$	$x$	$b$	$r$	$N(2, r)$	$x$	$b$	$r$	$N(2, r)$	$x$	$b$	$r$	$N(2, r)$	$x$	$b$
1	129	4	18	14	1398	19	47	27	7953	54	101	40	23679	100	169
2	129	4	18	15	1723	21	52	28	8677	57	105	41	25348	104	174
3	129	4	18	16	1991	24	54	29	9538	61	109	42	27208	108	180
4	129	4	18	17	2312	26	58	30	10394	63	114	43	29093	112	186
5	198	6	22	18	2673	28	62	31	11559	67	120	44	31229	116	193
6	238	6	23	19	3048	31	65	32	12603	71	125	45	33298	120	199
7	331	8	26	20	3493	34	69	33	13744	74	130	46	35290	123	205
8	383	9	27	21	4094	36	75	34	14864	78	135	47	37654	127	212
9	528	10	32	22	4614	39	79	35	16253	81	141	48	40043	132	218
10	648	12	33	23	5139	42	83	36	17529	85	146	49	42488	135	225
11	889	14	39	24	5719	44	87	37	18958	89	151	50	45024	140	231
12	989	15	41	25	6380	48	91	38	20482	92	158				
13	1178	17	44	26	7124	51	96	39	22043	96	163				

**Table 1.** For each  $r$  listed,  $N(2, r) - 1$  is  $(2, r)$ -bad, and the list of consecutive integers  $\{N(2, r), \dots, b^2\}$  is  $(2, r)$ -good. The three necessary conditions of Theorem 2.2 are satisfied by  $x$ .

*Proof.* For  $1 \leq r \leq 4$ ,  $N(2, r) = 129$  by [Johnson and Laughlin 2011] and Theorem 2.1. For each  $r$ ,  $N(2, r) - 1$  has been shown to be  $(2, r)$ -bad. There exist  $b$  and  $x$  such that the consecutive integers  $\{N(2, r), \dots, b^2\}$  are  $(2, r)$ -good, and  $x$  satisfies the four conditions of Theorem 2.2. □

**Theorem 2.4.** Table 2 is a list of  $N(3, r)$  for integers  $1 \leq r \leq 30$ .

*Proof.* For each  $r$ ,  $N(3, r) - 1$  has been shown to be  $(3, r)$ -bad. There exist  $b$  and  $x$  such that the consecutive integers  $\{N(3, r), \dots, b^3\}$  are  $(3, r)$ -good, and  $x$  satisfies the four conditions listed in Theorem 2.2. □

$r$	$N(3, r)$	$x$	$b$	$r$	$N(3, r)$	$x$	$b$	$r$	$N(3, r)$	$x$	$b$	$r$	$N(3, r)$	$x$	$b$
1	12759	5	32	9	16224	6	33	17	56076	11	47	25	179520	18	67
2	12759	5	32	10	18149	6	35	18	66534	12	50	26	201921	19	69
3	12759	5	32	11	22398	7	37	19	75912	12	52	27	227400	20	72
4	12759	5	32	12	24855	7	38	20	87567	13	54	28	256254	22	73
5	12759	5	32	13	28887	8	39	21	101093	14	56	29	289869	23	76
6	15279	6	33	14	36951	9	42	22	122064	15	60	30	325590	24	79
7	15279	6	33	15	39660	9	43	23	138696	16	62				
8	15279	6	33	16	49083	10	46	24	156498	17	64				

**Table 2.** For each  $r$  listed,  $N(3, r) - 1$  is  $(3, r)$ -bad, and the list of consecutive integers  $\{N(3, r), \dots, b^3\}$  is  $(3, r)$ -good. The three necessary conditions of Theorem 2.2 are satisfied by  $x$ .

### 3. Future work

The list of values of  $N(k, r)$  can be extended indefinitely for any value of  $k$ . Currently we are only limited by our computing speed. A natural direction for further research would be to attempt to find an explicit formula for  $N(k, r)$  for a specific  $k$ . In [Johnson and Laughlin 2011], it was noticed that  $N(1, r) = r(r+1)/2$ . However, we have not found a formula for  $N(2, r)$  or  $N(3, r)$ .

Another area that seems natural is to attempt to find  $N(k, r)$  for values of  $k$  greater than 3. We have attempted to use our current software to find  $N(4, 1)$  and  $N(5, 1)$ , but our methods appear to be too inefficient. At this point, all that can be said confidently is that  $N(4, 1)$  is greater than 4.3 million,  $N(5, 1)$  is greater than 26.25 million, and perhaps they are both much larger.

It is also clear that  $N(k, i) \leq N(k, j)$  when  $i \leq j$ , and it seems natural to conjecture that  $N(x, r) \leq N(y, r)$  when  $x \leq y$ . Since  $N(1, r) = (r(r+1))/2$ ,  $N(1, r) \leq S_k(r) \leq N(k, r)$  for any integer  $k \geq 1$ . However, it is possible for an integer that it is  $(k, r)$ -bad to be  $(l, r)$ -good with  $k < l$ . For example, 9 is  $(2, 2)$ -bad but  $(3, 2)$ -good. Thus, a proof of this conjecture eludes us currently.

**Note.** After finishing this paper, it was brought to our attention that [Deering and Jamieson] had recently been submitted for publication. This paper has some of the same results as ours. In particular, our method of discovering  $N(k, r)$ , with proof, is very much like that of Deering and Jamieson. However, we feel that our method is sufficiently different and easier to use to merit publication.

### References

- [Deering and Jamieson] J. Deering and W. Jamieson, “On anti-Waring numbers”, to appear in *J. Combin. Math. Combin. Comput.*
- [Johnson and Laughlin 2011] P. Johnson and M. Laughlin, “An anti-Waring conjecture and problem”, *Int. J. Math. Comput. Sci.* **6**:1 (2011), 21–26. MR 2012f:11013 Zbl 05954412
- [Looper and Saritzky 2012] N. Looper and N. Saritzky, “An anti-Waring theorem and proof”, presentation at the MAA undergraduate poster session, Boston, January 2012.
- [Weisstein] E. Weisstein, “Waring’s problem”, resource available at <http://mathworld.wolfram.com/WaringsProblem.html>.

Received: 2013-04-24      Revised: 2013-07-10      Accepted: 2013-07-24

cfuller@cumberland.edu

*Department of Mathematics, Cumberland University,  
Lebanon, TN 37087, United States*

prier001@gannon.edu

*Department of Mathematics, Gannon University,  
Erie, PA 16541-0001, United States*

vasconi002@gmail.com

*1440 Heinz Avenue, Sharon, PA 16146, United States*

## Guidelines for Authors

Authors may submit manuscripts in PDF format on-line at the Submission page at the *Involve* website.

**Originality.** Submission of a manuscript acknowledges that the manuscript is original and is not, in whole or in part, published or under consideration for publication elsewhere. It is understood also that the manuscript will not be submitted elsewhere while under consideration for publication in this journal.

**Language.** Articles in *Involve* are usually in English, but articles written in other languages are welcome.

**Required items.** A brief abstract of about 150 words or less must be included. It should be self-contained and not make any reference to the bibliography. If the article is not in English, two versions of the abstract must be included, one in the language of the article and one in English. Also required are keywords and subject classifications for the article, and, for each author, postal address, affiliation (if appropriate), and email address.

**Format.** Authors are encouraged to use  $\LaTeX$  but submissions in other varieties of  $\TeX$ , and exceptionally in other formats, are acceptable. Initial uploads should be in PDF format; after the refereeing process we will ask you to submit all source material.

**References.** Bibliographical references should be complete, including article titles and page ranges. All references in the bibliography should be cited in the text. The use of Bib $\TeX$  is preferred but not required. Tags will be converted to the house format, however, for submission you may use the format of your choice. Links will be provided to all literature with known web locations and authors are encouraged to provide their own links in addition to those supplied in the editorial process.

**Figures.** Figures must be of publication quality. After acceptance, you will need to submit the original source files in vector graphics format for all diagrams in your manuscript: vector EPS or vector PDF files are the most useful.

Most drawing and graphing packages (Mathematica, Adobe Illustrator, Corel Draw, MATLAB, etc.) allow the user to save files in one of these formats. Make sure that what you are saving is vector graphics and not a bitmap. If you need help, please write to [graphics@msp.org](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

2014

vol. 7

no. 2

An interesting proof of the nonexistence of a continuous bijection between $\mathbb{R}^n$ and $\mathbb{R}^2$ for $n \neq 2$	125
HAMID REZA DANESHPAJOUH, HAMED DANESHPAJOUH AND FERESHTE MALEK	
Analysing territorial models on graphs	129
MARIE BRUNI, MARK BROOM AND JAN RYCHTÁŘ	
Binary frames, graphs and erasures	151
BERNHARD G. BODMANN, BIJAN CAMP AND DAX MAHONEY	
On groups with a class-preserving outer automorphism	171
PETER A. BROOKSBANK AND MATTHEW S. MIZUHARA	
The sharp log-Sobolev inequality on a compact interval	181
WHAN GHANG, ZANE MARTIN AND STEVEN WARUHIU	
Analysis of a Sudoku variation using partially ordered sets and equivalence relations	187
ANA BURGERS, SHELLY SMITH AND KATHERINE VARGA	
Spanning tree congestion of planar graphs	205
HIU FAI LAW, SIU LAM LEUNG AND MIKHAIL I. OSTROVSKII	
Convex and subharmonic functions on graphs	227
MATTHEW J. BURKE AND TONY L. PERKINS	
New results on an anti-Waring problem	239
CHRIS FULLER, DAVID R. PRIER AND KARISSA A. VASCONI	



1944-4176(2014)7:2;1-7