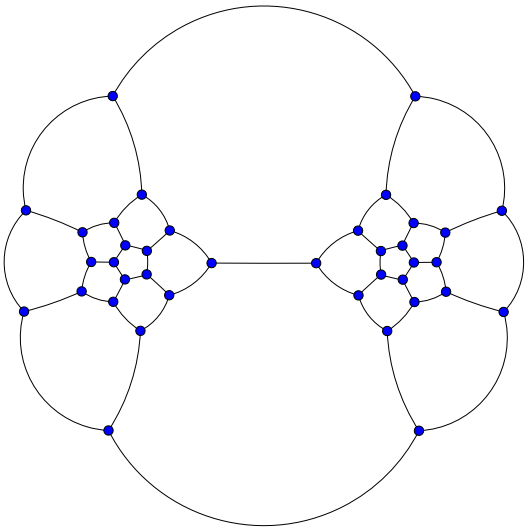
Tomislav Došlić, Ivica Martinjak (Eds.)

Proceedings

of the



2nd Croatian Combinatorial Days

Zagreb, September 27 – 28, 2018

Faculty of Civil Engineering, University of Zagreb

Editors: Tomislav Došlić, Ivica Martinjak

Publisher: Faculty of Civil Engineering
University of Zagreb

Copies: 55

Zagreb, 2019

ISBN 978-953-8168-33-8

DOI: https://doi.org/10.5592/CO/CCD.2018

The cover art is the planar Lombardi drawing of a 38-vertex graph with girth five and cyclic edge connectivity three by David Eppstein (arxiv:1206.6142v1), used with permission of the author.

CIP zapis dostupan u računalnom katalogu
Nacionalne i sveučilišne knjižnice u Zagrebu
pod brojem 001035467.

Sponsors:

Faculty of Civil Engineering, University of Zagreb

Croatian Academy of Sciences and Art

Croatian Science Foundation

This volume is dedicated to

Professor Darko Veljan

on the occasion of his 70^{th} birthday

Organizing and program committee:

Tomislav Došlić, Zagreb
Alan Filipin, Zagreb
Snježana Majstorović, Osijek
Ivica Martinjak, Zagreb
Philippe Nadeau, Lyon
Riste Škrekovski, Ljubljana
Damir Vukičević, Split
Ivana Zubac, Mostar
Petra Žigert Pleteršek, Maribor

Local organizing committee:

Nikola Adžaga
Tomislav Došlić
Alan Filipin
Ivica Martinjak

Table of Contents

Preface

Writing a preface is always a pleasure to me. It usually means that a more or less protracted and demanding process of writing or editing a volume is coming to a satisfactory end. It is also an opportunity to acknowledge efforts of all authors, referees, and other participants in the process, and to thank them for their contributions. This is always a pleasure, and I am thanking heartfully to all who made the present volume possible, either by contributing, refereeing, helping in production or supporting it financially.

This preface, however, gives me two more reasons for a pleasure. The first reason is that the present volume is the Proceedings of the **second** Croatian Combinatorial Days. It means that what started as a modest gathering of a small community and its guests and friends has acquired some momentum and a life of its own. This is attested not only by wider attendance at the second meeting, but also by more contributions to this volume. It is my hope that both increasing tendencies will persist and help to strengthen both our discrete mathematics and its ties to a wider community.

The other reason, and the most important one, is that this volume is dedicated to **Professor Darko Veljan** on the occasion of his 70th birthday. This is a small but a heartfelt sign of our appreciation for a colleague, a teacher and a friend who played an important role in education and in professional development of almost all Croatian mathematicians working in the area of discrete mathematics and combinatorics. Thank you, Darko, and keep sharing your knowledge with us!

Both the conference and the Proceedings were made possible by generous help of our sponsors, the Faculty of Civil Engineering, the Croatian Academy of Sciences and Arts and the Croatian Science Foundation (IP-2016-06-1142).

I hope you have enjoyed our last meeting. I also hope you will enjoy the present volume. May we all (and many others) meet again at the third (already traditional?) Croatian Combinatorial Days at the end of September 2020.

Zagreb, July 8, 2019 Tomislav Došlić

Detecting communities in directed acyclic networks using modified LPA algorithms

Suzana Antunović
University of Split, Faculty of Civil Engineering, Architecture and Geodesy, Split, Croatia

Damir Vukičević
University of Split, Faculty of Science, Split, Croatia

Abstract

Networks (or graphs) appear as dominant structures in different domains, including sociology, biology, neuroscience and computing. In most cases, these graphs are directed which changes the semantics of the edges that are no longer symmetrical in the sense that the beginning vertex transfers some property or value to the end vertex, but not vice versa. Detecting community structure in complex networks is an interdisciplinary topic with many relevant areas of application. In order to detect communities in directed acyclic networks, apart from the direction of the edge, the requirement for topological ordering of the vertices should be taken into account. In other words, if the vertices are topologically order is such a way that $x_1 < x_2 < ... < x_n$ we are interested in dividing the network into communities $C_1, C_2, ..., C_k$ in such a way that:

$$\text{if } x_i < x_j,\ x_i \in C_i, x_j \in C_j \text{ then } C_i < C_j \text{ or } C_i = C_j$$

We present an algorithm derived from LPA algorithms which are commonly used in network detection, mostly because of their quick computational time and fairly good results. They were originally developed for undirected networks, but have been modified for this purpose.

Keywords: community detection, directed acyclic network, label propagation algorithm

MSC: 90C35, 90C27, 05C82, 05C85, 97K30

DOI: https://doi.org/10.5592/CO/CCD.2018.01

1 Introduction

In complex networks theory, community detection problem refers to grouping vertices into communities according to their similarity, which usually considers either topological features or other characteristics related to the vertices and edges of the graph. Finding communities in directed networks is a challenging task with several important applications in a wide range of domains. However, the problem has mainly been considered and studied for the case of undirected networks. The problem of community detection in directed networks is considered to be a more challenging task as compared to the undirected case [9]. In order to detect communities in directed acyclic networks, apart from the direction of the edge, the requirement for topological ordering of the vertices should be taken into account. Topological ordering of a directed graph is a linear ordering of its vertices such that for every directed edge uv from vertex u to vertex v, u comes before v in the ordering. A topological ordering is possible if and only if the graph has no directed cycles, that is, if it is a directed acyclic graph [1].

Let G be a simple directed graph with no directed cycles and $P(G)$ the set of all bijections $p : V(G) \to \{1, ..., n\}$ such that $p(u) < p(v)$ for every directed edge $uv \in E(G)$. There is at least one function p in the set P [5]. One of the possible interpretations for this is that vertices represent educational units and for each unit u, $p(u)$ represents the order in which the lesson is learned. All vertices that point to u can be thought of as prerequisites for learning lesson u and for all vertices v such that there is a directed edge uv, lesson u is a prerequisite for learning and understanding lesson v.

For vertex $v \in V$ we define in–neighbor of v as vertex $u \in V$ such that there is a directed edge $uv \in E$. Analogly, we define out–neighbor of v as vertex $u \in V$ such that there is a directed edge $vu \in E$.

2 LPA algorithms

Label Propagation Algorithm uses only the network structure as input data and is relatively fast so it is one of frequently used algorithms for community detection. In practice, the algorithm proved to be very efficient. However, since it involves random processes, in different executions, the algorithm provides different network partitions [7]. The basic idea of the algorithm is that, in each step of the algorithm, vertex x selects the community to which the largest number of its neighbors belong. The propagation process is repeated iteratively until label changes are no longer

possible. The most significant feature of LPA is its computer complexity (near the linear time $O(m)$) [11]. The problem is that the LPA is not stable: the algorithm is susceptible to the order in which the vertices update the labels in each iteration, which is why the solution may be different in the various initiations of the algorithm [7] . Sometimes LPA can end up with a trivial solution - all vertices are identified in the same community [2]. Barber and Clark expanded the LPA by modifying the update rule and proposed a new algorithm called LPAm [2]. Instead of choosing the label that most often appears among his neighbors, vertex x selects a label that will result in maximum modularity increase. Modularity is commonly used measure for evaluating the quality of network division. It measures the actual ratio of edges within the community reduced by the expected value in the null model, where the division is the same, but the edges are placed uniformly and random [8]. The implementation of LPAm brings a monotonous increase in modularity and avoids the possibility of forming a trivial solution. Additionally, LPAm has the same effective speed as LPA. However, the tendency is to get stuck at a low local maximum of modularity [8]. Liu and Murata overlook this problem by joining communities that maximize modularity the most. In this way, the local modularity maximum is avoided. However, it is not certain that the new local maximum that we will get is good enough (though better than the previous one). For this reason it is necessary to repeat this process as long as it is no longer possible to increase modularity. The proposed algorithm is called LPAm+.

3 OLPAm+ algorithm

The algorithm that we propose to detect communities in directed acyclic networks is a heuristic algorithm based on label propagation and the maximization of the modularity Q_d, defined in Eq. (1), which is defined for directed networks.

The basic idea is similar to the idea of LPA algorithms. Vertices take one of their neighbors' labels, which maximally increases the modularity Q_d, but does not violate the requirement for partial arrangement of a set of the communities. Each step selects the optimal solution found on the basis of currently available information in the hope that the final solution will approach the global optimum. This algorithmic paradigm is known as a "greedy algorithm" [4] and is often used in optimization problems.

Let us define modularity for directed networks. Let there be a directed network with n vertices and m directed edges represented by adjacency matrix $\mathbf{A}$. Let $d^{in}(i)$ and $d^{out}(i)$ be in–degree and out–degree of a vertex $i \in V(G)$. Let vertex i belong

to a community l_i. Modularity for directed networks is defined as [6]

$$Q_d = \frac{1}{m} \sum_{1 \leq i,j \leq n} \left[A_{ij} - \frac{d^{in}(j) d^{out}(i)}{m} \right] \delta(l_i, l_j) \tag{1}$$

where $\delta(l_i, l_j)$ is Kronecker's delta.

The algorithm can be described as follows. Each vertex $i \in V(G)$ is assigned a unique numeric label $l_i = p(i)$. The vertices are placed in random order and the process of label propagation begins. For each vertex in that order, the change in modularity caused by the change of label is calculated. From equation (1) it follows that modularity increase can be caused by those pairs of vertices $i, j \in V(G)$ for which $A_{ij} \neq 0$ or $A_{ji} \neq 0$ so specific vertex can update its label to one of its neighbors' labels (including in–neighbors and out–neighbors). When vertex i changes the label, i.e. changes the community to which it belongs, it causes modification of the modularity that can be calculated in the following way. Let i change the existing label l_i to the new label l_j. The change in modularity caused by this change follows from equation (1) and is calculated as

$$\Delta Q_d(ij) = \frac{d_i^j}{m} - \left[\frac{d^{out}(i) S_{in}(j) + d^{in}(i) S_{out}(j)}{m^2} \right] \tag{2}$$

where

- d_i^j is the number of all neighbors of i with label l_j
- $S_{in}(j)$ is the total in–degree of vertices with label l_j
- $S_{out}(j)$ is the total out–degree of vertices with label l_j

The label update process is asynchronous. If the vertex i changes the label at iteration t, some of its neighbors $j \in V(G)$ have already changed the label in the current iteration and have the label $l_j(t)$ while some neighbors $k \in V(G)$ still have a label from the previous iteration $l_k(t-1)$. The vertex i makes a decision based on the state it has found at iteration t, i.e. selects the label that causes the maximum increase in modularity without disturbing the community order. If such label does not exist, the vertex retains its label. To ensure that the vertices (and the resulting communities) are in a valid order, vertex i can choose the largest among the labels of its in–neighbors or the smallest among the labels of its out–neighbors. Otherwise, the order will be disrupted.

When every vertex in the order is considered exactly once, the vertices are put in random order and the process is repeated. The algorithm stops when by changing the label it is no longer possible to get a positive change in modularity. At the end of the algorithm, we identify communities as groups of vertices that have the same label. The described algorithm was called *Orientation Respecting LPAm (OLPAm).* Pseudocode is available in the Algorithm 1.

As described in the last chapter, the LPAm algorithm is susceptible to the vertex order that is random in each iteration and is inclined to stuck in the poor local maximum modularity. The same goes for OLPAm. When this happens, we calculate the change in modularity that is generated by joining community pairs and merging the pair of communities resulting in the greatest increase in modularity and not disturbing the community order. If we merge communities with labels l_i and l_j, the modularity change caused by merging can be calculated as

$$\Delta Q_d(l_i l_j) = \frac{E_{ij}}{m} - \left[\frac{S_{out}(i)S_{in}(j) + S_{in}(i)S_{out}(j)}{m^2}\right] \tag{3}$$

where E_{ij} is the number of edges between communities l_i and l_j. The equation (3) is derived from (2) by summing over all the vertices in the community l_i. From equation (3) it is clear that it is sufficient to compute $\Delta Q_d(l_i l_j)$ for pairs of communities that are connected because only connected communities can make a positive change in modularity. Although merging communities increases modularity, it is not certain that the maximum is at the same time global, so again we call OLPAm and repeat the process as far as it is possible for the communities to get increased modularity without disturbing the community order.

In order to maintain a valid order of communities, consider the following. If there is at least one vertex x_i in the community A_i and at least one vertex x_j in the community A_j such that there exists an edge $x_i x_j \in E(G)$ we say that the community A_i *points to the community* A_j. In the same way as with vertices, the community labeled A can be merged with community labeled A_i that has the largest label l_{max} among the community labels that point to A or with the community A_j with the smallest label l_{min} among the community labels to which A points. Otherwise, the condition from the beginning of the chapter will not be met. The complete algorithm (OLPAm with merging communities) was called *Orientation Respecting LPAm+ (OLPAm+).* Pseudocode can be found in the Algorithm 2.

As far as the OLPAm computational complexity is concerned, one step of label propagation in OLPAm has the complexity of $O(n)$ since for each vertex we calculate two values so the total complexity of the $OLPAm$ algorithm is equal to

$O(rn)$ where r is the maximal number of steps label propagation needs to find the maximal modularity value. Furthermore, a method of merging communities in OLPAm+ algorithm has the complexity of $O(n)$. Namely, n is the maximal number of communities obtained so for calculating all values $\Delta Q_d(l_i l_j)$, the time required is $O(n)$. Determining the maximal value of a string of n elements has the complexity of $O(n)$, which gives the total complexity of one merging of communities $O(n+n) = O(n)$. Let h the number of times the communities merge. The total computational complexity of the algorithm is $O(rn) + h\,[O(n) + O(rn))] = O(n)$. The value of the h parameter can not be accurately estimated because it depends on the quality of the solution obtained in the OLPAm algorithm. Even the value of parameter r can not be fully predicted.

Table 1: **Estimated values of r and h parameters in OLPAm+.** Table shows the average number of steps r needed for the OLPAm to converge and the average number of times h of merging communities in OLPAm+.

Network	n	m	r	h
Number set $\mathbb{Q}$	47	254	7.31	3.25
Elementary functions	84	502	6.09	2.64
Integral	223	656	11.02	5.51
Data processing	54	197	6.09	2.64
Primary production	28	93	7.11	1.23
Physics	31	49	4.77	1.99

When joining communities in the LPAm+ algorithm, we merged only those two communities whose merger results in the largest increase in modularity without distorting the order. Following the example of [8], we consider modifying the OLPAm+ algorithm. When OLPAm stops at the local maximum (further label propagation does not increase modularity), we calculate the modularity changes caused by merging communities and joining those pairs of communities that maximize modularity the most. The pseudo-code of this modified OLPAm+ algorithm is given in the Algorithm 3.

Algorithm 1 Orientation Respecting LPAm **(OLPAm)**

Require: Edge list
Ensure: Community division, modularity

to each vertex i assign a unique numerical label $l_i(0) = p(i)$
set $t = 1$
repeat
  put vertices in random order X
  for each vertex $i \in X$ **do**
    among in–neighbors $x_{i_1}, x_{i_2}, ..., , x_{i_k}$ of vertex i with labels $l_{i_1}, l_{i_2}, ..., l_{i_k}$ find the largest label l_{max}
    among out–neighbors $x_{i_{k+1}}, x_{i_{k+2}}, ..., , x_{i_n}$ of vertex i with labels $l_{i_{k+1}}, l_{i_{k+2}}, ..., l_{i_n}$ find the smallest label l_{min}
    calculate $\Delta Q_d(i, max)$ and $\Delta Q_d(i, min)$
    if $\Delta Q_d(i, max) > \Delta Q_d(i, min)$ and $\Delta Q_d(i, max) > 0$ **then**
      set $l_i(t) = l_{max}$
    else if $\Delta Q_d(i, min) > \Delta Q_d(i, max)$ and $\Delta Q_d(i, min) > 0$ **then**
      set $l_i(t) = l_{min}$
    else if $\Delta Q_d(i, min) = \Delta Q_d(i, max) > 0$ **then**
      uniformly at random pick l_{max} or l_{min} and set it for l_i
    end if
    set $t = t + 1$
  end for
  if neither of vertices $i \in X$ changes its label **then**
    end algorithm
  else
    set $t = t + 1$
  end if
until neither vertex in the iteration changes its label

Algorithm 2 Orientation Respecting LPAm+ **(OLPAm+)**

assign to each vertex a unique numeric label
using OLPAm algorithm maximize modularity Q_d
while there are communities A_i and A_j such that $\Delta Q_d(l_i l_j) > 0$ **do**
  for each community A_i **do**
    calculate $\Delta Q_d(l_i l_{max})$ and $\Delta Q_d(l_i l_{min})$
  end for
  find the maximal value of all $\Delta Q_d(l_i l_j) > 0$
  merge communities A_i and A_j such that $\Delta Q_d(l_i l_j) > 0$ is maximal
  maximize modularity Q_d using OLPAm algorithm
end while

Algorithm 3 Modified OLPAm+ with multiple merging of communities

assign to each vertex a unique numeric label
using OLPAm algorithm maximize modularity Q_d
while $\exists$ pair of communities (A_i, A_j) such that $\Delta Q(l_i, l_j) > 0$ **do**
  for each pair of connected communities (A_i, A_j) where $\Delta Q(l_i, l_j) > 0$ **do**
    if there is no community A labeled l such that $\Delta Q(l, l_i) > \Delta Q(l_i, l_j)$ and $\Delta Q(l, l_j) > \Delta Q(l_i, l_j)$ **then**
      merge communities A_i and A_j
    end if
  end for
  maximize modularity Q_d using OLPAm algorithm
end while

3.1 Experiments and results

The proposed algorithm is implemented in *Microsoft Visual Studio 2015* program tool. Since random processes are used to update the label, the results may vary in each execution of the algorithm. For this reason, based on the example of [8], we ran the algorithm 100 times for each of the networks described below.

3.1.1 Data sets

The OLPAm+ algorithm was originally developed for curriculum networks, directed acyclic networks where vertices represent educational units and directed edge from

vertex u to v means that unit u is necessary for learning and understanding unit v. In order to better analyze the structure and community division, experts from different fields in science and education (mathematics, physics, computer science and biology) were asked to create a network for this purpose. The networks are named by the key concept whose understanding is set as a learning objective for that area. Some basic statistics of curriculum networks can be found in Tables 2 and 3.

"Number set $\mathbb{Q}$***"*** is a curriculum network with 47 vertices and 254 directed edges. Vertex labeled 1 denotes the term *natural number*, vertex labeled 47 represents the term *set of rational numbers*.

"Elementary functions" is a network of terms needed for understanding the term elementary functions. Network has 84 vertices and 502 directed edges. Vertex labeled 1 is *set*, vertex labeled 82 is *elementary function*.

"Integral " is a curriculum network of terms needed for passing the introductory course in mathematics. Network has 223 vertices and 655 directed edges. Vertex labeled 1 is *set*, vertex labeled 223 is *improper integral*.

"Physics" is curriculum network of topics covered in 7th grade. Network has 31 vertices and 49 directed edges. Vertex labeled 1 represents the term *length*, vertex labeled 31 is *the strength of the lens*.

"Data processing" is curriculum network of topics covered in the introductory college course. Network has 54 vertices and 197 directed edges. Vertex labeled 1 represents the term *data*, vertex labeled 31 is *data processing*.

"Primary production" is a curriculum network of terms needed for understanding the process of primary production in oceans. Network has 28 vertices and 93 directed edges. Vertex labeled 1 is *photosynthesis*, vertex labeled 28 is *primary production*.

Table 2: **Basic statistics for curriculum networks.** *Notation:* number of vertices n, number of directed edges m, largest in–degree d_{in}, largest out–degree d_{out}, average degree d_{avg}, average shortest path length l for pairs of connected vertices, clustering coefficient C. Measures used are defined in [10]

Network	n	m	d_{in}	d_{out}	d_{avg}	l	C
Number set $\mathbb{Q}$	47	254	17	26	5.404	2.011	0.254
Elementary functions	84	502	27	51	5.976	2.132	0.255
Integral	223	655	15	28	2.941	3.899	0.084
Physics	31	49	4	8	1.581	1.575	0.049
Primary production	28	93	9	14	3.321	2.135	0.183
Data processing	54	197	12	22	3.648	1.744	0.338

Table 3: **Other basic statistics for curriculum networks.** *Notation:*
diameter $diam$, density D, maximal value of betweenness centrality c, maximal value of hub centrality h, maximal value of authority centrality a.

Network	**diam**	**D**	**c**	**h**	**a**
Number set $\mathbb{Q}$	5	0.117	117.940	0.1219	0.0461
Elementary functions	6	0.072	309.429	0.1125	0.0313
Integral	10	0.013	2041.796	0.0881	0.0489
Physics	4	0.053	12.000	0.2402	0.1066
Primary production	5	0.123	66.089	0.1764	0.0885
Data processing	5	0.069	79.000	0.1463	0.0682

3.1.2 Results

In all following tables, n denotes the number of vertices and m denotes the number of directed edges.

We have compared the results obtained with the OLPAm+ algorithm with community division proposed by the authors of each of the networks. For the proposed division for each network, we calculated the modularity values of Q_d. The results can be found in Table 4. For each network, the algorithm gives higher values of modularity or better division into communities.

Table 4: **Comparation of results from OLPAm+ algorithm with community division proposed by the creators of the networks.** *Notation*: Q_{max} is the maximal value of modularity obtained, N_c is the number of communities which gives the value Q_{max}, Q_d is the value of modularity calculated for community division proposed by the creators.

			$OLPAm+$		Author	
	n	m	Q_{max}	N_c	Q_d	N_c
Number set $\mathbb{Q}$	47	254	0.377	4	0.311	5
Elementary functions	84	502	0.354	4	0.239	6
Integral	223	655	0.468	7	0.455	10
Data processing	54	197	0.426	5	0.389	6
Primary production	28	93	0.293	3	0.237	3
Physics	31	49	0.476	6	0.238	6

Since we were not familiar with any algorithm that requires the arrangement of a set of obtained communities, we compared the values obtained by the proposed OLPAm+ algorithm with the values obtained with the LPAm+ algorithm [8] and the Girvan - Newman algorithm [3] as follows. The LPAm+ algorithm was run 100 for each of the networks and the maximum value of the modularity Q was observed. For the resulting community division, we calculated the modularity values for the directed networks Q_d in the formula (1). The same was done for division by the Girvan–Newman algorithm. Although this is a disregard for the community order requirement, we can compare how much that demand affects the value of modularity and network division quality. The comparison of results is shown in Table 5.

Table 5: **Comparation of the results obtained from several community detection algorithms** Table shows results obtained from running LPAm+ i OLPAm+ algorithms 100 times and Girvan–Newman algorithm for each of the networks. *Notation*: Q_{max} is the maximal value of modularity Q_d obtained, N_c is the number of communities which gives the value Q_{max}, $OLPAm+(m)$ refers to multiple merging of communities.

	$LPAm+$		GN		$OLPAm+$		$OLPAm+(m)$	
	Q_d	N_c	Q_d	N_c	Q_d	N_c	Q_d	N_c
Number set $\mathbb{Q}$	0.376	4	0.367	4	0.377	4	0.377	4
El. functions	0.361	4	0.223	4	0.354	4	0.337	5
Integral	0.567	7	0.542	12	0.468	7	0.470	10
Data processing	0.477	4	0.438	4	0.426	5	0.426	5
Pr. production	0.297	3	0.099	5	0.293	3	0.293	3
Physics	0.457	7	0.377	8	0.476	6	0.467	5

It can be seen that OLPAm+ gives quite good results compared to the other two algorithms mentioned, although, in part, is limited by the requirement for a valid set of established communities.

Conclusion

$OLPAm+$ is a greedy algorithm for detecting communities in directed acyclic networks under the following condition: if the vertices are topologically ordered in such a way that $x_1 < x_2 < ... < x_n$ algorithm divides the network into communities $C_1, C_2, ..., C_k$ in such a way that:

$$\text{if } x_i < x_j,\ x_i \in C_i, x_j \in C_j \text{ then } C_i < C_j \text{ or } C_i = C_j$$

Algorithm has been tested on curriculum networks described in Section 3.1.1. and given results were compared with other community detection algorithms. We conclude that the $OLPAm+$ algorithm gives very good results in terms of modularity for directed networks Q_d defined in Eq. (1). It is also very efficient. The total computational complexity of the algorithm is $O(rn) + h\,[O(n) + O(rn)] = O(n)$.

The value of the h and r parameters can not be accurately estimated because it depends on the quality of the solution obtained in the $OLPAm$ algorithm as explained in Section 3.

References

[1] Bang-Jensen, J., Acyclic Digraphs in *Digraphs: Theory, Algorithms and Applications*, Springer Monographs in Mathematics (2nd ed.) (2008), Springer-Verlag, pp. 3234

[2] Barber, M. J.; Clark, J. W., Detecting network communities by propagating labels under constraints, *Phys. Rev. E*, **80** (2009), DOI:10.1103/PhysRevE.80.026129.

[3] Girvan, M.; Newman, M.E.J., Community structure in social and biological networks, *Proc. Natl. Acad. Sci. USA* **99** (2002), pp. 7821–7826.

[4] Hazewinkel, M., Greedy algorithm in *Encyclopedia of Mathematics*, Springer (2001)

[5] Kahn, A. B., Topological sorting of large networks, *Communications of the ACM* **5 (11)**, (1962), pp. 558–562.

[6] Leicht, E. A.; Newman, M. E. J., Community structure in directed networks,*Phys. Rev. Lett.*, **100** (2008), DOI : 10.1103/PhysRevLett.100.118703.

[7] Leung, . X. Y.; Hui, P.; Lió, P.; Crowcroft, J., Towards real-time community detection in large networks, *Phys. Rev. E* **79** (2009), DOI : 10.1103/PhysRevE.79.066107.

[8] Liu, X.; Murata, T.; Advanced modularity-specialized label propagation algorithm for detecting communities in networks, *Physica A* **389** (2010), pp. 1493–1500.

[9] Malliaros, F.D., Vazirgiannis, M., Clustering and Community Detection in Directed Networks: A Survey,(2013), arXiv:1308.0971, DOI: 10.1016/j.physrep.2013.08.002

[10] Newman, M.E.J., Measures and metrics in *Networks: An Introduction*, Oxford University Press (2010.), New York

[11] Raghavan, U.N.; Albert, R.; Kumara, S., Near linear time algorithm to detect community structures in large-scale networks, *Phys .Rev. E* **76** (2007), DOI : 10.1103/PhysRevE.76.036106.

Relating Brunn-Minkowski and Rogers-Shephard inequalities with the Minkowski asymmetry measure

Katherina von Dichter
Technische Universität München, Department of Mathematics
Boltzmannstraße 3, 85748 Garching by Munich, Germany
dichter@ma.tum.de

Abstract

In this paper we explain how the Minkowski asymmetry measure sharpens several classic results. Especially, we were able to tighten the Brunn-Minkowski and the Rogers-Shephard inequalities in terms of the Minkowski asymmetry measure using some stability results on those inequalities.

Keywords: Brunn-Minkowski inequality, Rogers-Shephard inequality, Minkowski asymmetry measure

MSC: 52A10, 52A38

1 Introduction

Let $\mathrm{vol}(K)$ be the n-dimensional volume (or Lebesgue measure) of K. We will write $\mathrm{vol_n}(K)$ whenever it is necessary to specify the dimension. Let $\mathcal{K}^n$ be the set of all convex compact sets in $\mathbb{R}^n$, $\mathbb{B}_2^n$ be the Euclidean ball of radius 1, $K + L$ be the Minkowski sum of convex bodies K and C, i.e. $K + L = \{a + b | a \in K,\, b \in L\}$.
The Brunn-Minkowski inequality establishes that for any convex compact sets K, L holds

$$\mathrm{vol}(K + L)^{\frac{1}{n}} \geq \mathrm{vol}(K)^{\frac{1}{n}} + \mathrm{vol}(L)^{\frac{1}{n}}.$$

Equality holds if and only if one of the three following cases are true (see [FiMaPr]):

DOI: https://doi.org/10.5592/CO/CCD.2018.02

(i) in case for sets K and L with positive volume if and only if K is a homothet of L,

(ii) in case when one of them has volume 0, namely $\mathrm{vol}(K) = 0$ and $\mathrm{vol}(L) > 0$, then if and only if K is a singleton,

(iii) in case when both of them have volume 0, then if and only if K and L are contained in parallel hyperplanes.

The Brunn-Minkowski inequality was proved in the 19th century by Brunn for compact convex sets in low dimensions ($n \leq 3$) (see [Bru]), and Minkowski for general compact convex sets in $\mathbb{R}^n$ (see [Mi]). With the time the Brunn-Minkowski inequality became the starting point of the Brunn-Minkowski theory and a powerful tool for problems involving metric quantities such as volume, surface area and mean width.

The Rogers-Shephard inequality (see [RoSh, Thm. 1]), which can be regarded as a reverse inequality to the Brunn-Minkowski inequality, yields

$$\mathrm{vol}(K + L) \leq \binom{2n}{n} \frac{\mathrm{vol}(K)\mathrm{vol}(L)}{\mathrm{vol}(K \cap (-L))}.$$

Moreover, equality holds if and only if $K = -L$ is an n-dimensional simplex.

The Minkowski asymmetry $s(K)$ of a convex compact sets K is the smallest rescalation of K which contains a translation of $-K$. It is known that $1 \leq s(K) \leq n$ with $s(K) = 1$ if and only if K is centrally symmetric, while $s(K) = n$ if and only if K is an n-dimensional simplex (see [BrKo, Cor. 2.7]).
Computing the volume is always a computationally hard task. However, computing the Minkowski asymmetry measure may be a computationally easy task (see [BrKo]). We wonder whether we could improve the Brunn-Minkowski and the Rogers-Shephard inequalities for prescribed $s(K) = s \in [1, n]$.

In order to start finding values of $c(s)$ and $C(s)$, we need to recall the proofs of the Brunn-Minkowski and the Rogers-Shepard inequalities. Secondly, we will show also the characterizations of their equality cases. Finally, we will show the ideas in order to obtain some stability results, in sense of near-equalities, of the Brunn-Minkowski and the Rogers-Shepard inequalities, which will help in providing bounds for $c(s)$ and $C(s)$ from some particular values of $s \in [0, 1]$.

The results presented in this paper are based on the master thesis (see [Di]).

1.1 Minkowski asymmetry measure

For an n-dimensional convex and compact set K, its Minkowski asymmetry (measure) $s(K)$ (cf. Minkowski 1911) is the smallest $\rho > 0$ such that $-K \subset x + \rho K$ for some $x \in \mathbb{R}^n$, i.e.

$$s(K) := \inf\{\lambda > 0 | K \subset x + \lambda(-K) \quad \text{for some} \quad x \in \mathbb{R}^n\}.$$

Note that for Minkowski asymmetry $s(K)$ holds that $K \subset x + s(K)(-K)$ for some $x \in \mathbb{R}^n$. Figure 1 shows two examples of computing the Minkowski asymmetry measure.

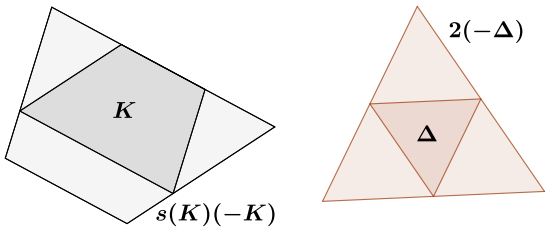


Figure 1: $K \subset x + s(K)(-K)$ for some $x \in \mathbb{R}^n$ and $\triangle \subset y + 2(-\triangle)$ for some $y \in \mathbb{R}^n$

1.2 The Brunn-Minkowski inequality

Figure 2 shows the Minkowski sum of two 0-centered triangles.

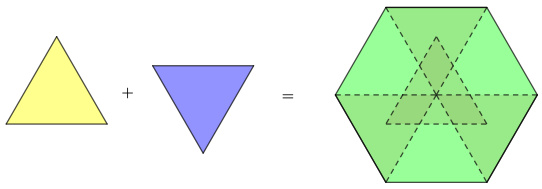


Figure 2: Minkowski sum of the regular triangles $\triangle$ and $\triangledown$

We are now ready to state the Brunn-Minkowski inequality for compact sets, which gives a lower bound for $\text{vol}(K + L)$ in terms of $\text{vol}(K)$ and $\text{vol}(L)$.

Theorem 1 (Brunn 1887, Minkowski 1896)**.** *For any compact sets K, L holds*

$$\mathrm{vol}(K+L)^{\frac{1}{n}} \geq \mathrm{vol}(K)^{\frac{1}{n}} + \mathrm{vol}(L)^{\frac{1}{n}}. \tag{1}$$

The Brunn-Minkowski inequality (1) implies a well-known fact from Analysis, namely the Isoperimetric inequality.

Theorem 2 (Isoperimetric inequality)**.** *Let K be a convex set in $\mathbb{R}^n$ and $P(K)$ its surface area. Then*

$$P(K) \geq n\mathrm{vol}(\mathbb{B}_2^n)^{\frac{1}{n}}\mathrm{vol}(K)^{\frac{n-1}{n}}.$$

The Isoperimetric inequality implies that among sets with a fixed surface area, Euclidean balls maximize the volume.

Next let us state the Rogers-Shephard inequality, which gives us an upper bound for $\mathrm{vol}(K+L)$ in terms of $\mathrm{vol}(K)$, $\mathrm{vol}(L)$ and $\mathrm{vol}(K \cap (-L))$. Note that it is in general impossible to find an upper bound for $\mathrm{vol}(K+L)$ just in terms of $\mathrm{vol}(K)$ and $\mathrm{vol}(L)$.

Theorem 3 (Rogers and Shephard, 1958)**.** *Let K, L be convex bodies. Then*

$$\mathrm{vol}(K+L)\mathrm{vol}(K \cap (-L)) \leq \binom{2n}{n}\mathrm{vol}(K)\mathrm{vol}(L). \tag{2}$$

There exist also another version of the Rogers-Shephard inequality, namely with projections and intersections.

Theorem 4 (Rogers and Shephard, 1958)**.** *Let K be a compact body, H an i-dim subspace of $\mathbb{R}^n$. Then*

$$\mathrm{vol}_i(P_H K) \max_{x \in \mathbb{R}^n} \mathrm{vol}_{n-i}(K \cap (x + H^{\perp})) \leq \binom{n}{i}\mathrm{vol}_n(K). \tag{3}$$

An application of the Rogers-Shepard inequality (3) can be the following. One could approximate the volume $\mathrm{vol}(K)$ of a compact set $K \in \mathbb{R}^n$ (for instance a brain tumor) by measuring the volume of the projection $\mathrm{vol}_i(P_H K)$ and the volume of the biggest section $\max_{x \in \mathbb{R}^n} \mathrm{vol}_{n-i}(K \cap (x + H^{\perp}))$.

In order to see that (3) implies (2), we introduce the set $M := \{(z_1, z_2)^T \in \mathbb{R}^{2n} | z_2 \in K, z_1 \in z_2 + L\}$. Using the definition of the set M it can be easily seen that $P_H M = K + L$, $M \cap H^{\perp} = K \cap (-L)$ and $\mathrm{vol}_{2n}(M) = \mathrm{vol}_n(K)\mathrm{vol}_n(L)$. And therefore two versions of the Rogers-Shepard inequality (2) and (3) are equivalent.

Choosing $L = -K$ and combining the Rogers-Shephard and the Brunn-Minkowski inequalities leads to

$$2^n \leq \frac{\mathrm{vol}(K-K)}{\mathrm{vol}(K)} \leq \binom{2n}{n}.$$

We aim to find constants $c(s), C(s)$ with $2^n \leq c(s) \leq C(s) \leq \binom{2n}{n}$, such that for every convex body K such that $s(K) = s \in [1, n]$ holds

$$c(s) \leq \frac{\mathrm{vol}(K-K)}{\mathrm{vol}(K)} \leq C(s)\,.$$

In order to do it we study the equality cases of the Brunn-Minkowski inequality (1) and the Rogers-Shephard inequalities (2) and (3), their stability results and then involve the Minkowski asymmetry measure in that inequalities.

1.3 Equality case of the Brunn-Minkowski inequality

The equality case of the Brunn-Minkowski inequality (1) is the following.

Theorem 5 (Klain, 2011, Kneser, Süss, 1932)**.** *Let K and L be convex compact sets with positive volumes. Then*

$$\mathrm{vol}(K+L)^{\frac{1}{n}} = \mathrm{vol}(K)^{\frac{1}{n}} + \mathrm{vol}(L)^{\frac{1}{n}}$$

holds if and only if K and L are homothets.

Moreover, if exactly one of the sets is lower dimensional, namely $\mathrm{vol}(K) = 0$, and $\mathrm{vol}(L) > 0$, then equality holds if and only if K is a singleton; and if $\mathrm{vol}(K) = \mathrm{vol}(L)$, then equality holds if and only if K and L are contained in parallel hyperplanes.

1.4 Stability of the Brunn-Minkowski inequality

We consider the stability as a sharpening of the inequality at the near-quality case. Let $K \triangle L$ be the symmetric difference of K and L, i.e. $K \triangle L = (K \setminus L) \cup (L \setminus K)$. For the Brunn-Minkowski inequality (1) it holds that

Theorem 6 (Figalli, 2009)**.** *Let K and L be convex compact sets in $\mathbb{R}^n$. Then*

$$\mathrm{vol}(K+L)^{\frac{1}{n}} \geq \left(\mathrm{vol}(K)^{\frac{1}{n}} + \mathrm{vol}(L)^{\frac{1}{n}}\right)\left(1 + \frac{A(K,L)^2}{c(n)\sigma(K,L)^{\frac{1}{n}}}\right)$$

with $A(K,L) := \inf_{x\in\mathbb{R}^n}\left\{\frac{\mathrm{vol}(K\triangle(x+\lambda L))}{\mathrm{vol}(K)}\right\}$, $\lambda = \left(\frac{\mathrm{vol}(K)}{\mathrm{vol}(L)}\right)^{\frac{1}{n}}$, $\sigma(K,L) := max\left\{\frac{\mathrm{vol}(K)}{\mathrm{vol}(L)}, \frac{\mathrm{vol}(L)}{\mathrm{vol}(K)}\right\}$ *and* $c(n) = 14n^2 4^{n-1}$.

One can see that due to the stability result in case of near-equality of (1) we have that K and L are almost homothets. Therefore $\mathrm{vol}(K \triangle (x+\lambda L)) \approx 0$ and therefore $A(K,L) \approx 0$.

1.5 Equality case of the Rogers-Shephard inequality

The equality case of the Rogers-Shephard inequality (3) can be stated as following.

Theorem 7 (Rogers and Shepard, 1958)**.** *Let K be a convex compact sets in $\mathbb{R}^n$, H an i-dim subspace of $\mathbb{R}^n$ and $x \in \mathbb{R}^n$. Then*

$$\mathrm{vol}_i(P_H K) \max_{x\in\mathbb{R}^n} \mathrm{vol}_{n-i}(K \cap (x+H^\perp)) = \binom{n}{i}\mathrm{vol}_n(K)$$

if and only if

$$K \cap (x+H^\perp) =_h K \cap H^\perp \quad \forall\, x \in \mathbb{R}^n$$

and

$$\mathrm{vol}_{n-i}(K \cap (x+H^\perp)) = (1-||x||_K)^{n-i}\mathrm{vol}_{n-i}(K \cap H^\perp).$$

Next we present a more recent result that is the equality case of the Rogers-Shephard inequality (2).

Theorem 8 (Alonso-Gutierrez, Jimenez, Villa, 2013)**.** *Let K and L be convex compact sets in $\mathbb{R}^n$. Then*

$$\mathrm{vol}_n(K+L)\mathrm{vol}_n(K \cap (-L)) = \binom{2n}{n}\mathrm{vol}_n(K)\mathrm{vol}_n(L)$$

if and only if K and L are simplices such that $K = -L$.

1.6 Stability of the Rogers-Shephard inequality

Let $d_{BM}(K,L) := \min\{\rho \geq 1 | K \subset A(L) \subset x+\rho K \text{ for some } x \in \mathbb{R}^n\}$ with $A(L)$ being any affine transformation of L be the Banach-Mazur distance of K with respect to L and T be an n-simplex.

Theorem 9 (Boroczsky, 2005)**.** *Let K be a convex compact sets in $\mathbb{R}^n$. If*

$$\mathrm{vol}(K-K) = (1-\varepsilon)\binom{2n}{n}\mathrm{vol}(K),$$

then

$$1+\frac{1}{n}\varepsilon \le d_{BM}(K,T) \le 1+n^{50n^2}\varepsilon.$$

1.7 Results

Let us recall that the Brunn-Minkowski and Rogers-Shephard inequalities for convex compact sets K and $-K$ state that

$$2^n \le \frac{\mathrm{vol}(K-K)}{\mathrm{vol}(K)} \le \binom{2n}{n}.$$

Moreover, $\frac{\mathrm{vol}(K-K)}{\mathrm{vol}(K)} = 2^n$ if and only if K is symmetric and $\frac{\mathrm{vol}(K-K)}{\mathrm{vol}(K)} = \binom{2n}{n}$ if and only if K is an n-dimensional simplex.

But at the same time we have that

$$1 \le s(K) \le n.$$

Moreover, $s(K)=1$ if and only if K is symmetric and $s(K)=n$ if and only if K is an n-dimensional simplex.

Combining those facts enable us to involve the asymmetry measure into these inequalities. We state some improvements on the Brunn-Minkowski and the Rogers-Shephard inequalities by means of the Minkowski asymmetry measure (see [Di]).

Theorem 10. *Let K be a convex compact set in $\mathbb{R}^n$ and $s=s(K)$. Then*

$$c(s) \ge 2^n\left(1+\frac{1}{n\,4^{n-1}}\left(\frac{(s-1)^n\mathrm{vol}_{n-1}(\mathbb{B}_2^{n-1})}{2^{n-1}n^{2n}\mathrm{vol}_n(\mathbb{B}_2^n)}\right)^2\right)^n \tag{4}$$

and

$$C(s) \le (1+s)^n. \tag{5}$$

Moreover, if $n-\frac{1}{4n} < s < n$, then

$$c(s) \ge \binom{2n}{n}(1-4n^2(n-s)) \quad \textit{and} \quad C(s) \le \binom{2n}{n}\left(1-\frac{n-s}{n^{1+50n^2}}\right). \tag{6}$$

It is worth mentioning that (4) and (5) (resp. (6)) are specially good when $s(K) \approx 1$ (resp. $s(K) \approx n$).

Theorem 11. *The diagram $f(\mathcal{K}^n)$, where $f : \mathcal{K}^n \to [1, n] \times \left[2^n, \binom{2n}{n}\right]$ is given by $f(K) := \left(s(K), \frac{\mathrm{vol}(K-K)}{\mathrm{vol}(K)}\right)$, is simply connected, contains $(1, 2^n)$ and $(n, \binom{2n}{n})$.*

We now investigate $f(\mathcal{K}^2)$.

Remark 1. *Let $T = \mathrm{conv}(\{(0, 1)^T, (\pm\sqrt{3}/2, -1/2)^T\})$, let $s \in [1, 2]$, and let*

$$K_s := T \cap (-sT) \quad \textit{and} \quad C_s := \mathrm{conv}(T \cup (-sT)).$$

Then $s(K_s) = s(C_s) = s$, $\mathrm{vol}(K_s) = \frac{2s-(s-1)^2}{4}$, $\mathrm{vol}(K_s - K_s) = \frac{(s+1)^2}{2}$, $\mathrm{vol}(C_s) = \frac{3\sqrt{3}}{2}s$, and $\mathrm{vol}(C_s - C_s) = 3\sqrt{3}s(1 + s)$.

We finally provide upper and lower bounds for the constants $c(n), C(n)$ in the planar case derived from the Theorem 10 and Remark 1.

Corollary 1. *Let $K \in \mathcal{K}^2$ and let $s = s(K)$. Then*

$$4\left(1 + \frac{(s-1)^4}{2^{11}\pi^2}\right)^2 \leq c(s) \leq 2\frac{(s+1)^2}{2s-(s-1)^2}$$

and

$$2(s + 1) \leq C(s) \leq (1 + s)^2.$$

Moreover, if $s > \frac{8}{7}$, then

$$6(16s - 31) \leq c(s) \quad \textit{and} \quad C(s) \leq 6\left(1 - \frac{2-s}{2^{201}}\right).$$

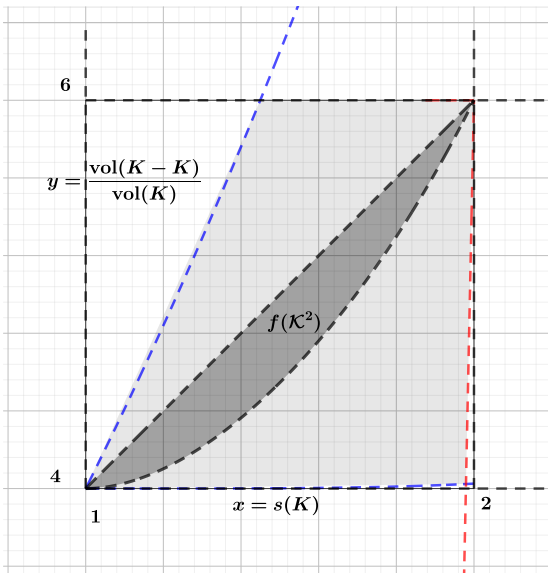


Figure 3: The diagram in Theorem 11 and the bounds obtained in Theorem 11 (dashed blue and red lines, respectively) and Corollary 1 (dashed black lines) in the case of $n = 2$; the light grey area is obtained in Theorems 1 and 3.

References

[AlArGoJiVi] D. Alonso-Gutiérrez, S. Artstein-Avidan, B. González Merino, C. H. Jiménez, R. Villa, Rogers-Shephard and local Loomis-Whitney type inequalities, arXiv:1706.01499, 2017.

[AJV] D. Alonso-Gutiérrez, C.H. Jiménez, R. Villa, Brunn–Minkowski and Zhang inequalities for convolution bodies, Adv. Math., 238 (2013), 50–69.

[BiGaGr] G. Bianchi, R.J. Gardner, P. Gronchi, Symmetrization in Geometry, Adv. Math., 36 (2017), 51–88.

[Bl] W. Blaschke, Eine Frage über konvexe Körper, Jahresber. Deutsch. Math.-Verein., 25 (1916), 121–125.

[Bo] F. Bohnenblust, Convex regions and projections in Minkowski spaces, Ann. of Math. 39 (1938), no. 2, 301–308.

[Bor] K. Boroczsky, The stability of the Rogers-Shephard inequality and some related inequalities, Adv. Math., 190 (2005), no. 1, 47–76.

[BrG] R. Brandenberg, B. González Merino, A complete 3-dimensional Blaschke-Santaló diagram, Math. Ineq. Appl. 20 (2017), no. 2, 301–348.

[BrG2] R. Brandenberg, B. González Merino, The asymmetry of complete and constant width bodies in general normed spaces and the Jung constant, Israel J. Math. 218 (2017), no. 1, 489–510.

[BrG3] R. Brandenberg, B. González Merino, Minkowski concentricity and complete simplices, J. Math. Anal. Appl. 454 (2017), no. 2, 981–994.

[BrKo] R. Brandenberg and S. König, No dimension-independent core-sets for containment under homothetics, Discrete Comput. Geom., 49 (2013), no 1, 3–21.

[Bru] H. Brunn, Über Ovale und Eiflächen, Inaugural Dissertation, München (1887).

[BrK2] R. Brandenberg, S. König, Sharpening geometric inequalities using computable symmetry measures, Mathematika, 61 (2015), no. 3, 559-–580.

[Da] P.J. Davis, 6. Gamma function and related functions, in Abramowitz, Milton; Stegun, Irene A., Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, New York: Dover Publications, 1972.

[Di] K. von Dichter, Volume estimates via the Asymmetry Measure of Minkowski, Master Thesis (2018).

[Fe] H. Federer, Geometric measure theory, Die Grundlehren der mathematischen Wissenschaften, Band 153, New York: Springer-Verlag New York Inc., pp. xiv+676, 1969.

[FiMaPr] A. Figalli, F. Maggi, A. Pratelli, A refined Brunn–Minkowski inequality for convex sets, Ann. Inst. H. Poincaré Anal. Non Linéaire, 26 (2009), no. 6, 2511–2519.

[FiMaPr2] A. Figalli, F. Maggi, A. Pratelli, A mass transportation approach to quantitative isoperimetric inequalities, Invent. Math., 182 (2010), no. 1, 167–211.

[Gl] E. D. Gluskin, Diameter of the Minkowski compactum is approximately equal to n, Funct. Anal. Appl. 15 (1981), 57–58.

[GoHe] B. González Merino, M. A. Hernández Cifre, El teorema del elipsoide de John, http://hdl.handle.net/10201/30278, Universidad de Murcia, 2009.

[Gr] B. Grünbaum, Measures of symmetry for convex sets, Proc. Sympos. Pure Math. 7 (1963), 233–270.

[GuJ] Q. Guo, H. Jin, Asymmetry of convex bodies of constant width, Discrete Comput. Geom. 47 (2012), no. 2, 415–423.

[GuK] Q. Guo, S. Kaijser, On asymmetry of some convex bodies, Discrete Comput. Geom. 27 (2002), no. 2, 239–247.

[HaOh] H. Hadwiger, D. Ohmann, Brunn-Minkowskischer Satz und Isoperimetrie, Math. Z., 66 (1956), no. 1, 1–8.

[Kl] D. A. Klain, On the equality conditions of the Brunn-Minkowski theorem, Proc. Amer. Math. Soc., 139 (2011), no. 10, 3719–3726.

[La] M. Lassak, Approximation of convex bodies by inscribed simplices of maximum volume, Beitr. Algebra Geom. 52 (2011), no. 2, 389–394.

[McGu] R. J. McCann, N. Guillen, Five lectures on optimal transportation: geometry, regularity and applications. Analysis and geometry of metric measure spaces,

Lecture notes of the séminaire de Mathématiques Supérieure (SMS) Montréal, 145–180, 2011.

[Mc] P. McMullen, Inequalities between intrinsic volumes, Monatsh. Math., 111 (1991), no. 1, 47-–53.

[Mi] H. Minkowski, Geometrie der Zahlen, Leipzig: Teubner, 1896.

[RoSh] C. A. Rogers, G. C. Shephard, Convex bodies associated with a given convex body, J. Lond. Math. Soc., 1 (1958), no. 3, 270–281.

[Ru] M. Rudelson, Distance between non-symmetric convex bodies and the MM*-estimate, Positive 4 (2000), no. 2, 161–178.

[Sa] L. Santaló, Sobre los sistemas completos de desigualdades entre tres elementos de una figura convexa plana, Math. Notae 17 (1961), 82–104.

[Sch] R. Schneider, Convex bodies: the Brunn–Minkowski theory, no. 151, Cambridge university press, 2014.

[So] V. Soltan, A characterization of homothetic simplices, Discr. Comput. Geom., 22 (1999), no. 2, 193–200.

[St] P. Steinhagen, Über die größte Kugel in einer konvexen Punktmenge, Abh. Hamb. Sem. Hamburg 1 (1921), 15–26.

[TJ] N. Tomczak-Jaegermann, Banach-Mazur Distances and Finite-Dimensional Operator Ideals, Pitman Monographs and Surveys in Pure and Applied Mathematics 38, Longman Scientifical and Technical, New York (1989).

Directed packings of circles in the plane

Tomislav Došlić
University of Zagreb, Faculty of Civil Engineering,
Kačićeva 26, 10000 Zagreb, Croatia

Mirjam Škegro
Erste & Steiermärkische Bank d.d.,
Ivana Lučića 2, 10000 Zagreb, Croatia

Abstract

We consider sequential packings of families of circles in the plane whose curvatures are given as members of a sequence of non-negative real numbers. Each such packing gives rise to a sequence of circle centers that might diverge to infinity or remain bounded. We examine the behavior of the sequence of circle centers as a function of the growth rate of the sequence of curvatures. In several special cases we obtain explicit formulas for the coordinates of the limit, while in other cases we obtain accurate estimates.

Keywords: packing of circles

MSC: 05B40, 51M04, 52C15

1 Introduction and motivation

Packings of geometric objects have been attracting attention of researchers and engineers since the antiquity [1]. Probably the best known example is the celebrated Kepler conjecture on sphere packings in space, centuries old, and solved only recently

DOI: https://doi.org/10.5592/CO/CCD.2018.03

[2, 3]. Many other problems of efficient use of (not necessarily spatial) resources can be successfully modeled by packings. Among the examples are accretion processes of impenetrable particles subject to attractive forces. A toy model of one such process is considered in this paper. We construct it by considering dynamic (sequential) packing of circles in the plane subject to the following conditions.

In the beginning (at the time zero), we have two circles touching externally at the origin of the plane coordinate system. Both their centers are on the y-axis, and we allow that one of the circles has an infinitely large radius (hence the curvature equal to zero). At each successive moment a circle arrives and settles in the plane so that it touches externally two circles that immediately precede it. Out of the two possible location for the circle center, we choose the one that is farther from the origin. We are interested in the behavior of the sequence of centers of circles in such packings. More precisely, we wish to decide whether the sequence of centers diverges to infinity or converges to a limit, i.e., to a point at a finite distance from the origin. In the latter case, we would like also to determine the coordinates of the limit, or, at least, to give some estimate of its location.

It is clear that the behavior of the sequence of circle centers depends on the radii of accreting circles. We assume that the inverse values of the radii (hence the circle curvatures) are given as elements of a sequence $(a_n)_{n\geqslant 0}$ of non-negative real numbers. Our goal is to determine how behavior of the sequence $(S_n)_{n\geqslant 0}$ of circle centers depends on the sequence $(a_n)_{n\geqslant 0}$. In particular, we aim at finding the necessary and sufficient conditions for convergence of $(S_n)_{n\geqslant 0}$ in terms of growth properties of sequence of curvatures. In case of convergence, we will try to determine the exact or approximate coordinates of the limit.

2 Definitions and preliminaries

Let I be an arbitrary index set. A family $\{K_i, i \in I\}$ of sets in the plane is a **packing** if $intK_i \cap intK_j \neq \emptyset, \forall i \neq j \in I$.

In our case, $I = \mathbb{N}_0$ and each K_i is a circle of radius r_i centered at S_i. It is clear that

the efficiency of packing will be the greatest if the circles touch each other (since their interiors cannot overlap).

Let us take a sequence $(a_n)_{n\geqslant 0}$ of non-negative real numbers such that $a_n > 0$ for all $n \geqslant 1$ and two circles, K_0 and K_1 so that they touch externally at the origin of the plane coordinate system. We assume that the center of K_0 is on the negative part of the y-axis and the center of K_1 is on the positive part of the y-axis. We denote these centers by S_0 and S_1, respectively. The radii r_0 and r_1 of K_0 and K_1 are given as the reciprocal values of a_0 and a_1, respectively. Clearly, if $a_0 = 0$, K_0 becomes the x-axis; such a situation is shown in Fig. 1. Now take a circle K_2 of

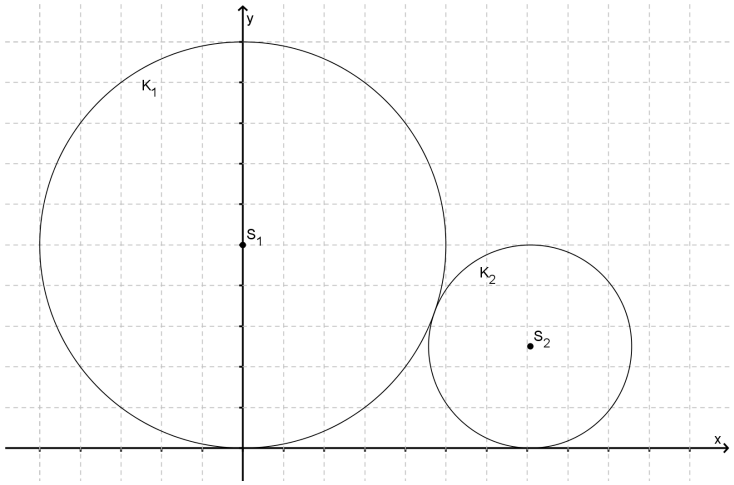


Figure 1: The case $a_0 = 0$

radius $r_2 = \frac{1}{a_2}$ and position it in the right half-plane so that it touches externally both K_0 and K_1. Denote its center by S_2. Now, for each $n \geqslant 3$ take a circle K_n of radius $r_n = \frac{1}{a_n}$ and determine its center S_n so that K_n touches externally both K_{n-1} and K_{n-2}. In each step there will be two possible locations of S_n; we take the one more distant from the origin. In the rest of this paper we will study how the behavior of so constructed sequence $(S_n)_{n\geqslant 0}$ depends on properties of $(a_n)_{n\geqslant 0}$.

Let $S((a_n))$ denote the limit (if it exists) of the sequence of circle centers for a given sequence (a_n), and let $\rho((a_n))$ denotes its distance from the origin. When (S_n) diverges, we set $\rho((a_n)) = \infty$. Our problem can be now formulated in terms

of $\rho((a_n))$: whether it is finite or infinite for a given (a_n), and, if finite, what is its exact or approximate value.

We start by some simple results for the case of constant and of decreasing sequences $(a_n)_{n\geqslant 0}$. In the case of constant sequence we can assume $a_n = 1$ for all $n \geqslant 0$. The situation is shown in Fig. 2. It is a simple exercise to compute the coordinates (p_n, q_n) of the sequence of centers S_n [4].

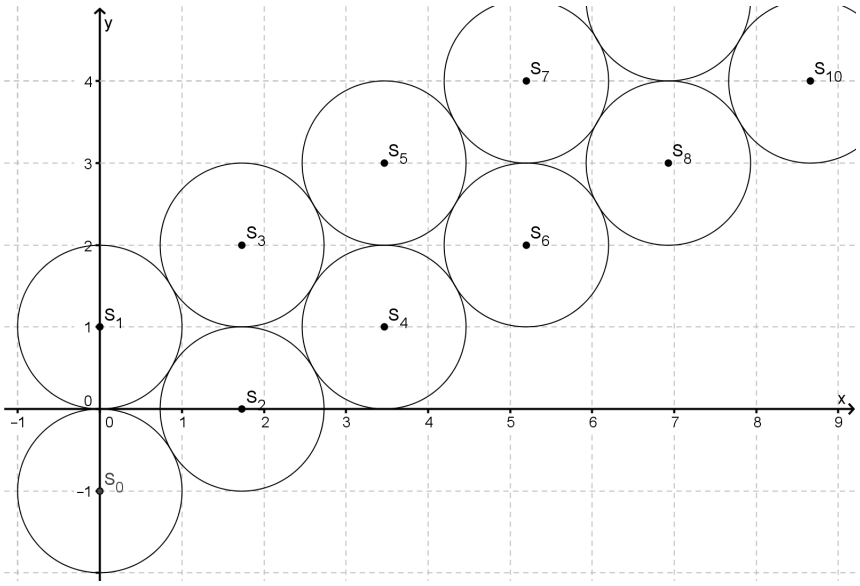


Figure 2: The packing for $a_n = 1$

$$p_n = \lfloor \frac{n}{2} \rfloor \sqrt{3}, \qquad q_n = 3\lfloor \frac{n+1}{2} \rfloor - n - 1.$$

It is clear from both the figure and from the explicit expression for the coordinates of S_n that the sequence (S_n) diverges.

Another simple situation arises when the sequence $(a_n)_{n\geqslant 0}$ is decreasing. An example of such a situation with $a_0 = 0$, $a_n = \frac{1}{n}$ for $n \geqslant 1$, is shown in Fig. 3. It is clear that the sequence (S_n) diverges. Hence, it makes sense to restrict our attention on

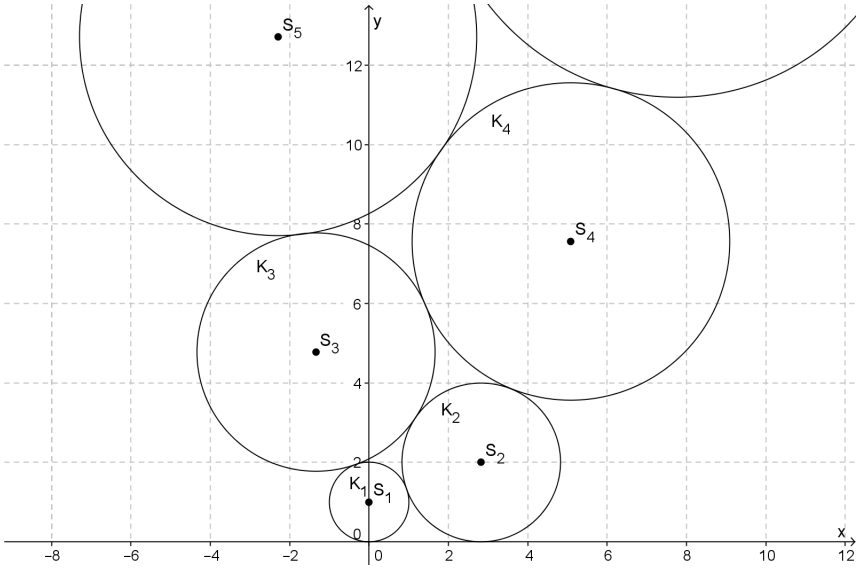


Figure 3: A packing with an inadmissible sequence (a_n)

(at least weakly) increasing sequences that grow without bound. From now on we consider only such sequences. We call them the **admissible sequences**.

By using induction on n and passing to the limit we can establish the following property of admissible sequences.

Proposition 1 *Let $(a_n)_{n\geqslant 0}$ and $(b_n)_{n\geqslant 0}$ be two admissible sequences with $a_n \leqslant b_n$ for all $n \in \mathbb{N}_0$. Then $\rho((a_n)) \geqslant \rho((b_n))$.*

We close this section by considering the case when the sequence of radii converges, hence when $\sum_{n=0}^{\infty} \frac{1}{a_n} = M < \infty$.

Proposition 2 *If the series $\sum \frac{1}{a_n}$ is convergent, then the sequence (S_n) converges to the limit $S((a_n))$ at a finite distance from the origin.*

Proof. Let us denote by d_n the distance between S_n and the origin. It is clear that the difference $d_n - d_{n-1}$ cannot exceed the sum of the radii of the corresponding circles. Hence, $d_n - d_{n-1} \leqslant \frac{1}{a_{n-1}} + \frac{1}{a_n}$ for all $n \geqslant 2$. The claim now follows by summing over all $n \geqslant 2$. ■

Corollary 3 *Let $(a_n)_{n\geqslant 0}$ be an admissible sequence and let $a_n \in \Omega(n^s)$ for some $s > 1$. Then $\rho((a_n)) < \infty$.*

The above result settles the case of sequences (a_n) growing faster than the sequence of natural numbers. It leaves open the case when (a_n) is exactly the sequence of natural numbers, as well as all admissible sequences of slower growth. Also, it provides no information on the actual values of $S((a_n))$ and $\rho((a_n))$ for convergent sequences (S_n).

3 Numerical experiments

In order to facilitate our investigation and to gather some information on behavior of various sequences, we have designed a recursive algorithm for computing the coordinates of centers S_n for a given sequence (a_n). It is based on the procedure for finding intersections of two circles with given centers and radii. We have implemented the algorithm in *Matlab*; the code is given in the Appendix.

3.1 Algorithm

We start from a given sequence $(a_n)_{n\geqslant 0}$, and the sequence of reciprocal values $(r_n)_{n\geqslant 0} = (\frac{1}{a_n})_{n\geqslant 0}$. We take circles K_0 and K_1 centered at $S_0 = (0, -\frac{1}{a_0})$ and $S_1 = (0, \frac{1}{a_1})$, respectively, touching externally at the origin. For $n \geqslant 2$ we solve the system

$$\begin{cases} (x - p_{n-2})^2 + (y - q_{n-2})^2 = (r_{n-2} + r_n)^2 \\ (x - p_{n-1})^2 + (y - q_{n-1})^2 = (r_{n-1} + r_n)^2. \end{cases}$$

for p_n and q_n, the coordinates of S_n. The solutions represent two intersection of circles centered at S_{n-2} and S_{n-1} with radii $r_{n-2}+r_n$ and $r_{n-1}+r_n$, respectively, and at each step we choose the one farther from the origin. The first step is shown in Fig. 4. In case of $a_0 = 0$, the first step is modified so as to acknowledge the fact

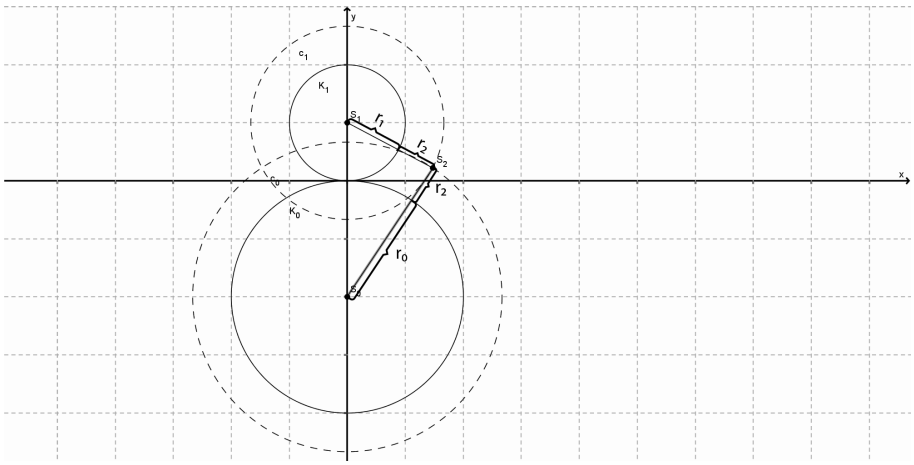


Figure 4: Start of the algorithm.

that one circle is actually a line, the x-axis. The modification is shown in Fig. 5.

For $n \geqslant 2$ (or $n \geqslant 3$ if $a_0 = 0$) we define

$$\begin{aligned}
k_{n-2} &:= \frac{p_{n-2}-p_{n-1}}{q_{n-1}-q_{n-2}} \\
l_{n-2} &:= \frac{r_{n-2}^2 - r_{n-1}^2 + (2r_{n-2}-2r_{n-1})r_n - p_{n-2}^2 + p_{n-1}^2 - q_{n-2}^2 + q_{n-1}^2}{2q_{n-1}-2q_{n-2}}. \\
a'_{n-2} &= 1 + k_{n-2}^2 \\
b_{n-2} &= 2k_{n-2}l_{n-2} - 2p_{n-2} - 2q_{n-2}k_n \\
c_{n-2} &= -r_{n-2}^2 - 2r_{n-2}r_n - r_n^2 + p_{n-2}^2 - 2q_{n-2}l_n + q_{n-2}^2 + l_n^2.
\end{aligned}$$

After some tedious, but otherwise quite straightforward, algebraic manipulations, we obtain a recursive formula for computing sequences (p_n) and (q_n) of center co-

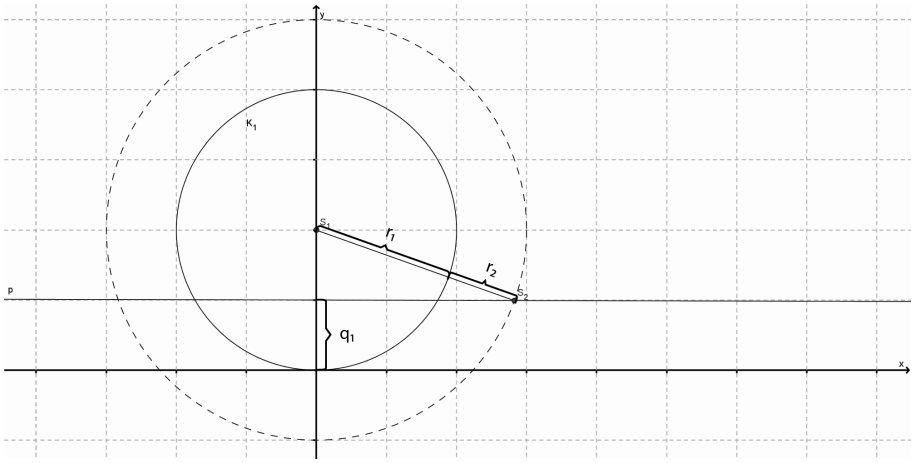


Figure 5: Start of the algorithm for $a_0 = 0$.

ordinates.

$$\begin{aligned} p_n &= \frac{-b_{n-2} + \sqrt{b_{n-2}^2 - 4a_{n-2}^{'}c_{n-2}}}{2a_{n-2}^{'}}, \\ q_n &= k_n p_n + l_n. \end{aligned}$$

In the next subsection we show the results of the algorithm for two sequences with convergent sequences of centers.

3.2 Examples

Our first example is representative for all geometric sequences q^n for $q > 1$.

Example 4 $a_n = 2^n$.

The values for the first ten iterations are shown in Table 1.

n	p_n	q_n	r_n
0	0	-1	1
1	0	0.5	0.5
2	0.62361	0.083333	0.25
3	0.62361	0.458333	0.125
4	0.779512	0.354167	0.0625
5	0.779512	0.447917	0.03125
6	0.818488	0.421875	0.015625
7	0.818488	0.445312	0.007813
8	0.828231	0.438802	0.003906
9	0.828231	0.444661	0.001953
10	0.830667	0.443034	0.000977
...	...	...	...

We observe that $p_{2k} = p_{2k+1}$ for $k \geqslant 0$; the pattern will be rigorously established later. By increasing the number of iterations and zooming the corresponding pictures we can obtain approximate coordinates of $S((2^n))$. The process is illustrated in Figures 6 and 7.

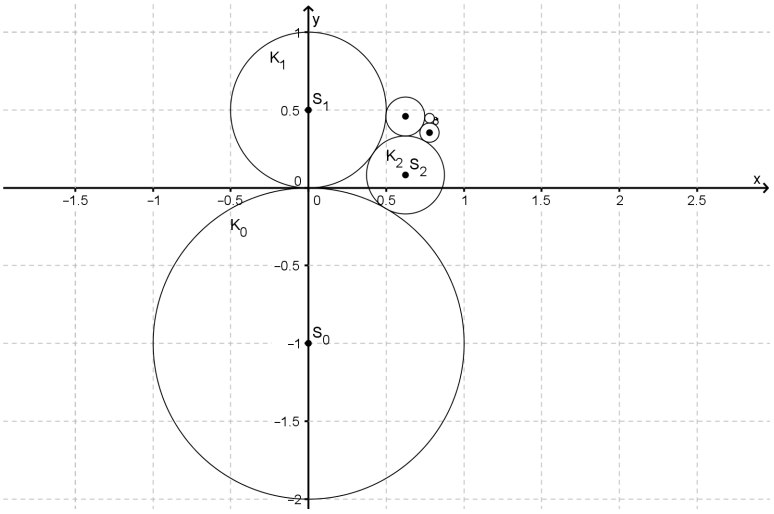


Figure 6: $a_n = 2^n$

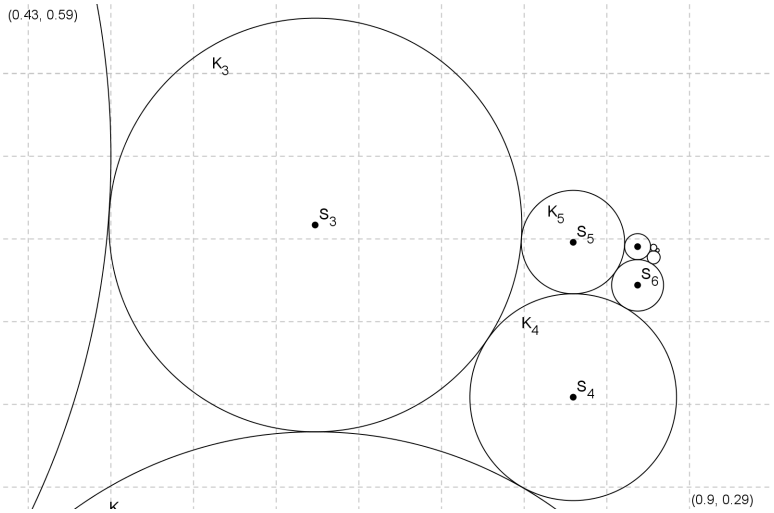


Figure 7: Zoom for $a_n = 2^n$

Example 5 $a_n = n^2$.

Here we omit the table and show only graphical representation (see Figure 8).

4 Exact approach

In this section we first introduce and study the mesh of triangles defined by the circle centers. The mesh is shown in Fig. 9. It is fully enclosed between two piecewise linear curves made of segments connecting the centers of successive even- and odd-numbered circles in the packing. For a given admissible sequence (a_n), let us denote by T_n the triangle with vertices S_{n-1}, S_n, S_{n+1} for $n \geqslant 1$ if $a_0 \neq 0$ and for $n \geqslant 2$ if $a_0 = 0$. Its sides have lengths $\frac{1}{a_n} + \frac{1}{a_{n+1}}$, $\frac{1}{a_{n-1}} + \frac{1}{a_{n+1}}$, and $\frac{1}{a_n} + \frac{1}{a_{n-1}}$. By using Heron's formula we obtain the area of T_n as

$$P_n = \frac{\sqrt{a_n a_{n-1} + a_{n-1} a_{n+1} + a_n a_{n+1}}}{a_{n-1} a_n a_{n+1}}.$$

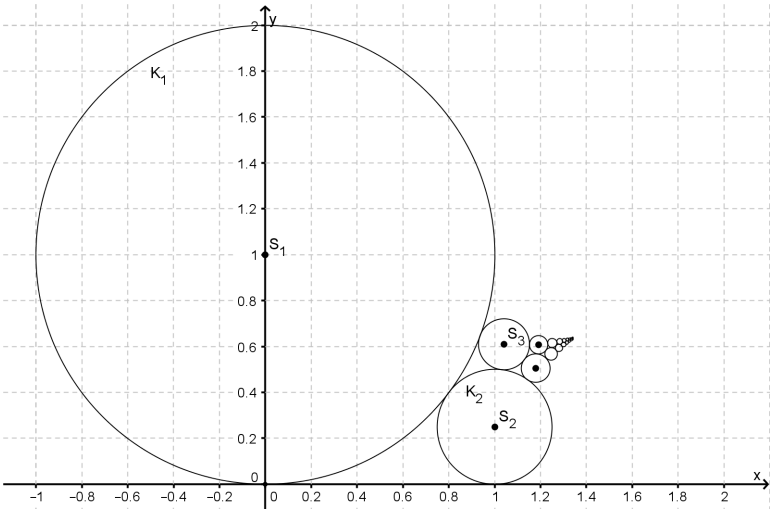


Figure 8: $a_n = n^2$

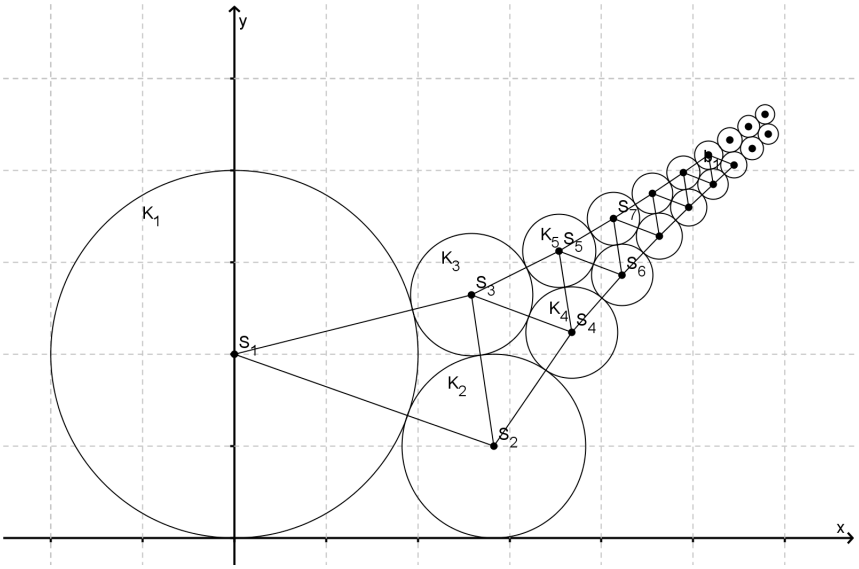


Figure 9: Triangle mesh.

Now we look at the sum of all triangle areas. If the series $\sum_{n=1}^{\infty} P_n$ diverges, then the sequence S_n must also diverge, since an infinite area cannot be enclosed by a circle of a finite radius. This immediately yields a lower bound for the growth rate of sequences with convergent center sequences.

Proposition 6 *Let (a_n) be an admissible sequence and $a_n \in O(\sqrt{n})$. Then $\rho((a_n)) = \infty$.*

Proof. $a_n \in O(\sqrt{n})$ means that there is $C > 0$ such that $a_n \leqslant C\sqrt{n}$, $\forall n$. It is easy to see that for the sequence $b_n = C\sqrt{n}$ the corresponding sequence P_n diverges. Hence $\rho((b_n)) = \infty$, and then, by Proposition 1, also $\rho((a_n)) = \infty$. ■

We notice that convergence of $\sum_{n=1}^{\infty} P_n$ does not imply the convergence of the sequence of circle centers, as a finite area can be enclosed between two curves of infinite length. Similarly, the divergence of the sum of distances between the centers of successive circles does not imply the divergence of the sequence of centers, since a curve of infinite length can be enclosed within a circle of a finite diameter. A nice example is provided by the Koch curve.

4.1 $a_n = n$

In this subsection we consider and settle the case $a_n = n$. The mesh of triangles is shown in Fig. 10. It is easy to see that both $|\overline{S_2S_4}|+|\overline{S_4S_6}|+|\overline{S_6S_8}|+... = \frac{1}{2}+\sum_{k=2}^{\infty}\frac{1}{k}$ and $|\overline{S_1S_3}| + |\overline{S_3S_5}| + |\overline{S_5S_7}| + ... = 1 + 2 \cdot \sum_{k=1}^{\infty}\frac{1}{2k+1}$ diverge. It does not mean, however, that the sequence of centers diverges. The divergence will follow only if we show that the piecewise linear curves bounding our mesh do not vary wildly, i.e., that they have certain monotonicity properties. In order to establish that monotonicity, look at Fig. 11 and consider angles α_1, α_2, α_3. We would like to establish a lower bound on the sum of those angles $\alpha_1 + \alpha_2 + \alpha_3$. If that lower bound exceeds π, it means that the slopes of successive segments are non-increasing and remain positive. That will further imply that the piecewise linear curve is a graph of a function and hence cannot achieve its infinite length over a finite segment.

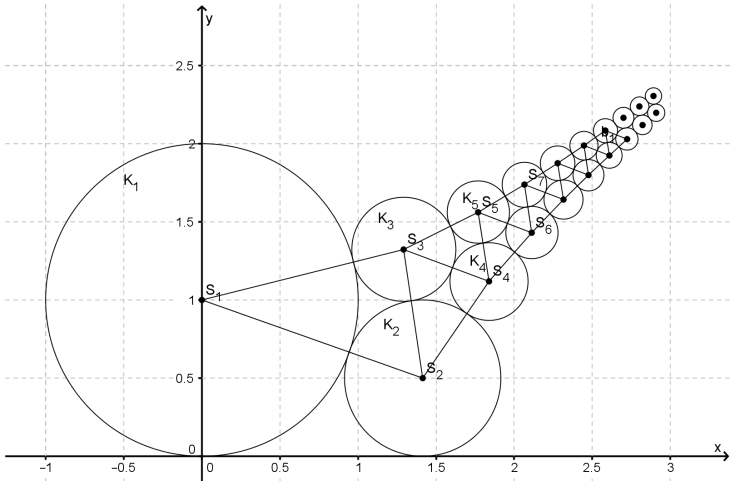


Figure 10: Triangle mesh for $a_n = n$

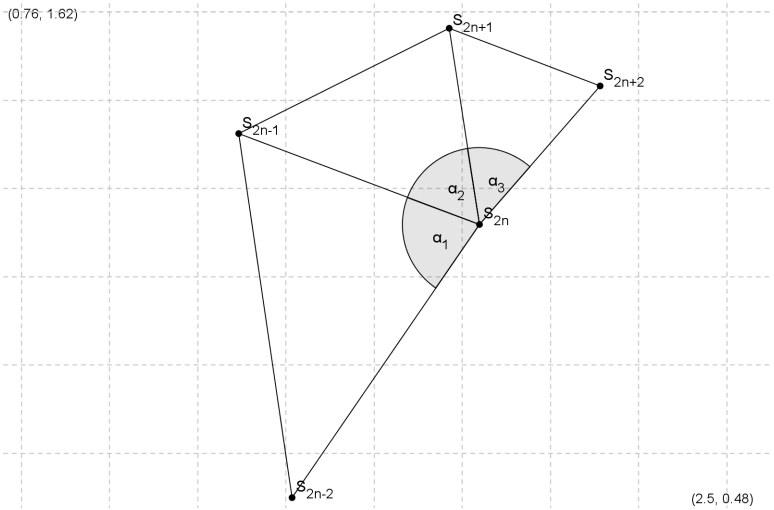


Figure 11: Angles at the center of an even-numbered circle for $a_n = n$

Consider the triangle $S_{2n-2}S_{2n-1}S_{2n}$. By Heron's formula, its area is given by

$$P_{2n-1} = \frac{\sqrt{12n^2 - 12n + 2}}{2n(2n-1)(2n-2)}.$$

From that, by formula $P = \frac{ab\sin\gamma}{2}$, we obtain

$$\sin\alpha_1 = \frac{4n\sqrt{12n^2-12n+2}}{16n^2-12n+2}.$$

Consider now the function $f(n) := \frac{4n\sqrt{12n^2-12n+2}}{16n^2-12n+2}$. It tends toward $\frac{\sqrt{3}}{2}$ as $n \to \infty$. By taking derivative with respect to n it is easy to see that f is decreasing for $n > \frac{5+\sqrt{13}}{4}$. Hence, $f(n)$ is approaching the value $\frac{\sqrt{3}}{2}$ from above for $n > \frac{5+\sqrt{13}}{4} \approx 2.15$. From there it follows $\sin\alpha_1 > \frac{\sqrt{3}}{2}$ and hence $\alpha_1 > \frac{\pi}{3}$.

In a similar way, by considering the triangle $S_{2n-1}S_{2n}S_{2n+1}$ with area

$$P_{2n} = \frac{\sqrt{12n^2-1}}{2n(2n-1)(2n+1)},$$

we obtain

$$\sin\alpha_2 = \frac{4n\sqrt{12n^2-1}}{16n^2-1}.$$

The function $g(n)$ defined by $g(n) := \frac{4n\sqrt{12n^2-1}}{16n^2-1}$ also approaches the value $\frac{\sqrt{3}}{2}$ as $n \to \infty$ remaining above this value. From there we have $\sin\alpha_2 > \frac{\sqrt{3}}{2}$ and hence $\alpha_2 > \frac{\pi}{3}$.

Finally, consider the triangle $S_{2n}S_{2n+1}S_{2n+2}$. Its area is given by

$$P_{2n+1} = \frac{\sqrt{12n^2+12n+2}}{2n(2n+1)(2n+2)},$$

from where it follows

$$\sin\alpha_3 = \frac{4n\sqrt{12n^2+12n+2}}{16n^2+12n+2}.$$

Here, however, the function $h(n) := \frac{4n\sqrt{12n^2+12n+2}}{16n^2+12n+2}$ approaches its limit $\frac{\sqrt{3}}{2}$ from below. Hence, $\alpha_3 < \frac{\pi}{3}$, and we cannot conclude that the sum of three angles in the center of an even-numbered circle exceeds π.

The situation can be saved, however, if we manage to show that a sum of α_3 and one of the remaining angles is at least $\frac{2\pi}{3}$. It will follow if we prove, for example, $\arcsin f(n) + \arcsin h(n) > \frac{2\pi}{3}$. By starting from

$$\arcsin f(n) + \arcsin h(n) = \pi - \arcsin\left[f(n)\sqrt{1-h(n)^2} + h(n)\sqrt{1-f(n)^2}\right]$$

and plugging in the expressions for $f(n)$ and $h(n)$ we obtain that $f(n)\sqrt{1-h(n)^2}+h(n)\sqrt{1-f(n)^2}$ remains below $\frac{\sqrt{3}}{2}$ and tends increasingly toward this value. As this is the argument of the arcsin function on the right-hand side of the above expression, it means that the value of this arcsin function is below $\pi/3$, and then the whole right-hand side exceeds the value of $\frac{2\pi}{3}$, as desired.

Now we can establish the main result of this subsection.

Theorem 7 *The sequence of circle centers for the packing with $a_n = n$ diverges.*

Proof. Consider the piecewise linear curve $L = S_2S_4S_6....$ It is a graph of a continuous function $\varphi(x)$ defined on some interval starting at p_2, the abscissa of S_2. The length of L is then given by

$$l(L) = \int_{p_2}^{M} \sqrt{1+\varphi'(x)^2}dx,$$

where the integral is taken over the domain of the function $\varphi(x)$. Since the integrand is bounded on the whole domain, it follows that the infinite length of L can be achieved only if $M = \infty$. ■

Corollary 8 *Let (a_n) be an admissible sequence with $a_n \in O(n)$. Then $\rho((a_n)) = \infty$.*

Hence we have established divergence of circle centers for all packings with curvatures growing slower than the sequence of natural numbers. This, together with Theorem 7 and Corollary 3, settles the problem of convergence. Now it remains to look at convergent cases and see what we can say about their limits.

4.2 $a_n = q^n$

We have observed in section 3 that the abscissas of centers in the packing of circles whose curvatures are given by a geometric sequence appear in pairs. That means

that the segments connecting the centers of circles whose indices have the opposite parity are vertical. This is not accidental; it is a consequence of the fact that for a geometric sequence of curvatures all triangles T_n are similar, as it is shown in Figure 12. The proof is straightforward and we omit it.

A direct consequence of similarity of all T_n is that the centers of all circles with

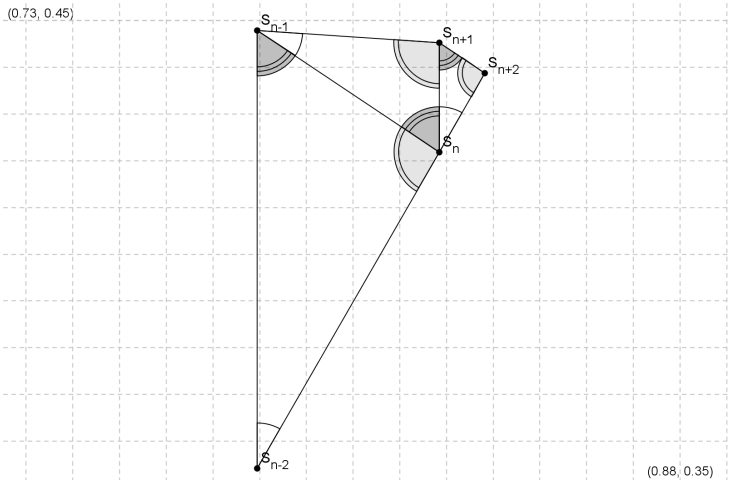


Figure 12: Triangle mesh for $a_n = q^n$

indices of the same parity lie on a line. From there, it follows that the sum of all lengths of segments connecting the center of even-numbered circles is actually the distance between S_0 and the limit. Since the same conclusion is valid for the sum of distances between the centers of odd-numbered circles, the limit must lie on the intersection of two circles with known centers and known radii. Hence, we have exact and explicit expressions for the coordinates of the limit $S((q^n))$. The proof is straightforward and we omit it.

Theorem 9 *Let (a_n) be a geometric sequence with quotient $q > 1$, $a_n = q^n$. Then*

$$S((q^n)) = \left(\frac{2\sqrt{q+q^2+q^3}}{(q+1)(q^2-1)}, \frac{2q}{(q+1)(q^2-1)} \right) \quad \text{and} \quad \rho((q^n)) = \frac{2\sqrt{q}}{q^2-1}.$$

For our example $a_n = 2^n$ this yields $S((2^n)) = \left(\frac{2\sqrt{14}}{9}, \frac{4}{9}\right)$ and $\rho((2^n)) = \frac{2\sqrt{2}}{3}$.

We have shown that a geometric sequence of curvatures forces similarity of triangles T_n. It can be shown that also the opposite is valid, hence that all T_n are similar if and only if the sequence of curvatures is geometric. It is unclear whether this means that the geometric sequences are the only ones allowing exact solutions for the limit.

5 Estimates for convergent sequences

We know from section 2 that the sequence of circle centers converges for all curvature sequences (a_n) such that $\sum \frac{1}{a_n}$ converges. In this section we present a method to estimate the position of the limit $S((a_n))$.

We start by finding the centers S_1 and S_2. Then we compute two sums. The first one is the total length of all segments between the centers of successive odd-numbered circles; the second one is the total length of all segments connecting the centers of successive even-numbered circles. Since $\sum \frac{1}{a_n}$ converges, both sums must be finite. Formally,

$$R_1 = \sum_{k=1}^{\infty} \left[\frac{1}{a_{2k-1}} + \frac{1}{a_{2k+1}} \right], \qquad R_2 = \sum_{k=1}^{\infty} \left[\frac{1}{a_{2k}} + \frac{1}{a_{2k+2}} \right].$$

In general case, R_1 and R_2 are lengths of piecewise linear curves connecting S_1 and S_2, respectively, with $S((a_n))$. Hence they can serve as estimates of the distances from S_1 and S_2, respectively, to $S((a_n))$. It means that $S((a_n))$ must be closer to both S_1 and S_2 than the point S' at the intersection of circles centered at S_1 and S_2 with radii R_1 and R_2, respectively.

We illustrate the quality of those estimates on two examples.

Example 10 $a_n = n^2$.

It follows immediately from the well known expression $\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$ that the expressions for sums of reciprocal values of squares of even and odd numbers are given by

$$\sum_{n=0}^{\infty} \frac{1}{(2n+1)^2} = \frac{\pi^2}{8}, \qquad \sum_{n=1}^{\infty} \frac{1}{(2n)^2} = \frac{\pi^2}{24}.$$

From there we readily obtain

$$R_1 = 2 \cdot \frac{\pi^2}{8} - 1 \approx 1.4674011, \qquad R_2 = 2 \cdot \frac{\pi^2}{24} - \frac{1}{4} \approx 0.572467.$$

By finding intersections of the circle centered at $(0, 1)$ of radius R_1 with the circle centered at $(1, \frac{1}{4})$ of radius R_2, we obtain two points. The one with both coordinates positive, $S' = (1.4218, 0.6370)$, can serve as an estimate of $S((n^2))$. We see from Fig. 13 that the estimate is not very good. This is a consequence of the fact that the piecewise linear curves connecting S_1 and S_2 to $S((n^2))$ are quite far from straight segments, leading thus to overestimates of the corresponding distances.

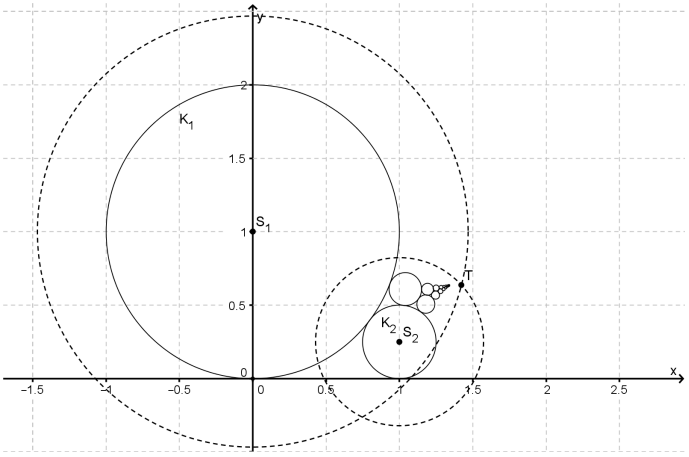


Figure 13: Estimate for $a_n = n^2$

Example 11 *Our last example is concerned with Fibonacci numbers,* $a_n = F_n$.

There are explicit expressions for sums of the reciprocal values of odd- and even-indexed Fibonacci numbers. For example,

$$\sum_{n=0}^{\infty} \frac{1}{F_{2n+1}} = \frac{1}{4}\sqrt{5}\,\vartheta_2\left(0, \frac{3-\sqrt{5}}{2}\right)^2 \approx 1.824515,$$

$$\sum_{n=1}^{\infty} \frac{1}{F_{2n}} = \frac{\sqrt{5}\left(2\psi^{(0)}_{\phi^{-4}}(1) - 4\psi^{(0)}_{\phi^{-2}}(1) + \ln 5\right)}{8\ln\frac{1+\sqrt{5}}{2}} \approx 1.535371.$$

(Here ϑ and ψ denote the theta and digamma functions, respectively, and ϕ is the golden ratio [5]. However, we are interested mostly in the approximate values.) Starting from the above expressions we can derive the approximate values $R_1 = 2.64903$ and $R_2 = 2.070742$ for the radii of circles centered at $(0,1)$ and $(2,1)$, respectively. Their intersection with coordinates $S' = (1.68235, 3.04623)$ yields a pretty accurate estimate of the limit $S((F_n))$. We do not have an exact expression, but the value of $S_{20} = (1.67851, 3.04503)$ given by our algorithm is both close to S' and closer to S_1 and S_2 than S' is to either of them, consistent with the fact that S' is an overestimate. Figures 14 and 15 show the quality of the estimate. We can

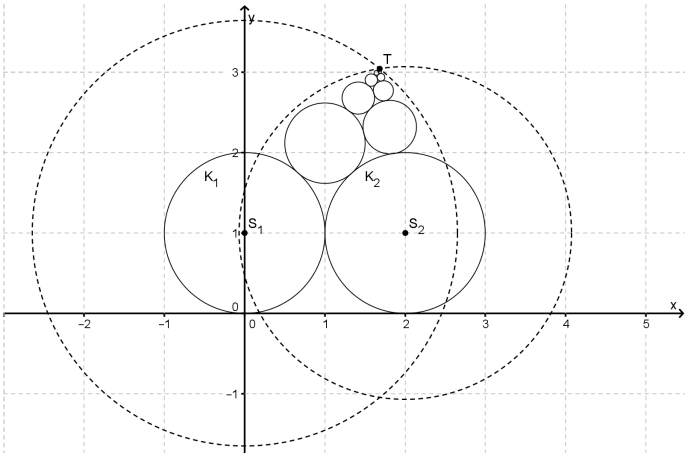


Figure 14: Estimate for $a_n = F_n$

observe that the estimate for Fibonacci numbers is much better than for the squares of natural numbers. The explanation is that the Fibonacci numbers asymptotically

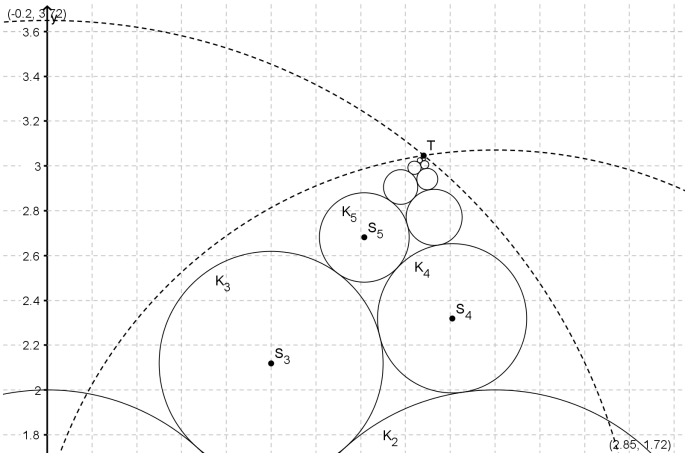


Figure 15: Zoom for $a_n = F_n$

behave as $F_n \sim \phi^n$ for large n. Hence they behave almost like geometric sequences, and for geometric sequences we have exact solutions.

6 Concluding remarks

In this paper we have analyzed behavior of directed packings of circles in the plane when their curvatures are given by a non-decreasing sequence of non-negative real numbers. We have established that the sequence of circle centers converges if and only if the growth rate of the sequence of curvatures strictly exceeds the growth rate of the sequence of natural numbers. For the case of geometric sequences $a_n = q^n$, $q > 1$, we obtained exact coordinates of the limit $S((q^n))$, while in some other convergent cases we obtained good quality upper bounds on its distance from the origin.

It would be interesting to investigate if some other classes of sequences also allow explicit expressions for coordinates of their limits. Another interesting thing to do would be to consider three- (and maybe even higher-) dimensional analogues. We

believe that at least in the three-dimensional case it should be possible to obtain some exact results.

References

[1] L. Fejes Toth, *Lagerungen in der Ebene auf der Kugel und im Raum*, Springer-Verlag, Berlin, Heidelberg, New York, 1972.

[2] T. C. Hales, S. P. Ferguson, *The Kepler Conjecture: The Hales-Ferguson Proof*, New York, 2011.

[3] T. Hales, M. Adams, G. Bauer, T. D. Dang, J. Harrison, L. T. Hoang, C. Kaliszyk, V. Magron, S. McLaughlin, T. T. Nguyen, Q. T. Nguyen, T. Nipkow, S. Obua, J. Pleso, J. Rute, A. Solovyev, T. H. A. Ta, N. T. Tran, T. D. Trieu, J. Urban, K. Vu, R. Zumkeller, A Formal Proof of the Kepler Conjecture, *Forum of Mathematics, Pi* 5: e2. doi:10.1017/fmp.2017.1

[4] *The On-Line Encyclopedia of Integer Sequences*, published electronically at https://oeis.org

[5] *Wolfram Alpha*, https://www.wolframalpha.com/

Appendix

```
m=20;   % number of circles
n=1:1:m;
a=n.^2; %a_{n} % setting the sequence
p(1)=0;
q(1)=1/a(1);
q(2)=1/a(2);
r(1)=abs(q(1));
r(2)=abs(q(2));
r=abs(1./a(n));  %r_{n}
p(2)=p(1)+sqrt((r(1)+q(1))*(2*r(2)+r(1)-q(1))); % if a_{0} = 0
%p(2)=1; % if a_{0} <> 0
for i=3:1:m
    k(i)=(p(i-2)-p(i-1))./(q(i-1)-q(i-2));   %kn
    l(i)=(r(i-2).^2-r(i-1).^2+2*r(i-2)*r(i)-2*r(i-1)*r(i)-p(i-2).^2+p(i-1).^2-q(i-2).^2+
    q(i-1).^2)./(2*q(i-1)-2*q(i-2)); %ln
    c(i)=-r(i-2).^2-2*r(i-2)*r(i)-r(i).^2+p(i-2).^2-2*q(i-2)*l(i)+q(i-2).^2+l(i).^2;  %cn
    a1(i)=1+k(i).^2;    %a'n
    b(i)=2*k(i)*l(i)-2*p(i-2)-2*q(i-2)*k(i);    %bn
    p(i)=(-b(i)+sqrt(b(i).^2-4*a1(i)*c(i)))./(2*a1(i));  %pn
    q(i)=k(i)*p(i)+l(i);    %qn
end
p=p';
q=q';
r=r';
for k=1:1:(m-2)
   area(k)=sqrt(r(k)*r(k+1)*r(k+2)*(r(k)+r(k+1)+r(k+2)));
end
area=area';
name='coordinates.xlsx';
xlswrite(name,p,'coordinates','B2')
xlswrite(name,q,'coordinates','C2')
xlswrite(name,r,'coordinates','D2')
xlswrite(name,area,'coordinates','E2')
```

Some results on Wiener index of a graph: an overview

Martin Knor[*] Snježana Majstorović[†] Riste Škrekovski[‡]

Abstract

The Wiener index $W(G)$ of a connected graph G is defined as the sum of distances between all pairs of vertices in G. In 1991, Šoltés [9] posed the problem of finding all graphs G such that equality $W(G) = W(G - v)$ holds for all vertices v in G. The only known graph with this property is the cycle C_{11}. Our main object of study is the relaxed version of this problem: find graphs for which Wiener index does not change when a particular vertex v is removed. This overview contains results which were obtained and published during the past two years concerning relaxed Šoltés's problem.

Keywords: Wiener index, transmission, Cartesian product, induced subgraph

MSC: 05C12, 05C90

1 Introduction

Average distance is one of the three most robust measures of network topology, along with its clustering coefficient and its degree distribution. Nowadays it has been frequently used in sociometry and the theory of social networks [4]. Wiener index, defined as the sum of distances between all (unordered) pairs of vertices in a graph, besides its crucial role in the calculation of average distance, is the most famous topological index in mathematical chemistry. It is named after Wiener [10], who introduced it in 1947 for the purpose of determining boiling points of alkanes. Since

[*]Slovak University of Technology in Bratislava, Faculty of Civil Engineering, Department of Mathematics, Bratislava, Slovakia. E-Mail: knor@math.sk

[†]Department of Mathematics, Josip Juraj Strossmayer University of Osijek, Osijek, Croatia. E-Mail: smajstor@mathos.hr

[‡]FMF, University of Ljubljana & Faculty of Information Studies, Novo Mesto & FAMNIT, University of Primorska, Slovenia. E-Mail: skrekovski@gmail.com

DOI: https://doi.org/10.5592/CO/CCD.2018.04

then Wiener index has become one of the most frequently used topological indices in chemistry, since molecules are usually modeled by undirected graphs. Other applications of this graph invariant can be found in crystallography, communication theory and facility location. Wiener index has also been studied in pure mathematics under various names: the gross status, the distance of a graph, the transmission of a graph etc. It seems that the first mathematical paper on Wiener index was published in 1976 [3]. Since then, a lot of mathematicians have studied this quantity very extensively. A great deal of knowledge on Wiener index is accumulated in survey papers [2, 5, 11]. Nowadays it has been frequently used in sociometry and the theory of social networks [4].

Throughout this paper all graphs will be finite, simple and undirected.

The *Wiener index* $W(G)$ of a connected graph G is defined as the sum of distances between all (unordered) pairs of vertices in G:

$$W(G) = \sum_{\{u,v\}\subseteq V(G)} d_G(u,v) = \frac{1}{2}\sum_{v\in V(G)} t_G(v), \tag{1}$$

where the *distance* $d_G(u,v)$ between vertices u and v is defined as the number of edges on a shortest path connecting these vertices in G, and the distance, or transmission, $t_G(v)$ of a vertex $v \in V(G)$ is the sum of distances between v and all other vertices of G.

In 1991, Šoltés [9] posed the following problem:

Problem 1. *Find all such graphs G that the equality $W(G) = W(G-v)$ holds for all their vertices v.*

Till now, only one such graph is known: it is a cycle with 11 vertices.

Motivated by Šoltés's problem, in [6] we constructed an infinite family of unicyclic graphs which preserve Wiener index after removal of a particular vertex. In fact, we proved that there are infinitely many unicyclic graphs with this property even when we fix the length of the cycle. Then we showed that for every graph G there are infinitely many graphs H such that G is an induced subgraph of H and $W(H) = W(H-v)$ for some vertex $v \in V(H)\setminus V(G)$. Our research is further extended to graphs in which vertex v is of arbitrary degree, see [7]. For $k \geq 3$ we showed that there are infinitely many graphs G with a vertex v of degree k for which $W(G) = W(G-v)$. Moreover, we proved the existence of such graphs when the degree is $n-1$ or $n-2$. Finally, we showed that dense graphs cannot be a solution of Problem 1.

Our contribution shows that the class of graphs, for which Wiener index does not change when a particular vertex is removed, is rich. This gives hope that Šoltes's problem may have another solution besides C_{11}.

2 Preliminaries

Let G be a connected graph. By $d_G(v)$ we denote the degree of vertex v. A *pendent vertex* is a vertex of degree one and a *pendent edge* is an edge incident with a pendent vertex. For a given vertex v of a graph G, the *eccentricity* of v, denoted by $ecc(v)$ is defined to be the greatest distance from v to any other vertex in G. A *diameter* $\mathrm{diam}(G)$ of a graph G is the value of the greatest eccentricity in G.
By K_n we denote an n-vertex *complete graph* and by S_n an n-vertex *star*. For more definitions and terminologies in graph theory, see [1]. For Wiener index of the path P_n and cycle C_n we have very simple formulae. Wiener index of path P_n is

$$W(P_n) = \binom{n+1}{3}, \tag{2}$$

and Wiener index of a cycle C_n is

$$W(C_n) = \begin{cases} \dfrac{n^3}{8} & \text{if } n \text{ is even} \\ \dfrac{n(n^2-1)}{8} & \text{if } n \text{ is odd.} \end{cases} \tag{3}$$

Proposition 2. *Let G be a connected graph and $v \in V(G)$ be a pendent vertex. Let uv be the corresponding pendent edge in G and $G' = G - v$. Then*

$$W(G) = W(G') + t_{G'}(u) + n(G'),$$

where $n(G')$ is the number of vertices in a graph G'.

The next statement was proved in [8].

Theorem 3. *Let G_u and G_v be two graphs with n_u and n_v vertices, respectively, and let $u \in V(G_u)$, $v \in V(G_v)$.*

(a) If G arises from G_u and G_v by connecting u and v by an edge, then

$$W(G) = W(G_u) + W(G_v) + n_u t_{G_v}(v) + n_v t_{G_u}(u) + n_u n_v.$$

(b) If G arises from G_u and G_v by identifying u and v, then

$$W(G) = W(G_u) + W(G_v) + (n_u - 1)t_{G_v}(v) + (n_v - 1)t_{G_u}(u).$$

3 Results for unicyclic graphs

Theorem 4. *[6] Let $c \geq 5$. There exists infinitely many unicyclic graphs G with a cycle of length c for which equality $W(G) = W(G - v)$ holds for some $v \in V(G)$.*

Proof. Our construction of unicyclic graphs G for which $W(G) = W(G - v)$ goes in the following way. Let C_c be a cycle of length c. We denote its vertices consecutively by $v_0, v_1, \ldots, v_{c-1}$. We add to C_c a pendent vertex, to obtain a new graph, then we add another pendent vertex (which may be connected to previously added vertex) and so on, until we get a unicyclic graph G with $W(G) = W(G - v_0)$. Then we continue with adding pendent vertices to create infinitely many graphs G with the property $W(G) = W(G - v_0)$. Since $G - v_0$ has to be connected, we cannot add pendent vertices to v_0. □

Most of our graphs are obtained from C_c by adding a path to v_{c-1} and a tree to v_1, that is, usually the vertices $v_2, v_3, \ldots, v_{c-2}$ will all have degree 2 in G. By studying the case when $c \in \{3, 4\}$, we conclude that there is no unicyclic graph G with a cycle of length c satisfying $W(G) = W(G - v)$ for some $v \in V(G)$. Justification for this conclusion lies in the fact that $W(G) = W(G - v)$ if and only if the following equality holds

$$t_G(v) = \sum_{\{u_1,u_2\} \subseteq (V(G) \setminus \{v\})} [d_{G-v}(u_1, u_2) - d_G(u_1, u_2)] \,. \tag{4}$$

If $c = 3, 4$, then removal of any vertex of degree two from C_c will not increase distance between any pair of vertices in $G - v$.

Additionally, we showed that a unicyclic graph G on n vertices for which $W(G) = W(G - v)$ exists if and only if $n \geq 9$.

4 Induced subgraphs

By using cycles of certain length, we showed that any graph (even a disconnected one) can be an induced subgraph of some graph H for which $W(H) = W(H - v)$. For this result we needed the following two lemmas.

Lemma 5. *[6] Let C_c be a cycle of even length, $c = 2a$, such that a is a square. Moreover, let G_m be a graph with a vertex u for which $t_{G_m}(u) = \frac{a}{3}[a^2 - 6a + 2]$. Let H be obtained from G_m and C_c by identifying u with v_i, where $i = a - \sqrt{a}$. Then $W(H) - W(H - v_0) = 0$.*

Lemma 6. *[6] Let C_c be a cycle of odd length, $c = 2a+1$, such that $4a+1$ is a square. Moreover, let G_m be a graph with a vertex u for which $t_{G_m}(u) = \frac{a}{6}[2a^2-9a-5]$. Let H be obtained from G_m and C_c by identifying u with v_i, where $i = \frac{1}{2}(2a+1-\sqrt{4a+1})$. Then $W(H) - W(H - v_0) = 0$.*

Now using Lemmas 5 and 6 we obtain the following result in which G does not need to be connected.

As a main tool we used Theorem 3b in which one graph contains G as an induced subgraph and the other one is a cycle.

Theorem 7. *[6] Let G be an arbitrary graph. Then there are infinitely many connected graphs H, containing G as an induced subgraph, and such that $W(H) = W(H - v_0)$ for some vertex $v_0 \in V(H) - V(G)$.*

5 Vertex of a fixed degree

Our first observation is that if a vertex v has degree 1 in G, then $W(G) > W(G-v)$. Since case $d_G(v) = 2$ was already studied through unicyclic graphs, we focus on v such that $d_G(v) \geq 3$. Our main result is the following theorem.

Theorem 8. *[7] For every $k \geq 3$ there exist infinitely many graphs G with vertex v such that $d_G(v) = k$ and $W(G) = W(G - v)$.*

Proof. In each case we show the existence of a graph G_1 with a vertex v such that $d_{G_1}(v) = k$ and $W(G_1) = W(G_1 - v)$. Then we construct an infinite class of graphs by attaching to G_1 a new graph G_2 according to Theorem 3, and by taking into a consideration necessary and sufficient condition given by (4) under which the resulting graph H satisfies $W(H) = W(H - v)$. □

If we consider graphs with $n \geq 7$ vertices, then we can study the case when a degree of v is close to n, that is $d(v) = n - 1$ or $d(v) = n - 2$. In this case we can show the existence of at least one graph G such that $W(G) = W(G - v)$. This is stated in the following theorem.

Theorem 9. *[7] Let $n \geq 7$. There exists an n-vertex graph G with vertex v so that $d_G(v) = n - 2$ and $W(G) = W(G - v)$.*

Proof. Let $d(v) = n - 2$. Cases $n = 7, 8, 9, 10$ are considered separately, see [7]. For $n \geq 11$ we take two stars S_3 and S_{n-4} and connect their central vertices with an

edge. We add edges between one pendent vertex of S_3 and $n-10$ pendent vertices of S_{n-4} and denote the resulting graph by G_0. We take a new vertex v and connect it with all vertices of G_0 except the central vertex of S_3. In the resulting graph G we have $d_G(v) = n-2$, $t_G(v) = n$ and $\mathrm{diam}(G) = 2$. From (4) it follows that $W(G) = W(G_0)$.

Let us now consider the case $d(v) = n-1$. Let H be a graph having $n-1$ vertices and m edges and let G be obtained from H by adding a new vertex v and connecting it by an edge to all vertices of H. Then $\mathrm{diam}(G) = 2$ and we have

$$\begin{aligned} W(G) &= t_G(v) + \sum_{\{u,w\}\subseteq V(H)} d_G(u,w) \\ &= n-1+2\binom{n-1}{2} - m \\ &= (n-1)^2 - m. \end{aligned}$$

Since $H = G - v$, we conclude that $W(G) = W(H)$ if and only if

$$W(H) = (n-1)^2 - m. \tag{5}$$

By a computer we checked that for $n-1 \leq 5$ there are no graphs on $n-1$ vertices satisfying (5). Hence we assume that $n-1 \geq 6$. By using (5), for each $n-1 \geq 6$ we managed to construct a graph H on $n-1$ vertices satisfying $W(G) = W(H)$. □

5.1 Graphs with large minimum degree

At last, we prove that dense graphs cannot be particular solutions of Problem 1. By dense graphs we mean those n-vertex graphs in which the minimum degree $\delta(G)$ is at least $n/2$. Our result relies on the observation that for $n \geq 3$ and $\delta(G) \geq n/2$, we have $\mathrm{diam}(G) \leq 2$.

6 Concluding remarks

The quest for graphs other than C_{11} which are solutions of Šoltes's problem has been completely unsuccessful so far. However, by studying a relaxed version of the problem or by focusing on some particular classes of graph, one could get a better insight into the original problem and find one more solution of it, or show that such graphs do not exist.

One can consider regular graphs. Note that asking for a graph to be vertex-transitive will be a as well a solution of the Šoltés's problem.

We can pose the following problems.

Problem 10. *Are there k-regular connected graphs G other than C_{11} for which the equality $W(G) = W(G - v)$ holds for at least one vertex $v \in V(G)$?*

We know that there are no such graphs for $k \geq n/2$.

One can go further and study graphs G for which equation $W(G) = W(G - S)$ holds for a subset S of the vertex set $V(G)$ consisting of at least 2 vertices.

Problem 11. *Find connected graphs G for which*

$$W(G) = W(G - S)$$

for any $S \subset V(G)$, with $|S| \geq 2$.

Our results show the existence of an infinite class of graphs G for which $W(G) = W(G-v)$ for a particular vertex v. It is natural to formulate the following conjecture.

Problem 12. *For a given r, find (infinitely many) graphs G for which*

$$W(G) = W(G - v_1) = W(G - v_2) = \cdots = W(G - v_r)$$

for any distinct vertices $v_1, \ldots, v_r \in V(G)$.

Acknowledgements. The authors acknowledge partial support by Slovak research grants VEGA 1/0026/16, VEGA 1/0142/17 and APVV–15–0220, National Scholarschip Programme of the Slovak Republic SAIA and Slovenian research agency ARRS, program no. P1–0383.

References

[1] R. Diestel, Graph Theory: Electronic Edition 2000 *Springer Verlag - New York* (1997, 2000).

[2] A. A. Dobrynin, R. Entringer and I. Gutman, Wiener index of trees: theory and application, *Acta Appl. Math.* **66** (2001) 211–249.

[3] R. C. Entringer, D. E. Jackson and D. A. Snyder, Distance in graphs, *Czechoslovak Math. J.* **26** (1976) 283–296.

[4] E. Estrada, The Structure of Complex Networks: Theory and Applications, *Oxford University Press* (2011).

[5] M. Knor, R. Škrekovski and A. Tepeh, Mathematical aspects on Wiener index, *Ars Math. Contemp.* **11** (2016) 327–352.

[6] M. Knor, S. Majstorović, R. Škrekovski, Graphs whose Wiener index does not change when a specific vertex is deleted, *Discrete Appl. Math.* **238** (2018) 126–132.

[7] M. Knor, S. Majstorović, R. Škrekovski, Graphs preserving Wiener index upon vertex removal, *Appl. Math. Comput.* **338** (2018) 25–32.

[8] O. E. Polansky, D. Bonchev, The Wiener number of graphs. I. General theory and changes due to some graph operations, *MATCH Commun. Math. Comput. Chem.* **21** (1986) 133–186.

[9] L'. Šoltés, Transmission in graphs: A bound and vertex removing, *Math. Slovaca* **41** (1991) 11–16.

[10] H. Wiener, Structural determination of paraffin boiling points, *J. Am. Chem. Soc.* **69** (1947) 17–20.

[11] K. Xu, M. Liu, K. C. Das, I. Gutman and B. Furtula, A survey on graphs extremal with respect to distance-based topological indices, *MATCH Commun. Math. Comput. Chem.* **71** (2014) 461–508.

QUESTIONS ON THE STRUCTURE OF PERFECT MATCHINGS INSPIRED BY QUANTUM PHYSICS

MARIO KRENN, XUEMEI GU, AND DÁNIEL SOLTÉSZ

Abstract. We state a number of related questions on the structure of perfect matchings. Those questions are inspired by and directly connected to Quantum Physics. In particular, they concern the constructability of general quantum states using modern photonic technology. For that we introduce a new concept, denoted as *inherited vertex coloring*. It is a vertex coloring for every perfect matching. The colors are inherited from the color of the incident edge for each perfect matching. First, we formulate the concepts and questions in pure graph-theoretical language, and finally we explain the physical context of every mathematical object that we use. Importantly, every progress towards answering these questions can directly be translated into new understanding in quantum physics.

Keywords: perfect matching, inherited vertex coloring, constructable quantum states

MSC: 05C15

1. Motivation

A bridge between quantum physics and graph theory has been uncovered recently [1, 2, 3]. It allows to translate questions from quantum physics – in particular about photonic quantum physical experiments – into a purely graph theoretical language. The question can then be analysed using tools from graph theory and the results can be translated back and interpreted in terms of quantum physics. The purpose of this manuscript is to collect and formulate a large class of questions that concern the generation of pure quantum states with photons with modern technology. This will hopefully allow and motivate experts in the field to think about these issues.

DOI: https://doi.org/10.5592/CO/CCD.2018.05

More concrete, the problems that we present here are concerned with the design of quantum experiments for producing high-dimensional and multi-partite entangled quantum states using state-of-the-art photonic technology [4]. We start by asking for the generation of Greenberger-Horne-Zeilinger (GHZ) states [5], and their high-dimensional generalizations [6, 7, 8, 9], and further generalize the questions to cover arbitrary pure quantum states.

The paper is organized as follows. In Section 2, we rigorously define the graph theoretic questions that turn out to be relevant in quantum physics. In Section 3 we discuss the correspondence between the all mathematical objects used in Section 2 and quantum experiments.

2. Concepts and Questions

The type of quantum experiments, that we are interested in, correspond to so-called bi-colored graphs, that are defined as follows.

Definition 2.1 (Edge bi-colored weighted graph)**.** Let $C = \{c_1, \ldots, c_d\}$ be the set of $d \geqslant 2$ distinct colors. An edge bi-colored weighted graph G=(V(G),E(G)), on n vertices with $d \geqslant 2$ colors is an undirected, loopless graph where there is a fixed ordering of the vertices $v_1, \ldots, v_n \in V(G)$ and to each edge $e \in E(G)$ a complex weight w_e and an ordered pair of (not necessarily different) colors from C is associated. We say that an edge is monochromatic if two associated colors are not different, otherwise the edge is bi-chromatic. Moreover, if e is an edge incident to the vertices $v_i, v_j \in V(G)$ with $i < j$ and the associated ordered pair of colors to e is $(c_1(e), c_2(e))$ then we say that e is colored c_1 at at the endpoint v_i and c_2 at the endpoint v_j.

For simplicity, for the rest of the manuscript we abbreviate *edge bi-colored weighted graph* by *bi-colored graph.*

The unusual property of bi-colored graphs (compared to other edge-colorings in graph theory) is that edges are allowed to have different colors at different endpoints. The next definition will establish a connection between perfect matchings and vertex colorings of a bi-colored graph.

Definition 2.2 (Inherited Vertex Coloring)**.** Let G be a bi-colored graph and let PM denote a perfect matching in G. We associate a coloring of the vertices of G with PM in the natural way: for every vertex v_i there is a

single edge $e(v_i) \in PM$ that is incident to v_i, let the color of v_i be the color of $e(v_i)$ at v_i. We call this coloring the inherited vertex coloring (IVC) of the perfect matching PM and denote it by c. When all vertices in IVC are colored with only one color, we call c a monochromatic coloring.

Now we are ready to define how constructive and destructive interference during an experiment is governed by perfect matchings of a bi-colored graph.

Definition 2.3 (Weight of Vertex Coloring)**.** Let G be a bi-colored graph. Let $\mathcal{M}$ be the set of perfect matchings of G which have the coloring c as their inherited vertex coloring. We define the weight of c as

$$w(c) := \sum_{PM \in \mathcal{M}} \prod_{e \in PM} w_e.$$

Moreover, if $w(c)$=1 we say that the coloring gets unit weight, and if $w(c)$=0 we say that the coloring cancels out.

An example for a bi-colored graph where some colorings of the vertices get unit weight and some other colorings cancel out can be seen in Figure 1.

Question 1: monochromatic graph

For which values of n and d are there bi-colored graphs on n vertices and d different colors with the property that all the d monochromatic colorings have unit weight, and every other coloring cancels out? We call such a graph *monochromatic*.

The only known values of n and d, for which the answer for Question 1 is affirmative, are $d = 2$ and n arbitrary even, and $d = 3$,$n = 4$. For $d = 2$ and n even an alternately colored (all edges are monochromatic) even cycle C_n suffices with all edge weights being one. For $d = 3$, $n = 4$ a suitable bi-colored graph can be constructed as follows. Decompose the edges of the complete graph K_4 into three disjoint perfect matchings, and let the edges of these matchings be monochromatic, and colored with different color, finally assign weight w_e =1 to each edge. It is easy to check that the resulting graph satisfies the conditions of Question 1, see Figure 2. Observe that in all known cases we can use weight 1 for each edge. It was shown by Ilya Bogdanov that no other examples are possible with the restriction that all edge weights are positive [10]. The graph in Figure 1 is not monochromatic.

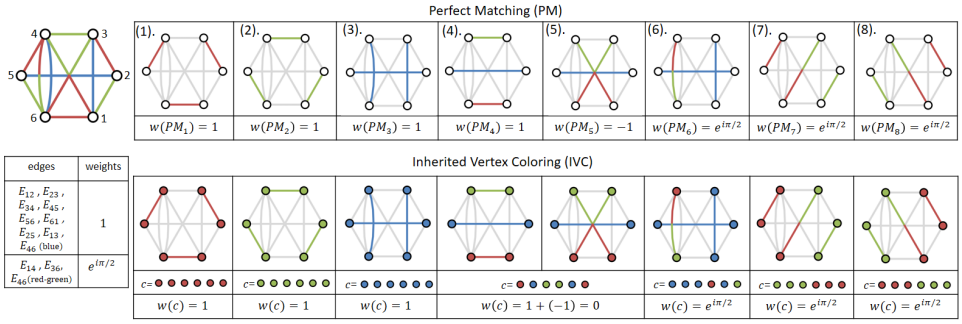


FIGURE 1. Example for inherited vertex coloring and coloring weight. A bi-chromatic weighted edge with one double edge between vertex 4 and 6 is shown on the top left, the edge weights E_{ij} are shown below. On the right top, its eight perfect matchings are shown, and $w(PM_i)$ denotes the product of the edge weights of the perfect matching PM_i. The perfect matching 4 and 5 have the same inherited vertex coloring. As $w(c) = w(PM_4) + w(PM_5) = 0$, we say this coloring cancels out. There are six remaining IVCs with nonzero weights.

In quantum experiments, one can use additional heralding photons in order to produce a certain state. This concept can be formulated in the following way.

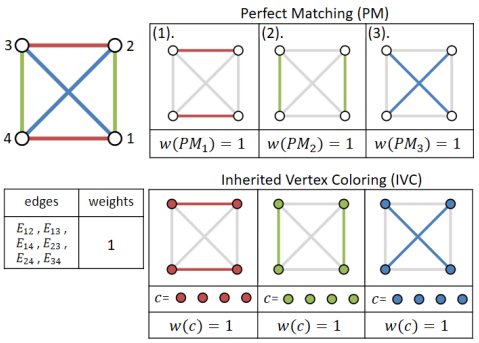


FIGURE 2. A bi-coloring and weight assignment of the edges of K_4 that demonstrated that the answer to Question 1 is affirmative for $d = 3, n = 4$.

Definition 2.4 (k-monochromatic colorings)**.** A coloring c is called k-monochromatic, if the first $k \leqslant |V|$ vertices have the same color, and all other vertices are colored (without loss of generality) red.

Question 2: k-monochromatic Graph

For which values of n, d and k are there bi-colored graphs on n vertices and d different colors with the property that all the d k-monochromatic colorings have unit weight, and every other coloring cancels out? We call such a graph *k-monochromatic*.

The only known example of a k-monochromatic graph with $k > 4$ and $d \geqslant 3$ is shown in Figure 3. There are three 6-monochromatic colorings, where each has $w(c) = 1$. All other colorings are non-6-monochromatic, and have a weight of $w(c) = 0$. We call this graph *Erhard graph*[1]. Note that increasing the number n while keeping k constant can be done straight forwardly. However, increasing k or d seems to be very difficult.

Since it is possible that for large values of n and d, there are no monochromatic graphs, we introduce a measure of monochromaticness on bi-colored graphs as follows.

Definition 2.5 (monochromatic Fidelity)**.** Let N be

$$N = \sum_c |w(c)|^2 ,$$

let $\mathcal{C}^{mono}$ be the set of all monochromatic IVC of G, and d be the number of different colors of G. The *monochromatic fidelity* is defined as

$$F^{mono} := \frac{1}{d}\frac{1}{N}\left|\sum_{c\in\mathcal{C}^{mono}} w(c)\right|^2 .$$

As an example, we can calculate the monochromatic fidelity of the graph in Figure 1. It has d=3 monochromatic inherited vertex colorings and $N = 6$. Then we find that $F^{mono} = \frac{3}{6} = 0.5$. Furthermore, all monochromatic graphs reach the maximum of $F^{mono} = 1$.

[1]It is named after Manuel Erhard, who discovered the quantum mechanical technique which has inspired the construction of this graph.

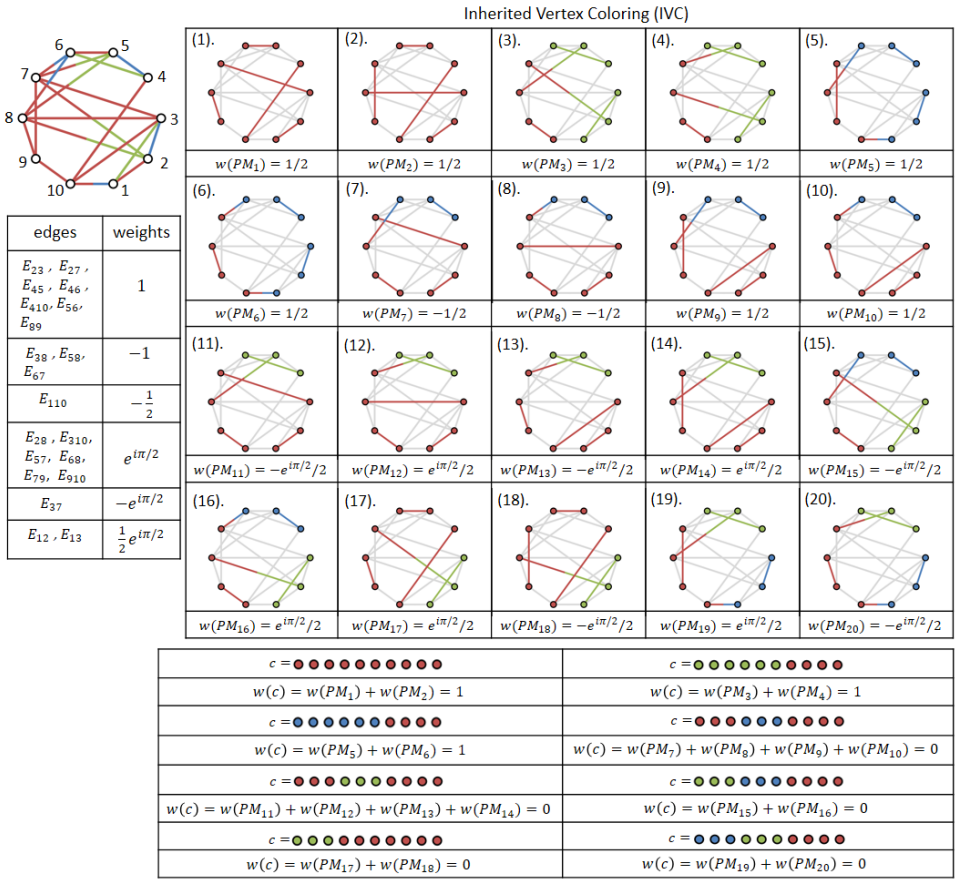


FIGURE 3. The *Erhard graph* is $6-$*monochromatic*. It is the only known example for $k > 4$ and $d \geqslant 3$ satisfying Question 2.

Question 3: approximative monochromatic graph

For every value of n and d, which bi-colored graphs G with n vertices and d different colors maximizes the monochromatic fidelity F^{mono}?

Even if one has access to $n-k$ heralding particles, it is possible that there are no k-monochromatic graphs with d different colors, therefore we can define a fidelity as follows.

Definition 2.6 (k-monochromatic Fidelity)**.** Let N be

$$N = \sum_c |w(c)|^2 ,$$

let $\mathcal{C}^{k-mono}$ be the set of all k-monochromatic IVC of G, and d be the number of different colors of G. The *k-monochromatic fidelity* is defined as

$$F^{k-mono} := \frac{1}{d}\frac{1}{N}\left|\sum_{c\in\mathcal{C}^{k-mono}} w(c)\right|^2 .$$

For k-monochromatic states, the fidelity is $F^{k-mono} = 1$. Naturally, we can ask what graph is closest to monochromatic.

Question 4: approximative k-monochromatic graph

For every value of n, d and k, which bi-colored graphs G with n vertices and d different colors minimizes the k-monochromatic fidelity F^{k-mono}?

Until now, we considered only monochromatic colorings, as they correspond to an important class of quantum states. However, in general we are interested in the total capability of photonic quantum experiments to create quantum states. For that, we generalize our questions such that we cover every pure quantum state.

Question 5: general inherited vertex colorings

Let $\mathcal{C}_p = \{C_i\}_{i=1}^t$ be a set of (prescribed) different colorings of n vertices and $\mathcal{W}_p = \{w_i\}_{i=1}^t$ be the set of (prescribed) weights. For every $\mathcal{C}_p$ and $\mathcal{W}_p$, is there a bi-colored graph G on the same n vertices as the colorings in $\mathcal{C}_p$ so that for each i, $w(C_i) = w_i$, and every coloring not in $\mathcal{C}_p$ cancels out?

A particularly interesting special case of this question is the case where $\mathcal{C}_p$ is restricted to contain only $d = 2$ colors. As an example, we consider the set of colorings $\mathcal{C}_p = ((g,r,r,r),(r,g,r,r),(r,r,g,r),(r,r,r,g))$ and weights $\mathcal{W}_p = (1,1,2,i)$. Is there a graph which is affirmative to Question 5 with these colorings and weights? We answer this question affirmatively, and show the solution in Figure 4.

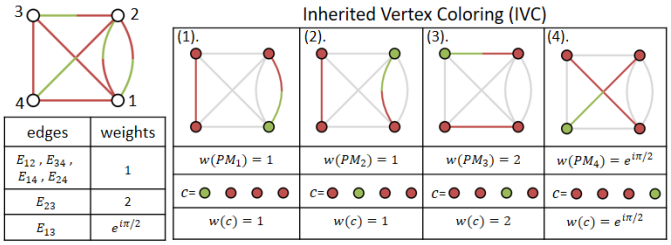


FIGURE 4. This multi-edge graph answers the Question 5 for a given $\mathcal{C}_p$ and $\mathcal{W}_p$.

Again, it might be the case that not every set of coloring and weight can be constructed, thus we define a fidelity that gives us a notion of distance between the target and the graph.

Definition 2.7 (general fidelity)**.** Let $\mathcal{C}_p = \{C_i\}_{i=1}^t$ be a set of (prescribed) different colorings (with up to d different colors), and $\mathcal{W}_p = \{w_i\}_{i=1}^t$ be the set of (prescribed) weights, let G be a bi-colored graph. Let N_1 and N_2 be

$$N_1 = \sum_{i=1}^{t} |w_i|^2 , \quad N_2 = \sum_{\forall c} |w(c)|^2 .$$

The *general fidelity* is defined as

$$F^{general} := \frac{1}{N_1 N_2} \left| \sum_{i=1}^{t} w_i \cdot w(C_i) \right|^2 .$$

Now a natural and most general question can be stated as follows.

Question 6: approximative general graph

For every $\mathcal{C}_p$ and $\mathcal{W}_p$, which bi-colored graphs G with n vertices minimizes the general fidelity $F^{general}$?

Question 6 contains Question 1-5 as special cases. Thus its resolution would resolve the question about the power of modern photonic quantum entanglement sources.

3. Quantum Mechanical Formulation

All of the concepts, questions and partial results in this paper can directly be translated into the language of quantum mechanics [1, 2, 3].

Undirected Graphs correspond to quantum optical experiments, using probabilistic photon-pair sources and linear optics.

Vertices correspond to single photon detectors in the output of some photon path.

Edges correspond to photon pairs that emerge from two photon paths.

Edge weights correspond to the amplitude of the corresponding photon pair.

Edge colors correspond to the mode number of the two photons in the path defined by the vertices at the endpoint of the edge. They can be bi-colored, as the two photons can have different mode numbers. A monochromatic edge corresponds to a photon pair with the same mode number.

Perfect matchings correspond to a multi-photon event where each single photon detector detects a photon. The coherent sum of all perfect matchings leads to the quantum state (conditioning on the click of each detector). Not every perfect matching necessarily leads

to an unique term in the quantum state. Different perfect matchings can lead to the same inherited vertex colorings, thus coherently sum up and constructively or destructively interfere.

Inherited vertex colorings corresponds to multi-photonic terms with different mode numbers in the quantum state. Terms with different IVCs are orthogonal.

Weights of vertex colorings $w(c)$ correspond to the amplitude of terms with mode numbers described by the inherited vertex colorings. More than one perfect matching can lead to the same inherited vertex colorings. As these terms can have opposite weights, it could be that the weight of an inherited vertex coloring is zero even though there are several perfect matchings leading to that coloring with nonzero weights.

Monochromatic vertex colorings lead to terms where every photon carries the same mode number. A graph with only monochromatic vertex colorings (with d different colors) corresponds to d-dimensional Greenberger-Horne-Zeilinger state. These states are of significant importance in quantum physics.

Question 1 asks which high-dimensional Greenberger-Horne-Zeilinger states can be created if general amplitudes $w_i \in \mathbb{C}$ can be used, but without trigger photons.

Monochromatic Graph corresponds to a high-dimensional multi--photonic Greenberger-Horne-Zeilinger state.

Bogdanov's Lemma states that Greenberger-Horne-Zeilinger states can be created only with $d = 3$ dimensions with $n = 4$ photons, or $d = 2$ dimensions for arbitrary even number of n photons, if all amplitudes are real valued (i.e. no destructive interference happens) and no additional trigger photons are used [10].

Figure 2 corresponds to a 4-photon 3-dimensional Greenberger-Horne-Zeilinger state.

k-monochromatic colorings correspond to quantum states where the first k photons have the same mode number, and the remaining $(n-k)$ photons have mode number zero (we can define red to be an arbitrary mode number). The $(n-k)$ red vertices can be used as trigger photons that herald an k-photon state where every photon has the same mode number.

Question 2 asks which high-dimensional Greenberger-Horne-Zeilinger states can be created if general amplitudes $w_i \in \mathbb{C}$ can be used, and $(n-k)$ trigger photons can be used.

Erhard graph is the only known example which corresponds to a quantum state that goes beyond Bogdanov's limit – it can produce a 6-photon 3-dimensional entangled GHZ state. Four heralding photons and complex weights are used to cancel out all non-monochromatic colorings. It is created using two copies of the graph in Figure 2, which are merged using a quantum technique discovered by Manuel Erhard.

Monochromatic fidelity stands for a quantum fidelity to a high-dimensional n-particle GHZ state.

Question 3 asks for every d-dimensional and n-particle state, what is the state that comes closest to the GHZ state, allowing only linear optics and probabilistic pair sources.

k-monochromatic fidelity stands for a quantum fidelity to a high-dimensional k-particle GHZ state, using $(n-k)$ trigger photons.

Question 4 asks for every d-dimensional and k-particle state with $(n-k)$ triggers, what is the state that comes closest to the GHZ state,

allowing only linear optics, probabilistic pair sources and heralding photons.

Question 5 asks in general, which high-dimensional multipartite pure quantum states can be created using these techniques?

Figure 4 is an example to produce a 4-particle W state.

General fidelity corresponds to a fidelity between a prescribed quantum state, and a quantum state that originates from a bi-colored graph.

Question 6 asks for an arbitrary pure quantum state, with which fidelity can it maximally be created?

4. Conclusion

Every progress in any of these purely graph theoretical questions can be immediately translated to new understandings in quantum physics. Apart from the intrinsic beauty of answering purely mathematical questions, we hope that the link to natural science gives additional motivation for having a deeper look on the questions raised above.

Acknowledgements

The authors thank Manuel Erhard, Anton Zeilinger, Tomislav Došlić and Roland Bacher for useful discussions and comments on the manuscript. M.K. acknowledges support from by the Austrian Academy of Sciences (ÖAW), by the Austrian Science Fund (FWF) with SFB F40 (FOQUS). X.G. acknowledges support from the National Natural Science Foundation of China (No.61771236) and its Major Program (No. 11690030, 11690032), the National Key Research and Development Program of China (2017YFA0303700), and from a Scholarship from the China Scholarship Council (CSC). D.S. acknowledges support by the National Research, Development and Innovation Office NKFIH, No. K-120706, No. KH-130371 and No. KH-126853.

References

[1] Mario Krenn, Xuemei Gu, and Anton Zeilinger. Quantum experiments and graphs: Multiparty states as coherent superpositions of perfect matchings. *Physical Review Letters*, 119(24):240403, 2017.

[2] Xuemei Gu, Manuel Erhard, Anton Zeilinger, and Mario Krenn. Quantum experiments and graphs ii: Quantum interference, computation, and state generation. *Proceedings of the National Academy of Sciences*, 116(10):4147–4155, 2019.

[3] Xuemei Gu, Lijun Chen, Anton Zeilinger, and Mario Krenn. Quantum experiments and graphs. iii. high-dimensional and multiparticle entanglement. *Physical Review A*, 99(3):032338, 2019.

[4] Jian-Wei Pan, Zeng-Bing Chen, Chao-Yang Lu, Harald Weinfurter, Anton Zeilinger, and Marek Żukowski. Multiphoton entanglement and interferometry. *Reviews of Modern Physics*, 84(2):777, 2012.

[5] Daniel M Greenberger, Michael A Horne, and Anton Zeilinger. *Going beyond Bell's theorem*. Springer, 1989.

[6] Junghee Ryu, Changhyoup Lee, Marek Żukowski, and Jinhyoung Lee. Greenberger-horne-zeilinger theorem for n qudits. *Physical Review A*, 88(4):042101, 2013.

[7] Junghee Ryu, Changhyoup Lee, Zhi Yin, Ramij Rahaman, Dimitris G Angelakis, Jinhyoung Lee, and Marek Żukowski. Multisetting greenberger-horne-zeilinger theorem. *Physical Review A*, 89(2):024103, 2014.

[8] Jay Lawrence. Rotational covariance and greenberger-horne-zeilinger theorems for three or more particles of any dimension. *Physical Review A*, 89(1):012105, 2014.

[9] Manuel Erhard, Mehul Malik, Mario Krenn, and Anton Zeilinger. Experimental greenberger–horne–zeilinger entanglement beyond qubits. *Nature Photonics*, 12(12):759, 2018.

[10] Ilya Bogdanov. Graphs with only disjoint perfect matchings. *https://mathoverflow.net/q/267013*, 2017.

Vienna Center for Quantum Science & Technology (VCQ), Faculty of Physics, University of Vienna, Boltzmanngasse 5, 1090 Vienna, Austria.

Institute for Quantum Optics and Quantum Information (IQOQI), Austrian Academy of Sciences, Boltzmanngasse 3, 1090 Vienna, Austria.

present address: Department of Chemistry, University of Toronto, Toronto, Ontario M5S 3H6, Canada.

present address: Vector Institute for Artificial Intelligence, Toronto, Canada.

Email address: mario.krenn@univie.ac.at

State Key Laboratory for Novel Software Technology, Nanjing University, 163 Xianlin Avenue, Qixia District, 210023, Nanjing City, China.

Institute for Quantum Optics and Quantum Information (IQOQI), Austrian Academy of Sciences, Boltzmanngasse 3, 1090 Vienna, Austria.

Email address: xmgu@smail.nju.edu.cn

Alfréd Rényi Institute of Mathematics, Hungarian Academy of Sciences, 13-15 Reáltanoda Street, 1053 Budapest, Hungary.

Email address: solteszd@renyi.hu

Lindström – Gessel – Viennot theorem as a common point of linear algebra and combinatorics

Edin Liđan
University of Bihać
Bosnia and Herzegovina
edin.lidan@unbi.ba

Abstract

Lindström – Gessel – Viennot theorem connects linear algebra and combinatorics with graph theory. We will present proof of LGV theorem and its application on Cauchy – Binnet and generalized Cauchy – Binnet formula as well as calculation of binomial determinants and some other specific determinants.

Keywords: LGV theorem, calculating determinants, weighted directed acyclic graph, path, Cauchy – Binnet formula

MSC: 05A15, 05A05

1 Introduction

Linear algebra and combinatorics are one of the oldest mathematical disciplines which even today significantly influenced further development of other disciplines and computer science. Although if we think about modern mathematics as a collection of many overlapping disciplines whose subjects may look far distant from each other, mathematics was always strongly integrated science with unexpected, mysterious and beautiful links among diverse subjects. Here we present one such deep result which connects determinants and graphs.

On the website of KAIST Math Problem of the Week (Weekly Math Challenges in KAIST) in December 2016 the following problem was posted.

DOI: https://doi.org/10.5592/CO/CCD.2018.06

Problem 1 (Koon and Yun Bum)**.** *Let* $S_n = (a_{ij})_{ij}$ *be an* $n \times n$ *matrix such that*

$$a_{ij} = \binom{2(i+j-1)}{i+j-1}.$$

Find $\det S_n$*.*

The solution of this problem was given by Koon and Yun Bum in 2017. Using the properties of linear algebra and binomial coefficients they reduced matrix an upper triangular matrix which determinant is $S_n = 2^n$. We sketch his solution:
Proof: Let L_n be the lower triangular matrix with entries given by

$$L_n := \begin{cases} \binom{2i-1}{i+j-1} & \text{if } i \geq j \\ 0 & \text{otherwise} \end{cases}$$

and let $U_n := L_n^T$. Note that

$$(L_n U_n)_{ij} = \sum_{k=1}^{2i-1} \binom{2i-1}{k} \binom{2j-1}{k+j-i} \text{for } i \geq j.$$

Observe the following identity

$$\binom{2(i+j-1)}{i+j-1} = \sum_{k=0}^{2i-1} \binom{2i-1}{k} \binom{2j-1}{k+j-i}.$$

As

$$2(L_n U_n)_{ij} = \sum_{k=0}^{2i-1} \binom{2i-1}{k} \binom{2j-1}{k+j-i},$$

so $2(L_n U_n)_{ij} = a_{ij}$, where $S_n = (a_{ij})$. Hence, $S_n = (2L_n)U_n$,

$$\det S_n = (\det 2L_n) \det U_n = 2^n \det L_n \det U_n = 2^n.$$

□

In this paper we will view this matrix as the matrix of path systems of some graph. We will say something about its determinant based on the Lindström – Gessel – Viennot theorem.

Theorem 1.1 (LGV theorem)**.** *Let G be a directed acyclic graph, with a weight function $\omega : E \to \mathbb{R}$, $\mathcal{A} = \{A_1, A_2, \ldots, A_n\}$ and $\mathcal{B} = \{B_1, B_2, \ldots, B_n\}$ be two (not necessarily disjoint) sets of vertices. Let M be the path matrix from $\mathcal{A}$ to $\mathcal{B}$, and let VD be the set of all vertex disjoint path systems of $\mathcal{A}$ to $\mathcal{B}$. Then*

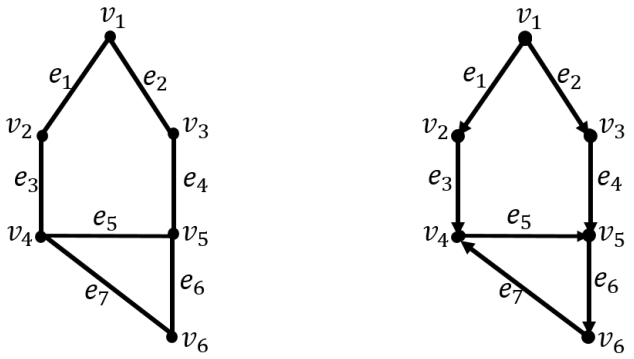
$$\det M = \sum_{\mathcal{P} \in VD} sign(\mathcal{P})\omega(\mathcal{P}).$$

In Section 2, we will define basic terms and properties related to directed weighted acyclic graphs. In Section 3, the proof of LGV theorem and its application in Cauchy-Binnet's and generalized Cauchy - Binnet's formula will be provided. Thus, we will show how to apply LGV theorem to a binomial determinant. In the last section the solution of Problem 1 will be provided using LGV theorem for $n = 2$.

2 Weighted graphs and directed paths in graphs

In this section we review basics of graphs and explain path systems in a graph in order to state Lindström – Gessel – Viennot theorem.

Definition 2.1. *Graph G is pair of sets $G = (V, E)$, where V is a set of vertices, and E is a set of edges, formed by pairs of vertices.*



(a) Graph G

(b) Directed graph G

Figure 1: Example of graphs

For example in Figure 1(a) we have a graph $G = (V, E)$ with the finite set of vertices $V = \{v_1, \ldots, v_6\}$, and finite set of edges $E = \{e_1, e_2, \ldots, e_7\}$. If we give a directions to edges (Figure 1(b)) then we we call a graph directed.

Definition 2.2. *A walk is a sequence v_0, e_1, v_1, ..., v_k of graph vertices v_i and graph edges e_i such that for $1 \leq i \leq k$, the edge e_i has endpoints v_{i-1} and v_i. The length of a walk is its number of edges.*

Definition 2.3. *A path in G is a walk with with all distinct vertices in sequence.*

Definition 2.4. *A trail is a walk v_0, e_1, v_1, ..., v_k with no repeated edge. The length of a trail is its number of edges.*

A u, v trail is a trail with first vertex u and last vertex v, where u and v are known as the endpoints. A walk of trail is closed if the first vertex is equal to last vertex and is the only vertex that is repeated.

Definition 2.5. *A cycle of a graph G is a subset of the edge set of G that forms a path such that the first node of the path corresponds to the last.*

Definition 2.6. *A directed acyclic graph is a graph with directed edges containing no cycles.*

Throughout the paper we consider only simple graphs (no loops and no multiple edges). Let us suppose that for each edge e of graph G it is associated a real number $w(e)$ called its *weight*. Then graph G together with these weights is called a *weighted graph*. If their edges are directed, the graph is called *directed acyclic graph G*. For us, paths of directed weighted acyclic graph G will be the most interesting.

Definition 2.7. *A path system $\mathcal{P}$ is given by a permutation $\sigma \in S_n$ and n paths $P_1 : A_1 \to B_{\sigma(1)}$, $P_2 : A_2 \to B_{\sigma(2)}$, ..., $P_n : A_n \to B_{\sigma(n)}$. Weight of a path system $\mathcal{P}$ is given by*

$$\omega(\mathcal{P}) = \prod_{i=1}^{n} \omega(P_i)$$

where $w(P)$ is the weight of path P and $sign(\mathcal{P}) = sign(\sigma)$. Weight of a path P is defined by the product of the edges in the path

$$\omega(P) = \prod_{e \in P} \omega(e).$$

For a trivial path P (from a vertex v to itself), we define $\omega(P) = 1$.

For two vertices A and B of G we define weight from A to B, with

$$\omega(A,B) = \sum_{P:A\to B} \omega(P). \tag{1}$$

Example 2.1. *We now illustrate the weights of paths from v_1 to v_6 in the following weighted graph G:*

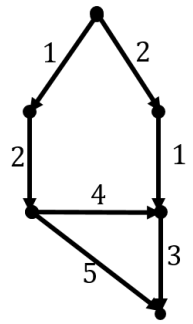


Figure 2: Directed weighted acyclic graph G

There are three such paths $P_1 : e_2e_4e_6$, path $P_2 : e_1e_3e_5e_6$ and path $P_3 : e_1e_3e_7$. Determine now the weight for all possible paths which we noticed and then we obtain that their weights are

$$\begin{aligned}
\omega(P_1) &= \omega(e_2)\omega(e_4)\omega(e_6) = 2\cdot 1\cdot 3 = 6,\\
\omega(P_2) &= \omega(e_1)\omega(e_3)\omega(e_5)\omega(e_6) = 1\cdot 2\cdot 4\cdot 3 = 24,\\
\omega(P_3) &= \omega(e_1)\omega(e_3)\omega(e_7) = 1\cdot 2\cdot 5 = 10.
\end{aligned}$$

Now we deduce that

$$\omega(v_1,v_6) = \omega(P_1)+\omega(P_2)+\omega(P_3) = 6+24+10 = 40.$$

□

Let $\mathcal{A} = \{A_1, A_2, \ldots, A_n\} \subset V$ and $\mathcal{B} = \{B_1, B_2, \ldots, B_n\} \subset V$ be a two subsets of V having the same cardinality n.

Definition 2.8. *Vertex - disjoint path system $\mathcal{P} : \mathcal{A} \to \mathcal{B}$ is a collection of all paths where in every collection there are no two paths P_i, $P_j \in \mathcal{P}$ with a common vertex.*

Definition 2.9. *The matrix of paths* $M = [m_{i,j}]_{i,j=1}^{n}$ *from* $\mathcal{A}$ *to* $\mathcal{B}$ *is defined by*

$$m_{i,j} = \sum_{P:A_i \to B_j} \omega(P) = \omega(A_i, B_j).$$

3 Lindström – Gessel – Viennot theorem

Lindström - Gessel - Viennot Lemma (LGV theorem) or the nonintersecting paths theorem gives some characterization of determinant of the matrix of paths in terms of special path systems. The proof of this theorem was presented by Bernt Lindström (1973) in the context of matroid theory [4], but all beauty of this theorem was revealed by Ira Gessel and Gerard Viennot in their paper *Binomial Determinants, Paths, and Hook Length Formulae* ([2]). In this paper it is described how to apply the theorem to the combinatorics problems. However, we must mention that similar idea appeared earlier in the work of Karlin and McGregor (1959) in a probabilistic framework ("Slater determinant") in quantum mechanics ([3]). Now we will give the proof of LGV theorem.

Proof of LGV theorem: Determinant of $n \times n$ matrix is defined as

$$\det(M) = \sum_{\sigma \in S_n} \left(\operatorname{sign}(\sigma) \prod_{i=1}^{n} m_{i\sigma(i)} \right).$$

Consider $\sigma \in S_n$, where σ is a permutation of set $\{1, 2, \ldots, n\}$

$$\operatorname{sign}(\sigma) \prod_{i=1}^{n} m_{i,\sigma(i)} = \operatorname{sign}(\sigma) m_{1\sigma(1)} m_{2\sigma(2)} \cdots m_{n\sigma(n)},$$

where $m_{i\sigma(i)}$ is the sum of weights of collection path system from A_i to $B_{i\sigma(i)}$. Now apply the definition of weight from some vertex to some other vertex within the graph to get that

$$\begin{aligned} & \operatorname{sign}(\sigma) m_{1\sigma(1)} m_{2\sigma(2)} \cdots m_{n\sigma(n)} \\ = \; & \operatorname{sign}(\sigma) \left(\sum_{P_1 : A_1 \to B_{\sigma(1)}} \omega(P_1) \right) \cdots \left(\sum_{P_n : A_n \to B_{\sigma(n)}} \omega(P_n) \right) \\ = \; & \sum_{\mathcal{P} : P_1 : A_1 \to B_{\sigma(1)}, \ldots, P_n : A_n \to B_{\sigma(n)}} \operatorname{sign}(\mathcal{P}) \omega(\mathcal{P}). \end{aligned}$$

If we make the sum over all σ, we get

$$\det M = \sum_{\mathcal{P}} (\text{sign}(\mathcal{P}_\sigma))\, \omega(\mathcal{P}_\sigma),$$

where $\mathcal{P} = (P_1, P_2, \ldots, P_n)$ is collection of all path systems which run from $\mathcal{A}$ to $\mathcal{B}$ and

$$\mathcal{P}_\sigma = \{\mathcal{P} : \mathcal{P} \text{ system of paths } \mathcal{A} \text{ to } \mathcal{B} \text{ given with } \sigma\}.$$

From the sum over all path systems $\mathcal{P}$ from $\mathcal{A}$ to $\mathcal{B}$ we obtain

$$\det M = \sum_{\mathcal{P}} \text{sign}(\mathcal{P})\omega(\mathcal{P}). \tag{2}$$

Let ND be collection of all path systems which have at least two common vertices. Then we can show the right side of equality (2) as

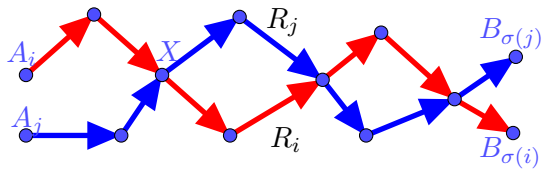
$$\sum_{\mathcal{P}} \text{sign}(\mathcal{P})\omega(\mathcal{P}) = \sum_{\mathcal{P}\in VD} \text{sign}(\mathcal{P})\omega(\mathcal{P}) + \sum_{\mathcal{P}\in ND} \text{sign}(\mathcal{P})\omega(\mathcal{P}).$$

The goal is to show that we have

$$\sum_{\mathcal{P}\in ND} (\text{sign}\mathcal{P})\omega(\mathcal{P}) = 0.$$

For a path system $\mathcal{R} = (R_1, R_2, \ldots, R_n) \in ND$, define

- i to be the smallest index such that R_i intersected with some R_j,
- X to be the first vertex at which R_i intersects some other path $\mathcal{R}$,
- j to be the smallest index of all the paths in $\mathcal{R}$ that intersects R_i u X (equivalently the smallest index of all paths such that $X \in P_i \cap P_j$, $(j > i)$,
- L_{iX} to be part of path R_i from A_i to X, and R_{iX} part of path R_i from X to $B_{\sigma(i)}$, so it is $\omega(R_i) = \omega(L_{iX}) \cdot \omega(R_{iX})$,
- L_{jX} to be part of path R_j from A_j to X, and R_{jX} part of the path R_j from X to $B_{\sigma(j)}$, so it is $\omega(R_j) = \omega(L_{jX}) \cdot \omega(R_{jX})$.

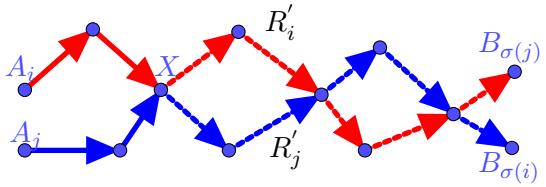


Now, we define an involution φ on ND by setting

$$\varphi : ND \to ND \quad \varphi(\mathcal{R}) = \mathcal{T} = (T_1, T_2, T_3, \ldots, T_n),$$

where $T_k = R_k$ when $k \neq i, j$, and T_i and T_j are defined as

- T_i is the path from A_i using the edges L_{iX} to X, after that we use the edges from R_{jX} to $B_{\sigma(j)}$, so that $\omega(T_i) = \omega(L_{iX}) \cdot \omega(R_{jX})$,
- T_j is the path from A_j using the edges L_{jX} to X, after that we use the edges from R_{iX} to $B_{\sigma(i)}$, so that $\omega(T_j) = \omega(L_{jX}) \cdot \omega(R_{iX})$.



$\mathcal{T} = (R'_1, R'_2, \ldots, R'_n)$ $\mathcal{T}$ have $\sigma' \in S_n$ where is $\sigma' = \sigma \circ (i, j)$

$$\begin{aligned} \text{sign } \sigma' &= \text{sign } \sigma \text{ sign}(i, j) \\ \text{sign } \sigma' &= -\text{sign } \sigma \end{aligned}$$

From which we obtained

$$\begin{aligned} \text{sign } \mathcal{R} &= -\text{sign } \mathcal{T} \\ -\text{sign } \mathcal{R} &= \text{sign } \mathcal{T} \end{aligned}$$

Both path systems $\mathcal{T}$ and $\mathcal{R}$ are contained from the same set of the edges, so

$$\begin{aligned}\omega(\mathcal{T}) &= \prod_{i=1}^{n}\omega(T_i) = \omega(T_1)\omega(T_2)\cdots\omega(T_n)\\ &= \left(\prod_{k\in\{1,2,\ldots,n\}\setminus\{i,j\}}\omega(T_k)\right)\omega(T_i)\omega(T_j)\\ &= \left(\prod_{k\in\{1,2,\ldots,n\}\setminus\{i,j\}}\omega(R_k)\right)\omega(T_i)\omega(T_j).\end{aligned}$$

Based on above obtained paths T_i and T_j it follows that

$$\begin{aligned}\omega(T_i)\omega(T_j) &= (\omega(L_{iX})\cdot\omega(R_{jX}))(\omega(L_{jX})\cdot\omega(R_{iX}))\\ &= (\omega(L_{iX})\cdot\omega(R_{iX}))\cdot(\omega(L_{jX})\cdot\omega(R_{jX}))\\ &= \omega(R_i)\omega(R_j).\end{aligned}$$

Thus, $\omega(\mathcal{T}) = \omega(\mathcal{R})$. From definition φ is an involution so it follows that $\varphi = \varphi^{-1}$, i.e. φ is bijection. Thus we found $1-1$ correspondence for matching pairs of system paths $(\mathcal{R}, \varphi(\mathcal{R}))$ in ND where every pair of system paths is

$$\begin{aligned}\omega(\mathcal{R}) &= \omega(\varphi(\mathcal{R}))\\ \operatorname{sign}(\mathcal{R}) &= -\operatorname{sign}(\varphi(\mathcal{R})).\end{aligned}$$

It follows that

$$\sum_{\mathcal{P}\in ND}\operatorname{sign}(\mathcal{P})\omega(\mathcal{P}) = 0.$$

The Theorem is proved.

□

Apart from the connection with linear algebra, it offers a nice connection between graph theory and combinatorics, which will be illustrated on the theorem and examples.

Proposition 3.1. *For $m, n \in \mathbb{N}$*

$$\det M = \begin{bmatrix} \binom{m}{0} & \binom{m}{1} & \cdots & \binom{m}{n-1} \\ \binom{m+1}{0} & \binom{m+1}{1} & \cdots & \binom{m+1}{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ \binom{m+n-1}{0} & \binom{m+n-1}{1} & \cdots & \binom{m+n-1}{n-1} \end{bmatrix} = 1.$$

Proof: This Proposition we will prove using the LGV theorem. The idea consists of creating the directed weighted acyclic graph which weights of paths is equal to the appropriate binomial coeficient in determinant. On the other words

$$\begin{bmatrix} \binom{m}{0} & \binom{m}{1} & \dots & \binom{m}{n-1} \\ \binom{m+1}{0} & \binom{m+1}{1} & \dots & \binom{m+1}{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ \binom{m+n-1}{0} & \binom{m+n-1}{1} & \dots & \binom{m+n-1}{n-1} \end{bmatrix} = \begin{bmatrix} \omega(A_1, B_1) & \omega(A_1, B_2) & \dots & \omega(A_1, B_n) \\ \omega(A_2, B_1) & \omega(A_2, B_2) & \dots & \omega(A_2, B_n) \\ \vdots & \vdots & \ddots & \vdots \\ \omega(A_n, B_1) & \omega(A_n, B_2) & \dots & \omega(A_n, B_n) \end{bmatrix}.$$

That graph we can construct in the following way (Figure 3). The edges of matrix M are directed on the right (horizontally) and upwards (vertically).

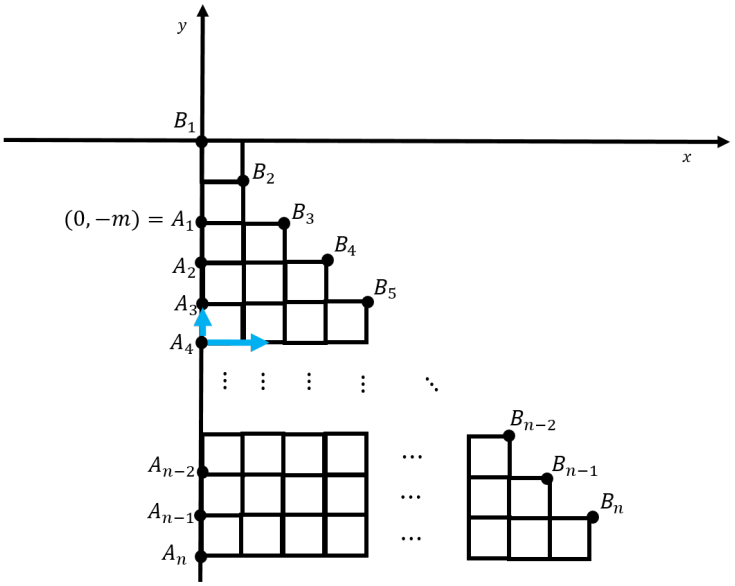


Figure 3: Directed weighted acyclic graph of determinant M

Consider now all disjoint path system in graph constructed graph. If we observe the path from A_1 to B_1, we notice that we have only one such path. In case we start from A_2 upwards, paths would be intersected. Thus, one option is to go right so

that the path from A_2 to B_2 is determined by only one possible path. The analogy applies to all other paths in the graph, which means that only one disjoint paths system exists.

$$\det M = \#\text{number of disjoint path systems} = 1.$$

□

Theorem 3.1 (Chauchy - Binnet's formula)**.** *For every two $n \times n$ square matrices M_1 and M_2*

$$\det(M_1 M_2) = \det(M_1)\det(M_2).$$

Proof: Let us take the following sets of vertices

$$\begin{aligned} \mathcal{A} &= \{A_1, A_2, \ldots, A_n\}, \\ \mathcal{B} &= \{B_1, B_2, \ldots, B_n\}, \\ \mathcal{C} &= \{C_1, C_2, \ldots, C_n\}. \end{aligned}$$

Now we construct directed graph with vertices $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$, where the edges are directed from $\mathcal{A}$ to $\mathcal{B}$ and from $\mathcal{B}$ to $\mathcal{C}$ with weights $\omega(A_i, B_j) = m_1[i,j]$ and $\omega(B_j, C_k) = m_2[j,k]$. If $M = M_1 M_2$, then

$$m[i,j] = \sum_{k=1}^{k} m_1[i,j] m_2[j,k].$$

Consider random system of paths $\mathcal{P}$ in which verices are disjoint from $\mathcal{A}$ to $\mathcal{C}$. $\mathcal{P}$ must go through $\mathcal{B}$. Every system of paths from $\mathcal{A}$ to $\mathcal{C}$ is divided in two parts $\mathcal{Q}$ and $\mathcal{R}$, where $\mathcal{Q}$ is the system of disjoint paths from $\mathcal{A}$ to $\mathcal{B}$, and $\mathcal{R}$ is the system of disjoint paths from $\mathcal{B}$ to $\mathcal{C}$. $\mathcal{W}$ is the set of all disjoint paths systems from $\mathcal{A}$ to $\mathcal{B}$, and $\mathcal{Z}$ is set of all disjoint path systems from $\mathcal{B}$ to $\mathcal{C}$. Now consider

$$\begin{aligned} \det(M_1)\det(M_2) &= \sum_{\mathcal{Q}\in\mathcal{W}} \operatorname{sign}(\mathcal{Q})\omega(\mathcal{Q}) \sum_{\mathcal{R}\in\mathcal{Z}} \operatorname{sign}(\mathcal{R})\omega(\mathcal{R}) \\ &= \sum_{\mathcal{P}\in\mathcal{W}\times\mathcal{Z}} \operatorname{sign}(\mathcal{R})\operatorname{sign}(\mathcal{Q})\omega(\mathcal{R})\omega(\mathcal{Q}), \end{aligned}$$

where $\mathcal{W} \times \mathcal{Z}$ is set of ordered pairs $(\mathcal{Q}, \mathcal{R})$ suitable disjoint paths $P_i : A_i \to B_{\sigma(\mathcal{Q}(i))} \to C_{\sigma(\mathcal{R}(\sigma(\mathcal{Q}(i))))}$, and $\sigma_{\mathcal{Q}} \circ \sigma_{\mathcal{R}} = \sigma$. Then weight of random disjoint path system $\mathcal{P}$ is

$$\omega(\mathcal{P}) = \omega(\mathcal{Q})(\mathcal{R}), \tag{3}$$

and

$$\text{sign}(\sigma) = \text{sign}(\sigma_{\mathcal{Q}} \circ \sigma_{\mathcal{R}}) = \text{sign}(\sigma_{\mathcal{Q}})\text{sign}(\sigma_{\mathcal{R}}), \tag{4}$$

from which it follows

$$\text{sign}\mathcal{P} = \text{sign}(\mathcal{Q})\text{sign}(\mathcal{R}) \tag{5}$$

If we now apply LGV theorem, we get

$$\begin{aligned} \det(M_1)\det(M_2) &= \sum_{\mathcal{P}\in\mathcal{W}\times\mathcal{Z}} \text{sign}(\mathcal{R})\text{sign}(\mathcal{Q})\omega(\mathcal{R})\omega(\mathcal{Q}) \\ &= \sum_{\mathcal{P}} \text{sign}(\mathcal{P})\omega(\mathcal{P}) = \det(M_1 M_2). \end{aligned}$$

□

Theorem 3.2 (Generalized Cauchy - Binet's formula)**.** *Let M_1 be a $n \times r$ matrix and let M_2 be a $r \times n$ matrix where $n \leq r$. Then we have*

$$\det(M_1 M_2) = \sum_{X\subset\{1,2,\ldots,r\},|X|=n} \det(M_1[X])\det(M_2[X]),$$

where $M_1[X]$ is square submatrix formed by columns matrix limited to columns indexed as X and $M_2[X]$ is matrix limited on rows indexed as X.

Proof: Construct directed graph $G = \{\mathcal{A} \cup \mathcal{B} \cup \mathcal{C}$,E$\}$, where is

$$\begin{aligned} \mathcal{A} &= \{A_1, A_2, \ldots, A_n\}, \\ \mathcal{B} &= \{B_1, B_2, \ldots, B_n\}, \\ \mathcal{C} &= \{C_1, C_2, \ldots, C_n\}, \\ E &= \{(A_i, B_j) : i \in \{\overline{1,n}\}, j \in \{\overline{1,r}\}\} \cup \{(B_j, C_k) : j \in \{\overline{1,r}\}, k \in \{\overline{1,n}\}\}. \end{aligned}$$

Define weights of edges with

$$\begin{aligned} \omega(A_i, B_j) &= m_1[i,j], \\ \omega(B_i, C_j) &= m_2[j,k]. \end{aligned}$$

If $M = M_1 M_2$, then

$$m[i,j] = \sum_{k=1}^{r} m_1[i,k] m_2[j,k].$$

Fix some arbitrary $X \subset \{1,2,\ldots,r\}$. Let $\mathcal{P}_{\mathcal{AX}}$ be set of all disjoint paths from $\mathcal{A}$ to $B[\mathcal{X}]$, where is $B[\mathcal{X}]$ subset of B limited with indexes of X, and $\mathcal{P}_{\mathcal{XB}}$ the set of all disjoint path systems from $B[\mathcal{X}]$ to $\mathcal{C}$. Consider now

$$\begin{aligned} \det(M_1)\cdot\det(M_2) &= \sum_{Q\in\mathcal{P}_{AX}} \operatorname{sign}(\mathcal{Q})\omega(\mathcal{Q}) \sum_{R\in\mathcal{P}_{XB}} \operatorname{sign}(\mathcal{R})\omega(\mathcal{R}) \\ &= \sum_{\mathcal{P}\in\mathcal{P}_{AX}\times\mathcal{P}_{XB}} \operatorname{sign}(\mathcal{P})\omega(\mathcal{P}), \end{aligned}$$

where $\mathcal{P}_{AX} \times \mathcal{P}_{XB}$ contains set of all disjoint path systems from $\mathcal{A}$ to $\mathcal{C}$ which go through all vertices $\mathcal{B}[X]$. Now, in principle, we get

$$\sum_{X\subset\{1,2,\ldots,r\},|X|=n} \det(M_1[X])\det(M_2[X]).$$

This sum also gives us a sum over all system paths from $\mathcal{A}$ to $\mathcal{C}$. Furthermore, we have

$$\begin{aligned} &\sum_{X\subset\{1,\ldots,r\},|X|=n} \det(M_1[X])\det(M_2[X]) \\ = &\sum_{X\subset\{1,\ldots,r\},|X|=n} \left(\sum_{\mathcal{P}\in\mathcal{P}_{AX}\times\mathcal{P}_{XB}} \operatorname{sign}(\mathcal{P})\omega(\mathcal{P}) \right) \\ = &\sum_{\mathcal{P}} \operatorname{sign}(\mathcal{P})\omega(\mathcal{P}) = \det(M_1\cdot M_2). \end{aligned}$$

□

4 LGV and Problem 1

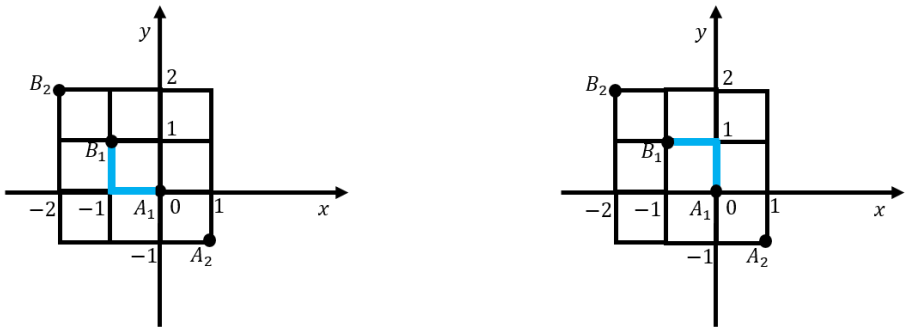
We calculate matrix using LGV theorem in the case $n = 2$.

$$\begin{bmatrix} \binom{2}{1} & \binom{4}{2} \\ \binom{4}{2} & \binom{6}{3} \end{bmatrix} = \begin{bmatrix} 2 & 6 \\ 6 & 20 \end{bmatrix}$$

The Idea is same as like in the Proposition 3.1. We want to create the directed weighted acyclic graph in which weights of paths are equal to the value in our matrix, i.e.

$$\begin{bmatrix} \omega(A_1, B_1) & \omega(A_1, B_2) \\ \omega(A_2, B_1) & \omega(A_2, B_2) \end{bmatrix} = \begin{bmatrix} 2 & 6 \\ 6 & 20 \end{bmatrix}$$

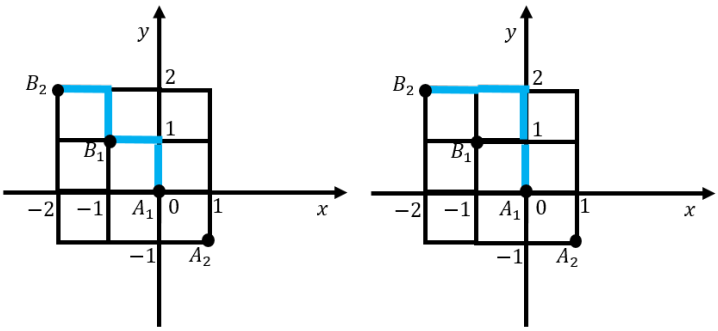
So consider our case of matrix S_2 and create a directed weighted graph of matrix S_2. First, we will consider the possible paths from A_1 to B_2 (Figure 4).



(a) Path P_1 (b) Path P_2

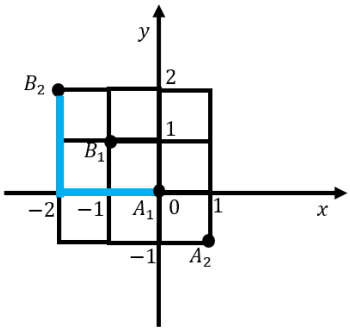
Figure 4: Paths from A_1 to B_1

Check now the weight of path from A_1 to B_1. Path P_1 and P_2 in our graph have a weight one, using (2) we get that the weight of path form A_1 to B_1 is 2. Now, consider all possible paths from A_1 to B_2 (Figure 5).

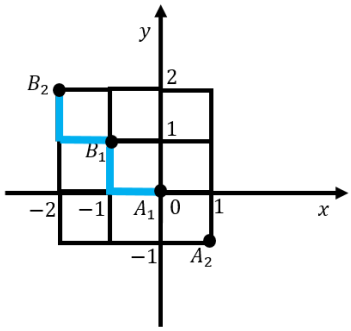


(a) Path P_3

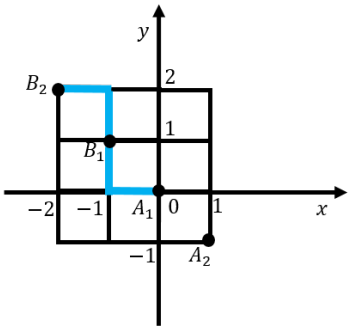
(b) Path P_4



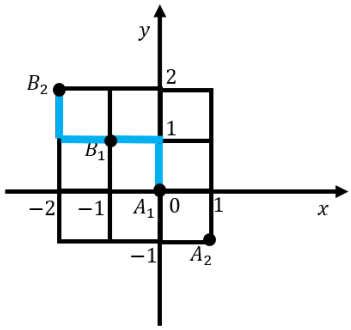
(c) Path P_5



(d) Path P_6



(e) Path P_4

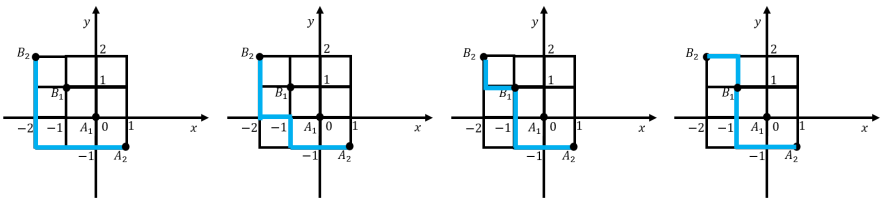


(f) Path P_8

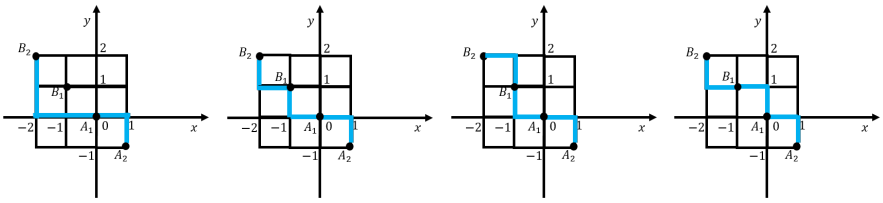
Figure 5: Paths from A_1 to B_2

Every paths P_i, $i = 1, \ldots, 6$ have a weight one. Using the (2) we obtain that the

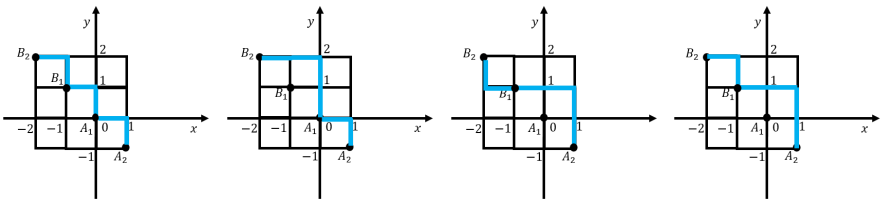
weight of path from A_1 to B_2 is equal to 6. Analogously we will obtained the weight of paths from A_2 to B_1. Now, we will consider the possible paths from A_2 to B_2 (Figure 6).



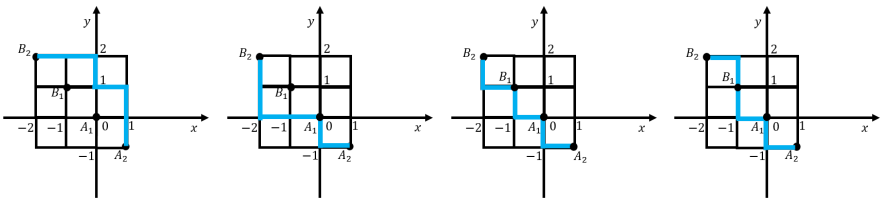
(a) Path P_{15} (b) Path P_{16} (c) Path P_{17} (d) Path P_{18}



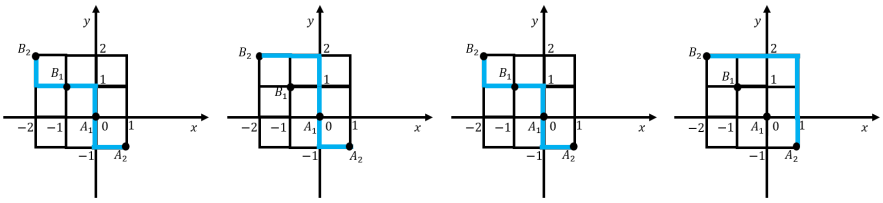
(e) Path P_{19} (f) Path P_{20} (g) Path P_{21} (h) Path P_{22}



(i) Path P_{23} (j) Path P_{24} (k) Path P_{25} (l) Path P_{26}



(m) Path P_{27} (n) Path P_{28} (o) Path P_{29} (p) Path P_{30}



(q) Path P_{31} (r) Path P_{32} (s) Path P_{33} (t) Path P_{34}

Figure 6: Paths from A_2 to B_2

Using (2) we obtained $\omega(A_2, B_2) = 20$. If we now consider disjoint path systems in the graph and their signs we conclude that $\det S_2 = 4$ (Figure 7).

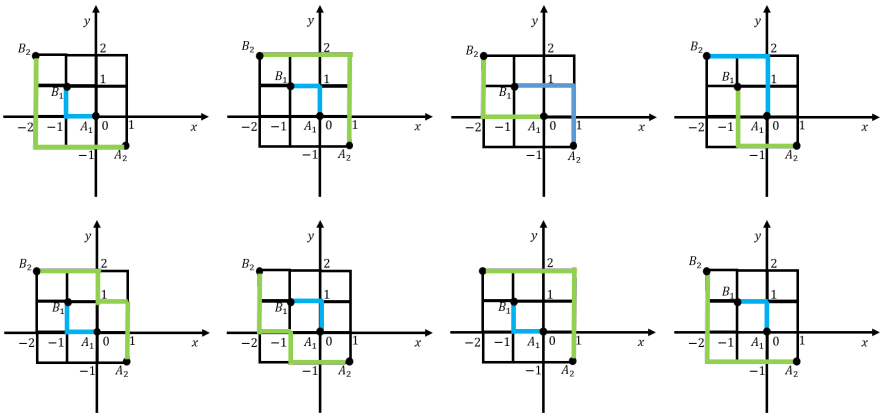


Figure 7: Disjoint system of paths

In the case of S_n, we use the same idea and create the graph of matrix S_n in the following way (Figure 8). It's known that in rectangular dimensions of $m \times n$ we have a $\frac{(m+n)!}{m!n!}$ different nonintersecting paths.

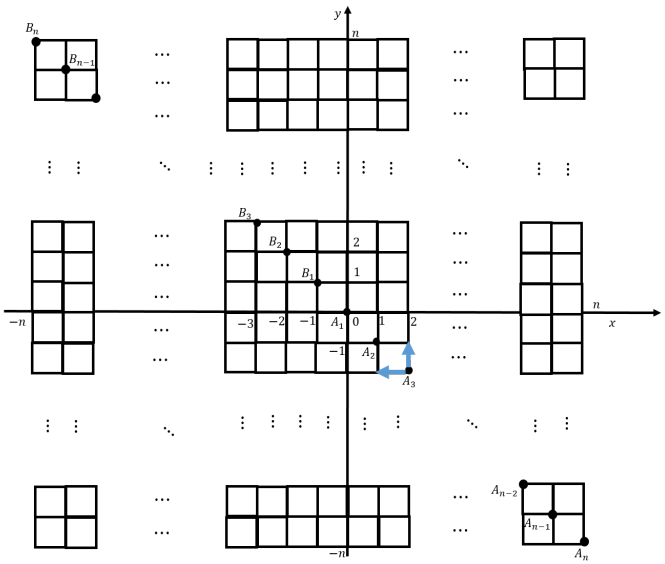


Figure 8: Directed weighted acyclic graph of matrix s_n

From Koon and Yun Bum's solution we know that the difference between the number of non-intersecting paths with positive sign and the number of non-intersecting paths with negative sign is 2^n. However, in general the number of all non-intersecting path systems is large, even for $n = 3$ so deducing the result directly from the LGV theorem requires this result, which seems non-trivial.

Problem 2. *Find a combinatorial argument that the difference between the number of non-intersecting paths with positive sign and the number of non-intersecting paths with negative sign from* $\{A_1, \ldots, A_n\}$ *and* $\{B_1, \ldots, B_n\}$ *in the graph on the Figure 8 is* 2^n.

5 Conclusion

We illustrated some possibilities of applying LGV theorem. This theorem can be implemented many problems in combinatorics and some other mathematicians area, for example we can LGV theorem apply on: Dyck paths, Motzkin numbers, Hankel determinants, Catalan numbers, rhombus tilings and many others problems.

Acknowledgment

The author is grateful to Đorđe Baralić for suggesting this topic and useful comments and discussions.

References

[1] M. Aigner *A Course in Enumeration*, Springer – Verlag, Berlin, 2007.

[2] I. Gessel and G. Viennot, *Binomial Determinants, Paths, and Hook Length Formulae.* Advances in Mathematics 58 (1985), no. 3, 300–321.

[3] S. Karlin and J. McGregor, *Coincidence probabilities.* Pacific J. Math. 9 (1959), no. 4, 1141–1164.

[4] B. Lindström, *On the vector representations of induced matroids*, Bull. London Math. Soc. 5 (1973) 85–90.

[5] https://mathsci.kaist.ac.kr/pow/2016/12/

Periodicity of the Greatest Common Divisors within Generalized Divisibility Sequences

Ivica Martinjak
Faculty of Science, University of Zagreb
Bijenička cesta 32, HR-10000 Zagreb, Croatia
and
Sanja Tipurić - Spužević
University of Mostar, Faculty of Science and Education,
Matice Hrvatske b.b., 88000 Mostar, Bosnia and Herzegovina

Abstract

This paper addresses divisibility properties of some families of sequences arising from partial sums of a strong divisibility sequence. In particular, we demonstrate periodicity of greatest common divisor within 1-fibonacci numbers. We also present congruences within this sequence modulo a prime number p where $p \equiv \pm 1 \pmod 5$ and $p \equiv \pm 2 \pmod 5$.

Keywords: divisibility sequence, elliptic divisibility sequences, hypersequence, Fibonacci numbers, Mersenne numbers, Somos sequence, elliptic curves

MSC: 11A05, 11B39

1 Introduction

A *divisibility sequence* is an integer sequence $(d_n)_{n\geq 0}$ with the property that an index n being a multiple of index m imply the term d_n is a multiple of d_m,

$$m \mid n \implies d_m \mid d_n \tag{1}$$

for all natural numbers m, n. If for the sequence $(d_n)_{n\geq 0}$ we have

$$\gcd(d_m, d_n) = d_{\gcd(m,n)}, \tag{2}$$

DOI: https://doi.org/10.5592/CO/CCD.2018.07

then such sequence is called *strong divisibility sequence*. It is immediately seen that a strong divisibility sequence is also divisibility sequence.

An important class of these sequences are *elliptic divisibility sequences (EDS)*. An elliptic divisibility sequence is a sequence of integers $(Wn)_{n\geq 0}$ satisfying the recursive relation

$$W_{n+m}W_{n-m}W_1^2 = W_{n+1}W_{n-1}W_m^2 - W_{m+1}W_{m-1}W_n^2 \tag{3}$$

and such that W_n divides W_m whenever n divides m. It is known that if the initial conditions for (3) satisfies

i) $W_1 = 1$,

ii) $W_2, W_3, W_4/W_2 \in \mathbb{Z} \setminus 0$

then W_n is an integer for every n. As a further basic property of EDS we have that if the sequence $(W_n)_{n\geq 0}$ is a solution of (3) then we have

$$\begin{aligned} W_{2n+1} &= W_{n+2}W_n^3 - W_{n-1}W_{n+1}^3, \ \ n \geq 1 \\ W_{2n}W_2 &= W_n(W_{n+2}W_{n-1}^2 - W_{n-2}W_{n+1}^2, \ \ n \geq 2. \end{aligned}$$

An example of such sequences of numbers is the sequence

$$1, 1, 1, -1, -2, -3, -1, 7, 11, 20, -19, -87, -191, -197, 1018, \ldots$$

(the sequence A050512 in the OEIS). As another example let as mention the sequence $(G_n)_{n\geq 0}$ consisting of every second Fibonacci numbers is a EDS, $G_n = F_{2m}$. We let $(h_n)_{h\geq 0}$ denote the sequence defined by $h_n = (n/3)$ where $n \in \mathbb{N}$ and (a/p) denote the *Legendre symbol*, for the prime number p.

Among notable representatives of divisibility sequences we have *Mersenne numbers* defined by the explicit formula

$$M_n = 2^n - 1, \ \ n \geq 0,$$

as well as the Fibonacci numbers,

$$F_n = \frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^n - \left(\frac{1-\sqrt{5}}{2}\right)^n\right].$$

Both of these sequences appears in various number theoretical and combinatorial context. In addition to certain families of *Dyck paths*, the n-th Mersenne number

appears as the number of nonempty subsets of a set with n elements, as a q-binomial coefficient, a rank of matroids, etc. (the sequence A000225 in the OEIS). Recall that Fibonacci numbers appears as the solutions of the Diophantine equation

$$x^2 - 5y^2 = 4(-1)^n, \tag{4}$$

i.e. we have the Fibonacci sequence $(F_n)_{n\geq 0}$ and the Lucas sequence $(L_n)_{n\geq 0}$ as the solutions $(x, y) = (L_n, F_n)$ of (4), where the Lucas numbers are defined by the same recurrence relation as the Fibonacci numbers but with the initial conditions $L_0 = 2, L_1 = 1$. One can also use the Diophantine equation (4) as a definition of these two sequences of numbers.

This work aim at finding divisibility properties of some families of generalized divisibility sequences. We were curious to establish how the properties (1) and (2) are inherited within such sequences.

2 Previous results and motivation

A complete characterization of divisibility sequences arising from linear recurrences is done by Bézivin, Pethő, and van der Poorten [2]. Recent development is done by Ingram [11], Silverman [18] and Gezer and Bizim [6]. There are further generalizations and extensions of this notion. A natural generalization of divisibility sequences is through divisibility of ideals in a ring. One can find more on this in a work of Silverman [17]. Results on *matrix divisibility sequences* (a sequence of matrices with properties analogue to (1)) are found by Cornelissen and Reynolds [4] as well as Górnisiewicz [10]. Among other classes of divisibility sequences, let mention a class of sequences defined as

$$d_n(\alpha) = \max\{d \in \mathbb{Z} : \alpha^n \equiv 1 \pmod{d}\},$$

where α is an *algebraic integer*. Moreover, such sequences also satisfy property (2) i.e. they are strong divisibility sequences, which is proved by Silverman [16].

Among many remarkable properties of the Fibonacci sequence $(F_n)_{n\geq 0}$, $F_{n+2} = F_{n+1} + F_n$, $F_0 = 0$, $F_1 = 1$ we have that when m divide n, then F_m divide F_n,

$$m \mid n \implies F_m \mid F_n. \tag{5}$$

There is also identity

$$F_{m+n} = F_{m+1}F_n + F_mF_{n-1}$$

and as a consequence of these two facts one can derive that the greatest common divisor of the Fibonacci numbers F_m and F_n is again Fibonacci number, that one whose index is $\gcd(m,n)$,

$$\gcd(F_m, F_n) = F_{\gcd(m,n)}. \tag{6}$$

Details on proof of this one can find in [1]. Among further divisibility properties of Fibonacci numbers is a well known fact that

$$F_p \equiv \left(\frac{p}{5}\right) \pmod{p} \tag{7}$$

$$F_{p\pm 1} \equiv \frac{1 \pm \left(\frac{p}{5}\right)}{2} \pmod{p} \tag{8}$$

where p is an odd prime.

In what follows we present divisibility properties of the sequences arising from partial sums of a family of strong divisibility sequence. In particular, the hyperfibonacci sequence of the rth generation $(F_n^{(r)})_{n\geq 0}$, is defined by the recurrence relation

$$F_n^{(r)} = \sum_{k=0}^{n} F_k^{(r-1)}, \;\; F_n^{(0)} = F_n, \;\; F_0^{(r)} = 0, \;\; F_1^{(r)} = 1, \tag{9}$$

where $r \in \mathbb{N}$ and F_n is the nth term of the Fibonacci sequence $(F_n)_{n\geq 0}$. These sequences are introduced by Dil and Mező, in a study of a symmetric algorithm for hyperharmonic, Fibonacci and some other integer sequences [8]. Several number theoretical, combinatorial and algebraical properties of hyperfibonacci sequences is already known [3, 13, 14, 20]. An alternative definition of $(F_n^{(r)})_{n\geq 0}$ is by the recurrence relation

$$F_{n+2}^{(r)} = F_{n+1}^{(r)} + F_n^{(r)} + \binom{n+r}{r-1}, \;\; n \geq 0 \tag{10}$$

where initial values are $F_0^{(r)} = 0$, $F_1^{(r)} = 1$. The proof of this one can find in [5]. For $r = 1$ this relation gives the sequence of numbers

$$0, 1, 2, 4, 7, 12, 20, 33, 54, 88, 143, 232, 376, \ldots,$$

for $r = 2$ we have the sequence

$$0, 1, 3, 7, 14, 26, 46, 79, 133, 221, 364, 596, 972, \ldots$$

etc. When $r = 1$, relation (10) reduces to

$$F_{n+2}^{(1)} = F_{n+1}^{(1)} + F_n^{(1)} + 1. \tag{11}$$

We shall present divisibility properties of hypefibonacci numbers of the first generation $(F_n^{(1)})_{n\geq 0}$. Throughout the paper, the hyperfibonacci sequence of the 1st generation we shall also call hyperfibonacci sequence, in short.

3 The main result

Every two consecutive Fibonacci numbers are relatively prime. In Lemma 1 we generalize this property on the case of hyperfibonacci numbers.

Lemma 1. *Every three consecutive hyperfibonacci numbers $F_n^{(1)}$, $F_{n+1}^{(1)}$, $F_{n+2}^{(1)}$, $n \geq 0$ are relatively prime,*

$$\gcd\left(F_n^{(1)}, F_{n+1}^{(1)}, F_{n+2}^{(1)}\right) = 1. \tag{12}$$

Proof. Using basic properties of the gcd function and the recurrence relation (11) we obtain

$$\begin{aligned}
\gcd\left(F_n^{(1)}, F_{n+1}^{(1)}, F_{n+2}^{(1)}\right) &= \gcd\left(F_n^{(1)}, \gcd\left(F_{n+1}^{(1)}, F_{n+1}^{(1)} + F_n^{(1)} + 1\right)\right) \\
&= \gcd\left(\gcd\left(F_n^{(1)}, F_n^{(1)} + 1\right), F_{n+1}^{(1)}\right) \\
&= \gcd\left(1, F_n^{(1)}, F_{n+1}^{(1)}\right) = 1.
\end{aligned}$$

□

In a similar fashion one can prove that every four consecutive hyperfibonacci numbers of the 2-nd generation are relatively prime. Furthermore, for $r \geq 3$ one can use the obvious equality for binomial coefficients

$$\binom{n+q}{p} - \binom{n+q-1}{p} = \binom{n+q-1}{p-1} \tag{13}$$

when applying Euclid algorithm. In particular, for $r = 3$ we have

$$\begin{aligned}
& \gcd\left(F_n^{(3)}, F_{n+1}^{(3)}, \ldots, F_{n+4}^{(3)}\right) \\
= \; & \gcd\left(F_n^{(3)}, \ldots, F_{n+3}^{(3)}, F_{n+3}^{(3)} + F_{n+2}^{(3)} + \binom{n+5}{2}\right) \\
= \; & \gcd\left(F_n^{(3)}, \ldots, F_{n+3}^{(3)}, \binom{n+5}{2}\right) \\
= \; & \gcd\left(F_n^{(3)}, \ldots, F_{n+2}^{(3)} + F_{n+1}^{(3)} + \binom{n+4}{2}, \binom{n+5}{2}\right) \\
= \; & \gcd\left(F_n^{(3)}, F_{n+1}^{(3)}, F_{n+2}^{(3)}, \binom{n+4}{2}, \binom{n+5}{2}\right) \\
= \; & \gcd\left(F_n^{(3)}, F_{n+1}^{(3)}, \binom{n+3}{2}, \binom{n+4}{2}, \binom{n+5}{2}\right) \\
= \; & \gcd\left(F_n^{(3)}, F_{n+1}^{(3)}, \binom{n+3}{2}, n+4, 1\right) \\
= \; & 1.
\end{aligned}$$

We formalize these arguments in the proof of the following Theorem 1.

Theorem 1. *For $n \geq 0$, every $(r+2)$-tuple of consecutive hyperfibonacci numbers of r-th generation are relatively prime,*

$$\gcd\left(F_n^{(r)}, F_{n+1}^{(r)}, \ldots, F_{n+r+1}^{(r)}\right) = 1. \tag{14}$$

Proof. When applying basic properties of the gcd function we use relations (11) and (13) to get

$$\begin{aligned}
& \gcd(F_n^{(r)}, F_{n+1}^{(r)}, \ldots, F_{n+r+1}^{(r)}) \\
= \; & \gcd\left(F_n^{(r)}, F_{n+1}^{(r)}, \binom{n+r}{r-1}, \binom{n+r+1}{r-1}, \ldots, \binom{n+2r-1}{r-1}\right) \\
= \; & \gcd\left(F_n^{(r)}, F_{n+1}^{(r)}, \binom{n+r}{r-1}, \binom{n+r}{r-2}, \ldots, \binom{n+2r-2}{r-2}\right) \\
= \; & \gcd\left(F_n^{(r)}, F_{n+1}^{(r)}, \binom{n+r}{r-1}, \binom{n+r}{r-2}, \ldots, n+r, n+r+1\right) \\
= \; & \gcd\left(F_n^{(r)}, F_{n+1}^{(r)}, \binom{n+r}{r-1}, \binom{n+r}{r-2}, \ldots, n+r, 1\right) = 1.
\end{aligned}$$

□

In addition, we have that the greatest common divisor of some pairs of hyperfibonacci numbers is a Fibonacci number, as stated in Theorem 2.

Theorem 2. *For $m, n \in \mathbb{N}$ the greatest common divisor of the $(4m-3)$-th and $(4m-1)$-st hyperfibonacci numbers is equal to F_{2m},*

$$\gcd\left(F^{(1)}_{4m-3}, F^{(1)}_{4m-1}\right) = F_{2m}. \tag{15}$$

Proof. Using the fact that the gcd of two numbers does not change if the largest number is replaced by its difference with the smaller one and applying the recurrence relation (11) we obtain

$$\gcd\left(F^{(1)}_{4m-3}, F^{(1)}_{4m-1}\right) = \gcd\left(F^{(1)}_{4m-3}, F^{(1)}_{4m-2}+1\right) = \gcd\left(F^{(1)}_{4m-5}-1, F^{(1)}_{4m-4}+2\right).$$

When we continue to diminish the larger number this way, resulting number is always represented as a sum of a hyperfibonacci number and an integer, $F_{4m-q}+a_q$ and $F_{m-q+1}+a_{q-1}$. According to the initial terms $a_3 = 0$ and $a_2 = 1$, the absolute value of the n-th number in sequence of these integers differentiate from the n-th Fibonacci number for 1. More precisely, we have

$$\begin{aligned} & \gcd\left(F^{(1)}_{4m-3}, F^{(1)}_{4m-1}\right) \\ = \; & \gcd\left(F^{(1)}_{4m-q} + (-1)^q(F_{q-2}+1), F^{(1)}_{4m-q+1} + (-1)^{q+1}(F_{q-3}+1)\right) \end{aligned}$$

where $3 \leq q \leq 4m$. Now, according to this fact we obtain

$$\begin{aligned} & \gcd\left(F^{(1)}_{4m-3}, F^{(1)}_{4m-1}\right) = \gcd\left(F^{(1)}_{4m-3} - (F_1-1), F^{(1)}_{4m-2} + (F_0+1)\right) \\ = \; & \gcd\left(F^{(1)}_{4m-4} + (F_2+1), F^{(1)}_{4m-3} - (F_1-1)\right) \\ = \; & \gcd\left(F^{(1)}_{2m-2} + (F_{2m}+1), F^{(1)}_{2m-1} - (F_{2m-1}-1)\right) \\ = \; & \gcd\left(2F_{2m}, F_{2m+1} - F_{2m-1}\right) = \gcd\left(2F_{2m}, F_{2m}\right) = F_{2m}, \end{aligned}$$

which completes the proof. □

For an alternative proof of Theorem 2 we have the following. By the product expansion formula we have

$$F_{m+n} = F_m L_n + (-1)^{n+1} F_{m-n}$$

and from it we get

$$F_{4m} = F_{2m}L_{2m} \text{ and } F_{4m-1} = F_{2m}L_{2m-1} + (-1)^{2m} F_1 = F_{2m}L_{2m-1} + 1$$

We now have:

$$\begin{aligned} \gcd\left(F^{(1)}_{4m-3}, F^{(1)}_{4m-1}\right) &= \gcd\left(F_{4m-1} - 1, F_{4m+1} - 1\right) = \gcd\left(F_{4m-1} - 1, F_{4m}\right) = \\ &= \gcd\left(F_{2m}L_{2m-1}, F_{2m}L_{2m}\right) = F_{2m}\gcd\left(L_{2m-1}, L_{2m}\right) = F_{2m} \end{aligned}$$

As an example, let consider the case when $m = 3$. According to Theorem 2 the greatest common divisor of numbers $F^{(1)}_9 (= 88)$ and $F^{(1)}_{11} (= 232)$ is equal to

$$\begin{aligned} & \gcd\left(F^{(1)}_9 - (F_1 - 1), F^{(1)}_{10} + (F_0 + 1)\right) \\ = \ & \gcd\left(F^{(1)}_8 + (F_2 + 1), F^{(1)}_9 - (F_1 - 1)\right) \\ = \ & \gcd\left(F^{(1)}_7 - (F_3 - 1), F^{(1)}_8 + (F_2 + 1)\right) \\ = \ & \gcd\left(F^{(1)}_4 + (F_6 + 1), F^{(1)}_5 - (F_5 - 1)\right) \\ = \ & \gcd\left(2F_6, F_6\right) = F_6. \end{aligned}$$

Indeed, $\gcd(88, 232) = 8$ which is the 6-th number in the Fibonacci sequence.

When applying (6) we have an obvious consequence of Theorem 2, stated in Corollary 1.

Corollary 1. *The greatest common divisor of the 4-tuple of hyperfibonacci numbers* $F^{(1)}_{4m-1}$, $F^{(1)}_{4m-3}$, $F^{(1)}_{4n-1}$, $F^{(1)}_{4n-3}$, $m, n \in \mathbb{N}$ *is equal to the* $\gcd(2m, 2n)$*-th Fibonacci number,*

$$\gcd\left(F^{(1)}_{4m-1}, F^{(1)}_{4m-3}, F^{(1)}_{4n-1}, F^{(1)}_{4n-3}\right) = F_{\gcd(2m,2n)}.$$

In Corollary 2 we list further periodicity in relatively prime pairs and the greatest common divisor, for hyperfibonacci numbers.

Corollary 2. *For the hyperfibonacci sequence* $\left(F^{(1)}_n\right)_{n\geq 0}$ *we have*

i) $\gcd\left(F^{(1)}_{6n+4}, F^{(1)}_{6n+5}\right) = 1,$

ii) $\gcd\left(F^{(1)}_{6n+2}, F^{(1)}_{6n+3}\right) = 2$,

iii) $\gcd\left(F^{(1)}_{6n+6}, F^{(1)}_{6n+7}\right) = 1$.

Proof. i) We employ recurrence relation (11) to get

$$\begin{aligned}
& \gcd\left(F^{(1)}_{6n+4}, F^{(1)}_{6n+5}\right) = \gcd\left(F^{(1)}_{6n+3} + 1, F^{(1)}_{6n+4}\right) \\
= \; & \gcd\left(F^{(1)}_{6n+2}, F^{(1)}_{6n+3} + 1\right) = \gcd\left(F^{(1)}_{6n+1} + 2, F^{(1)}_{6n+2}\right) \\
= \; & \gcd\left(F^{(1)}_{6n} - 1, F^{(1)}_{6n+1} + 2\right) \\
= \; & \gcd\left(F^{(1)}_{6n-1} + (F_4 + 1), F^{(1)}_{6n} - (F_3 - 1)\right) \\
= \; & \gcd\left(F^{(1)}_{6n-2} - (F_5 - 1), F^{(1)}_{6n+5} + (F_4 + 1)\right)
\end{aligned}$$

When iteratively applying relation (11) and the basic properties of the gcd function we obtain

$$\gcd\left(F^{(1)}_{6n+4}, F^{(1)}_{6n+5}\right) = \gcd\left(F^{(1)}_{3n+1} - (F_{3n+2} - 1), F^{(1)}_{3n+2} + (F_{3n+1} + 1)\right). \tag{16}$$

From the fact that the sum of the first n numbers in Fibonacci sequence $(F_n)_{n\geq 0}$ is equal to $F_{n+2} - 1$, we immediately have

$$F^{(1)}_n = F_{n+2} - 1. \tag{17}$$

We substitute (17) into r.h.s. of relation (16) to get

$$\gcd\left(F^{(1)}_{6n+4}, F^{(1)}_{6n+5}\right) = \gcd\left(F_{3n+1}, F_{3n+4}\right). \tag{18}$$

Having in mind that

$$\begin{aligned}
\gcd(3n+1, 3n+4) &= \gcd(3n+1, 3) \\
&= 1
\end{aligned}$$

we finally have

$$\begin{aligned}
\gcd\left(F^{(1)}_{6n+4}, F^{(1)}_{6n+5}\right) &= \gcd\left(F_{3n+1}, F_{3n+4}\right) \\
&= F_{\gcd(3n+1,3n+4)} = F_1 \\
&= 1.
\end{aligned}$$

□

Proof. ii) Once having equality

$$\gcd\left(F^{(1)}_{6n+2}, F^{(1)}_{6n+3}\right) = \gcd\left(F_{3n+3}, F_{3n}\right)$$

we get

$$\begin{aligned}\gcd(3n+3, 3n) &= (3n, 3)\\ &= 3\end{aligned}$$

and finally

$$\begin{aligned}\gcd\left(F^{(1)}_{6n+2}, F^{(1)}_{6n+3}\right) &= \gcd\left(F_{3n+3}, F_{3n}\right)\\ &= F_3 = 2.\end{aligned}$$

iii) Having in mind that

$$\begin{aligned}\gcd(3n+2, 3n+5) &= \gcd(3n+2, 3)\\ &= \gcd(2, 3) = 1\end{aligned}$$

we obtain

$$\begin{aligned}\gcd\left(F^{(1)}_{6n+6}, F^{(1)}_{6n+7}\right) &= \gcd\left(F_{3n+5}, F_{3n+2}\right)\\ &= F_{\gcd(3n+5,3n+2)}\\ &= F_1 = 1.\end{aligned}$$

which completes the proof. □

We consider the alternative way of calculating $\gcd(F^{(1)}_n, F^{(1)}_{n+1})$. The $\gcd(F^{(1)}_n, F^{(1)}_{n+1})$ can be written as

$$\begin{aligned}\gcd(F^{(1)}_n, F^{(1)}_{n+1}) &=\\ \gcd(F_{n+2}-1, F_{n+3}-1) &= \gcd(F_{n+2}-F_{-1}, F_{n+3}+F_{-2}) =\\ \gcd(F_{n+2}-F_{-1}, F_{n+1}+F_0) &= \gcd(F_n-F_1, F_{n+1}+F_0) =\\ \gcd(F_n-F_1, F_{n-1}+F_2) &= \gcd(F_{n-2}-F_3, F_{n-1}+F_2) =\\ \gcd(F_{n-2}-F_3, F_{n-3}+F_4) &= \gcd(F_{n-4}-F_5, F_{n-3}+F_4) = \ldots =\\ &\quad \gcd\left(F_{n-2k}-F_{2k+1}, F_{n-2k+1}+F_{2k}\right)\end{aligned}$$

Now it follows:

(*a*) $n = 4m,\ k = m,\ \gcd(F^{(1)}_{4m}, F^{(1)}_{4m+1}) = \gcd\left(F_{2m} - F_{2m+1}, F_{2m+1} + F_{2m}\right) = \gcd\left(-F_{2m-1}, F_{2m+2}\right) = \gcd\left(F_{2m-1}, F_{2m+2}\right) = F_{\gcd(2m-1,2m+2)} = F_{\gcd(m+1,3)} = F_{\gcd(n+1,3)}$

(*b*) $n = 4m+1,\ k = m,\ \gcd(F^{(1)}_{4m+1}, F^{(1)}_{4m+2}) = \gcd\left(F_{2m+1} - F_{2m+1}, F_{2m+2} + F_{2m}\right) = \gcd\left(0, L_{2m+1}\right) = L_{\frac{n+1}{2}}$

(*c*) $n = 4m+2,\ k = m,\ \gcd(F^{(1)}_{4m+2}, F^{(1)}_{4m+3}) = \gcd\left(F_{2m+2} - F_{2m+1}, F_{2m+3} + F_{2m}\right) = \gcd\left(F_{2m}, F_{2m+3} + F_{2m}\right) = \gcd\left(F_{2m}, F_{2m+3}\right) = F_{\gcd(2m,2m+3)} = F_{\gcd(m,3)} = F_{\gcd(n,3)}$

(*d*) $n = 4m+3,\ k = m+1,\ \gcd(F^{(1)}_{4m+3}, F^{(1)}_{4m+4}) = \gcd\left(F_{2m+1} - F_{2m+3}, F_{2m+2} + F_{2m+2}\right) = \gcd\left(-F_{2m+2}, 2F_{2m+2}\right) = F_{2m+2} = F_{\frac{n+1}{2}}$

4 Congruences for $F^{(1)}_{p-1}$, $F^{(1)}_{p-2}$ and $F^{(1)}_{p-3}$

Once having relation (17), we immediately obtain congruences for $F^{(1)}_{p-2}$, by substitution into (7). We present these congruences in the following Theorem 3, where we also give more detailed proof.

Theorem 3. *Let p be an odd prime. Then for the hyperfibonacci sequence we have*

$$F^{(1)}_{p-2} \equiv \left(\frac{p}{5}\right) - 1 \pmod{p}. \tag{19}$$

Proof. When applying the binomial theorem to the Binet's formula we get

$$\begin{aligned} F^{(1)}_{p-2} &= \frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^p - \left(\frac{1-\sqrt{5}}{2}\right)^p\right] - 1 \\ &= \frac{1}{2^p\sqrt{5}}\sum_{k=0}^{p}\binom{p}{k}\left((\sqrt{5})^k - (-\sqrt{5})^k\right) - 1 \\ &= \frac{1}{2^{p-1}}\sum_{k=0,\ 2\nmid k}^{p}\binom{p}{k}5^{\frac{k-1}{2}} - 1 \end{aligned}$$

Having in mind an obvious fact that $p \mid \binom{p}{k},\ \ k = 1, 2, \ldots, p-1$ we obtain

$$1 + 2^{p-1}F^{(1)}_{p-2} \equiv 5^{\frac{p-1}{2}} \pmod{p}$$

and furthermore from the Euler's criterion

$$1 + 2^{p-1}F^{(1)}_{p-2} \equiv \left(\frac{5}{p}\right) \pmod{p}.$$

□

In the similar fashion one can prove Theorem 4. Both congruences in Theorem 4 also follows by substitution of (17) into (8).

Theorem 4. *Let p be an odd prime. Then for the hyperfibonacci sequence we have*

$$F_{p-3}^{(1)} \equiv \frac{-1-\left(\frac{p}{5}\right)}{2} \pmod{p} \tag{20}$$

and

$$F_{p-1}^{(1)} \equiv \frac{-1+\left(\frac{p}{5}\right)}{2} \pmod{p}. \tag{21}$$

According to the quadratic reciprocity theorem we get equality

$$\left(\frac{5}{p}\right) = \left(\frac{p}{5}\right)$$

which gives

$$\left(\frac{p}{5}\right) = \begin{cases} 1 & \text{if } p \equiv \pm 1 \pmod 5 \\ -1 & \text{if } p \equiv \pm 2 \pmod 5. \end{cases}$$

when we employ basic properties of the Legendre symbol. Now we have immediate consequences of Theorems 3 and 4. Corollary 3 follows from the congruence (20) while Corollary 4 follows from congruences (19) and (21).

Corollary 3. *Let p be a prime such that $p \equiv \pm 2 \pmod 5$. Then $p \mid F_{p-3}^{(1)}$.*

Corollary 4. *Let p be a prime such that $p \equiv \pm 1 \pmod 5$. Then $p \mid F_{p-2}^{(1)}$ and $p \mid F_{p-1}^{(1)}$.*

5 Concluding remarks and open questions

We believe that results obtained in this paper can be extended to other families of strong divisibility sequences. It would be of interest to find periodicity of gcd for other generation of hyperfibonacci numbers and possibly to give statements in full generality. There are also a few other generalization of recursive sequences of numbers ([9, 12, 15]) and it is of interest to see an extension of property (2) within these sequences. Some further generalizations and extensions of these sequences would be of interest as well.

Acknowledgment

The authors thank the referee for careful reading, valuable help and valuable suggestions that improved the final version of the paper.

References

[1] A.T. Benjamin, J.J. Quinn, Proofs that Really Count, The Mathematical Association of America, 2003.

[2] J.P. Bézivin, A. Pethő, A.J. van der Poorten, A full characterization of divisibility sequences, Am. J. Math. 112 (6) (1990), 985-1001.

[3] N. N. Cao and F. Z. Zhao, Some properties of Hyperfibonacci and Hyperlucas Numbers, *J. Integer Seq.* 13 (2010), Article 10.8.8.

[4] G. Cornelissen, J. Reynolds, Matrix divisibility sequences, Acta Arith., 156 (2012), 177-188.

[5] L. Cristea, I. Martinjak, I. Urbiha, Hyperfibonacci Sequences and Polytopic Numbers, J. Integer Seq., 19/7 (2016), Article 16.7.6, 13pp.

[6] B. Gezer, O. Bizim, Squares in elliptic divisibility sequences, Acta Arith., 144.2 (2010), 125-134.

[7] M. Glunčić, I. Martinjak, A Class of S-Restricted Compositions, Int. Journal of Number Theory, to appear, https://doi.org/10.1142/S1793042119500180

[8] A. Dil and I. Mező, A symmetric algorithm for hyperharmonic and Fibonacci numbers, Appl. Math. Comput., 206 (2008), 942–951.

[9] T. Došlić, I. Martinjak, R. Škrekovski, Total Positivity of Toeplitz Matrices of Recursive Hypersequences, preprint

[10] K. Górnisiewicz, New Examples of Divisibility Sequences, Integers 16 (2006), #A56, 5pp.

[11] P. Ingram, Elliptic divisibility sequences over certain curves, J. Number Theory 123 (2) (2007), 473-486.

[12] C. Krattenthaler and A.M. Oller-Marcén, *A Determinant of Generalized Fibonacci Numbers*, J. Combin. Number Theory 5(2) (2003), article 2 1–7.

[13] R. Liu and F. Zhao, On the sums of reciprocal Hyperfibonacci numbers and Hyperlucas numbers, J. Integer Seq. 15 (2012), Article 12.4.5.

[14] I. Martinjak, I. Urbiha, A New Generalized Cassini Determinant, Colloq. Math. 145 (2), 2016, 209-218

[15] E. Miles, *Generalized Fibonacci numbers and related matrices*, Amer. Math. Monthly 67 (1960), 745-752.

[16] J. H. Silverman, Divisibility Sequences and Powers of Algebraic Integers, Documenta Math., Extra Volume Coates (2006), 711-727.

[17] J.H. Silverman, Generalized greatest common divisors, divisibility sequences, and Vojta's conjecture for blowups, Monatsh. Math., 145 (2015), 333-350.

[18] J.H. Silverman, Common divisors of elliptic divisibility sequences over function fields, Manuscripta Math., 114 (2004), 432-446.

[19] Z. H. Sun, Z. W. Sun, Fibonacci numbers and Fermat's last theorem, Acta Arith., 60 (4) (1992), 371-388.

[20] L. Zheng, R. Liu, and F. Zhao, On the log-concavity of the hyperfibonacci numbers and the hyperlucas numbers, J. Integer Seq., 17 (2014), Article 14.1.4

On the Bodlaender Sequence

Daniele Parisse
Airbus Defence and Space GmbH
Rechliner Str., 85077 Manching, Germany
daniele.parisse@t-online.de

Abstract

We will present two equivalent solutions of the Bodlaender sequence $g : \mathbb{N} \longrightarrow \mathbb{Z}$ first introduced recursively by him et al. and used by them to determine the edge ranking number of the complete graphs. These solutions rely on the binary representation of $m \in \mathbb{N}$ either in the form $m = 2^{a_0} + \cdots + 2^{a_l}$ with the strictly decreasing integer sequence $a_0 > a_1 > \cdots > a_l \geq 0$, $l \geq 0$, or using the binary expansion of $m = \sum_{k=0}^{n} b_k 2^k$, where $n \in \mathbb{N}_0$ and $b_0, \ldots, b_n \in \{0, 1\}$. In addition, we will determine sharp bounds of the solution and we will give some properties of related sequences such as $a(m) := \frac{1}{3}\big(m^2 + g(m)\big)$, $m \in \mathbb{N}$, $d(m) := g(m + 1) + 1$, $m \in \mathbb{N}_0$, and the sequence of the partial sums of $g(m)$, $m \in \mathbb{N}$.

Keywords: sequences, divide-and-conquer recurrences, edge ranking number of a graph

MSC: Primary 11B37; Secondary 05C15.

1 Introduction

In order to determine the edge ranking number of complete graphs Hans L. Bodlaender et al. introduced in [1, Chapter 7] the following sequence defined recursively for any $m \in \mathbb{N}$ by the rules

$$\left.\begin{aligned} g(1) &= -1 \\ g(2m) &= g(m) \\ g(2m+1) &= g(m+1) + m \end{aligned}\right\} \tag{1.1}$$

DOI: https://doi.org/10.5592/CO/CCD.2018.08

with the first few values

$$(-1,-1,0,-1,2,0,2,-1,6,2,5,0,8,2,6,-1,14,6,11,2,15,5,11,0,20,\ldots).$$

(This sequence is not available in the On-Line Encyclopedia of Integer Sequences (OEIS ®)[8].)

We briefly recall the definition of the edge ranking number of a graph G (for more details we refer to [1, Definition 2] or [7, p.1067].) Let $G=(V,E)$ be a (simple) graph and t a positive integer. An *edge-t-ranking* is an edge coloring $c' : E \longrightarrow \{1,2,\ldots,t\}$ such that for any two edges of the same color, every path between them contains an intermediate edge with a larger color value. The *edge ranking number* denoted by $\chi'_r(G)$ is the smallest value of t such that the graph G has an edge-t-ranking.

Remark 1.1 *In [4, p.1068] and [7, Corollary 4.9] the authors called it Bodlaender function g, but since (1.1) is defined only for all $m \in \mathbb{N}$, it is more appropriate to denote it Bodlaender sequence.*

Note that for $m-1$ instead of m we obtain from (1.1) $g\big(2(m-1)+1\big) = g(2m-1) = g(m)+m-1$ and therefore $g(2m-1)-g(2m) = m-1$ for all $m \in \mathbb{N}$.

The recurrence relation (1.1) is a special case of the general recurrence relation defined for all $m \in \mathbb{N}$

$$\left.\begin{aligned} f(1) &= \zeta \\ f(2m) &= \alpha f(m) + g(m) \\ f(2m+1) &= \gamma f(m) + \delta f(m+1) + h(m) \end{aligned}\right\} \tag{1.2}$$

with the parameters $\alpha,\gamma,\delta,\zeta \in \mathbb{Z}$ and the integer functions $g,h : \mathbb{N} \longrightarrow \mathbb{Z}$. In our case it is $\alpha=1, \gamma=0, \delta=1, \zeta=-1$ and $g(m)=0$, $h(m)=m$ for all $m \in \mathbb{N}$.

Eq.(1.2) can also be written for all $m \in \mathbb{N}_2 := \{2,3,4,\ldots\}$ as follows

$$f(m) = a(m) f\big(\lfloor m/2 \rfloor\big) + b(m) f\big(\lceil m/2 \rceil\big) + c(m),\ f(1)=\zeta \tag{1.3}$$

with

$$a(m) := \frac{\alpha+\gamma}{2} + (-1)^m \frac{\alpha-\gamma}{2}, \quad b(m) := \frac{1-(-1)^m}{2}\delta$$

and

$$c(m) := \frac{g(m)+h(m)}{2} + (-1)^m \frac{g(m)-h(m)}{2}.$$

Recurrence relations of this form are called *(binary) divide-and-conquer recurrences* and appear often in computer science, because algorithms based on the technique of *divide et impera* (divide and conquer) often reduce a problem of size m to

the solution of two problems of approximately equal sizes $\lfloor m/2 \rfloor$ and $\lceil m/2 \rceil$, where $m = \lfloor m/2 \rfloor + \lceil m/2 \rceil, m \in \mathbb{N}_0$. The solutions of the two subproblems are then used to solve the original problem.

A prominent example is given by the problem to sort m records, $m > 1$. One method is called *mergesort* [3, p.79] and consists in dividing the m records into two approximately equal parts, one of size $\lfloor m/2 \rfloor$ and the other of size $\lceil m/2 \rceil$. After each part has been sorted separately by the same method, the records are merged into their final order by doing at most $m - 1$ further comparisons. The total number of comparisons performed is at most $f(m)$, where $f(1) = 0$ and $f(m) = f\big(\lfloor m/2 \rfloor\big) + f\big(\lceil m/2 \rceil\big) + m - 1$, $m > 1$. This is a special case of (1.3) with $a(m) = b(m) = 1$, $c(m) = m - 1$ and $\zeta = 0$.

Another well-known example is given by Stern's diatomic sequence (this is the sequence A002487 in the OEIS [8]), defined by $s(1) = 1$, $s(2m) = s(m)$ and $s(2m + 1) = s(m) + s(m + 1)$ for all $m \in \mathbb{N}$, that is $\alpha = \gamma = \delta = \zeta = 1$ and $g(m) = h(m) = 0$ for all $m \in \mathbb{N}$. Note that the value $s(0)$ has to be 0, since from the recurrence relation for odd indices for $m = 0$ we get $s(1) = s(0) + s(1)$, that is $s(0) = 0$.

Adding 1 and starting with the index 0 we get from $g(m)$ the sequence $d(m) := g(m + 1) + 1, m \in \mathbb{N}_0$, satisfying for all $m \geq 1$ the recurrence relation

$$\left.\begin{aligned} d(0) = g(1) + 1 = -1 + 1 = 0,\ d(1) = g(2) + 1 = 0 \\ d(2m) = g(2m + 1) + 1 = g(m + 1) + m + 1 = d(m) + m \\ d(2m + 1) = g(2m + 1 + 1) + 1 = g(m + 1) + 1 = d(m) \end{aligned}\right\} \tag{1.4}$$

This is the sequence A233931 in the OEIS [8] with the first few values

$$(0, 0, 1, 0, 3, 1, 3, 0, 7, 3, 6, 1, 9, 3, 7, 0, 15, 7, 12, 3, 16, 6, 12, 1, 21, 9, 16, 3, \ldots)$$

and a special case of (1.2) with $\alpha = 1, \gamma = 1, \delta = 0, \zeta = 0$, $g(m) = m$ and $h(m) = 0$.

In this paper we shall determine the solution of (1.1) by writing $m \in \mathbb{N}$ either in the form $m = 2^{a_0} + \cdots + 2^{a_l}$ with the strictly decreasing integer sequence $a_0 > a_1 > \cdots > a_l \geq 0$, $l \geq 0$, or using the binary expansion of $m = (b_n \ldots b_0)_2 := \sum_{k=0}^{n} b_k 2^k$, where $n \in \mathbb{N}_0$ and $b_0, \ldots, b_n \in \{0, 1\}$. Moreover, we shall also give sharp lower and upper bounds for $g(m)$.

Finally, we shall explore some properties of the sequence $a(m) := \frac{1}{3}\big(m^2 + g(m)\big)$, $m \in \mathbb{N}$, giving the edge ranking number of complete graphs and investigate the sequence of the partial sums of $g(m)$, $m \in \mathbb{N}$, and the sequence $d(m) := g(m + 1) + 1$, $m \in \mathbb{N}_0$.

2 Preliminaries

We start with a lemma which gives the values of the Bodlaender sequence for some special numbers.

Lemma 2.1 *Let $a, b, c \in \mathbb{N}_0$, then*

$$g(2^a) = -1, \qquad a \geq 0 \tag{2.1}$$
$$g(2^a - 1) = 2^{a-1} - 2, \qquad a \geq 1 \tag{2.2}$$
$$g(2^a + 1) = 2^a - 2, \qquad a \geq 0 \tag{2.3}$$
$$g(2^a + 2^b) = 2^{a-b} - 2, \qquad a > b \geq 0 \tag{2.4}$$
$$g(3 \cdot 2^a) = 0, \qquad a \geq 0 \tag{2.5}$$
$$g(2^a + 2^b - 1) = 2^{a-1} + 2^{b-1} + 2^{a-b} - 3, \qquad a > b \geq 1 \tag{2.6}$$
$$g(2^a + 2^b + 1) = 2^a + 2^b - 2^{a-b-1} - 3, \qquad a > b \geq 0 \tag{2.7}$$
$$g(2^a + 2^b + 2^c) = 2^{a-c} + 2^{b-c} + 2^{a-b-1} - 3, \qquad a > b > c \geq 0 \tag{2.8}$$

Proof. a) By repeated use of (1.1) we have $g(2^a) = g(1) = -1$.
b) By (1.1) and Eq.(2.1) we have $g(2^a - 1) = g\big(2 \cdot (2^{a-1} - 1) + 1\big) = g(2^{a-1} - 1 + 1) + 2^{a-1} - 1 = -1 + 2^{a-1} - 1 = 2^{a-1} - 2$.
c) Let $f(a) := g(2^a + 1)$, $a \geq 0$, then by (1.1) the sequence $\big(f(a)\big)_{a \in \mathbb{N}_0}$ satisfies the recurrence relation

$$f(a+1) = g(2^{a+1} + 1) = g(2 \cdot 2^a + 1) = g(2^a + 1) + 2^a = f(a) + 2^a, \; a \geq 0,$$

with $f(0) = g(2) = -1$. The solution of this linear first-order recurrence relation can be obtained by backward substitution. After a substitutions it follows $f(a) = f(a-1) + 2^{a-1} = f(a-2) + 2^{a-2} + 2^{a-1} = \cdots = f(0) + 2^0 + 2^1 + \cdots + 2^{a-1} = -1 + (2^a - 1) = 2^a - 2$ and this proves the formula (2.3).
d) By (1.1) we have $g(2^a + 2^b) = g\big(2^b(2^{a-b} + 1)\big) = g(2^{a-b} + 1) = 2^{a-b} - 2$, where the last equation follows from (2.3).
e) By Eq.(2.4) we have $g(3 \cdot 2^a) = g(2^{a+1} + 2^a) = 2^{a+1-a} - 2 = 0$.
f) By (1.1) and Eq.(2.4) we have $g(2^a + 2^b - 1) = g\big(2 \cdot (2^{a-1} + 2^{b-1} - 1) + 1\big) = g(2^{a-1} + 2^{b-1} - 1 + 1) + 2^{a-1} + 2^{b-1} - 1 = 2^{a-b} - 2 + 2^{a-1} + 2^{b-1} - 1$ and this is exactly formula (2.6).
g) By (1.1) we have $g(2^a + 2^b + 1) = g\big(2 \cdot (2^{a-1} + 2^{b-1}) + 1\big) = g(2^{a-1} + 2^{b-1} + 1) + 2^{a-1} + 2^{b-1}$. Let $f(a, b) := g(2^a + 2^b + 1)$, then the above equation means that

$f(a,b)$ satisfies the recurrence relation

$$f(a,b) = f(a-1,b-1) + 2^{a-1} + 2^{b-1},$$

which can be solved by backward substitution. After b substitutions we obtain $f(a,b) = f(a-b,0) + 2^{a-b} + 2^0 + 2^{a-b+1} + 2^1 + \cdots + 2^{a-1} + 2^{b-1}$. The first term is by (1.1) and (2.3) equal to $f(a-b,0) = g(2^{a-b} + 2^0 + 1) = g\big(2 \cdot (2^{a-b-1} + 1)\big) = g(2^{a-b-1} + 1) = 2^{a-b-1} - 2$. The other term is the sum of three geometric series, namely $(1 + 2 + \cdots + 2^{b-1}) + (1 + 2 + \cdots + 2^{a-1}) - (1 + 2 + \cdots + 2^{a-b-1}) = (2^b - 1) + (2^{a-b} - 1) - (2^{a-b} - 1) = 2^a + 2^b - 2^{a-b} - 1$. Summing up the two terms it follows $2^{a-b-1} - 2 + 2^a + 2^b - 2^{a-b} - 1 = 2^a + 2^b - 2^{a-b-1} - 3$, as asserted.
h) By (1.1) and Eq.(2.7) we have $g(2^a + 2^b + 2^c) = g\big(2^c \cdot (2^{a-c} + 2^{b-c} + 1)\big) = g(2^{a-c} + 2^{b-c} + 1) = 2^{a-c} + 2^{b-c} + 2^{a-b-1} - 3$. □

The next lemma shows that it is sufficient to consider only the case $a_l = 0$, that is m is an odd number.

Lemma 2.2 *Let $m = 2^{a_0} + \cdots + 2^{a_l}$, where $a_0 > a_1 > \cdots > a_l \geq 0$, $l \geq 0$, is a strictly decreasing integer sequence, then*
for $a_l \neq 0$:

$$g(m) = g(2^{-a_l} m) \tag{2.9}$$

and for $a_l = 0$:

$$g(m) = g\big(2^{-a_{l-1}}(m-1) + 1\big) + (1 - 2^{-a_{l-1}})(m-1) \tag{2.10}$$

Proof. a) Let $a_l \neq 0$, then by (1.1)

$$\begin{aligned} g(m) &= g(2^{a_0} + \cdots + 2^{a_l}) = g\big(2^{a_l}(2^{a_0-a_l} + \cdots + 2^{a_{l-1}-a_l} + 1)\big) \\ &= g(2^{a_0-a_l} + \cdots + 2^{a_{l-1}-a_l} + 1) = g\big(2^{-a_l}(m - 2^{a_l}) + 1\big) = g(2^{-a_l} m) \end{aligned}$$

b) Now let $a_l = 0$, then by (1.1) and since $\sum_{k=0}^{l-1} 2^{a_k} = m - 1$ we obtain

$$\begin{aligned} g(m) &= g(2^{a_0} + \cdots + 2^{a_{l-1}} + 1) = g\big(2 \cdot (2^{a_0-1} + \cdots + 2^{a_{l-1}-1}) + 1\big) \\ &= g(2^{a_0-1} + \cdots + 2^{a_{l-1}-1} + 1) + 2^{a_0-1} + \cdots + 2^{a_{l-1}-1} \\ &= g\big(2 \cdot (2^{a_0-2} + \cdots + 2^{a_{l-1}-2}) + 1\big) + \frac{1}{2}(m-1) \\ &= g(2^{a_0-2} + \cdots + 2^{a_{l-1}-2} + 1) + \frac{1}{4}(m-1) + \frac{1}{2}(m-1) \end{aligned}$$

Repeating this procedure a_{l-1} times and noting that $\sum_{k=1}^{a_{l-1}} 2^{-k} = \sum_{k=0}^{a_{l-1}} 2^{-k} - 1 = 2 \cdot \left(1 - 2^{-a_{l-1}-1}\right) - 1 = 1 - 2^{-a_{l-1}}$ we obtain

$$\begin{aligned} g(m) &= g(2^{a_0-a_{l-1}} + \cdots + 2^{a_{l-1}-a_{l-1}} + 1) + (m-1)\sum_{k=1}^{a_{l-1}} 2^{-k} \\ &= g\left(2^{-a_{l-1}}(m-1)+1\right) + (1-2^{-a_{l-1}})(m-1) \end{aligned}$$

and this proves Eq.(2.10). □

Note that in Eq.(2.10) the argument of g on the left-hand side has $l+1$ summands, whereas on the right-hand side it has only l summands.

3 Main Result

We can now prove our main result.

Theorem 3.1 *Let $m = 2^{a_0} + \cdots + 2^{a_l}$, where $a_0 > a_1 > \cdots > a_l \geq 0$, $l \geq 0$, is a strictly decreasing integer sequence, then*

$$g(m) = 2^{-a_l}(m - 2^{a_l}) - (l+1) - \sum_{k=1}^{l} 2^{a_{l-k}} \left(\sum_{j=1}^{k-1} 2^{-a_{l-j}-1} \right) \tag{3.1}$$

In particular, for $a_l = 0$

$$g(m) = m - 1 - (l+1) - \sum_{k=1}^{l} 2^{a_{l-k}} \left(\sum_{j=1}^{k-1} 2^{-a_{l-j}-1} \right) \tag{3.2}$$

Proof. It is sufficient to prove Eq.(3.2), since for $a_l \neq 0$ we have $g(2^{a_0} + \cdots + 2^{a_l}) = g(m) = g(2^{-a_l} m) = g(2^{a_0-a_l} + \cdots + 2^{a_{l-1}-a_l} + 1)$ by Eq.(2.9). We obtain Eq.(3.1) by simply writing $a_k - a_l$ instead of a_k, $k = 0, 1, \ldots, l$ in Eq.(3.2) and noting that $m - 1 = 2^{a_0-a_l} + \cdots + 2^{a_{l-1}-a_l}$ and $a_{l-k} - a_l - (a_{l-j} - a_l) - 1 = a_{l-k} - a_{l-j} - 1$. Thus the double sum and the term $l+1$ in (3.1) do not change. For the first term we obtain $2^{a_0-a_l} + \cdots + 2^{a_{l-1}-a_l} = 2^{-a_l}(2^{a_0} + \cdots + 2^{a_{l-1}} + 2^{a_l} - 2^{a_l}) = 2^{-a_l}(m - 2^{a_l})$ thus obtaining Eq.(3.1).
We now prove the case $a_l = 0$.
By (2.10) we have

$$g\left(\sum_{k=0}^{l-1} 2^{a_k} + 1\right) = g\left(\sum_{k=0}^{l-2} 2^{a_k - a_{l-1} - 1} + 1\right) + (1 - 2^{-a_{l-1}}) \sum_{k=1}^{l} 2^{a_{l-k}} \tag{3.3}$$

where the argument of g on the left-hand side consists of $l+1$ terms and that on the right-hand side of l terms. Setting $\beta_k := a_k - a_{l-1} - 1$ for any $k = 0, 1, \ldots, l-2$, and applying again Eq.(2.10) to the first term in the above equation (3.3), we obtain

$$g\Big(\sum_{k=0}^{l-2} 2^{\beta_k} + 1\Big) = g\Big(\sum_{k=0}^{l-3} 2^{\beta_k - \beta_{l-2} - 1} + 1\Big) + (1 - 2^{-\beta_{l-2}}) \sum_{k=2}^{l} 2^{\beta_{l-k}}$$

or, in terms of a_k, noting that for $k = 0, 1, \ldots, l-3$ we have
$\beta_k - \beta_{l-2} - 1 = a_k - a_{l-1} - 1 - (a_{l-2} - a_{l-1} - 1) - 1 = a_k - a_{l-2} - 1$

$$g\Big(\sum_{k=0}^{l-2} 2^{a_k - a_{l-1} - 1} + 1\Big) = g\Big(\sum_{k=0}^{l-3} 2^{a_k - a_{l-2} - 1} + 1\Big) + \big(1 - 2^{-(a_{l-2} - a_{l-1} - 1)}\big) \sum_{k=2}^{l} 2^{a_{l-k} - a_{l-1} - 1}$$

or, substituting this equation into (3.3)

$$g\Big(\sum_{k=0}^{l-1} 2^{a_k} + 1\Big) = g\Big(\sum_{k=0}^{l-3} 2^{a_k - a_{l-2} - 1} + 1\Big) + (1 - 2^{-a_{l-1}}) \sum_{k=1}^{l} 2^{a_{l-k}} + \big(1 - 2^{-(a_{l-2} - a_{l-1} - 1)}\big) \sum_{k=2}^{l} 2^{a_{l-k} - a_{l-1} - 1}$$

Note that the argument of g on the right-hand side has now $l-1$ terms. Repeating this procedure $l-1$ times we finally obtain

$$g(2^{a_0} + \cdots + 2^{a_{l-1}} + 1) = g\big(2^{a_0 - a_1 - 1} + 1\big) + (1 - 2^{-a_{l-1}}) \sum_{k=1}^{l} 2^{a_{l-k}} + \big(1 - 2^{-(a_{l-2} - a_{l-1} - 1)}\big) \sum_{k=2}^{l} 2^{a_{l-k} - a_{l-1} - 1} + \cdots + \big(1 - 2^{-(a_1 - a_2 - 1)}\big) \sum_{k=l-1}^{l} 2^{a_{l-k} - a_2 - 1}$$

Hence, by Eq.(2.3) and since $g\big(2^{a_0-a_1-1}+1\big) = 2^{a_0-a_1-1} - 2 = \big(1-2^{-(a_0-a_1-1)}\big)\cdot 2^{a_0-a_1-1} - 1 = \big(1-2^{-(a_0-a_1-1)}\big)\sum_{k=l}^{l} 2^{a_{l-k}-a_1-1} - 1$ it follows

$$\begin{aligned} g(2^{a_0}+\cdots+2^{a_{l-1}}+1) = -1 &+ (1-2^{-a_{l-1}})\sum_{k=1}^{l} 2^{a_{l-k}} \\ &+ \big(1-2^{-(a_{l-2}-a_{l-1}-1)}\big)\sum_{k=2}^{l} 2^{a_{l-k}-a_{l-1}-1} \\ &+ \cdots + \big(1-2^{-(a_1-a_2-1)}\big)\sum_{k=l-1}^{l} 2^{a_{l-k}-a_2-1} \\ &+ \big(1-2^{-(a_0-a_1-1)}\big)\sum_{k=l}^{l} 2^{a_{l-k}-a_1-1} \end{aligned}$$

A further simplification of the right-hand side of this equation leads to

$$\begin{aligned} g\Big(\sum_{k=0}^{l-1} 2^{a_k}+1\Big) &= -1 + \sum_{k=1}^{l} 2^{a_{l-k}} - 1 - \sum_{k=2}^{l} 2^{a_{l-k}-a_{l-1}} \\ &\quad + \sum_{k=2}^{l} 2^{a_{l-k}-a_{l-1}-1} - 1 - \sum_{k=3}^{l} 2^{a_{l-k}-a_{l-2}} \\ &\quad + \cdots + \sum_{k=l-1}^{l} 2^{a_{l-k}-a_2-1} - 1 - 2^{a_0-a_1} + 2^{a_0-a_1-1} - 1 \\ &= \sum_{k=1}^{l} 2^{a_{l-k}} - \sum_{k=2}^{l} 2^{a_{l-k}-a_{l-1}-1} - \sum_{k=3}^{l} 2^{a_{l-k}-a_{l-2}} \\ &\quad - \cdots - \sum_{k=l}^{l} 2^{a_{l-k}-a_1-1} - \sum_{k=1}^{l} 1 - 1 \end{aligned}$$

and finally

$$g\Big(\sum_{k=0}^{l-1} 2^{a_k}+1\Big) = 2^{a_{l-1}} + 2^{a_{l-2}}\big(1-2^{-a_{l-1}-1}\big) + 2^{a_{l-3}}\big(1-\sum_{j=1}^{2} 2^{-a_{l-j}-1}\big)$$
$$+\cdots+2^{a_0}\big(1-\sum_{j=1}^{l-1} 2^{-a_{l-j}-1}\big)-(l+1)$$
$$=\sum_{k=1}^{l}\big\{2^{a_{l-k}}\big(1-\sum_{j=1}^{k-1} 2^{-a_{l-j}-1}\big)-1\big\}-1$$

and this is Eq.(3.2), since $m-1=\sum_{k=1}^{l} 2^{a_{l-k}}$ and $\sum_{k=1}^{l} 1 = l$. □

Remark 3.2 *Note that $s_2(m) := l+1$ is the number of nonzero digits of the binary expansion of m and $m-(l+1)$ is the exponent of the highest power of 2 which divides $m!$ according to a theorem of Legendre [6, pp.10-12] (for the special case $p=2$).*

For example, let $m = 13 = 2^3+2^2+2^0$, that is $l=2, a_0=3, a_1=2, a_2=0$, then by (3.2) we have $g(13) = 13-1-(2+1)-2^{a_0-a_1-1} = 9-2^{3-2-1} = 8$.
A consequence of Theorem 3.1 is the

Proposition 3.3 *For all $m \in \mathbb{N}$ we have*

$$-1 \le g(m) \le m-1 \tag{3.4}$$

and, therefore, for all $m \in \mathbb{N}_0$

$$0 \le d(m) \le m+1 \tag{3.5}$$

Proof. Let $m = 2^{a_0}+\cdots+2^{a_l}$, where $a_0 > a_1 > \cdots > a_l \ge 0$, $l \ge 0$, is a strictly decreasing integer sequence, then $l+1$ and the double sum in (3.1) are always ≥ 0 and therefore $g(m) \le 2^{-a_l}(m-2^{a_l}) = 2^{-a_l}m - 1 \le m-1$, since $2^{a_l} \ge 1$. This proves the upper bound.

In order to prove that -1 is a lower bound we note first that by Eq.(2.1) this value is attained for $m = 2^n$, $n \ge 0$, and secondly by Eq.(2.9) it is sufficient to consider only odd numbers m, that is $a_l = 0$. As shown in the proof of Theorem 3.1 the formula (3.2) can be written as

$$g(m) = \sum_{k=1}^{l}\big\{2^{a_{l-k}}\big(1-\sum_{j=1}^{k-1} 2^{-a_{l-j}-1}\big)-1\big\}-1,$$

therefore we have to prove that $2^{a_{l-k}}\big(1-\sum_{j=1}^{k-1}2^{-a_{l-j}-1}\big)-1 = 2^{a_{l-k}}-1-\sum_{j=1}^{k-1}2^{a_{l-k}-a_{l-j}-1} \geq 0$. This can be proved as follows: By definition $a_{l-j} \geq j$ for any $j = 0, 1, \ldots, l$ or $-a_{l-j} \leq -j$ and, therefore, $\sum_{j=1}^{k-1}2^{-a_{l-j}} \leq \sum_{j=1}^{k-1}2^{-j} = 1 - 2^{-(k-1)}$. It follows $1 - \frac{1}{2}\sum_{j=1}^{k-1}2^{-a_{l-j}} \geq 1 - \frac{1}{2}\cdot\big(1-\frac{1}{2^{k-1}}\big) = \frac{1}{2} + \frac{1}{2^k} \geq \frac{1}{2^k}$ and, multiplying both sides by $2^{a_{l-k}}$ and adding to both sides -1, we obtain $2^{a_{l-k}}\big(1-\sum_{j=1}^{k-1}2^{-a_{l-j}-1}\big)-1 \geq 2^{a_{l-k}-k}-1 \geq 0$, since $a_{l-k}-k \geq 0$ for any $k = 1, \ldots, l$. This proves Eq.(3.4). Finally, the double inequality (3.5) follows immediately from the definition of the sequence $\big(d(m)\big)_{m\in\mathbb{N}_0}$. □

By means of $h(m) := g(m+1), m \in \mathbb{N}_0$, $h(0) = g(1) = -1$, we can give another representation of the solution (3.1) which will lead to an improvement of Proposition 3.3.

Proposition 3.4 *The sequence* $\big(h(m)\big)_{m\in\mathbb{N}_0}$ *satisfies for all* $m \geq 1$ *the recurrence relation*

$$\left.\begin{aligned} h(0) &= -1 \ \big(h(1) = -1\big) \\ h(2m) &= h(m) + m \\ h(2m+1) &= h(m) \end{aligned}\right\} \tag{3.6}$$

Proof. By (1.1) it follows $h(0) = g(1) = -1$ and $h(1) = g(2) = -1$. Further

$$\begin{aligned} h(2m) &= g(2m+1) = g(m+1) + m = h(m) + m \\ h(2m+1) &= g(2m+2) = g(m+1) = h(m) \end{aligned}$$

and this proves the proposition. □

Note that (3.6) is a special case of (1.2) with $\alpha = 1, \gamma = 1, \delta = 0, \zeta = -1$, $g(m) = m$ and $h(m) = 0$.

Using the binary expansion of $m = (b_n \ldots b_0)_2$, where $b_0, \ldots, b_n \in \{0, 1\}$, we can derive an alternative solution of (1.1).

Theorem 3.5 *Let* $m = (b_n \ldots b_0)_2 \in \mathbb{N}_0$, *then the sequence* $\big(h(m)\big)_{m\in\mathbb{N}_0}$ *satisfies the recurrence relation*

$$h\big((b_n \ldots b_0)_2\big) = h\big((b_n \ldots b_1)_2\big) + (1 - b_0)\cdot(b_n \ldots b_1)_2 \tag{3.7}$$

with the solution

$$g\big((b_n \ldots b_0)_2 + 1\big) = h\big((b_n \ldots b_0)_2\big) = -1 + \sum_{k=0}^{n-1}(1 - b_k)(b_n \ldots b_{k+1})_2 \tag{3.8}$$

and, therefore, for all $m \in \mathbb{N}_0$

$$d\big((b_n \dots b_0)_2\big) = g\big((b_n \dots b_0)_2 + 1\big) + 1 = \sum_{k=0}^{n-1} (1 - b_k)(b_n \dots b_{k+1})_2 \tag{3.9}$$

Proof. Let $m = (b_n \dots b_0)_2 \in \mathbb{N}_0$, then

$$\begin{aligned} h\big((b_n \dots b_0)_2\big) &= b_0 h\big((b_n \dots b_1)_2\big) + (1 - b_0)\Big(h\big((b_n \dots b_1)_2\big) + (b_n \dots b_1)_2\Big) \\ &= h\big((b_n \dots b_1)_2\big) + (1 - b_0) \cdot (b_n \dots b_1)_2 \end{aligned}$$

since for $b_0 = 0$, that is m is even, we have $m/2 = (b_n \dots b_1)_2$ and by (3.6) $h(m) = h(m/2) + m/2 = h\big((b_n \dots b_1)_2\big) + (b_n \dots b_1)_2$ and for $b_0 = 1$, that is m is odd, we have $(m-1)/2 = (b_n \dots b_1)_2$ and by (3.6) $h(m) = h\big((m-1)/2\big) + m/2 = h\big((b_n \dots b_1)_2\big)$. This proves Eq.(3.7).

Repeating this procedure $n-1$ times we finally obtain $h\big((b_n \dots b_0)_2\big) = h(b_n) + \sum_{k=0}^{n-1}(1 - b_k)(b_n \dots b_{k+1})_2$ and this proves Eq.(3.8), since $b_n = 1$ and $h(1) = -1$. Finally, by definition it follows Eq.(3.9). □

For example, let $m = 10 = (1010)_2$, that is $n = 3$, $b_0 = 0$, $b_1 = 1$, $b_2 = 0$ and $b_3 = 1$. Then

$$\begin{aligned} g(11) &= -1 + \sum_{k=0}^{2} (1 - b_k)(b_3 \dots b_{k+1})_2 \\ &= -1 + (1 - b_0)(b_3 b_2 b_1)_2 + (1 - b_1)(b_3 b_2)_2 + (1 - b_2)(b_3)_2 \\ &= -1 + (101)_2 + 0 \cdot (10)_2 + 1 = -1 + 5 + 0 + 1 = 5. \end{aligned}$$

From Eq.(3.8) one can deduce that h attains its greatest value at $m = 2^n = (10 \dots 0)_2$. In this case we have $b_k = 0$ for all $k = 0, 1, \dots, n-1$. Hence $1 - b_k = 1$ for all $k = 0, 1, \dots, n-1$. Therefore $g(2^n+1) = h(2^n) = -1 + \sum_{k=0}^{n-1}(b_n \dots b_{k+1})_2 = -1 + \sum_{k=0}^{n-1} 2^{n-1-k} = -1 + 2^n - 1 = 2^n - 2$ (see also Eq.(2.3)). Hence, $2^n + 1 - (2^n - 2) = 3$, that is for all $m \geq 2$ we have $m - g(m) \geq 3$.

Similarly, h attains its smallest value at $m = 2^n - 1 = (11 \dots 1)_2$. In this case we have $b_k = 1$ for all $k = 0, 1, \dots, n-1$ and hence $1 - b_k = 0$ for all $k = 0, 1, \dots, n-1$. Therefore $g\big((1 \dots 1)_2 + 1\big) = g(2^n) = -1$ (see also Eq.(2.1)). Hence, $g(m) \geq -1$ for all $m \geq 1$.

Thus, we have shown the desired improvement of Proposition 3.3.

Proposition 3.6 *For all $m \in \mathbb{N}_2$ we have*

$$-1 \leq g(m) \leq m-3, \tag{3.10}$$

where both bounds are sharp, since for $m = 2^n, n \geq 0$, it is $g(m) = -1$ and for $m = 2^n + 1, n \geq 0$, it is $g(m) = m - 3$.
Therefore, for all $m \in \mathbb{N}$ we have

$$0 \leq d(m) \leq m-1 \tag{3.11}$$

4 Some Consequences

In [1, Theorem 25] Bodlaender et al. showed that the edge ranking number of the complete graphs on p vertices, $p \in \mathbb{N}$, is given by $\chi'_r(K_p) = a(p)$, where $a(p) := \frac{1}{3}\big(p^2 + g(p)\big)$, and in [7, Theorem 7, Corollary 8] Lin, Juan and Wang showed that the edge ranking number of the Sierpiński graphs is given by $\chi'_r(S_p^n) = n\chi'_r(K_p) = \frac{n}{3}\big(p^2 + g(p)\big)$, $n, p \in \mathbb{N}_2$. (For a definition of the Sierpiński graphs we refer the reader to the seminal paper of Klavžar and Milutinovič [5] and to the survey paper on Sierpiński-type graphs by Hinz, Klavžar and Zemljič [4].)

Proposition 4.1 *The sequence $a(m) := \frac{1}{3}\big(m^2 + g(m)\big)$, $m \in \mathbb{N}$, satisfies for all $m \in \mathbb{N}$ the recurrence relation*

$$\left.\begin{aligned} a(1) &= 0 \\ a(2m) &= a(m) + m^2 \\ a(2m+1) &= a(m+1) + m(m+1) \end{aligned}\right\} \tag{4.1}$$

In particular,

$$3|(m^2 + g(m)). \tag{4.2}$$

Note that (4.1) is a special case of (1.2) with $\alpha = 1, \gamma = 0, \delta = 1, \zeta = 0$, $g(m) = m^2$ and $h(m) = m(m+1)$.

Proof. By definition $a(1) = \frac{1}{3}(1^2 + g(1)) = \frac{1}{3}(1-1) = 0$ and by (1.1) it follows that

$$\begin{aligned} a(2m) &= \frac{4m^2 + g(2m)}{3} = \frac{m^2 + g(m) + 3m^2}{3} = a(m) + m^2 \\ a(2m+1) &= \frac{(2m+1)^2 + g(2m+1)}{3} = \frac{4m^2 + 4m + 1 + g(m+1) + m}{3} \\ &= \frac{(m+1)^2 + g(m+1) + 3m^2 + 3m}{3} = a(m+1) + m(m+1) \end{aligned}$$

This proves (4.1) and shows that $a(m) \in \mathbb{N}_0$, thus proving (4.2). □

Note that for $m-1$ instead of m in Eq.(4.1) we obtain $a\big(2(m-1)+1\big) = a(2m-1) = a(m)+(m-1)m$ and therefore $a(2m)-a(2m-1) = m^2-m(m-1) = m$ for all $m \in \mathbb{N}$.

The first few values of $(a(m))_{m\in\mathbb{N}}$ (not available in the OEIS [8]) are

$$(0, 1, 3, 5, 9, 12, 17, 21, 29, 34, 42, 48, 59, 66, 77, 85, 101, 110, 124, 134, \ldots).$$

The next lemma gives the values of this sequence for some special numbers.

Lemma 4.2 *Let $n \in \mathbb{N}_0$, then*

$$a(2^n) = \frac{4^n-1}{3} \tag{4.3}$$

$$a(2^n-1) = \frac{4^n-1}{3} - 2^{n-1}, \qquad n \geq 1 \tag{4.4}$$

$$a(2^n+1) = \frac{4^n-1}{3} + 2^n \tag{4.5}$$

Proof. By definition and using Eqs.(2.1), (2.2) and (2.3), we have $a(2^n) = \frac{1}{3}\big(4^n + g(2^n)\big) = \frac{1}{3}(4^n-1)$, $a(2^n-1) = \frac{1}{3}\big((2^n-1)^2 + g(2^n-1)\big) = \frac{1}{3}(4^n-1) - 2^{n-1} = a(2^n)-2^{n-1}$ and $a(2^n+1) = \frac{1}{3}\big((2^n+1)^2+g(2^n+1)\big) = \frac{1}{3}(4^n-1)+2^n = a(2^n)+2^n$. □

We notice that (4.3) is the sequence A002450, (4.5) is the sequence A079319, whereas the sequence (4.4) is not available in the OEIS [8].

Note that for $m = 2^n$, $n \geq 0$, we have $a(m) = \frac{1}{3}(m^2-1)$ and for $m = 2^n+1$, $n \geq 0$, we have $a(m) = \frac{1}{3}(m^2+m-3)$. By Proposition 3.6 we obtain in this way

Corollary 4.3 *For all $m \in \mathbb{N}_2$ we have*

$$\frac{m^2-1}{3} \leq a(m) \leq \frac{m^2+m-3}{3}, \tag{4.6}$$

where both bounds are sharp.

To conclude we mention some properties of the sequence of the partial sums of $g(m)$.

Proposition 4.4 *Let* $t(m) := \sum_{k=1}^{m} g(k)$, $m \in \mathbb{N}$, *then*

$$\left.\begin{aligned} t(1) &= -1 \\ t(2m) &= 2t(m) + \binom{m}{2} \\ t(2m+1) &= t(m) + t(m+1) + \binom{m+1}{2} \end{aligned}\right\} \tag{4.7}$$

Note that (4.7) is a special case of (1.2) with $\alpha = 2, \gamma = 1, \delta = 1, \zeta = -1$, $g(m) = \binom{n}{2}$ and $h(m) = \binom{m+1}{2}$.

Proof. It is $t(1) = g(1) = -1$ and by definition we have

$$\begin{aligned} t(2m) = \sum_{k=1}^{2m} g(k) &= \sum_{k=1}^{m} g(2k) + \sum_{k=1}^{m-1} g(2k+1) + g(1) \\ &= \sum_{k=1}^{m} g(k) + \sum_{k=1}^{m-1} \big(g(k+1) + k\big) + g(1) \\ &= t(m) + \sum_{k=1}^{m-1} g(k+1) + \sum_{k=1}^{m-1} k + g(1) \\ &= t(m) + t(m) - g(1) + \binom{m}{2} + g(1) = 2t(m) + \binom{m}{2} \end{aligned}$$

Similarly,

$$\begin{aligned} t(2m+1) = \sum_{k=1}^{2m+1} g(k) &= \sum_{k=1}^{m} g(2k) + \sum_{k=1}^{m} g(2k+1) + g(1) \\ &= \sum_{k=1}^{m} g(k) + \sum_{k=1}^{m} g(k+1) + \sum_{k=1}^{m} k + g(1) \\ &= t(m) + t(m+1) - g(1) + \binom{m+1}{2} + g(1) \\ &= 2t(m) + \binom{m+1}{2} \end{aligned}$$

and this proves (4.7). □

The first few values of $\big(t(m)\big)_{m\in\mathbb{N}}$ (not available in the OEIS [8]) are

$$(-1,-2,-2,-3,-1,-1,1,0,6,8,13,13,21,23,29,28,42,48,59,61,76,\ldots).$$

The next lemma gives the values of this sequence for some special numbers.

Lemma 4.5 *Let $n\in\mathbb{N}_0$, then*

$$t(2^n)=\frac{4^n-5\cdot 2^n-n\cdot 2^n}{4} \tag{4.8}$$

$$t(2^n-1)=\frac{4^n-5\cdot 2^n-n\cdot 2^n}{4}+1, \qquad n\geq 1 \tag{4.9}$$

$$t(2^n+1)=\frac{4^n-5\cdot 2^n-n\cdot 2^n}{4}+2^n-2, \tag{4.10}$$

Proof. a) Let $f(n):=t(2^n)$, $n\geq 0$, then by (4.7) the sequence $\big(f(n)\big)_{n\in\mathbb{N}_0}$ satisfies the recurrence relation

$$f(n+1)=t(2^{n+1})=t(2\cdot 2^n)=2t(2^n)+\binom{2^n}{2}=2f(n)+2^{n-1}(2^n-1),\ n\geq 0,$$

with $f(0)=t(1)=-1$. The solution of this linear first-order recurrence relation can be obtained again by backward substitution. After n substitutions it follows

$$\begin{aligned}
f(n)&=2f(n-1)+2^{n-2}(2^{n-1}-1)\\
&=2\big(2f(n-2)+2^{n-3}(2^{n-2}-1)\big)+2^{n-2}(2^{n-1}-1)\\
&=2^2f(n-2)+2\cdot 2^{n-3}(2^{n-2}-1)+2^{n-2}(2^{n-1}-1)\\
&=\cdots\cdots\cdots\cdots\\
&=2^nf(0)+\sum_{k=0}^{n-1}2^k\cdot 2^{n-2-k}\cdot(2^{n-1-k}-1)\\
&=-2^n+2^{n-2}(2^n-1-n)=\frac{4^n-5\cdot 2^n-n\cdot 2^n}{4}
\end{aligned}$$

and this proves (4.8).
b) By definition and using Eqs.(4.8) and (2.1) we get $t(2^n-1)=\sum_{k=1}^{2^n-1}g(k)=\sum_{k=1}^{2^n}g(k)-g(2^n)=t(2^n)-(-1)=\frac{1}{4}(4^n-5\cdot 2^n-n\cdot 2^n)+1$ and this proves (4.9).
c) By definition and using Eqs.(4.8) and (2.3) we get $t(2^n+1)=\sum_{k=1}^{2^n+1}g(k)=\sum_{k=1}^{2^n}g(k)+g(2^n+1)=t(2^n)+2^n-2=\frac{1}{4}(4^n-5\cdot 2^n-n\cdot 2^n)+2^n-2$ and this

proves (4.10). □

Note that for $m = 2^n$, $n \geq 0$, that is $n = \log_2(m)$, we obtain from (4.8) the sequence $\underline{t}(m) := \frac{m}{4}(m - 5 - \log_2(m))$, $m \in \mathbb{N}$. Numerical results suggest that $\underline{t}(m)$ is a lower bound of $t(m)$, that is $\underline{t}(m) \leq t(m),\ m \in \mathbb{N}$.

An upper bound $\bar{t}(m)$ for $t(m)$ could be obtained from (3.10) as follows: It is $g(k) \leq k-3$ for all $k \in \mathbb{N}_2$ and, therefore, $t(m) = g(1)+\sum_{k=2}^{m} g(k) \leq -1+\sum_{k=2}^{m}(k-3) = -1+\sum_{k=2}^{m} k - 3\sum_{k=2}^{m} 1 = -1+\binom{m+1}{2}-1-3(m-1) = \frac{1}{2}(m^2-5m+2) =: \bar{t}(m)$. Since $\frac{1}{2}(1^2 - 5 \cdot 1 + 2) = -1 = t(1)$ we obtain

$$\underline{t}(m) \leq t(m) \leq \bar{t}(m),\ m \in \mathbb{N}.$$

We notice that $\bar{t}(m+4) = \underline{A034856}(m),\ m \in \mathbb{N}$.

Remark 4.6 *All the sequences encountered so far, namely* $g(m)$, $d(m) := g(m+1)+1$, $h(m) := g(m+1)$, $a(m) := \frac{1}{3}\big(m^2 + g(m)\big)$ *and* $t(m) := \sum_{k=1}^{m} g(k)$ *are of* divide-and-conquer type*, that is their ordinary generating functions defined as the power series* $G(s) := \sum_{m=1}^{\infty} g(m)s^m$, $D(s) := \sum_{m=0}^{\infty} d(m)s^m$, $H(s) := \sum_{m=0}^{\infty} h(m)s^m$, $A(s) := \sum_{m=1}^{\infty} a(m)s^m$ *and* $T(s) := \sum_{m=1}^{\infty} t(m)s^m$ *satisfy a functional equation known as* Mahlerian equation *(cf.[2])*

$$a_0(s)F(s) + a_1(s)F(s^2) + \cdots + a_n(s)F(s^{2^n}) = f(s) \tag{4.11}$$

in which $n \in \mathbb{N}$, $f(s)$ *is a formal series and* $a_0(s)$, $a_1(s), \ldots, a_n(s)$ *are polynomials not all zero. If* $f(s) = 0$, *then the solution of (4.11) is said to be a* Mahlerian series.

Indeed, applying standard generating function techniques to the recurrence relations from Eqs.(1.1), (1.4), (3.6), (4.1) and (4.7) we have the functional equations

$$\begin{aligned}
G(s) &= \Big(1+\frac{1}{s}\Big)G(s^2) + \frac{s^3}{(1-s^2)^2} \\
D(s) &= (1+s)D(s^2) + \frac{s^2}{(1-s^2)^2} \\
H(s) &= (1+s)H(s^2) + \frac{s^2}{(1-s^2)^2} \\
A(s) &= \Big(1+\frac{1}{s}\Big)A(s^2) + \frac{s^2}{(1+s)(1-s)^3} \\
T(s) &= s\Big(1+\frac{1}{s}\Big)^2 T(s^2) + \frac{s^3}{(1-s)(1-s^2)^2}
\end{aligned} \tag{4.12}$$

and the relations

$$D(s) = \frac{G(s)}{s} + \frac{1}{1-s},\ H(s) = \frac{G(s)}{s},\ A(s) = \frac{1}{3}\Big(G(s) + \frac{t(1+s)}{(1-s)^3}\Big),\ T(s) = \frac{G(s)}{1-s}.$$

Acknowledgments

The author would like to thank Andreas M. Hinz for careful reading of the manuscript, Sandi Klavžar for some advices and an anonymous referee who helped to improve the paper with useful remarks.

References

[1] H. Bodlaender, J. S. Deogun, K. Jansen, T. Kloks, D. Kratsch, H. Müller and Zs. Tuza, Ranking of graphs, *SIAM J. Discrete Mathematics* 11 (1998) 168-181.

[2] Ph. Dumas, Algebraic aspects of B-regular series, in: Automata, languages and programming (Lund, 1993), *Lectures Notes in Comput. Sci.* 700, Springer-Verlag, Berlin (1993) 457-468.

[3] R. L. Graham, D. E. Knuth, O. Patashnik, Concrete Mathematics, Addison-Wesley, Reading, 1990.

[4] A. M. Hinz, S. Klavžar, S. S. Zemljič, A survey and classification of Sierpiński-type graphs, *Discrete Appl. Math.* 217 (2017) 565-600.

[5] S. Klavžar, U. Milutinovič, Graphs $t(n, k)$ and a variant of the Tower of Hanoi problem, *Czechoslovak Math. J.* 47 (122) (1997) 95-104.

[6] A. M. Legendre, Theórie des nombres, Tome I, Troisième Édition, Firmin Didot frères, Paris, 1830.

[7] Y. L. Lin, J. S. Juan, Y. L. Wang, Finding the edge ranking number through vertex partitions, *Discrete Appl. Math.* 161 (2013) 1067-1071.

[8] OEIS Foundation Inc. (2011), The On-Line Encyclopedia of Integer Sequences, `https://oeis.org/`

Three models for resilient network design and a genetic algorithm to approach them

Jelena Sedlar[1], Martina Milat[1]

[1]Faculty of civil engineering, architecture and geodesy, University of Split,
Matice hrvatske 15, HR-21000 Split, Croatia
(jelena.sedlar@gradst.hr, martina.milat@gradst.hr)

Abstract

This paper examines the types of directed networks with one source and one sink. The problem of resilient network design is studied with respect to such networks. The upper and the lower bound of the capacity are given for each edge in the network, while the cost of each edge is given as the function of edge capacity. Said problem of network design consists of selecting a subset of edges in the given network, which induces an optimal subnetwork to be resilient after disruptive event. The restoration behaviour of each edge in a network N after the disruptive event is described by using a non-linear function that enables the modelling of three components affecting resilience: the remaining capacity of the edge after the disruption, the degree to which capability can be recovered and the recovery speed. Three different models for designing a resilient network are proposed and then formulated as problems of non-linear optimisation. A simple genetic algorithm using stochastic ranking, which can be used to approach all three proposed network design problems, is proposed. One numerical example is used to illustrate the proposed procedure and the effectiveness of the proposed algorithm.

Keywords: flow networks, design, resilience, genetic algorithm

MSC: 90C27

DOI: https://doi.org/10.5592/CO/CCD.2018.09

1 Introduction

For the proper functioning of everyday life, we often rely on the regular functioning of many networks supporting our daily routine; from infrastructure networks, to traffic networks or information networks. Since disruptive events may occur which affect the functioning of those networks, there is an ever-increasing demand for ensuring their regular functioning by minimising the adverse effects of disruptive events. One possible solution is by increasing system redundancy, which is often the focus of reliability engineering [10]. This approach focusses on increasing the probability of a system to properly operate for a specified period of time. On the other hand, there is a resilience engineering approach which focuses more on improving the system's capability to recuperate from disruptive events in a sense that a desired level of performance is quickly recovered after the disruption [4]. The concept of resilience was first introduced in [8] which confirmed its significant role in maintaining the stability of ecological systems. Since then, this research topic has received increasing attention, and substantial effort has been dedicated to defining and to measuring system resilience [3]. In [2], an indicator has been developed to measure the component importance by quantifying its adverse impact on system resilience when the disruption affected that component.

There are many definitions of resilience [9] and all these definitions aim at better understanding of system resilience in different contexts. However, the methods for resilient infrastructure system design have not been so extensively studied. There are several related studies ([5], [12]), however, since in reality many infrastructure systems exist in the form of networks, it is of the utmost interest to study the resilient network design methods ([1], [6], [7]). One such study of resilient network design is [13] where the restoration behaviour is described by using a non-linear function that enables the modelling of more refined attributes of restoration at the component level. Three components influencing resilience are especially considered: the remaining capacity (absorptive ability), the degree to which capability can be recovered (restoration ability) and the recovery speed. The method for designing a resilient network is consequently formulated as the problem of non-linear optimisation.

A network is given as a directed graph with one source and one sink. When constructing a real-life infrastructure system, due to various technical reasons it is often impracticable to construct the connection between all nodes. Since the network considered here is a mathematical model of the infrastructure system intended for construction, the edges included in the network represent the connections which

are technically feasible in the real-life. Generally, not every technically feasible connection is actually constructed in the infrastructure system, as it would often be too expensive and unnecessary. A satisfactory network can be obtained by constructing only some of the possible connections. Therefore, in designing the infrastructure system the engineer has to choose which feasible connections will be constructed. Mathematically, that means we have to choose a subset of edges in a given network which induces an optimal subnetwork with respect to the capacity, the cost and the resilience. In [13] said problem is approached by using a probabilistic solution discovery algorithm combined with stochastic ranking.

The proposed resilient network design model from [13] can be improved by allowing variable capacities of the edges and by introducing a variable cost of the edge construction defined as the function of the edge capacity. The aim of this paper is to implement these improvements in the resilient network design model and, consequently, to propose a simple genetic algorithm approach to the problem, since the probabilistic solution discovery algorithm proposed in [13] can hardly be extended to the model with variable edge capacities. Three different resilient network design models are proposed and then formulated as the problems of non-linear optimisation, and subsequently approached by a single algorithm. One numerical example is used to illustrate the proposed procedure and the effectiveness of the proposed method.

The present paper is structured as follows: the following section describes the preliminaries and introduces the basic notation. In Section 3, three different resilient network design models are proposed and then formulated as problems of non-linear optimisation. Section 4 presents a simple genetic algorithm combined with stochastic ranking which may be used to approach all three problems stated in Section 3. Finally, in Section 5, the algorithm presented in Section 4 is applied to one particular network N, and all three design problems are solved by using said algorithm. The effectiveness of the proposed method on the given example is verified by comparing the obtained results with the results obtained by exact calculation.

2 Preliminaries

Let $G = (V, E)$ be a directed graph on the set of vertices V and the set of directed edges E. Directed edge $(v, w) \in E$ will often be denoted by abbreviation vw.

Definition 1 *Flow network N is defined by $N = (G, u, s, r)$, where $G = (V, E)$ is a directed graph, $u : E \to \mathbb{R}^+$ is a non-negative edge capacity function, $s \in V$ is the*

source and $r \in V$ is the sink vertex.

The value $u(vw)$ of the edge capacity function u in the network N represents the capacity of the edge $vw \in E$ and will often be shortly denoted by u_{vw}.

Definition 2 *Flow in the flow network $N = (G, u, s, r)$ is a non-negative function $\psi : E \to \mathbb{R}^+$ such that $\psi(vw) \leq u(vw)$ for every $vw \in E$ and $\sum_{(w,v)\in E} \psi(w, v) = \sum_{(v,z)\in E} \psi(v, z)$ for every vertex $v \in V \backslash \{s, r\}$. The value of the flow ψ in the network N is defined as $\psi(N) = \sum_{(s,v)\in E} \psi(s, v)$.*

The standard problem in the flow networks theory is to find a flow in the network with the maximum possible value. Since the flow networks are often used to model various real-life problems, the designing or studying such networks involves the cost of building the network or similar costs. In addition, once it has been designed, the flow network exists in time, where a disruptive event can occur and diminish the edge capacity of several or all edges in the network. Since the original network is usually designed so that the maximum flow equals the amount of the actually required flow, it is of utmost interest to repair the network after the disruptive event as soon as possible. The value of the flow which is lost in a disruptive event e^d is denoted by $Loss(t_d)$, while the value of flow which can be recovered (of the flow which was lost) in the time t after t_d is denoted by $Recovery(t)$. The value of $Loss(t_d)$ and $Recovery(t)$ is usually given by real-life circumstances and is therefore predefined in the analysed problem. Now we can define resilience function $R(t)$ of the network N in the recovery time t as

$$R(t) = \frac{Recovery(t)}{Loss(t_d)}.$$

Note that $R(t)$ is the percentage of the recovered flow in the recovery time t, hence $R(t) \leq 1$. Therefore, it is now possible to consider the maximum flow in the network, the cost of the network and the resilience of the network in the recovery time t.

One context in which we can consider the maximum flow, the cost and the resilience of a network is the context of network design. In [13] the authors studied the problem of designing a network with the lowest cost such that the designed network satisfies a given lower bound on resilience in the recovery time t. The concept of resilience is introduced by comparing two networks, the original network denoted by N and the disrupted network after the recovery time t which is denoted by $N^*(t)$. Both N and $N^*(t)$ are based on the same directed graph G where the vertices are

denoted by $\{1, \ldots, n\}$ with vertex 1 being the source and vertex n being the sink, they differ only in the edge capacity function.

The edge capacity function in the original undisrupted network N is denoted by u, while the edge capacity function in the disrupted network $N^*(t)$ after the recovery time t is denoted by $u^*(t)$. The values of the function u for each edge of the network N are predefined, while the value of the function $u^*(t)$ on the edge $ij \in E$ is denoted by $u^*_{ij}(t)$ and defined as the function

$$u^*_{ij}(t) = u_{ij}(a_{ij} + \lambda_{ij}(1 - a_{ij})(1 - e^{-b_{ij}t})).$$

In this expression, u_{ij} is the capacity of an edge ij in the original network, a_{ij} is the percentage of capacity u_{ij} which remains operative in a disruptive event (hence, $1 - a_{ij}$ is the percentage of capacity u_{ij} lost in a disruptive event). Further, λ_{ij} represents the percentage of the lost capacity $(1 - a_{ij})u_{ij}$ which can be recovered by repair. Finally, b_{ij} denotes the speed of recovery of the edge ij. The values of a_{ij}, b_{ij} and λ_{ij} are predefined. Note that for $t = 0$ we have $u^*_{ij}(t) = u_{ij}a_{ij}$, therefore the given function presupposes that $t = 0$ is the time of the disruptive event, so $u^*_{ij}(t)$ denotes the capacity of the edge ij in the recovery time t.

Terms $\psi(N)$ and $\psi(N^*(t))$ denote the value of the maximum flow in networks N and $N^*(t)$ respectively. The resilience function of the network N is now defined as

$$R(N, t) = \frac{\psi(N^*(t)) - \psi(N^*(0))}{\psi(N) - \psi(N^*(0))}.$$

After the concept of resilience in a network is introduced, the problem of network design is postulated as follows. Network N is a mathematical model of the infrastructure system an engineer has to construct, where connections between some nodes are technically feasible, while some other connections are technically impossible to construct. Therefore, the edges in the network N represent the connections of the infrastructure system which are technically feasible. Since not all possible connections will be constructed, the engineer has to choose which connections to construct in order to obtain the optimal infrastructure system with respect to the capacity, the cost and the resilience. Mathematically, that means we have to find the subset of edges in N which induces the optimal subnetwork of N. For a given network N, one such subset of edges is defined by the edge inclusion function $\delta : E \to \{0, 1\}$ where

$$\delta(ij) = \begin{cases} 0, & \text{if the edge } ij \text{ is not included in the constructed network,} \\ 1, & \text{if the edge } ij \text{ is included in the constructed network.} \end{cases}$$

We commonly use the abbreviated notation δ_{ij} instead of $\delta(ij)$. Consequently, the subnetwork of N, which is induced by the set of edges $E_\delta = \{ij \in E : \delta_{ij} = 1\}$, is denoted by N_δ.

Finally, the cost of the edge construction in the network N is defined as the non-negative function $c : E \to \mathbb{R}^+$, where the cost $c(ij)$ of the edge $ij \in E$ is often denoted by abbreviation c_{ij}. The authors propose a network design model in which the subnetwork of the minimum cost is sought for, such that it has a minimum resilience required. In other words, they propose a network design model which is defined as the problem of nonlinear optimisation given by

$$\text{minimize } f(N_\delta) := \sum_{ij \in E} \delta_{ij} c_{ij} \text{ subject to } R(N_\delta, t) \geq \kappa.$$

They also propose an improved probabilistic solution discovery algorithm for solving the proposed problem of nonlinear optimization. Finally, they provide two numerical examples on which their algorithm has been tested.

Since in real-life problems it is often possible to connect vertices i and j by the edge ij of various capacities, it would be more realistic if in a proposed model one could choose which capacity u_{ij} of the edge ij to construct given the upper and lower capacity bound $\overline{u}_{ij}$ and $\underline{u}_{ij}$ respectively. In addition, the cost of construction of the edge ij usually depends on the capacity, therefore a model would be more realistic if the cost were given as the function of edge capacity. Finally, although the authors of [13] mentioned that, in the problem of nonlinear optimisation an additional constraint can be added on the lower bound of the required value of the flow, they did not implement it in their example and therefore did not test the algorithm on said problem. The inherent problem of the model without the constraint on the value of the flow, but only with the minimized cost, is that we can obtain an optimal network which is only slightly cheaper than the rest, but the value of the maximum flow in it is significantly lower.

For instance, we may observe the network N in Figure 1 (a), where vertex 1 is the source, while vertex 4 is the sink. The edge labels are the edge capacities and the cost of every edge equals 1. As for the recovery parameters a_{ij}, b_{ij} and λ_{ij}, suppose that $a_{ij} = 0.3$ and $\lambda_{ij} = 0.9$ for every edge ij, while $b_{ij} = 0.7$ for $ij = 13$, 34 and $b_{ij} = 0.9$ for $ij = 12, 14, 24$. In other words these recovery parameters mean that the edges $12, 14$ and 24 recover considerably faster than the edges 13 and 34. Therefore, the both flow directions $1 \to 4$ and $1 \to 2 \to 4$ are substantially more resilient than the flow direction $1 \to 3 \to 4$. Let us consider two subnetworks of N, the first denoted by N_δ and shown in Figure 1 (b), the other is denoted by $N_{\delta'}$ and

shown in Figure 1 (c). If we apply the model from [13], where we minimise the cost given the constraint $R(N_\delta, 6) > 0.8$, we will obtain network N_δ as optimal, since it is the cheapest with the cost equal to 1 and the required resilience $R(N_\delta, 6) = 0.896$. However, the value of the maximum flow in that network equals $\psi(N_\delta) = 1$. On the other hand, if we consider the network $N_{\delta'}$, we note that at slightly greater cost 2, we will obtain a similarly resilient network ($R(N_{\delta'}, 6) = 0.896$), with a significantly greater value of the maximum flow ($\psi(N_{\delta'}) = 10$). Therefore, from the economic perspective, the network $N_{\delta'}$ is more frequently regarded as optimal.

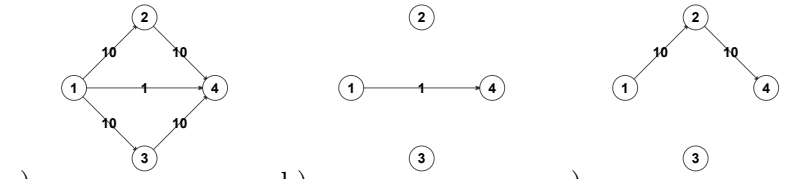


Figure 1: a) The network N, b) the cheapest resilient subnetwork N_δ, b) a slightly more expensive resilient subnetwork $N_{\delta'}$ with a significantly greater value of the maximum flow.

The aim of this paper is to propose three models of resilient network design considering all the mentioned improvements and to propose a simple genetic algorithm for problem-solving. The proposed algorithm is tested on one numerical example for each of the models.

3 Resilient network design models

In order to describe the improved network design models, slightly different notation should be introduced. Let $G = (V, E)$ be a directed graph where vertices from V are denoted by integers $1, \ldots, n$, with vertex 1 being the source and vertex n being the sink. Let $\underline{u} : E \to \mathbb{R}^+$ and $\overline{u} : E \to \mathbb{R}^+$ be two non-negative functions such that $\underline{u}(ij) \leq \overline{u}(ij)$ for every $ij \in E$. The values $\underline{u}(ij)$ and $\overline{u}(ij)$ are often denoted by abbreviations $\underline{u}_{ij}$ and $\overline{u}_{ij}$ respectively. Let us denote by N_u the network $(G, u, 1, n)$ where the edge capacity function u satisfies the condition $\underline{u}_{ij} \leq u_{ij} \leq \overline{u}_{ij}$ for every $ij \in E$. If a disruptive event occurs in the network N_u and diminishes the capacities of edges in N_u, then $N_u^*(t)$ denotes the disrupted network after the recovery time t.

Therefore, networks N_u and $N_u^*(t)$ differ only in the edge capacity function, where the capacity $u_{ij}^*(t)$ of the edge ij in the network $N_u^*(t)$ is defined by

$$u_{ij}^*(t) = u_{ij}(a_{ij} + \lambda_{ij}(1 - a_{ij})(1 - e^{-b_{ij}t}))$$

as in the previous section, and the values of the recovery parameters a_{ij}, b_{ij} and λ_{ij} are predefined.

The value of the maximum flow in networks N_u and $N_u^*(t)$ is denoted by $\psi(N_u)$ and $\psi(N_u^*(t))$ respectively. Now, the resilience $R(N_u, t)$ of the network N_u after the recovery time t is also defined as in the previous section, i.e.

$$R(N_u, t) = \frac{\psi(N_u^*(t)) - \psi(N_u^*(0))}{\psi(N_u) - \psi(N_u^*(0))}.$$

Finally, the cost of the edge construction in the network N_u is defined as the function $c : E \to \mathbb{R}^+$ such that $c = h(u)$, where h can be any function which gives the dependence of the edge cost and the edge capacity. The cost $c(N_u)$ of the network N_u is then defined by

$$c(N_u) = \sum_{ij \in E} c_{ij}.$$

We can now propose three models of resilient network design. In the first model the cost is minimised given the constraint on the value of the flow and the resilience, in the other model the value of the flow is maximised given the constraints on the resilience and the cost, and, finally, in the third model the resilience is maximised given the constraints on the flow value and the cost.

In order to formally describe the design models and define them as the problems of non-linear optimisation, we should primarily describe how the edge capacity function u of the network N_u can be written in the vector form. Let m denote the number of edges in the network N_u. Since vertices in N_u are denoted by integers $1, \ldots, n$, the edges $ij \in E$ can be lexicographically ordered. Therefore, we can define the edge capacity vector $\mathbf{u}$ by

$$\mathbf{u} = (\ldots, u_{ij}, \ldots) \in \mathbb{R}^m$$

where the order of capacities u_{ij} in the $m-$tuple corresponds to the lexicographic order of the edges $ij \in E$. By analogy, we can define the vector $\mathbf{u}^*(t)$ for the capacity function $u^*(t)$ in the network $N_u^*(t)$. Since all networks N_u are based on the same

graph G and differ only in the capacity function u, we can denote $\psi(N_u)$, $R(N_u, t)$ and $c(N_u)$ by $\psi(\mathbf{u})$, $R(\mathbf{u}, t)$ and $c(\mathbf{u})$ and regard them as functions in the variable $\mathbf{u} \in \mathbb{R}^m$.

We will now consider the general non-linear programming problem formulated as

$$\text{minimise } f(\mathbf{x}) \text{ subject to } \mathbf{x} \in \mathcal{S} \cap \mathcal{F},$$

where $f(\mathbf{x})$ is the objective function, $\mathcal{S}$ is the subset of $\mathbb{R}^m$ which consists of $\mathbf{x} = (x_1, \ldots, x_m) \in \mathbb{R}^m$ satisfying the following constraints

$$\underline{x}_i \leq x_i \leq \overline{x}_i$$

and $\mathcal{F}$ is the subset of $\mathbb{R}^m$ consisting of $\mathbf{x} \in \mathbb{R}^m$ satisfying conditions $g_j(\mathbf{x}) \leq 0$ for $j = 1, \ldots, p$, i.e.

$$\mathcal{F} = \{\mathbf{x} \in \mathbb{R}^m : g_1(\mathbf{x}) \leq 0 \wedge \ldots \wedge g_p(\mathbf{x}) \leq 0\}.$$

We will further define $g_j^+(\mathbf{x}) = \max\{0, g_j(\mathbf{x})\}$ for every $j = 1, \ldots, p$, therefore, by using a penalty function approach the constraint violations can be treated as a single penalty function

$$\phi(\mathbf{x}) = \sum_{j=1}^{p} (g_j^+(\mathbf{x}))^2.$$

Now, a simple genetic algorithm can be developed in order to solve this nonlinear programming problem which uses stochastic ranking, as presented in the following section. We have to define an objective function $f(\mathbf{x})$ we minimise and the corresponding penalty function $\phi(\mathbf{x})$ for all three proposed models.

We can now propose three models of resilient network design.

Model 1. In this model, we minimise the cost subject to constraints on the value of flow and the resilience of the network. Therefore, the problem can be written as

$$\begin{array}{ll} \min & c(\mathbf{u}) \\ \text{s.t.} & \psi(\mathbf{u}) \geq \psi_0 \text{ and} \\ & R(\mathbf{u},t) \geq R_0 \end{array}$$

Therefore, in this model the objective function is $f(\mathbf{u}) = c(\mathbf{u})$ subject to constraints $g_1(\mathbf{u}) = \psi_0 - \psi(\mathbf{u}) \leq 0$ and $g_2(\mathbf{u}) = R_0 - R(\mathbf{u},t) \leq 0$. Now the penalty function $\phi(\mathbf{u})$ is defined by

$$\phi(\mathbf{u}) = \sum_{j=1}^{2} (g_j^+(\mathbf{u}))^2$$

where $g_1^+(\mathbf{u}) = \max\{\psi_0 - \psi(\mathbf{u}), 0\}$ and $g_2^+(\mathbf{u}) = \max\{R_0 - R(\mathbf{u},t), 0\}$.

Model 2. In this model, we maximise the value of the flow subject to constraints on the cost and the resilience of the network. Therefore, the problem can be written as

$$\begin{array}{ll} \max & \psi(\mathbf{u}) \\ \text{s.t.} & c(\mathbf{u}) \leq c_0 \text{ and} \\ & R(\mathbf{u},t) \geq R_0 \end{array}$$

In order to apply the same algorithm to all models, we will convert this problem to minimising, i.e. we define the objective function as $f(\mathbf{u}) = -\psi(\mathbf{u})$ and then we minimise it subject to constraints $g_1(\mathbf{u}) = c(\mathbf{u}) - c_0 \leq 0$ and $g_2(\mathbf{u}) = R_0 - R(\mathbf{u},t) \leq 0$. Now the penalty function $\phi(\mathbf{u})$ is defined by

$$\phi(\mathbf{u}) = \sum_{j=1}^{2} (g_j^+(\mathbf{u}))^2$$

where $g_1^+(\mathbf{u}) = \max\{c(\mathbf{u}) - c_0, 0\}$ and $g_2^+(\mathbf{u}) = \max\{R_0 - R(\mathbf{u},t), 0\}$.

Model 3. In this model, we maximise the resilience of the network in the recovery time t subject to constraints on the cost and the value of the flow in the network. Therefore, the problem can be written as

$$\begin{array}{ll} \max & R(\mathbf{u},t) \\ \text{s.t.} & c(\mathbf{u}) \leq c_0 \text{ and} \\ & \psi(\mathbf{u}) \geq \psi_0 \end{array}$$

Again, in order to apply the same algorithm to all models, we will convert this problem to minimising, i.e. we define objective function as $f(\mathbf{u}) = -R(\mathbf{u},t)$ and then minimize it subject to constraints $g_1(\mathbf{u}) = c(\mathbf{u}) - c_0 \leq 0$ and $g_2(\mathbf{u}) = \psi_0 - \psi(\mathbf{u}) \leq 0$. Now the penalty function $\phi(\mathbf{u})$ is defined by

$$\phi(\mathbf{u}) = \sum_{j=1}^{2} (g_j^+(\mathbf{u}))^2$$

where $g_1^+(\mathbf{u}) = \max\{c(\mathbf{u}) - c_0, 0\}$ and $g_2^+(\mathbf{u}) = \max\{\psi_0 - \psi(\mathbf{u}), 0\}$.

After proposing three different models of resilient network design, defined as the problems of non-linear optimisation, we will propose an efficient algorithm for solving said problems. Therefore, in the next section we will describe a simple genetic algorithm for solving these nonlinear optimization problems and then apply it to the numerical example in the following section.

4 Algorithm for solving resilient network design problem

We want to propose a simple genetic algorithm for solving the non-linear optimisation problems to which a resilient network design models reduce. Since all considered networks are based on the same directed graph G and differ only in the edge capacity function u, we will represent the network N_u by its edge capacity vector $\mathbf{u}$ throughout the rest of the paper. In the first step of the algorithm, a population of λ networks $\mathcal{G}^{(1)} = \{\mathbf{u}_{1,1}, \ldots, \mathbf{u}_{\lambda,1}\}$ is generated at random, so that the constraint $\underline{u}_{ij} \leq u_{ij} \leq \overline{u}_{ij}$ is satisfied for every edge ij in each individual $\mathbf{u}_{k,1}$. This population represents the first generation of individuals. In addition, a generation counter g is set to 0, i.e. $g \leftarrow 0$. Until the stopping criterion has been satisfied, the following steps are repeated. Firstly, the generation counter is set to $g+1$, i.e. $g \leftarrow g+1$. The individuals from population $\mathcal{G}^{(g)} = \{\mathbf{u}_{1,g}, \ldots, \mathbf{u}_{\lambda,g}\}$ are ranked using the procedure of stochastic ranking (to be subsequently explained) in order to obtain a new order of those individuals where I_j is the ranking of j-th individual $\mathbf{u}_{j,g}$ from generation g. The μ best ranked individuals $\mathbf{u}_{j,g}$ are denoted by $\mathbf{u}'_{I_j,g}$ and selected to be the set of genitors $\mathcal{P}^{(g)} = \{\mathbf{u}'_{1,g}, \ldots, \mathbf{u}'_{\mu,g}\}$ of the next generation. The next generation consists of all genitors from the previous generation (i.e. $\mathcal{P}^{(g)} \subseteq \mathcal{G}^{(g+1)}$) and all children generated by all parents from $\mathcal{P}^{(g)}$.

In order for $\mathcal{G}^{(g+1)}$ to have λ individuals, besides μ parents from $\mathcal{G}^{(g)}$ which are included in $\mathcal{G}^{(g+1)}$, $\lambda - \mu$ children have to be generated. Therefore, for $k = 1, \ldots, \lambda - \mu$, the child $\mathbf{u}_{k,g+1}$ is generated from the genitor $\mathbf{u}'_{i,g}$ where $i = (k-1) \bmod \mu + 1$. In order to explain how the child $\mathbf{u}_{k,g+1}$ is generated from the genitor $\mathbf{u}'_{i,g}$, let us recall that $\mathbf{u}'_{i,g} \in \mathbb{R}^m$ where m is the number of edges in the network. Each of the m coordinates in $\mathbf{u}'_{i,g}$ represents the capacity of the corresponding edge ij and can be regarded as a gene. Mutation can occur in every gene in the sense that the capacity of that edge can be changed which results in a new 'mutated' network.

If the number of mutations is lower and the size of mutation (i.e. the amount of capacity change in a gene which represents a certain edge ij) is smaller, the mutated child $\mathbf{u}_{k,g+1}$ is more similar to its parent $\mathbf{u}'_{i,g}$. Therefore, the child $\mathbf{u}_{k,g+1}$ of the genitor $\mathbf{u}'_{i,g}$ is created so that a random number n_{mut} of mutations is generated with half-normal distribution along the integers from $[1, 2n/3]$. The n_{mut} coordinates in $\mathbf{u}'_{i,g}$ where a mutation occurs are selected at random. Now, if a mutation is to occur in the coordinate corresponding to the edge ij in $\mathbf{u}'_{i,g}$, then the size δ_{ij} of the mutation is selected as a random number with normal distribution along the interval $[-(\overline{u}_{ij} - \underline{u}_{ij}), \overline{u}_{ij} - \underline{u}_{ij}]$. Assuming that u'_{ij} denotes the value of the ij coordinate in $\mathbf{u}'_{i,g}$, then if $u'_{ij} + \delta_{ij}$ doesn't satisfy the constraint $\underline{u}_{ij} \leq u'_{ij} + \delta_{ij} \leq \overline{u}_{ij}$, the process is repeated until it does. Furthermore, the genitors from $\mathcal{P}^{(g)}$ are included in $\mathcal{G}^{(g+1)}$ by setting $\mathbf{u}_{k,g+1} = \mathbf{u}'_{k-(\lambda-\mu),g}$ for $k = \lambda-\mu+1, \ldots, \lambda$. The entire process is repeated with $\mathcal{G}^{(g+1)}$ and so on until the stopping criterion has been satisfied. The pseudocode of this algorithm is shown in Figure 2, while the process of offspring generation is given as the separate procedure whose pseudocode is shown in Figure 3.

Algorithm 1 Simple genetic algorithm for resilient network design.

```
generate G^(1) = {u_{1,1}, ..., u_{λ,1}} at random
g ← 0
while (stopping criterion not satisfied) do
    g ← g + 1
    order G^(g) = {u_{1,g}, ..., u_{λ,g}} using stochastic ranking
    select P^(g) = {u'_{1,g}, ..., u'_{μ,g}} as μ best ranked individuals from G^(g)
    generate G^(g+1) from P^(g) using the procedure of offspring generation
end while
return P^(g)
```

Figure 2: A simple genetic algorithm for resilient network design.

In each generation of λ individuals, only μ best ranked individuals are selected to become parents of the next generation. The ranking of the λ individuals is done by the stochastic ranking procedure [11] which is a bubble-sort-like procedure where a

Algorithm 2 The procedure of offspring generation

procedure OffspringGen($\mathcal{P}^{(g)} = \{\mathbf{u}'_{1,g}, \ldots, \mathbf{u}'_{\mu,g}\}$)
  for $k = 1, \ldots, \lambda - \mu$ **do**
    $i \leftarrow (k-1) \bmod \mu + 1$
    $n_{mut} \leftarrow \left\lceil \frac{2m}{9} \cdot |N(0,1)| \right\rceil$ (repeat until $n_{mut} \leq \frac{2m}{3}$)
    $\delta \leftarrow \mathbf{U}(0,1)$ such that $|\delta| = n_{mut}$
    for $ij \in E$ **do**
      if $\delta_{ij} = 1$ **then**
        $\delta_{ij} \leftarrow \delta_{ij} \cdot (\overline{u}_{ij} - \underline{u}_{ij}) \cdot N(0,1)/3$ (repeat until $\underline{u}_{ij} \leq u'_{ij} + \delta_{ij} \leq \overline{u}_{ij}$)
      end if
    end for
    $\mathbf{u}_{k,g+1} \leftarrow \mathbf{u}'_{i,g} + \delta$
  end for
  for $k = \lambda - \mu + 1, \ldots, \lambda$ **do**
    $\mathbf{u}_{k,g+1} \leftarrow \mathbf{u}'_{k-(\lambda-\mu),g}$
  end for
  $\mathcal{G}^{(g+1)} \leftarrow \{\mathbf{u}_{1,g+1}, \ldots, \mathbf{u}_{\lambda,g+1}\}$
  return $\mathcal{G}^{(g+1)}$
end procedure

Figure 3: The procedure of offspring generation, $N(0,1)$ is a random number generated with normal distribution, while $\mathbf{U}(0,1)$ is randomly generated vector from $\{0,1\}^m$ with uniform distribution.

probability P_f of using only the objective function for comparing individuals in the infeasible region of the search space is introduced. Namely, when two individuals are compared in order to determine which one is more suitable and should therefore be better ranked, if both individuals are feasible (the penalty function ϕ equals zero for both individuals), then the probability to rank those two individuals according to the value of the objective function is 1, otherwise that probability is P_f. The procedure provides a convenient way of balancing the dominance in a ranked set. The pseudocode of the procedure of stochastic ranking is shown in Figure 4. Since we, eventually, want to obtain only feasible solutions, the probability P_f should be set to be less than 0.5 so that there is a pressure against infeasible solutions.

The value of P_f is usually set to be 0.45 as the previous studies have shown that $P_f = 0.45$ is often sufficient.

Algorithm 3 Stochastic ranking

```
I_j = j  ∀j ∈ {1, ..., λ}
for i = 1, ..., λ do
  for j = 1, ..., λ − 1 do
    sample u ∈ U(0, 1)
    if (φ(x_{I_j}) = φ(x_{I_{j+1}}) = 0) or (u < P_f) then
      if (f(x_{I_j}) > f(x_{I_{j+1}})) then
        swap(I_j, I_{j+1})
      end if
    else
      if (φ(x_{I_j}) > φ(x_{I_{j+1}})) then
        swap(I_j, I_{j+1})
      end if
    end if
  end for
end for
```

Figure 4: Stochastic ranking procedure, $P_f = 0.45$.

5 Numerical example

Finally, we applied the proposed resilient network design models on graph G and used the proposed genetic algorithms to solve those models. Firstly, we define the directed graph G to which we will apply our models. Let $G = (V, E)$ be a graph with 6 vertices and 9 directed edges as shown in Figure 5.

Vertex 1 is the source and vertex 6 is the sink. In addition, let

$$\overline{\mathbf{u}} = (4, 3, 2, 2, 2, 3, 1, 3, 4)$$

be the vector whose coordinates $\overline{u}_{ij}$ are the upper bounds on the edge capacities u_{ij}. The coordinates $\overline{u}_{ij}$ in the vector $\overline{\mathbf{u}}$ are ordered according to the lexicographic order of the edges, i.e. $\overline{\mathbf{u}} = (\overline{u}_{12}, \overline{u}_{13}, \overline{u}_{23}, \overline{u}_{24}, \overline{u}_{34}, \overline{u}_{35}, \overline{u}_{45}, \overline{u}_{46}, \overline{u}_{56})$. The same edge order will be used throughout this section. The lower bounds on edge capacities

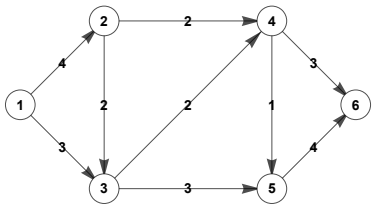


Figure 5: The directed graph G which to which the proposed network design models will be applied.

ij	12	13	23	24	34	35	45	46	56
a_{ij}	0.37	0.21	0.47	0.28	0.43	0.25	0.33	0.44	0.31
b_{ij}	0.92	0.78	0.88	0.81	0.87	0.81	0.79	0.91	0.82
λ_{ij}	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9

Table 1: Recovery parameters a_{ij}, b_{ij} and λ_{ij} of the edge ij.

are defined by $\underline{u}_{ij} = 0$ for every $ij \in E$, i.e. $\underline{\mathbf{u}} = \mathbf{0}$. In order to verify the optimal solution obtained by the algorithm, we want to be able to calculate the optimum exactly, so will allow only integer capacity values between $\underline{u}_{ij}$ and $\overline{u}_{ij}$. Therefore, this example has 86 400 different possible networks in total. The cost of an individual edge of capacity u_{ij} is defined as $c(u_{ij}) = 0.3 + u_{ij}$, while the cost of the network N_u represented by the vector $\mathbf{u}$ is consequently defined as

$$c(\mathbf{u}) = \sum_{ij \in E, u_{ij} \neq 0} c(u_{ij}).$$

Finally, the recovery parameters a_{ij}, b_{ij} and λ_{ij} of the edge ij are presented in the Table 1. We want to design an optimal network according to each of the proposed three resilient network design models by using the proposed genetic algorithm.

Model 1. Let us recall that in this model the cost is minimised subject to constraints on the value of flow and the resilience of the network. Therefore, the problem can be written as

$$\begin{array}{ll} \min & c(\mathbf{u}) \\ \text{s.t.} & \psi(\mathbf{u}) \geq \psi_0 \text{ and} \\ & R(\mathbf{u},t) \geq R_0 \end{array}$$

t	n_{gen}	$\mathbf{u}_{opt}$	$\psi(\mathbf{u}_{opt})$	$c(\mathbf{u}_{opt})$	$R(\mathbf{u}_{opt}, t)$
1	33	$(2,3,0,2,2,1,0,3,1)$	4	16.1	0.590
2	51	$(2,3,0,2,2,2,0,3,1)$	4	17.1	0.786
3	49	$(2,3,0,2,1,2,0,2,2)$	4	16.1	0.841
4	31	$(2,3,0,2,1,2,0,2,2)$	4	16.1	0.879
5	44	$(2,3,0,2,2,1,0,3,1)$	4	16.1	0.904
6	50	$(2,3,0,2,1,2,0,2,2)$	4	16.1	0.902

Table 2: The computational results of Model 1 by using the proposed genetic algorithm.

We set that $\psi_0 = 4$ and R_0 equals 0.58, 0.779, 0.838, 0.878, 0.892, 0.9 for $t = 1, \ldots, 6$ respectively. By using these set values, we apply the proposed genetic algorithm to the problem of non-linear optimisation. When applying the genetic algorithm, the concrete values of the algorithm parameters, namely the size λ of each generation, the number μ of genitors in every generation and the stopping criterion, must be selected. In all these models, the same values and stopping criterion are used. Firstly, the number of individuals in each generation is set to $\lambda = 400$, the number of the most suitable individuals to become genitors of the next generation is set to $\mu = 50$, the stopping criterion is set in a way that the number of generations is at most $g_{\max} = 200$ or that the set of different genitors in $\mathcal{P}^{(g)}$ is the same in 10 consecutive generations. The results obtained by using the proposed genetic algorithm are shown in Table 2, where n_{gen} denotes the number of generations produced by the algorithm before the stopping criterion was satisfied. Since $n_{gen} \leq 200 = g_{\max}$ for every t, it follows that for each t the execution of the algorithm stopped because the population of genitors was stable throughout the last 10 generation. By $\mathbf{u}_{opt}$ the best ranked individual from the last generation generated by the algorithm is denoted. The last three columns of Table 2 show the value of the flow, the cost and the resilience of the network represented by $\mathbf{u}_{opt}$. The comparison of the obtained $\mathbf{u}_{opt}$ with the results of exact calculations confirms that the proposed algorithm indeed resulted in the optimum solution for every $t = 1, \ldots, 6$.

Model 2. Recall that in this model the value of the flow is maximised subject to constraints on the cost and the resilience of the network. Therefore, the problem

t	n_{gen}	$\mathbf{u}_{opt}$	$\psi(\mathbf{u}_{opt})$	$c(\mathbf{u}_{opt})$	$R(\mathbf{u}_{opt}, t)$
1	15	$(1,3,0,2,2,0,1,3,0)$	3	13.8	0.584
2	19	$(2,2,0,2,2,0,0,3,0)$	3	12.5	0.796
3	20	$(1,3,0,1,2,1,0,3,0)$	3	12.8	0.848
4	19	$(1,3,0,1,2,0,1,2,1)$	3	13.1	0.884
5	18	$(1,3,0,1,2,0,0,3,2)$	3	13.8	0.900
6	19	$(1,3,1,1,2,0,0,3,0)$	3	12.8	0.902

Table 3: The computational results of Model 2 by using the proposed genetic algorithm.

can be written as

$$\begin{array}{ll} \max & \psi(\mathbf{u}) \\ \text{s.t.} & c(\mathbf{u}) \leq c_0 \text{ and} \\ & R(\mathbf{u},t) \geq R_0 \end{array}$$

We set that $c_0 = 14$ and R_0 equals 0.58, 0.779, 0.838, 0.878, 0.892, 0.9 for $t = 1, \ldots, 6$ respectively. Setting the same $\lambda = 400$ and $\mu = 50$ and by using the same stopping criterion as in the previous model, the proposed genetic algorithm yields the results presented in Table 3. The comparison of the obtained $\mathbf{u}_{opt}$ with the results of exact calculations confirms that the proposed algorithm indeed resulted in the optimum solution for every $t = 1, \ldots, 6$.

Model 3. Recall that in this model the resilience of the network in the time t is maximised subject to constraints on the cost and the value of the flow in the network. Therefore, the problem can be written as

$$\begin{array}{ll} \max & R(\mathbf{u},t) \\ \text{s.t.} & c(\mathbf{u}) \leq c_0 \text{ and} \\ & \psi(\mathbf{u}) \geq \psi_0 \end{array}$$

We set that $c_0 = 18$ and $\psi_0 = 4$. Setting the same $\lambda = 400$ and $\mu = 50$ and using the same stopping criterion as in previous models, the proposed genetic algorithm yields the results presented in Table 4. The comparison of the obtained $\mathbf{u}_{opt}$ with the results of exact calculations confirms that the proposed algorithm indeed resulted in the optimum solution for every $t = 1, \ldots, 6$.

t	n_{gen}	$\mathbf{u}_{opt}$	$\psi(\mathbf{u}_{opt})$	$c(\mathbf{u}_{opt})$	$R(\mathbf{u}_{opt}, t)$
1	21	$(2, 3, 0, 2, 2, 1, 1, 3, 1)$	4	17.4	0.601
2	18	$(2, 3, 0, 2, 2, 1, 1, 3, 1)$	4	17.4	0.786
3	23	$(2, 3, 0, 2, 2, 1, 1, 3, 1)$	4	17.4	0.862
4	17	$(2, 3, 0, 2, 2, 1, 1, 3, 1)$	4	17.4	0.893
5	23	$(2, 3, 0, 2, 2, 1, 1, 3, 1)$	4	17.4	0.906
6	19	$(2, 3, 0, 2, 2, 1, 1, 3, 1)$	4	17.4	0.912

Table 4: The computational results of Model 3 by using the proposed genetic algorithm.

6 Acknowledgements

This research is partially supported through project KK.01.1.1.02.0027, a project co-financed by the Croatian Government and the European Union through the European Regional Development Fund - the Competitiveness and Cohesion Operational Programme.

References

[1] A. Asadabadi, E. Miller-Hooks, Optimal transportation and shoreline infrastructure investment planning under a stochastic climate future, Transp. Res. Part B 100, 156–74, 2017.

[2] K. Barker, J. E. Ramirez-Marquez, C. M. Rocco, Resilience-based network component importance measures, Reliab. Eng. Syst. Saf. 117, 89–97, 2013.

[3] H. Baroud, K. Barker, J. E. Ramirez-Marquez, C. M. Rocco, Inherent costs and interdependent impacts of infrastructure network resilience, Risk. Anal. 35(4), 642–662, 2015.

[4] M. Chertoff, National infrastructure protection plan, Department of Homeland Security (DHS), Washington, DC 2009.

[5] M. Christopher, H. Peck, Building the resilient supply chain, Int. J. Logist. Manage. 15(2), 1–14, 2004.

[6] Y. Fang, N. Pedroni, E. Zio, Optimization of cascade-resilient electrical infrastructures and its validation by power flow modeling, Risk. Anal. 35(4), 594–607, 2015.

[7] R. Faturechi, E. Miller-Hooks, Travel time resilience of roadway networks under disaster, Transp. Res. Part B 70, 47–64, 2014.

[8] C. S. Holling, Resilience and stability of ecological systems, Annu. Rev. Ecol. Syst. 1–23, 1973.

[9] S. Hosseini, K. Barker, J. E. Ramirez-Marquez, A review of definitions and measures of system resilience, Reliab. Eng. Syst. Saf. 145, 47–61, 2016.

[10] J. A. Nachlas, Reliability engineering: probabilistic models and maintenance methods, CRC Press, 2005.

[11] T. P. Runarsson, X. Yao, Stochastic ranking for constrained evolutionary optimization, IEEE Transactions on Evolutionary Computation 4(3), 284–294, 2000.

[12] N. Yodo, P. Wang, Resilience allocation for early stage design of complex engineered systems, J. Mech. Des. 138(9), 091402, 2016.

[13] X. Zhang, S. Mahadevan, S. Sankararaman, K. Goebel, Resilience-based network design under uncertainty, Reliab. Eng. Syst. Saf. 169, 364–79, 2018.

Verification and Strengthening of the Atiyah–Sutcliffe Conjectures for Several Types of Almost Collinear Configurations in Euclidean and Hyperbolic Plane

Dragutin Svrtan
Department of Mathematics, University of Zagreb,
Bijenička cesta 30, 10000 Zagreb, Croatia,
dsvrtan@math.hr

Igor Urbiha
Department of Informatics, Polytechnic of Zagreb, University of Zagreb,
Konavoska 2, 10000 Zagreb, Croatia,
urbiha@vtszg.hr

Abstract

In 2001 Sir M. F. Atiyah formulated a conjecture C1 and later with P. Sutcliffe two stronger conjectures C2 and C3. These conjectures, inspired by physics (spin-statistics theorem of quantum mechanics), are geometrically defined for any configuration of points in the Euclidean three space. The conjecture C1 is proved for $n = 3, 4$ and for general n only for some special configurations (M. F. Atiyah, M. Eastwood and P. Norbury, D.Đoković). In [11] and [12] we have verified the conjectures C2 and C3 for parallelograms, cyclic quadrilaterals and some infinite families of tetrahedra, and have proposed a strengthening of conjecture C3 for configurations of four points (Four Points Conjectures). All three Euclidean four-point conjectures have been proved in [14] (2010.) (see also [15] for more information) then independently four years later in [10].

For almost collinear configurations of type A (with all but one point on a line) we propose in [12] several new conjectures (some for symmetric functions) which imply C2 and C3. By using computations with multi-Schur functions we can do verifications up to $n = 9$ of our conjectures. For a type B (and some special dihedral configurations) we verify a stronger conjecture of Đoković which implies C2 for his nonplanar configurations with dihedral symmetry. Recently

DOI: https://doi.org/10.5592/CO/CCD.2018.10

we have observed that our generalizations from [12] are capable to imply hyperbolic C2 for type A and B configurations and we make these connections clear in Sections 1.2 and 1.3.

Finally we mention that by minimizing a geometrically defined energy, figuring in these conjectures, one gets a connection to some complicated physical theories, such as Skyrmions and Fullerenes.

Keywords: Atiyah–Sutcliffe conjecture, almost collinear configuration

MSC: 74H05, 11B37, 26A18, 05A15, 11Y55, 11Y65

1 Almost collinear configurations. Đoković's approach

1.1 Type A configurations

By a type A configurations of N points $x_1, \ldots, x_N$ we shall mean the case when $N-1$ of the points $x_1, \ldots, x_N$ are collinear. Set $n = N-1$. In ([6]) Đoković has proved, for configurations of type A, both the Atiyah conjecture (Theorem 2.1) and the first Atiyah–Sutcliffe conjecture (Theorem 3.1). By using Cartesian coordinates, with $x_i = (a_i, 0)$, $a_1 < a_2 < \cdots < a_n$ and $x_N = x_{n+1} = (0, b)$ (with $b = 1$), the normalized Atiyah matrix $M_{n+1} = M_{n+1}(\lambda_1, \ldots, \lambda_n)$ (denoted by P in [6] when $b = -1$) is given by

$$M_{n+1} = \begin{bmatrix} 1 & \lambda_1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \lambda_2 & \cdots & 0 & 0 \\ 0 & 0 & 1 & & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & & & 1 & \lambda_n \\ (-1)^n e_n & (-1)^{n-1} e_{n-1} & \cdots & \cdots & -e_1 & 1 \end{bmatrix}$$

where $\lambda_1 = a_1 + \sqrt{a_1^2 + b^2} < \lambda_2 = a_2 + \sqrt{a_2^2 + b^2} < \cdots < \lambda_n = a_n + \sqrt{a_n^2 + b^2}$ (with $b = 1$) are positive real numbers and where $e_k = e_k(\lambda_1, \ldots, \lambda_n)$, $1 \le k \le n$, is the k–th elementary symmetric function of $\lambda_1, \lambda_2, \ldots, \lambda_n$. Its determinant satisfies the inequality

$$\begin{aligned} D_n &= 1 + \lambda_n e_1 + \lambda_n \lambda_{n-1} e_2 + \cdots + \lambda_n \lambda_{n-1} \cdots \lambda_1 e_n \\ &\ge 1 + e_1(\lambda_1^2, \ldots, \lambda_n^2) + e_2(\lambda_1^2, \ldots, \lambda_n^2) + \cdots + e_n(\lambda_1^2, \ldots, \lambda_n^2) \\ &= \textstyle\prod_{i=1}^n (1 + \lambda_i^2) \end{aligned}$$

equivalent to the first Atiyah–Sutcliffe conjecture ([4],Conjecture 2). The second Atiyah–Sutcliffe conjecture ([4],Conjecture 3) for configurations of type A is equivalent to the following inequality

$$[D_{n+1}(\lambda_1,\dots,\lambda_n)]^{n-1} \geq \prod_{k=1}^{n} D_n(\lambda_1,\dots,\lambda_{k-1},\lambda_{k+1},\dots,\lambda_n) \tag{1.1}$$

For $n=2$ this inequality takes the form

$$1+\lambda_2 e_1(\lambda_1,\lambda_2)+\lambda_1\lambda_2 e_2(\lambda_1,\lambda_2) \geq (1+\lambda_2 e_1(\lambda_2))(1+\lambda_1 e_1(\lambda_1)$$

i.e.

$$1+\lambda_2 e_1(\lambda_1,\lambda_2)+\lambda_1\lambda_2 e_2(\lambda_1,\lambda_2) \geq (1+\lambda_2^2)(1+\lambda_1^2). \tag{1.2}$$

This reduces to $(\lambda_2-\lambda_1)\lambda_1 \geq 0$, so it is true.

1.2 Type A configurations - hyperbolic case

Let $H^2=\{(x,y)\in\mathbb{R}^2 | y>0\}$ be the upper half plane model of hyperbolic plane. A type A hyperbolic configuration denoted by $A_{n,1}=A(\mathbf{x}_1,\mathbf{x}_2,\dots,\mathbf{x}_n;\mathbf{x}_{n+1})$ consists of $N=n+1$ points in H^2 where n points are collinear, $(\mathbf{x}_1,\mathbf{x}_2,\dots,\mathbf{x}_n)$, $\mathbf{x}_i=(0,b_i)$, $b_1>b_2>\cdots>b_n>0$ and $\mathbf{x}_{n+1}=(a,b)$.
Let us abbreviate the directions t_{ij} from $\mathbf{x}_i$ to $\mathbf{x}_j$ (viewed as points on the absolute $\mathbb{R}\cup\{\infty\}$)

$$\begin{array}{ll} t_{i,n+1}=X_i, 1\leq i\leq n & t_{ij}=0 \text{ for } 1\leq i<j\leq n\ , \\ t_{n+1,i}=-\xi_i, 1\leq i\leq n & t_{ji}=\infty \text{ for } 1\leq i<j\leq n\ . \end{array}$$

Then we have $t_{i,n+1}\cdot t_{n+1,i}=-b_i^2$ i.e. $\xi_i X_i=b_i^2$.
The Atiyah polynomials associated to points $\mathbf{x}_i$ $(1\leq i\leq n+1)$ defined by

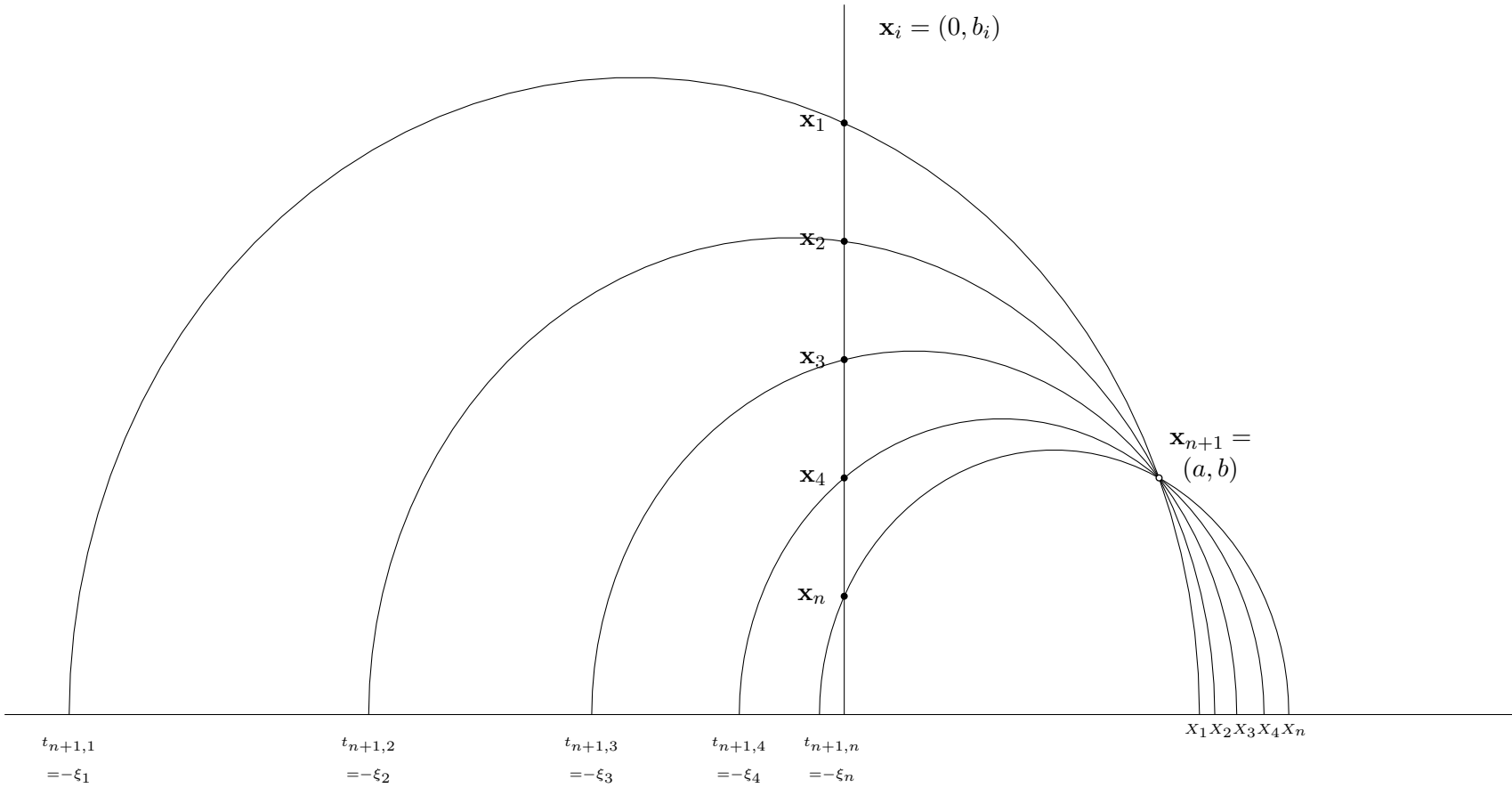
$$p_i=\prod_{j\neq i}(z-t_{ij})$$

(with $z-t_{ij}$ interpreted as 1 if $t_{ij}=\infty$) are given explicitly as follows

$$\begin{aligned} p_1 &= z^{n-1}(z-X_1)\\ p_2 &= z^{n-2}(z-X_2)\\ &\vdots\\ p_{n-1} &= z(z-X_{n-1})\\ p_n &= z-X_n\\ p_{n+1} &= (z+\xi_1)(z+\xi_2)\cdots(z+\xi_n)=z^n+e_1z^{n-1}+\cdots+e_n\ , \end{aligned}$$

where $e_1, \ldots, e_n$ denotes the elementary symmetric functions of $\xi_1, \ldots, \xi_n$.
The Atiyah determinant (the determinant of the matrix of coefficients of p_i's) $D^{\mathrm{hyp}}_{n,1} = D^{\mathrm{hyp}}_{n,1}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n; \mathbf{x}_{n+1})$ is then equal to

$$D^{\mathrm{hyp}}_{n,1} = \begin{vmatrix} 1 & -X_1 & & & & \\ 0 & 1 & -X_2 & & & \\ \vdots & \vdots & \vdots & \ddots & & \\ & & & & 1 & -X_n \\ 1 & e_1 & e_2 & \ldots & e_{n-1} & e_n \end{vmatrix} = X_1 \cdots X_n + X_2 \cdots X_n e_1 + \cdots + e_n = \sum_{i=0}^{n} e_i X_{i+1} \cdots X_n \ .$$



$t_{i,j} = t_{ij}$ = direction from $\mathbf{x}_i$ to $\mathbf{x}_j$; $0 < a < X_1 < X_2 < \cdots < X_n$

Note that

$$X_1 \cdots X_n \cdot D^{\mathrm{hyp}}_{n,1} \Big|_{X_i \to X_i^{-1}}$$

is a specialization of the polynomial

$$\Psi^{12\ldots n}_{12\ldots n} \in Q[X_1, \ldots, X_n, \xi_1, \ldots, \xi_n]$$

defined in the section 1.4 which is symmetric in $\xi_1, \ldots, \xi_n$. The properties of polynomials Ψ^I_J will enable us to study simultaneously both the euclidean and the hyperbolic configurations of type A (and B, defined later). Since

$$\begin{vmatrix} 1 & -t_{j,n+1} \\ 1 & -t_{n+1,j} \end{vmatrix} = t_{j,n+1} - t_{n+1,j} = X_j + \xi_j, 1 \leq j \leq n,$$

we obtain for the normalized Atiyah determinant the following formula

$$D_{n,1} = \frac{\sum_{i=0}^{n} e_i X_{i+1} \cdots X_n}{\prod_{j=1}^{n} (X_j + \xi_j)} \quad (\text{with } \xi_i X_i = b_i^2) \ .$$

Lemma 1.1 *For $0 < X_1 \leq X_2 \leq \ldots \leq X_n$ and $\xi_1, \ldots, \xi_n$ commuting indeterminates, the inequality*

$$\prod_{j=1}^{n} (X_j + \xi_j) \leq \sum_{i=0}^{n} e_i(\xi_1, \ldots, \xi_n) X_{i+1} \cdots X_n$$

holds true coefficientwise.

Proof .

$$\begin{aligned} \prod_{j=1}^{n} (X_j + \xi_j) = \sum_{S \subseteq [n]} \prod_{j \in S} \xi_j \cdot \prod_{k \in S^C} X_k \ &\leq \ \sum_{S \subseteq [n]} \prod_{j \in S} \xi_j \cdot X_{|S|+1} \cdots X_n \\ &= \ \sum_{i=0}^{n} e_i \cdot X_{i+1} \cdots X_n \end{aligned}$$

■

Corollary 1.2 *For any n points on a line and one point outside it in a hyperbolic plane (or space) we have that the second Atiyah–Sutcliffe conjecture holds true:*

$$D^{hyp}_{n,1}(\mathbf{x}_1,\dots,\mathbf{x}_n,\mathbf{x}_{n+1}) \geq \prod_{1\leq i<j\leq n+1} D_2(\mathbf{x}_i,\mathbf{x}_j) = \prod_{j=1}^{n} D_2(\mathbf{x}_j,\mathbf{x}_{n+1}) \ .$$

Proof .
We have that l.h.s. is equal to $\sum e_i X_{i+1}\cdots X_n$ and the r.h.s. reduces to the product $\prod_{j=1}^{n}(X_j+\xi_j)$. ∎

1.3 Type B configurations - hyperbolic case

Let $\mathbf{x}_i = (0,b_i)$, $b_1 > b_2 > \cdots > b_r > c > b_{r+1} > \cdots > b_n > 0$, $(c=\sqrt{a^2+b^2})$, $\mathbf{x}_{n+1}=(a,b)$ and $\mathbf{x}_{n+2}=(-a,b)$ be a type B hyperbolic configuration of $N=n+2$ points in H^2 denoted by $B_{n,2}=B(\mathbf{x}_1,\mathbf{x}_2,\dots,\mathbf{x}_n;\mathbf{x}_{n+1},\mathbf{x}_{n+2})$.
The hyperbolic directions from x_i to x_j we abbreviate as follows

$$\begin{array}{ll} t_{i,n+1}=X_i, 1\leq i\leq n & t_{ij}=0 \text{ for } 1\leq i<j\leq n \ , \\ t_{n+1,i}=-\xi_i, 1\leq i\leq n & t_{ji}=\infty \text{ for } 1\leq i<j\leq n \ , \\ t_{i,n+2}=-X_i, 1\leq i\leq n & t_{n+1,n+2}=-c, \\ t_{n+2,i}=\xi_i, 1\leq i\leq n & t_{n+2,n+1}=c. \end{array}$$

The Atiyah polynomials associated to points $\mathbf{x}_i$ $(1\leq i\leq n+2)$ defined (in [1] - [4]) by

$$p_i=\prod_{j\neq i}(z-t_{ij})$$

(with $z-t_{ij}$ interpreted as 1 if $t_{ij}=\infty$) associated to points $\mathbf{x}_i$ $(1\leq i\leq n+2)$ are then

$$\begin{aligned} p_1 &= z^{n-1}(z-X_1)(z+X_1) \\ p_2 &= z^{n-2}(z-X_2)(z+X_2) \\ &\vdots \\ p_{n-1} &= z(z-X_{n-1})(z+X_{n-1}) \\ p_n &= (z-X_n)(z+X_n) \\ p_{n+1} &= (z+\xi_1)(z+\xi_2)\cdots(z+\xi_n)(z+c) = z^{n+1}+e_1z^n+\cdots+e_{n+1} \\ p_{n+2} &= (z-\xi_1)(z-\xi_2)\cdots(z-\xi_n)(z-c) = z^{n+1}-e_1z^n+\cdots+(-1)^{n+1}e_{n+1} \ . \end{aligned}$$

$$(-1)^{n+1}D^{\rm hyp}_{n,2}=(-1)^{n+1}\begin{vmatrix} 1 & 0 & -X_1^2 & 0 & & & 0 \\ 0 & 1 & 0 & -X_2^2 & & & 0 \\ \vdots & & & & \ddots & & \\ 0 & 0 & 0 & & 1 & 0 & -X_n^2 \\ 1 & e_1 & e_2 & e_3 & \cdots & e_n & e_{n+1} \\ 1 & -e_1 & e_2 & -e_3 & \cdots & & (-1)^{n+1}e_{n+1} \end{vmatrix} =$$

We first add $n+1$-st row to the $n+2$-nd row and then subtract a half of the $n+2$-nd row from the $n+1$-st row.

$$=(-1)^{n+1}2\begin{vmatrix} 1 & 0 & -X_1^2 & 0 & & & 0 \\ 0 & 1 & 0 & -X_2^2 & & & 0 \\ \vdots & & & & \ddots & & \\ 0 & 0 & 0 & & 1 & 0 & -X_n^2 \\ 0 & e_1 & 0 & e_3 & \cdots & 0 & e_{n+1} \\ 1 & 0 & e_2 & 0 & \cdots & e_n & 0 \end{vmatrix} (\ n \text{ even }) = \dots$$

Then we add $-1\cdot 1^{\rm st}$ row $-(X_1^2+e_2)\cdot 3^{\rm rd}$ row $+X_3^2(X_1^2+e_2)\cdot 5^{\rm th}$ row $+\cdots$ to the $n+2$-nd row and we also add $-e_1\cdot 2^{\rm nd}$ row $+(-(X_2^2e_1+e_3))\cdot 4^{\rm th}$ row $+\cdots$ to the $n+1$-st row.

$$=2\begin{vmatrix} 1 & 0 & -X_1^2 & 0 & 0 & & & 0 \\ 0 & 1 & 0 & -X_2^2 & 0 & & & 0 \\ 0 & 0 & 1 & 0 & -X_3^2 & & & 0 \\ 0 & 0 & 0 & 1 & 0 & -X_4^2 & & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & -X_5^2 & 0 \\ \vdots & & & & \vdots & & & \vdots \\ 0 & 0 & 0 & X_2^2e_1+e_3 & 0 & e_5 & \cdots & \\ 0 & 0 & X_1^2+e_2 & 0 & e_4 & \cdots & & \end{vmatrix} = \cdots$$

At the end we get a triangular matrix whose determinant is equal to $D^{\rm hyp}_{n,2}=2f_0f_1$ where we have for n even
$f^0_{n,2}=X_1^2X_3^2\cdots X_{n-1}^2+X_3^2\cdots X_{n-1}^2e_2+\cdots+X_{n-1}^2e_{n-2}+e_n$
$f^1_{n,2}=X_2^2X_4^2\cdots X_n^2e_1+X_4^2\cdots X_n^2e_3+\cdots+X_n^2e_{n-1}+e_{n+1}$,
and for n odd
$f^0_{n,2}=X_1^2X_3^2\cdots X_n^2+X_3^2\cdots X_n^2e_2+\cdots+X_n^2e_{n-1}+e_{n+1}$
$f^1_{n,2}=X_2^2X_4^2\cdots X_{n-1}^2e_1+X_4^2\cdots X_{n-1}^2e_3+\cdots+X_{n-1}^2e_{n-2}+e_n$.

For example

$$
\begin{aligned}
D^{\mathrm{hyp}}_{1,2} &= 2(X_1^2 + e_2)e_1 \\
D^{\mathrm{hyp}}_{2,2} &= 2(X_1^2 + e_2)(X_2^2 e_1 + e_3) \\
D^{\mathrm{hyp}}_{3,2} &= 2(X_1^2 X_3^2 + X_3^2 e_2 + e_4)(X_2^2 e_1 + e_3) \\
D^{\mathrm{hyp}}_{4,2} &= 2(X_1^2 X_3^2 + X_3^2 e_2 + e_4)(X_2^2 X_4^2 e_1 + X_4^2 e_3 + e_5)
\end{aligned}
$$

Now we shall verify Atiyah-Sutcliffe Conjecture 2 for the hyperbolic configuration $\mathrm{B}_{n,2}$ of the type B which reads as follows

$$
D^{\mathrm{hyp}}_{n,2} \geq (X_1 + \xi_1)^2 (X_2 + \xi_2)^2 \cdots (X_n + \xi_n)^2 2c \qquad (\star)
$$

Case 1. **(n even)** We first rewrite $f^0_{n,2}$ and $f^1_{n,2}$ in terms of c and elementary symmetric functions e'_k $(1 \leq k \leq n)$ of $\xi_1, \ldots \xi_n$:

$$
\begin{aligned}
f^0_{n,2} &= X_1^2 X_3^2 \cdots X_{n-1}^2 + X_3^2 \cdots X_{n-1}^2 e'_2 + \cdots + X_{n-1}^2 e'_{n-2} + e'_n \\
&\qquad + X_3^2 \cdots X_{n-1}^2 e'_1 c + \cdots + X_{n-1}^2 e'_{n-3} c + e'_{n-1} c \\
f^1_{n,2} &= X_2^2 X_4^2 \cdots X_n^2 e'_1 + X_4^2 \cdots X_n^2 e'_3 + \cdots + X_n^2 e'_{n-1} + e'_{n+1} \\
&\qquad X_2^2 X_4^2 \cdots X_n^2 c + X_4^2 \cdots X_n^2 e'_2 c + \cdots + X_n^2 e_{n-2} c + e'_n c
\end{aligned}
$$

Now by Cauchy-Schwartz

$$
\begin{aligned}
&D^{\mathrm{hyp}}_{n,2} = 2 f^0_{n,2} f^1_{n,2} \geq \\
&\geq 2(X_1 X_3 \cdots X_{n-1} \cdot X_2 X_4 \cdots X_n \sqrt{c} + X_3 X_5 \cdots X_{n-1} \cdot X_2 X_4 \cdots X_n e'_1 \sqrt{c} + \cdots + e'_n \sqrt{c})^2 = \\
&= (X_1 X_2 \cdots X_n + X_2 X_3 \cdots X_n e'_1 + \cdots + X_n e'_{n-1} + e'_n)^2 \cdot 2c = \\
&= (D^{\mathrm{hyp}}_{n,1})^2 \cdot 2c
\end{aligned}
$$

Now by Corollary 1.2 for $D^{\mathrm{hyp}}_{n,1}$ the inequality $(\star)$ follows.

Case 2. **(n odd)** Is similar to Case 1.

This concludes the verification of the Atiyah-Sutcliffe Conjecture C2 for hyperbolic type B configurations.

Now we state the strongest Atiyah-Sutcliffe conjecture C3 for type A and type B hyperbolic configurations.

Conjecture 3.

(i) $D^{\mathrm{hyp}}_{n,1}(\mathbf{x}_1,\dots,\mathbf{x}_n;\mathbf{x}_{n+1})^{n-1} \geq \prod_{k=1}^{n} D^{\mathrm{hyp}}_{n-1,1}(\mathbf{x}_1,\dots\widehat{\mathbf{x}}_k,\dots,\mathbf{x}_n;\mathbf{x}_{n+1})$
(a hyperbolic analogue of formula (1.1))

(ii) $D^{\mathrm{hyp}}_{n,2}(\mathbf{x}_1,\dots,\mathbf{x}_n;\mathbf{x}_{n+1},\mathbf{x}_{n+2})^{n} \geq D^{\mathrm{hyp}}_{n,1}(\mathbf{x}_1,\dots,\mathbf{x}_n;\mathbf{x}_{n+1})\cdot$
$\cdot D^{\mathrm{hyp}}_{n,1}(\mathbf{x}_1,\dots,\mathbf{x}_n;\mathbf{x}_{n+2}) \prod_{k=1}^{n} D^{\mathrm{hyp}}_{n-1,2}(\mathbf{x}_1,\dots\widehat{\mathbf{x}}_k,\dots,\mathbf{x}_n;\mathbf{x}_{n+1},\mathbf{x}_{n+2})$

For example for $n=2$ we have

$$\left(D^{\mathrm{hyp}}_{2,2}\right)^2 = 4\left[X_1^2+\xi_1\xi_2+(\xi_1+\xi_2)c\right]^2\left[X_2^2(\xi_1+\xi_2)+(X_2^2+\xi_1\xi_2)c\right]^2 \geq$$
$$\geq (X_1X_2+X_2(\xi_1+\xi_2)+\xi_1\xi_2)^2\cdot(X_1^2+\xi_1c)(c+\xi_2)\cdot(X_2^2+\xi_2c)(c+\xi_1)$$

which can be proved by considering the following three cases:

Case 1: $0<X_1<X_2<c<\xi_2<\xi_1$

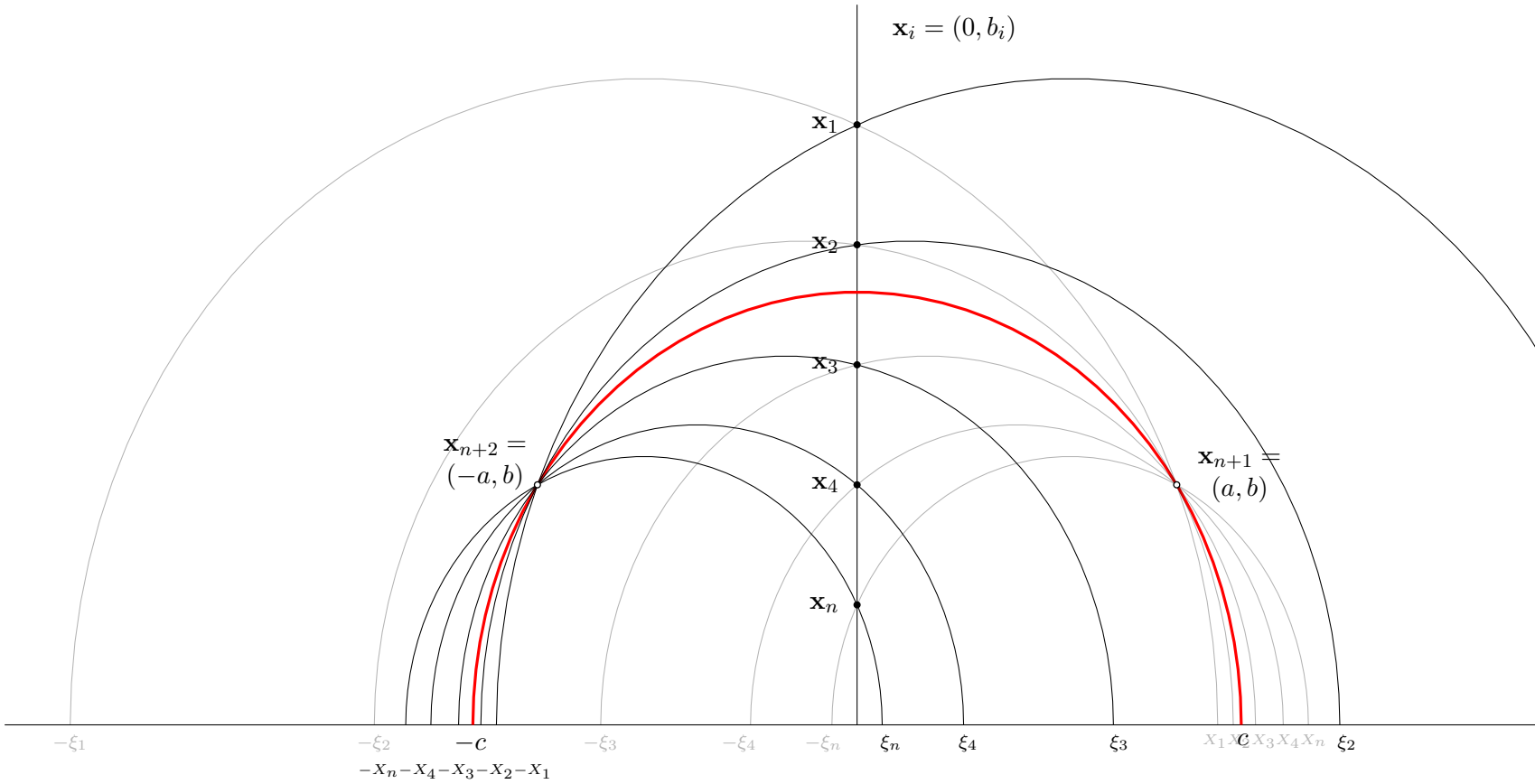
Case 2: $0<\xi_2<X_1<c<X_2<\xi_1$

Case 3: $0<\xi_2<\xi_1<c<X_1<X_2$

and we get coefficient-wise inequality in terms of variables representing increments e.g. in Case 1: $h_1=X_2-X_1$, $h_2=c-X_2$, $h_3=\xi_2-c$, $h_4=\xi_1-\xi_2$ and X_1.

We expect that the general proof will not be so elementary.

For hyperbolic planar four points case the conjectures C1 and C2 were verified and reported in [16] and the conjecture C1 for non planar case of four points in H^3 were treated in [9] therefore conjecture C1 was verified for any four points in H^3.



$t_{i,j} = t_{ij} =$ direction from $\mathbf{x}_i$ to $\mathbf{x}_j$; $0 < a < X_1 < X_2 < \cdots < X_n$

1.4 Type (A) configurations (cont.)

Even for $n = 3$ the inequality (1.1) is quite messy thanks to nonsymmetric character of both sides.

Knowing that sometimes it is easier to solve a more general problem we followed that path (although we didn't solve the problem in full generality). So let us start with the case $n = 2$. If we look at the following inequality

$$1 + X_1(\xi_1 + \xi_2) + X_1X_2\xi_1\xi_2 \geq (1 + X_1\xi_1)(1 + X_2\xi_2)$$

which is clearly true if $X_1 \geq X_2 \geq 0$ and $\xi_1, \xi_2 \geq 0$ we obtain the inequality (1.2) simply by a specialization $X_1 = \xi_1 = \lambda_2$, $X_2 = \xi_2 = \lambda_1$. So we proceed as follows:

Let $\xi_1, \ldots, \xi_n, X_1, \ldots, X_n, n \geq 1$ be two sets of commuting indeterminates. For any $l, 1 \leq l \leq n$ and any sequences $1 \leq i_1 \leq \cdots \leq i_l \leq n, 1 \leq j_1, \ldots, j_l \leq n$ we define polynomials $\Psi^I_J = \Psi^{i_1 \ldots i_l}_{j_1 \ldots j_l} \in \mathbb{Q}[\xi_1, \ldots, \xi_n, X_1, \ldots, X_n]$ as follows:

$$\Psi^I_J := \sum_{k=0}^{l} e_k(\xi_{j_1}, \xi_{j_2}, \ldots, \xi_{j_l}) X_{i_1} X_{i_2} \cdots X_{i_k},\ (l \geq 1),\ \Psi^{\emptyset}_{\emptyset} := 1\ (j = 0)$$

where e_k is the k-th elementary symmetric function.

In particular we have

$$\begin{aligned}
\Psi^i_j &= 1 + \xi_j X_i, \\
\Psi^{i_1 i_2}_{j_1 j_2} &= 1 + (\xi_{j_1} + \xi_{j_2}) X_{i_1} + \xi_{j_1}\xi_{j_2} X_{i_1} X_{i_2}, \\
\Psi^{i_1 i_2 i_3}_{j_1 j_2 j_3} &= 1 + (\xi_{j_1} + \xi_{j_2} + \xi_{j_3}) X_{i_1} + (\xi_{j_1}\xi_{j_2} + \xi_{j_1}\xi_{j_3} + \xi_{j_2}\xi_{j_3}) X_{i_1} X_{i_2} + \\
&\quad + \xi_{j_1}\xi_{j_2}\xi_{j_3} X_{i_1} X_{i_2} X_{i_3}, \\
&\text{etc.}
\end{aligned}$$

The polynomials Ψ^I_J are symmetric w.r.t. $\xi_{j_1}, \xi_{j_2}, \ldots, \xi_{j_l}$, but nonsymmetric w.r.t. $X_{i_1}, X_{i_2}, \ldots, X_{i_l}$. By specializing X_i's to assume real values such that $X_{i_1} \geq X_{i_2} \geq \ldots \geq X_{i_l} \geq 0$ then we obtain polynomials in ξ_j's satisfying the following simple but important property.

Proposition 1.3 *(Partition property)*
Let $(I_1, \ldots, I_s)$ and $(J_1, \ldots, J_s)$ be ordered set partitions of respective sets $I = \bigcup_{p=1}^{s} I_p$ and $J = \bigcup_{p=1}^{s} J_p$ such that $|I_p| = |J_p|$, $1 \leq p \leq s$. Then the inequality

$$\Psi^I_J \geq \prod_{p=1}^{s} \Psi^{I_p}_{J_p}$$

holds coefficientwise w.r.t. ξ_j's.

Proof .
Proof is evident from the definition of Ψ^I_J and the monotonicity of X_i's. ■

For the powers $\left(\Psi^I_J\right)^m$ we have the following conjecture.

Conjecture 1.4 *(Weighted Multiset Partition Conjecture)*
For given natural number m and sets I and J, $|I| = |J|$, of natural numbers let $(I_1, \ldots, I_s)$ and $(J_1, \ldots, J_s)$ be the partitions of the multiset I^m consisting of m copies of all elements of I and similarly for J^m.

(i) *Then the inequality*

$$\left(\Psi^I_J\right)^m \geq \prod_{p=1}^{s} \Psi^{I_p}_{J_p}$$

holds coefficientwise w.r.t. ξ_j's.

(ii) *The difference*

$$\left(\Psi^I_J\right)^m - \prod_{p=1}^{s} \Psi^{I_p}_{J_p}$$

is multi–Schur positive with respect to partial alphabets corresponding to the atoms of the intersection lattice of the set system $\{J_1, \ldots, J_s\}$.

For example, by Partition property, we have the following inequalities

$$\Psi^{1\ldots n}_{1\ldots n} \geq \Psi^k_k \Psi^{1..\widehat{k}..n}_{1..\widehat{k}..n},\ (1 \leq k \leq n)$$

which imply the following inequality

$$\left(\Psi^{1\ldots n}_{1\ldots n}\right)^n \geq \prod_{k=1}^{n} \Psi^k_k \prod_{k=1}^{n} \Psi^{1..\widehat{k}..n}_{1..\widehat{k}..n}$$

By Partition property we also have the following inequality

$$\Psi^{1\ldots n}_{1\ldots n} \geq \prod_{k=1}^{n} \Psi^k_k$$

The last two inequalities suggest the validity of the following inequality

$$\left(\Psi_{1\ldots n}^{1\ldots n}\right)^{n-1} \geq \prod_{k=1}^{n} \Psi_{1..\widehat{k}..n}^{1..\widehat{k}..n}$$

which is far from obvious (see Conjecture 1.5 below) although it would be a simple consequence of our Weighted Multiset Partition Conjecture.

This last conjectural inequality is interesting because it generalizes some special cases of not yet proven conjectures of Atiyah and Sutcliffe on configurations of points in three dimensional Euclidean space.

Our conjecture reads as follows:

Conjecture 1.5 *For any $n \geq 1$, let $X_1 \geq X_2 \geq \ldots \geq X_n \geq 0$, $\xi_1, \xi_2, \ldots, \xi_n \geq 0$, be nonnegative real numbers. Then we have coefficientwise (w.r.t. $\xi_1, \xi_2, \ldots, \xi_n$) inequality*

$$\left(\Psi_{12\cdots n}^{12\cdots n}\right)^{n-1} \geq \prod_{k=1}^{n} \Psi_{12\cdots\hat{k}\cdots n}^{12\cdots\hat{k}\cdots n}$$

where $12\cdots\hat{k}\cdots n$ denotes the sequence $12\cdots(k-1)(k+1)\cdots n$. The equality obviously holds true iff $X_1 = X_2 = \cdots = X_n$.

This Conjecture implies the strongest Atiyah–Sutcliffe's conjecture for almost collinear configurations of points (all but one point are collinear, called type(A) in [6]).

To illustrate the Conjecture (1.5) we consider first the cases $n = 2$ and $n = 3$.

Case $n = 2$: We have

$$\begin{aligned}\Psi_{12}^{12} &= 1 + (\xi_1 + \xi_2)X_1 + \xi_1\xi_2 X_1 X_2 = \\ &= 1 + \xi_1 X_1 + \xi_2 X_2 + \xi_1\xi_2 X_1 X_2 + (X_1 - X_2)\xi_2 = \\ &= (1 + \xi_1 X_1)(1 + \xi_2 X_2) + \xi_2(X_1 - X_2) \geq \\ &\geq (1 + \xi_1 X_1)(1 + \xi_2 X_2) = \Psi_1^1 \Psi_2^2.\end{aligned}$$

Case $n = 3$: We first write Ψ_{123}^{123} in two different ways:

$$\Psi_{123}^{123} = \xi_2(X_1 - X_2) + \widehat{\Psi}_{1\underline{2}3}^{123} \quad \text{and} \quad \Psi_{123}^{123} = \xi_3(X_1 - X_2) + \widehat{\Psi}_{12\underline{3}}^{123}.$$

Note that $\widehat{\Psi}^{123}_{1\underline{2}3}$ is obtained from Ψ^{123}_{123} by replacing the linear term $\xi_2 X_1$ by $\xi_2 X_2$, hence all its coefficients are nonnegative.

The left hand side of the Conjecture (1.5) L_3 can be rewritten as follows:

$$\begin{aligned} L_3 = (\Psi^{123}_{123})^2 &= (\xi_2(X_1 - X_2) + \widehat{\Psi}^{123}_{1\underline{2}3})\Psi^{123}_{123} \\ &= \xi_2(X_1 - X_2)\Psi^{123}_{123} + \widehat{\Psi}^{123}_{1\underline{2}3}\Psi^{123}_{123} \\ &= \xi_2(X_1 - X_2)\Psi^{123}_{123} + \widehat{\Psi}^{123}_{1\underline{2}3}(\xi_3(X_1 - X_2) + \widehat{\Psi}^{123}_{12\underline{3}}) \\ &= L'_3(X_1 - X_2) + \widehat{\Psi}^{123}_{1\underline{2}3}\widehat{\Psi}^{123}_{12\underline{3}} \end{aligned}$$

where $L'_3 = \xi_2\Psi^{123}_{123} + \xi_3\widehat{\Psi}^{123}_{1\underline{2}3}$ is a positive polynomial.

Now we have

$$L_3 \geq \widehat{L}_3 := \widehat{\Psi}^{123}_{1\underline{2}3}\widehat{\Psi}^{123}_{12\underline{3}}.$$

By using the formula

$$\widehat{\Psi}^{123}_{1\underline{2}3} = \Psi^{12}_{13} + \xi_2 X_2 \Psi^{13}_{13} = (\Psi^2_2 - 1)\Psi^{13}_{13} + \Psi^{12}_{13}$$

we can rewrite $\widehat{L}_3$ as

$$\begin{aligned} \widehat{L}_3 &= \left[(\Psi^{12}_{13} - \Psi^{13}_{13}) + \Psi^2_2\Psi^{13}_{13}\right]\widehat{\Psi}^{123}_{12\underline{3}} \\ &= \xi_1\xi_3 X_1(X_2 - X_3)\widehat{\Psi}^{123}_{12\underline{3}} + \Psi^{13}_{13}(\Psi^2_2\widehat{\Psi}^{123}_{12\underline{3}}) \end{aligned}$$

The last term in parenthesis can be written as

$$\begin{aligned} \Psi^2_2\widehat{\Psi}^{123}_{12\underline{3}} &= \Psi^{12}_{12}\Psi^{23}_{23} + \Psi^1_2(\Psi^{22}_{23} - \Psi^{23}_{23}) \\ &= \Psi^{12}_{12}\Psi^{23}_{23} + \xi_2\xi_3 X_2(X_2 - X_3)\Psi^1_2, \end{aligned}$$

so we get

$$\widehat{L}_3 = L''_3(X_2 - X_3) + \Psi^{12}_{12}\Psi^{13}_{13}\Psi^{23}_{23}$$

where L''_3 denotes the positive polynomial

$$L''_3 = \xi_1\xi_3 X_1\widehat{\Psi}^{123}_{12\underline{3}} + \xi_2\xi_3 X_2\Psi^1_2\Psi^{13}_{13}.$$

We now have an explicit formula for L_3:

$$L_3 = L'_3(X_1 - X_2) + L''_3(X_2 - X_3) + \Psi^{12}_{12}\Psi^{13}_{13}\Psi^{23}_{23}$$

with L_3', L_3'' positive polynomials, which together with $X_1 \geq X_2 \geq X_3 (\geq 0)$ implies that

$$L_3 \geq R_3 := \Psi_{12}^{12}\Psi_{13}^{13}\Psi_{23}^{23}$$

and the Conjecture (1.5) ($n = 3$) is proved.

In fact we have proven an instance $n = 3$ $\widehat{L}_3 \geq R_3$ of a stronger conjecture which we are going to formulate now. Let $2 \leq k \leq n$. We define the modified polynomials $\widehat{\Psi}_{12\ldots\underline{k}\ldots n}^{12\ldots k\ldots n}$ as follows:

$$\widehat{\Psi}_{12\ldots\underline{k}\ldots n}^{12\ldots k\ldots n} := \xi_k(X_2 - X_1) + \Psi_{12\ldots n}^{12\ldots n}$$

obtained from $\Psi_{12\ldots n}^{12\ldots n}$ by replacing only one term $\xi_k X_1$ by $\xi_k X_2$, hence $\widehat{\Psi}_{12\ldots\underline{k}\ldots n}^{12\ldots k\ldots n}$ are still positive. Let us introduce the following notation:

$$\widehat{L}_n := \prod_{k=2}^{n} \widehat{\Psi}_{12\ldots\underline{k}\ldots n}^{12\ldots k\ldots n} \; ; \quad R_n := \prod_{k=1}^{n} \Psi_{12\ldots\hat{k}\ldots n}^{12\ldots\hat{k}\ldots n}.$$

Then clearly $L_n := (\Psi_{12\ldots n}^{12\ldots n})^{n-1} \geq \widehat{L}_n$. Now our stronger conjecture reads as

Conjecture 1.6

$$\widehat{L}_n \geq R_n \ (n \geq 1)$$

with equality iff $X_2 = X_3 = \cdots = X_n$.

More generally, we conjecture that the difference $\widehat{L}_n - R_n$ is a polynomial in the differences $X_2 - X_3$, $X_3 - X_4$, …, $X_{n-1} - X_n$ with coefficients in $\mathbb{Z}_{\geq 0}[X_1, \ldots, X_n, \xi_1, \ldots, \xi_n]$.

Proposition 1.7

$$L_n = L_n'(X_1 - X_2) + \widehat{L}_n$$

for some positive polynomial L_n'.

Proof of Proposition 1.7.

$$
\begin{aligned}
L_n &= (\Psi^{12\cdots n}_{12\cdots n})^{n-1} = (\xi_2(X_1 - X_2) + \widehat{\Psi}^{12\cdots n}_{1\underline{2}\cdots n})(\Psi^{12\cdots n}_{12\cdots n})^{n-2} \\
&= \xi_2(X_1 - X_2)(\Psi^{12\cdots n}_{12\cdots n})^{n-2} + \widehat{\Psi}^{12\cdots n}_{1\underline{2}\cdots n}(\xi_3(X_1 - X_2) + \widehat{\Psi}^{123\cdots n}_{12\underline{3}\cdots n})(\Psi^{12\cdots n}_{12\cdots n})^{n-3} \\
&= \xi_2(X_1 - X_2)(\Psi^{12\cdots n}_{12\cdots n})^{n-2} + \xi_3(X_1 - X_2)\widehat{\Psi}^{12\cdots n}_{1\underline{2}\cdots n}(\Psi^{12\cdots n}_{12\cdots n})^{n-3} + \\
&\quad + \widehat{\Psi}^{12\cdots n}_{1\underline{2}\cdots n}\widehat{\Psi}^{123\cdots n}_{12\underline{3}\cdots n}(\Psi^{12\cdots n}_{12\cdots n})^{n-3} \\
&\vdots \\
&= (\textstyle\sum_{k=1}^{n-1} \xi_{k+1}(\prod_{j=2}^{k} \widehat{\Psi}^{12\ldots j\ldots n}_{12\ldots\underline{j}\ldots n})(\Psi^{12\ldots n}_{12\ldots n})^{n-k})(X_1 - X_2) + \textstyle\prod_{j=2}^{n} \widehat{\Psi}^{12\ldots j\ldots n}_{12\ldots\underline{j}\ldots n}.
\end{aligned}
$$

■

Now we turn to study the quotient

$$\frac{L_n}{R_n} = \frac{(\Psi^{1\ldots n}_{1\ldots n})^{n-1}}{\prod_{k=1}^{n} \Psi^{1\ldots\widehat{k}\ldots n}_{1\ldots\widehat{k}\ldots n}}$$

by studying the growth behaviour of quotients of its factors $\Psi^{1\ldots n}_{1\ldots n}/\Psi^{1\ldots\widehat{k}\ldots n}_{1\ldots\widehat{k}\ldots n}$ w.r.t. any of its arguments X_r, $1 \le r \le n$.

In the following theorem we obtain an explicit formula for the numerators of the derivatives w.r.t. $X_r, (1 \le r \le n, r \ne k)$ of the quantities $\Psi^{1\ldots n}_{1\ldots n}/\Psi^{1\ldots\widehat{k}\ldots n}_{1\ldots\widehat{k}\ldots n}$. From this formulas we get some monotonicity properties which enable us to state some new (refined) conjectures later on.

Theorem 1.8 *Let*

$$\Delta_r := \partial_{X_r}\Psi^{1\ldots n}_{1\ldots n} \cdot \Psi^{1\ldots\widehat{k}\ldots n}_{1\ldots\widehat{k}\ldots n} - \Psi^{1\ldots n}_{1\ldots n} \cdot \partial_{X_r}\Psi^{1\ldots\widehat{k}\ldots n}_{1\ldots\widehat{k}\ldots n}, \quad (1 \le r \le n). \tag{1.3}$$

Then we have the following explicit formulas

(i) *for any r, $1 \le r < k(\le n)$ we have*

$$
\begin{aligned}
\Delta_r = \; & \xi_k \sum_{0\le i<r\le j\le n} s^{(k)}_{(j-1,i)'} X_1^2 \cdots X_i^2 X_{i+1} \cdots \widehat{X}_k \cdots X_j + \\
& + \sum_{0\le i<r,k\le j<n} e_i e^{(k)}_j X_1^2 \cdots X_i^2 X_{i+1} \cdots \widehat{X}_r \cdots \widehat{X}_k \cdots X_j (X_k - X_{j+1})
\end{aligned}
$$

(ii) *for any* r, $(1 \leq)k < r \leq n$ *we have*

$$\Delta_r = -\left(\sum_{0\leq i<r\leq j\leq n} s^{(k)}_{(j-1,i)'} X_1^2 \cdots X_i^2 X_{i+1} \cdots \widehat{X}_k \cdots \widehat{X}_r \cdots X_j + \right.$$
$$\left. + \sum_{0\leq i<k, r\leq j<n} e_i^{(k)} e_j X_1^2 \cdots X_i^2 X_{i+1} \cdots \widehat{X}_k \cdots \widehat{X}_r \cdots X_j (X_{j+1} - X_k) \right)$$

where $s^{(k)}_{\lambda}$ *denotes the* λ*–th Schur function of* $\xi_1, \ldots, \xi_{k-1}, \xi_{k+1}, \ldots, \xi_n$ *(*ξ_k *omitted).*

Proof of Theorem 1.8.

(i) For any r, $1 \leq r < k(\leq n)$ we find explicitly a formula as follows. We shall use notations $X_{1..i} := X_1 X_2 \cdots X_i$, for multilinear monomials and $e_i := e_i(\xi_1, \ldots, \xi_n)$, $e_i^{(k)} = e_i(\xi_1, \ldots, \widehat{\xi_k}, \ldots \xi_n)$ for the elementary symmetric functions (here k is fixed). Then we can rewrite our basic quantities

$$\Psi^{1\ldots n}_{1\ldots n} := \sum_{i=0}^{n} e_i X_{1..i} \tag{1.4}$$

$$\begin{aligned} \Psi^{1\ldots\widehat{k}\ldots n}_{1\ldots\widehat{k}\ldots n} &:= \sum_{i=0}^{k-1} e_i^{(k)} X_{1..i} + \frac{1}{X_k} \sum_{i=k}^{n-1} e_i^{(k)} X_{1..i+1} = \\ &= \sum_{i=0}^{n-1} e_i^{(k)} X_{1..i} + \frac{1}{X_k} \sum_{i=k}^{n-1} e_i^{(k)} X_{1..i} (X_{i+1} - X_k) \end{aligned} \tag{1.5}$$

For the derivatives we get immediately

$$\partial_{X_r} \Psi^{1\ldots n}_{1\ldots n} = \frac{1}{X_r} \sum_{i=r}^{n} e_i X_{1..i} = \frac{1}{X_r} \left(\Psi^{1\ldots n}_{1\ldots n} - \sum_{i=0}^{r-1} e_i X_{1..i} \right) \tag{1.6}$$

$$\partial_{X_r} \Psi^{1\ldots\widehat{k}\ldots n}_{1\ldots\widehat{k}\ldots n} = \frac{1}{X_r} \sum_{i=r}^{n-1} e_i^{(k)} X_{1..i} + \frac{1}{X_k X_r} \sum_{i=k}^{n-1} e_i^{(k)} X_{1..i} (X_{i+1} - X_k) \tag{1.7}$$

$$= \frac{1}{X_r} \left(\Psi^{1\ldots\widehat{k}\ldots n}_{1\ldots\widehat{k}\ldots n} - \sum_{i=0}^{r-1} e_i^{(k)} X_{1..i} \right) \tag{1.8}$$

By plugging (1.6) and (1.8) into (1.3) we obtain

$$X_r\Delta_r = \Psi^{1\dots n}_{1\dots n}\left(\sum_{i=0}^{r-1} e_i^{(k)} X_{1..i}\right) - \Psi^{1\dots\widehat{k}\dots n}_{1\dots\widehat{k}\dots n}\left(\sum_{i=0}^{r-1} e_i X_{1..i}\right) =$$

and after simple cancelation, by invoking (1.5) we get

$$= \left(\sum_{j=r}^{n} e_j X_{1..j}\right)\left(\sum_{i=0}^{r-1} e_i^{(k)} X_{1..i}\right) - \left(\sum_{j=r}^{n-1} e_j^{(k)} X_{1..j} + \frac{1}{X_k}\sum_{j=k}^{n-1} e_j^{(k)} X_{1..j}(X_{j+1}-X_k)\right)\left(\sum_{i=0}^{r-1} e_i X_{1..i}\right)$$

i.e.

$$X_r\Delta_r = \sum_{0\le i<r\le j\le n} (e_j e_i^{(k)} - e_i e_j^{(k)}) X_{1..i}X_{1..j} + \frac{1}{X_k}\sum_{0\le i<r,k\le j<n} e_i e_j^{(k)} X_{1..i}X_{1..j}(X_k - X_{j+1})$$

If we use a simple identity $e_j = e_j^{(k)} + \xi_k e_{j-1}^{(k)}$, we can identify the quantity

$$e_j e_i^{(k)} - e_i e_j^{(k)} = (e_j^{(k)} + \xi_k e_{j-1}^{(k)})e_i^{(k)} - (e_i^{(k)} + \xi_k e_{i-1}^{(k)})e_j^{(k)} =$$

$$= \begin{vmatrix} e_{j-1}^{(k)} & e_j^{(k)} \\ e_{i-1}^{(k)} & e_i^{(k)} \end{vmatrix} \xi_k = s^{(k)}_{2^i 1^{j-i-1}}\xi_k$$

Thus in this case $(1 \le r < k)$ we obtain a formula

$$\Delta_r = \xi_k \sum_{0\le i<r\le j\le n} s^{(k)}_{(j-1,i)'} X_1^2\cdots X_i^2 X_{i+1}\cdots\widehat{X}_r\cdots X_j + \sum_{0\le i<r,k\le j<n} e_i e_j^{(k)} X_1^2\cdots X_i^2 X_{i+1}\cdots\widehat{X}_r\cdots\widehat{X}_k\cdots X_j(X_k - X_{j+1})$$

(where $e_j^{(k)} = e_j(\xi_1,\dots,\widehat{\xi_k},\dots,\xi_n)$) in terms of Schur functions (of arguments $\xi_1,\dots,\widehat{\xi_k},\dots,\xi_n$) corresponding to a transpose $(j-1,i)' = (2^i 1^{j-i-1})$ of a partition $(j-1,i)$ (cf. Jacobi–Trudi formula, I 3.5 in [8]).

(ii) For any r, $(1\le)k < r \le n$. In this case we use

$$\partial_{X_r}\Psi^{1\dots\widehat{k}\dots n}_{1\dots\widehat{k}\dots n} = \frac{1}{X_k X_r}\sum_{j=r-1}^{n-1} e_j^{(k)} X_{1..j+1}$$

$$\Psi^{1\ldots\widehat{k}\ldots n}_{1\ldots\widehat{k}\ldots n} = \sum_{i=0}^{k-1} e_i^{(k)} X_{1..i} + \frac{1}{X_k}\sum_{i=k}^{n-1} e_i^{(k)} X_{1..i+1} =$$
$$= \frac{1}{X_k}\left(\sum_{i=0}^{k-1} X_{1..i}(X_k - X_{i+1}) + \sum_{i=0}^{n-1} e_i^{(k)} X_{1..i}\right)$$

By plugging this into (1.3) we get

$$\begin{aligned} X_k X_r \Delta_r &= \left(\sum_{j=r}^{n} e_j X_{1..j}\right)\left(\sum_{i=0}^{k-1} e_i^{(k)} X_{1..i}(X_k - X_{i+1}) + \sum_{i=0}^{n-1} e_i^{(k)} X_{1..i+1}\right) - \\ &\quad - \left(\sum_{j=0}^{r-1} e_j X_{1..j} + \sum_{j=r}^{n} e_j X_{1..j}\right)\left(\sum_{i=r-1}^{n-1} e_i^{(k)} X_{1..i+1}\right) \\ &= \left(\sum_{i=0}^{r-2} e_i^{(k)} X_{1..i+1}\right)\left(\sum_{j=r}^{n} e_j X_{1..j}\right) - \left(\sum_{i=0}^{r-1} e_i X_{1..i}\right)\left(\sum_{j=r-1}^{n-1} e_j^{(k)} X_{1..j+1}\right) + \\ &\quad + \sum_{i=0}^{k-1}\sum_{j=r}^{n} e_i^{(k)} e_j X_{1..i} X_{1..j}(X_k - X_{i+1}) \\ &= \left(\sum_{i=1}^{r-1} e_{i-1}^{(k)} X_{1..i}\right)\left(\sum_{j=r}^{n} e_j X_{1..j}\right) - \left(\sum_{i=0}^{r-1} e_i X_{1..i}\right)\left(\sum_{j=r}^{n} e_{j-1}^{(k)} X_{1..j}\right) + \\ &\quad + \sum_{i=0}^{k-1}\sum_{j=r}^{n} e_i^{(k)} e_j X_{1..i} X_{1..j}(X_k - X_{i+1}) \end{aligned}$$

By using a formula for elementary symmetric functions ($e_i = e_i^{(k)} + \xi_k e_{i-1}^{(k)}$) we can write in terms of Schur functions (of arguments $\xi_1, \ldots, \xi_{k-1}, \xi_{k+1}, \ldots, \xi_n$), where λ' is a conjugate of λ.

$$e_{i-1}^{(k)} e_j - e_i e_{j-1}^{(k)} = e_{i-1}^{(k)} e_j^{(k)} - e_i^{(k)} e_{j-1}^{(k)} = -\begin{vmatrix} e_{j-1}^{(k)} & e_j^{(k)} \\ e_{i-1}^{(k)} & e_i^{(k)} \end{vmatrix} = -s^{(k)}_{2^i 1^{j-i-1}} = -s^{(k)}_{(j-1,i)'}$$

Thus we obtain a formula

$$\begin{aligned} \Delta_r = \; & -\left(\sum_{0\le i<r\le j\le n} s^{(k)}_{(j-1,i)'} X_1^2 \cdots X_i^2 X_{i+1} \cdots \widehat{X}_k \cdots \widehat{X}_r \cdots X_j + \right. \\ & \left. + \sum_{0\le i<k, r\le j<n} e_i^{(k)} e_j X_1^2 \cdots X_i^2 X_{i+1} \cdots \widehat{X}_k \cdots \widehat{X}_r \cdots X_j (X_{j+1} - X_k)\right) \end{aligned}$$

Corollary 1.9 *(X_r–monotonicity)*
Let $X_1 \geq \cdots \geq X_n \geq 0$, $\xi_1, \ldots, \xi_n \geq 0$ be as before. Then

(i) for any r, $1 \leq r < k$ $(\leq n)$ we have

$$\frac{\Psi^{1\ldots n}_{1\ldots n}}{\Psi^{1\ldots\widehat{k}\ldots n}_{1\ldots\widehat{k}\ldots n}} \geq \frac{\Psi^{1\ldots\ r+1\ r+1\ \ldots n}_{1\ldots\ \ \ r\ \ \ r+1\ \ \ldots n}}{\Psi^{1\ldots\ r+1\ r+1\ \ldots\widehat{k}\ldots n}_{1\ldots\ \ \ r\ \ \ r+1\ \ \ldots\widehat{k}\ldots n}}$$

(ii) for any r, $(1 \leq)$ $k < r$ $(\leq n)$ we have

$$\frac{\Psi^{1\ldots n}_{1\ldots n}}{\Psi^{1\ldots\widehat{k}\ldots n}_{1\ldots\widehat{k}\ldots n}} \geq \frac{\Psi^{1\ldots\ r-1\ r-1\ \ldots n}_{1\ldots\ r-1\ \ \ r\ \ \ \ldots n}}{\Psi^{1\ldots\widehat{k}\ldots\ r-1\ r-1\ \ldots n}_{1\ldots\widehat{k}\ldots\ r-1\ \ \ r\ \ \ \ldots n}}$$

Now we illustrate how to use Corollary 1.9 to prove our Conjecture 1.5 for $n = 2, 3, 4$ and 5.

Case $n = 2$

$$Q_2 := \frac{\Psi^{12}_{12}}{\Psi^1_1\Psi^2_2} \geq \frac{\Psi^{22}_{12}}{\Psi^2_1\Psi^2_2} = 1 \text{ (by } (i))$$

Case $n = 3$

$$\begin{aligned} Q_3 := \frac{\Psi^{123}_{123}\Psi^{123}_{123}}{\Psi^{12}_{12}\Psi^{13}_{13}\Psi^{23}_{23}} &\geq \frac{\Psi^{\mathbf{2}23}_{123}\Psi^{123}_{123}}{\Psi^{\mathbf{2}2}_{12}\Psi^{13}_{13}\Psi^{23}_{23}} \geq \frac{\Psi^{223}_{123}\Psi^{\mathbf{2}23}_{123}}{\Psi^{22}_{12}\Psi^{\mathbf{2}3}_{13}\Psi^{23}_{23}} \text{ (by } (i)) \\ &\geq \frac{\Psi^{22\mathbf{2}}_{123}\Psi^{223}_{123}}{\Psi^{22}_{12}\Psi^{2\mathbf{2}}_{13}\Psi^{23}_{23}} \geq \frac{\Psi^{222}_{123}\Psi^{22\mathbf{2}}_{123}}{\Psi^{22}_{12}\Psi^{22}_{13}\Psi^{\mathbf{2}2}_{23}} = 1 \text{ (by } (ii)) \end{aligned}$$

Case $n = 4$

$$Q_4 := \frac{(\Psi^{1234}_{1234})^3}{\Psi^{123}_{123}\Psi^{124}_{124}\Psi^{134}_{134}\Psi^{234}_{234}} \geq \cdots \geq \frac{\Psi^{2244}_{1234}(\Psi^{2224}_{1234})^2}{\Psi^{224}_{123}\Psi^{224}_{124}\Psi^{224}_{134}\Psi^{224}_{234}} \quad (\geq 1)$$

This last inequality follows from the following symmetric function identity:

$$\begin{aligned} &\Psi^{2244}_{1234}(\Psi^{2224}_{1234})^2 - \Psi^{224}_{123}\Psi^{224}_{124}\Psi^{224}_{134}\Psi^{224}_{234} = \\ &X_2^2X_4^4m_{2222} + 2X_2^2X_4^3m_{2221} + X_2^2X_4^2m_{222} + 3X_2^2X_4^2m_{2211} + X_2^2X_4m_{221} \\ &+4X_2^2X_4m_{2111} + X_2^2m_{211} + X_2(3X_2 + 2X_4)m_{1111} + X_2m_{111} \end{aligned}$$

where $m_\lambda = m_\lambda(\xi_1, \xi_2, \xi_3, \xi_4)$ are the monomial symmetric functions.
Case $n = 5$

$$Q_5 := \frac{(\Psi^{1\dots5}_{1\dots5})^4}{\prod_{k=1}^{5} \Psi^{1\dots\widehat{k}\dots5}_{1\dots\widehat{k}\dots5}} \geq \cdots \geq \frac{(\Psi^{22244}_{12345}\Psi^{22444}_{12345})^2}{\Psi^{2244}_{1234}\Psi^{2244}_{1235}\Psi^{2244}_{1245}\Psi^{2244}_{1345}\Psi^{2244}_{2345}} \quad (\geq 1)$$

The last inequality is equivalent to an explicit symmetric function identity with all coefficients (w.r.t. monomial basis) positive.

Now we state our stronger conjecture.

Conjecture 1.10 *(for symmetric functions)*
Let $X_1 \geq X_2 \geq \cdots \geq X_n \geq 0$ *and* $\xi_1, \dots, \xi_n \geq 0$. *Then the inequalities*

(a) For n *even*

$$\boldsymbol{\Psi}^{\mathbf{2\ 2\ 4\ 4\dots n\ n}}_{\mathbf{1\ 2\dots n{-}1\ n}} \left(\prod_{k=1}^{n/2} \boldsymbol{\Psi}^{\mathbf{2\ 2\ 4\ 4\dots 2k\ 2k\ 2k\dots n{-}2\ n{-}2\ n}}_{\mathbf{1\ 2\ 3\ 4\ \dots\ n{-}1\ n}} \right)^{\mathbf{2}} \geq \prod_{k=1}^{n} \boldsymbol{\Psi}^{\mathbf{2\ 2\ 4\ 4\dots n{-}2\ n{-}2\ n}}_{\mathbf{1\ 2\ \dots\ \widehat{k}\ \dots\ n{-}1\ n}}$$

(b) For n *odd*

$$\left(\prod_{k=1}^{\lfloor n/2 \rfloor} \boldsymbol{\Psi}^{\mathbf{2\ 2\ 4\ 4\dots 2k\ 2k\ 2k\dots n{-}1\ n{-}1}}_{\mathbf{1\ 2\ 3\ 4\ \dots\ n{-}1\ n}} \right)^{\mathbf{2}} \geq \prod_{k=1}^{n} \boldsymbol{\Psi}^{\mathbf{2\ 2\ 4\ 4\dots n{-}1\ n{-}1}}_{\mathbf{1\ 2\ \dots\ \widehat{k}\ \dots\ n}}$$

hold true coefficientwise (m–positivity).

Now we motivate another inequalities for symmetric functions which also refine the strongest Atiyah–Sutcliffe conjecture for configurations of type (A). Let $n = 3$. We apply Corollary 1.9 by using steps (ii) only.

$$Q_3 := \frac{\Psi^{123}_{123}\Psi^{123}_{123}}{\Psi^{12}_{12}\Psi^{13}_{13}\Psi^{23}_{23}} \geq \frac{\Psi^{113}_{123}\Psi^{123}_{123}}{\Psi^{12}_{12}\Psi^{13}_{13}\Psi^{13}_{23}} \geq \frac{\Psi^{112}_{123}\Psi^{123}_{123}}{\Psi^{12}_{12}\Psi^{12}_{13}\Psi^{13}_{23}} \geq \frac{\Psi^{112}_{123}\Psi^{122}_{123}}{\Psi^{12}_{12}\Psi^{12}_{13}\Psi^{12}_{23}} \geq 1$$

The last inequality is equivalent to nonnegativity of the expression

$$\Psi^{112}_{123}\Psi^{122}_{123} - \Psi^{12}_{12}\Psi^{12}_{13}\Psi^{12}_{23} \quad (= X_1(X_1 - X_2)^2\xi_1\xi_2\xi_3 \geq 0).$$

Similarly, for $n = 4$, the symmetric function inequality stronger than $Q_4 \geq 1$ would be the following

$$\Psi^{1123}_{1234}\Psi^{1223}_{1234}\Psi^{1233}_{1234} \geq \Psi^{123}_{123}\Psi^{123}_{124}\Psi^{123}_{134}\Psi^{123}_{234}$$

Now we state a general conjecture for symmetric functions which imply the strongest Atiyah–Sutcliffe conjecture for almost collinear type (A) configurations.

Conjecture 1.11 *Let $X_1 \geq \cdots \geq X_n \geq$, $\xi_1, \ldots \xi_n \geq 0$. Then the following inequality for symmetric functions in $\xi_1, \ldots, \xi_n$*

$$\Psi^{\mathbf{11}2\ldots n-1}_{123\ldots n}\Psi^{1\mathbf{22}3\ldots n-1}_{1234\ldots n}\cdots\Psi^{12\ldots n-2\ \boldsymbol{n-1}\ \boldsymbol{n-1}}_{12\ldots n-2\ n-1\ n} \geq \Psi^{1\ 2\ldots n-1}_{1\ 2\ldots n-1}\Psi^{1\ 2\ldots n-1}_{1\ 2\ldots n-2\ n}\cdots\Psi^{1\ 2\ldots n-1}_{2\ 3\ldots n-1}$$

i.e.

$$\prod_{k=1}^{n-1}\Psi^{1\ \mathbf{2\ldots k}\ \boldsymbol{k}\ \ldots n}_{1\ 2\ldots k\ k+1\ldots n} \geq \prod_{k=1}^{n}\Psi^{1\ 2\ \ldots\ n-1}_{1\ 2\ldots\widehat{k}\ldots n}$$

holds true coefficientwise (m–positivity, even s–positivity).

Remark 1.12 *Conjectures 1.10 and 1.11 seems to hold also for the Schur basis of symmetric functions in $\xi_1, \ldots, \xi_n$.*

We have checked this Conjecture 1.11 up to $n = 5$ by using `Maple` and symmetric function package `SF` of J. Stembridge. For n bigger than five the computations are extremely intensive and hopefully in the near future would be possible by using more powerful computers.

Note that the right hand side of the Conjecture 1.11 involves symmetric functions of partial alphabets $\xi_1, \xi_2, \ldots, \xi_{k-1}, \xi_{k+1}, \ldots, \xi_n$. But the left hand side doesn't have this "defect". Our objective now is to give explicit formula for the right hand side in terms of the elementary symmetric functions of the full alphabet $\xi_1, \xi_2, \ldots, \xi_n$. This we are going to achieve by using resultants as follows.

Lemma 1.13 *For any k, $(1 \leq k \leq n)$, we have*

$$\Psi^{1\ldots k\ldots n-1}_{1\ldots\widehat{k}\ldots n} = \sum_{j=0}^{n-1} a_j \xi_k^{n-1-j}$$

where

$$\begin{aligned}
&a_{n-1} = 1 + X_1e_1 + X_1X_2e_2 + \ldots + X_1\cdots X_{n-1}e_{n-1},\\
&a_{n-2} = -X_1 - X_1X_2e_1 - \ldots - X_1\cdots X_{n-1}e_{n-2},\\
&\cdots\\
&a_0 = (-1)^{n-1}X_1\cdots X_{n-1}
\end{aligned}$$

i.e.

$$a_{n-1-j} = (-1)^j\sum_{i=j}^{n-1} X_1\cdots X_i e_{i-j}$$

Proof of Lemma 1.13.
By definition we have

$$\Psi^{1\dots n-1}_{1\dots\widehat{k}\dots n} = \sum_{i=0}^{n-1} X_1 \cdots X_i e_i^{(k)} \tag{1.9}$$

where $e_i^{(k)}$ is the i–th elementary function of $\xi_1, \dots, \xi_{k-1}, \xi_{k+1}, \dots, \xi_n$. Now from the decomposition

$$(1+\xi_k t)^{-1} \prod_{j=1}^{n}(1+\xi_j t) = \prod_{j\neq k}(1+\xi_j t) = \sum_{i=0}^{n-1} e_i^{(k)} t^i$$

we get

$$e_i^{(k)} = e_i - e_{i-1}\xi_k + e_{i-2}\xi_k^2 - \cdots + (-1)^i \xi_k^i$$

By substituting this into equation (1.9) the Lemma 1.13 follows. ∎

Then, by Lemma 1.13, the right hand side of the Conjecture 1.11

$$R_n = \prod_{k=1}^{n} \Psi^{1\ 2\ \dots\ k\ \dots\ n-1}_{1\ 2\ \dots\ \widehat{k}\ \dots\ \ n} = \prod_{k=1}^{n} \left(\sum_{j=0}^{n-1} a_j \xi_k^{n-1-j} \right)$$

can be interpreted as a resultant $R_n = Resultant(f, g)$ of the following two polynomials

$$f(x) = \sum_{j=0}^{n-1} a_j x^{n-1-j}$$
$$g(x) = \prod_{i=1}^{n}(x-\xi_i) = \sum_{j=0}^{n}(-1)^j e_j x^{n-j}$$

Then Sylvester formula

$$R_n = \begin{vmatrix} 1 & -e_1 & e_2 & -e_3 & \dots & (-1)^n e_n & & \\ & 1 & -e_1 & e_2 & -e_3 & \dots & & \\ & & \ddots & & & & & \\ & & & 1 & -e_1 & \cdots & & \\ a_0 & a_1 & a_2 & \cdots & a_n & & & \\ & a_0 & a_1 & a_2 & \cdots & a_n & & \\ & & \ddots & & & & & \\ & & & a_0 & a_1 & a_2 & \cdots & a_n \end{vmatrix} \left(=: \begin{vmatrix} A & B \\ C & D \end{vmatrix}\right)$$

can be simplified as

$$= |A| \cdot |D - CA^{-1}B| = |D - CA^{-1}B|.$$

The entries of the $n \times n$ matrix $\Delta := D - CA^{-1}B$ are given by

$$\delta_{ij} = \begin{cases} (-1)^{j-i-1} \displaystyle\sum_{k=j+1}^{n} X_1 \cdots X_{k+i-j} e_k, & 0 \le i < j \le n-1 \\ (-1)^{j-i} \displaystyle\sum_{k=0}^{j} X_1 \cdots X_{k+i-j} e_k, & 0 \le j \le i \le n-1 \end{cases}$$

For example, for $n = 3$

$$\Delta_3 = \begin{vmatrix} 1 & X_1 e_2 + X_1 X_2 e_3 & -X_1 e_3 \\ -X_1 & 1 + X_1 e_1 & X_1 X_2 e_3 \\ X_1 X_2 & -X_1 - X_1 X_2 e_1 & 1 + X_1 e_1 + X_1 X_2 e_2 \end{vmatrix}$$

By elementary operations (including multiplication of 2^{nd} row and column by -1) we get

$$\Delta_3 = \begin{vmatrix} 1 & * & * \\ 0 & \Psi^{112}_{123} & X_1(X_2 - X_1)e_3 \\ 0 & X_2 - X_1 & \Psi^{122}_{123} \end{vmatrix} = \begin{vmatrix} \Psi^{112}_{123} & X_1(X_2 - X_1)e_3 \\ X_2 - X_1 & \Psi^{122}_{123} \end{vmatrix} =$$

$$= \begin{vmatrix} \Psi^{112}_{123} & X_1(X_1 - X_2)e_3 \\ X_1 - X_2 & \Psi^{122}_{123} \end{vmatrix}.$$

Similarly, for $n = 4$ we obtain

$$\Delta_4 = \begin{vmatrix} \Psi^{1123}_{1234} & -X_1(X_1-X_2)e_3 - X_1X_2(X_1-X_3)e_4 & X_1(X_1-X_2)e_4 \\ -(X_1-X_2) & \Psi^{1223}_{1234} & -X_1X_2(X_2-X_3)e_4 \\ X_1(X_2-X_3) & -(X_1-X_3) - X_1(X_2-X_3)e_1 & \Psi^{1233}_{1234} \end{vmatrix} =$$

$$= \begin{vmatrix} \Psi^{1123}_{1234} & X_1(X_1-X_2)e_3 + X_1X_2(X_1-X_3)e_4 & X_1(X_1-X_2)e_4 \\ (X_1-X_2) & \Psi^{1223}_{1234} & X_1X_2(X_2-X_3)e_4 \\ X_1(X_2-X_3) & (X_1-X_3) + X_1(X_2-X_3)e_1 & \Psi^{1233}_{1234} \end{vmatrix} .$$

By using abbreviations $X_{ij} = X_i - X_j$ it can be rewritten as

$$\Delta_4 = \begin{vmatrix} \Psi^{1123}_{1234} & X_1X_{12}e_3 + X_1X_2X_{13}e_4 & X_1X_{12}e_4 \\ X_{12} & \Psi^{1223}_{1234} & X_1X_2X_{23}e_4 \\ X_1X_{23} & X_{13} + X_1X_{23}e_1 & \Psi^{1233}_{1234} \end{vmatrix} .$$

In general

$$\Delta_n = \det(\delta'_{ij})_{1\le i,j\le n-1}$$

where

$$\delta'_{ij} = \begin{cases} \displaystyle\sum_{k=j+1}^{n} X_1 \cdots X_{k+i-j-1}(X_i - X_{k+i-j})e_k \ , \ 1 \le i < j \le n-1 \\ \Psi^{1 \ \dots \ \boldsymbol{i} \ \boldsymbol{i} \ \dots \ n}_{1 \ 2 \ \dots \ n} \qquad\qquad , \ i = j \\ \displaystyle\sum_{k=0}^{j} X_1 \cdots X_{k+i-j-1}(X_{k+i-j} - X_i)e_k \quad , \ 1 \le j < i \le n-1 \end{cases}$$

Note that all δ'_{ij} are symmetric polynomials with non negative coefficients.

Corollary 1.14 *The conjecture 1.11 is equivalent to a Hadamard type inequality, holding coefficientwise (m–positivity, even s–positivity), for the (non Hermitian) matrix* $(\delta'_{ij})_{1\le i,j\le n-1}$*, i.e.*

$$\prod_{i=1}^{n-1} \delta'_{ii} \ge \det(\delta'_{ij}) \, .$$

2 Verification of the Đoković's strengthening of the Atiyah–Sutcliffe Conjecture (C2) for some nonplanar configurations with dihedral symmetry

Here we basically follow Đoković's [7], where he considered the following dihedral configurations $C_{m,n}$ (type D).

Let $N = m + n$ points be such that

1. The first m points $x_1, \ldots, x_m$ lie on a line L.
2. The remaining n points $y_j = x_{m+j+1}$ $(j = 0, 1, ..., n-1)$ are the vertices of a regular n–gon whose plane is perpendicular to L and whose centroid lies on L.

He only proved Atiyah conjecture C1. We make some additional refinements including a proof of Atiyah–Sutcliffe conjecture C2 for such configurations.

We may assume $L = \mathbb{R} \times \{0\} \subset \mathbb{R} \times \mathbb{C} = \mathbb{R}^3$ and write $x_i = (a_i, 0)$, $1 \leq i \leq m$, $a_1 \leq \ldots \leq a_m$ and $y_j = (0, b_j)$, $b_j = -\xi^j$, $\xi = e^{2\pi i/n}$, $0 \leq j \leq n-1$.

We set

$$\lambda_i = a_i + \sqrt{1 + a_i^2}$$

Recall that $a_1 < \cdots < a_m$ and, consequently $0 < \lambda_1 < \cdots < \lambda_m$. Then the associated polynomials p_i (up to scalar factors) are given by

$$p_i(x, y) = x^{m-i} y^{i-1} (x^n - \lambda_i^n y^n),\ 1 \leq i \leq m$$
$$p_{m+j+1}(x, y) = \prod_{s \neq j} \left(x + \frac{\overline{b_s} - \overline{b_j}}{|b_s - b_j|} y \right) \cdot \prod_{i=1}^{m} (y - \lambda_i b_j x), 0 \leq j < n$$

By noting that

$$b_s - b_j = 2i\xi^{\frac{j+s}{2}} \sin \frac{\pi(j-s)}{n}$$

(in Đoković ξ^{j+s} should be replaced by $\xi^{\frac{j+s}{2}}$) we obtain

$$x + \frac{\overline{b_s} - \overline{b_j}}{|b_s - b_j|} y = \left(-\overline{b_j} y - i\xi^{\frac{s-j}{2}} \operatorname{sgn}(s-j) \right) \frac{1 - \overline{b_s} b_j}{|b_s - b_j|}$$

and

$$y - \lambda_i b_j x = -b_j(-\overline{b_j} y + \lambda_i x)$$

Note also that

$$\{\xi^{\frac{s-j}{2}}\mathrm{sgn}(s-j) | s = 1, \ldots, j-1, j+1, \ldots, n\} = \{e^{\pi i k/n} | k = 1, \ldots, n-1\}$$

Thus, after dehomogenizing the polynomials p_i by setting $x = 1$, we obtain (up to scalar factors) the following polynomials:

$$\widetilde{P}_i(y) = y^{i-1}(1 - \lambda_i^n y^n),\ 1 \le i \le m;$$
$$\widetilde{P}_{m+j+1}(y) = f(\xi^{-1} y),\ 0 \le j < n$$

where

$$f(y) = \prod_{s=1}^{n-1}(y - ie^{\pi i s/n}) \prod_{i=1}^{m}(y + \lambda_i)$$

(in Đoković the last n polynomials are reordered)

The main result of Đoković is the Theorem *3.1* where he proved Atiyah conjecture for configurations described above, by explicitly computing the determinant of the coefficients matrix $\widetilde{P}$ of the polynomials $\{\widetilde{p}_k(y) | k = 1, \ldots, \underbrace{m+n}_{N}\}$ in terms of the coefficients of

$$f(y) = \sum_{k=0}^{N-1} \widetilde{E}_k y^{N-1-k}$$

His formula reads as follows:

$$\left|\det(\widetilde{P})\right| = n^{n/2} \prod_{k=0}^{n-1} f_k$$

where

$$f_k = \sum_{s \ge 0} \left(\prod_{j=1}^{s} \lambda_{N-jn-k}^n \right) \widetilde{E}_{k+sn},\ 0 \le k < n.$$

We shall now present an amazingly simple formula for coefficients of the polynomial

$$h(y) := \prod_{s=1}^{n-1}(y - ie^{\pi i s/n}) = \sum_{j=0}^{n-1} c_j y^{n-1-j}$$

Proposition 2.1 *let* $\gamma_k := \cot\left(\frac{k\pi}{2n}\right)$. *Then*

$$c_0 = 1,\ c_j = \prod_{k=1}^{j} \gamma_k \qquad (1 \le j \le n-1)$$

Proof .
Put $\xi_k = -ie^{\pi ik/n}$, $k = 1, \ldots, n-1$. Then

$$\begin{aligned} c_j &= \text{the j–th elementary symmetric function of } \xi_1, \ldots, \xi_{n-1} \\ &= e_j(\xi_1, \ldots, \xi_{n-1}) \end{aligned}$$

Let us first compute the power sums

$$p_s = \sum_{k=1}^{n-1} \xi_k^s = (-i)^s \sum_{k=1}^{n-1} e^{\pi isk/n} = (-i)^s (e^{\pi is/n} - e^{\pi is})/(1 - e^{\pi is})$$
$$= \begin{cases} (-1)^{\frac{s}{2}-1}, s \text{ even} \\ (-1)^{\frac{s-1}{2}} \cot(\frac{s\pi}{2n}) = (-1)^{\frac{s-1}{2}} \gamma_s, s \text{ odd} \end{cases}$$

The proof will be by induction. For $j = 1$ we have $c_1 = \xi_1 + \cdots + \xi_{n-1} = p_1 = \gamma_1$. Suppose that the proposition is true for all $k < i$. Then by Newton formula for symmetric functions

$$je_j = \sum_{k=1}^{j} (-1)^{k-1} p_k e_{j-k} = \sum_{l=1}^{\lceil j/2 \rceil} (p_{2l-1} e_{j-2l+1} - p_{2l} e_{j-2l})$$

we obtain by writing $c_{j-2l+1} = c_{j-2l}\gamma_{j-2l+1}$

$$\begin{aligned}
je_j &= \sum_{l=1}^{\lceil j/2\rceil} \left((-1)^{l-1}\gamma_{2l-1}\gamma_{j-2l+1} - (-1)^{l-1}\right) c_{j-2l} \\
&= \sum_{l=1}^{\lceil j/2\rceil} (-1)^{l-1}(\gamma_{2l-1}\gamma_{j-2l+1} - 1)c_{j-2l} \\
&\overset{*}{=} \sum_{l=1}^{\lceil j/2\rceil} (-1)^{l-1}(\gamma_{2l-1} + \gamma_{j-2l+1})\gamma_j c_{j-2l} \\
&= \sum_{l=1}^{\lceil j/2\rceil} (p_{2l-1}c_{j-2l} - p_{2l-2}\gamma_{j-2l+1}c_{j-2l})\gamma_j \quad (\text{here } p_0 := -1) \\
&= \sum_{l=1}^{\lceil j/2\rceil} (p_{2l-1}c_{j-2l} - p_{2l-2}c_{j-2l+1})\gamma_j \\
&= \sum_{l=1}^{\lceil j/2\rceil} (p_{2l-1}c_{j-1-(2l-1)} - p_{2l-2}c_{j-1-(2l-2)})\gamma_j \\
&= (-p_0 c_{j-1} + \sum_{l=1}^{\lceil (j-1)/2\rceil} (p_{2l-1}c_{j-1-(2l-1)} - p_{2l}c_{j-1-2l)})\gamma_j \\
&\overset{**}{=} (c_{j-1} + (j-1)c_{j-1})\gamma_j \\
&= jc_{j-1}\gamma_j = jc_j
\end{aligned}$$

Here in $(*)$ we have used the cotangent addition formula $\cot(\alpha)\cot(\beta) - 1 = (\cot\alpha + \cot\beta)\cot(\alpha+\beta)$ and in $(**)$ Newton formula for $i-1$ which holds by induction hypothesis. The proposition is thus proved. ∎

For our dihedral configurations we can state the stronger conjecture of Atiyah and Sutcliffe ([7], Conjecture *2.*) as follows

$$n^{\frac{n}{2}} \prod_{k=0}^{n-1} f_k \geq 2^{\binom{n}{2}} \prod_{i=0}^{n} (1+\lambda_i^2)^n \tag{2.10}$$

where

$$f_k = \sum_{s\geq 0} \left(\prod_{j=1}^{s} \lambda_{N-jn-k}^n \widetilde{E}_{k+sn}, \quad (0 \leq k < n) \right) \tag{2.11}$$

From the factorization

$$f(y) = h(y)\prod_{i=1}^{m}(y+\lambda_i)$$

we can write

$$\widetilde{E}_k = \sum_{i=0}^{n-1} c_i E_{k-i}$$

in terms of elementary symmetric functions $E_k = e_k(\lambda_1, \ldots, \lambda_m)$ of our positive quantities $0 < \lambda_1 < \cdots < \lambda_m$ with coefficients c_i given in Proposition 2.1 (note that $c_0 = 1 \leq c_1 \leq \cdots \leq c_{\lfloor \frac{n-1}{2} \rfloor} \geq \cdots \geq c_{n-1} = 1$ (unimodality) and $c_i = c_{n-1-i}$ (symmetry)).

Now we shall prove a generalization of the Đoković's conjecture which apparently strengthens (2.10).

Theorem 2.2 *We have:*

1. $$\prod_{k=0}^{n-1} f_k \geq \prod_{k=0}^{n-1} c_k \left(\sum_{l=0}^{m} \left(\prod_{j=0}^{l-1} \lambda_{m-j} E_l \right) \right)^n$$

2. $$\prod_{k=0}^{n-1} f_k \geq \prod_{k=0}^{n-1} c_k \prod_{i=1}^{m} (1+\lambda_i^2)^n$$

Proof .
Let us write

$$f_k = \sum_{l=0}^{m} \varphi_{kl} E_l$$

Let us substitute $\widetilde{E}_{k+sn} = \sum_{i=0}^{n-1} c_i E_{k-i+sn}$ into (2.11). Then for fixed k $(0 \leq k < n-1)$ and given l $(0 \leq l \leq m)$ we seek $s \geq 0$ and i, $0 \leq i < n$ such that $l = k-1+sn$, i.e. $l-k = sn-i$, $0 \leq i < n$. We conclude that s and i are uniquely determined by a division algorithm (with nonpositive remainder):

$$s_k := \left\lceil \frac{l-k}{n} \right\rceil, \quad i_k = s_k n - l - k.$$

Hence

$$\varphi_{kl} = \prod_{j=1}^{s_k} \lambda^n_{N-jn-k} c_{i_k}$$

with s_k and i_k just defined. It is easy to see that

$$s_k = s_0 \left(= \left\lceil \frac{l}{n} \right\rceil \right) \text{ and } i_k = i_0 + k \text{ for } 0 \le k \le n - i_0 - 1$$

and

$$s_k = s_0 - 1 \text{ and } i_k = i_0 + k - n \text{ for } n - i_0 \le k \le n - 1.$$

Lemma 2.3 *For each l, $0 \le l \le m$, we have*

$$\prod_{k=0}^{n-1} \varphi_{kl} = \prod_{j=0}^{l-1} \lambda^n_{m-j} \prod_{j=0}^{n-1} c_j$$

Proof (of Lemma).

$$\begin{aligned}\prod_{k=0}^{n-1} \varphi_{kl} &= \prod_{k=0}^{n-i_0-1} \left(\prod_{j=1}^{s_0} \lambda^n_{N-jn-k} \prod_{k=i_0}^{n-1} c_k \right) \prod_{k=n-i_0}^{n-1} \prod_{j=1}^{s_0-1} \lambda^n_{N-jn-k} \prod_{k=0}^{i_0-1} c_k \\ &= \prod_{k=0}^{n-1} \prod_{j=1}^{s_0-1} \lambda^n_{N-jn-k} \prod_{k=0}^{n-i_0-1} \lambda^n_{N-s_0n-k} \prod_{k=0}^{n-1} c_k\end{aligned}$$

We put now $N = n + m$

$$\begin{aligned}&= \lambda^n_m \lambda^n_{m-1} \cdots \lambda^n_{m+n-s_0n-(n-i_0-1)} \prod_{k=0}^{n-1} c_k \\ &= \lambda^n_m \lambda^n_{m-1} \cdots \lambda^n_{m-l+1} \prod_{k=0}^{n-1} c_k\end{aligned}$$

■

Proof (of Theorem).

We shall use the Hölder inequality

$$\begin{aligned}\prod_{k=0}^{n-1} f_k = \prod_{k=0}^{n-1}\left(\sum_{l=0}^{m}\varphi_{kl}E_l\right) \geq & \left(\sum_{l=0}^{m}\left(\prod_{k=1}^{n-1}\varphi_{kl}E_l\right)^{\frac{1}{n}}\right)^n \\ = & \left(\sum_{l=0}^{m}\prod_{j=0}^{l-1}\lambda_{m-j}\left(\prod_{j=0}^{n-1}c_j\right)^{\frac{1}{n}}E_l\right)^n \quad \text{(by lemma)} \\ = & \left(\prod_{j=0}^{n-1}c_j\right)\left(\sum_{l=0}^{m}\prod_{j=0}^{l-1}\lambda_{m-j}E_l\right)^n\end{aligned}$$

Thus *1.* is proved. To obtain *2.* we apply Đoković proof of Atiyah conjecture for type A configurations

$$\sum_{l=0}^{m}\prod_{j=0}^{l-1}\lambda_{m-j}E_l \geq \prod_{i=1}^{m}\left(1+\lambda_i^2\right)$$

(c.f. section 3.) ∎

3 Appendix

After the first version of this paper was finished, in the meantime, we have discovered formulas for the partial derivatives, of the quantities $\Psi^{1\ldots n}_{1\ldots n}/\Psi^{1\ldots\widehat{k}\ldots n}_{1\ldots\widehat{k}\ldots n}$, with respect to variables ξ_r (Note that in Theorem 1.8 we have given formulas w.r.t. variables X_r!).

Lemma 3.1 *For $2 \leq r \leq n$ the partial derivative w.r.t. ξ_r of the quotient $\Psi^{1\ldots n}_{1\ldots n}/\Psi^{2\ldots n}_{2\ldots n}$ is given by*

$$\left(\Psi^{2\ldots n}_{2\ldots n}\right)^2\partial_{\xi_r}\left(\frac{\Psi^{1\ldots n}_{1\ldots n}}{\Psi^{2\ldots n}_{2\ldots n}}\right) = \sum_{i\geq j}s'_{ij}X_1(X_2\cdots X_j)^2X_{j+1}\cdots X_{i+1}(X_{j+1}-X_{i+2})$$

where s'_{ij} is the conjugated Schur function $s_{ij}=s_{ij}(\xi_2,\ldots,\xi_{r-1},\xi_{r+1},\ldots,\xi_n)$ corresponding to a two–rowed partition $\lambda=(i\geq j)$.
In particular for $X_1\geq\cdots\geq X_n>0$ the function $\Psi^{1\ldots n}_{1\ldots n}/\Psi^{2\ldots n}_{2\ldots n}$ is monotonically increasing w.r.t. the variable ξ_r (for $r=1$, too).

Proof .
By using the formula $\Psi^{1\dots n}_{1\dots n} = \Psi^{1\dots n-1}_{1\dots \widehat{r}\dots n} + X_1\xi_r\Psi^{2\dots n}_{1\dots \widehat{r}\dots n}$ we get

$$\begin{aligned}
&\partial_{\xi_r}(\Psi^{1\dots n}_{1\dots n})\Psi^{2\dots n}_{2\dots n} - \Psi^{1\dots n}_{1\dots n}\partial_{\xi_r}(\Psi^{2\dots n}_{2\dots n}) = \\
&= X_1\Psi^{2\dots n}_{1\dots \widehat{r}\dots n}\left(\Psi^{2\dots n-1}_{2\dots \widehat{r}\dots n} + X_2\xi_r\Psi^{3\dots n}_{2\dots \widehat{r}\dots n}\right) - \left(\Psi^{1\dots n-1}_{1\dots \widehat{r}\dots n} + X_1\xi_r\Psi^{2\dots n}_{1\dots \widehat{r}\dots n}\right)X_2\Psi^{3\dots n}_{2\dots \widehat{r}\dots n} \\
&= X_1\Psi^{2\dots n}_{1\dots \widehat{r}\dots n}\Psi^{2\dots n-1}_{2\dots \widehat{r}\dots n} - X_2\Psi^{1\dots n-1}_{1\dots \widehat{r}\dots n}\Psi^{3\dots n}_{2\dots \widehat{r}\dots n} \\
&= X_1\left(\Psi^{2\dots n-1}_{2\dots \widehat{r}\dots n} + X_2\xi_1\Psi^{3\dots n}_{2\dots \widehat{r}\dots n}\right)\Psi^{2\dots n-1}_{2\dots \widehat{r}\dots n} - X_2\left(\Psi^{1\dots n-2}_{2\dots \widehat{r}\dots n} + X_1\xi_1\Psi^{2\dots n-1}_{2\dots \widehat{r}\dots n}\right)\Psi^{3\dots n}_{2\dots \widehat{r}\dots n} \\
&= X_1\left(\Psi^{2\dots n-2}_{2\dots \widehat{r}\dots n}\right)^2 - X_2\Psi^{1\dots n-2}_{2\dots \widehat{r}\dots n}\Psi^{3\dots n}_{2\dots \widehat{r}\dots n}
\end{aligned}$$

With $e_i = e_i^{(1r)} = e_i(\xi_2,\dots,\xi_{r-1},\xi_{r+1},\dots,\xi_n)$ denoting the i–th elementary symmetric function of the truncated alphabet $A^{(1r)} = \{\xi_2,\dots,\xi_{r-1},\xi_{r+1},\dots,\xi_n\}$ we have further

$$\begin{aligned}
&= X_1\left(\sum_{i,j} e_ie_jX_{2\dots i+1}X_{2\dots j+1}\right) - X_2\left(\sum_{i,j} e_ie_jX_{1\dots i}X_{3\dots j+2}\right) \\
&= \sum_{i,j} e_ie_jX_{1..i+1}X_{2..j+1} - \sum_{i,j} e_ie_jX_{1..i}X_{2..j+2} \\
&= \sum_{i,j}\begin{vmatrix} e_i & e_{i+1} \\ e_{j-1} & e_j \end{vmatrix} X_{1..i+1}X_{2..j+1} \\
&= \sum_{i\geq j}\begin{vmatrix} e_i & e_{i+1} \\ e_{j-1} & e_j \end{vmatrix} X_1(X_{2..j})^2X_{j+1}\cdots X_{i+1}(X_{j+1}-X_{i+2})
\end{aligned}$$

Now by Jacobi–Trudy formula we can write $\begin{vmatrix} e_i & e_{i+1} \\ e_{j-1} & e_j \end{vmatrix}$ as the conjugated Schur function $s'_{ij} = s'^{(1r)}_{ij}$ corresponding to a partition $(i \geq j)$. ∎

Corollary 3.2 *(ξ_n–monotonicity)*
We have the following inequality:

$$\frac{\Psi^{1\dots n}_{1\dots n}}{\Psi^{2\dots n}_{2\dots n}} \geq \frac{\Psi^{1\dots n-1}_{1\dots n-1}}{\Psi^{2\dots n-1}_{2\dots n-1}}$$

Proof .
By Lemma 3.1 by letting $\xi_n \downarrow 0$ we get

$$\Psi^{1\dots n}_{1\dots n}/\Psi^{2\dots n}_{2\dots n} \geq \Psi^{1\dots n}_{1\dots n}/\Psi^{2\dots n}_{2\dots n}\big|_{\xi_n=0} = \Psi^{1\dots n-1}_{1\dots n-1}/\Psi^{2\dots n-1}_{2\dots n-1}$$

■

By using this Corollary we state a strengthening of our Conjecture 1.5:

Conjecture 3.3

$$\left(\Psi^{1\dots n}_{1\dots n}\right)^{n-2} \geq \Psi^{2\dots n-1}_{2\dots n-2}\prod_{k=2}^{n-1}\Psi^{1\dots\widehat{k}\dots n}_{1\dots\widehat{k}\dots n}$$

We also have formulas for partial derivative of the quotient $\Psi^{1\dots n}_{1\dots n}/\Psi^{1\dots\widehat{k}\dots n}_{1\dots\widehat{k}\dots n}$ w.r.t. variable ξ_r, $2\leq r\leq n$, which are more complicated than for $k=1$ (given in Lemma 3.1). Without loss of generality we take $r=n$ and proceed as follows:

$$\begin{aligned}
&\partial_{\xi_n}(\Psi^{1\dots n}_{1\dots n})\Psi^{1\dots\widehat{k}\dots n}_{1\dots\widehat{k}\dots n}-\Psi^{1\dots n}_{1\dots n}\partial_{\xi_n}(\Psi^{1\dots\widehat{k}\dots n}_{1\dots\widehat{k}\dots n})=\\
&=X_1\Psi^{2\dots n}_{1\dots n-1}\Psi^{1\dots\widehat{k}\dots n}_{1\dots\widehat{k}\dots n}-X_1\Psi^{1\dots n}_{1\dots n}\Psi^{2\dots\widehat{k}\dots n}_{1\dots\widehat{k}\dots n-1}\\
&=X_1\Psi^{2\dots n}_{1\dots n-1}\left(\Psi^{1\dots\widehat{k}\dots n}_{1\dots\widehat{k}\dots n}+X_1\xi_n\Psi^{2\dots\widehat{k}\dots n}_{1\dots\widehat{k}\dots n-1}\right)-X_1\left(\Psi^{1\dots n-1}_{1\dots n-1}+X_1\xi_n\Psi^{2\dots n}_{1\dots n-1}\right)\Psi^{2\dots\widehat{k}\dots n}_{1\dots\widehat{k}\dots n-1}\\
&=X_1\left(\Psi^{2\dots n}_{1\dots n-1}\Psi^{1\dots\widehat{k}\dots n-1}_{1\dots\widehat{k}\dots n-1}-\Psi^{1\dots n-1}_{1\dots n-1}\Psi^{2\dots\widehat{k}\dots n}_{1\dots\widehat{k}\dots n-1}\right)\\
&=X_1\left[\left(\Psi^{2\dots n-1}_{1\dots\widehat{k}\dots n-1}+X_2\xi_k\Psi^{3\dots n}_{1\dots\widehat{k}\dots n-1}\right)\Psi^{1\dots\widehat{k}\dots n-1}_{1\dots\widehat{k}\dots n-1}-\right.\\
&\qquad\left.-\left(\Psi^{1\dots n-2}_{1\dots\widehat{k}\dots n-1}+X_1\xi_k\Psi^{2\dots n-1}_{1\dots\widehat{k}\dots n-1}\right)\Psi^{2\dots\widehat{k}\dots n}_{1\dots\widehat{k}\dots n-1}\right]\\
&=X_1\left[\Psi^{2\dots n-1}_{1\dots\widehat{k}\dots n-1}\Psi^{1\dots\widehat{k}\dots n-1}_{1\dots\widehat{k}\dots n-1}-\Psi^{1\dots n-2}_{1\dots\widehat{k}\dots n-1}\Psi^{2\dots\widehat{k}\dots n}_{1\dots\widehat{k}\dots n-1}+\right.\\
&\qquad\left.+\xi_k\left(X_2\Psi^{3\dots n}_{1\dots\widehat{k}\dots n-1}\Psi^{1\dots\widehat{k}\dots n-1}_{1\dots\widehat{k}\dots n-1}-X_1\Psi^{2\dots n-1}_{1\dots\widehat{k}\dots n-1}\Psi^{2\dots\widehat{k}\dots n}_{1\dots\widehat{k}\dots n-1}\right)\right]\\
&=X_1\left[I_1-\xi_kI_2\right]
\end{aligned}$$

Now we first compute

$$I_1 = \Psi^{2\ldots n-1}_{1\ldots\widehat{k}\ldots n-1}\Psi^{1\ldots\widehat{k}\ldots n-1}_{1\ldots\widehat{k}\ldots n-1} - \Psi^{1\ldots n-2}_{1\ldots\widehat{k}\ldots n-1}\Psi^{2\ldots\widehat{k}\ldots n}_{1\ldots\widehat{k}\ldots n-1} =$$
$$\left(\sum_{i=0}^{k-2} e_i X_{2..i+1} + \sum_{i=k-1}^{n-2} e_i X_{2..i+1}\right)\left(\sum_{j=0}^{k-1} e_j X_{1..j} + \sum_{j=k}^{n-2} e_j X_{1..\widehat{k}..j+1}\right) -$$
$$-\left(\sum_{j=0}^{k-1} e_j X_{1..j} + \sum_{j=k}^{n-2} e_j X_{1..j}\right)\left(\sum_{i=0}^{k-2} e_i X_{2..i+1} + \sum_{i=k-1}^{n-2} e_i X_{2..\widehat{k}..i+2}\right) =$$
$$= \sum_{i=k-1}^{n-2}\sum_{j=0}^{k-1} e_i e_j \left(X_{2..i+1}X_{1..j} - X_{2..\widehat{k}..i+1}X_{1..j}\right) +$$
$$+\sum_{j=k}^{n-2}\sum_{i=0}^{k-2} e_j e_i \left(X_{1..\widehat{k}..j+1}X_{2..i+1} - X_{1..j}X_{2..i+1}\right) +$$
$$+\sum_{i=k-1}^{n-2}\sum_{j=k}^{n-2} e_i e_j \left(X_{2..i+1}X_{1..\widehat{k}..j+1} - X_{1..j}X_{2..\widehat{k}..i+2}\right)$$

By replacing, in the middle sum, j with $i+1$ and i with $j-1$, and observing that then $X_{1..\widehat{k}..i+2}X_{2..j} - X_{1..i+1}X_{2..j} = -(X_{2..i+1}X_{1..j} - X_{2..\widehat{k}..i+2}X_{1..j})$ the contribution of the first two sums is

$$\sum_{i=k-1}^{n-2}\sum_{j=0}^{k-1}\begin{vmatrix} e_i & e_{i+1} \\ e_{j-1} & e_j \end{vmatrix} X_{2..\widehat{k}..i+1}(X_k - X_{i+2})X_{1..j}$$

The third sum can similarly be transformed to the following form:

$$\sum_{k\le j\le i\le n-2}\begin{vmatrix} e_i & e_{i+1} \\ e_{j-1} & e_j \end{vmatrix} X_{2..\widehat{k}..i+1}(X_{j+1} - X_{i+2})X_{1..j}$$

Hence

$$I_1 = \sum_{0\le j,\max\{j,k-1\}\le i\le n-2} s'_{ij} X_{2..\widehat{k}..i+1}(X_{\max\{j+1,k\}} - X_{i+2})X_{1..j} \quad (\ge 0)$$

By a similar manipulation we can obtain the expression for the quantity

$$I_2 = X_1\Psi^{2\ldots n-1}_{1\ldots\widehat{k}\ldots n-1}\Psi^{2\ldots\widehat{k}\ldots n}_{1\ldots\widehat{k}\ldots n-1} - X_2\Psi^{3\ldots n}_{1\ldots\widehat{k}\ldots n-1}\Psi^{1\ldots\widehat{k}\ldots n-1}_{1\ldots\widehat{k}\ldots n-1} =$$
$$= X_1 - X_2 + \sum_{i=1}^{n-1}\sum_{j\le min\{k-1,i\}} s'_{ij} X_{2..\widehat{k}..i+2}X_{1..j}(X_{j+1} - X_k) \ge 0$$

where s'_{ij} is conjugated Schur function $s'_{ij} = s'^{(kn)}_{ij}$. We see that

$$\left(\Psi^{1\ldots\widehat{k}\ldots n}_{1\ldots\widehat{k}\ldots n}\right)^2 \partial_{\xi_n}\left(\frac{\Psi^{1\ldots n}_{1\ldots n}}{\Psi^{1\ldots\widehat{k}\ldots n}_{1\ldots\widehat{k}\ldots n}}\right) = X_1\left[I_1 - \xi_k I_2\right]$$

has both positive and negative terms. And we have not been able to apply it so far.

Now we illustrate use of ξ–monotonicity (in addition to X–monotonicity) for proving once more the case $n = 4$ of our Conjecture 1.5:

$$\begin{aligned}
&\frac{(\Psi^{1234}_{1234})^3}{\Psi^{234}_{234}\Psi^{134}_{134}\Psi^{124}_{124}\Psi^{123}_{123}} = \frac{\Psi^{1234}_{1234}}{\Psi^{234}_{234}\Psi^{123}_{123}}\frac{\Psi^{1234}_{1234}}{\Psi^{134}_{134}}\frac{\Psi^{1234}_{1234}}{\Psi^{124}_{124}} \geq \ (by\ \xi_4\text{–monotonicity})\\
&\geq \frac{1}{\Psi^{23}_{23}}\frac{\Psi^{1234}_{1234}}{\Psi^{134}_{134}}\frac{\Psi^{1234}_{1234}}{\Psi^{124}_{124}} \geq \ (by\ X_1\text{–monotonicity twice and } X_4\text{–monotonicity})\\
&\geq \frac{1}{\Psi^{23}_{23}}\frac{\Psi^{2234}_{1243}}{\Psi^{234}_{143}}\frac{\Psi^{2233}_{1234}}{\Psi^{223}_{124}} \geq \ (by\ \xi_3\text{–monotonicity})\\
&\geq \frac{1}{\Psi^{23}_{23}}\frac{\Psi^{223}_{124}}{\Psi^{23}_{14}}\frac{\Psi^{2233}_{1234}}{\Psi^{223}_{124}} = \frac{\Psi^{2233}_{1234}}{\Psi^{23}_{23}\Psi^{23}_{14}} \geq 1
\end{aligned}$$

Similarly the cases $n = 5, 6, 7$ of Conjecture 1.5 would be, by using ξ–monotonicity and X–monotonicity, consequences of the following inequalities

$$\widetilde{Q}_n \geq 1$$

where

$$\begin{aligned}
\widetilde{Q}_5 &= \Psi^{22344}_{12345}\Psi^{22344}_{12345}/\Psi^{234}_{234}\Psi^{234}_{135}\Psi^{2244}_{1245}\\
\widetilde{Q}_6 &= \Psi^{223445}_{123456}\Psi^{233455}_{123456}/\Psi^{2345}_{2345}\Psi^{2345}_{1346}\Psi^{2345}_{1256}\\
\widetilde{Q}_7 &= \Psi^{2234556}_{1234567}\Psi^{2334566}_{1234567}\Psi^{2344566}_{1234567}/\Psi^{23456}_{23456}\Psi^{23456}_{13457}\Psi^{23456}_{12467}\Psi^{234566}_{123567}
\end{aligned}$$

3.1 Computer verification of the Conjecture 1.5 (and hence of the Atiyah–Sutcliffe conjecture C3) for almost collinear $9+1$ configuration.

Let us now explain our computer verification of the inequality $\widetilde{Q}_9 \geq 1$ where

$$\widetilde{Q}_9 = \frac{\Psi^{223456778}_{123456789}\Psi^{233456788}_{123456789}\Psi^{223456678}_{123456789}\Psi^{234456788}_{123456789}}{\Psi^{2345678}_{2345678}\Psi^{2345678}_{1345679}\Psi^{2345678}_{1245689}\Psi^{2345678}_{1235789}\Psi^{22346788}_{12346789}}$$

which refines the case $n = 9$ of the Conjecture 1.5. We have observed first that $\widetilde{Q}_9$ is symmetric in partial alphabets

$$A_1 = \{\xi_1, \xi_2, \xi_8, \xi_9\},\ A_2 = \{\xi_3, \xi_4, \xi_6, \xi_7\},\ A_3 = \{\xi_5\}$$

then by introducing the elementary symmetric functions $\{e_1, e_2, e_3, e_4\}$ of A_1 and $\{f_1, f_2, f_3, f_4\}$ of A_2 we first computed the products

$$\Psi^{2345678}_{2345678}\Psi^{2345678}_{1345679} \text{ and } \Psi^{2345678}_{1245689}\Psi^{2345678}_{1235789}$$

in terms of $\{e_1, e_2, e_3, e_4, f_1, f_2, f_3, f_4, \xi_5\}$. Then by successive application of Stembridge's `Maple SF` package we expressed the difference $\Delta := numer(\widetilde{Q}_9) - denom(\widetilde{Q}_9)$ in terms of the Schur functions of both alphabets A_1 and A_2. Then we factored each coefficient in such a multi–Schur expansion and into non-monomial factors we substituted $X_2 = X_3 + h_2$, $X_3 = X_4 + h_3$, ..., $X_7 = X_8 + h_7$. Then the computation showed that the coefficients of all monomials in $X_8, h_2, \ldots, h_7$ were nonnegative. The factoring out the trivial monomial factors in $X_2, \ldots, X_8$ (which are trivially nonnegative) was crucial because otherwise the expansion of multi–Schur function coefficients in terms of increments $h_2, \ldots, h_7$ may not be feasible.

References

[1] M. Atiyah. The geometry of classical particles. *Surveys in Differential Geometry* (International Press) **7** (2001).

[2] M. Atiyah. Configurations of points. *Phil. Trans. R. Soc. Lond. A* **359** (2001), 1375–1387.

[3] M. Atiyah and P. Sutcliffe. Polyhedra in Physics, Chemistry and Geometry, *Milan Journal of Mathematics*, Vol 71, Number 1/September 2003, 33–58

[4] M. Atiyah and P. Sutcliffe. The geometry of point particles. *Royal Society of London Proceedings Series A*, vol. 458, Issue 2021., 1089.–1116.

[5] M. Eastwood and P. Norbury, A proof of Atiyah's conjecture on configurations of four points in Euclidean three-space. *Geometry & Topology* **5** (2001), 885–893.

[6] D.Ž. Đoković, Proof of Atiyah's conjecture for two special types of configurations, arXiv:math.GT/0205221 v4, 11 June 2002. *Electron. J. Linear Algebra* **9** (2002), 132–137.

[7] D.Ž. Đoković, D. Ž., Verification of Atiyah's conjecture for some nonplanar configurations with dihedral symmetry.In *Publ. Inst. Math., Nouv. Ser.* 72(86), (2002) 23–28.

[8] I. G. Macdonald *Symmetric functions and Hall polynomials* 2nd edition, Oxford University Press, 1995.

[9] Joseph Malkoun, On the Atiyah problem on Hiperbolic Configurations of Four Points, arXive:150201364, 2015.; Geometrie Dedicata 2016., Vol 180, pp 287-292

[10] Mazen Bou Khusan, Michael Johnson On the conjecture regarding the four-point Atiyah determinant, SIGMA 10 (2014.), 070, 9 pages

[11] D. Svrtan, I. Urbiha Atiyah-Sutcliffe Conjectures for Almost Collinear Configurations and Some New Conjectures for Symmetric Functions, arXiv:math/0406386 (23 pages)

[12] D. Svrtan, I. Urbiha Verification and Strengthening of the Atiyah-Sutcliffe Conjectures for several types of Configurations, arXiv:math/0609174 (49 pages)

[13] D. Svrtan, Intrinsic Formula for Five Point (in euclidean plane) Atiyah Determinant, https://www.bib.irb.hr/553790 (2010)

[14] D. Svrtan, A proof of All three Euclidean Four Point Atiyah-Sutcliffe Conjectures, https://emis.de/journals/SLC/wpapers/s73vortrag/svrtan.pdf

[15] D. Svrtan, A proof of All three Atiyah-Sutcliffe 4-point Conjectures, International Congress of Mathematicians, Seoul 2014, Abstracts Short Communication Poster Sessions, 136-137, S. Korea

[16] D. Svrtan, A progress on Atiyah-Sutcliffe geometric conjectures, Math-Chem-Comp, 2007., Dubrovnik, June 11.-16.

Two inequalities: a geometric and a combinatorial

Darko Veljan
Department of Mathematics, University of Zagreb,
Bijenička cesta 30, 10000 Zagreb, Croatia,
darko.veljan@gmail.com

Abstract

We present two interesting inequalities: one geometric and one combinatorial. The geometric one involves symmetric functions of side lengths of a triangle. It simultaneously improves Euler's inequality and isoperimetric inequality for triangles and has non-Euclidean versions. As a consequence, in combinatorics we apply it to degenerate (Fibonacci) triangles. We discuss similar inequalities for simplices in higher dimensions. The combinatorial inequality deals with the following question. What is more probable among maps: an injection or a surjection? For maps between finite sets, the answer is surjection. We present several proofs and provide a brief discussion on open problems for continuous maps for metric and other spaces.

Keywords: triangle inequality, tetrahedron and volume inequality, Euler's inequality in 2D and 3D, combinatorial inequality, injective proof

MSC: 51M04, 51M09, 51M16, 05A20, 60C05

Introduction

The paper has two separate parts. The first part contains sections 1-4, and deals with geometric symmetric functions-inequalities for triangles and simplices; the second part is section 5, devoted to a combinatorial inequality which answers an intriguing question: what is more probable – surjections or injections?

DOI: https://doi.org/10.5592/CO/CCD.2018.11

1 Symmetric functions-inequality for side lengths of a triangle

We shall prove an interesting and somewhat unusual inequality for side lengths of a triangle. It is symmetric in all three sides. Therefore, it can better be comprehended in terms of symmetric functions in three variables.

In standard notations, let a, b and c be side lengths of a triangle (even degenerate) and let $e_1 = a + b + c$, $e_2 = ab + bc + ca$, $e_3 = abc$ be the elementary symmetric functions of a, b and c. Then the following symmetric inequality in a, b, c holds.

Theorem 1.1

$$e_1^6 + 12e_1^3e_3 + 12e_1^2e_2^2 + 36e_3^2 \geq 7e_1^4e_2 + 40e_1e_2e_3. \tag{1}$$

Equality holds if and only if the triangle is equilateral, $a = b = c$.

Proof .
Let S, $2s$, R and r be the area, perimeter, circumradius and inradius, respectively, of the triangle with side lengths a, b and c. Then we have

$$\begin{aligned}\frac{R}{r} &= \frac{abcs}{4S^2} = \frac{abc}{4(s-a)(s-b)(s-c)} = \frac{2abc}{(-a+b+c)(a-b+c)(a+b-c)} \\ &\geq \frac{abc + a^3 + b^3 + c^3}{2abc} \geq 2.\end{aligned} \tag{2}$$

The second inequality in (2) is the AM-GM inequality for a^3, b^3 and c^3, and the first inequality is proved below. Note that it is an improvement of Euler's inequality $R/r \geq 2$ from 1765. For more details see [7], and for more on AM-GM inequality see [8].

To prove the first inequality in (2), let x, y and z be the tangent segments from vertices to the incircle, so $a = y + z$, $b = z + x$, and $c = x + y$. Then it is easy to see that the first inequality in (2) is equivalent to

$$\begin{aligned}2x^2y^2(x-z)(y-z) + 2y^2z^2(y-x)(z-x) + 2z^2x^2(x-y)(z-y) \\ + x^4(y-z)^2 + y^4(z-x)^2 + z^4(x-y)^2 \geq 0.\end{aligned} \tag{3}$$

To prove (3) it suffices to prove that the sum of the first three summands in (3) is non negative. Without loss of generality, we may assume that $x \leq y \leq z$. Then the half of the sum of the first three summands in (3) is greater or equal than

$x^2y^2(x-z)(y-z)+z^2x^2(y-z)(x-y)+z^2x^2(z-y)(x-y)=x^2y^2(x-z)(y-z)$, and this number is greater or equal than 0. This proves (3). The first inequality in (2) has on both sides fractions of symmetric functions in a, b, c. By expressing numerators and denominators in terms of e_1, e_2 and e_3, it is not hard to show that this is equivalent to the inequality (1). Finally, since the equality $R/r=2$ holds if and only if the triangle is equilateral, it follows that equality in (1) holds if and only if $a=b=c$. ∎

We can write (2) also in the form

$$\frac{R}{r}=\frac{(a+b+c)abc}{8S^2}\geq\frac{abc+a^3+b^3+c^3}{2abc}.$$

From Heron's formula

$$16S^2=e_1(4e_1e_2-e_1^3-8e_3),$$

we thus obtain the following inequality equivalent to (1):

$$4S^2(e_1^3-3e_1e_2+4e_3)\leq e_1e_3^2. \tag{4}$$

Equality in (4) again occurs if and only if the triangle is equilateral. The standard isoperimetric inequality for triangles reads as follows

$$S\leq e_1^2\frac{\sqrt{3}}{36}, \tag{5}$$

with equality if and only if the triangle is equilateral. By comparing (4) and (5), we shall show that (4), which is equivalent to (1), in fact improves (5). Namely,

$$4S^2=e_1(4e_1e_2-e_1^3-8e_3)/4\leq e_1e_3^2/(e_1^3-3e_1e_2+4e_3)\leq e_1^4/2^23^3. \tag{6}$$

Here, the equality is Heron's formula, the first inequality is (1), and the second inequality is equivalent to

$$e_1^3(e_1^3+4e_3)\geq 3(e_1^4e_2+36e_3^2).$$

But this follows by applying Newton's inequalities (see [8]), once as $e_1^2\geq 3e_2$, and once as $e_1^3\geq 27e_3$. So, we have proved the following.

Theorem 1.2 *The inequality (1) improves not only Euler's inequality that the circumcircle of a triangle is at least twice longer than its incircle, but also improves the standard isoperimetric inequality (5) for triangles. This improvement becomes equality if and only if the triangle is equilateral.*

We can also give a lower bound for the area S. Again we start with Heron's formula written as

$$S^2 = s\left\{[(s-a)(s-b)(s-c)]^{1/3}\right\}^3,$$

and apply the geometric-harmonic inequalities to the last three factors of the previous expression to obtain

$$S^2 \geq s\left\{\frac{3(s-a)(s-b)(s-c)}{(s-a)(s-b)+(s-b)(s-c)+(s-a)(s-c)}\right\}^3. \tag{7}$$

Now we express both the numerator and denominator in terms of e_1, e_2 and e_3. Then we can summarize inequalities (4), (5), (6) and (7) in the following chain of inequalities.

Theorem 1.3 *The squared area S^2 of a triangle is bounded in terms of e_i's of side lengths as*

$$(27e_1/4)[(4e_1e_2-e_1^3-8e_3)/(4e_2-e_1^2)]^3 \leq 4S^2 \leq e_1e_3^2/(e_1^3-3e_1e_2+4e_3) \leq e_1^4/108.$$

2 Symmetric functions-inequalities for non-Euclidean triangles

The spherical and hyperbolic versions of Euler's inequality $R/r \geq 2$, respectively, are the following inequalities (see [6]):

$$\tan(R)/\tan(r) \geq 2 \quad \text{and} \quad \tanh(R)/\tanh(r) \geq 2. \tag{8}$$

As proved in [2], non-Euclidean Euler's inequalities (8) can be strengthened in a symmetric way via side-lengths, but not in the sense analogous to (2). It seems (2) is too strong in these cases. Still, improvements to non-Euclidean cases can be done by using the following Lemma proved in [2].

Lemma 2.1 *If $f(a,b,c) \geq 0$ is an inequality which holds for all Euclidean triangles with side lengths a, b, c, then $f(s(a),s(b),s(c)) \geq 0$ for all spherical or hyperbolic triangles with side lengths a, b, c, where $s(x) = x/2$ in Euclidean geometry, $s(x) = \sin(x/2)$ in spherical geometry and $s(x) = \sinh(x/2)$ in hyperbolic geometry.*

By using the above Lemma and Theorem 1 we conclude that the following theorem holds true.

Theorem 2.2 *Inequality (1) holds also for non-Euclidean triangles with side lengths a,b and c but with symmetric functions in corresponding quantities $s(a)$, $s(b)$ and $s(c)$ as in Lemma. Equalities hold again if and only if $a = b = c$. These inequalities are simultaneous improvements of Euler's inequalities (8) and isoperimetric inequalities as (5) in both spherical and hyperbolic geometry.*

3 Degenerate triangles and Fibonacci numbers

Theorem 1 and all of its equivalent forms hold also in the case of a degenerate triangle, for instance if $a + b = c$. A natural example of such a degenerate triangle is given by a Fibonacci triple (F_{n-1}, F_n, F_{n+1}). Recall Heron's formula in the form

$$16S^2 = (a^2 + b^2 + c^2)^2 - 2(a^4 + b^4 + c^4).$$

Since degenerate triangle has area $S = 0$, it follows that the equality $a + b = c$ can be replaced by the totally symmetric expression

$$(a^2 + b^2 + c^2)^2 = 2(a^4 + b^4 + c^4), \tag{9}$$

known as *Candido' s identity* (from 1950). In fact, it was first noted just for Fibonacci numbers. However, it is a pure algebraic result which holds in any commutative ring for any of its two elements a, b, and their sum c. This can easily be checked directly.

A little historical remark on Fibonacci numbers is due to M. Bhargava. He said in an interview that they should actually be called Virahanka-Fibonacci numbers after Indian mathematician Virahanka who discovered them back in 7th century, in the sense that F_{n+1} is the number of ways to write n as an ordered sum (composition) of 1's and 2's. On different aspects of Candido's identity see in [9].

Not only that Candido's identity (9) and inequality (1) hold for Fibonacci numbers but they also hold for any combinatorially interesting numbers satisfying recurrence of the form $c = a + b$. Examples include binomial coefficients (due to Pascal's formula), *Padovan numbers* (P_n) defined by $P_0 = P_1 = P_2 = 1$ and $P_{n+2} = P_n + P_{n-1}$, where P_n is the number of ways to write $n+2$ as an ordered sum of 2's and 3's, and many other. The inequality (1) in the degenerate case $a + b = c$ is equivalent to the trivial inequality $a^2b^2 \geq 0$, for any two real numbers a, b, which reveals the fact that a square of any real number is nonnegative and the product of such numbers is also nonnegative.

4 Inequalities for tetrahedra and simplices

Let $T = ABCD$ be a tetrahedron or a 3-dimensional simplex with edge lengths a, b, c, a', b', c', where a, b, c form a triangle, a is opposite to a' etc. Let $V = vol(T)$ be the volume, S the surface area of T, and R, and r, respectively, the circumradius and inradius of T, and C the area of the *Crelle triangle* of T whose side lengths are products aa', bb', cc' of opposite edge lengths of T. The Crelle formula $C = 6RV$, the fact $3V = rS$, and Euler's inequality $R/r \geq 3$ imply

$$54V^2 \leq C \cdot S. \tag{10}$$

By applying the standard isoperimetric inequality to both C and S, from (10) we infer

$$25 \cdot 36V^2 \leq (aa'+bb'cc')^2[(a+b+c)^2+(a+b'+c')^2+(a'+b+c')^2+(a'+b'+c)^2]. \tag{11}$$

Substituting V^2 in (11) in terms of $a, a', ..., c'$, for instance Euler's formula (from 1752):

$$\begin{aligned} 144V^2 = \; & (2a'b'c')^2 \\ & - a'^2(b'^2 + c'^2 - a^2)^2 - b'^2(c'^2 + a'^2 - b^2)^2 - c'^2(a'^2 + b'^2 - c^2)^2 \\ & + (a'^2 + b'^2 - c^2)(b'^2 + c'^2 - a^2)(c'^2 + a'^2 - b^2), \end{aligned}$$

we get an inequality of degree 6 for edge lengths of T, in a way analogous to (1) for a triangle, with equality if T is regular. However, (11) is not as sharp as the inequality

$$72V^2 \leq aa'\, bb'\, cc'$$

with equality if and only if T is regular (see [10],[11]). Inequality (11) can be improved by applying (4) to C^2 and all four face areas S_i^2, with $S = S_1+S_2+S_3+S_4$ and using the arithmetic-quadratic inequality for S_i in the form $S^2 \leq 4(S_1^2 + S_2^2 + S_3^2 + S_4^2)$. The obtained inequality becomes equality if and only if T is equifacial, i.e. $S_1 = S_2 = S_3 = S_4$, or equivalently $a = a'$, $b = b'$, $c = c'$. It is in the range of the known inequality (see [4])

$$72V^2 \leq (aa' + bb' - cc')(bb' + cc' - aa')(cc' + aa' - bb'),$$

with equality if and only if T is equifacial.

Let us briefly explain Euler's inequality $R/r \geq 3$ for a tetrahedron T. Let T' be the tetrahedron whose vertices are centroids of faces of T. Then T' is similar to T

with the similarity coefficient 3. Hence, the circumradius R' of T' is one third of R, that is, $R = 3R'$. But, $R' \geq r$, because the smallest ball that touches all faces of T is just the inscribed ball of T. So, $R \geq 3r$. The equality is attained if and only if T is regular. The same argument works for any n-dimensional simplex (and of course, $R \geq nr$ in that case).

Unlike this simple argument, there is no evident argument for yet another inequality relating R, r and the distance d between the circumcenter and incenter of a tetrahedron. It is known as *Grace-Danielsson's inequality* (from 1949). It reads as follows

$$d^2 + (2r)^2 \leq (R - r)^2.$$

Euler's formula for triangles $d^2 + r^2 = (R - r)^2$ can easily be proved via inversion, but for Grace-Danielsson's inequality, a short elegant proof in the sense of CEEG (Classical Euclidean Elementary Geometry) is still missing. There is a proof using quantum information theory [5] and computer aided proof [3]. Corresponding formulas or inequalities for simplices in higher dimensions are also not known as well as non-Euclidean versions.

Recall that for triangles we found a rational non-constant symmetric function $f(a, b, c)$ such that (see (2)):

$$\frac{R}{r} \geq f(a, b, c) \geq 2,$$

and this essentially refined Euler's inequality. A similar problem is for tetrahedron T: find a non-constant simple enough symmetric function Θ in edge lengths $a, a', ..., c'$ and in aa', bb', cc' such that

$$\frac{R}{r} = \frac{C \cdot S}{18V^2} \geq \Theta \geq 3.$$

We can express C, S and V in terms of edges of T and get the corresponding inequality for edges of T, but this is not quite adequate because we have sum of four square roots, so it is not good in the above sense. In fact, in [7] we proved an interpolation of $(R/r)^2 \geq 9$ in terms of symmetric functions of aa', bb' and cc'.

For general simplices, besides exact volume formulas, the Cayley-Menger formula, some known inequalities may be of interest such as ([10],[11]):

$$(n!V)^2 n^n \leq (n+1)^{n+1} R^{2n}, \qquad (n!V)^2 2^n \leq (n+1)(\prod a_{ij})^{4/(n+1)}. \qquad (12)$$

Both of them attain equality if and only if the simplex is regular. For $n = 2$, the right inequality (12) is not as good as our Theorem 1 (or equivalently, the inequality (4)).

For $n = 3$, we mentioned it earlier. For higher dimensions, no essential improvement of Euler's inequality is known.

5 Are surjections more probable than injections? Yes – for finite sets!

Problem 11957 of the American Mathematical Monthly (AMM) [12] was proposed by É. Pité, Paris, France, in vol. 124, February 2017. It is as follows (a bit rephrased).

Let $n \geq k \geq 1$ be integers. Prove that

$$n^k S(n,k) \geq k^n \binom{n}{k}, \tag{13}$$

where $S(n,k)$ is the Stirling number of the second kind, equal to the number of partitions of an n-set into k blocks (nonempty subsets).

We shall survey several proofs of (13) and provide some comments. Two solutions appeared in November 2018 issue of AMM ([13]) and one solution was published online earlier. This earlier solution by R. Tauraso, Rome, Italy, is by double induction on $n \geq k \geq 1$ using some known identities involving Stirling numbers and binomial coefficients and using convexity proving that the second derivative of the function $f_k(x) = (1 + x/k)^{-k}$ is positive for $x > 0$. We skip this proof and concentrate on combinatorial proofs.

But first some comments. For $k = 1$ or $k = n$, the inequality (13) becomes equality, as well as for $k > n$, when $S(n,k) = \binom{n}{k} = 0$. For $k = 2, S(n,2) = 2^{n-1} - 1$, so (13) reduces to the well known inequality $2^{n-1} \geq n$. For $k = n - 1$, since $S(n, n-1) = \binom{n}{2}$, (13) reduces to the also well known inequality $n^{n-1} \geq 2(n-1)^{n-1}$.

In fact, (13) is not very surprising, because $S(n,k)$ for large n behaves as $k^n/k!$, so (13) is then of the type

$$n^k \geq n^{\underline{k}} = n(n-1)(n-2)...(n-k+1),$$

which is obviously true.

Let us first prove a weaker inequality (interesting by itself):

$$n^k S(n,k) \geq k^{n-k} \binom{n}{k}. \tag{14}$$

Proof was given by M. Wildon, London, UK, in mathoverflow on 4 July 2017. (Recall, $[n] = \{1, 2, ..., n\}$.)

The right-hand side of (14) counts the set

$$Y = \{(K, f) | K \subseteq [n] \text{ is a } k - \text{set, and } f : [n] \backslash K \to K \text{ is a function}\},$$

while the left-hand side of (14) counts the set

$$X = \{(P, g) | P \text{ is a } k - \text{partition of } [n], \text{ and } g : P \to [n] \text{ a function}\}.$$

Define the function $h : Y \to X$ by $h(K, f) = (P, g)$, where $P = \{\{x\} \cup f^{-1}(x) | x \in K\}$, and $g(x \cup f^{-1}(x)) = x$. Function g essentially picks a point (element) from each block, and serves to uniquely reconstruct the subset K and the function $f : [n] \backslash K \to K$. Hence, h is an injection which proves (14).

One more (even simpler) proof of (14) is this. Let $H : X \to Y$, given by $H(P, g) = (K, f)$ be defined as follows. Let $P = \{B_1, ..., B_k\}$ and $g : P \to [n]$. Then let $K = \{x_1, ..., x_k\}$, $x_i \in B_i, i = 1, ..., k$, be any selection of one element from each block of P and define $f : [n] \backslash K \to K$. Let $x \in [n] \backslash K$, specifically $x \in B_i \backslash x_i$ and assume $g(B_i) \in B_j \subseteq [n]$. Then let $f(x) := x_j$. The map H is clearly surjective, proving (14). Perhaps (13) can also be proved in a similar manner.

We now prove (13). The following elegant proof was given by T. Horine, Indiana Univ. SE, New Albany, IN. To partition the set $[n]$ into k blocks first choose k elements (points) and place one in each block. This choice can be done in $\binom{n}{k}$ ways. The rest, that is $n - k$ elements, can be put to those k blocks in k^{n-k} ways to complete a partition. But a partition with block sizes $s_1, ..., s_k$ has been counted $s_1 s_2 ... s_k$ times, since each block can be initiated by any of its s_i elements. In each case, $s_1 + ... + s_k = n$. By the AM-GM inequality $s_1 s_2 ... s_k \leq (n/k)^k$ for each partition. Hence we have

$$(n/k)^k S(n, k) = [(s_1 + ... + s_k)/k]^k S(n, k) \geq (s_1 ... s_k) S(n, k) \geq \binom{n}{k} k^{n-k}.$$

This proves (13).

A similar proof of (13) was provided by Filip Nikšić (Germany) in mathoverflow on 4 July 2017.

Given a k-partition P of $[n]$ and a k-set S of $[n]$, we say that P splits S if every block of P contains exactly one point of S, that is, $B \cap S \neq \emptyset$ for every $B \in P$. We prove equivalent inequality to (13):

$$k^{n-k} \binom{n}{k} \leq (n/k)^k S(n, k). \tag{15}$$

Let M be a (0,1)-matrix whose rows are indexed by k-partitions P of $[n]$ and columns by k-subsets S of $[n]$. (So, M is of the size $S(n,k) \times \binom{n}{k}$.) The entry $M(P,S) = 1$ if and only if P splits S.

We count the number of ones in two ways. The number of ones in a column indexed by S is the number of partitions that split S. Such a partition is uniquely determined by a map $[n]\backslash S \to S$ that maps $x \in [n]\backslash S$ to $y \in S$ if x and y are in the same block of the partition. Hence the number of ones in the column is k^{n-k} and the total number of ones in M is $k^{n-k}\binom{n}{k}$. On the other hand, the number of ones in a row indexed by $P = \{B_1, ..., B_k\}$ is the number of k-sets split by P. Such a set is uniquely determined by a choice of one element from each block. Hence, the number of ones in the row is $s_1 s_2 ... s_k$, where $s_i = |B_i|$, $i = 1, ..., k$. The total number of ones in M is therefore the sum of such products over all k-partitions of $[n]$. By the AM-GM inequality (as in the previous proof) we have $s_1 s_2 ... s_k \leq (n/k)^k$. Finally, we conclude

$$k^{n-k}\binom{n}{k} = \sum s_1 s_2 ... s_k \leq (n/k)^k S(n,k).$$

This proves (15), and hence (13).

One more combinatorial proof of (13) was provided by M. Wildon. By considering $|\mathrm{Im}(f)| = r$ and $|\mathrm{Im}(g)| = r \leq k$, for various r, (13) reduces to show the following interesting inequality

$$S(n,k)S(k,r) \geq \binom{n-r}{k-r} S(n,r),$$

for all $r \leq k \leq n$.

To prove this inequality, start with given r-partition Q of $[n]$. Let $M(Q)$ be the r-subset of $[n]$ consisting of the largest element in each block of Q. Choose also a set T of $k-r$ elements in the set $[n]\backslash M(Q)$. The right-hand side counts all such pairs (Q,T).

Given such a pair (Q,T), define a k-partition P of $[n]$ that refines Q. Simply extract each element of T from its block in Q and make a new singleton block.

Next consider all pairs (Q,T), where Q is an r-partition and P a k-partition of $[n]$ and P a refinement of Q. We can build such pairs by first choosing P and then grouping the blocks of P into a partition with r blocks. Hence, there are $S(n,k)S(k,r)$ such pairs. So, it suffices to show that the map $(Q,T) \mapsto (Q,P)$ is 1-1.

If (Q,P) arises from (Q,T) by this map, then P has at least $k-r$ singleton blocks. The element x of a singleton block lies in T if and only if $x \notin M(Q)$. Thus

we can uniquely reconstruct (Q, T) from (Q, P) and the map is injective as desired.

Finally, let us explain the title of this paragraph. Namely, the main inequality (13) is equivalent to

$$k!S(n,k)/k^n \geq n^{\underline{k}}/n^k. \tag{16}$$

The right-hand side of (16) is the probability that a uniformly chosen random function $[k] \to [n]$ is injective, while the left-hand side of (16) is the probability that a uniformly chosen random function $[n] \to [k]$ is surjective. So, surjections are more probable than injections among maps between finite sets.

Small numerical examples of (16): for $n = 7$ and $k = 4$, $Prob([7] \to [4]surj) = 0.51269... \geq 0.34985... = Prob([4] \to [7]inj)$, for $n = 4$ and $k = 3$, $Prob([4] \to [3]surj) = 4/9 \approx 0.44 \geq 0.375 = Prob([3] \to [4]inj)$. So, the event that a uniformly chosen random map $[4] \to [3]$ is surjective is not very likely to happen, because its probability is less than 0.5.

Note that for fixed k, both sides of (16) tend to 1, when $n \to \infty$. But for $k = n$, or $k = n - 1$, both sides of (16) tend to 0, when $n \to \infty$.

Finally, let us make some comments and raise some problems. It would be interesting to find a direct probabilistic argument for (16). Next, what is a "continuous" version of (16)?

For instance, if X and Y are compact metric spaces and, say, $Y \subseteq X$, is it more probable that a uniformly and randomly chosen (continuous) map $X \to Y$ is surjective (covering) than a uniformly and randomly chosen (continuous) map $Y \to X$ injective (embedding)? Of course, this requires a choice of an appropriate probability measure (i.e. a positive regular Borel measure m with $m(X) = 1$ and if $f \in L^1(X, m)$, then $m(f) = \int f dm$, and similarly on Y). A related but more complex problem is to find probability distribution of injections among all maps $Y \to X$ (and similarly for surjections among all maps $X \to Y$), or at least find good bounds. We can restrict questions to, say, manifolds (Riemannian, smooth or topological), or generalize further to more general (topological) spaces or even to more general categorical framework. Even in the case $X = Y = [0, 1]$ nothing essential in this respect is known. Another "concrete" examples are: $Y = I = [0, 1]$, and $X = I^3$ (a sort of self-avoiding walk problem) and $Y = S^1$ (circle) and $X = S^3$ (3-sphere), problems in knot theory. Perhaps a good starting point to think about such general problems is Grothendieck's inequality (see e.g. [1]).

References

[1] S. Friedland, L.-H. Lim and J. Zhang, An elementary and unified proof of Grothendieck's inequality, online 29 Dec 2018.

[2] R. Guo, E. Black and C. Smith, Strengthened Euler's inequality in spherical and hyperbolic geometries, *arXiv: 1704.05373 [math.MG]*, 17 Apr 2017.

[3] L. László, On the Grace-Danielsson inequality for tetrahedra, *Discrete Appl. Math.* 256 (2019) 83–90.

[4] M. Mazur, An inequality for the volume of a tetrahedron, *Amer. Math. Monthly* 125 (2018) 273–275.

[5] A. Milne, The Euler and Grace-Danielsson inequalities for nested triangles and tetrahedra: a derivation and generalization using quantum information theory, *J. Geom.* 106 (2015) 455–463.

[6] D. Svrtan and D. Veljan, Non Euclidean versions of some classical triangle inequalities, *Forum Geom.* 12 (2012) 197–209.

[7] D. Veljan, Symmetric functions-inequalities for triangles and simplices: Euler's case, submitted.

[8] D. Veljan, The AM-GM inequality from different viewpoints, *Elem. Math.* 72 (2017) 24–34.

[9] D. Veljan, A note on Candido's identity and Heron's formula, *Proceedings of the 1st Croatian Combinatorial. Days, Zagreb, Sept. 29-30, 2016* (Eds. T. Došlić and I. Martinjak), Fac. Civil Eng., Univ. Zagreb, 2017, pp. 95–105.

[10] D. Veljan, The sine theorem and inequalities for volumes of simplices and determinants, *Lin. Alg. Appl.* 219 (1996) 79–91.

[11] V. Volenec, D. Veljan and J. Pečarić, Inequalities for volumes of simplices in terms of their faces, *Geom. Dedicata* 37 (1998), 57–62.

[12] Problem 11957, *Amer. Math. Monthly* 124 (2017) 179.

[13] Problem 11957, *Amer. Math. Monthly* 125 (2018) 858-859.

Highly resistant multicoloring with 3 attackers and 1 malfunctioning vertex

Tanja Vojković*, Damir Vukičević

Department of Mathematics, Faculty of Science,
Split, Croatia

Abstract

In this paper we explore a way of securing a secret inside a graph by observing pieces of the secret as colors assigned to the graph vertices. If a graph allows a highly (a, b)-resistant k-multicoloring then a secret can be divided into k parts and sets of those parts distributed to the vertices of the graph so that no a attackers can steal the secret, and when a attackers and b malfuntioning vertices leave the graph, the secret is still whole in the remaining graph. In this paper we explore how many vertices a graph must have in order to allow a highly $(3, 1)$-resistant k-multicoloring, and what is the minimal number of colors, for graphs that do allow such multicoloring.

Keywords: graph theory, graph coloring, multicoloring, secret sharing

MSC: 05C82, 05C15, 68R10, 94A62

1 Introduction

In paper Multicoloring of graphs to secure a secret, [7], we were motivated by a problem of securing a secret by dividing it into parts and distributing them to the participants of some network. This is a known method of securing a secret [6]. In our model, there are a attacker vertices in the network, trying to read the secret or disable the group from reading it. Here we make an additional assumption,

*Corresponding author: tanja@pmfst.hr

DOI: https://doi.org/10.5592/CO/CCD.2018.12

that besides the a attacker vertices, there are b vertices that are malfunctioning and leaving the network. The secret is secure if a attackers didn't steal the secret and if the group is still able to read the whole secret after a attacker vertices, their neighbors, and b malfunctioning vertices are removed from the network. We represented the network with graph and parts of the secret with colors assigned to the vertices. Coloring and multicoloring of graphs are often used to model some real-life problem, like scheduling or frequency allocation and there are many new colorings defined with different coloring conditions [2, 3, 4, 5, 9]. This prompted us to define a highly (a, b)-resistant multicoloring with the conditions that make the secret secure.

Our goal is to analyze minimal number of vertices a graph must have to allow a highly (a, b)-resistant multicoloring, for given a and b, and if such a coloring exists for a graph G, to determine what is the minimal number of colors. The results for $a = 1$, $b \in \mathbb{N}$ and $a = 2$, $b = 1, 2, 3$ are presented in paper Highly (a, b)-resistant multicoloring of graphs, [8], which is not yet published at this point. It can be obtained from the authors by request, however, it is not necessary for understanding this paper.

In this manuscript we analyze what graphs will allow a highly (a, b)-resistant multicoloring for $a = 3$ and $b = 1$.

2 Preliminaries

We will mostly use standard definitions and notations of graph theory from [1], and the rest we present in this section. For graph G and $u \in V(G)$, with $N(u) = N_G(u)$ we denote the set of neighbors of u in G, and $M(u) = M_G(u) = N(u) \cup \{u\}$. Naturally, for $A \subseteq V(G)$ we denote:

$$N(A) = N_G(A) = \bigcup_{u \in A} N_G(u);$$

$$M(A) = M_G(A) = \bigcup_{u \in A} M_G(u).$$

First, let us give the formal definition of highly (a, b)-resistant multicoloring.

Definition 1. *Let G be a graph, and $a, b, k \in \mathbb{N}_0$. Vertex k-multicoloring κ of G is called a* ***highly (a, b)-resistant vertex k-multicoloring*** *if for each $A, B \subseteq V(G)$, where $|A| = a$ and $|B| = b$, the following holds:*

1. *There exists a component H of the graph $G\backslash(M_G(A) \cup B)$ such that*

$$\bigcup_{u \in V(H)} \kappa(u) = \{1, ..., k\}.$$

2. $\bigcup_{u \in A} \kappa(u) \neq \{1, ..., k\}$.

We will say that graph G allows a highly (a, b)-resistant multicoloring if a multicoloring function κ exists that is highly (a, b)-resistant. We will denote by $HR_{a,b}(n) = k$ the fact that there exists a graph G with n vertices that allows a highly (a, b)-resistant k-multicoloring, where k is the minimal number of colors needed.

It can be easily proven that if a graph G with n vertices allows a highly (a, b)-resistant k-multicoloring than the same graph allows a highly (a, b)-resistant $(k+1)$-multicoloring, and there exists a graph G' with $n + 1$ vertices that allows a highly (a, b)-resistant k-multicoloring.

In our proofs, we will make use of the notion of an l-separable graph.

Definition 2. *Let G be a graph and $a, b, l \in \mathbb{N}_0$. We will say that G is l-**separable with** (a, b) **vertices**, if subsets $A, B \subseteq V(G)$ exist, with $|A| = a$ and $|B| = b$, such that all the components in graph $G\backslash(M_G(A) \cup B)$ have at most l vertices.*

For a graph to allow a highly $(3, 1)$-resistant multicoloring, no 3 vertices can have all the colors, so we will need the notion of a 3-separable graph.

3 Highly $(3, 1)$-resistant multicoloring

Our main theorem answers two questions:
1. What is the minimal number of vertices a graph must have in order to allow a highly $(3, 1)$-resistant multicoloring?
2. What is the minimal number of colors needed for a highly $(3, 1)$-resistant multicoloring in all the graphs that allow such a coloring?

In order to shorten the proof of the theorem, we will first prove several lemmas.

Lemma 1. *Let G be a graph.*
i) If G has at most 8 vertices it is 3-separable with $(1, 1)$ vertices.
ii) If G has at most 10 vertices it is 3-separable with $(2, 0)$ vertices.
iii) If G has at most 12 vertices it is 3-separable with $(2, 1)$ vertices.
iv) If G has at most 13 vertices it is 3-separable with $(3, 0)$ vertices.

Proof. All the claims will be proven for connected graphs, and from that it easily follows that they also hold for disconnected graphs.

i) Let G be a connected graph with at most 8 vertices. If G contains a vertex u of degree at least 3 the claim is obvious, and if that is not the case then let u be any vertex of degree 2 in G. $G\backslash M(u)$ has at most 5 vertices and it is a union of paths so it is 3-separable with $(0,1)$ vertices.

ii) Let G be a connected graph with at most 10 vertices. If there exists a vertex u of degree at least 3 in G then $G\backslash M(u)$ has at most 6 vertices. If there exists a vertex of degree at least 2 in $G\backslash M(u)$, then $G\backslash M(u)$ is 3-separable with $(1,0)$ vertices, and if all the vertices have the degree at most 1 then all the components already have at most 3 vertices. On the other hand, if all the vertices in G have the degree at most 2, and u is any vertex of degree 2 then $G\backslash M(u)$ has at most 7 vertices and it is a union of paths so it is easily seen that it is 3- separable with $(1,0)$ vertices.

iii) Let G be a connected graph with at most 12 vertices. If there exists a vertex u of degree at least 3 in G then $G\backslash M(u)$ has at most 8 vertices and the claim now follows from i). On the other hand, if the highest degree in G is at most 2 then by observing any vertex u of degree 2, $G\backslash M(u)$ remains with at most 9 vertices and it is a union of paths so the claim again easily follows.

iv) This case is proven in paper [7]. □

Lemma 2. *i) A graph G with 9 vertices, $\Delta(G) \leq 3$ and $\delta(G) = 1$ is 3-separable with $(1,1)$ vertices.*

ii) A graph G with 13 vertices, $\Delta(G) \leq 3$ and $\delta(G) = 1$ is 3-separable with $(2,1)$ vertices.

Proof. i) Let G be a connected graph with 9 vertices, $\Delta(G) \leq 3$ and $\delta(G) = 1$. Let x be a vertex of degree 1 in G, and let us denote its only neighbor by y. If y has the degree 2 then let us denote the other neighbor of y by w. $G\backslash\{w\}$ has 8 vertices and at most 6 vertices in its largest component. If we denote any vertex u of degree at least 2 in that component then $G\backslash(M(u)\cup\{w\})$ is a graph with all components of size at most 3. On the other hand, if y has the degree 3, let us denote its other 2 neighbors by w_1 and w_2. At least one of them has another neighbor, not in $\{x, y, w_1, w_2\}$. Let us assume that w_1 has another neighbor and let us denote it by u. If u has the degree 3 than $G\backslash M(u)$ has 5 vertices and either its largest component has at most 4 vertices (if u was adjacent to w_2), so it is 3-separable with $(0,1)$ vertices, or its largest component has at most 5 vertices and by removing w_2 obtain a graph with all components of size at most 3. If u has the degree 2, then either it is adjacent

to w_2 and then by removing the third neighbor of w_2 and its neighbors, and w_1, we obtain a graph with components of size at most 3, or $G\backslash(M(u) \cup \{w_2\})$ is such a graph. If G is disconnected it can be easily seen that the claim also stands.

ii) Let G be a connected graph with 13 vertices, $\Delta(G) \leq 3$ and $\delta(G) = 1$. Let x be a vertex of degree 1, and y its only neighbor. If y has the degree 2 let us denote by w its other neighbor. $G\backslash\{w\}$ has 1 component of 2 vertices, x and y, and the remaining graph has 10 so it is 3-separable with $(2, 0)$ vertices by Lemma 1 ii). If y has the degree 3 let us denote its other 2 neighbors by w_1 and w_2. Now, if there exists a vertex u in $G\backslash\{y\}$ of degree 3 in G then $G\backslash M(u)$ has 9 vertices, $\Delta(G\backslash M(u)) \leq 3$ and $\delta(G\backslash M(u)) \leq 1$, so the claim follows from i) (if $\delta(G\backslash M(u)) = 1$) or from Lemma 1 i) (if $\delta(G\backslash M(u)) = 0$). If this is not the case then the only vertex with degree 3 in G is y and $G\backslash M(y)$ has 9 vertices in a union of paths and it is easy to see that the claim also follows. If G is disconnected the claim also holds. □

Lemma 3. *Let G be a graph with 9 vertices, $\Delta(G) = 3$ and $\delta(G) \geq 2$. G is 3-separable with $(1, 1)$ vertices.*

Proof. It is easy to see that G cannot be 3-regular, so there is at least 1 vertex of degree 2, let us denote it by x. Let us denote the two neighbors of x by u and v. If u or v have a neighbor of degree 3, let us denote it by y. $G\backslash M(y)$ has 5 vertices and x is of degree 0 or 1 so that graph is 3-separable by the $(0, 1)$ vertices. On the other hand, if neither u nor v have a neighbor of degree 3 then at least 1 of them has a neighbor of degree 2, not in $\{u, x, v\}$ and we distinguish three subcases:

1) One of them, say u, has a neighbor y of degree 2 and v is adjacent only to vertices in $\{x, M(y)\backslash\{y\}\}$.

Let us denote the other neighbor of y by z. Now $G\backslash M(y)$ has 6 vertices, one component contains vertices x and v and the graph of 4 remaining vertices is 3-separable with $(0, 1)$ vertices.

2) One of them, say u, has a neighbor y of degree 2 and v is adjacent only to x and y. This means that u must have another neighbor, say z, of degree 2. Graph $G\backslash M(z)$ has 6 vertices, and one component contains only vertices x, v and y so the claim easily follows.

3) Both of u and v have neighbors, say y and w of degree 2, and y has another neighbor, z (not necessarily different from w).

If w has a neighbor in $\{u, y, z\}$ then let us observe $G\backslash M(y)$.

3.1.) v has the degree 2 in G. Now $G\backslash M(y)$ has 6 vertices, x and v form one component and the remaining graph of 4 vertices is 3-separable by $(0, 1)$ vertices.

3.2.) v has the degree 3 in G. Let us denote the neighbor of v different from x and w by v_1. Now $G\setminus(M(y)\cup\{v_1\})$ has all the components with at most 3 vertices.

On the other hand, if non v neighbor of w is not in $\{u,y,z\}$ then $G\setminus(M(y)\cup\{v\})$ has 5 vertices in two components so the claim again stands.

If G is not connected it is easy to see that the claim also holds. □

Lemma 4. *Let $k\leq 5$ and let it hold that no three vertices can have all the colors. Then:*

i) A graph G with at most 6 vertices doesn't allow a highly $(0,1)$-resistant k-multicoloring.

ii) A graph G with at most 7 vertices doesn't allow a highly $(1,0)$-resistant k-multicoloring.

iii) A graph G with at most 10 vertices doesn't allow a highly $(1,1)$-resistant k-multicoloring.

iv) A graph G with at most 11 vertices doesn't allow a highly $(2,0)$-resistant k-multicoloring.

Proof. We will prove all the claims for connected graphs and the claims for disconnected graphs follow from there. Without the loss of generality we can assume $k=5$.

i) Let G be a connected graph with at most 6 vertices. Each color must be assigned at least 2 times so at least one vertex will have at least two colors. Moreover, no vertex can have three or more colors because no three vertices can have all the colors. Let us assume that one of the vertices has the set of colors $\{1,2\}$. Now neither of the remaining vertices can have neither of the sets $\{3,4\}$, $\{3,5\}$, $\{4,5\}$. But then it is impossible to assign the colors two times each.

ii) Let G be a connected graph with at most 7 vertices. If there is a vertex of degree at least 3 in G the claim is easily seen, so let $\Delta(G)=2$. If G is a path the claim can again be easily seen so let us assume G is a cycle. No vertex can have 3 or more colors and every color must be assigned at least twice so at least one vertex must have two colors. Let us assume one of the vertices has the set of colors $\{1,2\}$ and let us denote that vertex by u_1. Further, let us denote the rest of the vertices in the cycle by $u_2,...,u_7$, starting from u_1 in any direction. Now, no other vertex can have neither of the sets $\{3,4\}$, $\{3,5\}$, $\{4,5\}$, and since each of those colors must appear at least twice we assign the colors $3,4,5$ to the remaining 6 vertices, one to each. It is obvious that the colors $3,4,5$ will not appear anywhere else so there is no point in assigning them in such a way that one color is given to two vertices that are on a distance 2 or less, since then both instances of that single color could be

easily removed. So without the loss of generality we may assume the multicoloring function is the following:

u_1	u_2	u_3	u_4	u_5	u_6	u_7
$1,2$	3	4	5	3	4	5

Now we must assign the colors 1 and 2 once more each. It is easy to see that if we assign the colors 1 and 2 to two vertices that have different colors, three vertices could be found that have all the colors, so the only option is to assign 1 and 2 only to u_2 and u_5, or u_3 and u_6, or u_4 and u_7, one to each. However in each of those possibilities a vertex can be chosen to remove both instances of one of the colors.

iii) Let G be a connected graph with at most 10 vertices. Each color must be assigned at least 3 times so at least one vertex will have 2 or more colors. If one vertex would have 3 or more colors it would be easy to find three vertices that have all the colors, so let us assume that one of the vertices has exactly 2 keys, the set $\{1,2\}$. Now none of the other vertices can have neither of the sets $\{3,4\}$, $\{3,5\}$, $\{4,5\}$ and since each of those colors must be assigned at least 3 times each, we must assign them to the remaining 9 vertices so that each vertex has exactly one color from the set $\{3,4,5\}$. The colors 1 and 2 must be assigned twice more each so obviously two of the 9 vertices that have different colors from $\{3,4,5\}$ will have different colors from $\{1,2\}$. But then again 3 vertices can be found that have all the colors.

iv) Let G be a connected graph with at most 11 vertices. If there is a vertex of degree 3 in G then let us denote it by u. $G\backslash M(u)$ has 7 vertices and the claim follows from ii). Let us assume that the highest degree in G is 2. If G is a path the claim is easy to see so let us assume G is a cycle. Let u be any vertex in G. $G\backslash M(u)$ is a path of 8 vertices and by denoting one of the central vertices with v we can see that $G\backslash(M(u)\cup M(v))$ has all the components of size at most 3. □

Lemma 5. *Let $k \leq 5$ and let it hold that no three vertices can have all the colors. Then:*

i) A graph G with 7 vertices that is either a path or it is disconnected doesn't allow a highly $(0,1)$-resistant k-multicoloring.

ii) A graph G with 11 vertices that is either disconnected or it has a minimal degree 1 doesn't allow a highly $(1,1)$-resistant k-multicoloring.

iii) A disconnected graph G with 15 vertices doesn't allow a highly $(2,1)$-resistant k-multicoloring.

Proof. i) If G is a disconnected graph with 7 vertices the claim follows from Lemma 4 i), and if G is a path it is 3-separable with $(0, 1)$ vertices.

ii) Let G be a disconnected graph with 11 vertices. The claim follows from Lemma 4 i), ii) and iii). Now let G be a connected graph with 11 vertices such that $\delta(G) = 1$. Let us denote with x a vertex of degree 1 and with y its only neighbor. We distinguish two cases.

1) y has the degree 2. We follow the path starting in y until we reach a vertex of degree 3. If such a vertex doesn't exist G is a path and it is 3-separable with $(1, 1)$ vertices. On the other hand, if such a vertex exists, let us denote it by u. $G\backslash M(u)$ has 7 vertices and it is either disconnected or it is a path so the claim follows from i).

2) y has the degree 3. If any neighbor of y has the degree 3 let us denote it by u. $G\backslash M(u)$ is disconnected (x is isolated) with 7 vertices so the claim follows from i). Let us assume both neighbors of y have the degree at most 2. If they have a common neighbor other then y, it must have the degree 3 so by removing it along with its neighbors we again obtain a disconnected graph with 7 vertices and the claim follows as before. If one of those neighbors has the degree 1 then let us observe the other neighbor of y, of degree 2. By following the path starting in that neighbor, not containing y, let us denote by w the first vertex of degree 3 we find. If $G\backslash M(w)$ is disconnected the claim follows from i), and if it is connected then $G\backslash M(w)$ is a path of 5 vertices with two leaves attached to one end of it (vertex y). If we denote by z the neighbor of y that has the degree 2, then G is 3-separable by w and z.

Let us assume both neighbors of y have the degree 2 and each has its own neighbor. Let us denote those neighbors by u and v. We consider three subcases.

2.1.) u and v are adjacent. In that case at least one of them must have the degree 3, without the loss of generality let us assume $d(u) = 3$. $G\backslash M(u)$ is disconnected with 7 vertices so the claim follows from i).

2.2.) u and v are not adjacent and at least one of them has the degree 3. We may assume $d(u) = 3$. Now $G\backslash(M(u) \cup \{v\})$ has all the components with at most 3 vertices.

2.3.) u and v both have the degree 2. If they have a common neighbor it must have the degree 3, so the claim follows as before. Let us assume each of them has its own neighbor, let us denote them by u_1 and v_1, respectively. If u_1 and v_1 are adjacent the claim follows as in 2.1. and if they are not adjacent and at least one of them has the degree 3 then it is easy to see that G is 3-separable, similarly as in 2.2. Let us assume u_1 and v_1 both have the degree 2, and let us denote their

neighbors by u_2 and v_2, respectively. If u_2 and v_2 are adjacent, at least one of them must have the degree 3, without the loss of generality we may assume $d(u_2) = 3$. Now $G\backslash(M(u_2) \cup \{y\})$ has all the components with at most 3 vertices. If any of u_2, v_2 has the degree 1 then $G\backslash M(y)$ has 7 vertices and is disconnected so the claim follows from i). And if they are not adjacent and both have the degree 2, they are both adjacent to the 1 remaining vertex in the graph. But now $G\backslash M(y)$ is a path of 7 vertices and the claim again follows from i).

iii) Let G be a graph with 15 vertices and at least 2 components.

If the smallest component in G has 1 vertex it obviously cannot have all the colors so let us observe the remaining 14 vertices. If there exist a vertex of degree 3 among those 14 vertices let us denote it by u. $G\backslash M(u)$ a graph with 10 vertices which is doesn't allow a highly $(1,1)$-resistant k-multicoloring by Lemma 4 iii). The same reasoning follows if the smallest component in G has 2 or 3 vertices.

If the smallest component in G has 4 vertices then that component is 3-separable with $(0,1)$ vertices and the graph remaining 11 vertices doesn't allow a highly $(2,0)$-resistant k-multicoloring by Lemma 4 iv).

If the smallest component in G has 5, 6 or 7 vertices then by Lemma 4 ii) it doesn't allow a highly $(1,0)$-resistant k-multicoloring and the graph of remaining 10, 9 or 8 vertices doesn't allow a highly $(1,1)$-resistant k-multicoloring by Lemma 4 iii). □

Lemma 6. *Let G be a connected graph with 19 vertices, $\Delta(G) = 3$, and $\Delta(G\backslash M(u)) = 3$, for any vertex u with degree 3 in G. Then one of the following holds:*

a) There exists a vertex u in G such that $G\backslash M(u)$ is disconnected with 15 vertices.

b) There are vertices u and v in G such that $G\backslash(M(u) \cup M(v))$ has 11 vertices and it is either disconnected or it has the minimal degree 1.

Proof. We will prove the claim through two cases, depending on the minimal degree in G. Obviously G cannot be 3-regular so its minimal degree is either 1 or 2.

1) $\delta(G) = 1$. Let us denote by x the vertex of degree 1 and by y its only neighbor. If y has the degree 2 we follow the path starting in y, not containing x, and denote by u the first vertex of degree 3 in that path. $G\backslash M(u)$ has 15 vertices and it is disconnected, so a) holds. Let us assume y has the degree 3 and let us denote the other two neighbors of y by u and v. If any of them has the degree 3 then by removing it and its neighbors, we obtain a disconnected graph (x is isolated) with 15 vertices and again a) holds. If any of u and v has the degree 1 the claim is easy to see, similarly as when y has the degree 2. So let us assume u and v both

have the degree 2. If u and v have a common neighbor it must have the degree 3 so by removing and its neighbors, we again obtain a disconnected graph with 15 vertices. Let us assume u and v have each its own new neighbor and let us denote them by u_1 and v_1, respectively. If any of them has the degree 3, without the loss of generality we may assume that $d(u_1) = 3$, then y has the degree 2 in $G\backslash M(u_1)$ so when we remove another vertex of degree 3 from $G\backslash M(u_1)$ along with its neighbors, the remaining graph will have 11 vertices and x will have the degree 1 in it, so b) holds. If any of u_1 and v_1 has the degree 1, let us assume $d(u_1) = 1$, then whatever two vertices of degree 3 we remove along with their neighbors, u_1 will have the degree 0 or 1 in the remaining graph of 11 vertices so b) holds. On the other hand, if u_1 and v_1 both have the degree 2 we observe the rest of the graph. There must be at least one vertex besides y with degree 3. Let us remove it and its neighbors. Now by removing y and its neighbors at least one of u_1 and v_1 remains either isolated or with the degree 1 in the remaining graph of 11 vertices, so again b) holds.

2) $\delta(G) = 2$. Let us distinguish two subcases.

2.1.) Two vertices of degree 2 are adjacent in G.

Let us denote those vertices by x and y. If x and y have a common neighbor of degree 3 let us denote it by u and let us observe the path starting in u not containing x nor y. Let us denote by w the first vertex of degree 3 on that path. If $G\backslash M(w)$ is disconnected then a) holds and if it is connected then u is the only vertex of degree 3 in the remaining graph and by removing it and its neighbors, we obtain a path of 11 vertices so b) holds. Let us assume x and y have one more neighbor each. If they are adjacent at least one of them must have the degree 3 and by removing it and its neighbors, either x or y will remain isolated so a) holds. Let us assume the neighbors are not adjacent and let us denote them by x_1 and y_1. If at least one of those neighbors has the degree 2, without the loss of generality we may assume that it is x_1, then let us follow the path starting in x, not passing through y, and let us denote by u the first vertex of degree 3 in that path (not necessarily different from y_1) (such a vertex must exist because $\delta(G) > 1$), and let us denote the predecessor of the predecessor of u by v (note that v can be x if we found u on a distance 2 from x). Now vertex v has the degree 1 in $G\backslash M(u)$ and since its only neighbor is surely of degree 2 then by removing the other vertex of degree 3 and its neighbors (one more vertex of degree 3 must exist) from $G\backslash M(u)$, v will remain with the degree at most 1 and b) holds.

So let us assume that both neighbors, of x and y, have the degree 3. Let us again denote the neighbors by x_1 and y_1, respectively. If they have a common neighbor then let us remove x_1 and its neighbors. Now y has the degree 1 and y_1 has the

degree 2 in $G\backslash M(x_1)$ so after removing the other vertex of degree 3 and its neighbors (one more vertex of degree 3 must exist), y will have the degree at most 1 and b) holds.

Let us assume x_1 and y_1 have no common neighbors but instead have 2 more neighbors each. If any of them has the degree 3, without the loss of generality we may assume that it is the neighbor of x_1, then by removing it and its neighbors, x remains with the degree 1 and it will have the degree at most 1 after removing the other vertex of degree 3 along with its neighbors, by the same reasoning as before. Let us assume that all the neighbors of x_1 and y_1 have the degree 2.

If two of them are adjacent, both of them from x_1 (or y_1), then it is easy to see that 1 vertex of degree 3 can be found to leave the graph disconnected with 15 vertices. And if one neighbor of x_1 is adjacent to a neighbor of y_1 then we observe the other neighbors of x_1 and y_1 and let us denote them by x_2 and y_2, respectively. They obviously cannot be adjacent and if they have a common neighbor it must be of degree 3, so by removing it and its neighbors, we obtain a disconnected graph with 15 vertices. Let us assume x_2 and y_2 have each its own neighbor and let us denote them by x_3 and y_3, respectively. If any of x_3 and y_3 has the degree 3, without the loss of generality let us assume it is x_3, then $G\backslash(M(x_3)\cup M(y_1))$ is disconnected with 11 vertices so b) holds. Let us assume x_3 and y_3 have the degree 2. Now if we remove x_1 and its neighbors, y and x_3 have the degree 1 in $G\backslash M(x_1)$. If we remove y_1 and its neighbors, then x_3 will have the degree 1 in the remaining graph of 11 vertices and otherwise y will remain with degree 1 so in each case a) holds.

On the other hand, if none of the neighbors of x_1 and y_1 are adjacent let us denote them by x_{11}, x_{12}, y_{11}, y_{12}. If x_{11} and x_{12} (or y_{11} and y_{12}) have a common neighbor of degree 2 or 3 it is easy to obtain a disconnected graph. Let us assume that some other two have a common neighbor and without the loss of generality let us assume those are x_{12} and y_{11}. (Figure 1 a) If that neighbor has the degree 2 then $G\backslash(M(x_1)\cup M(y_1))$ is disconnected with 11 vertices and if it has the degree 3 then that vertex will have the degree at most 1 in $G\backslash(M(x_1)\cup M(y_1))$ so b) holds. Let us assume none of x_{11}, x_{12}, y_{11}, y_{12} have a common neighbor and let us denote their neighbors by x_{21}, x_{22}, y_{21}, y_{22} (Figure 1 b)).

If any of x_{21}, x_{22}, y_{21}, y_{22} has the degree 3, without the loss of generality let us assume x_{21} then $G\backslash(M(x_{21})\cup M(y_1))$ has 11 vertices and minimal degree 1 ($d(x)=1$), so let us assume they all have the degree 2. But now y, x_{21} and x_{22} have the degree 1 in $G\backslash M(x_1)$ and when we remove any vertex of degree 3 and its neighbors, at least one of them will have the degree at most 1 in the remaining graph of 11 vertices.

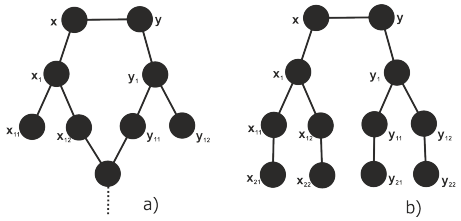


Figure 1: Subcases of 2.1.

2.2.) There are no adjacent vertices of degree 2.

Let x be a vertex of degree 2 and u and v its neighbors of degree 3. Let us consider two subcases.

2.2.1.) u and v are adjacent.

If they have a common neighbor it must have the degree 3 and by removing it and its neighbors, we obtain a disconnected graph with 15 vertices.

If u and v have one more neighbor each and at least one of them has the degree 3, without the loss of generality let us assume it is the neighbor of u, then by removing it and its neighbors, x remains with degree 1 and v with degree 2, so by removing the other vertex of degree 3 and its neighbors from the graph the degree of x will be at most 1 and b) will hold. Let us assume both neighbors have the degree 2 and let us denote them by u_1 and v_1. If u_1 and v_1 have a common neighbor the claim is again easy to see so let us assume they have one more neighbor each and let us denote them by u_2 and v_2. They must have the degree 3 because there are no adjacent vertices of degree 2 in the graph.

If u_2 and v_2 are adjacent then by removing any of them and its neighbors, we obtain a disconnected graph on 15 vertices and if they have a common neighbor of degree 3 then by removing it and its neighbors, we obtain a disconnected graph on 15 vertices. If u_2 and v_2 have a common neighbor of degree 2 then $G\backslash(M(u_2) \cup M(v))$ has 11 vertices and v_2 has the degree at most 1 so b) holds. And if they have no common neighbors then $G\backslash(M(u_2) \cup M(v_2))$ is disconnected with 11 vertices.

2.2.2.) u and v are not adjacent.

If they have a common neighbor of degree 3 the claim is easy to see. Let us first assume they have a common neighbor of degree 2 and let us denote their other neighbors by u_1 and v_1 (they cannot have 2 common neighbors of degree 2). If at least one of them has the degree 3, without the loss of generality let us assume that it is u_1, then in $G\backslash M(u_1)$ x has the degree 1. If v is not the only vertex of degree 3 in

$G\backslash M(u_1)$ then by removing that other vertex we obtain a graph with 11 vertices in which x has the degree 1. And if v is the only vertex with the degree 3 in $G\backslash M(u_1)$ then that means v_1 has the degree at most 2 in $G\backslash M(u_1)$ and by removing v and its neighbors, the other (non v) neighbor of v_1 remains with the degree 1 in a graph of 11 vertices or we have obtained a disconnected graph. So let us assume that u_1 and v_1 have the degree 2. They are obviously not adjacent and if they have a common neighbor of degree 3 the claim is easily seen, so let us assume they each have a new neighbor of degree 3, let us denote them by u_2 and v_2. Now $G\backslash(M(u_2) \cup M(v))$ is disconnected (u is isolated) with 11 vertices.

If u and v have no common neighbors let us denote their neighbors by u_1, u_2, v_1, v_2. If any of them has the degree 3, without the loss of generality let us assume it is u_1, then x has the degree 1 in $G\backslash M(u_1)$. If there exists a vertex of degree 3 in $G\backslash M(u_1)$ different from v then by removing it as the other agent x has the degree at most 1 in the remaining graph of 11 vertices. And if v is the only vertex with degree 3 in $G\backslash M(u_1)$ then v_1 and v_2 have the degree at most 2 in $G\backslash M(u_1)$ and by removing v and its neighbors we either obtain a disconnected graph or at least one neighbor of v_1 and v_2 has the degree at most 1 in $G\backslash(M(u_1) \cup M(v))$. The only remaining case to consider is when u_1, u_2, v_1, v_2 all have the degree 2. Neither two of them may be adjacent and they cannot have common neighbors of degree 2. If any two of them have a common neighbor of degree 3 the claim is easy to see and if they have no common neighbors then let us remove the neighbor of u_1 and its neighbors, and $M(v)$. The remaining graph has 11 vertices and u has the degree 1 so the claim is proven. □

Theorem 7. *1. A graph G with at most 17 vertices doesn't allow a highly $(3,1)$-resistant multicoloring.*

2. $6 \leq HR_{3,1}(19) \leq HR_{3,1}(18) \leq 7$.

3. $HR_{3,1}(20) = 4$.

Proof. First we observe that that if G allows a highly $(3,1)$-resistant k-multicoloring, then $k \geq 4$. Let us first prove that a graph G with at most 17 vertices doesn't allow a highly $(3,1)$-resistant multicoloring. Let us assume the opposite, that there exists a graph G with 17 vertices and $k \in \mathbb{N}$, such that G allows a highly $(3,1)$-resistant k-multicoloring. Components of at most 3 vertices cannot have all the colors so we disregard them. We will prove the claim through four cases, depending on the number of components with at least 4 vertices in G.

1) G has 4 components with at least 4 vertices.

The largest component in G has at most 5 vertices and it is 3-separable with $(1,0)$ vertices. The second and third largest component are also 3-separable with $(1,0)$ vertices, and the fourth largest component is 3-separable with $(0,1)$ vertices. This means that G is 3-separable by $(3,1)$ vertices and therefore doesn't allow the aforementioned coloring.

2) G has 3 components with at least 4 vertices.

If the largest component in G has 9 or 8 vertices then it is 3-separable with $(2,0)$ vertices by Lemma 1 ii), the second largest component is 3-separable with $(1,0)$ vertices and the remaining component with $(0,1)$ vertices.

If the largest component in G has 7 vertices it is 3-separable with $(1,1)$ vertices by Lemma 1 i) and the two remaining components are 3-separable by $(1,0)$ vertices each.

3) G has 2 components with at least 4 vertices.

If the largest component in G has 13 vertices it is 3-separable with $(3,0)$ vertices by Lemma 1 iv), and the remaining component of at most 4 vertices is 3-separable with $(0,1)$ vertices.

If the largest component in G has 11 or 12 vertices it is 3-separable with $(2,1)$ vertices by Lemma 1 iii) and it is easily seen that the remaining component is 3-separable with $(1,0)$ vertices.

If the largest component has 9 or 10 vertices it is 3-separable with $(2,0)$ vertices by Lemma 1 ii) and the remaining component is 3-separable with $(1,1)$ vertices by Lemma 1 i).

4) G has exactly 1 component with at least 4 vertices.

Let G be a connected graph with 17 vertices. We can assume this because all other cases are implied by this solution. We consider 3 subcases, depending on the highest degree in G.

4.1.) The highest degree in G is 4.

Let u be a vertex of degree 4 in G. $G\backslash M(u)$ has 12 vertices and it is 3-separable with $(2,1)$ vertices by Lemma 1 iii).

4.2.) The highest degree in G is 2.

G is obviously a cycle or a path and by choosing a vertex u of degree 2, $G\backslash M(u)$ is a union of paths and has 14 vertices. It can now be easily seen that the claim holds.

4.3.) The highest degree in G is 3.

Let us denote any vertex of degree 3 by u. $G\backslash M(u)$ has 13 vertices. If the highest degree in $G\backslash M(u)$ is at most 2 the claim easily follows so let us assume the highest degree in $G\backslash M(u)$ is 3. We distinguish 2 possibilities.

a) There is a vertex of degree 1 in $G\backslash M(u)$. The claim now follows from Lemma 2 ii).

b) The minimal degree in $G\backslash M(u)$ is 2. Let v be a vertex in $G\backslash M(u)$ with degree 3. $G\backslash(M(u) \cup M(v))$ has 9 vertices. If the highest degree in $G\backslash(M(u) \cup M(v))$ is at most 2 the claim can easily be seen, so let us assume the highest degree in $G\backslash(M(u) \cup M(v))$ is 3. Also, if there exists a vertex of degree 1 in $G\backslash(M(u) \cup M(v))$, the claim follows from Lemma 2 i). The only case left to consider is if all the vertices in $G\backslash(M(u) \cup M(v))$ are of degree 2 or 3. But in this case the claim follows from Lemma 3. We have proven that a graph G with at most 17 vertices doesn't allow a highly $(3, 1)$-resistant multicoloring.

A graph with 18 vertices that that allows a highly $(3, 1)$-resistant 7-multicoloring is given in Figure 2.

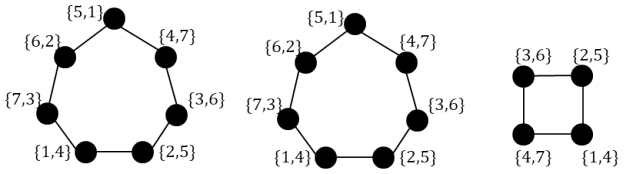


Figure 2: A graph with 18 vertices and a highly $(3, 1)$-resistant 7-multicoloring

From this it follows that $HR_{3,1}(19) \leq 7$.

Let us prove that $HR_{3,1}(19) \geq 6$. Let us assume the opposite, that there exists a graph G with 19 vertices that allows a highly $(3, 1)$-resistant 5-multicoloring.

Since no three vertices can have all the colors it is enough to observe only components with 4 or more vertices. We distinguish four possibilities:

1) G has four components with at least 4 vertices.

The largest component has at most 7 vertices so by Lemma 4 ii) it doesn't allow a highly $(1, 0)$-resistant 5-multicoloring. The second and third largest component also don't allow a highly $(1, 0)$-resistant 5-multicoloring and the smallest of the 4 observed components can have 4 vertices at most so it is 3-separable by $(0, 1)$ vertices.

2) G has three components with at least 4 vertices.

Possible component sizes are different, depending on how many vertices in G are in components with less then 4 vertices. However, the largest component can have 11 vertices at most so by Lemma 4 iv) it doesn't allow a highly $(2, 0)$-resistant

5-multicoloring. Second largest component can have 7 vertices at most so by Lemma 4 ii) it doesn't allow a highly $(1,0)$-resistant 5-multicoloring and the third largest component has at most 6 vertices so by Lemma 4 i) it doesn't allow a highly $(0,1)$-resistant 5-multicoloring. This holds for all possible sizes of the components.

3) G has two components with at least 4 vertices.

The largest component has 15 vertices at most and the fact that a graph of 15 vertices doesn't allow a highly $(3,0)$-resistant 5-multicoloring follows from the results in [7]. The second largest component has 4 vertices and it is 3-separable by $(0,1)$ vertices. The case when the largest component has 14 or 13 vertices and the second largest 5 or 6 follows in the same way.

If the largest component has 12 vertices then it is 3-separable by $(2,1)$ vertices which follows from Lemma 1 iii). The second largest component has at most 7 vertices and it doesn't allow a highly $(1,0)$-resistant 5-multicoloring, by Lemma 4 ii).

If the largest component has 11 or 10 vertices then it doesn't allow a highly $(2,0)$-resistant 5-multicoloring and the second largest component then has at most 8 or 9 vertices and it doesn't allow a highly $(1,1)$-resistant 5-multicoloring, which follows from Lemma 4 iv) and iii).

4) G has exactly one component with at least 4 vertices.

Without the loss of generality we may assume that G is connected with 19 vertices because all other cases are implied by this solution.

4.1.) The highest degree in G is 2. Let u be any vertex with degree 2. $G\backslash M(u)$ has 16 vertices and it is a union of paths. Let us denote by v the center of the largest path in $G\backslash M(u)$. Now $G\backslash(M(u)\cup M(v))$ has 13 vertices and it is a union of at least 2 paths of which the largest one has at most 7 vertices and it is doesn't allow a highly $(1,0)$-resistant 5-multicoloring by Lemma 4 ii) The remaining grapf of 6 vertices doesn't allow a highly $(0,1)$-resistant 5-multicoloring by Lemma 4 i).

4.2.) The highest degree in G is 3. Let u be any vertex with degree 3. $G\backslash M(u)$ has 15 vertices.

If all of them have the degree at most 2 then let v be any vertex of degree 2. $G\backslash(M(u)\cup M(v))$ has 12 vertices and it is a union of paths. If $G\backslash(M(u)\cup M(v))$ is not connected it is easy to see that it is 3-separable with $(1,1)$ vertices and if it is one path of 12 vertices then let w be a vertex on a distance 4 from the end of that path. $G\backslash(M(u)\cup M(v)\cup M(w))$ is a union of two paths, of lengths three and six, and the larger one is 3-separable with $(0,1)$ vertices.

On the other hand, if there exists a vertex of degree 3 in $G\backslash M(u)$ than the claim follows from Lemma 6 and Lemma 5 ii) and iii).

4.3.) The highest degree in G is at least 4. Let u be a vertex in G with the highest degree. $G \backslash M(u)$ has at most 14 vertices and it doesn't allow a highly $(2,1)$-resistant 5-multicoloring by Lemma 5 iii). This proves our claim that $HR_{3,1}(19) \geq 6$.

It remains to prove that $HR_{3,1}(20) = 4$. It is easily seen that $HR_{3,1}(20) \geq 4$ must hold, and the graph G with 20 vertices that allows a highly $(3,1)$-resistant 4-multicoloring is given in Figure 3.

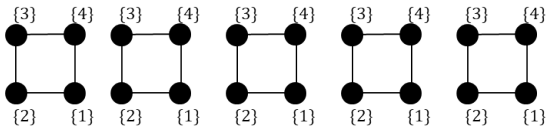


Figure 3: A graph with 20 vertices and a highly $(3,1)$-resistant 4-multicoloring

□

4 Acknowledgements

Partial support of the Croatian Ministry of Science and Education is gratefully acknowledged.

References

[1] B. Bollobas, Modern Graph Theory, Springer, New York, 1998.

[2] Halldórsson, M. M. and Kortsarz, G. (2002). Tools for multicoloring with applications to planar graphs and partial k-trees. Journal of Algorithms, 42(2), 334-366.

[3] Halldórsson, M. M., and Kortsarz, G. (2004, August). Multicoloring: Problems and techniques. In International Symposium on Mathematical Foundations of Computer Science, pp. 25-41.

[4] Harary, F. (1985). Conditional colorability in graphs. Graphs and applications (Boulder, Colo., 1982), Wiley-Intersci. Publ, 127-136.

[5] Marx, D. (2004). Graph colouring problems and their applications in scheduling. Periodica Polytechnica Electrical Engineering, 48(1-2), 11-16.

[6] Shamir, A. (1979) How to share a secret. Communications of the ACM 22, 11, 612-613.

[7] Vojković, T., Vukičević, D. and Zlatić, V. (2018). Multicoloring of Graphs to Secure a Secret. Rad HAZU, Matematičke znanosti, 22, 1-22.

[8] Vojković, T. and Vukičević, D. (2018). Highly (a, b)-resistant multicoloring of graphs. Submitted to Annales Mathematicae et Informaticae.

[9] Zhang, Z., Liu, L. and Wang, J. (2002). Adjacent strong edge coloring of graphs. Applied Mathematics Letters, 15(5), 623-626.

A note on maximal matchings in rooted products of paths and short cycles

Ivana Zubac

Faculty of Mechanical Engineering, Computing and Electrical Engineering,
University of Mostar
Matice hrvatske bb, BA-88000 Mostar, Bosnia and Herzegovina

Abstract

A matching M in a graph G is maximal if it cannot be extended to a larger matching in G. The enumerative properties of maximal matchings are much less known and researched than for maximum and perfect matchings. In this paper we present the recurrences and generating functions for the sequences enumerating maximal matchings in rooted products of paths and short cycles. We also analyze the asymptotic behavior of those sequences.

Keywords: maximal matching, rooted product of graphs

MSC: 05C30, 05C70, 05C76

1 Introduction

A matching in a graph is a collection of its edges such that no two edges in this collection have a vertex in common. Many problems in natural, technical and social sciences can be successfully modeled by matchings in graphs. Today the matching theory is a well developed part of graph theory, strongly influenced by chemical applications. Particularly well researched are perfect and maximum matchings. For a general background on matching theory and terminology we refer the reader to the classical monograph by Lóvasz and Plummer [4].

A matching is perfect if its edges are collectively incident to all vertices of G. It is clear that perfect matchings are as large as possible and that no other matching in G

DOI: https://doi.org/10.5592/CO/CCD.2018.13

can be " larger" than a perfect one. In all other applications we are also interested mostly in large matchings. If we are using the number of edges as the measure of "largeness", we get the maximum matchings. For them we have well developed structural theory and many enumerative results. No such theory, however, exists for another class of large matchings - maximal matchings. A matching M is **maximal** if it cannot be extended to a larger matching in G. Obviously, every maximum matching is also maximal, but the opposite is generally not true. Maximal matchings are much less researched, but still very useful. That goes both for their structural and their enumerative aspects. The crucial difference is non-locality. That means that we cannot split the set of all maximal matchings into those containing an edge e and those not containing it without taking into account the edge-neighborhood of e. The main goal of this work is to increase our knowledge about the enumerative aspects of maximal matchings.

We begin by establishing the principal terminology and notation which we will use throughout the article. It is taken from [7]. All graphs G considered in this paper will be finite and simple, with vertex set $V(G)$ and set of edges $E(G)$. We will denote by $n = |V(G)|$ the number of vertices and by $m = |E(G)|$ the number of edges in G. As usual, the cycle and complete graph on n vertices are denoted by C_n and K_n, respectively. However, by P_n we denote the path of length n, i.e., on $n + 1$ vertices. Let $\Psi(G)$ denote the number of maximal matchings in graph G.

2 Rooted product of graphs

Many interesting graphs arise from simpler building blocks via some binary operations known as graph products. In 1978, Godsil and McKay [2] introduced a new product of two graphs G_1 and G_2, called the rooted product, and denoted by $G_1 \odot G_2$. The rooted product graphs are used, for example, in internet networking for connecting local networks to a wider frame. One can view such a situation as a combination of many local networks (copies of graph H) having a server (the root vertex of graph H). These servers are themselves connected through a global network (the graph G). So, one motivation to count the maximal matchings can be securing networks with as few sensors as possible.

The main goal of this paper is to study the enumerative aspects of maximal matchings in some classes of rooted products. We establish recurrences for the enumerating sequences and, in some cases, we use generating functions to determine

their asymptotic behavior. We consider operations of rooted product of paths and cycles, and enumerate maximal matchings in the resulting graphs.

Let $V(G) = \{u_i | 1 \leq i \leq n\}$ and $v \in V(H)$. **The rooted product** $G \odot H$ of a graph G and a rooted graph H with respect to the root v is defined as follows: take $n = |V(G)|$ copies of H, and for every vertex u_i of G, identify u_i with the root v of the i-th copy of H. Obviously,

$$\begin{aligned} |V(G \odot H)| &= |V(G)|\,|V(H)| \\ |E(G \odot H)| &= |E(G)| + |V(G)|\,|E(H)|\,. \end{aligned}$$

Hence, we take as many copies of graph H as we have vertices in graph G. If H or G is the singleton graph, then $G \odot H$ is equal to G or H, respectively. In this sense, to obtain the rooted product $G \odot H$, hereafter we will only consider graphs G and H of orders greater than or equal to two. We note that the rooted product is not commutative.

The rooted product is especially relevant for trees, as a rooted product of two trees is another tree. For instance, [3] used rooted products to find graceful numberings for a wide family of trees. Also, the rooted product of two graphs is a subgraph of the Cartesian product of those two graphs.

3 Rooted product graph $P_n \odot P_m$

We first consider the case when we have a rooted product of two paths, always taking a leaf as a root. In all cases we obtain a tree. An example is shown in Figure 1.

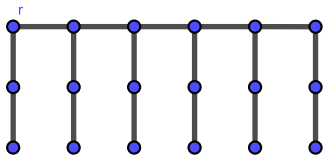


Figure 1: Example of $P_n \odot P_2$ rooted product

The next lemma is a direct consequence of Proposition 6.1. from article [1].

Lemma 1. *The number of maximal matchings in rooted product graph $P_n \odot P_1$ is equal to the value of the $(n+2)$-nd Fibonacci number, i.e. $\Psi(P_n \odot P_1) = F_{n+2}$.*

Proof. It is clear that every vertex of the P_n must be covered by an edge of a maximal matching. If vertex n is covered by the edge $v_{n-1}v_n$, the remaining edges of a maximal matching must form a valid maximal matching in $P_{n-2} \odot P_1$, and hence are counted by $\Psi(P_{n-2} \odot P_1)$. If v_n is covered by the pendent edge of P_1, the remaining maximal matchings are counted by $\Psi(P_{n-1} \odot P_1)$. Hence, the number of maximal matchings in $P_n \odot P_1$ satisfies the recurrence

$$\Psi(P_n \odot P_1) = \Psi(P_{n-1} \odot P_1) + \Psi(P_{n-2} \odot P_1),$$

with initial conditions $\Psi(P_0 \odot P_1) = 1, \Psi(P_1 \odot P_1) = 2$.
The sequence F_n of Fibonacci numbers is defined by the same recurrence relation and the same initial conditions, so the claim follows. □

We now give another result for a rooted product of two paths, this time for the case $P_n \odot P_2$ shown in Figure 1. We will denote the number of maximal matchings in the graph $P_n \odot P_2$ with t_n.

Proposition 1. *The sequence t_n satisfies the recurrence $t_n = t_{n-1} + 2t_{n-2} + t_{n-3}$ with the initial conditions $t_0 = 2, t_1 = 4$.*

In the proof we will repeat argument from paper [6] – we split graph into two parts: counting the number of matchings containing the root and counting the matchings which do not contain the root. We must take into account the edge-neighborhood of e.

Proof. Each maximal matching in $P_n \odot P_2$ either covers the rightmost root or does not cover it. In the first case, the remaining edges must form either a valid matching in $P_{n-1} \odot P_2$ (if is root covered by edge of P_2) or a valid maximal matching in $P_{n-2} \odot P_2$ (if the root is covered by edge of P_n).

Maximal matchings that do not cover that root must cover both its neighbors. The neighbor in P_n can be covered in two ways, one of them counted by the number of maximal matchings in $P_{n-2} \odot P_2$, another by the number of maximal matchings in $P_{n-3} \odot P_2$. The claim now follows by adding the two contributions. □

The following proposition and corollary give the generating function and asymptotic behavior of t_n.

Proposition 2. *The generating function $T(x)$ for the sequence t_n is given by*

$$T(x) = \frac{x^2 + 2x + 2}{1 - x - 2x^2 - x^3}.$$

Corollary 1. *The asymptotic behavior of t_n is given by $t_n \sim 1.92485 \cdot 0.51952^n$.*

The sequence t_n provides a new combinatorial interpretation of sequence $A141015$ from the OEIS [5].

4 Rooted product graph $P_n \odot C_m$

In this section we will show some results about rooted product graph $P_n \odot C_m$. Figure 2 shows the first case - the rooted product $P_n \odot C_3$.

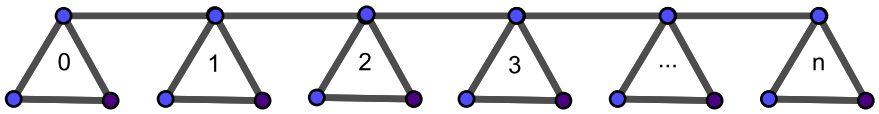


Figure 2: Example of $P_n \odot C_3$ rooted product

The next lemma is a consequence of Proposition 7.3. from article [1]. We will denote the number of maximal matchings in the rooted product graph $P_n \odot C_3$ with p_n.

Lemma 2. *The sequence p_n satisfies the recurrence*

$$p_n = 2p_{n-1} + 3p_{n-2} + p_{n-3},$$

with the initial conditions $p_0 = 3, p_1 = 9, p_2 = 28$.

The sequence p_n appears as $A084084$ in the OEIS.

An example for the next case, the rooted product $P_n \odot C_4$, is shown in Figure 3. We will denote the number of maximal matchings in this case with $\Psi(P_n \odot C_4) = r_n$.

Proposition 3. *The sequence r_n satisfies the recurrence*

$$r_n = 2r_{n-1} + 4r_{n-2},$$

with the initial conditions $r_0 = 2, r_1 = 8$.

Proof. Each maximal matching in $P_n \odot C_4$ must cover the root of the copy of C_4 rooted at the rightmost vertex of P_n. If the root is covered by one of the two edges of C_4 incident with it, the remaining edges of this maximal matching must form a valid maximal matching in $P_{n-1} \odot C_4$. If the root is covered by the edge incident with it in P_n, the remaining edges must form a valid maximal matching in $P_{n-2} \odot C_4$. In that case, there are four ways to form maximal matchings in the rest of the two rightmost copies of C_4. The claim now follows by adding the two contributions. □

Proposition 4. *The generating function $R(x)$ for the sequence r_n is given by*

$$R(x) = \frac{1}{1 - 2x - 4x^2}.$$

Corollary 2. *The asymptotic behavior of r_n is given by $r_n \sim 1 + \sqrt{5}$, twice the golden ratio.*

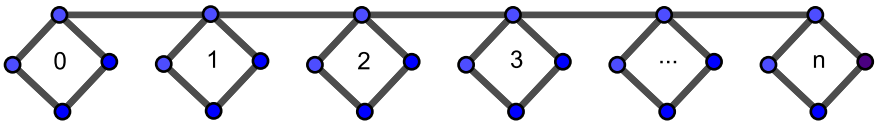


Figure 3: Example of $P_n \odot C_4$ rooted product

The sequence r_n is known as the Horadam sequence $a_n = s \cdot a_{n-1} + r \cdot a_{n-2}$ for $s = 2, r = 4$. It appears as $A085449$ in the OEIS and it counts, among other things, the number of ways to tile an n-board with two types of colored squares and four types of colored dominoes.

5 Concluding remarks

In this paper we have counted maximal matchings in rooted products of some graphs. We have addressed only the combination of paths and short cycles. It would be probably too ambitious to hope for general enumerative results for rooted products. However, many interesting cases should be within the reach. Another interesting thing in such graphs would be to consider their saturation number and to compute their maximal matching polynomials.

References

[1] T. Došlić, I. Zubac, Counting maximal matchings in linear polymers, *Ars Math. Contemp.* 11 (2016) 255-276.

[2] C. D. Godsil., B. D. McKay, A new graph product and its spectrum, *Bull. Aust. Math. Soc.* 18 (1) (1978) 21-28.

[3] K. M. Koh, D. G. Rogers, T. Tan, Products of graceful trees, *Discrete Math.* 31 (3) (1980) 279-292.

[4] L. Lovasz, M. D. Plummer, *Matching Theory,* North-Holland, Amsterdam, 1986.

[5] N. J. A. Sloane, The On-Line Encyclopedia of Integer Sequences, *Notices Amer. Math. Soc.* 50/8 (2003) 912–915.

[6] S. G. Wagner, On the number of matchings of a tree, *European J. Combin.* 28 (2007), 1322-1330.

[7] D. B. West, *Introduction to Graph Theory*, Prentice Hall, Upper Saddle River, 1996.