Acta Universitatis Sapientiae

Informatica

Volume 6, Number 1, 2014

Sapientia Hungarian University of Transylvania
Scientia Publishing House

Contents

Acta Univ. Sapientiae, Informatica, **6,** 1 (2014) 5–20

Some more algorithms for Conway's universal automaton

Boris MELNIKOV
Russia, Togliatti State University
email: B.Melnikov@tltsu.ru

Vasily DOLGOV
Russia, Togliatti State University
email: terenga74@mail.ru

Abstract. In this paper authors propose algorithms for constructing so-called automaton $\mathcal{COM}(\mathrm{L})$ and prove that this automaton coincides, up to re-denoting states, with Conway's universal automaton. We give some algorithms of constructing this automaton and consider some examples.

1 Introduction

In this paper authors propose algorithms for constructing so-called automaton $\mathcal{COM}(\mathrm{L})$. The definition of this automaton could be constructed by [8]. Below, we shall consider this definition.We also prove that this automaton coincides, up to re-denoting states, with Conway's universal automaton $\mathcal{U}_{\mathrm{L}}$ ([4] etc.). We hope that in future it will be possible to show that "in average" (for specially defined "average" notion) this algorithms are more effective than algorithms which construct $\mathcal{U}_{\mathrm{L}}$ by definition, i.e., using (sub)factorisations etc.

Although we do not consider classic complexity problems here, it is obvious that for constructing $\mathcal{U}_{\mathrm{L}}$ (Section 5) complexity is proportional to the number of grids, which is considered in Section 7.

The contents of this article is as follows.

In Section 2 we introduce some notions: binary relation #, pseudo-grid, grid, covering subset of the grids. This notions appeared in [5, 6].

Computing Classification System 1998: F.4.3
Mathematics Subject Classification 2010: 68Q45
Key words and phrases: nondeterministic finite automata, Conway's universal automaton, constructing algorithms
DOI:10.2478/ausi-2014-0015

In Section 3 we give definition of automaton $\mathcal{COM}(L)$ and prove that this automaton coincides with $\mathcal{U}_L$.

In Section 4 we define "covering automata" by automaton $\mathcal{COM}(L)$.

In Section 5 we give two algorithms of constructing automaton $\mathcal{U}_L$, the second of them being "the mirror image" of the first. This algorithms rely on the algorithm, described in [6]. Also we give a simplified version of the first algorithm.

In Section 6 we give a detailed example of the work of algorithms, described in the previous section. An example of *equivalent* covering automaton is also given.

In Section 7 we consider examples which shows how fast a number of grids (i.e., the size of the universal automaton) can grow if we are given the sizes of two canonical automata (i.e., $\widetilde{L}$ and $\widetilde{L^R}$).

In Section 8 we give some ideas for further research.

2 Preliminaries

We shall use designations of [5, 6]. Let us repeat the main of them.

The language of nondeterministic finite automaton

$$K = (Q, \Sigma, \delta, S, F)$$

will be defined by $\mathcal{L}(K)$. For a state q of this automaton, we shall denote the language of automaton $(Q, \Sigma, \delta, S, \{q\})$ by $\mathcal{L}_K^{in}(q)$. $\widetilde{L}$ is the canonical automaton defining regular language L, we shall consider canonical automata without the useless ("dead") state. Let automata $\widetilde{L}$ and $\widetilde{L^R}$ for the given language L be as follows:

$$\widetilde{L} = (Q_\pi, \Sigma, \delta_\pi, \{s_\pi\}, F_\pi) \quad \text{and} \quad \widetilde{L^R} = (Q_\rho, \Sigma, \delta_\rho, \{s_\rho\}, F_\rho).$$

Binary relation $\# \subseteq Q_\pi \times Q_\rho$ is defined in the following way. For some states $A \in Q_\pi$ and $X \in Q_\rho$, condition $A\#X$ holds if and only if there exist some words $u \in \mathcal{L}^{in}_{\widetilde{L}}(A)$ and $v \in \mathcal{L}^{out}_{\widetilde{L^R}}(X)$, such that $uv^R \in \mathcal{L}(K)$. In [6], we considered a simple algorithm for constructing this relation.

If for some pair $P \subseteq Q_\pi$ and $R \subseteq Q_\rho$ we have

$$(\forall A \in P)\,(\forall X \in R)\,(A\#X),$$

then $\mathcal{B} = (P, R)$ is a pseudo-grid. For it, we shall write $\alpha(\mathcal{B}) = P$ and $\beta(\mathcal{B}) = R$.

If for some pseudo-grid $\mathcal{B} = (P, R)$, there exists

- neither $A \in Q_\pi \setminus P$ such that $\Big((P \cup \{A\}), R\Big)$ is also a pseudo-grid,
- nor $X \in Q_\rho \setminus R$ such that $\Big(P, (R \cup \{X\})\Big)$ is also a pseudo-grid,

then $\mathcal{B}$ is a grid.

For the given regular language L, we shall consider its set of grids; in the next sections we shall denote it by $Q_{\mathcal{COM}}$. Some its subset $\mathcal{Q} \subseteq Q_{\mathcal{COM}}$ will be called by *covering* subset of the grids, if for any $A \in Q_\pi$ and $X \in Q_\rho$ such that $A \# X$, there exists a grid $\mathcal{B} \in \mathcal{Q}$, such that $A \in \alpha(\mathcal{B})$ and $X \in \beta(\mathcal{B})$.[1]

As we said before, we shall also use the universal automaton for the given regular language L; by [4, Def. 2.4], we shall denote it by $\mathcal{U}_L$. However, its elements we shall mark by subscripts $_{\mathcal{U}_L}$; e.g., its transition function will be denoted by $\delta_{\mathcal{U}_L}$.

3 Definition of automaton $\mathcal{COM}(L)$ (an alternative definition of automaton $\mathcal{U}_L$)

Automaton $\mathcal{COM}(L)$ for the given regular language L has to be defined in the usual way, i.e., by a quintet. For now, we only have the set of states $Q_{\mathcal{COM}}$ (as we said before, it is the set of grids for the given *language*) and alphabet Σ; let us define sets of initial and final states and the transition function.

Thus, considering automata $\widetilde{L}$ and L^R, we define automaton

$$\mathcal{COM}(L) = (Q_{\mathcal{COM}}, \Sigma, \delta_{\mathcal{COM}}, S_{\mathcal{COM}}, F_{\mathcal{COM}}),$$

where:

- $S_{\mathcal{COM}} = \Big\{ \mathcal{B} \in Q_{\mathcal{COM}} \;\Big|\; \alpha(\mathcal{B}) \ni s_\pi \Big\}$;
- $F_{\mathcal{COM}} = \Big\{ \mathcal{B} \in Q_{\mathcal{COM}} \;\Big|\; \beta(\mathcal{B}) \ni s_\rho \Big\}$;
- for some pair $\mathcal{B}_1, \mathcal{B}_2 \in Q_{\mathcal{COM}}$ (condition $\mathcal{B}_1 = \mathcal{B}_2$ is possible) and some $a \in \Sigma$, we set
$$\delta_{\mathcal{COM}}(\mathcal{B}_1, a) \ni \mathcal{B}_2 \quad (\text{i.e., } \mathcal{B}_1 \xrightarrow[\delta_{\mathcal{COM}}]{a} \mathcal{B}_2)$$
if and only if both the following conditions hold:
$$\Big(\forall A \in \alpha(\mathcal{B}_1)\Big)\ \Big((\delta_\pi(A, a) \neq \emptyset)\ \&\ (\delta_\pi(A, a) \subseteq \alpha(\mathcal{B}_2))\Big); \qquad (1)$$
$$\Big(\forall Y \in \beta(\mathcal{B}_2)\Big)\ \Big((\delta_\rho(Y, a) \neq \emptyset)\ \&\ (\delta_\rho(Y, a) \subseteq \beta(\mathcal{B}_1))\Big). \qquad (2)$$

[1] We shall not consider *algorithms* for constructing such subsets.

Let us remark, that considering canonical automaton *having* possible "dead" state, we obtain *the same* values of φ^{in} and φ^{out} (because these values cannot contain these "dead" states). Then the definition of $\mathcal{COM}(\mathsf{L})$ for the given language L is independent of considering "dead" state.

Also let us remark, that we can write (1) in the following way:

$$\Big(\forall \mathsf{A} \in \alpha(\mathcal{B}_1)\Big)\ \Big((\delta_\pi(\mathsf{A}, \mathfrak{a}) = \{\mathsf{B}\})\,\&\,(\mathsf{B} \in \alpha(\mathcal{B}_2))\Big)$$

(similarly for (2)). But we can*not* write it in the following way:

$$\Big(\forall \mathsf{A} \in \alpha(\mathcal{B}_1)\Big)\ \Big(\delta_\pi(\mathsf{A}, \mathfrak{a}) \subseteq \alpha(\mathcal{B}_2)\Big) \tag{3}$$

(because the value of $\delta_\pi(\mathsf{A}, \mathfrak{a})$ can be $\emptyset$). But considering canonical automaton *having* possible "dead" state (such an automaton is total) we can write it by (3). And considering canonical automaton having transition function of the type

$$\delta : \mathsf{Q} \times \Sigma \to \mathsf{Q}$$

(like, for example, [1]), we can write (3) in the following simple way:

$$\Big(\forall \mathsf{A} \in \alpha(\mathcal{B}_1)\Big)\ \Big(\delta_\pi(\mathsf{A}, \mathfrak{a}) \in \alpha(\mathcal{B}_2)\Big)\,.$$

The following theorem formulates the correctness of both the definitions given before.

Theorem 1 $\quad \mathcal{COM}(\mathsf{L}) = \mathcal{U}_{\mathsf{L}}$.[2]

Proof. Firstly, let us prove for each state $\mathcal{B} \in \mathsf{Q}_{\mathcal{COM}}$, that the pair

$$\Big(\mathcal{L}^{in}_{\mathcal{COM}}(\mathcal{B})\,,\ \mathcal{L}^{out}_{\mathcal{COM}}(\mathcal{B})\Big) \tag{4}$$

is a factorization of L. For it let us suppose, that (4) is only a subfactorization of L (not a factorization). Then we would have:

- either some word $\mathfrak{u} \in \Sigma^*$ (where $\mathfrak{u} \notin \mathcal{L}^{in}_{\mathcal{COM}}(\mathcal{B})$), such that

  $$\Big(\mathcal{L}^{in}_{\mathcal{COM}}(\mathcal{B}) \cup \{\mathfrak{u}\},\ \mathcal{L}^{out}_{\mathcal{COM}}(\mathcal{B})\Big)$$

  is also a subfactorization (or a factorization) of L;

[2] Up to re-denoting states.

- or some word $v \in \Sigma^*$ (where $v \notin \mathcal{L}^{out}_{\mathcal{COM}}(\mathcal{B})$), such that

$$\left(\mathcal{L}^{in}_{\mathcal{COM}}(\mathcal{B})\,,\ \mathcal{L}^{out}_{\mathcal{COM}}(\mathcal{B}) \cup \{v\}\right)$$

  is also a subfactorization (or a factorization) of L.

Without loss of generality, we shall consider the first case. [3]

Besides, we would have a grid $\mathcal{B}' \in Q_{\mathcal{COM}}$, for which

$$\mathcal{L}^{in}_{\mathcal{COM}}(\mathcal{B}') \ni \mathcal{L}^{in}_{\mathcal{COM}}(\mathcal{B}) \cup \{u\}\,.$$

I.e., for the state A of automaton $\widetilde{L}$, such that $\mathcal{L}^{in}_{\widetilde{L}}(A) \ni u$, we would have that

$$A\#X \quad \text{for each} \quad X \in \beta(\mathcal{B})\,.$$

Then $(\alpha(\mathcal{B}) \cup \{A\}) \times \beta(\mathcal{B})$ is a pseudo-grid (or a grid, see [6]), and for $\mathcal{B}$, we obtain a contradiction with the definition of the grid.

Vice versa, let $(\mathcal{X}, \mathcal{Y})$ be a state of automaton $\mathcal{U}_L$. Consider the sets $P \subseteq Q_\pi$ and $R \subseteq Q_\rho$ defined in the following way:

$$P = \left\{ A \in Q_\pi \,\middle|\, (\exists u \in \mathcal{X})\,(\mathcal{L}^{in}_{\widetilde{L}}(A) \ni u) \right\},$$
$$\text{and} \quad R = \left\{ X \in Q_\rho \,\middle|\, (\exists v \in \mathcal{Y}^R)\,(\mathcal{L}^{in}_{\widetilde{L^R}}(X) \ni v) \right\}.$$

Since $(\mathcal{X}, \mathcal{Y})$ is a factorization of L, then for each pair of states $A \in P$ and $X \in Q$, we have $A\#X$. Therefore $P \times Q$ is a pseudo-grid.

And if $P \times Q$ is *not* a grid, then we would add some words to $\mathcal{X}$ or $\mathcal{Y}$ satisfying the definition of (sub) factorization; then $(\mathcal{X}, \mathcal{Y})$ would be not a factorization.

Let us prove the coincidence of the sets of edges. By the definition of $\mathcal{U}_L$,

$$(\mathcal{X}', \mathcal{Y}') \in \delta_{\mathcal{U}_L}((\mathcal{X}, \mathcal{Y}), a) \quad \text{holds if} \quad \mathcal{X}\, a\, \mathcal{Y}' \subseteq L\,.$$

The same condition holds also for $\delta_{\mathcal{COM}}$, i.e.,

$$\mathcal{B}' \in \delta_{\mathcal{COM}}(\mathcal{B}, a) \quad \text{holds if} \quad \mathcal{L}^{in}_{\mathcal{COM}}(\mathcal{B})\, a\, \mathcal{L}^{out}_{\mathcal{COM}}(\mathcal{B}') \subseteq L\,,$$

where $\mathcal{B}'$ corresponds to $(\mathcal{X}', \mathcal{Y}')$ and $\mathcal{B}$ corresponds to $(\mathcal{X}, \mathcal{Y})$. □

[3] (Sub) factorization $\left(\mathcal{L}^{in}_{\mathcal{COM}}(\mathcal{B}) \cup \{u\},\ \mathcal{L}^{\text{out}}_{\mathcal{COM}}(\mathcal{B}) \cup \{v\}\right)$ is also possible.

4 Covering automata

Using automaton $\mathcal{COM}(L)$ and *the given* covering subset of the grids $\mathcal{Q} \subseteq Q_{\mathcal{COM}}$, let us also define corresponding *covering automaton*. We define it in the following way:

$$\mathcal{COM}_{\mathcal{Q}}(L) = (\mathcal{Q}, \Sigma, \delta_{\mathcal{Q}}, S_{\mathcal{Q}}, F_{\mathcal{Q}}),$$

where:

- $S_{\mathcal{Q}} = \mathcal{Q} \cap S_{\mathcal{COM}}$;
- $F_{\mathcal{Q}} = \mathcal{Q} \cap F_{\mathcal{COM}}$;
- $\delta_{\mathcal{Q}} = \left\{ \mathcal{B}_1 \xrightarrow{a} \mathcal{B}_2 \;\middle|\; a \in \Sigma,\ \mathcal{B}_1, \mathcal{B}_2 \in \mathcal{Q},\ \mathcal{B}_1 \xrightarrow[\delta_{\mathcal{COM}}]{a} \mathcal{B}_2 \right\}$.

For now, we do not consider the equivalence of automata $\mathcal{COM}(L)$ and $\mathcal{COM}_{\mathcal{Q}}(L)$ (i.e., whether automaton $\mathcal{COM}_{\mathcal{Q}}(L)$ defines the given language L); some examples will be considered in the next sections.

5 Algorithms for constructing automaton $\mathcal{U}_L$

In [6], we considered a possible algorithm of constructing automaton $\widetilde{L^R}$; using this algorithm, we obtained *at the same time* values of functions φ^{in} and φ^{out} and binary relation $\#$. In this section, we shall obtain automaton $\mathcal{U}_L$ using *the same algorithm*.

Thus, let us suppose that we already have all these objects. Simply by definitions of grids and automaton $\mathcal{COM}(L)$, and also by Theorem 1, we obtain considering all the subsets of the set Q_π the following

Algorithm 1 (Constructing automaton $\mathcal{U}_L$)
Input: *automata* $\widetilde{L}$, $\widetilde{L^R}$, *binary relation* $\#$.
Output: *automaton* $\mathcal{U}_L$.
Step 1. *Consider array* `U[index]`, *where* `index` *can be each element of* $\mathcal{P}(Q_\pi)$ *(except* $\emptyset$*), and its values can be elements of* $\mathcal{P}(Q_\rho)$. *For each possible* `index`, *we set*

$$\texttt{U[index]} := \bigcap_{A \in \texttt{index}} \left\{ X \in Q_\rho \;\middle|\; A \# X \right\}.$$

Step 2. *Consider Boolean array* `B[index]`, *where* `index` *can be each element of* $\mathcal{P}(Q_\pi)$ *(except* $\emptyset$*). For each possible* `index`, *we set*

$$\texttt{B[index]} := (\texttt{U[index]} \neq \emptyset).$$

Step 3. *For each possible* `index`*, such that condition* `B[index]` *holds, if*

$$(\exists\, \mathtt{ind} \in \mathcal{P}(Q_\pi))\,((\mathtt{index} \subset \mathtt{ind})\,\&\,(\mathtt{U[ind]} = \mathtt{U[index]})),$$

then we set `B[index]:=false`*.*
Step 4. *We select the following set of grids:*

$$Q_{\mathcal{COM}} = \left\{ \mathtt{index} \times \mathtt{U[index]} \,\middle|\, \mathtt{index} \in \mathcal{P}(Q_\pi))\,\&\,\mathtt{B[index]} \right\}.$$

Step 5. $\delta_{\mathcal{COM}}$*,* $S_{\mathcal{COM}}$ *and* $F_{\mathcal{COM}}$ *are defined by definition of automaton* $\mathcal{COM}(L)$ *given before.* □

Let us formulate the "mirror image" of this algorithm, where we at first consider subsets of Q_ρ.

Algorithm 2 (Constructing automaton $\mathcal{U}_L$)
Input: *automata* $\widetilde{L}$, $\widetilde{L^R}$*, binary relation* $\#$*.*
Output: *automaton* $\mathcal{U}_L$*.*
Step 1. *Consider array* `U[index]`*, where* `index` *can be each element of* $\mathcal{P}(Q_\rho)$ *(except* $\emptyset$*), and its values can be elements of* $\mathcal{P}(Q_\pi)$*. For each possible* `index`*, we set*

$$\mathtt{U[index]} := \bigcap_{X \in \mathtt{index}} \left\{ A \in Q_\pi \,\middle|\, A \# X \right\}.$$

Step 2. *Consider Boolean array* `B[index]`*, where* `index` *can be each element of* $\mathcal{P}(Q_\rho)$ *(except* $\emptyset$*). For each possible* `index`*, we set*

$$\mathtt{B[index]} := (\mathtt{U[index]} \neq \emptyset).$$

Step 3. *For each possible* `index`*, such that condition* `B[index]` *holds, if*

$$(\exists\, \mathtt{ind} \in \mathcal{P}(Q_\rho))\,((\mathtt{index} \subset \mathtt{ind})\,\&\,(\mathtt{U[ind]} = \mathtt{U[index]})),$$

then we set `B[index]:=false`*.*
Step 4. *We select the following set of grids:*

$$Q_{\mathcal{COM}} = \left\{ \mathtt{U[index]} \times \mathtt{index} \,\middle|\, \mathtt{index} \in \mathcal{P}(Q_\rho))\,\&\,\mathtt{B[index]} \right\}.$$

Step 5. $\delta_{\mathcal{COM}}$, $S_{\mathcal{COM}}$ *and* $F_{\mathcal{COM}}$ *are defined by definition of automaton* $\mathcal{COM}(L)$ *given before.* □

Both these algorithms have evident simplified modifications. For obtaining them, let us consider the following directed graph of subsets of the set Q_π:

- for each element $\widetilde{Q} \subseteq Q_\pi$ except $\emptyset$, we have a vertex labeled by $\widetilde{Q}$; this label symbolizes the union of corresponding elements of Q_π;
- we have the edge from $\widetilde{Q}'$ to $\widetilde{Q}''$ (we shall write $\widetilde{Q}' \underset{\mathcal{DG}}{\longrightarrow} \widetilde{Q}''$) if and only if for some $A \in Q_\pi$, we have $\widetilde{Q}' = \widetilde{Q}'' \cup \{A\}$.

Let us denote this directed graph by $\mathcal{DG}(Q_\pi)$. For each its vertex $\widetilde{Q} \in \mathcal{P}(Q_\pi)$, let us define its level by

$$|Q_\pi| - |\widetilde{Q}|;$$

for example, vertex Q_π has level 0, and for each vertex $A \in Q_\pi$, vertex $\{A\}$ has level $|Q_\pi|-1$. Thus, by definitions of grids we obtain the following simplification of Algorithm 2.

Algorithm 3 (Constructing automaton $\mathcal{U}_L$)
Step 1. *Consider Boolean array* `B[index]`, *where* `index` *can be each element of* $\mathcal{P}(Q_\pi)$ *(except* $\emptyset$*). For each possible* `index`, *we set*

$$\texttt{B[index]} := (\exists X \in Q_\rho)\left(\texttt{index} = \left\{ A \in Q_\pi \,\middle|\, A \# X \right\}\right).$$

Step 2.
`for i:=1 to` $|Q_\pi| - 1$ `do`
  `for each` *vertex of level* `i` *(let this vertex be* `index`*)*
    `execute following Step 3`
Step 3.
  `if` *for there exist 2 (or more) vertices* `ind` *of level* `i-1`,
    such that condition `B[ind]` *holds and* $\texttt{ind} \underset{\mathcal{DG}}{\longrightarrow} \texttt{index}$
  `then B[index]:=true`
Step 4. *We select the following set of grids:*

$$Q_{\mathcal{COM}} = \left\{ \texttt{index} \times \bigcap_{A \in \texttt{index}} \left\{ X \in Q_\rho \,\middle|\, A \# X \right\} \,\middle|\, \texttt{index} \in \mathcal{P}(Q_\pi) \,\&\, \texttt{B[index]} \right\}.$$

Step 5. $\delta_{\mathcal{COM}}$, $S_{\mathcal{COM}}$ *and* $F_{\mathcal{COM}}$ *are defined by definition of automaton* $\mathcal{COM}(L)$ *given before.*

Let us remark, that considering subsets of Q_π and Q_ρ (as indexes of arrays), we could consider also element $\emptyset$; in this case, $\emptyset$ would correspond to the possible "dead" state of the equivalent canonical automaton.

For Algorithm 3, we shall not formulate its "mirror image".

6 The detailed example

Let us continue to consider the example of language of [6, Section 3]. For it, let us depict once again automata $\widetilde{L}$ and $\widetilde{L^R}$ for that language (Figure 1 and 2):

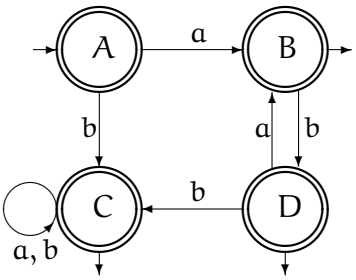


Figure 1

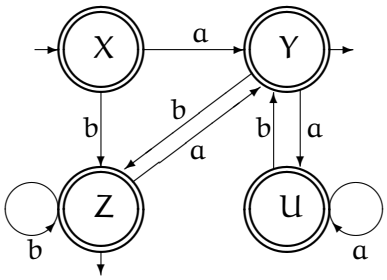


Figure 2

and also its binary relation (Table 1):

#	X	Y	Z	U
A	–	#	#	–
B	#	–	#	–
C	#	#	#	#
D	#	#	#	–

Table 1

Let us remark, that in [6, Section 3] we simply *indicate* the set of grids; and in this paper, we *use the algorithm* of their constructing. Thus, consider using Algorithm 3.

The directed graph $\mathcal{DG}$ for all nonempty subsets of Q_π is given in Figure 3 (the subsets are marked here simply by the strings consisting of their elements). For this figure, we have the following comments. Sets marked by 3 ovals (i.e.,

$\{A, B, C, D\}$, $\{A, C, D\}$, $\{B, C, D\}$ and $\{D\}$) were selected by Step 1 of Algorithm 3.

Using Steps 2 and 3, we consider other subsets (i.e., vertices of graph $\mathcal{DG}$). Considering them, we have the only vertex (i.e., $\{C, D\}$), for which there exist at least 2 vertices, such that we have edges from them in $\{C, D\}$; we marked this "new suitable" vertex by many ovals. Thus, all the 5 mentioned vertices (and only they) are elements $\alpha(\mathcal{B})$ for some grid $\mathcal{B}$.

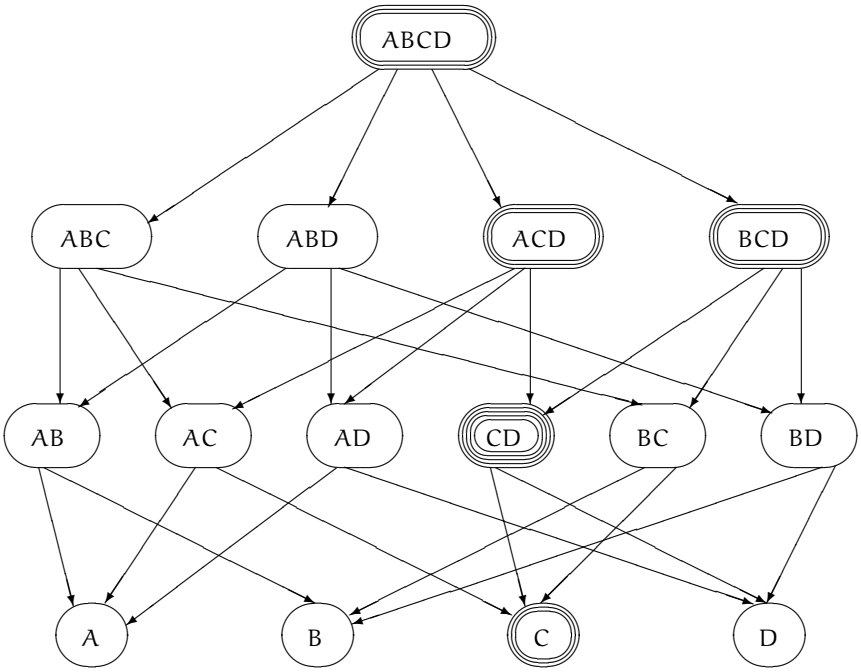


Figure 3

By Step 4 of Algorithm 3, we select for them the following set of grids: [4]

$$\zeta = \{A, C, D\} \times \{Y, Z\},\ \eta = \{A, B, C, D\} \times \{Z\},\ \vartheta = \{B, C, D\} \times \{X, Z\},$$
$$\nu = \{C\} \times \{X, Y, Z, U\}, \quad \xi = \{C, D\} \times \{X, Y, Z\}.$$

These letters (i.e., ζ, η, ϑ, ν and ξ) will symbolize states of automaton $\mathcal{U}_L$.

[4] Remark once again, that in [6] we simply indicate the set of grids. And in this section, we used Algorithm 3 for their construction.

By Step 5 of Algorithm 3, we simply obtain the following sets of inputs and outputs of $\mathcal{U}_L$:

$$S_{\mathcal{COM}} = \{\zeta, \eta\}, \quad F_{\mathcal{COM}} = \{\vartheta, \nu, \xi\}.$$

And also by Step 5 of Algorithm 3 (i.e., by definition of automaton $\mathcal{COM}(L)$ of Section 3), we obtain transition function $\delta_{\mathcal{COM}}$ in the following way.

First at all, consider the following Table 2:

grids ($\mathcal{B}$)	$\alpha(\mathcal{B})$	$\xrightarrow[\delta_\pi]{a}$	ζ	η	ϑ	ν	ξ
ζ	ACD	BC	–	+	+	–	–
η	ABCD	–	–	–	–	–	–
ϑ	BCD	–	–	–	–	–	–
ν	C	C	+	+	+	+	+
ξ	CD	BC	–	+	+	–	–

Table 2

We use here simplified notation as we did before. I.e., elements of the 2nd and the 3rd columns of this table are the sets of elements of Q_π. However, we have to explain this notation detailed, because the symbol "–" in the 3rd column does *not* symbolize the empty set.

Elements of the 2nd column are elements $\alpha(\mathcal{B})$ for the considered grid $\mathcal{B}$ (which is in the 1-st column). If *for each* element of this subset (let this state be A) there exists the transition $A \xrightarrow[\delta_\pi]{a} B$ for some $B \in Q_\pi$, then in this line, the corresponding set of the 3rd column is the union of all such B. [5] Otherwise, i.e., if for some element A (of the 2nd column) there exists *no* transition $A \xrightarrow[\delta_\pi]{a} B$ for some $B \in Q_\pi$, then in the 3rd column, we set the symbol "–".

The right part of this table (i.e., since the 4th column) is a square matrix. For each its element in the line marked ζ and in the column marked η, we set + if and only if:

- the set of 3rd column of the line ζ is not "–";
- and, besides, is a subset of the set of the 3rd column of the line η;

otherwise we set "–". [6] Thus, the right part of the table forms the square

[5] Let us remind, that we consider the letter a. Remind also, that we consider canonical automata *without* "dead" states.

[6] Let us especially remark, that the symbol "–" in the 3rd column implies "–" in each cell of this line in the right part of the table.

matrix (i.e., matrix

$$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 \end{pmatrix} \tag{5}$$

in the considered example), which formulates, in fact, the condition (1) for the letter $\mathfrak{a}$.

And for this letter $\mathfrak{a}$ and automaton $\widetilde{\mathrm{L}^{\mathrm{R}}}$, we obtain the following Table 3:

grids ($\mathcal{B}$)	$\beta(\mathcal{B})$	$\xrightarrow[\delta_\rho]{\mathfrak{a}}$	ζ	η	ϑ	ν	ξ
ζ	YZ	YU	−	−	−	+	−
η	Z	Y	+	−	−	+	+
ϑ	XZ	Y	+	−	−	+	+
ν	XYZU	YU	−	−	−	+	−
ξ	XYZ	YU	−	−	−	+	−

Table 3

Similarly, the right part of this table formulates, in fact, the condition (2) for the same letter $\mathfrak{a}$.

Then the elementwise conjunction of the first matrix (in the considered example, that is matrix (5)) and the transposed matrix of Table 3 gives the matrix for the existence of $\mathfrak{a}$-transitions. Let us remark, that in the considered examples two matrices for elementwise conjunction are the same;[7] however, there exist examples where these matrices are different.

Let us consider such tables for letter $\mathfrak{b}$. For automaton $\widetilde{\mathrm{L}}$, we obtain the following Table 4:

grids ($\mathcal{B}$)	$\alpha(\mathcal{B})$	$\xrightarrow[\delta_\pi]{\mathfrak{b}}$	ζ	η	ϑ	ν	ξ
ζ	ACD	C	+	+	+	+	+
η	ABCD	CD	+	+	+	−	+
ϑ	BCD	CD	+	+	+	−	+
ν	C	C	+	+	+	+	+
ξ	CD	C	+	+	+	+	+

Table 4

And for automaton $\widetilde{\mathrm{L}^{\mathrm{R}}}$, we obtain the following Table 5:

[7] For the letter $\mathfrak{b}$, two corresponding matrices also are the same. See below.

grids ($\mathcal{B}$)	$\beta(\mathcal{B})$	$\xrightarrow[\delta_\rho]{a}$	ζ	η	ϑ	ν	ξ
ζ	YZ	Z	+	+	+	+	+
η	Z	Z	+	+	+	+	+
ϑ	XZ	Z	+	+	+	+	+
ν	XYZU	YZ	+	−	−	+	+
ξ	XYZ	Z	+	+	+	+	+

Table 5

Therefore, we obtain the following matrix of b-transitions for automaton $\mathcal{COM}(L)$:

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix} \tag{6}$$

(the order of the grids is previous, i.e., ζ, η, ϑ, ν, ξ).

Using matrices (5) and (6), we simply obtain the following automaton $\mathcal{COM}(L)$ for the considered language:

	$\mathcal{COM}(L)$	a	b
$\rightarrow$	ζ	η, ϑ	$\zeta, \eta, \vartheta, \nu, \xi$
$\rightarrow$	η	−	$\zeta, \eta, \vartheta, \xi$
$\leftarrow$	ϑ	−	$\zeta, \eta, \vartheta, \xi$
$\leftarrow$	ν	$\zeta, \eta, \vartheta, \nu, \xi$	$\zeta, \eta, \vartheta, \nu, \xi$
$\leftarrow$	ξ	η, ϑ	$\zeta, \eta, \vartheta, \nu, \xi$

Table 6

For this automaton, let us also consider its covering subset of grids. One of them [8] is the following one: $\{\zeta, \vartheta, \nu\}$. And using definition of the covering automaton (Section 3), we obtain the following *corresponding* covering automaton $\mathcal{COM}_{\{\zeta,\vartheta,\nu\}}(L)$ (i.e., the covering automaton for the set $\{\zeta, \vartheta, \nu\}$, see Table 7 and Figure 4). It is easy to prove, that the last automaton does define the given language.

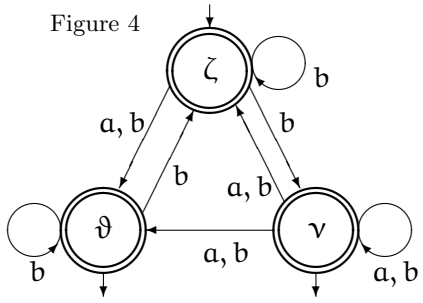
Thus, the considered example gives the *equivalent* covering automaton. However, there are examples of languages, where there exist covering automata which does *not* define the given languages. We shall continue to consider such examples in our following papers.

[8] As we said before, we shall not consider algorithms of constructing such subsets for arbitrary automaton. For our example there is evidently, that there exist *the only* covering subset containing no more than 3 grids.

	$\mathcal{COM}_{\{\zeta,\vartheta,\nu\}}(L)$	a	b
$\rightarrow$	ζ	ϑ	ζ,ϑ,ν
$\leftarrow$	ϑ	–	ζ,ϑ
$\leftarrow$	ν	ζ,ϑ,ν	ζ,ϑ,ν

Table 7

Figure 4



7 A series of examples

In [4, Th. 5.1], tight upper bound $D(k)$ on the size of the universal automaton was obtained. $D(k)$ happens to be the kth Dedekind number ([2] etc.), where k is the number of states of automaton accepting given language L.

In this section we consider examples which shows how fast a number of grids (i.e. the size of the universal automaton) can grow if we are given the sizes of two *canonical* automata (i.e., $\widetilde{L}$ and $\widetilde{L^R}$). Authors think that these examples supplement the examples considered in [4].

First of all, let us consider the following matrix of dimension 12×12:

$$\begin{pmatrix}
1 & \mathbf{0} & \mathbf{0} & \mathbf{0} & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
\mathbf{0} & 1 & \mathbf{0} & \mathbf{0} & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
\mathbf{0} & \mathbf{0} & 1 & \mathbf{0} & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
\mathbf{0} & \mathbf{0} & \mathbf{0} & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & \mathbf{0} & \mathbf{0} & \mathbf{0} & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & \mathbf{0} & 1 & \mathbf{0} & \mathbf{0} & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & \mathbf{0} & \mathbf{0} & 1 & \mathbf{0} & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & \mathbf{0} & \mathbf{0} & \mathbf{0} & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & \mathbf{0} & 1 & \mathbf{0} & \mathbf{0} \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & \mathbf{0} & \mathbf{0} & 1 & \mathbf{0} \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & \mathbf{0} & \mathbf{0} & \mathbf{0} & 1
\end{pmatrix} \tag{7}$$

Evidently, we can consider such matrices for each $x \geq 2$ ($x \times x$ is the dimensions of the "block of zeros") and each $n \geq 2$, such that $n \bmod x = 0$ ($n \times n$ is the dimensions of the matrix). In the above example (7) we have $x = 4$ and $n = 12$. (Moreover, we can take blocks of zeros of different sizes, see below). We shall not write the strict formulas for the elements of such matrices.

For each matrix of this type we can consider corresponding automaton having the same table of #-relation. And by [7], such an automaton always exists.

Constructing *some* grids for the example (7), we can select exactly one line of $\{1,2,3,4\}$, then exactly one line of $\{5,6,7,8\}$, and also exactly one line of $\{9,10,11,12\}$; obviously, the numbers of columns must be the same as the numbers of lines. Thus, in the example (7) we have at least $4^3 = 64$ grids.

Next, let

$$n = a_1 + a_2 + \ldots + a_k,$$

and we have blocks of zeros of the sizes a_1, a_2, ..., a_k; then corresponding number of blocks (states of canonical automaton) will be no less than

$$a_1 \cdot a_2 \cdot \ldots \cdot a_k$$

. Our task is to find the maximum value of this expression with fixed n. To do so, we should take into account that

$$4 = 2 + 2, \quad 2 \cdot 2 = 4 \geq 4;$$
$$5 = 2 + 3, \quad 2 \cdot 3 = 6 > 5;$$
$$6 = 2 + 2 + 2 = 3 + 3, \quad 3 \cdot 3 > 2 \cdot 2 \cdot 2 = 6 \quad \text{etc.}$$

i.e. each summand, greater or equal 4, can be split into 2's and 3's, while keeping the value of product at least as big. It is also obvious, that splitting the summands we should prefer 3's (see the last inequality).

So, the sought-for maximum has one of the following types:

$$3 \cdot 3 \cdot \ldots \cdot 3$$
$$3 \cdot 3 \cdot \ldots \cdot 3 \cdot 2 \cdot 2$$
$$3 \cdot 3 \cdot \ldots \cdot 3 \cdot 2,$$

depending on the n modulo 3. For simplicity's sake we shall limit ourselves with the case $n \bmod 3 = 0$.

To sum up, our examples shows that the size of the universal automaton can grow exponentially (with base $3^{1/3}$) with regard to the size of canonical automata, i.e. with regard to $n = \min(|Q_\pi|, |Q_\rho|)$. Let us repeat that our examples *do not* give the exact number of blocks which may form if we are given only the value of $n = \min(|Q_\pi|, |Q_\rho|)$ for automata $\widetilde{L}$ and $\widetilde{L^R}$.

Note that we must not combine this examples with an obvious bound $n = 2^k$ for the number of states of canonical automata, since the resulting function $3^{n/3} = 3^{2^k/3}$ grows much faster than $D(k)$. This only means that automata with #-relation similar to (7) cannot have equivalent automata with number of states much less then n.

8 Conclusion

In the next paper we are going to consider:

- the loops of the basis automaton $\mathcal{BA}(\mathsf{L})$ and of automaton $\mathcal{COM}(\mathsf{L})$;
- the consideration of the covering automaton, does *not* define the given language (unlike Section 6); i.e., in fact, the consideration of automaton Waterloo ([3]) from the point of view of the basis automaton;
- the *constructive* proof of the following fact: examples like Waterloo can be constructed *for each* table of relation # having the following additional property: there exists the *proper* covering subset.

9 Acknowledgement

Authors would like to express the gratitude to Professor Anton Klyachko (Moscow State University) for making valuable comments.

References

[1] A. Aho, J. Ullman, *The Theory of Parsing, Translation, and Compiling, Vol. 1, Parsing.* Prentice Hall, 1972. ⇒8

[2] R. Dedekind, *Über Zerlegungen von Zahlen durch ihre größten gemeinsamen Teiler.* Gesammelte Werke, Vol. **2**, 103–148. ⇒18

[3] T. Kameda, P. Weiner, On the state minimization of nondeterministic finite automata, *IEEE Trans. on Comp.* **C-19,** 7 (1970) 617–627. ⇒20

[4] S. Lombardy, J. Sakarovitch, The Universal Automaton, in: *Logic and Automata, Texts in Logic and Games* Amsterdam Univ. Press. Vol. **2** (2008) 457–504. ⇒ 5, 7, 18

[5] B. Melnikov, Extended nondeterministic finite automata, *Fundamenta Informaticae* **104,** 3 (2010) 255–265. ⇒5, 6

[6] B. Melnikov, Once more on the edge-minimization of nondeterministic finite automata and the connected problems, *Fundamenta Informaticae* **104,** 3 (2010) 267–283. ⇒5, 6, 9, 10, 13, 14

[7] B. Melnikov, A. Melnikova, Some more on the basis finite automaton, *Acta Univ. Sapientiae, Inform.* **5**, 2 (2013) 227–244 ⇒19

[8] B. Melnikov, N. Sciarini-Guryanova, Possible edges of a finite automaton defining a given regular language, *The Korean Journal of Computational and Applied Mathematics* **9,** 2 (2002) 475–485. ⇒5

Received: June 9, 2013 • Revised: August 2, 2013

Acta Univ. Sapientiae, Informatica, **6,** 1 (2014) 21–32

On linear programming duality and Landau's characterization of tournament scores

Allan B. CRUSE
Mathematics Department
University of San Francisco
San Francisco, California 94117
email: cruse@usfca.edu

Abstract. H. G. Landau has characterized those integer-sequences $S = (s_1, s_2, \ldots, s_n)$ which can arise as score-vectors in an ordinary round-robin tournament among n contestants [17]. If $s_1 \leq s_2 \leq \cdots \leq s_n$, the relevant conditions are expressed simply by the inequalities:

$$\sum_{i=1}^{k} s_i \geq \binom{k}{2}, \tag{1}$$

for $k = 1, 2, \ldots, n$, with equality holding when $k = n$. The necessity of these conditions is fairly obvious, and proofs of their sufficiency have been given using a variety of different methods [1, 2, 4, 10, 22, 23]. The purpose of this note is to exhibit Landau's theorem as an instance of the "duality principle" of linear programming, and to point out that this approach suggests an extension of Landau's result going beyond the well-known generalizations due to J. W. Moon [20, 19].

1 Background

In an ordinary round-robin tournament, there are n contestants, each of whom plays exactly one game against each other contestant, and no game is permitted to end in a tie. For a survey of results on tournaments and their generalizations, the reader is referred to [12] and [21].

Computing Classification System 1998: G.2.2.
Mathematics Subject Classification 2010: 05C85, 68R10
Key words and phrases: Landau's theorem, tournament scores, linear programming
DOI:10.2478/ausi-2014-0016

The results in such a tournament can be represented by an $n \times n$ matrix $T = (t_{ij})$ of zeros and ones, called a *tournament matrix*, in which $t_{ij} = 1$ if the i-th contestant defeats the j-th contestant, and $t_{ij} = 0$ otherwise. It is easy to see that the set of all $n \times n$ tournament matrices is identical to the set of all *integer* solutions to the following system of linear relations (in which i and j represent arbitrary *distinct* indices):

$$t_{ij} \geq 0, \tag{2}$$

$$t_{ii} = 0, \tag{3}$$

and

$$t_{ij} + t_{ji} = 1. \tag{4}$$

The i-th contestant's *score* s_i is the total number of games played in which the i-th contestant is the victor. Note that s_i may be obtained by summing up the entries in the i-th row of the matrix T:

$$s_i = \sum_{j=1}^{n} t_{ij}. \tag{5}$$

The sequence $S = (s_1, s_2, \ldots, s_n)$ consisting of all the contestants' scores is called the *score-vector* for the tournament. Clearly, the sum of all the scores in a score-vector must equal the total number of games played in the tournament:

$$\sum_{i=1}^{n} s_i = \binom{n}{2}. \tag{6}$$

Furthermore, any subset of the contestants taken together must score a total number of wins at least as large as the number of games they play with each other; hence, for $k = 1,\ 2,\ \ldots, n$, the inequality

$$\sum_{i \in K} s_i \geq \binom{k}{2} \tag{7}$$

must hold for each k-element subset K of $\{1, 2, \ldots, n\}$.

Landau's theorem, referred to in the introduction, asserts that these relations (6)–(7) completely characterize those integer-sequences which are score-vectors.

Theorem 1 (H. G. Landau [17]) *For an arbitrary integer-sequence* $S = (s_1, s_2, \ldots, s_n)$ *to be the score-vector of some round-robin tournament among* n *contestants, it is necessary and sufficient that, for* $1 \le k \le n$*, the inequality*

$$\sum_{i \in K} s_i \ge \binom{k}{2} \tag{8}$$

shall hold for each k-element subset K *of* $\{1, 2, \ldots, n\}$*, and moreover that strict equality shall hold when* $k = n$*.*

We remark that, in order to test a sequence $S = (s_1, s_2, \ldots, s_n)$ according to the criteria in Theorem 1, it might seem that a check of 2^n inequalities is required; but, as Landau himself pointed out, if the elements $s_1, s_2, \ldots, s_n$ are first arranged in nondescending order, then only n relations actually have to be examined. Thus the crucial relations are indicated in the following

Corollary 2 (Landau [17]) *For the integer-sequence* $S = (s_1, s_2, \ldots, s_n)$ *to be a score-vector, where* $s_1 \le s_2 \le \cdots \le s_n$*, it is both necessary and sufficient that the inequality*

$$\sum_{i=1}^{k} s_i \ge \binom{k}{2} \tag{9}$$

shall hold for each $k \le n$*, with strict equality when* $k = n$*.*

Since it is always possible, in at most $\binom{n}{2}$ steps, to rearrange the elements of any n-term sequence so that they appear in non-descending order, this corollary provides the basis for an efficient algorithm to detect score-vectors. To deduce Corollary 2 from Theorem 1 is easy: simply observe that the sum of any k elements chosen from a finite set must be as least as large as the sum of the k smallest elements in that set.[1]

J. W. Moon in [20] has extended Landau's theorem by referring to arbitrary real solutions of the system (2)–(4) as *generalized tournaments*. Scores for the contestants in a generalized tournament are defined by (5) and need not be integers, although such scores still must satisfy the relations (6)–(7), since these relations actually are algebraic combinations of (2)–(5). Moon's result, which closely parallel's Landau's theorem, may be phrased as follows:

[1] For an interesting discussion of algorithmic efficiency, the reader may consult the popular survey article [18].

Theorem 3 ((J. W. Moon [20]) *For an arbitrary real-sequence* $S = (s_1, s_2, \ldots, s_n)$ *to be the score-vector of some generalized tournament of size* $n \times n$*, it is necessary and sufficient that, for* $1 \leq k \leq n$*, the inequality*

$$\sum_{i \in K} s_i \geq \binom{k}{2} \tag{10}$$

shall hold for each k*-element subset* K *of* $\{1, 2, \ldots, n\}$*, and moreover that strict equality shall hold when* $k = n$*.*

Interestingly, from the point of view of linear programming, as we shall see, Landau's original theorem may be regarded as a somewhat deeper result than the apparently more general theorem due to Moon. The explanation for this opinion is that Landau's theorem rests with greater weight upon a special property of the linear constraint-system (2)–(5) known as "total unimodularity". The significance of this property for integer linear programming is indicated in the next section.

The structure of this paper is as follows: in Section 1 we present the terminology and notation for restating the theorems of Landau (for tournaments) and Moon (for generalized tournaments) which did not originally rely on linear programming methods; in Section 2 we state the "duality" and "unimodularity" principles which allow us to see afresh the theorems of Landau and Moon as special instances of linear programming principles; in Section 3 we show how proofs for their theorems can be given via "duality" and "unimodularity"; and in Section 4 we point out the advantage of a linear programming perspective, namely, a natural extension of their theorems to more general structures, called C-tournaments, and we speculate that the linear programming point-of-view offers a potential for discovering new results regarding other combinatorial objects. We conclude in Section 5 with a summary and some acknowledgements.

2 Duality and unimodularity

In the argument which follows we shall employ the so-called "duality principle" of linear programming. Complete discussions of this principle may be found in most standard textbooks, such as [9] or [11]. The version needed for our purposes relates the following pair of optimization problems built out of the same data, namely, a $p \times q$ matrix $A = (a_{ij})$, a p-vector $B = (b_i)$, and a q-vector $C = (c_j)$:

Maximum problem	Minimum problem
Maximize $\sum_{j=1}^{q} c_j x_j$	Minimize $\sum_{i=1}^{p} b_i y_i$
constrained by	constrained by
$x_j \geq 0$	$\sum_{i=1}^{p} a_{ij} y_i \geq c_j \quad (1 \leq j \leq q)$
$\sum_{j=1}^{q} a_{ij} x_j \leq b_i$	$y_i \geq 0 \quad (1 \leq i \leq p)$.

The *duality principle* asserts that, if the maximum problem is solvable, then the minimum problem also is solvable, and the constrained maximum of $\sum c_j x_j$ equals the constrained minimum of $\sum b_i y_i$.

Besides the duality principle we shall also require certain further facts from linear programming. As is well known, in any linear programming problem, the optimal value of the objective function (if it exists) is attained at a vertex, or "extreme point," of the polyhedral convex set of feasible solutions (see, for example [9] or [11]). Each such vertex arises as a *basic solution* of the linear inequalities which define the feasible region; that is, by choosing an appropriate subset of the inequalities, and then solving these simultaneously as if they were linear equations. Accordingly, as a consequence of Cramer's Rule, if the constraint-matrix A in the above pair of optimization problems happens to be *totally unimodular* (i.e., every square submatrix of A of every order has a determinant equal to 0, $+1$, or -1), and if the given vectors B and C are composed of integers, then in both problems the optimal value of the objective functions will be attained at integral solution-vectors $X = (x_j)$ and $Y = (y_i)$ (see [14]).

In general, it is not easy to tell whether a given matrix A is totally unimodular, although an obvious requirement in view of the definition is that the individual entries a_{ij} must themselves be equal to $0, +1$, or -1. A complete characterization of totally unimodular matrices (in terms of forbidden submatrices) has been given by P. Camion in [6]. A simpler criterion which is often useful is the following sufficient condition due to Heller and Tompkins.

Theorem 4 (Heller and Tompkins [13]) *In order for the matrix* $A = (a_{ij})$ *to be totally unimodular, the following three conditions are sufficient:*

(1) Each entry a_{ij} *is* $0, +1$*, or* -1*.*

(2) At most two nonzero entries appear in any column of A*.*

(3) The rows of A *can be partitioned into two subsets* R_1 *and* R_2 *such that:*

(i) If a column contains two nonzero entries with the same sign, then one entry is in a row of R_1 *and one entry is in a row of* R_2*.*

(ii) If a column contains two nonzero entries of opposite sign, then both entries are in rows of R_1 *, or both entries are in rows of* R_2*.*

For later reference we note here, in stating this criterion, the words "row" and "column" could be interchanged throughout, since it is obvious from the definition that a matrix A is totally unimodular if and only if its transpose A^t is also totally unimodular. With these results freshly in mind, we proceed to our proof of the Moon and Landau theorems.

3 Proof of Landau's theorem

Assume that $S = (s_1, s_2, \ldots, s_n)$ is an arbitrary real sequence satisfying the relations (6)–(7). We wish to show, first, that there exists a real $n \times n$ matrix $T = (t_{ij})$ satisfying (2)–(5); and further, that if S happens to be composed on integers, then T may be assumed to consist of integers as well. The first statement yields Moon's theorem (Theorem 3), while the second assertion gives the original theorem of Landau (Theorem 1). To achieve their proofs, we consider the following linear programming problem:

$$\text{Maximize} \quad z = \sum_{i=1}^{n} \sum_{j=1}^{n} x_{ij} \tag{11}$$

subject to the constraints

$$x_{ij} \geq 0, \quad \text{for} \quad 1 \leq i, j \leq n \tag{12}$$

$$x_{ii} \leq 0, \quad \text{for} \quad 1 \leq i \leq n \tag{13}$$

$$x_{ij} + x_{ji} \leq 1, \quad \text{for} \quad 1 \leq i < j \leq n \tag{14}$$

$$\sum_{j=1}^{n} x_{ij} \leq s_i, \quad \text{for} \quad 1 \leq i \leq n. \tag{15}$$

Notice that these constraints have at least one feasible solution (e.g., the zero matrix) since the inequalities (7) imply that the numbers s_i are all nonnegative;

and since the set of all feasible solutions is evidently closed and bounded, an optimal solution must exist. Indeed, by adding all inequalities of type (15), we can see from (6) that $\max z \le \binom{n}{2}$. Let us now show that in fact $\max z = \binom{n}{2}$. It is for this purpose that we utilize the principle of duality.

Consider the following minimum problem (which is the dual of the maximum problem above):

$$\text{Minimize} \quad \sum_{i=1}^{n} s_i u_i + \sum_{1 \le i < j \le n} v_{ij} \tag{16}$$

subject to the constraints

$$u_i \ge 0, \quad \text{for} \quad 1 \le i \le n, \tag{17}$$

$$v_{ij} \ge 0, \quad \text{for} \quad 1 \le i \le j \le n, \tag{18}$$

$$u_i + v_{ij} \ge 1, \quad \text{for} \quad 1 \le i < j \le n, \tag{19}$$

$$u_j + v_{ij} \ge 1, \quad \text{for} \quad 1 \le i < j \le n, \tag{20}$$

$$u_i + v_{ii} \ge 1, \quad \text{for} \quad 1 \le i \le n. \tag{21}$$

By the fundamental duality principle of linear programming, we know that $\min y = \max z$. So let us now show that $\min y < \binom{n}{2}$ is impossible.

Suppose, on the contrary, that we did have $\min y < \binom{n}{2}$. Then there would have to exist a solution-vector $(\overline{u}_i, \overline{v}_{ij})$ satisfying the constraints (17)–(21) such that

$$y = \sum_{i=1}^{n} s_i \overline{u}_i + \sum_{1 \le i < j \le n} \overline{v}_{ij} < \binom{n}{2}, \tag{22}$$

and we may assume (as explained in the preceeding section) that this vector $(\overline{u}_i, \overline{v}_{ij})$ is an extreme point of the polyhedral convex set defined by the constraints (17)–(21). For this polyhedron, the extreme points are particularly easy to describe.

Lemma 5 *If $(\overline{u}_i, \overline{v}_{ij})$ is any extreme point of the convex polyhedron defined by* (17)–(21)*, then:*

(i) The components of $(\overline{u}_i, \overline{v}_{ij})$ are zeros and ones.

(ii) If $K = \{i : \overline{u}_i = 1\}$, then $\overline{v}_{ij} = 0$ if and only if $i \in K$ and $j \in K$.

Proof. It is evident from an inspection of the constraints (17)–(21) that the vector $(\overline{u}_i, \overline{v}_{ij})$ cannot be extremal if it contains any entries larger than 1, since all such entries can be either increased or decreased by a small amount

without violating the constraints. Thus to prove (i) it suffices to show that $(\overline{u}_i, \overline{v}_{ij})$ cannot be extremal unless it is composed entirely of integers. But this fact follows at once from Cramer's Rule and the "unimodular property" of the constraints (17)–(21). (The criterion of Theorem 4 may be used to detect this property.) Alternatively, a direct argument inspired by Hoffman and Kuhn [15] may be given as follows.

Suppose $(\overline{u}_i, \overline{v}_{ij})$ is a vector satisfying (17)–(21) which contains some non-integer entries. Then for $e \neq 0$ let $(u'_i, v'_{ij})^e$ be the vector defined by

$$u'_i = \begin{cases} \overline{u}_i & \text{if } \overline{u}_{ij} \text{ is an integer,} \\ \overline{u}_i + e & \text{otherwise,} \end{cases} \tag{23}$$

$$v'_{ij} = \begin{cases} \overline{v}_{ij} & \text{if } \overline{v}_{ij} \text{ is an integer,} \\ \overline{v}_{ij} - e & \text{otherwise.} \end{cases} \tag{24}$$

Evidently both $(u'_i, v'_{ij})^e$ and $(u'_i, v'_{ij})^{-e}$ will satisfy(17)–(21) for a sufficiently small choice of $e > 0$, since the sum of two numbers cannot equal an integer if exactly one of them is non-integer. Now since $(\overline{u}_i, \overline{v}_{ij})$ can be written as

$$\overline{u}_i, \overline{v}_{ij}) = \frac{1}{2}(u'_i, v'_{ij})^e + \frac{1}{2}(u'_i, v'_{ij})^{-e},$$

we see that the vector $(\overline{u}_i, \overline{v}_{ij})$ is not extremal, which proves *(i)*.

Property *(ii)* is now apparent from an inspection of the constraints. □

From this Lemma we see that inequality (22) may be rewritten as

$$y = \sum_{i \in K} s_i + \left[\binom{n}{2} - \binom{k}{2} \right] < \binom{n}{2}, \tag{25}$$

where k denotes the cardinality of the set $K = \{i : \overline{u}_i = 1\}$. But this relation is clearly inconsistent with our hypothesis that the given sequence $S = (s_1, s_2, \ldots, s_n)$ satisfies (7). This contradiction shows that $\min y \geq \binom{n}{2}$, and so we must indeed have $\min y = \max z = \binom{n}{2}$.

Having established the existence of an $n \times n$ matrix $X = (x_{ij})$ satisfying (11)–(14) with $\sum_{i=1}^{n} \sum_{j=1}^{n} x_{ij} = \binom{n}{2}$, we note that this matrix X must satisfy all of the constraints (12)–(14) as actual equations. For otherwise, if any one of the relations (12)–(14) holds for X as a strict inequality, then addition of all of those constraints would yield the relation

$$2\left(\sum_{i=1}^{n} \sum_{j=1}^{n} x_{ij} \right) < n(n-1),$$

thereby contradicting the choice of X. Finally, since we may also assume that the matrix X is an extreme point of the convex polygon defined by (11)–(15), then in case the sequence $S = (s_1, s_2, \ldots, s_n)$ was composed entirely of integers, we may invoke the "unimodular property" once again to infer that X is actually a matrix of zeros and ones. This shows that X represents a (generalized) tournament having the given sequence S as its score-vector. The proof of Landau's theorem is now complete, and with it the proof of Moon's generalization.[2]

4 Generalization to C-tournaments

The approach taken in the preceding argument may be followed in a more general setting. Let $C = (c_{ij})$ be any upper-triangular $n \times n$ matrix of non-negative integers, and consider the set of all integer solutions $T = [t_{ij}]$ to the following linear system:

$$t_{ij} \geq 0, \quad \text{for} \quad 1 \leq i, j \leq n, \tag{26}$$

$$t_{ii} = c_{ii}, \quad \text{for} \quad 1 \leq i \leq n, \tag{27}$$

$$t_{ij} + t_{ji} = c_{ij}, \quad \text{for} \quad 1 \leq i < j \leq n. \tag{28}$$

Such an integer solution-matrix T will be called a C-*tournament* since it is plausible to interpret T as a record of the wins and losses in an expanded type of tournament competition where the i-th contestant plays an arbitrarily predetermined number of games against the j-th contestant. For example, C-tournaments include the so-called "n-partite tournaments" introduced by Moon in [21]. (An n-partite tournament differs from an ordinary tournament in that there are n nonempty sets of players P_1, P_2, ..., P_n, and two of the players compete if and only if they do not belong to the same set P_i.) Scores for the contestants in a C-tournament are defined in the same way as for ordinary tournaments, and it is clear that by modifying our proof of Landau's theorem in only a few details, we can immediately obtain a characterization for the score-vectors which may arise from a given choice of the matrix C.

[2]The argument presented here is similar in spirit and in certain details to the proof via linear programming of a theorem on systems of distinct representatives due to Hoffman and Kuhn [15], and to a proof by the author of a theorem due to D. R. Fulkerson which characterizes permutation matrices [7]. Still other combinatorial theorems whose proofs follow this same pattern are treated by various authors in [16].

Theorem 6 *For an arbitrary integer sequence* $S = (s_1, s_2, \ldots, s_n)$ *to be the score-vector of some* C*-tournament among* n *contestants, where* C *is a given* $n \times n$ *upper-triangular matrix of nonnegative integers, it is both necessary and sufficient that the inequality*

$$\sum_{i \in K} s_i \geq \sum_{i \in K} \sum_{j \in K} c_{i,j}$$

shall hold for each subset K *of* $\{1, 2, \ldots, n\}$*, and moreover that strict equality shall hold when* $K = \{1, 2, \ldots, n\}$*.*

Since our proof of Theorem 6 is practically the same as the proof just given, we do not repeat those details.

We remark that Theorem 6, which is the main new result of this note, reduces to Landau's theorem (Theorem 1) in case C is the upper-triangular matrix of zeros and ones in which $c_{ij} = 1$ if and only if $i < j$. Theorem 6 also encompasses the characterization of score-vectors for n-partite tournaments which was obtained with different methods by Moon in [19]. Finally we mention that Theorem 6 extends to "generalized" C-tournaments simply by dropping the requirement that the matrices C and T must be composed of integers.

It seems likely that several other problems which arise in the theory of tournaments may be amenable to the methods of linear-programming illustrated here. One example which suggests itself is the problem of determining the number of "upsets" which can occur in a tournament having a prescribed score-vector S. (An *upset* occurs when one contestant defeats another whose record of wins is better (or at least no worse).) For ordinary tournaments this problem was completely solved by D. R. Fulkerson in [10].

Earlier H. J. Ryser in [23] had obtained an explicit formula for the minimum number of upsets that must occur in any tournament with score-vector $S = (s_1, s_2, \ldots, s_n)$, where $s_1 \leq s_2 \leq \cdots \leq s_n$, namely,

$$\sum_{i \in J} [s_i - (i - 1)], \tag{29}$$

where $J = \{i : s_i \geq i - 1\}$. Although Ryser's methods were completely combinatorial, one can hardly help noticing that this problem asks for the optimum of a certain linear function defined over a convex polyhedron which, in view of the "unimodular property." can have only integral vertices.

5 Summary

Landau's pioneering investigation of round-robin tournaments can be seen as a special instance of linear programming constrained to integer variables, and Moon's generalization as a "relaxation" of the integrality constraint. Placing combinatorial studies within that broader linear algebra setting not only yields immediate new generalizations, such as our 6, where conditions for recognizing the score-vectors of C-tournaments are deduced as straightforward consequences of "duality" and "unimodularity", but suggest a tantalizing way to explore various other seemingly unrelated questions.

In January 2014 we learned of the recent work by R. Brualdi and E. Fritscher [5] in which an algorithm is presented for constructing one (or more) C-tournaments having a prescribed score-vector in all cases where that is possible, or else exhibiting a specific constraint among those in our 6 which is violated.

Acknowledgements. The original version of this paper was written during Spring 1978 while the author was Visiting Associate Professor in the Mathematics Department of Emory University. The author expresses thanks to Professors C. M. Bang and H. Sharp, Jr. for introducing him to the subject of tournaments and providing a pre-publication copy of [3].

The author also expresses thanks to Professor Antal Iványi of Eötvös Loránd University for encouraging him to seek a polynomial-time algorithm based on Theorem 6 which has been posted online at this URL: `http://www.cs.usfca.edu/~cruse/recog.html` [8].

References

[1] G. G. Alway, Matrices and sequences, *Math. Gazette* **46** (1962) 208–213. ⇒21

[2] C. M. Bang, H. Sharp, An elementary proof of Moon's theorem on generalized tournaments, *J. Comb. Theory, Ser. B* **22,** 3 (1977) 299–301. ⇒21

[3] C. M. Bang, H. Sharp, Score vectors of tournaments, *J. Comb. Theory, Ser. B* **26,** 1, (1979) 81–84. ⇒31

[4] A. Brauer, I. C. Gentry, K. Shaw, A new proof of a theorem by H. G. Landau on tournament matrices, *J. Comb. Theory* **5,** 3 (1968) 289–292. ⇒21

[5] R. A. Brualdi, E. Fritscher, Tournaments associated with multigraphs, *Discrete Math.* 2014, 17 pages (submitted). ⇒31

[6] P. Camion, Characterization of totally unimodular matrices, *Proc. Amer. Math. Soc.* **16,** 5 (1965) 1068–1073. ⇒25

[7] A. B. Cruse, A proof of Fulkerson's characterization of permutation matrices, *Lin. Alg. Appl.* **12,** 1 (1975) 21–28. ⇒29
[8] A. B. Cruse, Polynomial algorithm for testing potential C-tournaments. http://www.cs.usfca.edu/ cruse/recog.html. ⇒31
[9] G. B. Dantzig, *Linear Programming and Extensions*, Princeton University Press, Princeton, NJ, 1963. ⇒24, 25
[10] D. R. Fulkerson, Upsets in round robin tournaments, *Canad. J. Math.* **17** (1965) 957–969. ⇒21, 30
[11] G. Hadley, *Linear Programming*, Addison-Wesley, Reading, MA, 1962. ⇒ 24, 25
[12] F. Harary, L. Moser, The theory of round robin tournaments, *Amer. Math. Monthly* **73,** 3 (1966) 231–246. ⇒21
[13] I. Heller, C. B. Tompkins, An extension of a theorem of Dantzig's, in: G. B. Dantzig, *Linear Programming and Extensions*, Princeton University Press, Princeton, NJ, 1963., pp. 247–254. ⇒25
[14] A. J. Hoffman, J. B. Kruskal, Integral boundary points of convex polyhedra, in: *Linear Inequalities and Related Systems* (Study No. **38** of the Princeton Annals of Mathematics, ed. H. W. Kuhn and A. W. Tucker), Princeton University Press, Princeton, NJ, 1957, pp. 223–246. ⇒25
[15] A. J. Hoffman, H. W. Kuhn, Systems of distinct representatives and linear programming, *Amer. Math. Monthly* **63** (1956) 455–460. ⇒28, 29
[16] H. W. Kuhn, A. W. Tucker (eds.), *Linear Inequalities and Related Systems* Ann. Math. Studies **38**, Princeton University Press, Princeton, NJ, 1956. ⇒29
[17] H. G. Landau, On dominance relations and the structure of animal societies. III. The condition for a score structure, *Bull. Math. Biophys.* **15** (1953) 143–148. ⇒ 21, 23
[18] H. L. Lewis, C. H. Papadimitriou, The efficiency of algorithms, *Scientific American*, **238** (1978) 96–109. ⇒23
[19] J. W. Moon, On the score sequence of an $\mathfrak{n}$-partite tournament, *Canad. Math. Bull.* **5,** 1 (1962) 51–58. ⇒21, 30
[20] J. W. Moon, An extension of Landau's theorem on tournaments, *Pacific J. Math.* **13,** 4 (1963) 1343–1345. ⇒21, 23, 24
[21] J. W. Moon, *Topics on Tournaments*, Holt, Rinehart, and Winston, New York, 1968. ⇒21, 29
[22] H. J. Ryser, Matrices of zeros and ones. *Bull. of Amer. Math. Soc.* **66,** 6 (1960) 442–464. ⇒21
[23] H. J. Ryser, Matrices of zeros and ones in combinatorial mathematics, in: *Recent Advances in Matrix Theory*, Proc. Advances Matrix Theory (ed. H. Schneider), University of Wisconsin Press, Madison, WI, 1964, pp. 103–124. ⇒21, 30

Received: January 25, 2014 • Revised: February 28, 2014

Acta Univ. Sapientiae, Informatica, **6,** 1 (2014) 33–55

Survey on privacy preserving data mining techniques in health care databases

Tamás Zoltán GÁL
U1 Research
2 INFOPARK Gábor Dénes
Budapest, Hungary
email: galt@u1research.org

Gábor KOVÁCS
U1 Research
2 INFOPARK Gábor Dénes
Budapest, Hungary
email: kovacsg@u1research.org

Zsolt T. KARDKOVÁCS
U1 Research
2 INFOPARK Gábor Dénes
Budapest, Hungary
email: kardkovacs@u1research.org

Abstract. In health care databases, there are tireless and antagonistic interests between data mining research and privacy preservation, the more you try to hide sensitive private information, the less valuable it is for analysis. In this paper, we give an outlook on data anonymization problems by case studies. We give a summary on the state-of-the-art health care data anonymization issues including legal environment and expectations, the most common attacking strategies on privacy, and the proposed metrics for evaluating usefulness and privacy preservation for anonymization. Finally, we summarize the strength and the shortcomings of different approaches and techniques from the literature based on these evaluations.

Computing Classification System 1998: H.1.1. H.2.8. K.4.1.
Mathematics Subject Classification 2010: 97-02, 97R50, 97R40, 62P10, 68P20
Key words and phrases: anonymization, privacy preservation, health care database, survey, data mining, sensitive data
DOI:10.2478/ausi-2014-0017

1 Introduction

Databases that contain useful information about people have always been in the focus of research. Researchers apply various methods to extract valuable information from data sets to understand people and make predictions for their future behaviour. While legal systems may vary over countries, democratic regulations protect privacy at the highest level, usually in their constutition. Specially, a specific type of personal data called sensitive data, e.g. ethnicity, religious affiliation, medical condition, can only be accessed, transfered or handled by entities explicitly stated in regulations, and with the consent of the data subject.

Health care databases have an especially strict regulation because of the large number of sensitive data contained. For instance, pharmaceutical research must work with accurate data, but that retains all sensitive patient data as well, hence researchers working with such databases stumble very early in the legal limitations. Records of health care databases hold sensitive information from which one may be able to reveal medical condition of a person. Medical conditions may relate to e.g. food consumption preferences, life expectancy, drug taking habits, and other personal strengths or weaknesses. In wrong hands, e.g. decisions over employments [27], or mortgages might depend on such information which would be very unethical to use, and it must be avoided at all costs. On the other hand, health care databases also serve as the basis for better health care services, drug developments, and cost efficiency which also are in the focus of public interests. Therefore before publishing any piece of information from the database, it has to go through an anonymization procedure to hide sensitive data. Hence researchers must be aware of the legal requirements, the methods applicable to meet these requirements and the level protection these techniques provide.

One may take into account that neither well-known personal identifiers like birth name, social security number nor sensitive data on medical statements on their own harm privacy; only making connections between these pieces of information. That is, the main task called data anonymization is to prevent from establishing of connection between individuals and their data. Data anonymization can be forced by physically limit the data access by means of security policies, deletion, data perturbations, or by guaranteeing that any piece of data could be connected to more than one individual using repetition, sampling, aggregations, etc. As a consequence, data quality is reduced.

Reduced data quality for data analysis means somekind of loss in usefulness which directly affects the performance of data analyis. For example, data

mining procedures as the most commonly used data analysis tools aim at discovering valid, previously unknown (novel or hidden), potentially useful, understandable (actionable) patterns, information, or relationships in statistically large databases [12]. Data mining tools highly dependent from data quality, i.e., poor data quality may result in invalid, useless, empty or non-comprehensible knowledge discovery. We propose a novel metric, the accuracy to be used for evaluating the usefulness of the anonymized data instead of information loss metrics.

The question arises how to enable the extraction of useful and beneficial information from health care databases while maximizing the protection of privacy. In this paper, we review different aspects of data mining related data anonymization and privacy preserving data mining, and we analyze the question from legalislative, privacy intruder, and data owner points of view. We also investigate what level of protection existing health care anonymization methods provide by comparing them to general techniques, and point out their limitations showing additional aspects to be covered when protecting health data.

The paper is organized as follows. Section 2 makes an overview of past events, which helps to give and overview on the importance and motivations of data anonymization. We discuss legal regulations and limitation regarding sensitive data management in Section 3. Section 4 defines key terms used in this article, and it presents a novel classification of the different approaches to data anonymization by analyzing possible attacking techniques and motivations. We make a short summary on the most analyzed data anonymization techniques and illustrate them on example databases in Section 5. Section 6 gives a brief introduction to theoretical indicators of privacy preservation and data utility metrics. We also evaluate how data anonymization techniques perform on our examples.

2 Leaking examples: threats and motivations

In theory, private data are the most protected, only those who have specific permission can access them. Nevertheless, in United States the 35% of the employers had a deep insight of health records of employees to make decisions about them according to [27] made in the 1990's [27]. There is also an urban legend [13] about a bank officer in Maryland who cross-referenced a list of patients with cancer against a list of people with recallable mortgages [27]. But health care practioners have many another ways to harm patients' privacy

even after several regulations on the topic [6].

The problem is not new, the first census in the United States faced similar problems. For example, for electoral registry four typical data is given: sex, date of birth, postal code, ethnicity. In Cambridge, MA in 1997 there were 54.804 voters, and 12% of whom was uniquely identified by date of birth, 29% by date of birth and sex, 69% by date of birth and not full postal code, and 97% by date of birth and full length postal code [27]. Similar results are observed in other countries, as well [18, 10].

But this is not the only way to get sensitive data. In 2005, Netflix published a believed-to-be de-identified movie preference database of 480.000 customers. Database contained information on pseudo-anonymized ids, movie ids and titles, dates, and preferences. For those who had more than eight evaluations for movies there was a 99% match against publicly available IMDB (Internet Movie Database) evaluations [25]. Login names and real names are overlapping hence a huge number of customers could have been directly identified using these information. In 2006, AOL published another database on pseudo-anonymized web search queries. [4] found out that web search queries tend to contain the surfers' name, social security number, or other private information about them. In this article, they retrieved a picture of a 62-year-old web surfer believed to be anonym.

Obviously, storing, transferring, and handling private data is strictly forbidden in general. Nevertheless, it is unavoidable in many cases specially for public services. This implies that unencrypted or decrypted personal data can be accessed through hacking or other intrusion techniques.

For example, researchers fights for new discoveries as the basis of their promotions and livelihood. They are sometimes careless on handling properly personal data and it is not infrequent that data is not deleted after a research is completed. Can personal data be provided or be available for research? If ethical norms can be passed for some reasons there is no strict boundary which leads to gradual destruction and degradation of privacy norms [7].

Sometimes, if data subjects give specific permission to an organization to handle their private data could lead to privacy issues. For instance, customers allow a supermarket to handle their data on custom behaviors. Data mining on customer baskets, transactions are commonly used for data mining in order to reveal customer preferences, and to increase revenue with proper marketing communication and logistical strategies. Collected data may be shared legally with suppliers to help product development. In this case, data sharing may lead to revenue maximization on the supplier side only if data indicate clear preference over the competitive products. In addition, shared data based mar-

keting strategy could increase the profit in all but the data sharer company [30].

People in democratic countries have a strong need for privacy which must be protected at all costs. Researchers and other data consumers need a uniformly available dataset for innovation, and obviously a clear, ethical way to acquire those valueble information. What motivates data publishers in the use data anonymization? Data sharing is essential in the information era for product development, innovation and research, i.e. information and databases are just another types of raw materials. Firstly, data sharers wants to comply with regulations in order to re-sell databases. Secondly, they want to avoid responsibility for non-intensional loss in privacy. Thirdly, they try to keep the balance between revenues of data sharing and the risk of misuse. Finally, there are also anti-trust worries, that is why data sharing among competitors is regulated.

These examples clearly illustrate that there is a strong need for data publishing, for protecting privacy and data providers in a hand.

3 Privacy, public interests and legal regulations

In order to understand, what may harm privacy, we need to defined the notion privacy itself. In democratic countries, people have rights to be left alone, their actions and data shall be handled confidentially unless other, higher or public interests do not require otherwise. From this point of view, every action against an individual's specific will, or which disturbs private life including unreasonable publicity harm privacy.

Nevertheless, boundaries of the term ,,public interest" may vary in countries and it is hard to say where begins or ends private and social life. That is why interpretation of the right to privacy is not straightforward. It is a fact that birthday is a private event while wedding, hence marital status, is not because there are public interests about this information, e.g. in case of conflicts of interest, post incompatibility, bigamy. The right to health, and the right to research are often constitutional rights as well and as such they are part of the public interests, which may interfere with and limit the right to privacy in a reasonable and a proportional way. Health information are sensitive data, i.e., data which must be protected for privacy and they can only be handled according to the the regulations of records. That is, the right to health may require detailed analyis of health data while the right to privacy explicitly forbids the access of data out of medical condition related health care services.

Legislative solution for this problem is data anonymization. What does data

anonymization mean? In United States, the HIPAA (Health Insurance Portability and Accountability Act), in the European Community the EC/95/46 directive regulates data publishing, hence data anonymization policies. According to EC/95/46 directive, data is anonymized if data subject is no longer identifiable and retained in a form in which identification of the data subject is no longer possible[1]. HIPAA's approach is slightly different: anonymized data ,,does not identify an individual and if the covered entity has no reasonable basis to believe it can be used to identify an individual"[2]. The Hungarian jurisdiction defines anonymization as ,,a technical procedure that ensures connections between data and data subjects are no longer possible"[3]. Definition 1 extends legal definitions by combining their different approaches to the problem. The definition means informally that any Φ transformation is an anonymization function for a given database Δ, if the probability of finding a Ψ inverse transformation, which may use background knowledge not present in the original database, is close to zero.

Definition 1 (Data Anonymization) *Let $\Phi : \mathcal{DB} \to \mathcal{DB}$ be a database transformation function, and $\Delta \in \mathcal{DB}$ be a database. We say the transformed database $\Phi(\Delta)$ is anonymized if*

$$\forall\Psi\forall\Delta_1 \ldots \forall\Delta_N \quad \Pr(\Psi(\Phi(\Delta), \Delta_1, \ldots, \Delta_N) \not\subseteq \Delta) \approx 0, \tag{1}$$

where $\Psi : \mathcal{DB}^{N+1} \to \mathcal{DB}$ ($N = 0, \ldots, \infty$) is an arbitrary function, and Δ_i, $i = 1, \ldots, N$ are database representations of all available background knowledge. If $\Phi(\Delta)$ is anonymized for any $\Delta \in \mathcal{DB}$ then we say Φ is a data anonymization function.

Legal regulations explicitly state what is allowed and what is forbidden, but with the exception of HIPAA's Safe Harbour method they do not define how to achieve anonymity. Safe Harbour is about to pseudonymize or completely remove the following values in databases: name; date different from years including death, birth, discharge dates, etc., ages above 89 years; fax numbers; social security numbers; medical record numbers; health plan beneficiary numbers; account numbers; certificate/license numbers; vehicle identifiers and serial numbers, including license plate numbers; device identifiers and serial numbers; URLs; IP addresses; biometric identifiers; full-face or comparable photos or images; and any other unique identifying number, codes. In

[1] Article 26 of EC/95/46
[2] Section 164.514(a) of HIPAA
[3] Act of 2003/III. 1.§.(2) and Act of 1995/CXIX. 2.§.(1)

other words, Safe Harbour requires to eliminate direct and indirect identifiers, and has no advice on unintentional or partial, i.e. non-unique identifiers. In Section 5, we determine the level of anonymity Safe Harbour provides, and show its relation to our definition.

4 Approaches to data anonymization

Section 3 points out that data anonymization is a challenge for research, business and even for regulatory activities. Anonymization is a congitive process, practitioners must understand what could lead to the identification of an individual besides the obvious; direct data access either phyisically or logically, carelessness, etc.

Each person has some natural identifiers, i.e. data which characterizes an individual; name, social security number, passport number, cellular phone number, vehicle plate number, biometric identifiers etc. Some of them may not identify a person uniquely, but in proper contexts they shall be assumed to be unique identifiers. In this paper, we use the term direct identifiers for these kinds of data. There are a set of natural identifiers called indirect identifiers which together provide a unique identification, e.g. birth date, mother's name, address. One must notice that personal data and data which enable identification of a human being are not necessarily different things. For example, thoughts, forms of expression, activities, friends, medical case history, etc., may also identify people as well; we call them unintentional identifiers.

In practice, direct and indirect identifiers are replaced with one-way hash functions, i.e., functions that cannot allow original data to be restored since they have no inverse. Such a non-reversible value replacement of direct identifiers is called de-identification method. De-identification not necessarily assumes the complete removal of all but direct identifiers. If a de-identification method maps every identical direct/indirect identifier into an identical (but non-reversible) value then it is called pseudo-anonymization or pseudonymization for short. Later on this paper, we use the term re-identification for a procedure or method which processes one or more datasets to determine the identify of data subjects.

To protect privacy one must understand the potential threats, i.e. possible re-identification strategies:

- Direct re-identification. Data themselves without any further action reveal data subject identity.
- Re-identification through linking. Sometimes, data set is believed to be

de-identified while using publicly available or legally accessible databases enables re-identification of data subjects. For example, Netflix prize award data set contained pseudonymized user ids and movie ratings. Netflix ratings were easily correlated to IMDB ratings [25] where user ids are often personal names; re-identification was made possible through linking the preferences.

- Publishing anonymization algorithms, or settings for predictive algorithms. Publishing always makes ways to de-construct or to invert applied functions using direct re-mapping, guessing, etc. If we are looking for an information on an individual, one may deduce the future medical condition using medical predictive function based on their observed symptoms.
- Re-identification through extremities. Outlier values, rare or very unusual behavior are specific by definition to a very limited number of individuals, which may lead to re-identification. For example, if inhabitants of a small town suffer from the same disease, it is easy to infer the medical condition of an individual from that town.
- Background knowledge based re-identification. Sometimes, not structured, not stored, or single fact or knowledge known or accessed by a limit number of individuals is applied to retrieve someone's identity, e.g. custom habits of neighbors when and how they leave, or activities and photos published on a social network portal. Background knowledge is one of the most probable attacking strategy in our social network era.
- Re-identification through event sequencing. Frequencies or the ordering of data items also may uniquely identify certain individuals. A company based on a sick-leave registry may easily reveal employee medical condition if published health care database contains only dates and medical conditions.
- Information misuse. We are talking about information misuse whenever published database makes alternative, possibly harmful use possible. See, for example, the problem of sharing customer transactions in Section 2.

Anonymization techniques against these attacks get more complex in the order above. A question may arise: is it possible to eliminate all threats by a data anonymization algorithm while data still have enough value for data analysis? The following strategies have been applied for anonymization in the literature [30, 1, 31]:

1. Access limitations.

   (a) Limitation of data access. The most common procedure to limit the number of queries and to be run over a controlled environment through proper authentication and information hiding [24, 23].

   (b) Ciphering algorithms. Change of data values in a way that makes impossible to retrieve original values.

2. Obfuscation. These algorithms cut or aggregate parts of the database in order to avoid re-identification.

   (a) Dynamic sampling. Limited number of data elements are published, which meet the functional anonymization criteria.

   (b) Aggregation oriented anonymization. Enforcing data aggregations and micro-aggregation (aggregation over a subset of data elements) to achieve functional anonymization [29].

3. Functional anonymization. It aims at reducing the confidence about a piece of information related to a specific individual. It is widely discussed in the literature, there are several approaches to achieve this goal, e.g. by adding random noise [3, 2], using random data permutation [11], or by redundancy oriented data perturbations [29, 22, 14].

The first strategy is a very protective but not data publishing friendly solution. The second one is efficient for data publishing, however, they reduce dramatically data quality, which implies very limited use for data mining. Functional anonymization is the most discussed solution in the literature and it reveals the depth of data anonymization problems, as well. We make a brief overview on these techniques in Section 5.

5 Functional data anonynimization techniques

Let us consider Table 1 as a database to illustrate anonymization and related problems. The relation itself currently contains one sensitive information, the list of the salaries. Additionally, it can be directly connected with individuals as the ID attribute is present. Publishing such database could be strongly resisted with respect to data protection. Although this database is not

The easiest de-identification is omitting ID column; the result can be seen in Table 2. Note that, selecting any two of the *Date of birth*, *Sex* and *Postal code* attributes can identify the set of individuals in that relation. In general, these attribute pairs are not enough for unique identification, however,

ID	Date of birth	Sex	Postal code	Salary
Annie	21-01-76	Female	1107	40 000
Bill	24-03-76	Male	1107	45 000
Cecile	27-02-76	Female	1117	50 000
Dennis	21-01-76	Female	1117	55 000
Elise	24-03-76	Male	1127	60 000
Fred	27-02-76	Male	1127	65 000

Table 1: Contents of the `basictable` relation

in different countries they can identify a very large percent of the whole population [29, 18, 10], i.e. they quasi identify people. In hungarian a health care database, an individual who lives in a town with less than 50000 inhabitants can be identified with this triple with 94.2% probability. With more than 50000 inhabitants this value falls to 40.4%, which is still an unacceptably high value.

Definition 2 ([29] Quasi-identifier) *Given a set of individuals* I *and relation* $\mathtt{r}(\mathtt{R})$ *on the* $\mathtt{R}(\mathtt{A}_1, \ldots, \mathtt{A}_n)$ *schema, and let* $\mathtt{f}_c : \mathtt{t}[\mathtt{Q}_r] \to \mathtt{r}(\mathtt{R})$ *and let* $\mathtt{f}_d : \mathtt{r}(\mathtt{R}) \to \mathrm{I}'$, *where* $\mathrm{I}' \subseteq \mathrm{I}$, $\mathtt{t} \in \mathtt{r}(\mathtt{R})$ *and* $\mathtt{Q}_r \subseteq \{\mathtt{A}_1, \ldots, \mathtt{A}_n\}$. $\mathtt{Q}_r$ *is a quasi-identifier in relation* $\mathtt{r}(\mathtt{R})$, *if* $\exists \mathtt{p}_i \in \mathrm{I}$ *such that* $\exists \mathtt{t}_i \in \mathtt{r}[\mathtt{R}]$ *for which* $\mathtt{f}_d(\mathtt{f}_c(\mathtt{t}_i)) = \mathtt{p}_i$.

Sweeney's definition means informally that any tuple $\mathtt{t}[\mathtt{Q}_r]$ of a relation $\mathtt{r}$ is a quasi-identifier in that relation, if the subset of attributes $\mathtt{Q}_r$ is unique for some individuals $\mathtt{p}_i$.

Date of birth	Sex	Postal code	Salary
21-01-76	Female	1107	40 000
24-03-76	Male	1107	45 000
27-02-76	Female	1117	50 000
21-01-76	Female	1117	55 000
24-03-76	Male	1127	60 000
27-02-76	Male	1127	65 000

Table 2: Contents of the `de-identified` relation

By using background knowledge on quasi-identifiers and by having information about individuals from public sources, researchers can join records

on quasi-identifiers to the sensitive data items. A solution for this problem is to make data values ambiguous. Either one can delete some of the quasi-identifiers and/or sensitive data values as shown in Table 3 or one can add noise to data values as shown in Table 4.

Date of birth	**Sex**	**Postal code**	**Salary**
*	*	1107	40 000
24-03-76	Male	1107	45 000
27-02-76	Female	1117	50 000
*	*	*	*
24-03-76	Male	1127	60 000
27-02-76	Male	1127	65 000

Table 3: Contents of the `deleteddata` relation

If one deletes some values then it can be overwritten with any (other) value to inhibit data linking through external sources. This means, that if one finds a possible data re-connection through quasi-identifier values, one cannot be certain that quasi-identifier values or data linking restore any part of the original database. On the other hand, a clear disadvantage of this approach is that the usability for analysis is degrading fast with the number of data perturbations.

Date of birth	**Sex**	**Postal code**	**Salary**
21-01-76	Female	1107	50 000
24-03-76	Male	1107	35 000
27-02-76	Female	1117	50 000
21-01-76	Female	1117	55 000
24-03-76	Male	1127	65 000
27-02-76	Male	1127	60 000

Table 4: Contents of the `noisytable` relation

Another way is to add noise to sensitive data. In this case, the attacker cannot be certain about the real data value in any but all particular record. This solution inhibits exploring sensitive data thus linking external data sources provide no further information. As a special case noise can be added by using

microaggregation on data to lower the possibility of re-identification as shown in Table 5. Nevertheless, noise specially aggregations significantly reduce data quality.

Date of birth	Sex	Postal code	Salary
--76	Female	11**	40 000
--76	Male	11**	45 000
--76	Female	11**	50 000
--76	Female	11**	55 000
--76	Male	11**	60 000
--76	Male	11**	65 000

Table 5: Contents of the `aggregated` relation

The concept of k-anonymity limits the applicability of attack using external relationships. Instead of identifying individuals, for any key or keylike attribute there must be at least k with the same quasi-identifier in the database. This is usually achieved by generalizations, for instance deleting some numbers from an IP address or a postal code.

Definition 3 ([29] k-anonymity) *Given a relation* $r(R)$ *over the schema* $R(A_1, \ldots, A_n)$*, and* Q_r *is a quasi-identifier in* $r(R)$*, then* $r(R)$ *is* k*-anonym if for any* Q_r $t[Q_r]$ *value occurs at least* k *times in* $r(R)$*.*

The relation in Table 5 is 3-anonym. In this case, the most concrete knowledge of a record necessarily involves the uncertainity that for that record there are at least 3 other candidates. However, in general it is quite difficult to determine about a relation whether it is k-anonym.

The computational complexity has been shown to be at least of the order of $O(2^{|Q_r|})$ [5, 28] independently from whether the model allows deletion or not. If deletion, local rewrite and multidimensional partitioning is allowed, then finding the minimal k-anonym is NP-hard [19, 20]. Generally, it has $O(n^{2k})$ complexity, but an $O(n \log n)$ approximation has also been proposed with certain restrictions, assuming multidimensional clustering [18].

[22] has shown that k-anonymity is not sufficient hence re-identification is also possible through sensitive data linking as well. In other words, not only the entropy of quasi-identifiers, but the entropy of sensitive values should exceed a particular threshold. The database shown in Table 6 is 2-anonym. Note that,

there are no further constraints on sensitive data values, which are identical for different birth dates. As a consequence, one can easily reveal sensitive data in Table 6 database by knowing only birth dates.

Date of birth	**Sex**	**Postal code**	**Salary**
21-01-76	*	11**	40 000
24-03-76	*	11**	45 000
27-02-76	*	11**	50 000
21-01-76	*	11**	40 000
24-03-76	*	11**	45 000
27-02-76	*	11**	50 000

Table 6: Contents of the `diversity` relation

Definition 4 ([22] l-diversity) *Given relation* $r(R)$ *over the schema* $R(A_1, \ldots, A_n)$*, the relation* $r(R)$ *is* l*-diversive, if for any attribute* A_i *to be protected at least* l *different* A_i *values are assigned to any particular* $t[R \setminus \{A_i\}]$ *value. Formally* l*-diversity exists, if*

$$-\sum_{v \in t[A_i]} p(t, v) \log p(t, v) \geq \log l, \tag{2}$$

where

$$p(t, v) = \frac{|t'[R \setminus \{A_i\}] = t[R \setminus \{A_i\}] \wedge t'[A_i] = v|}{|t'[R \setminus \{A_i\}] = t[R \setminus \{A_i\}]|}, \tag{3}$$

where p *is the apriori probability of* v *value.*

It's important to see that l-diversity implies k-anonymity using $k = l$. According to Definiton 4 the 2-anonym relation shown in Table 7 is 2-diversive at the same time. Computational complexity of l-diversity is greater than the computation complexity of k-anonymity as additional attributes have to be handled.

In addition, re-identification threats are still alive on an l-diversive microaggregated Table [21]. Consider the database represented in Table 8, which is 2-diversive. The exact sensitive data cannot be retrieved by linking quasi-identifiers, however, the difference between data values within the same quasi-identifier determined cluster is marginal. [21] therefore recommends to extend k-anonymity criterion with a diversity related constraint. If distance between

Date of birth	Sex	Postal code	Salary
--76	*	1107	40 000
--76	*	1107	45 000
--76	*	1117	50 000
--76	*	1117	40 000
--76	*	1127	45 000
--76	*	1127	50 000

Table 7: Contents of the `diversity2` relation

sensitive attribute values with the same quasi-identifier values and values of the entire relation are very different then the sensitive attribute values can be estimated with statistical probing. For simplicity, let us say that sensitive data values that share the same quasi-identifier values are in the same equivalence class. That is, there can be found several distributions of different equivalence classes.

Date of birth	Sex	Postal code	Salary
21-01-76	*	11**	40 000
24-03-76	*	11**	45 000
27-02-76	*	11**	50 000
21-01-76	*	11**	40 001
24-03-76	*	11**	45 001
27-02-76	*	11**	50 001

Table 8: Contents of the `tclosed` relation

The concept of t-closure investigates whether there is a `t` threshold, which is not exceeded by a distance measure.

Definition 5 ([21] t-closure) *An equivalence class is* t*-closed, if the distance between the distribution of the sensitive data within that class and the distribution of the entire relation within that class does not exceed a* `t` *threshold. The relation is* t*-closed, if any equivalence class contained is* t*-closed.*

Note that the definition does not define the distance function to be used; that is, it can be applied for various data types including textual, categorical, etc.

It is practical to boost uncertainity with permutation of sensitive data instead of modifying them. It has been already mentioned before that aggregation significantly degrades usability, based on Table 7 one can deduce almost nothing from the data. A permutation approach is to put sensitive data with the same quasi-identifier into hash buckets, then iteratively replacing the current value with one from the bucket with the highest cardinality [32]. Another approach [33] proposes the ordering of sensitive data and selecting the candidate for permutation from an e-wide interval with at least k cardinality.

Permutation provides the same level of privacy preservation as generalization, however, aggregate values are accurate in this case. For instance, the salary of individuals who work in a certain field, or were born in a certain year is a valid and usable value. On the other hand, permutation changes dramatically sensitive data and thus their hidden patterns, which leads to a completely different or alternative result after an analysis. However, if there is only one sensitive data attribute, permutation within the same equivalence class does not pose this problem.

Date of birth	**Sex**	**Postal code**	**Salary**
21-01-76	Female	1107	45 000
24-03-76	Male	1107	40 000
27-02-76	Female	1117	50 000
21-01-76	Female	1117	55 000
24-03-76	Male	1127	65 000
27-02-76	Male	1127	60 000

Table 9: Contents of the `permuted` relation

After reviewing theoretical anonymization techniques, we can see that the Safe Harbour method of HIPAA does provide k-anonymity according to its pseudonymization procedure as it eliminates all possible quasi-identifiers, however it does not modify data records themselves hence it can neither provide l-diversity, nor give protection against t-closure based probing. This means that additional measures have to be completed even after the Safe Harbour method to achive the level of anonymity required by the legal environment.

6 Metrics for privacy and data utility

Ethical data mining aims to meet legal requirements to meet privacy and eliminate stereotype conclusions. At the same time, any change of data decreases the efficiency of data mining. Is privacy measureable?

Let $\mathrm{PP}_{\mathfrak{i}} : \mathcal{DB} \to [0,1] \in \mathbb{R}$ be a membership function that maps to every database a measure which proportional to privacy preservation, where 1 stands for complete privacy protection, and 0 for no privacy. Assume that $\mathrm{PP}_{\mathfrak{i}}(\emptyset) = 1$.

Let $\mathrm{RV}_{\mathfrak{i}} : \mathcal{DB} \times \mathcal{A} \to [0,1] \in \mathbb{R}$ denote the relative usefulness (fitness, accuracy, etc.) measure of applying data mining model on a database. The value 0 indicates that the data cannot be used for data mining purposes, and 1 stands for model application is the best one can achieve on the database. Assume that $\mathrm{RV}_{\mathfrak{i}}(\emptyset) = 1$. Let κ, ν denote the acceptable threshold for privacy preservation measure, and relative usefulness, respectively. Anonymization as an optimization problem for a database $\mathfrak{d} \in \mathcal{DB}$, and for a data mining model $\mathfrak{a}$ can be formulated as follows:

$$\exists ? \Phi : \quad \kappa \leq \mathrm{PP}_{\mathfrak{i}}(\Phi(\mathfrak{d})) \ \wedge \ \nu(\mathfrak{a}) \leq \mathrm{RV}_{\mathfrak{i}}(\Phi(\mathfrak{d}), \mathfrak{a}), \tag{4}$$

where $\Phi : \mathcal{DB} \to \mathcal{DB}$ is data anonymization method. We call an anonymization method Φ optimal for a database $\mathfrak{d}$ and a data mining model $\mathfrak{a}$ denoted by $\hat{\Phi}_{\mathfrak{d},\mathfrak{a}}$, if and only if

$$1 = \mathrm{PP}_{\mathfrak{i}}(\hat{\Phi}_{\mathfrak{d},\mathfrak{a}}) \ \wedge \ 1 = \mathrm{RV}_{\mathfrak{i}}(\hat{\Phi}_{\mathfrak{d},\mathfrak{a}}(\mathfrak{d}), \mathfrak{a}), \tag{5}$$

For research studies the following questions arise:

- Is there an optimal data anonymization for a given database and a given data mining model?
- Is there an optimal data anonymization for any given database and data mining model tuple?
- Is there an optimal data anonymization for a given database independently for all data mining model?
- Is there an optimal data anonymization for all database and all data mining model?

While research focuses on the last two questions, the former two may suffice for industry applications.

6.1 Privacy measurements

Note that, this paper does not introduce how to calculate PP or RV. In the literature, there are several measures depending on the different aspects of the problem. The simplest indicator is the ratio of the number of the identifiable individuals and the number of the total individuals in the database [27].

In case of data perturbations, privacy can be described with an $H(A|B)$ conditional entropy variable, where A is the remaining protection after B has been published [2]. Hence, the probability of correct data is leaking out is

$$\Pr(A|B) = 1 - \frac{2^{H(A|B)}}{2^{H(A)}} = 1 - 2^{-I(A;B)}, \tag{6}$$

where $I(A;B) = H(A) - H(A|B)$ is the mutual information between variables A and B. Similarly, [16, 8] adapt Shannon entropy to describe privacy:

$$\Pr(A|B) = 2^{-\int f_{A,B}(a,b)\log_2 f_{A|B=b}(a)\,da\,db}. \tag{7}$$

The two metrics above are essentially indentical. [26] simplifies these equations so that privacy is defined with variance:

$$\Pr(A|B) = \frac{\mathrm{Var}(A-B)}{\mathrm{Var}(A)}. \tag{8}$$

The size of potential information leakage can be bounded using matrices. Let $[d_{ij}]$ be a Boolean matrix representing an initial database containing the apriori probabilities $p_0^{ij} = \Pr[d_{ij} = 1]$. Once an adversary asks Q queries to the anonymized database as above, and all other values of the database are provided, we can define the posterior probability p_Q^{ij} of d_{ij} taking the value 1. The change in belief is $\Delta = |c(p_Q^{ij}) - c(p_0^{ij})|$, where $c(x) = \log(x/(1-x))$ is a monotonically increasing function.

This model can be generalized by approximating $d_{ij} = 1$ with the Boole-function $f(d_{i1}, \ldots, d_{ik}) = 1$ with k arguments. In this case, the for any query Q and all function f Δ shall be minimal. It has been shown that $O(\sqrt{Q(n)}/\delta)$ changes of data is sufficient for protection [1, 9], where δ is the maximum change allowed, and $Q(n)$ is the number of queries executed on the database of size n.

The efficiency of privacy for a data set on i individuals with p_i public, s_i sensitive, u_i unpublished data that can potentially be used for identifying s_i can be interpreted as follows. Let C be a classifier on p_i and u_i data items, and let C_1 be a classifier built on the published data $t_i =< p_i, s_i >$ with a_1

accuracy. Data protection damaged if for any C_2 classifier with accuracy a_2 that has access to C with regard to the t_i data set $a_1 < a_2$ [15]. The same statement is defined in [17] by means of the distance function δ over probability distributions such that

$$\delta(\Pr_t(s_i = a), \Pr_t(s_i = a | \mathbb{A}(\mathcal{DB}))) < \kappa, \tag{9}$$

where $\kappa \in \mathbb{R}$ and $\Pr_t(s_i)$ is the apriori probability that the sensitive data item s_i has a specific value. If this statement holds, then the database $\mathbb{A}(\mathcal{DB})$ is anonym for the data item s_i.

6.2 Indicators of usability and loss of information

From the point of usability, three relevant aspects of data mining has to be considered: accuracy, completeness and consistency [8]. Accuracy is a measure of difference between the original and anonymized data items. Completeness is the amount of data omitted in the process. Consistency measures the maintainability of inner relationships.

Accuracy can be defined as the difference between the real and the modified information by calculating the loss from the frequency of relative errors, formally:

$$\Delta(r, r') = \frac{\sum_{i=1}^{n} |f_r(i) - f_{r'}(i)|}{\sum_{i=1}^{n} f_r(i)}, \tag{10}$$

where i is a data item and $f_r(i)$ is the frequency of occurence of that data item in the r relation. This formula is sufficiently general to cover not only data modification, but data omission and the generation of new records.

When the anonymization process uses microaggregations, then the loss can be modelled [28] alternatively with

$$\Delta(r, r') = \frac{\sum_{A \in R} \sum_{t \in r} \frac{h}{|DOM(h_A)|}}{|r| \cdot |A|}, \tag{11}$$

where relation r fits the schema $R(A_1, \ldots, A_n)$ and A is an attribute in that schema, $|A| = n$ is the number of the attributes in R, t is a record in r, $|r|$ is the size of the relation, and h means the measure of the microaggregation in the domain $DOM(h_A)$, where $DOM(h_A)$ is the domain of a possible hierarchy levels of an attribute A. Note that we can not assume the omission of data

items in this case, the size of protected and anonynimized databases are essentially the same. Also note that this formula does not calculate the concrete measure of loss, it is only proportional to the possible microaggregation levels, and additionally determining the size of the hierarchy is very promiscuous.

This concept can be generalized by taking interval frequencies used in the tranformations into account:

$$\Delta(r, r') = \frac{\sum_{A \in R} \sum_{t \in r} \frac{f(t[A])-1}{g(A)-1}}{|r| \cdot |A|}, \tag{12}$$

where function f returns the number of different $t[A]$ values from relation r that can be mapped to the aggregated $t'[A]$ value from relation r', and g gives the size of the codomain of A attribute. Note that, neither equations above considers that changing two attributes properly the original value can be restored, and whether there is a loss of information after the aggregation.

An alternative definition for accuracy is based on distribution functions [2]. Given two functions, f and g, where the domain of f is the original database and the domain of g is the anonymized database, and the codomains of them are the same. In this case the loss of information can be described with the measure of mutual information:

$$I(f, g) = \frac{1}{2} E\left[\int |f(t) - g(t)| dx\right]. \tag{13}$$

	$H(A\|B)$	$Pr(A\|B)$	$\Delta(r, r')$
r' =Table 2	0.232	80.4%	0
r' =Table 3	0.289	79.7%	0.00133
r' =Table 4	0.309	82.0%	0.00102
r' =Table 5	0.500	76.4%	0.00412
r' =Table 6	1.057	76.9%	0.00566
r' =Table 7	1.057	76.9%	0.00560
r' =Table 8	0.528	76.0%	0.00257
r' =Table 9	0.232	80.4%	0

Table 10: Data protection and information loss numerically using A = salary, $B = \{\text{date of birth}, \text{sex}, \text{postal code}\}$, and r = Table 2

Table 10 summarizes the different metrics for sample databases presented in this paper. It is easy to see that some of the quality metrics do not take into

account quality changes using non-aggregating form of data replacements. For example, permutation method seems to keep data quality which is only true for calculating aggregating on equivalence classes. Moreover, it cannot capture small semantic distance between different values, that is why t-closure seems to have lower privacy preservation capability than k-anonymity. A comparable, valid measurement for privacy or for data utility is still missing from the literature.

7 Conclusions

This paper presents in very brief the depth, importance, and issues of data anonymization. Data anonymization is one of the most imminent problems which directly affects basic rights like the right to privacy, health, or research, the innovation policy of organizations, and last but not least, the future computing environments. Deconstruction and data linking are always possible in our information era, but inhibiting such threats by data perturbation does not help to exploit the values stored deeply inside the databases, specially health care databases.

We demonstrated by examples and equations that maintaining data quality for data mining and preserving privacy has some limitations; there are too many possibilities to reveal private information using public databases or background knowledge. We showed that legislation covers just parts of the problem by stating what but how to do. Therefore, preserving privacy data mining on health databases presents a new challenge for information engineers, hence the possible number of linkable data is dramatically increasing, e.g. using photos, personal data, activities, hobbies published in social networks.

We presented a well-known set of functional anonymization methods, which aims at perturbing data only the least possible. We demonstrated that these methods reduce data quality significantly. In order to measure such an information loss and privacy we briefly outlined the most discussed evaluation metrics. This paper proved that those metrics cannot capture semantical differences, which makes different anonymization methods incomparable. Open questions still remain for the future: do there exist an optimal data anonymization method for all databases and data mining models, how to compare different data anonymization methods, and how to measure the possible information leakage of a published real-life database assuming excessive usage of public databases and background knowledge.

Acknowledgements

Publishing of this work and the research project were funded by the European Union and co-financed by the European Social Fund under the name of ,,MEDICSPHERE – Complex, multipurpose, ICT technology driven medical, economic, and educational use of clinical data" and grant number TÁMOP-4.2.2.A-11/1/KONV-2012-0009.

References

[1] C. C. Aggarwal, P. S. Yu, An introduction to privacy-preserving data mining. in: *Privacy-Preserving Data Mining*, (Eds.: C. C. Aggarwal and P. S. Yu) chapter 1, pp. 1–9. Springer-Verlag, 2008. ⇒40, 49

[2] D. Agrawal, C. C. Aggarwal, On the design and quantification of privacy preserving data mining algorithms. *Proc. 20th ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 247–255. ACM, 2002. ⇒41, 49, 51

[3] R. Agrawal, R. Srikant, Privacy-preserving data mining. *ACM Sigmod Record* **29,** 2 (2000) 439–450. ⇒41

[4] M. Barbaro, T. J. Zeller, S. Hansell, A face is exposed for aol searcher no. 4417749. *The New York Times*, 9 Aug. 1, 2006. ⇒36

[5] R. J. Bayardo, R. Agrawal, Data privacy through optimal k-anonymization. *Proc. 21st International Conference on Data Engineering*, ICDE '05, pp. 217–228, Washington, DC, USA, 2005. IEEE Computer Society. ⇒44

[6] D. Benatar, Indiscretion and other threats to confidentiality. *South African J. Bioethics and Law*, **3,** 2 (2010) 59–62. ⇒36

[7] J. J. Berman, Confidentiality issues for medical data miners. *Artificial Intelligence in Medicine* **26,** 1 (2001) 25–36. ⇒36

[8] E. Bertino, D. Lin, W. Jiang, *Privacy Preserving Data Mining*, chapter A survey of quantification of privacy preserving data mining algorithms, pp. 183–205. Springer, 2008. ⇒49, 50

[9] C. Dwork, K. Nissim, Privacy-preserving datamining on vertically partitioned databases. in: *Advances in Cryptology – CRYPTO 2004* pp. 134–138. Springer, 2004. ⇒49

[10] K. El Emam, D. Buckerdige, A. Neisa, E. Jonker, A. Verma, The re-identification risk of canadians from longitudinal demographics. *BMC Medical Informatics and Decision Making* **11,** 46 (2011) 1–12. ⇒36, 42

[11] A. V. Evfimievski, Randomization in privacy preserving data mining. *ACM SIGKDD Explorations Newsletter* **4,** 2 (2002) 43–48. ⇒41

[12] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, From data mining to knowledge discovery in databases. *AI Magazine* **17,** 3 (1996)37–54. ⇒35

[13] R. Gellman, The story of the banker, state commission, health records, and the called loans: An urban legend?, `http://bobgellman.com/rg-docs/rg-bankerstory.pdf` 2011. ⇒35

[14] A. Gionis, A. Mazza, T. Tassa, k-anonymization revisited. *Proc. IEEE 24th Int. Conf. Data Engineering ICDE 2008*, pp. 744–753, 2008. ⇒41

[15] M. Kantarcıoğlu, J. Jin, C. Clifton, When do data mining results violate privacy? *Proc. 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA*, pp. 599–604, 2004. ⇒50

[16] H. Kargupta, S. Datta, Q. Wang, K. Sivakumar, On the privacy preserving properties of random data perturbation techniques. *Third IEEE International Conference on Data Mining, 2003. ICDM 2003*, pp. 99–106. IEEE, 2003. ⇒49

[17] D. Kifer, Attacks on privacy and deFinetti's theorem. *Proc. 2009 ACM SIGMOD International Conference on Management of Data*, pp. 127–138. ACM, 2009. ⇒50

[18] M. R. Koot, G. van't Noordende, C. de Laat, A study on the re-identifiability of dutch citizens. in: *Workshop on 3rd Hot Topics in Privacy Enhancing Technologies, HotPETs 2010*, 2010. ⇒36, 42, 44

[19] K. LeFevre, D. J. DeWitt, R. Ramakrishnan, Incognito: Efficient full-domain k-anonymity. *Proc. 2005 ACM SIGMOD International Conference on Management of Data*, pp. 49–60. ACM, 2005. ⇒44

[20] K. LeFevre, D. J. DeWitt, R. Ramakrishnan, Mondrian multidimensional k-anonymity. *Proc. 22nd International Conference on Data Engineering, 2006. ICDE'06.*, pp. 25–25. IEEE, 2006. ⇒44

[21] J.-L. Lin, J. Y.-C. Liu, Privacy preserving itemset mining through fake transactions. *SAC*, pp. 375–379, 2007. ⇒45, 46

[22] A. Machanavajjhala, D. Kifer, J. Gehrke, M. Venkitasubramaniam, L-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data* **1,** 1(2007) 3–3. ⇒41, 44, 45

[23] G. Miklau, D. Suciu, A formal analysis of information disclosure in data exchange. *Proc. 2004 ACM SIGMOD International Conference on Management of Data*, SIGMOD '04, pp. 575–586, New York, NY, USA, 2004. ACM. ⇒41

[24] M. Miller, J. Seberry, Relative compromise of statistical databases. *Austral. Computer J.* **21,** 2 (1989) 56–61. ⇒41

[25] A. Narayanan, V. Shmatikov, Robust de-anonymization of large sparse datasets. *Proc. 2008 IEEE Symposium on Security and Privacy*, SP '08, pp. 111–125, Washington, DC, USA, 2008. IEEE Computer Society. ⇒36, 40

[26] S. R. M. Oliveira, O. R. Zaïane, A privacy-preserving clustering approach toward secure and effective data analysis for business collaboration. *Computers & Security* **26,** 1 (2007) 81–93. ⇒49

[27] L. Sweeney, Datafly: A system for providing anonymity in medical data. *Proc IFIP TC11 WG11.3 Eleventh International Conference on Database Securty XI: Status and Prospects*, pp. 356–381, 1997. ⇒34, 35, 36, 49

[28] L. Sweeney, Achieving k-anonymity privacy protection using generalization and suppression. *Intern. J. Uncertainty, Fuzziness and Knowledge-Based Systems* **10,** 5 (2002) 571–588. ⇒44, 50

[29] L. Sweeney, k-anonymity: A model for protecting privacy. *Intern. J. Uncertainty, Fuzziness and Knowledge-Based Systems* **10,** 5 (2002) 557–570. ⇒41, 42, 44

[30] J. Vaidya, Y. Zhu, C. W. Clifton, Privacy preserving data mining. *Advances in Information Security* **19** (2006) 1–121. ⇒37, 40

[31] K. Wahlstrom, J. F. Roddick, R. Sarre, V. Estivill-Castro, D. de Vries, *Encyclopedia of Data Warehousing and Mining*, volume 2, chapter Legal and technical issues of privacy preservation in data mining, pp. 1158–1163. IGI Publishing, 2nd edition, 2008. ⇒40

[32] X. Xiao, Y. Tao, Anatomy: Simple and effective privacy preservation. *Proc. 32nd International Conference on Very Large Data Bases (VLDB)*, pp. 139–150. VLDB Endowment, 2006. ⇒47

[33] Q. Zhang, N. Koudas, D. Srivastava, T. Yu, Aggregate query answering on anonymized tables. *IEEE 23rd International Conference on Data Engineering, 2007. ICDE 2007*, pp. 116–125, Istanbul, Turkey, 2007. IEEE Xplore. ⇒47

Received: October 9, 2013• Revised: March 4, 2014

Acta Univ. Sapientiae, Informatica, **6,** 1 (2014) 56–70

Bluffing computer? Computer strategies to the Perudo game

Norbert BOROS
Széchenyi University, Győr
email: bnorby@sze.hu

Gábor KALLÓS
Széchenyi University, Győr
email: kallos@sze.hu

Abstract. In this paper we present the elaborated computer strategies which we developed to a less-known dice game, Perudo. Moreover, the related analysis is presented, with the results of our machine players against human players. The main contribution of our paper is a strategy which involves bluffing. An important goal of this work is the remembrance to John v. Neumann; with our means we pay a tribute to the activity of the "founding father"; on the occasion of his 110^{th} birthday.

1 Computer and human thinking

At the time of designing the first electronic calculators it came up as a straightforward possibility to try to copy/build in some—that time already known—modus operandi of the human brain (John v. Neumann et al., about 1946). To make the process operational it was necessary to prepare a simplified model of the brain and to analyse several very important fields like data storage (capacity), data transfer, arithmetic precision, and error-free operation (in computers and in brain); in wider meaning similarities and differences in human and computer thinking, respectively.

John v. Neumann turned to this field from the 1940s with special interest. He presented in this topic several memorable talks and published valuable

Computing Classification System 1998: I.2.1
Mathematics Subject Classification 2010: 68T20
Key words and phrases: games, machine strategies, bluff, Perudo
DOI:10.2478/ausi-2014-0018

papers. Unfortunately—because of his serious illness and early death—he was not able to finish one of his most interesting books (The Computer and the Brain [8]), but this work is even so extremely remarkable and can serve as a source of interesting ideas even for today's researchers.

The observations can be summarized in brief as follows. The human brain—contrary to computers—does not use the "language of mathematics", its operation is not digital, but essentially statistical; its important characteristic is the lower arithmetic accuracy, and extremely high logical reliability ([8], 1959).

The examination of similarities and differences in human and computer thinking is even today an actively investigated multidisciplinary top-field. In this "front" an interested researcher has obvious opportunities. In this paper within the main trend we "venture" to a smaller branch—which is proved to be very interesting, however—we present the strategies (and the connected analysis) which were elaborated to a less-known dice game, Perudo (first related publication by the first author: [2]).

2 Games and strategies

Playfulness has been always a characteristic of human nature. Challenge, victory, competition, tempting the fortune—these are factors which affect every human. We can take it for granted that for achievement of victory an interested, thinking player can take some theoretical considerations—over and above the given game—, and of course, it was the situation even for centuries/thousand years.

In this work we have not intended to describe the theory of games in detail; we always present the necessary background to the chosen game(s) (and to the problems which have actually arisen). For the readers interested in game theory deeply a rich literature is available (e.g.: [5, 7]).

The world of the game-area—separated partly from the reality of the real world—can be generally characterized as follows. A game "consists of" players, rules and results [6]. Usually we can assume that keeping the rules is obligatory for every actor (i.e. deception is not allowed); and also that players know the rules. The positive result (victory, winning the round or other favourable outcome) involves for the players winning (money, token, other real or fictitious objects); this is called simply *payoff*. So, the goal of the game is to reach the most favourable payoff.

By evaluating the results the players—according to their knowledge/skills—

set up a value order among the possible outcomes foreseen (step by step, on longer run or even globally) and make decisions in this way. When a player makes such a plan which prescribes the answer in every decision situation, then we talk about *strategy*.

Formally, a game can be identified by a vector (or list) $G = (S_1, S_2, \ldots, S_n; \varphi_1, \varphi_2, \ldots, \varphi_n)$, where n is the number of players, $S_1, S_2, \ldots, S_n$ denote the strategy-sets of the players, and $\varphi_1, \varphi_2, \ldots, \varphi_n$ are the payoff functions—here the list above is called as the normal form representation [5]. The domain of the payoff function is $S_1 \times S_2 \times \cdots \times S_n$, its range is a certain subset of real numbers, so $\varphi_i(x_1, x_2, \ldots, x_n)$ can be computed assuming that all players have a fixed strategy. (It should be noted that a player usually does not know the strategy of the others.)

The best possible strategy is the *winning strategy*. If in a particular game such a strategy exists, and the player who moves first knows it and adheres to it, then—regardless of the moves of the other players—he/she will always win.[1] The other extremity is given in games where we have no information about the will of the other players; in this case obviously we cannot perform any analysis. An example of such a game is rock-paper-scissors; in this case we can trust essentially only in luck. Card games represent an interesting middle course, where information is usually not complete, but we have possibility for certain deductions.

While a lot of games can be described as deterministic processes, and can be evaluated so, for economic processes this property usually does not hold. The optimal solution here depends mostly on a series of external factors which cannot be necessarily foreseen and we have to apply here certain simplifications in order to reach the good resolution [3]. Essentially, this is the main problem of the game theory. The game theory science was established by John v. Neumann, who realized the minimax theorem by analysing the two-player zero-sum games, according to which with a reasonable game style of the players, one actor always gets an average ν winning. The fact that game theory is widely applicable was made clear by his book published with Oskar Morgenstern [6].[2] [3]

[1] We note that in many "complicated" deterministic classic games the winning strategy is usually not known or—based on our knowledge for the time being—we are not able even to decide if such exists [3] (chess, chequers).

[2] An excellent round-up was presented about the activity of John v. Neumann by L. Á. Kóczy in [6].

[3] We note that in connection with the 2008 financial crisis it was heard from many experts that economic game theory is essentially unusable. This is, of course, a strong exaggeration,

3 Classification of games

Analysing well-known games we find that devising a strategy good enough is usually not an easy task. At the beginning of this process we have to answer e.g. the following questions:

- Do we know completely the standing of the other players, their move or choice, and objective circumstances of their decisions?
- Does luck play a role in the given game?
- Is the collaboration of the players allowed?

We know that many times in the deterministic case the extremely complicated game-space causes a big problem; while in games which comprise random elements we cannot reach even theoretically the complete mastery (however, some additional knowledge can help even in this case).

By an extensive analysis of the factors presented above, we can group the games into the following categories (here we used the book [5] as a source):

- Strategy game: fortune plays no role in the game, the decisions of the players can be analysed in a fully deterministic way (example: chess, noughts and crosses);
- Game of chance (gamble): the result of the game is determined by pure accidental factors; these games can be managed with the means of probability theory (example: lottery);
- Mixed game (partly strategy, partly gamble): between the two categories listed above (example: poker);
- Zero-sum game: a given player can increase his/her winnings only to the detriment of the other players (example: nim, noughts and crosses);
- Non zero-sum game: players can use even external sources to increase their winnings (example: roulette);
- Cooperative game: players can cooperate in coalitions to reach a common goal (example: monopoly);
- Non-cooperative game: keeping the game-rules players have no possibility to cooperate (example: chess);
- Game with complete information: players see the standing and the moves of the opponent(s) exactly, and—assuming appropriate game skills and knowledge—they can conclude precisely the reasons of the decisions of the other actor(s) (namely the payoff function of every player is a common knowledge for all players; example: chess);

but it is a fact where lessons should be deducted.

- Game with incomplete information: for one or more player(s) there exists some kind of uncertainty about the payoff function of the other players—standing, strategy (example: poker, rock-paper-scissors);
- Static game: the strategy of the players is fixed, does not change during the game (example: lottery);
- Dynamic game: the momentary strategy of the players is an element of a big strategy-set, the choice is determined by the current game-situation (example: poker, chess).

The classification according to the categories—with possible connections—is presented in Figure 1.

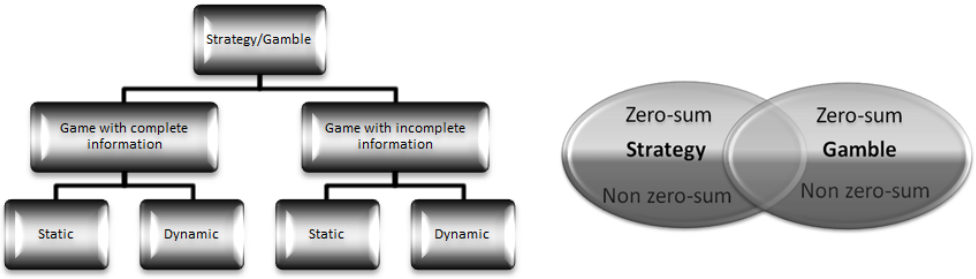


Figure 1: Classification of the games

4 Analysis of the game of poker

In the following—according to our game, Perudo—we will be interested in games located on the borderline of strategy games and gambles which are zero-sum, non-cooperative and dynamic with incomplete information. We will use poker as a kind of standard which is similar to Perudo in several aspects, but is better known and much more analyzed.

John v. Neumann liked to play poker himself, and he was deeply interested in finding the way of reaching victory with a good chance—regardless of dealing luck which cannot be influenced assuming a fair game style [6]. Achieving this, bluff can help most of all and the understanding of the connected human behaviour. He tried to describe the bluff methods with the means of mathematics and later he generalized this analysis into several directions (game theory, machine and human thinking, see the introduction section).

Poker is a game which is even nowadays very actively analysed—because its goal is essentially "making money". It is thought-provoking that some kinds

of people are in this game more successful and the winners of big competitions come from a well-limited group. About the theory of poker David Sklansky published a "bible-like" detailed writing [9] in which—among others—he presents the mathematical background and the "fundamental theorem of poker" (not a real theorem, principle; it refers to strategy based gambles, too), the good game strategies, and the possibilities of recognition and evasion of bluff. The book was written mostly for professional players, it is not an easy reading; but shows the enquiring reader ins and outs of the game with a method of a high standard, trying to reach completeness.

In order to algorithmize the game of poker various approaches were applied; adaptive learning-, probability-, deterministic rule-based- and balanced equilibrium methods. The theory of the machine techniques is presented in detail by Darse Billings in his interesting PhD thesis [1]. However, each method has its shortcomings, too. In summary, we can say that for the time being there is no chance for the best machine players to beat the best human players (comparing with chess it can be a bit surprising).

For the implementation of the machine strategies for Perudo we took into account mostly probability rules. It is important to emphasize that although the algorithmization of Perudo (variants) was dealt with by other programmers already (example: online Perudo, http://www.perudo.com), but in these realizations machine strategies *have not been created* (they have made online games, where people can play against each other). So to the best of our knowledge, the results published in this paper are new (stand: July 2013).

5 Presentation of the game of Perudo

The game of Perudo—as we know it today—derives from the pre-Columbian South-America and was very popular among the Inca Indians. According to tradition, emperor Atahualpa had taught Pizarro to play Perudo (during the period of his imprisonment in 1532), and after it the game spread in Europe.[4] Perudo can be characterized as the exciting game of logic, bluff and guess.

The game can be played by two or more players. *Initially*, all of the players have 5 dices (under a cup). The game consists of *rounds*, and at the end of each round some of the players lose one dice. *Finally*, the play is won by the player, who has at least one dice at the end.

[4]We note that Perudo has several versions. Here we highlighted one version of them.

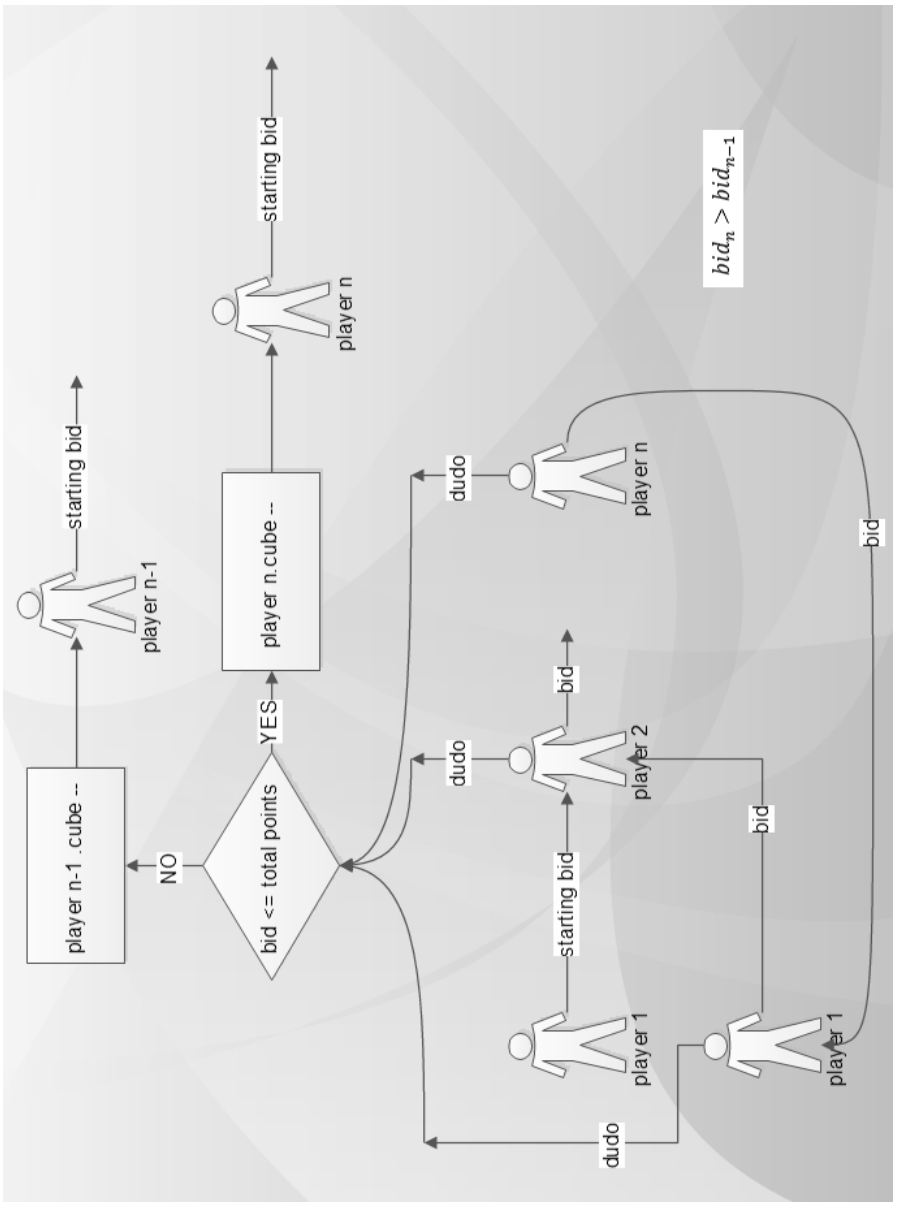


Figure 2: The course of the game of Perudo (without joker bid)

In one round the players—hidden from each other—roll their dices and then begins the continuous bid. Bidding consists of two parts; the number and the value of the dice belonging to it. The player who moves first estimates at least how many dices are on the table with that value. The number of ones is added to the number of the pieces, because they are counted as *joker*. The value of the bid is determined by the following formula:

$$\mathsf{L} = \textit{number of the dice(s)} \cdot 10 + \textit{value of the dice(s)}.$$

The next player has two choices: he/she can overcall the bid (at such time he/she can bluff, too), or says "dudo" (in Spanish: I doubt)—in this case he/she considers the former bid as a not real one. By the call of dudo, players reveal their dices and determine the number of the dices with the given value (including jokers). If the last bid was valid, then the player saying dudo has to throw away one of his/her dices, while if the player with the last bid bluffed then he/she looses one dice (Figure 2).

It is a twist in the game that the player whose turn it is can call a joker bid, too. At such time he/she estimates the number of jokers in the play currently (of course with this he/she can bluff, too). In a joker bid at least half of the number of pieces appearing in the last bid must be called (rounded upwards). The possibilities for the next player are: he/she can say dudo, can increase the number of jokers in the bid, or can return to normal bid (in this case the number of dices in the bid must be doubled and at least one additional dice must be added). Using formulas (for the number of the dices, ND):

$$\textit{Joker bid}_{\mathsf{ND}} \geq \lceil \textit{Normal bid}_{\mathsf{ND}}/2 \rceil, \text{ and } \textit{Normal bid}_{\mathsf{ND}} \geq \textit{Joker bid}_{\mathsf{ND}} \cdot 2 + 1.$$

6 The machine strategies implemented

It was our emphasized goal to create such machine strategies which possibly can cope successfully with human players, too. The realization was accomplished step by step, from the simplest strategies toward the more sophisticated ones. Machine and programming environment: table PC with Intel Core 2 Duo CPU and 4 GB RAM, Windows operating system, Netbeans integrated development environment, JAVA language.

The first programmed method was named as basic strategy. This strategy decides and says bid on the basis of analysing expected values. The strategy has an adjustable parameter, the tolerance, with which we can set the upper bound—exceeding the expected value—acceptable in bids. (In this case joker bid is said.)

Figure 3: The starting screen of the program

```
decision(bid_got)

1 expected = expected_value(bid_dice_value_got)
2 if (bid_dice_number_got < expected)
3       bid()                                   // lines 2–3: normal bid
4 else if (bid_dice_number_got <= expected + tolerance)
5       joker_bid()                             // lines 4–5: joker bid
6 else
7       dudo()
```

Following David Sklansky, the good player plays so that he/she would know the cards/dices of the other player(s)/opponent(s) [9]. In the extended strategy—based on basic strategy—this idea was applied: the code was supplemented with a method which deduces back—from the initial bids—to the dices of the players and stores them. The effect of this deduction resulted that the extended strategy has won—using a 100 thousand-round test—6-times more rounds than the basic strategy (Figure 4).

The reason for this is that basic strategy tried saying dudo roughly for half of the times as the opponent, and almost the half of its trials were unsuccessful; whilst the extended one said many times successful dudo, than unsuccessful.

The *foreseeing strategy* accomplished thirdly, —similarly to the extended one—has a `bidLoad()` method using which it keeps the records of dices of the

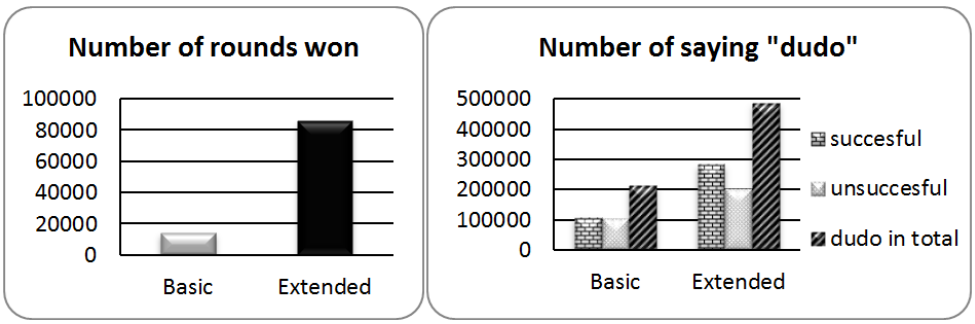


Figure 4: Basic vs. extended strategy

other players. However, by the deduction it takes into account all bids, so, it can calculate more exactly.

```
bidLoad(bid[ ], number_of_bids)
1 player[bid[0].player_id][bid[0].dice_value − 1]
        += number_of_dices(starting_bid)
2 for(i = 1; i < number_of_bids; i++)
3 player[bid[i].player_id][bid[i].dice_value − 1]
        += number_of_dices(bid[i], bid[i − 1])
```

It is true for the foreseeing strategy, too, that—due to the more refined deduction—it plays better than the basic one. Interestingly, the basic strategy against the extended one said only a little bit fewer successful dudos than unsuccessful ones, on the other hand, against the foreseeing strategy the number of saying unsuccessful dudos was definitely more (Figure 5).

Strategies were tested in various situations because we wanted to know whether the result of the game is influenced by such factors as which strategy starts the bid, how many times they can initiate a joker bid, or how the players sit side by side (seating), respectively. The modification of the first two factors had brought no change; however, we got surprising a result by altering the seating.

In the following the result of two complex simulations will be presented. In both examinations five machine players took part, four basic and one foreseeing. In the first case the seating changed randomly, in the second case it was fixed. In both simulations the foreseeing strategy won mostly, but the winning-distribution of the basic machines is interesting.

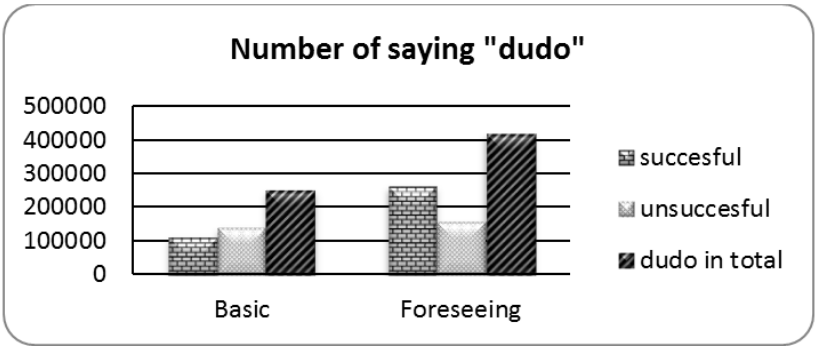


Figure 5: Basic strategy vs. the foreseeing one

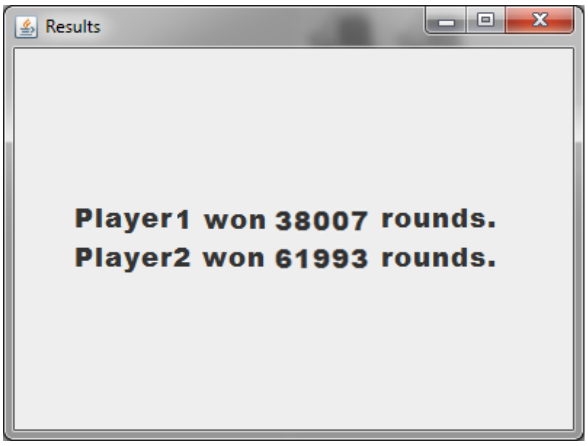


Figure 6: Displaying the result of a machine test (duel of strategies, program window)

With fixed seating—gradually—the basic machine following the foreseeing one won most often, and after it each less and less; the basic strategy right before the foreseeing one the least. The reason for this is that the foreseeing machine tends to say dudo at the right time to the strategy right before it, and the basic machines before the foreseeing say—often—at the wrong time dudo to the foreseeing strategy. So, the basic machine right before the foreseeing one drops out of the competition first. With random seating every basic machine won roughly the same number of rounds—this meets the expectations (Figure 7).

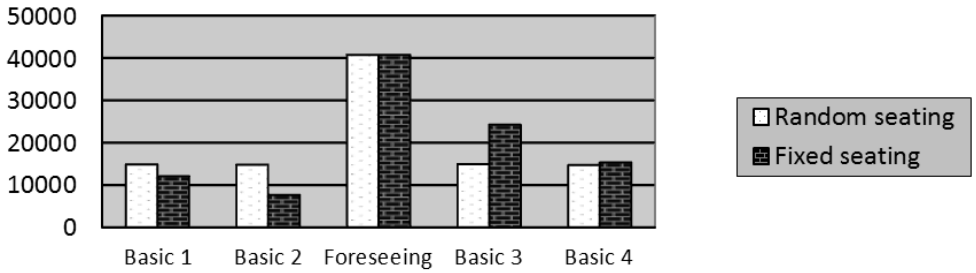


Figure 7: Basic machines vs. the foreseeing one: winnings with fixed and changing seating

After this, the tests with human players were organized. In the course of the trials 3 volunteers played 15 rounds against the machine players. The scenario of the games was similar: at the beginning almost exclusively the machine had won, later this equalized and slowly the human players gained the dominance. The reason of this phenomenon is that human players came to know the machine players. All three participants of the trial found out how the machine gave the bid and exploiting this they bluffed. The machine players played rationally, they believed the bluff and so finally lost.

7 Bluffing strategy

The parties against the human players have brought the clear experience that in certain situations it is impossible to win with rational behaviour; so our machine players—accomplished so far—has had a serious disability. As a "theorem", we can draw up that a good player has to be unrecognizable, because if he/she decides to always use the same strategy, then after a while the others saw through it and defeated him/her.

In the Perudo game bluff is an important factor since without it we would have essentially a simple gamble which would be won by the player who throws larger. Bluff is a form of inscrutableness, the "allowed lie" in the game, the tool of the deception of the opponent.[5] To the success of deception it is indispensably important to be disguised: in the behaviour of the player the "deceitful purpose" should not be visible. In the case of human players bluff should be said with the same poker face as the other bids; besides this a good player has

[5]As it was already mentioned, John v. Neumann came to the mathematical establishment of the game theory from the analysis of the bluff.

to be able to detect the deception of the others. The one, who wants to be really successful in this "genre", is worth getting to know thoroughly the psychological background, too: which are the usual telling/revealing characteristic of the lie [4].

Bluff must be set up by following a strategy, too (when and how we should bluff; concerning the poker game: [9]), winning is not for one game, but in the long run. Greedy strategy is not practical here: if a player always or almost always bluffs then although in the first few rounds he/she will win with better chance, but later when the opponents see through his/her habits, it is almost sure that he/she will lose. Similarly to poker—possibly—it is practical to make an analysis about the playmates: who in what manner reacts to the bluff, how credulous or doubting he/she is.

In machine environment clearly it should be implemented otherwise (at least partly) since—taking into account our current possibilities—considering the behaviour, gestures and facial features bluff should not be found out because a computer cannot recognize these, and such manifestations computer cannot own.

By creating the *bluffing* strategy, our goal was to evolve such an algorithm which can be efficient against both of the credulous and doubting player, too. To make itself less predictable, strategy "builds" from several elements. With bluffing it considers not always its best theoretical bid: it can happen that it bids with a dice from which it has minimum or just sends the value from which it has maximum, but playing out not all of them. Occasionally it can transmit a bid generated randomly which may not exist. In non-existing bids, however, we made sure it should be believable to every opponent (strategy), i.e. based upon probabilities theoretically it could be on the table.

By forming of the bids we have applied case separation, elaborating scenarios for various situations. Such cases are, for example, when the player or the opponent has only one dice; or the adversary has a lot of dices, but have told only a one-valued bid, and so on.

Similarly to the extended and foreseeing strategies, the bluffing strategy analyses the bids of the opponents and tries to deduce back to their dices. However, it takes the deduction not for sure (assuming that the other players can bluff, too). According to the current realization the strategy is efficient only in the case when it has only one opponent, but we are planning to implement the multi-opponent version, too.

With the bluffing strategy we performed three 100 000-round tests, against the basic, extended and foreseeing strategies. The following result was arrived at: the new method was successful against all of the older ones, however, in

varying degrees; it won the most parties against the foreseeing strategy, and the minimum against the basic strategy (Figure 8)!

Figure 8: The bluffing strategy vs. the basic and foreseeing strategies ("credulous" and "doubting" opponents)

The reason for this—maybe surprising—result can be that foreseeing strategy deduces back the better way to the dices of the opponents, that's why it can be cheated the most easily (so, the foreseeing machine can be considered as a "credulous" player). Basic strategy performs no such analysis, therefore it is not gullible (so, it can be counted as a "doubting" player). Analysis of course helps without bluff, too: that's why the bluffing strategy could won in majority against the "doubting" basic strategy.

The bluffing strategy was tested against human players, too. Several rounds were played with three human players—the Perudo skills of whom could be classified at average level. Now we cannot deduce how the human players came to know the machine player. Results are rather scattered, almost random (Table 1); definite superiority is not visible from either slide—so, the *bluffing strategy plays with a similar efficiency as a typical man/woman.*

Player	Winner
1	HHM MHM MMH MHM HMM
2	MMM HMH MMH MHH MHM
3	MHH MHH HHM HMM HHH

Table 1: Bluffing machine strategy (M) vs. human (H) players

We can conclude that predictability was removed from the machine play with this strategy, but still, our bluffing strategy—implemented so far—cannot be classified as a really successful and "clever" player.

We continue the work: our goal is to implement such a machine player that will be able to defeat clearly an average human player, and with the hope of success can enter the fight with real professional human players, too.

Acknowledgements

We thank Gábor Takács, our excellent colleague, who contributed significantly with valuable advices and useful comments to the writing of this paper.

References

[1] D. Billings, *Algorithms and assessment in computer poker*, PhD Thesis, Alberta University, 2006. ⇒61

[2] N. Boros, *Machine strategies in the game of Perudo*, Univ. Thesis (in Hungarian; supervisor: Gábor Takács), Széchenyi University, Győr, 2009. ⇒57

[3] M. Eigen, R. Winkler, *Das Spiel – Naturgesetze steuern den Zufall*, R. Piper & Co. Verlag, München, 1975. ⇒58

[4] A. Furnham, *50 Psychology Ideas You Really Need to Know*, Quercus Publishing, London, 2009. ⇒68

[5] R. Gibbons, *A Primer in Game Theory*, Pearson Acad. Publisher, Harlow, 1992. ⇒57, 58, 59

[6] L. Á. Kóczy, A Neumann-féle játékelmélet, *Közgazdasági Szemle* **53**, 1 (2006) 31–45. ⇒57, 58, 60

[7] S. Molnár, F. Szidarovszky, *Game Theory* (in Hungarian), ComputerBooks, Budapest, 2011. ⇒57

[8] J. von Neumann, *The Computer and the Brain*, Yale Univ. Press, York, PA, 1959. ⇒57

[9] D. Sklansky, *The Theory of Poker*, Creel Printing Co., Las Vegas, NV, 2007. ⇒61, 64, 68

Received: August 28, 2013 • Revised: March 24, 2014

Acta Univ. Sapientiae, Informatica, **6,** 1 (2014) 71–88

A computational model of outguessing in two-player non-cooperative games

Tamás László BALOGH
University of Debrecen
email: tamas.balogh@econ.unideb.hu

János KORMOS
University of Debrecen
email: kormos.janos@inf.unideb.hu

Abstract. Several behavioral game theory models aim at explaining why "smarter" people win more frequently in simultaneous zero-sum games, a phanomenon, which is not explained by the Nash equilibrium concept. We use a computational model and a numerical simulation based on Markov chains to describe player behavior and predict payoffs.

1 Introduction

Since the birth of experimental economics, thousands of experiments have been conducted to observe the behavior of decision makers in different situations (see e.g. [4]).

However, the most famous equilibrium concept—the Nash equilibrium [13]—has proved to be unable to explain the outcome of several game theoretical experiments, predicting that human thinking is more complicated than pure rationality.

Game theory has also proved to be a useful modelling tool for network situations, e.g. telecommunication problems. For a detailed survey in this field, we refer the reader to [15]. An application can be found in [1].

Computing Classification System 1998: G.3, I.6, J.4
Mathematics Subject Classification 2010: 60J20, 62P20, 68U20, 91A10
Key words and phrases: computational game theory, non-cooperative games, outguessing, Markov chains, reasoning, communication networks
DOI:10.2478/ausi-2014-0019

The phanomenon that human behavior is not purely rational in certain interactive network situations led researchers to construct behavioral game theory models. Recently, several models have been built to explain experimental results (e.g. [2, 7]).

A popular model class aiming at explaining how players outguess each other is the group of iterative reasoning models. Iterative reasoning has been applied in many settings to the Rock-Paper-Scissors game or the Beauty contest-type games ([3, 5, 8, 11, 12, 16]).

The concept of iterative reasoning and the corresponding main results are presented in [4, pages 205–236]. A simplified concept for non-cooperative, two-person, simultaneous games can be defined as follows. If Player A plays a certain action, while Player B plays the best response to this action, then we say that Player B outguessed Player A and played according to 1-reasoning. If now Player A outguesses Player B, then Player A plays according to 2-reasoning. Following this rule, the level of reasoning can be any k positive integer, where the concept is defined as k-reasoning.

In this paper we investigate simultaneous, two-person, zero-sum, repeated games that do not have a pure strategy Nash equilibrium and the players' decisions depend only on their actions in the previous round of the game. Here, the stochastic processes of the players' decisions and their expected payoffs can be described by Markov chains.

Our main goal is to point out why "smarter" people win more frequently in some well-known zero-sum games. There are several ways to define smartness. Our definition of smartness is connected to the concept of iterative reasoning and is introduced later on in Section 3.

We focus on modelling players' optimal strategy choices and expected payoffs. These are both stochastic processes given a certain bimatrix game and the level of iterative reasoning according to which players make their decisions.

We constructed a Matlab script that carries out the requested numerical analysis for any simultaneous, two-person bimatrix game. In our paper we present the relating analytical results, describe our concept and recall some numerical results and visualizations. Our Matlab script is also attached for testing and experimental purposes.

The rest of the paper is organized as follows. Section 2 recalls some important results in the field of Markov chains that are related to our topic. Section 3 describes our concept and provides numerical evidence. Section 4 describes the Matlab script. Finally, Section 5 concludes.

2 Markov chains—definitions and some important results

It is necessary to recall some basic results from the field of Markov chains that we use in the upcoming sections. For a more detailed analysis, we refer the reader to [6, 9, 10, 14]. This section is a brief summary of Chapter 4 in [6, pages 119–155], that is related to our concept. The proofs are always omitted.

Definition 1 *Let S be a countable (finite or countably infinite) set. An S-valued random variable X is a function from a sample space ω into S for which $\{X = x\}$ is an event for every $x \in S$.*

Here S need not be a subset of $\mathbb{R}$, so this extends the notion of a discrete random variable (or vector). The concepts of distribution, jointly distributed random variables, and so on, extend in the obvious way. The expectation of X, however, is not meaningful unless $S \subset R$. On the other hand, the conditioning random variables in a conditional expectation may be S-valued, and all of the results about conditional expectation generalize without difficulty.

Definition 2 *A matrix $P = (P(i,j))_{i,j\in S}$ with rows and columns indexed by S is called a one-step transition matrix if $P(i,j) \geq 0$ for all $i,j \in S$ and $\sum_{j\in S} P(i,j) = 1$ for all $i \in S$.*

In particular, the row sums of a one-step transition matrix are equal to 1. We call $P(i,j)$, the entry in row i and column j of the matrix P, a one-step transition probability.

Definition 3 *We say that $\{X_n\} n \geq 0$ is a Markov chain in a countable state space S with one-step transition matrix P if $X_0, X_1, \ldots$ is a sequence of jointly distributed S-valued random variables with the property that*

$$P(X_{n+1} = j|X_0, \ldots, X_n) = P(X_{n+1} = j|X_n) = P(X_n, j) \tag{1}$$

for all $n \geq 0$ and $j \in S$.

We assume that the sequence $X_0, X_1, \ldots$ is indexed by time, and if we regard time n as the present, the first equation in (1), known as the Markov property, says that the conditional distribution of the state of the process one time step into the future, given its present state as well as its past history, depends only on its present state. The second equation in (1) tells us that $P(X_{n+1} = j|X_n = i) = P(i,j)$ does not depend on n. This property is called time homogeneity.

The distribution of X_0 is called the initial distribution and is given by $(i) := P(X_0 = i), i \in S$.

A Markov chain can be described by specifying its state space, its initial distribution, and its one-step transition matrix.

Given a Markov chain $\{X_n\} n \geq 0$ in the state space S with one-step transition matrix P, it can be shown that, for every $m \geq 1, i_0, i_1, \ldots, i_m \in S$, and $n \geq 0$, $P(X_{n+1} = i_1, \ldots, X_{n+m} = i_m | X_n = i_0) = P(i_0, i_1)P(i_1, i_2) \ldots P(i_{m-1}, i_m)$.

Definition 4 *We define the m-step transition matrix* P^m *of the Markov chain by*

$$P^m(i, j) = \sum_{i_1, \ldots, i_{m-1} \in S} \cdots \sum P(i, i_1)P(i_1, i_2)P(i_{m-1}, j) \tag{2}$$

Notice that the superscript m can be interpreted as an exponent, this is, the m-step transition matrix is the mth power of the one-step transition matrix. This is valid both when S is finite and when S is countably infinite. It is easy to check that this allows us to generalize (2), obtaining $P(X_{n+m} = j | X_0, \ldots, X_n) = P(X_{n+m} = j | X_n) = P_m(X_n, j)$ for all $n \geq 0$, $m \geq 1$, and $j \in S$.

Given $i \in S$, let us introduce the notation $P_i(\cdot) = P(\cdot | X_0 = i)$, with the understanding that the initial distribution is such that $P(X_0 = i) > 0$. It can be shown that

$$P_i(X_1 = i_1, \ldots, X_m = i_m) = P(i, i_1)P(i_1, i_2) \cdots P(i_{m-1}, i_m) \tag{3}$$

for all $i_1, \ldots, i_m \in S$.

Given $j \in S$, let us introduce the notation T_j for the first hitting time of state j (or first return time if starting in state j) and N_j for the number of visits to state j (excluding visits at time 0). More precisely, $T_j = \min\{n \geq 1 : X_n = j\}$ and $N_j = \sum_{n=1}^{\infty} 1_{\{X_n = j\}}$, where $\min \emptyset = \infty$. If also $i \in S$, we define $f_{ij} = P_i(T_j < \infty) = P_i(N_j \geq 1)$. This is the probability that the Markov chain, starting in state i, ever visits state j (or ever returns to state i if $j = i$). We can now define transient and recurrent states.

Definition 5 *We define state* j *to be transient if* $f_{jj} < 1$ *and to be recurrent if* $f_{jj} = 1$.

Some important features are pointed out in the next propositions.

Theorem 6 *Letting* $m \to \infty$ *it can be shown that*

$$P_i(N_j = \infty) = \begin{cases} 0 & \text{if } j \text{ is transient} \\ f_{ij} & \text{if } j \text{ is recurrent.} \end{cases} \tag{4}$$

Theorem 7 *For a Markov chain in* S *with one-step transition matrix* P*, state* $j \in S$ *is*

$$\textit{transient if } \sum_{n=1}^{\infty} P_{j,j}^{n} < \infty, \tag{5}$$

$$\textit{recurrent if } \sum_{n=1}^{\infty} P_{j,j}^{n} = \infty. \tag{6}$$

What is more, given that $i, j \in S$ is distinct, if state i is recurrent and $f_{ij} > 0$, then state j is also recurrent and $f_{ji} = 1$.

Let us define irreducible Markov chains.

Definition 8 *A Markov chain in* S *with one-step transition matrix* P *to be irreducible if* $f_{ij} > 0$ *for all* $i, j \in S$.

By Proposition 2, if a Markov chain in S with one-step transition matrix P is irreducible, then either all states in S are transient or all are recurrent. This allows us to refer to an irreducible Markov chain as either transient or recurrent.

Now we turn to the analysis of the asymptotic behavior of Markov chains.

Let π be a probability distribution on S satisfying

$$\pi_j = \sum_{i \in S} \pi_i P(i, j), \quad j \in S. \tag{7}$$

Regarding π as a row vector, this condition is equivalent to $\pi = \pi P$. Iterating, we have

$$\pi = \pi P = \pi P^2 = \ldots = \pi P^n, \ n \geq 1. \tag{8}$$

In particular, if $\{X_n\} n \geq 0$ is a Markov chain in S with one-step transition matrix P and if X_0 has distribution π_0, then X_n has distribution π_n for each $n \geq 1$. For this reason, a distribution π satisfying (7) is called a stationary distribution for the Markov chain.

We need one more definition to state an important result.

Definition 9 *The period* $d(i)$ *of state* $i \in S$ *is defined to be* $d(i) = g.c.d.D(i)$, $D(i) = \{n \in N : P^n(i, i) > 0\}$*, where g.c.d. stands for greatest common divisor.*

We first notice that every state of an irreducible Markov chain has the same period.

Note that if $i, j \in S$ are such that $f_{ij} > 0$ and $f_{ji} > 0$, then $d(i) = d(j)$. This allows us to speak of the period of an irreducible Markov chain.

Definition 10 *If the period is 1, we call the chain aperiodic.*

We can now describe the asymptotic behavior of the n-step transition probabilities of an irreducible aperiodic Markov chain.

Theorem 11 *If an irreducible aperiodic Markov chain in* S *with one-step transition matrix* P *has a stationary distribution* π*, then it is recurrent and*

$$\lim_{n\to\infty} P^n(i,j) = \pi(j) \; i,j \in S. \tag{9}$$

Furthermore, $\pi(i) > 0$ *for all* $i \in S$*.*

It follows from the previous statement that if an irreducible aperiodic Markov chain in S with one-step transition matrix P has no stationary distribution, then

$$\lim_{n\to\infty} P^n(i,j) = 0 \; i,j \in S. \tag{10}$$

Thus, an irreducible aperiodic Markov chain in a finite state space S has a stationary distribution.

Theorem 12 *Let* $\{X_n\}_{n\geq 0}$ *be an irreducible aperiodic recurrent Markov chain in* S *with one-step transition matrix* P*. Then one of the following conclusions holds:*

(a) $E_i[T_i] < \infty$ *for all* $i \in S$*, and* P *has a unique stationary distribution* π *given by*

$$\pi(i) = \frac{1}{E_i[T_i]}, \; i \in S. \tag{11}$$

(b) $E_i[T_i] = \infty$ *for all* $i \in S$*, and* P *has no stationary distribution.*

If (a) holds, then the chain is said to be positive recurrent, and Equation (9) holds. If (b) holds, then the chain is said to be null recurrent, and Equation (10) holds.

In the following section we define our concept and point out its relationship with Markov chains.

3 The outguessing equilibrium

This section introduces our notion of outguessing equilibrium. We first define the family of games we analyze.

Definition 13 *In a two person simultaneous normal form game we denote the players by* $i = 1, 2$*. We denote by* S_i *the pure strategy set of player* i*, where* $s_i \in S_i$ *and* $S = \times_{i=1}^{2} S_i$ *The utility (or payoff) of any player* i *is given by* $u_i(s_i, s_{-i}) \in \mathbb{R}$*, where* s_{-i} *denotes the strategy chosen by the other player.*

Our main assumptions are as follows.

Assumption 1 *We restrict attention to generic games, i.e. where the best response correspondance is a function. That means that there exists only one best response for any action of any of the two players.*

Assumption 2 *The game does not have a pure strategy Nash equilibrium.*

Notice that if the game had a pure strategy Nash equilibrium, mixed strategies and probability distributions would not have to be dealt with.

Assumption 3 *The game is repeated, the rounds are denoted by 1,2,. . . ,*n*,. . .*

Assumption 4 *The players are assumed to keep in mind the strategy profile of the previous round (i.e. their own previous choice and their opponent's previous choice) and nothing else.*

Assumption 5 *Players are assumed to play according to 0-reasoning, 1-reasoning, 2-reasoning, . . . ,* k*-reasoning, or a according to a probability distribution of the different reasoning levels. The distributions are exogenously given and do not change among different rounds of the game.*

The definition of the different reasoning levels are discussed in Section 1. Besides, we define 0-reasoning by playing the same strategy as in the previous round.

The exogenously given distribution over the set of reasoning levels is defined as follows.

Definition 14 *For any player* i *and any reasoning level* k *let* P_{ik} *denote the probability of acting according to* k*-reasoning.*

A player is considered smarter than its opponent if his expected reasoning level is higher than that of his opponent. This is how we grab the difference in the complexity of human thinking and try to point out why smarter people may win more frequently in several strategic interactions.

We begin the analysis with the description of the equilibrium concept for the simplest case, where both players have two strategies each.

3.1 The 2-by-2 model

Initially, we restrict attention to two-player 2x2 games with the following general payoff matrix:

			Player 2	
			q	$1-q$
			Left	Right
Player 1	p	Top	$u_{TL}; v_{TL}$	$u_{TR}; v_{TR}$
	$1-p$	Bottom	$u_{BL}; v_{BL}$	$u_{BR}; v_{BR}$

Table 1: The 2-by-2 game

According to Table 1, Player 1's strategies are Top and Bottom, while Player 2 can choose between Left and Right. $p, 1-p, q, 1-q$ are the respective strategy choice probabilities. Finally, u_{ij}, v_{ij} (where $i \in \{T, B\}$ and $j \in \{L, R\}$) are the two players' payoff levels given a certain strategy pair.

According to Assumption 2, we assume that the game does not have a pure strategy Nash-equilibrium. A necessary and sufficient condition for this is

$$u_{TL} > u_{BL}, u_{BR} > u_{TR}, v_{TR} > v_{TL}, v_{BL} > v_{BR}. \tag{12}$$

This means that the best responses of both players are given for any action of their opponent. E.g. if Player 1 chooses Top, then Player 2's best response is Right, as $v_{TR} > v_{TL}$.

For games that do not have a pure strategy Nash equilibrium, the classical solution is the mixed strategy Nash equilibrium. As a reference point, we provide the formulas for calculating the Nash-equlibrium mixing probabilities of the two players for the game using the notations of Table 1:

$$p_{nash} = \frac{v_{BL} - v_{BR}}{v_{BL} - v_{BR} + v_{TR} - v_{TL}}, \tag{13}$$

$$q_{nash} = \frac{u_{BR} - u_{TR}}{u_{BR} - u_{TR} + u_{TL} - u_{TR}}. \tag{14}$$

However, the mixed strategy Nash equilibrium has been criticized, as several experiments pointed out that it does not describe player behavior properly (e.g. [2, 7]). As described in the introduction, these findings led researchers to construct behavioral game theory models that may explain the way of strategic thinking more precisely.

Our model tries to provide a mathematical framework for player behavior.

We introduce our concept of play history in the next definition.

Definition 15 *We use the notion history for the strategy profile of the previous round of the game.*

The history of the game described by Table 1 can be the following: (Top;Left), (Top;Right), (Bottom;Left) and (Bottom;Right).

Depending on the history, we can define four different games, where the strategies and the payoffs are the same. The only difference is that both players keep the history in mind and this has an influence on their decisions, i.e. their strategy mixing probabilities.

The payoff and probability matrices with the four different histories are as follows.

			Player 2	
			q_{TL} Left	$1-q_{TL}$ Right
Player 1	p_{TL}	Top	$u_{TL};v_{TL}$	$u_{TR};v_{TR}$
	$1-p_{TL}$	Bottom	$u_{BL};v_{BL}$	$u_{BR};v_{BR}$

Table 2: The game with (Top, Left) history

			Player 2	
			q_{TR} Left	$1-q_{TR}$ Right
Player 1	p_{TR}	Top	$u_{TL};v_{TL}$	$u_{TR};v_{TR}$
	$1-p_{TR}$	Bottom	$u_{BL};v_{BL}$	$u_{BR};v_{BR}$

Table 3: The game with (Top, Right) history

			Player 2	
			q_{BL} Left	$1-q_{BL}$ Right
Player 1	p_{BL}	Top	$u_{TL};v_{TL}$	$u_{TR};v_{TR}$
	$1-p_{BL}$	Bottom	$u_{BL};v_{BL}$	$u_{BR};v_{BR}$

Table 4: The game with (Bottom, Left) history

			Player 2	
			q_{BR}	$1-q_{BR}$
			Left	Right
Player 1	p_{BR}	Top	$u_{TL};v_{TL}$	$u_{TR};v_{TR}$
	$1-p_{BR}$	Bottom	$u_{BL};v_{BL}$	$u_{BR};v_{BR}$

Table 5: The game with (Bottom, Right) history

An example for the game in Table 1 can be as follows.

Example 16 *We assume that Player 1 chooses strategy Top, while Player B chooses strategy Left in the first round of the game. Thus, for the second round the history is (Top,Left). Let Player 1 play according to 0-reasoning with certainty and Player 2 according to 1-reasoning with certainty. Thus, Player 1 will remain at strategy Top, while Player 2 will choose his best response to Top with certainty, that is, Right. We arrived at the (Top, Right) strategy pair with certainty. Using the notations of Table 2, this means that* $p_{TL}=1$*, while* $q_{TL}=0$*. Clearly, even if the reasoning levels follow a more complicated distribution, then* $p_{TL}\in[0,1]$ *and* $q_{TL}\in[0,1]$*. As we arrived at (Top, Right) with certainty, (Top, Right) becomes the history for the third round of the game. Applying again that Player 1 plays according to 0-reasoning and Player 2 plays according to 1-reasoning with certainty, the (Top, Right) profile will occur in the third round of the game. With the same logic, the process can be continued till any* k*th round of the game.*

If we consider any 2-by-2 game that satisfies our assumptions and a first-round strategy profile and a distribution on the set of reasonng levels is given for both players, there emerge the following questions:

1. What is the ex ante strategy choice distribution of the two players if the number of rounds $n\to\infty$? Is there any limiting distribution?
2. What is the expected payoff of the players for each round if $n\to\infty$? Is the series of expected payoffs convergent?

To answer these questions, we will apply the theory of Markov chains.

3.2 The Markov chain model of the 2-by-2 game

The outguessing model can be interpreted as a Markov chain. According to Assumption 4, the players keep in mind only the actions of the previous round

of the game. Let us define the Markov chain of the described 2-by-2 game.

Proposition 17 *The strategy profile sequence of the repeated game represents a Markov chain.*

Proof. The proof comes directly from Definition 3 and Assumption 4 that show that the strategy profile sequence $\{X_n\}$ $(n \geq 1)$ has the Markov property. □

In a 2-by-2 game we have 4 different strategy profiles, in our example these are (Top;Left), (Top;Right), (Bottom;Left) and (Bottom;Right), or in a shorter form: TL, TR, BL, BR. Thus, we can define the four states as follows:

State no.	Strategy profile
1	(Top;Left)
2	(Top;Right)
3	(Bottom;Left)
4	(Bottom;Right)

Table 6: States of the Markov chain

The transition matrix of the Markov chain can be obtained by using the data of the general payoff and probability matrices from the previous subsection.

Proposition 18 *The 4-by-4 transition matrix can be written as follows:*

$$T = \begin{pmatrix} p_{TL}q_{TL} & p_{TL}(1-q_{TL}) & (1-p_{TL})q_{TL} & (1-p_{TL})(1-q_{TL}) \\ p_{TR}q_{TR} & p_{TR}(1-q_{TR}) & (1-p_{TR})q_{TR} & (1-p_{TR})(1-q_{TR}) \\ p_{BL}q_{BL} & p_{BL}(1-q_{BL}) & (1-p_{BL})q_{BL} & (1-p_{BL})(1-q_{BL}) \\ p_{BR}q_{BR} & p_{BR}(1-q_{BR}) & (1-p_{BR})q_{BR} & (1-p_{BR})(1-q_{BR}) \end{pmatrix}.$$

Proof. The elements of the transition matrix are the probabilities of getting into a given state from a given previous state, i.e. the probabilities that a certain strategy profile will emerge given the strategy profile of the previous round. Using the previously defined p_{ij} and q_{ij} probabilities, and knowing that strategic decisions are independent from each other in a simultaneous game, we obtain the formula in the statement. □

Clearly, the transition matrix depends directly only on the players' probabilities of choosing a certain strategy with a given history. The transition

matrix is independent from the construction of these probabilities. Thus, it remains the same for all models where the players' actions depend only on the previous round and a probability distribution is exogenously given for both players on the set of reasoning levels.

The following lemma indicates that only 0, 1, 2 and 3-reasoning levels are relevant for the given 2-by-2 game.

Lemma 19 *For 2-by-2 games and for all* $k \geq 4$*,* k*-reasoning is equivalent to* $(k-4)$*-reasoning.*

Proof. The proof comes directly from the inequalities in (12). □

We need one more definition to be able to state the key result of the paper.

Definition 20 *Let us denote the initial strategy distribution of the players by* π_0*.*

The distribution over the state space (the set of strategy pairs) in the nth round can obviously be calculated as follows:

$$\pi_n = T^n \pi_0. \tag{15}$$

The key result states that under certain conditions there exists a limiting distribution if $n \to \infty$.

Proposition 21 *If* $P_{ik} > 0$ *(see Definition 14) for every* $i \in \{1, 2\}$ *and every* $k \in \{0, 1, 2, 3\}$ *and if* $n \to \infty$*, then there exists a limiting distribution* π *over the state space of the* $\{X_n\}$ $(n \geq 1)$ *Markov chain.*

Proof. If $P_{ik} > 0$ for every $i \in \{1, 2\}$ and every $k \in \{0, 1, 2, 3\}$, then it can easily be verified according to Definitions 8 and 13 that $\{X_n\}$ is an irreducible aperiodic Markov chain. Thus, according to Theorem 11 it is recurrent and has a limiting distribution. □

We arrived at our equilibrium concept. The outguessing equilibrium is defined as the limiting distribution.

Definition 22 *We call the limiting distribution* π *the outguessing equilibrium.*

According to the proof of Proposition 21, the key result is supported by Theorem 11: the strategy profile sequence of the repeated 2-by-2 game represents a Markov chain that has a limiting distribution.

As far as the players' expected payoffs are concerned, they can easily be determined by multiplying π' (π vector transposed) with the vector of the corresponding payoff levels.

By running our script, the limiting distribution π and the long-term expected payoffs can be calculated and visualized. It becomes clear that the player with the higher expected reasoning level has the higher expected payoff on the long run.

3.3 Numerical experiment—the matching pennies

In the matching pennies game, both players have to announce "heads" or "tails" at the same time. If the announcements are the same, Player 1 wins 1 from Player 2, otherwise Player 2 wins 1 from Player 1. The payoff matrix of the well-known zero-sum game is as follows:

			Player 2	
			q	$1-q$
			Left	Right
Player 1	p	Top	$1;-1$	$-1;1$
	$1-p$	Bottom	$-1;1$	$1;-1$

The Nash-equilibrium mixing probabilities are 50%-50% for both players.

Let us assume that in our model the initial strategy choice probabilities are 0.5 each (in the first round when there is no history). For the distributions over the set of reasoning levels (see Definition 14) let us assume that $P_{10} = 0.4$, $P_{11} = 0.2$, $P_{12} = 0.2$, $P_{13} = 0.2$, while $P_{10} = 0.2$, $P_{10} = 0.2$, $P_{10} = 0.4$, $P_{10} = 0.4$. Clearly, Player 2 is considered smarter due to his higher expected reasoning level.

We ran our script and the process of the expected payoffs calculated from $(\pi_1, \pi_2, \ldots, \pi_n, \ldots)$ and T is depicted in Figure 1.

It can easily be seen that these processes converge to certain limit values, a numerical evidence for Proposition 21. It is also verified for the Matching pennies that the smarter Player 2 (the above "+" sequence) has a higher expected payoff (0.232) than Player 1 (-0.232). According to the mixed strategy Nash equilibrium concept, both players would have zero expected payoff.

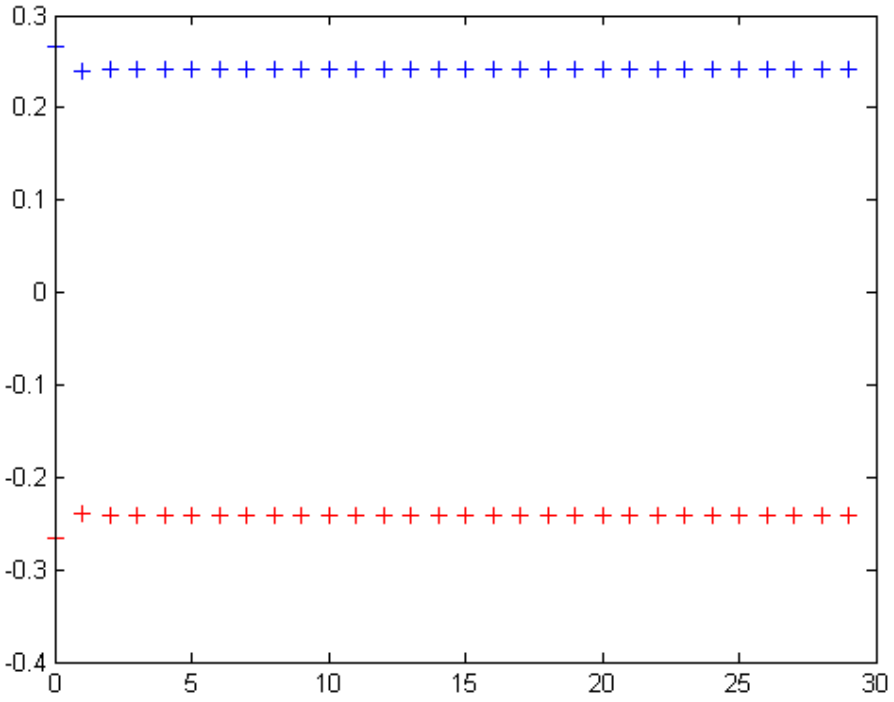


Figure 1: Long-term expected payoffs of the two players depending on the number of rounds; 2-by-2 case

3.4 The 3-by-3 case

If we consider a 3-by-3 game and keep all our assumptions, then the propositions trivially remain valid. The only exception is Lemma 19. The modified version for 3-by-3 games is as follows.

Lemma 23 *For 3-by-3 games and for all* $k \geq 6$*,* k*-reasoning is equivalent to* $(k-6)$*-reasoning.*

Proof. The proof comes directly from the modified version of inequalities in (12) for 3-by-3 games. □

What is important is that Proposition 21 remains valid if $k \in \{0, 1, 2, 3, 4, 5\}$. By running our script for 3-by-3 games, we can obtain numerical evidence for the results.

3.5 Numerical experiment—the Rock-paper-scissors game

The payoff matrix of the well-known Rock-paper-scissors game is as follows.

			Player 2		
			q_1 Rock	q_2 Paper	q_3 Scissors
	p_1	Rock	$0;0$	$-1;1$	$1;-1$
Player 1	p_2	Paper	$1;-1$	$0;0$	$-1,1$
	p_3	Scissors	$-1;1$	$1;-1$	$0;0$

We fixed the expected (average) reasoning level of Player 1 at 2.0 and that of Player 2 at 2.5 (not violating the conditions of Proposition 7). The expected payoffs are depicted in Figure 2 below.

Clearly, smarter Player 2 (crosses) "beats" Player 1 (dotted crosses) on the long run. Player 2's long term expected payoff lies at 0.038, while Player 1's is -0.038.

4 Notes about the script

Our script was written in Matlab. Its inputs are the following values:

- the payoff matrix of the corresponding 2-by-2 or 3-by-3 game
- the players' discrete probability distributions over the set of reasoning levels
- initial strategies (i.e. player behavior in the very first round–either a fixed strategy pair or an initial distribution)

The script works the following way. Firstly, from the given values the script calculates the transition matrix of the Markov chain. Then, the outguessing equilibrium (see Definition 13) and the long-term expected payoffs for both players are also calculated. Proposition 3 suggests that the "smarter" player (if there is one) beats its opponent on the long run.

Apart from the calculations, the power of the script is that the outguessing equilibrium concept can be tested for any 2-by-2 or 3-by-3 bimatrix game that does not have a Nash equilibrium on pure strategies.[1]

[1]Upon request, the authors provide the interested reader with the script with pleasure for testing purposes.

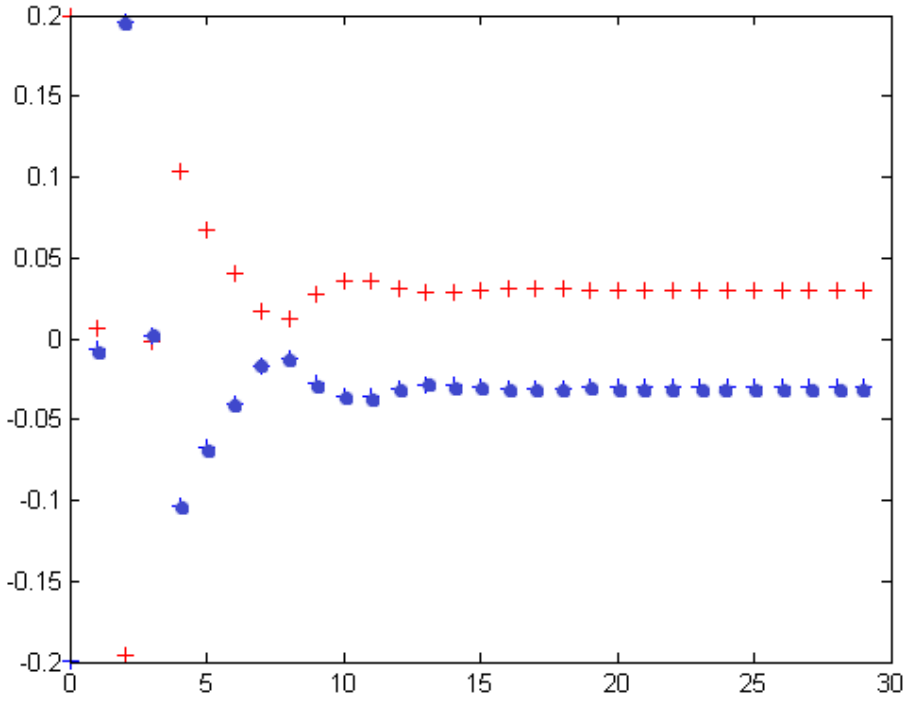


Figure 2: Long-term expected payoffs of the two players depending on the number of rounds; 3-by-3 case

5 Conclusions

Behavioral game theory has been dealing with the understanding of human behavior in strategic interactions. Among several different approaches, we have developed a behavioral model that aims at showing why "smarter" people outguess their opponents and win more frequently in some well-known zero-sum games.

Game theory is a useful modeling tool for network problems. We defined a behavioral model in a two-player non-cooperative network.

We used the concept of iterative reasoning to define smartness. The theory of Markov chains has proved to be a very useful technical tool to prove the

main result of the paper. Namely, an outguessing equilibrium according to our definition exists and can even be calculated.

A Matlab script supports the calculations and provides numerical evidence for our concept.

The authors wish to emphasize that the introduced model can not only be applied for the games recalled in the examples, but for any conflict situation that can be modeled by bimatrix games.

Although the theoretical results are proved, and numerical evidence is also provided, there have remained some interesting questions which are out of the scope of this paper. One of these questions is rather technical: what types of Markov chains (e.g. periodic, absorbing etc. . .) can emerge given a specific bimatrix game and initial strategy profile? Another one deals with the game theoretic assumptions: if either the number of players, or the simultanity of decisions were altered, or we allowed for non-generic games, how would the equilibrium outcome change? These problems are left for future research.

Acknowledgements

This paper was supported in part by the TÁMOP-4.2.2C-11/1/KONV-2012-0001 project supported by the European Union, co-financed by the European Social Fund.

References

[1] T. Alpcan, T. Basar, A Globally Stable Adaptive Congestion Control Scheme for Internet-Style Networks with Delay, *IEEE/ACM Trans. On Networking* **13** (2005) 1261–1274. ⇒71

[2] T. L. Beard, R. Beil, Do people rely on the self-interested maximization of others? An experimental test, *Management Science* **40,** 2 (1994) 252–262. ⇒72, 78

[3] S.C. Cabrera, C. M. Capra, R. Gomez, *The effects of common advice on one-shot traveller's dilemma games: Explaining behavior through an introspective model with errors*, Washington and Lee University working paper, 2002. ⇒72

[4] C.F. Camerer, *Behavioral Game Theory: Experiments in Strategic Interaction*, Princeton University Press, 2003, pp. 199–264. ⇒71, 72

[5] C. M. Capra, J.K. Goeree, R. Gomez, C. Holt, Anomalous behavior in traveller's dilemma?, *Amer. Economic Review* **89,** 3 (1999) 678–690. ⇒72

[6] S. N. Ethier, *The Doctrine of Chances*, Springer, 2010, pp. 119–155. ⇒73

[7] J. K. Goeree, C. Holt, Stochastic game theory: For playing games, not just for doing theory, *Proc. National Academy of Sciences* **96** (1999) 10564–10567. ⇒ 72, 78

[8] T. Ho, C. F. Camerer, K. Weigelt, Iterated dominance and iterated best-response in experimental "p-beauty contests", *Amer. Economic Review* **88** (1998) 947–969. ⇒72

[9] S. P. Meyn, R. L. Tweedie, *Markov chains and stochastic stability*, Cambridge University Press, 2009. ⇒73

[10] S. P. Meyn, *Control Techniques for Complex Networks*, Cambridge University Press, 2007. ⇒73

[11] H. Moulin, *Game Theory for the Social Sciences*, New York University Press, 1986. ⇒72

[12] R. Nagel, Unravelling in guessing games: an experimental study, *Amer. Economic Review* **85,** 5 (1995) 1313–1326. ⇒72

[13] J. Nash, Non-cooperative games, *Annals of Mathematics* **54,** 2 (1951) 286-295. ⇒71

[14] J. R. Norris, *Markov chains*, Cambridge University Press, 1998. ⇒73

[15] I. A. Shah, S. Jan, I. Khan, S. Qamar, An overview of game theory and its applications in communication networks *Int. J. Multidisciplinary Sciences and Engineering* **3,** 4 (2012) 5–11. ⇒71

[16] Y. Sovik, *Impossible bets: an experimental study*, University of Oslo working paper, 1999. ⇒72

Received: February 16, 2014 • Revised: March 31, 2014

Acta Univ. Sapientiae, Informatica, **6,** 1 (2014) 89–116

Energy, Laplacian energy of double graphs and new families of equienergetic graphs

Hilal A. GANIE
University of Kashmir
Srinagar, India
email: hilahmad1119kt@gmail.com

Shariefuddin PIRZADA
University of Kashmir
Srinagar, India
email:
pirzadasd@kashmiruniversity.ac.in

Antal IVÁNYI
Eötvös Loránd University
Faculty of Informatics
Budapest, Hungary
email: tony@inf.elte.hu

Abstract. For a graph G with vertex set $V(G) = \{v_1, v_2, \ldots, v_n\}$, the extended double cover G^* is a bipartite graph with bipartition (X, Y), $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_n\}$, where two vertices x_i and y_j are adjacent if and only if $i = j$ or v_i adjacent to v_j in G. The double graph $D[G]$ of G is a graph obtained by taking two copies of G and joining each vertex in one copy with the neighbors of corresponding vertex in another copy. In this paper we study energy and Laplacian energy of the graphs G^* and $D[G]$, L-spectra of G^{k*} the k-th iterated extended double cover of G. We obtain a formula for the number of spanning trees of G^*. We also obtain some new families of equienergetic and L-equienergetic graphs.

1 Introduction

Let G be finite, undirected, simple graph with n vertices and m edges having vertex set $V(G) = \{v_1, v_2, \ldots, v_n\}$. Throughout this paper we denote such a

Computing Classification System 1998: G.2.2
Mathematics Subject Classification 2010: 05C30, 05C50
Key words and phrases: double graph, spectra, energy, Laplacian energy, L-equienergetic, equienergetic
DOI:10.2478/ausi-2014-0020

graph by $G(n, m)$. The *adjacency matrix* $A = (a_{ij})$ of G is a $(0,1)$-square matrix of order n whose (i,j)-entry is equal to one if v_i is adjacent to v_j and equal to zero, otherwise. The spectrum of the adjacency matrix is called the A-*spectrum* of G. If $\lambda_1, \lambda_2, \ldots, \lambda_n$ is the adjacency spectrum of G, the *energy* of G is defined as $E(G) = \sum_{i=1}^{n} |\lambda_i|$. This quantity introduced by I. Gutman [16] has noteworthy chemical applications.

Let $D(G) = diag(d_1, d_2, \ldots, d_n)$ be the diagonal matrix associated to G, where d_i is the degree of vertex v_i. The matrices $L(G) = D(G) - A(G)$ and $L^+(G) = D(G) + A(G)$ are called *Laplacian* and *signless Laplacian matrices* and their spectras are respectively called *Laplacian spectrum* (L-spectrum) and *signless Laplacian spectrum* (Q-spectrum) of G. Being real symmetric, positive semi-definite matrices, let $0 = \mu_n \leq \mu_{n-1} \leq \cdots \leq \mu_1$ and $0 \leq \mu_n^+ \leq \mu_{n-1}^+ \leq \cdots \leq \mu_1^+$ be respectively the L-spectrum and Q-spectrum of G. It is well-known that $\mu_n = 0$ with multiplicity equal to the number of connected components of G (see [11]). Fiedler [11] showed that a graph G is connected if and only if its second smallest Laplacian eigenvalue is positive and called it as the *algebraic connectivity* of the graph G. Also it is well-known [8] that for a bipartite graph the L-spectra and Q-spectra are identical. The *Laplacian energy* of a graph G as put forward by Gutman and Zhou [17] is defined as

$$LE(G) = \sum_{i=1}^{n} \left| \mu_i - \frac{2m}{n} \right| .$$

This quantity, which is an extension of graph-energy concept has found remarkable chemical applications beyond the molecular orbital theory of conjugated molecules [22]. Both energy and Laplacian energy have been extensively studied in the literature (see [1, 2, 5, 9, 10, 12, 13, 14, 15, 18, 19, 20, 24, 25, 26, 27, 29, 30] and the references in them). Based on the above definition, the *signless Laplacian energy* of a graph G is defined as

$$LE^+(G) = \sum_{i=1}^{n} \left| \mu_i^+ - \frac{2m}{n} \right| ,$$

where μ_i^+ $(i = 1, 2, \ldots, n)$is the signless Laplacian spectra of G. It is easy to see that

$$tr(L(G)) = \sum_{i=1}^{n} \mu_i = \sum_{i=1}^{n-1} \mu_i = 2m \quad \text{and} \quad tr(LE^+(G)) = \sum_{i=1}^{n} \mu_i^+ = 2m,$$

where tr is the *trace* of the matrix.

Two graphs G_1 and G_2 of same order are said to be *equienergetic* if $E(G_1) = E(G_2)$ [3, 23]. In analogy to this two graphs G_1 and G_2 of same order are said to L-*equienergetic* if $LE(G_1) = LE(G_2)$ and Q-*equienergetic* if $LE^+(G_1) = LE^+(G_2)$. Since cospectral (Laplacian cospectral) graphs are always equienergetic (L-equienergetic), the problem of constructing equienergetic (L-equienergetic) graphs is only considered for non-cospectral (non Laplacian cospectral) graphs.

The *extended double cover* [6] of the graph $G(n, m)$ with vertex set $V(G) = \{v_1, v_2, \ldots, v_n\}$ is a bipartite graph G^* with bipartition (X, Y), $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_n\}$, where two vertices x_i and y_j are adjacent if and only if $i = j$ or v_i adjacent v_j in G. It is easy to see that G^* is connected if and only if G is connected and a vertex v_i is of degree d_i in G if and only if it is of degree $d_i + 1$ in G^*. Also the extended double cover G^* of the graph G always contains a perfect matching. The double graph $D[G]$ of G is a graph obtained by taking two copies of G and joining each vertex in one copy with the neighbors of corresponding vertex in another copy. The k-fold graph $D^k[G]$ [21] of the graph G is obtained by taking k copies of the graph G and joining each vertex in one of the copy with the neighbors of the corresponding vertices in the other copies. If T_n is the graph obtained from the complete graph K_n by adding a loop at each of the vertex, it is easy to see that $D^k[G] = G \otimes T_k$. In this paper we study energy, Laplacian energy of the graphs G^* and $D[G]$, the L-spectra of G^{k*} the k-th iterated extended double cover of G and obtain a formula for the number of spanning trees of G^*. We also obtain some new families of the equienergetic and L-equienergetic graphs.

We denote the complement of graph G by $\overline{G}$, the complete graph on n vertices by K_n, the empty graph by $\overline{K}_n$ and the complete bipartite graph with cardinalities of partite sets q and r by $K_{q,r}$. The rest of the paper is organized as follows. In Section 2, energy of the graphs G^* and $D^k[G]$ are obtained and some new families of equienergetic graphs are given, in Section 3 L-spectra of G^{k*} and a formula for the number of spanning tress of G^* is obtained and in Section 4 Laplacian energy of the graphs G^* and $D^k[G]$ and the construction of some new families of L-equienergetic graphs by using the graphs G^{k*} and $D^k[G]$ is presented.

2 Energy of double graphs

In this section we find the energy of the graphs G^* and $D^k[G]$. We also construct some new families of equienergetic graphs based on these graphs.

For the graphs G_1 and G_2 with disjoint vertex sets $V(G_1)$ and $V(G_2)$, the

Cartesian product is a graph $G = G_1 \times G_2$ with vertex set $V(G_1) \times V(G_2)$ and an edge $((u_1, v_1), (u_2, v_2))$ if and only if $u_1 = u_2$ and (v_1, v_2) is an edge of G_2 or $v_1 = v_2$ and (u_1, u_2) is an edge of G_1. The following result gives the A-spectra (L-spectra) of the Cartesian product of graphs.

Lemma 1 (Cvetkovic, Doob, Sachs, 1980 [7]) *If* $G_1(n_1, m_1)$ *and* $G_2(n_2, m_2)$ *are two graphs having* A*-spectra (*L*-spectra) respectively as,* $\mu_1, \mu_2, \ldots, \mu_{n_1}$ *and* $\sigma_1, \sigma_2, \ldots, \sigma_{n_2}$*, then the* A*-spectra (*L*-spectra) of* $G = G_1 \times G_2$ *is* $\mu_i + \sigma_j$*, where* $i = 1, 2, \ldots, n_1$ *and* $j = 1, 2, \ldots, n_2$*.*

The *conjunction* (*Kronecker product*) of G_1 and G_2 is a graph $G = G_1 \otimes G_2$ with vertex set $V(G_1) \times V(G_2)$ and an edge $((u_1, v_1), (u_2, v_2))$ if and only if (u_1, u_2) and (v_1, v_2) are edges in G_1 and G_2, respectively. The following result gives the A-spectra (L-spectra) of the Kronecker product of graphs.

Lemma 2 (Cvetkovic, Doob, Sachs, 1980 [7]) *If* $G_1(n_1, m_1)$ *and* $G_2(n_2, m_2)$ *are two graphs having* A*-spectra (*L*-spectra) respectively as* $\mu_1, \mu_2, \ldots, \mu_{n_1}$ *and* $\sigma_1, \sigma_2, \ldots, \sigma_{n_2}$*, then the* A*-spectra (*L*-spectra) of* $G = G_1 \otimes G_2$ *is* $\mu_i \sigma_j$*, where* $i = 1, 2, \ldots, n_1$ *and* $j = 1, 2, \ldots, n_2$*.*

The *join product* of G_1 and G_2 is a graph $G = G_1 \vee G_2$ with vertex set $V(G_1) \cup V(G_2)$ and an edge set consisting of all the edges of G_1 and G_2 together with the edges joining each vertex of G_1 with every vertex of G_2. The L-spectra of join product of graphs is given by the following result.

Lemma 3 (Cvetkovic, Doob, Sachs, 1980 [7]) *If* $G_1(n_1, m_1)$ *and* $G_2(n_2, m_2)$ *are two graphs having* L*-spectra respectively as* $\mu_1, \mu_2, \ldots, \mu_{n_1-1}, \mu_{n_1} = 0$ *and* $\sigma_1, \sigma_2, \ldots, \sigma_{n_2-1}, \sigma_{n_2} = 0$*, then the* L*-spectra of* $G = G_1 \vee G_2$ *is* $n_1 + n_2, n_1 + \sigma_1, n_1 + \sigma_2, \ldots, n_1 + \sigma_{n_2-1}, n_2 + \mu_1, n_2 + \mu_2, \ldots, n_2 + \mu_{n_1-1}, 0$*.*

The following result gives the A-spectra of G^*, the extended double cover of the graph G.

Theorem 4 (Cvetkovic, Doob, Sachs, 1980 [7], Chen, 2004 [6]) *If* $\lambda_1, \lambda_2, \ldots, \lambda_n$ *is the* A*-spectra of a graph* G*, then the* A*-spectra of the graph* G^* *is* $\pm(\lambda_1 + 1), \pm(\lambda_2 + 1), \ldots, \pm(\lambda_n + 1)$*.*

If $\lambda_1, \lambda_2, \ldots, \lambda_n$ is the A-spectra of the graph G, then by Lemma 1, the A-spectra of the graph $G \times K_2$ is $\lambda_i + 1, \lambda_i - 1$ for $1 \leq i \leq n$. It is clear from Theorem 4, that the graphs $G \times K_2$ and G^* are cospectral if and only if G is bipartite [6]. If $D^k[G]$ is the k-fold graph of the graph G, the A-spectra of $D^k[G]$ is given by the following result.

Theorem 5 (Cvetkovic, Doob, Sachs, 1980 [7], Marino, Salvi, 2007 [21]) *If* $\lambda_1, \lambda_2, \ldots, \lambda_n$ *is the* A*-spectra of a graph* G*, then the* A*-spectra of the graph* $D^k[G]$ *is* $k\lambda_1, k\lambda_2, \ldots, k\lambda_n, 0$ $((k-1)n\text{times})$.

If $\lambda_1, \lambda_2, \ldots, \lambda_n$ is the A-spectra of the graph G, then by Theorem 4, the A-spectra of the graph G^* is $\pm(\lambda_1+1), \pm(\lambda_2+1), \ldots, \pm(\lambda_n+1)$ and by Theorem 5, the A-spectra of $D^k[G]$ is $k\lambda_1, k\lambda_2, \ldots, k\lambda_n, 0$ ($(k-1)n$ times). Therefore,

$$E(G^*) = \sum_{i=1}^{n} |\lambda_i + 1| + \sum_{i=1}^{n} |-\lambda_i - 1| = 2\sum_{i=1}^{n} |\lambda_i + 1|,$$

and

$$E(D^k[G]) = \sum_{i=1}^{n} |2\lambda_i| = 2\sum_{i=1}^{n} |\lambda_| = kE(G).$$

If $\lambda_1, \lambda_2, \ldots, \lambda_n$ is the A-spectra of a graph G, then the A-spectra of the graph $(G \otimes K_2) \times K_2$ is $\lambda_i + 1, \lambda_i - 1, -\lambda_i + 1, -\lambda_i - 1$, $1 \le i \le n$. Therefore,

$$\begin{aligned} E((G\otimes K_2)\times K_2) &= 2\sum_{i=1}^{n} |\lambda_i + 1| + 2\sum_{i=1}^{n} |\lambda_i - 1| = 2\left(\sum_{i=1}^{n} |\lambda_i + 1| + \sum_{i=1}^{n} |\lambda_i - 1|\right) \\ &= 2E(G \times K_2) = E(2(G \times K_2)) = E((G \times K_2) \cup (G \times K_2)). \end{aligned}$$

From the above discussion, we observe that the graphs $(G \otimes K_2) \times K_2$ and $(G \times K_2) \cup (G \times K_2)$ are equienergetic. Moreover, if the graph G is a bipartite graph then the graphs $(G\otimes K_2)\times K_2$ and $G^* \cup G^*$ are also equienergetic graphs.

As seen above $E(D^k[G]) = k\sum_{i=1}^{n} |\lambda_i| = kE(G) = E(kG) = E(G \cup G \cup \cdots \cup G)$ (G is repeated k times). This shows that the graphs $D^k[G]$ and $(G \cup G \cup \cdots \cup G)$ (G is repeated k times) are non-cospectral equienergetic. However, we show for any graph G the graphs D[G] and $G \otimes K_2$ are always equienergetic non-cospectral graphs.

Theorem 6 *If* D[G] *is the double graph of the graph* G*, then the graphs* $G \otimes K_2$ *and* D[G] *are non-cospectral equienergetic graphs.*

Proof. Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the eigenvalues of the graph G, then by Lemma 2, the eigenvalues of the graph $G \otimes K_2$ are $\lambda_i, -\lambda_i$ for $1 \le i \le n$ and by Theorem 5 (for $k = 2$), the eigenvalues of the graph D[G] are $2\lambda_i$, 0 (n times) for $1 \le i \le n$. Therefore,

$$E(G \otimes K_2) = \sum_{i=1}^{n} |\lambda_i| + \sum_{i=1}^{n} |-\lambda_i| = 2\sum_{i=1}^{n} |\lambda_i|.$$

Also,

$$E(D[G]) = \sum_{i=1}^{n} |2\lambda_i| = 2\sum_{i=1}^{n} |\lambda_i|.$$

Clearly these graphs are non-cospectral, so the result follows. □

In general, if $D^k[G]$ be the k-fold graph of the graph G, we have the following observation.

Theorem 7 *If $D^k[G]$ is the k-fold graph of the graph G, then the graphs $D^k[G]$ and $G \otimes sK_2$ are non-cospectral equienergetic graphs if and only if $k = 2^s$.*

Proof. If $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the eigenvalues of the graph, then by Lemma 2, the eigenvalues of the graph $G \otimes sK_2$ are λ_i (2^{s-1} times), $-\lambda_i$ (2^{s-1} times) for $1 \leq i \leq n$ and by Theorem 5, the eigenvalues of the graph $D^k[G]$ are $k\lambda_i$, 0 $((k-1)n$ times) for $1 \leq i \leq n$. Therefore,

$$E(G \otimes sK_2) = 2^{s-1} \sum_{i=1}^{n} |\lambda_i| + 2^{s-1} \sum_{i=1}^{n} |-\lambda_i| = 2^s \sum_{i=1}^{n} |\lambda_i|. \tag{1}$$

Also,

$$E(D^k[G]) = \sum_{i=1}^{n} |k\lambda_i| = k\sum_{i=1}^{n} |\lambda_i|. \tag{2}$$

From (1) and (2) it is clear that $E(G \otimes sK_2) = E(D^k[G])$ if and only if $k = 2^s$. □

Let G^{**} be the extended double cover of the graph G^*. We have the following result.

Theorem 8 *If G is an n-vertex graph, then $E(G^* \otimes K_2) = E(G^{**})$, if $|\lambda_i| \geq 2$, for all non-zero eigenvalues of G. Moreover these graphs are non-cospectral with equal number of vertices.*

Proof. Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the eigenvalues of the graph G. By Theorem 4, the eigenvalues of the graph G^* are $\lambda_i+1, -(\lambda_i+1)$ for $1 \leq i \leq n$ and so of G^{**} are $\lambda_i + 2, \lambda_i, -(\lambda_i + 2), -\lambda_i$ for $1 \leq i \leq n$. Also by Lemma 2, the eigenvalues of the graph $G^* \otimes K_2$ are $\lambda_i + 1, -(\lambda_i + 1), \lambda_i + 1, -(\lambda_i + 1)$ for $1 \leq i \leq n$. Assume that $|\lambda_i| \geq 2$. Then

$$|\lambda_i + 1| = \begin{cases} |\lambda_i| + 1, & \text{if } \lambda_i \geq 0 \\ |\lambda_i| - 1, & \text{if } \lambda_i < 0 \end{cases}, \quad |\lambda_i + 2| = \begin{cases} |\lambda_i| + 2, & \text{if } \lambda_i \geq 0 \\ |\lambda_i| - 2, & \text{if } \lambda_i < 0. \end{cases}$$

Therefore,

$$\begin{aligned}
E(G^{**}) &= 2\sum_{i=1}^{n}|\lambda_i+2| + 2\sum_{i=1}^{n}|\lambda_i| = 2\left(\sum_{\lambda_i\geq 0}|\lambda_i+2| + \sum_{\lambda_i<0}|\lambda_i+2| + \sum_{i=1}^{n}|\lambda_i|\right)\\
&= 2\left(\sum_{\lambda_i\geq 0}|\lambda_i| + 2 + \sum_{\lambda_i<0}|\lambda_i| - 2 + \sum_{i=1}^{n}|\lambda_i|\right)\\
&= 2\left(\sum_{i=1}^{n}|\lambda_i| + \sum_{i=1}^{n}|\lambda_i| + 2(\sum_{\lambda_i\geq 0}1 - \sum_{\lambda_i<0}1)\right)\\
&= 4\sum_{i=1}^{n}|\lambda_i| + 4\theta,
\end{aligned}$$

where θ is the difference between the number of nonnegative and negative eigenvalues of G and

$$\begin{aligned}
E(G^*\otimes K_2) &= 2\left(\sum_{i=1}^{n}|\lambda_i+1| + \sum_{i=1}^{n}|-(\lambda_i+1)|\right) = 4\sum_{i=1}^{n}|\lambda_i+1|\\
&= 4\left(\sum_{\lambda_i\geq 0}|\lambda_i+1| + \sum_{\lambda_i<0}|\lambda_i+1|\right) = 4\left(\sum_{\lambda_i\geq 0}|\lambda_i| + 1 + \sum_{\lambda_i<0}|\lambda_i| - 1|\right)\\
&= 4\sum_{i=1}^{n}|\lambda_i| + 4\left(\sum_{\lambda_i\geq 0}1 - \sum_{\lambda_i<0}1\right) = 4\sum_{i=1}^{n}|\lambda_i| + 4\theta.
\end{aligned}$$

Clearly these graphs are noncospectral with same number of vertices. □

Let G be a bipartite graph. It is well-known that the spectra of G is symmetric about the origin, so half of the nonzero eigenvalues of G lies to the left and half lies to the right of origin. Therefore if G is a bipartite graph having all its eigenvalues nonzero, the number of positive and negative eigenvalues of G are same. Keeping this in mind we have the following result.

Theorem 9 *If G^* is the extended double cover of the bipartite graph G, then the graphs G^* and $D[G]$ are noncospectral equienergetic if and only if $|\lambda_i| \geq 1$ for all $1 \leq i \leq n$.*

Proof. Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the eigenvalues of the graph G. By Theorem 4, the eigenvalues of the graph G^* are $\lambda_i+1, -\lambda_i-1$ for $1 \leq i \leq n$ and by Theorem 5,

the eigenvalues of the graph $D[G]$ are $2\lambda_i,\ \ 0$ (n times) for $1 \le i \le n$. Suppose that $|\lambda_i| \ge 1$ for $i = 1, 2, \ldots, n$, then

$$|\lambda_i + 1| = \begin{cases} |\lambda_i| + 1, & \text{if } \lambda_i > 0 \\ |\lambda_i| - 1, & \text{if } \lambda_i < 0. \end{cases}$$

Therefore,

$$\begin{aligned} E(G^*) &= \sum_{i=1}^{n} |\lambda_i + 1| + \sum_{i=1}^{n} |-\lambda_i - 1| = 2\sum_{i-1}^{n} |\lambda_i + 1| \\ &= 2\left(\sum_{\lambda_i>0} |\lambda_i + 1| + \sum_{\lambda_i<0} |\lambda_i + 1|\right) = 2\left(\sum_{\lambda_i>0} (|\lambda_i| + 1) + \sum_{\lambda_i<0} (|\lambda_i| - 1)\right) \\ &= 2\left((\sum_{\lambda_i>0} |\lambda_i| + \sum_{\lambda_<0} |\lambda_i|) + (\sum_{\lambda_i>0} 1 - \sum_{\lambda_<0} 1)\right) = 2\sum_{i=1}^{n} |\lambda_i| = E(D[G]). \end{aligned}$$

Clearly these graphs are noncospectral with same number of vertices.

Conversely, suppose that the graphs G^* and $D[G]$ are noncospectral equienergetic. We will show that $|\lambda_i| \ge 1$ for all $1 \le i \le n$.

Assume to the contrary that $|\lambda_i| < 1$ for some i. Then for this i, $|\lambda_i + 1| = \lambda_i + 1$. Without loss of generality, suppose that the eigenvalues of G satisfy $|\lambda_i| \ge 1$, for $i = 1, 2, \ldots, k$ and $|\lambda_i| < 1$, for $i = k+1, k+2, \ldots, n$, since the eigenvalues are real and reordering does not effect the argument. We have the following cases to consider.

Case i. If $\lambda_i > 0$ for $i = 1, 2, \ldots, k$ and $\lambda_i \ge 0$ for $i = k+1, k+2, \ldots, n$, then

$$E(G^*) = 2\left(\sum_{i=1}^{k} |\lambda_i + 1| + \sum_{i=k+1}^{n} |\lambda_i + 1|\right) = 2\left(\sum_{i=1}^{n} |\lambda_i| + n\right).$$

Case ii. If $\lambda_i > 0$ for $i = 1, 2, \ldots, k$ and $\lambda_i \le 0$ for $i = k+1, k+2, \ldots, n$, then if θ_0 is the number of zero eigenvalues of G, we have

$$\begin{aligned} E(G^*) &= 2\left(\sum_{i=1}^{k} |\lambda_i + 1| + \sum_{i=k+1}^{n} |\lambda_i + 1|\right) = 2\left(\sum_{i=1}^{k} (|\lambda_i| + 1) + \sum_{i=k+1}^{n} (\lambda_i + 1)\right) \\ &> 2\left(\sum_{i=1}^{k} (|\lambda_i| + 1) + \sum_{i=k+1}^{n} (|\lambda_i| - 1)\right) = 2\left(\sum_{i=1}^{n} |\lambda_i| - \theta_0\right). \end{aligned}$$

Case iii. If $\lambda_i < 0$ for $i = 1, 2, \ldots, k$ and $\lambda_i \geq 0$ for $i = k+1, k+2, \ldots, n$, then

$$E(G^*) = 2\left(\sum_{i=1}^{k} |\lambda_i + 1| + \sum_{i=k+1}^{n} |\lambda_i + 1|\right) = 2\left(\sum_{i=1}^{k} (|\lambda_i| - 1) + \sum_{i=k+1}^{n} (|\lambda_i| + 1)\right)$$
$$= 2\left(\sum_{i=1}^{n} |\lambda_i| + \theta_0\right).$$

Case iv. If $\lambda_i < 0$ for $i = 1, 2, \ldots, k$ and $\lambda_i \leq 0$ for $i = k+1, k+2, \ldots, n$, then

$$E(G^*) = 2\left(\sum_{i=1}^{k} |\lambda_i + 1| + \sum_{i=k+1}^{n} |\lambda_i + 1|\right) = 2\left(\sum_{i=1}^{k} (|\lambda_i| - 1) + \sum_{i=k+1}^{n} (\lambda_i + 1)\right)$$
$$> 2\left(\sum_{i=1}^{k} (|\lambda_i| - 1) + \sum_{i=k+1}^{n} (|\lambda_i| - 1)\right) = 2\left(\sum_{i=1}^{n} |\lambda_i| - n\right).$$

Clearly in all these cases, we obtain $E(G^*) \neq E(D[G])$, a contradiction. Therefore the result follows. □

We can also prove Theorem 9 by using Theorem 6, the fact that the graphs G^* and $G \times K_2$ are cospectral if G is bipartite [6, Theorem 2] and the graphs $G \times K_2$ and $G \otimes K_2$ are equienergetic if an only if $|\lambda_i| \geq 1$ [4, Theorem 8].

3 The Laplacian spectra of G^{k*}

Let G^* be the extended double cover of the graph G, define $G^{**} = (G^*)^*$, and in general $G^{k*} = (G^{(k-1)*})^*$, $k \geq 1$, called the k-qtextitth iterated double cover graph of G. The A-spectra of G^{k*} is given in [6]. Here we obtain the L-spectra of the k-th iterated extended double cover G^{k*} of the graph G. Since the graph G^{k*} is always bipartite for $k \geq 1$, therefore its Laplacian (L-spectra) and signless Laplacian (Q-spectra) spectra are same.

For any complex square matrices A and B of same order, the following observation can be seen in ([28, page 41]).

Theorem 10 *If A and B are complex square matrices of same order, then*

$$\begin{vmatrix} A & B \\ B & A \end{vmatrix} = |A + B||A - B|,$$

where the symbol $|\ |$ denotes the determinant of a matrix.

We first obtain the L-spectra of G^*, the extended double cover of the graph G, in the following result.

Theorem 11 *Let* $G(n,m)$ *be an* n*-vertex graph having Laplacian and signless Laplacian spectra, respectively as* $0=\mu_n<\mu_{n-1}\le\ldots\le\mu_1$ *and* $0<\mu_n^+<\mu_{n-1}^+\le\ldots\le\mu_1^+$*. Then the Laplacian spectra of* G^* *is* $\mu_1,\mu_2,\ldots,\mu_n,\ \ \mu_1^++2,\mu_2^++2,\ldots,\mu_n^++2$*.*

Proof. Let $A(G)$ be the adjacency matrix of the graph G. By a suitable relabelling of vertices it can be seen that the adjacency matrix $A(G^*)$ of the graph G^* is

$$A(G^*)=\begin{pmatrix}0 & A(G)+I_n\\ A(G)+I_n & 0\end{pmatrix}.$$

Let $D(G)$ and $D(G^*)$ be respectively the degree matrices of the graphs G and G^*. It is easy to see that

$$D(G^*)=\begin{pmatrix}D(G)+I_n & 0\\ 0 & D(G)+I_n\end{pmatrix}.$$

Therefore, Laplacian matrix $L(G^*)$ of G^* is

$$L(G^*)=D(G^*)-A(G^*)=\begin{pmatrix}D(G)+I_n & -(A(G)+I_n)\\ -(A(G)+I_n) & D(G)+I_n\end{pmatrix}.$$

So the Laplacian characteristic polynomial of G^* is

$$C_{G^*}(\lambda)=|\lambda I_{2n}-L(G^*)|=\begin{vmatrix}(\lambda-1)I_n-D(G) & A(G)+I_n\\ A(G)+I_n & (\lambda-1)I_n-D(G)\end{vmatrix}$$

$$=|((\lambda-1)I_n-D(G))-(A(G)+I_n)|\,|((\lambda-1)I_n-D(G))+(A(G)+I_n)|$$

$$=|(\lambda-2)I_n-(D(G)+A(G))|\,|\lambda I_n-(D(G)-A(G))|$$

$$=Q_G(\lambda-2)C_G(\lambda).$$

From this the result follows. □

We now obtain the L-spectra of G^{k*} as follows.

Theorem 12 *Let* $G(n,m)$ *be a graph having* L*-spectra* μ_i*, and* Q*-spectra* μ_i^+*,* $1\le i\le n$*. The* L*-spectra of the graph* G^{k*} *is* $\mu_i\left(\binom{k}{0}\text{ times}\right)$*,* $\mu_i+2\left(\binom{k-1}{1}\text{ times}\right)$*,* $\mu_i^+ + 2\left(\binom{k-1}{0}\text{ times}\right)$*,* $\mu_i + 4\left(\binom{k-1}{2}\text{ times}\right)$*,* $\mu_i^+ + 4\left(\binom{k-1}{1}\times\right),\ldots,\mu_i+2(k-2)\left(\binom{k-1}{k-2}\text{ times}\right)$*,* $\mu_i^+ + 2(k-2)\left(\binom{k-1}{k-3}\text{ times}\right)$*,* $\mu_i + 2(k-1)\left(\binom{k-1}{k-1}\text{ times}\right)$*,* $\mu_i^+ + 2(k-1)\left(\binom{k-1}{k-2}\text{ times}\right)$*,* $\mu_i^+ + 2k\left(\binom{k}{k}\text{ times}\right)$*, where* $1\le i\le n$*.*

Proof. We prove this result by induction and we use induction on k. For $k = 1$, the result follows by Theorem 11. For $k = 2$, we have $G^{2*} = G^{**}$. Let $A(G^*)$ and $A(G^{**})$ be the adjacency matrices respectively of the graphs G^* and G^{**}. It is not difficult to see that

$$A(G^{**}) = \begin{pmatrix} 0 & A(G^*) + I_{2n} \\ A(G^*) + I_{2n} & 0 \end{pmatrix}.$$

Let $D(G^*)$ and $D(G^{**})$ be respectively the degree matrices of G^* and G^{**}. It can be seen that

$$D(G^{**}) = \begin{pmatrix} D(G^*) + I_{2n} & 0 \\ 0 & D(G^*) + I_{2n} \end{pmatrix}.$$

Therefore the Laplacian matrix of G^{**} is

$$L(G^{**}) = D(G^{**}) - A(G^{**}) =$$

$$\begin{pmatrix} D(G^*) + I_{2n} & -(A(G^*) + I_{2n}) \\ -(A(G^*) + I_{2n}) & D(G^*) + I_{2n} \end{pmatrix}.$$

So the Laplacian characteristic polynomial of G^{**} is

$$C_{G^{**}}(\lambda) = |\lambda I_{4n} - L(G^{**})| = \begin{vmatrix} (\lambda-1)I_{2n} - D(G^*) & A(G^*) + I_{2n} \\ A(G^*) + I_{2n} & (\lambda-1)I_{2n} - D(G^*) \end{vmatrix}$$

$$= |((\lambda-1)I_{2n} - D(G^*)) - (A(G^*) + I_{2n})|\,|((\lambda-1)I_{2n} - D(G^*)) + (A(G^*) + I_{2n})|$$

$$= |(\lambda-2)I_{2n} - (D(G^*) + A(G^*))|\,|\lambda I_{2n} - (D(G^*) - A(G^*))|$$

$$= Q_{G^*}(\lambda-2)C_{G^*}(\lambda).$$

From this it is clear that the L-spectra of G^{**} is $\mu_i,\ \mu_i + 2,\ \mu_i^+ + 2, \mu_i^+ + 4$, for $1 \le i \le n$, that is L-spectra of G^{**} is μ_i $\left(\binom{2}{0}\text{ times}\right)$, $\mu_i + 2$ $\left(\binom{1}{1}\text{ times}\right)$, $\mu_i^+ + 2$ $\left(\binom{1}{0}\text{ times}\right)$, and $\mu_i^+ + 4$ $\left(\binom{2}{2}\text{ times}\right)$. Therefore the result is true in this case. Assume that the result is true for $k = s - 1$. Then by induction hypothesis the L-spectra of $G^{(s-1)*}$ is μ_i $\left(\binom{s-1}{0}\text{ times}\right)$, $\mu_i + 2$ $\left(\binom{s-2}{1}\text{ times}\right)$, $\mu_i^+ + 2$ $\left(\binom{s-2}{0}\text{ times}\right)$, $\ldots$, $\mu_i + 2(s-2)$ $\left(\binom{s-2}{s-2}\text{ times}\right)$, $\mu_i^+ + 2(s-2)$ $\left(\binom{s-2}{s-3}\text{ times}\right)$, $\mu_i^+ + 2(s-1)$ $\left(\binom{s-1}{s-1}\text{ times}\right)$. Now for $k = s$, it can be seen by proceeding as in the case $k = 2$ the Laplacian matrix $L(G^{s*})$ of the graph G^{s*} is

$$L(G^{s*}) = D(G^{s*}) - A(G^{s*}) = \begin{pmatrix} D(G^{(s-1)*}) + I_{2^{s-1}n} & -(A(G^{(s-1)*}) + I_{2^{s-1}n}) \\ -(A(G^{(s-1)*}) + I_{2^{s-1}n}) & D(G^{(s-1)*}) + I_{2^{s-1}n} \end{pmatrix}.$$

Therefore, the Laplacian characteristic polynomial of G^{s*} is

$$\begin{aligned}
C_{G^{s*}}(\lambda) &= |\lambda I_{2^s n} - L(G^{s*})| \\
&= \begin{vmatrix} (\lambda-1)I_{2^{s-1}n} - D(G^{(s-1)*}) & A(G^{(s-1)*}) + I_{2^{s-1}n} \\ A(G^{(s-1)*}) + I_{2^{s-1}n} & (\lambda-1)I_{2^{s-1}n} - D(G^{(s-1)*}) \end{vmatrix} \\
&= \left|((\lambda-1)I_{2^{s-1}n} - D(G^{(s-1)*})) - (A(G^{(s-1)*}) + I_{2^{s-1}n})\right| \times \\
&\quad \left|((\lambda-1)I_{2^{s-1}n} - D(G^{(s-1)*})) + (A(G^{(s-1)*}) + I_{2^{s-1}n})\right| \\
&= \left|(\lambda-2)I_{2^{s-1}n} - (D(G^{(s-1)*}) + A(G^{(s-1)*}))\right| \\
&\quad \times \left|\lambda I_{2^{s-1}n} - (D(G^{(s-1)*}) - A(G^{(s-1)*}))\right| \\
&= Q_{G^{(s-1)*}}(\lambda-2) C_{G^{(s-1)*}}(\lambda).
\end{aligned}$$

Therefore, it follows that the L-spectra of the graph G^{s*} is $\mu_i \left(\binom{s-1}{0} \text{ times}\right)$, $\mu_i + 2\left(\binom{s-2}{1} \text{ times}\right)$, $\mu_i^+ + 2\left(\binom{s-2}{0} \text{ times}\right)$, …, $\mu_i + 2(s-2)\left(\binom{s-2}{s-2} \text{ times}\right)$, $\mu_i^+ + 2(s-2)\left(\binom{s-2}{s-3} \text{ times}\right)$, $\mu_i^+ + 2(s-1)\left(\binom{s-1}{s-1} \text{ times}\right)$, $\mu_i + 2\left(\binom{s-1}{0} \text{ times}\right)$, $\mu_i + 4(\binom{s-2}{1} \text{ times})$, $\mu_i^+ + 4(\binom{s-2}{0} \text{ times})$, …, $\mu_i + 2(s-1)(\binom{s-2}{s-2} \text{ times})$, $\mu_i^+ + 2(s-1)(\binom{s-2}{s-3} \text{ times})$, $\mu_i^+ + 2s(\binom{s-1}{s-1} \text{ times})$.

Using $\binom{k}{r} + \binom{k}{r-1} = \binom{k+1}{r}$, $0 \le r \le k$ and $\binom{s-1}{0} = \binom{s}{0} = \binom{s-1}{s-1} = \binom{s-2}{s-2} = 1$, we see that the L-spectra of G^{s*} is $\mu_i(\binom{s}{0} \text{ times})$, $\mu_i + 2(\binom{s-1}{1} \text{ times})$, $\mu_i^+ + 2(\binom{s-1}{0} \text{ times})$, $\mu_i + 4(\binom{s-1}{2} \text{ times})$, $\mu_i^+ + 4(\binom{s-1}{1} \text{ times})$, …, $\mu_i + 2(s-2)(\binom{s-1}{s-2} \text{ times})$, $\mu_i^+ + 2(s-2)(\binom{s-1}{s-3} \text{ times})$, $\mu_i + 2(s-1)(\binom{s-1}{s-1} \text{ times})$, $\mu_i^+ + 2(s-1)(\binom{s-1}{s-2} \text{ times})$, $\mu_i^+ + 2s(\binom{s}{s} \text{ times})$. Thus the result is true in this case as well hence by induction the result follows. □

If G is a bipartite graph, it is easy to see that under elementary transformation the Laplacian characteristic polynomial of G coincides with the signless Laplacian characteristic polynomial of G. Therefore the Laplacian and signless Laplacian spectra of G are same. We have the following observation.

Corollary 13 *If* $G(n, m)$ *is a bipartite graph having* L*-spectra* μ_i, $1 \le i \le n$, *then the* L*-spectra of* k*-th iterated double cover* G^{k*} *of* G *is* $\mu_i \left(\binom{k}{0} \textit{ times}\right)$, $\mu_i + 2\left(\binom{k}{1} \textit{ times}\right)$, …, $\mu_i + 2(k-2)\left(\binom{k}{k-2} \textit{ times}\right)$, $\mu_i + 2(k-1)\left(\binom{k}{k-1} \textit{ times}\right)$, $\mu_i + 2k\left(\binom{k}{k} \textit{ times}\right)$, *where* $1 \le i \le n$.

Proof. Since for a bipartite graph G the Laplacian and the signless Laplacian spectra are same, we have $\mu_i = \mu_i^+$ for all $1 \leq i \leq n$. Using this in Theorem 21, we obtain the L-spectra of G^{k*} as $\mu_i \left(\binom{k}{0} \text{ times}\right)$, $\mu_i + 2\left(\binom{k-1}{1} \text{ times}\right)$, $\mu_i + 2\left(\binom{k-1}{0} \text{ times}\right)$, $\mu_i + 4\left(\binom{k-1}{2} \text{ times}\right)$, $\mu_i + 4\left(\binom{k-1}{1} \text{ times}\right)$, $\ldots$, $\mu_i + 2(k-2)\left(\binom{k-1}{k-2} \text{ times}\right)$, $\mu_i + 2(k-2)\left(\binom{k-1}{k-3} \text{ times}\right)$, $\mu_i + 2(k-1)\left(\binom{k-1}{k-1} \text{ times}\right)$, $\mu_i + 2(k-1)\left(\binom{k-1}{k-2} \text{ times}\right)$, $\mu_i + 2k\left(\binom{k}{k} \text{ times}\right)$. Now using the fact $\binom{t}{r} + \binom{t}{r-1} = \binom{t+1}{r}$, $0 \leq r \leq t$, the result follows. □

In [6] three formulae are given for the number of spanning trees of G^* in terms of A-spectra of the corresponding graph G. We now obtain a formula for the number of spanning trees in terms of the L and Q-spectra of G^*.

Theorem 14 *The number of spanning trees* $\tau(G^*)$ *of the graph* G^* *is*

$$\tau(G^*) = \frac{1}{2}\tau(G)\prod_{i=1}^{n}(\mu_i^+ + 2).$$

Proof. Let $0 = \mu_n < \mu_{n-1} \leq \cdots \leq \mu_1$ and $0 < \mu_n^+ < \mu_{n-1}^+ \leq \cdots \leq \mu_1^+$ be respectively the L-spectra and the Q-spectra of the graph G. By Theorem 3.2, the L-spectra of the graph G^* is $\mu_i, \mu_i^+ + 2$ for $i = 1, 2, \ldots, n$. By using the fact that the number of spanning trees of a graph of order n is $\frac{1}{n}$ times the product of $(n-1)$ largest Laplacian eigenvalues of the graph, we have

$$\tau(G^*) = \frac{1}{2n}\prod_{i=1}^{n-1}\mu_i\prod_{i=1}^{n}(\mu_i^+ + 2) = \frac{1}{2}\tau(G)\prod_{i=1}^{n}(\mu_i^+ + 2).$$

□

In case G is bipartite, $\mu_i = \mu_i^+$, so we have

$$\tau(G^*) = \frac{1}{2n}\prod_{i=1}^{n-1}\mu_i\prod_{i=1}^{n}(\mu_i + 2) = \tau(G)\prod_{i=1}^{n-1}(\mu_i + 2).$$

In [6] it is shown that the graphs G^* and $G \times K_2$ are A-cospectral if and only if $G = K_1$ or G is bipartite. An analogous result holds for the L-spectra and is given below.

Theorem 15 *The graphs* G^* *and* $G \times K_2$ *are* L*-cospectral if and only if* $G = K_1$ *or* G *is bipartite.*

Proof. If $G = K_1$, the graphs G^* and $G \times K_2$ are both isomorphic to K_1, so are L-cospectral. Now if $G \neq K_1$, assume that G is bipartite. Then $\mu_i = \mu_i^+$ and so the L-spectra of G^* is $\mu_i, \mu_i + 2$ for $1 \leq i \leq n$ which is same as the L-spectra of $G \times K_2$. Conversely, suppose that the graphs G^* and $G \times K_2$ are L-cospectral. Then $\mu_i = \mu_i^+$, which is only possible if G is bipartite. Hence the result. □

An *integral graph* is a graph all of whose eigenvalues are integers. Following observation is a consequence of Theorem 12.

Theorem 16 *A graph G is Laplacian integral if and only if the graph G^{k*} is Laplacian integral graph.*

It is clear from Theorem 16, that given a Laplacian integral G it is always possible to construct an infinite sequence of Laplacian integral graphs. Indeed the graph G^{k*} is Laplacian integral for all $k \geq 1$.

Two graphs G_1 and G_2 are said to be co-spectral, if they are non-isomorphic and have the same spectra. We have the following result, which follows by Theorem 12.

Theorem 17 *Two graphs G_1 and G_2 are Laplacian cospectral if and only if the graphs G_1^{k*} and G_2^{k*} are Laplacian cospectral.*

Thus given two Laplacian co-spectral graphs G_1 and G_2, it is always possible to construct an infinite sequence of Laplacian co-spectral graphs. Indeed the graphs G_1^{k*} and Since the extended double cover G^* of the graph G is always bipartite, it follows by Theorem 6, the graphs G^{**} and $G^* \times K_2$ are L-cospectral and in general the graphs G^{s*} and $G^{(s-1)*} \times K_2$ are L-cospectral. Also it is easy to see that the graphs $(G \times K_2)^*$ and $G^* \times K_2$ are L-cospectral and in general the graphs $(G \times K_2)^{s*}$ and $G^{s*} \times K_2$ are both L-cospectral as well as Q-cospectral. Moreover, if G is bipartite then as seen in Theorem 6, the graphs G^* and $G \times K_2$ are L-cospectral. Using the same argument it can be seen that the graphs G^{**} and $G \times K_2 \times K_2$ are L-cospectral if and only if G is bipartite. A repeated use of the argument as used in Theorem 6, gives the graphs G^{s*} and $G \times K_2 \times K_2 \times \cdots \times K_2 = G \times sK_2 = G \times Q_s$ (where K_2 is repeated s times) are L-cospectral if and only if G is bipartite. From this discussion it follows that the graphs G^{s*}, $G^{(s-1)*} \times K_2$, $(G \times K_2)^{(s-1)*}$ and $G \times Q_{s-1}$ are mutually non-isomorphic L-cospectral graphs if and only G is bipartite, where Q_n is the hypercube.

4 Laplacian energy of double graphs

In this section, we study the Laplacian energy of the graphs $D[G]$, $D^k[G]$ and G^*. Using these graphs we obtain some new families of non Laplacian cospectral L-equienergetic graphs. Let $D[G]$ and G^* be respectively the double graph and the extended double cover of the graph G. Then the Laplacian spectra of the graph G^* is given by Lemma 2, and the Laplacian spectra of $D^k[G]$ is given by the following result.

Theorem 18 (Marino, Salvi, 2007 [21]) *Let G be a graph with n vertices having degrees $d_1, d_2, \ldots, d_n$ and let $\mu_1, \mu_2, \ldots, \mu_n$ be its Laplacian spectra. Then the Laplacian spectra of $D^k[G]$ is $k\mu_i, kd_i$ ($(k-1)n$ times) for $1 \leq i \leq n$.*

Let μ_i for $1 \leq i \leq n$ be the L-spectra of the graph G. Then by Theorem 11, the L-spectra of the extended double cover G^* of the graph G is $\mu_i, \mu_i^+ + 2$ for $1 \leq i \leq n$. Also the average vertex degree of G^* is $\frac{2m}{n} + 1$. Therefore,

$$LE(G^*) = \sum_{i=1}^{n} |\mu_i - \frac{2m}{n} - 1| + \sum_{i=1}^{n} |\mu_i^+ - \frac{2m}{n} + 1|.$$

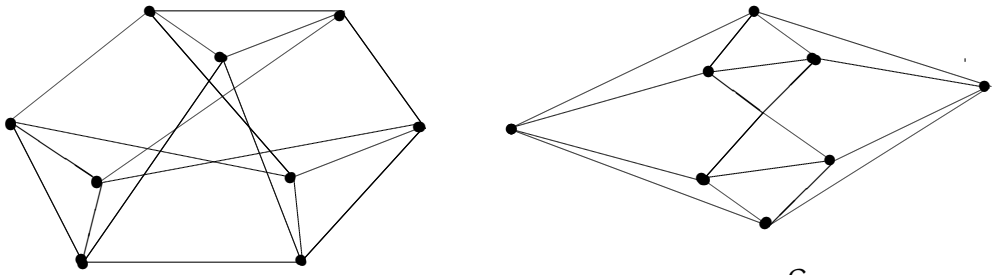
Since average vertex degree of $D^k[G]$ is $k\frac{2m}{n}$, we have

$$\begin{aligned} LE(D^k[G]) &= \sum_{i=1}^{n} \left| k\mu_i - k\frac{2m}{n} \right| + (k-1) \sum_{i=1}^{n} \left| kd_i - k\frac{2m}{n} \right| \\ &= k \sum_{i=1}^{n} \left| \mu_i - \frac{2m}{n} \right| + k(k-1) \sum_{i=1}^{n} \left| d_i - \frac{2m}{n} \right| \\ &= kLE(G) + k(k-1) \sum_{i=1}^{n} \left| d_i - \frac{2m}{n} \right|. \end{aligned}$$

From this it is clear that $LE(D^k[G]) = kLE(G)$, if G is regular. Also, since the k-fold graph of a regular graph is regular, it follows, if G_1 and G_2 are r-regular L-equienergetic graphs then their k-fold graphs $D^k[G_1]$ and $D^k[G_2]$ are also L-equienergetic. Let $\pounds(G)$ be the line graph of the graph G. It is shown in [23] that if G_1 and G_2 are r-regular graphs then their k-th ($k \geq 2$) iterated line graphs $\pounds^k(G_1)$ and $\pounds^k(G_2)$ are always equienergetic and so L-equienergetic. Therefore it follows that given any two r-regular graphs, we can always construct an infinite family of L-equienergetic graphs.

In case the given r-regular connected graphs are L-equienergetic, the k-fold graph forms a larger family of L-equienergetic graphs than the k-th iterated

line graph. As an example, consider the 4-regular graphs G_1 and G_2 shown in Figure 1 on 9-vertices. It can be seen that the L-spectra of G_1 and G_2 are respectively as $0, 3^4, 6^4$ and $0, 2, 3^2, 5^2, 6^3$ (where a^s means a occurs s times in the spectrum). Therefore $LE(G_1) = 16 = LE(G_2)$. This shows that the graphs G_1 and G_2 are regular L-equienergetic graphs, so their k-fold graphs $D^k[G_1]$ and $D^k[G_2]$ and their k-th ($k \geq 2$) iterated line graphs are also L-equienergetic. In fact the k-fold graph gives an infinite family of L-equienergetic graph pairs of order $n \equiv 0 \pmod 9$, whereas the k-th iterated line graph gives an infinite family of L-equienergetic graph pairs of orders $n = 542\,702\,430$, and so on, from this the assertion follows.



G_1 G_2

Figure 1

We have seen that the Laplacian energy of the graph $D[G]$ is twice the Laplacian energy of G when G is regular. But this need not be true for the graph G^* as seen from the Laplacian energy of G^* given above. However we have the following observation.

Theorem 19 *Let G^* be the extended double cover of the bipartite graph G. Then $LE(G^*) = 2LE(G)$ if and only if $\left|\mu_i - \frac{2m}{n}\right| \geq 1$ for $1 \leq i \leq n$.*

Proof. Let μ_i for $1 \leq i \leq n$ be the L-spectra of the graph G. Then by Corollary 13, the L-spectra of G^* is $\mu_i,\ \ \mu_i + 2$ for $1 \leq i \leq n$. Assume that $\left|\mu_i - \frac{2m}{n}\right| \geq 1$, for all $i = 1, 2, \ldots, n$. Then since average vertex degree of G^* is $\frac{2m}{n} + 1$, we have

$$\left|\mu_i - \frac{2m}{n} + 1\right| = \begin{cases} \left|\mu_i - \frac{2m}{n}\right| + 1, & \text{if } \mu_i \geq \frac{2m}{n} \\ |\mu_i - \frac{2m}{n}| - 1, & \text{if } \mu_i < \frac{2m}{n}, \end{cases}$$

$$\left|\mu_i - \frac{2m}{n} - 1\right| = \begin{cases} \left|\mu_i - \frac{2m}{n}\right| - 1, & \text{if } \mu_i \geq \frac{2m}{n} \\ \left|\mu_i - \frac{2m}{n}\right| + 1, & \text{if } \mu_i < \frac{2m}{n}. \end{cases}$$

Therefore,

$$\begin{aligned}
&LE(G^*)\\
&= \sum_{i=1}^{n} \left|\mu_i - \frac{2m}{n} - 1\right| + \sum_{i=1}^{n} \left|\mu_i - \frac{2m}{n} + 1\right|\\
&= \sum_{i=1}^{n} \left(\left|\mu_i - \frac{2m}{n} - 1\right| + \left|\mu_i - \frac{2m}{n} + 1\right|\right)\\
&= \sum_{\mu_i \geq \frac{2m}{n}} \left(\left|\mu_i - \frac{2m}{n} - 1\right| + \left|\mu_i - \frac{2m}{n} + 1\right|\right)\\
&+ \sum_{\mu_i < \frac{2m}{n}} \left(\left|\mu_i - \frac{2m}{n} - 1\right| + \left|\mu_i - \frac{2m}{n} + 1\right|\right)\\
&= \sum_{\mu_i \geq \frac{2m}{n}} \left(\left|\mu_i - \frac{2m}{n}\right| - 1 + \left|\mu_i - \frac{2m}{n}\right| + 1\right)\\
&+ \sum_{\mu_i < \frac{2m}{n}} \left(\left|\mu_i - \frac{2m}{n}\right| + 1 + \left|\mu_i - \frac{2m}{n}\right| - 1\right)\\
&= 2 \sum_{\mu_i \geq \frac{2m}{n}} \left|\mu_i - \frac{2m}{n}\right| + 2 \sum_{\mu_i < \frac{2m}{n}} \left|\mu_i - \frac{2m}{n}\right| = 2LE(G).
\end{aligned}$$

Conversely, suppose that $LE(G^*) = 2LE(G)$. We will show that $\left|\mu_i - \frac{2m}{n}\right| \geq 1$ for all $1 \leq i \leq n$. We prove this by contradiction. Assume that $\left|\mu_i - \frac{2m}{n}\right| < 1$, for some λ_j. Putting $\beta_i = \mu_i - \frac{2m}{n}$, and using the same argument as used in the converse of Theorem 8 in [4] we arrive at a contradiction. □

If G is a graph satisfying the conditions of Theorem 19, then clearly the graphs G^* and $G \cup G$ are L-equienergetic. We now obtain some new families of L-equienergetic graphs by means of the graphs G^*, G^{k*}, $D[G]$ and $D^k[G]$.

Theorem 20 *Let $G_1(n, m)$ be a graph having L-spectra and Q-spectra respectively as μ_i and μ_i^+ and let $G_2(n, m)$ be another graph having L-spectra and Q-spectra respectively as λ_i and λ_i^+ for $i = 1, 2, \ldots, n$. Then for $p \geq 2n + k$ and $m \leq \frac{(k-1)n}{2} + \frac{k^2}{4}$, $k \geq 3$, we have $LE(G_1^* \vee \bar{K_p}) = LE(G_2^* \vee \bar{K_p})$.*

Proof. Let G_1^* be the extended double cover of the graph G_1. Then by Theorem 11, the L-spectra of G_1^* is $\mu_i,\ \mu_i^+ + 2$ for $1 \leq i \leq n$ and so by Lemma 3, the L-spectra of $G_1^* \vee \bar{K_p}$ is $p+2n,\ p+\mu_i (1 \leq i \leq n-1), p+\mu_i^+ +2 (1 \leq i \leq n), 2n$

($(p-1)$ times), 0, with average vertex degree

$$\frac{2m'}{n'} = \frac{4m+4pn+2n}{p+2n}.$$

Therefore,

$$LE(G_1^* \vee \bar{K_p}) = \left|p+2n-\frac{2m'}{n'}\right| + \sum_{i=1}^{n-1}\left|p+\mu_i-\frac{2m'}{n'}\right| + \left|0-\frac{2m'}{n'}\right| + \sum_{i=1}^{n}\left|p+\mu_i^+ +2-\frac{2m'}{n'}\right| + (p-1)\left|2n-\frac{2m'}{n'}\right|.$$

Now, if $p \geq 2n+k$ and $m \leq \frac{(k-1)n}{2} + \frac{k^2}{4}$, $k \geq 3$, we have for $i = 1, 2, \ldots, n$,

$$p+\mu_i-\frac{2m'}{n'} = p+\mu_i-\frac{4m+4pn+2n}{p+2n} = \frac{p(p-2n)+(2n+p)\mu_i-4m-2n}{p+2n}$$

$$\geq \frac{k(2n+k)-2(k-1)n-k^2-2n}{p+2n} = 0,$$

and

$$p+\mu_i^+ +2-\frac{2m'}{n'} = p+\mu_i^+ +2-\frac{4m+4pn+2n}{p+2n}$$

$$= \frac{p(p-2n)+(2n+p)\mu_i^+ +2(p+n)-4m}{p+2n}$$

$$\geq \frac{k(2n+k)-2(k-1)n-k^2+2(3n+k)}{p+2n} = \frac{8n+2k}{p+2n} \geq 0.$$

So we have

$$LE(G_1^* \vee \bar{K_p}) = \left(p+2n-\frac{2m'}{n'}\right) + (n-1)\left(p-\frac{2m'}{n'}\right) + n\left(p+2-\frac{2m'}{n'}\right) + (p-1)\left(\frac{2m'}{n'}-2n\right) + \frac{2m'}{n'} + 4m = 6n + (p-2n)\frac{2m'}{n'} + 4m.$$

From this it is clear that the Laplacian energy of G_1^* depends only on the parameters p, m and n. Since these parameters are also same for G_2^*, it follows that $LE(G_1^* \vee \bar{K_p}) = LE(G_2^* \vee \bar{K_p})$. In fact all the graphs of the family $(G_i^* \vee \bar{K_p})$, $i = 1, 2, \ldots$, having the same parameters n, p and m satisfying the conditions in the hypothesis are mutually L-equienergetic. □

Let G^{t*} be the t-th iterated extended double cover of the graph G. We have the following generalization of Theorem 20.

Theorem 21 *Let* $G(n,m)$ *be a graph having* L*-spectra and* Q*-spectra respectively as* μ_i *and* μ_i^+ *for* $1 \le i \le n$. *For* $p \ge 2^t n + k$ *and* $m \le \frac{(k-t)n}{2} + \frac{k^2}{2^{t+1}}$, $k \ge t+2, t \ge 1$, *we have* $LE(G^{t*} \vee \bar{K_p}) = 2^t n(t+2) + (p - 2^t n)\frac{2m'}{n'} + 2^t(2m)$.

Proof. Let G^{t*} be the t-th iterated extended double cover of the graph G. Then by Theorem 12, the L-spectra of G^{t*} is μ_i $\left(\binom{t}{0}\text{ times}\right)$, $\mu_i + 2$ $\left(\binom{t-1}{1}\text{ times}\right)$, $\mu_i^+ + 2$ $\left(\binom{t-1}{0}\text{ times}\right)$, $\mu_i + 4$ $\left(\binom{t-1}{2}\text{ times}\right)$, $\mu_i^+ + 4$ $\left(\binom{t-1}{1}\text{ times}\right)$, $\ldots$, $\mu_i + 2(t-2)$ $\left(\binom{t-1}{t-2}\text{ times}\right)$, $\mu_i^+ + 2(t-2)$ $\left(\binom{t-1}{t-3}\text{ times}\right)$, $\mu_i + 2(t-1)$ $\left(\binom{t-1}{t-1}\text{ times}\right)$, $\mu_i^+ + 2(t-1)$ $\left(\binom{t-1}{t-2}\text{ times}\right)$, $\mu_i^+ + 2t$ $\left(\binom{t}{t}\text{ times}\right)$, where $1 \le i \le n$. So by Lemma 2.3, the L-spectra of $G^{t*} \vee \bar{K_p}$ is $0, p + 2^t n, 2^t n$ $(p-1$ times$)$, $p + \mu_i$ $\left(\binom{t}{0}\right.$ times$)$ $(1 \le i \le n-1)$, $p + \mu_i + 2$ $\left(\binom{t-1}{1}\right.$ times$)$, $p + \mu_i^+ + 2$ $\left(\binom{t-1}{0}\right.$ times$)$, $p + \mu_i + 4$ $\left(\binom{t-1}{2}\right.$ times$)$, $p + \mu_i^+ + 4$ $\left(\binom{t-1}{1}\right.$ times$)$, $\ldots$, $p + \mu_i + 2(t-2)$ $\left(\binom{t-1}{t-2}\right.$ times$)$, $p + \mu_i^+ + 2(t-2)$ $\left(\binom{t-1}{t-3}\right.$ times$)$, $p + \mu_i + 2(t-1)$ $\left(\binom{t-1}{t-1}\right.$ times$)$, $p + \mu_i^+ + 2(t-1)$ $\left(\binom{t-1}{t-2}\right.$ times$)$, $p + \mu_i^+ + 2t\left(\binom{t}{t}\right.$ times$)$, $1 \le i \le n$, with average vertex degree

$$\frac{2m'}{n'} = \frac{2^{t+1}m + 2^t tn + 2^{t+1}pn}{p + 2^t n}.$$

Therefore,

$$\begin{aligned}
&LE(G^{t*} \vee \bar{K_p}) \\
&= \sum_{i=1}^{n-1} \left| p + \mu_i - \frac{2m'}{n'} \right| + \sum_{r=1}^{t-1} \sum_{i=1}^{n} \binom{t-1}{r} \left| p + \mu_i + 2r - \frac{2m'}{n'} \right| \\
&+ \sum_{r=1}^{t-1} \sum_{i=1}^{n} \binom{t-1}{r-1} \left| p + \mu_i^+ + 2r - \frac{2m'}{n'} \right| + \sum_{i=1}^{n} \left| p + \mu_i + 2t - \frac{2m'}{n'} \right| \\
&+ |p + 2^t n - \frac{2m'}{n'}| + (p-1)|2^t n - \frac{2m'}{n'}| + \left| 0 - \frac{2m'}{n'} \right|.
\end{aligned}$$

Now, if $p \ge 2^t n + k$ and $m \le \frac{(k-t)n}{2} + \frac{k^2}{2^{t+1}}$, $k \ge t+2, t \ge 1$, we have for $i = 1, 2, \ldots, n$ and $r = 0, 1, \ldots, t$

$$p + \mu_i + 2r - \frac{2m'}{n'} = p + \mu_i + 2r - \frac{2^{t+1}m + 2^t tn + 2^{t+1}pn}{p + 2^t n}$$

$$= \frac{p(p - 2^t n) + 2r(p + 2^t n) + (p + 2^t n)\mu_i - 2^{t+1}m - 2^t tn}{p + 2^t n}$$

$$\geq \frac{k(2^t n + k) - k(2^t n + k) + 2^t tn - 2^t tn}{p + 2^t n} = 0.$$

Similarly, it can be seen that $p + \mu_i^+ + 2r - \frac{2m'}{n'} \geq 0$. So we have

$$\begin{aligned}
&LE(G^{t*} \vee \bar{K}_p) \\
&= (n-1)\left(p - \frac{2m'}{n'}\right) + \sum_{r=1}^{t-1}\left(n\left(p + 2r - \frac{2m'}{n'}\right) + 2m\right)\left[\binom{t-1}{r} + \binom{t-1}{r-1}\right] \\
&+ \left(p + 2^t n - \frac{2m'}{n'}\right) + (p-1)\left(\frac{2m'}{n'} - 2^t n\right) \\
&+ \left(n\left(p + 2t - \frac{2m'}{n'}\right) + 2m\right) + \frac{2m'}{n'} + 2m \\
&= 2^{t+1}n - pn(2^t - 1) + (p-n)\frac{2m'}{n'} \\
&+ \sum_{r=1}^{t}\binom{t}{r}\left(n\left(p + 2r - \frac{2m'}{n'}\right) + 2m\right) + 2m \\
&= 2^{t+1}n - pn(2^t - 1) + (p-n)\frac{2m'}{n'} + n(2^t - 1)\left(p - \frac{2m'}{n'}\right) \\
&+ (2^t - 1)2m + 2^t tn + 2m \\
&= 2^t n(t+2) + (p - 2^t n)\frac{2m'}{n'} + 2^t(2m),
\end{aligned}$$

where we have made use of the fact $\left[\binom{t-1}{r} + \binom{t-1}{r-1}\right] = \binom{t}{r}$ and $\sum_{r=1}^{t} r\binom{t}{r} = t2^{t-1}$.

Clearly the Laplacian energy of the graph $(G^{t*} \vee \bar{K}_p)$ depends only on the parameters p, m, t and n. Therefore all the graphs of the families $(G_i^{t*} \vee \bar{K}_p)$, where $t, i = 1, 2, \ldots$, with the same parameters p, m, t and n satisfying the conditions in the hypothesis are mutually L-equienergetic. □

Theorem 21 gives an infinite family of L-equienergetic graphs in various ways, firstly fix the value of t and allow p to vary we obtain families of L-equienergetic graphs with same t, secondly fix the value of p and allow t to vary we obtain families of L-equienergetic graphs with same p and so on.

Corollary 22 *Let* $G(n, m)$ *be a bipartite graph having* L*-spectra* μ_i *for* $1 \leq i \leq n$. *For* $p \geq 2^t n + k$ *and* $m \leq \frac{(k-t)n}{2} + \frac{k^2}{2^{t+1}}$, $k \geq t + 2, t \geq 1$, *we have* $LE(G^{t*} \vee \bar{K}_p) = 2^t n(t+2) + (p - 2^t n)\frac{2m'}{n'} + 2^t(2m)$.

Proof. The proof follows the proof of Corollary 13 and the same argument as in the proof Theorem 21. □

From Theorem 21, it is clear if G_1 and G_2 are any two graphs with the same parameters, then we can always find tripartite graphs $(G_1^* \vee \bar{K_p})$ and $(G_2^* \vee \bar{K_p})$ having the same Laplacian energy. Next we show the construction of L-equienergetic graphs by means of graphs $D[G]$ and $D^k[G]$.

Theorem 23 *Let* $D[G]$ *be the double graph of the graph* G*. Then, for* $p \geq 2n + k$ *and* $m \leq \frac{k(2n+k)}{8}$, $k \geq 4$, *we have*

$$LE(D[G] \vee \bar{K_p}) = 4n + (p - 2n)\frac{2m'}{n'} + 8m.$$

Proof. Let μ_i and d_i for $i = 1, 2, \ldots, n$ be respectively the L-spectra and the degree sequence of the graph G. Then by Theorem 21, the L-spectra of the graph $D^k[G]$ is $k\mu_i,\ kd_i$ ($(k-1)n$ times) and so by Lemma 3, the L-spectra of the graph $D^k[G] \vee \bar{K_p}$ is $p + kn,\ p + k\mu_i$ $(1 \leq i \leq n-1)$, $p + kd_i$ ($(k-1)n$ times) $(1 \leq i \leq n)$, kn $((p-1)$ times), 0, with average vertex degree

$$\frac{2m'}{n'} = \frac{2k^2m + 2pkn}{p + kn}.$$

So, if $p \geq kn + t$ and $m \leq \frac{t(kn+t)}{2k^2}$, $t \geq 2k, k \geq 2$, we have for $i = 1, 2, \ldots, n$

$$\begin{aligned} p + k\mu_i - \frac{2m'}{n'} &= p + k\mu_i - \frac{2k^2m + 2pkn}{p + kn} \\ &= \frac{p(p - kn) - 2k^2m + k(p + kn)\mu_i}{p + kn} \\ &\geq \frac{t(kn + t) - t(kn + t)}{p + kn} = 0. \end{aligned}$$

Similarly, we see that

$$p + 2d_i - \frac{2m'}{n'} \geq 0.$$

Therefore,

$$\begin{aligned} LE(D[G] \vee \bar{K_p}) &= \left|p + 2\mu_i - \frac{2m'}{n'}\right| + \sum_{i=1}^{n-1}\left|p + 2\mu_i - \frac{2m'}{n'}\right| + \sum_{i=1}^{n}\left|p + 2d_i - \frac{2m'}{n'}\right| \\ &\quad + (p - 1\left|2n - \frac{2m'}{n'}\right| + \left|0 - \frac{2m'}{n'}\right| \\ &= 4n + (p - 2n)\frac{2m'}{n'} + 8m. \end{aligned}$$

Clearly the Laplacian energy of the graph $D[G] \vee \bar{K}_p$ depends only on the parameters p, m and n. Therefore all the graphs of the family $(D[G_i] \vee \bar{K}_p)$, $i = 1, 2, \ldots$ with the same parameters p, m and n satisfying the conditions of the theorem, are mutually L-equienergetic. □

If $D^k[G]$ is the k-fold graph of the graph G, we have the following generalization of Theorem 23.

Theorem 24 *Let* $D^k[G]$ *be the* k*-fold graph of the graph* G*. Then for* $p \geq kn + t$ *and* $m \leq \frac{t(kn+t)}{2k^2}$, $t \geq 2k, k \geq 2$, *we have* $LE(D^k[G] \vee \bar{K}_p) = 2kn + (p - nk)\frac{2m'}{n'} + 2mk^2$.

Proof. Let μ_i and d_i for $i = 1, 2, \ldots, n$ be respectively the L-spectra and the degree sequence of the graph G. Then by Theorem 21, the L-spectra of the graph $D^k[G]$ is $k\mu_i$, kd_i ($(k-1)n$ times) and so by Lemma 3, the L-spectra of the graph $D^k[G] \vee \bar{K}_p$ is $p + kn$, $p + k\mu_i$ $(1 \leq i \leq n-1)$, $p + kd_i$ ($(k-1)n$ times) $(1 \leq i \leq n)$, kn ($(p-1)$ times), 0, with average vertex degree

$$\frac{2m'}{n'} = \frac{2k^2m + 2pkn}{p + kn}.$$

So, if $p \geq kn + t$ and $m \leq \frac{t(kn+t)}{2k^2}$, $t \geq 2k, k \geq 2$, we have for $i = 1, 2, \ldots, n$

$$p + k\mu_i - \frac{2m'}{n'} = p + k\mu_i - \frac{2k^2m + 2pkn}{p + kn}$$

$$= \frac{p(p - kn) - 2k^2m + k(p + kn)\mu_i}{p + kn} \geq \frac{t(kn + t) - t(kn + t)}{p + kn} = 0.$$

Similarly, we see that

$$p + kd_i - \frac{2k^2m + 2pkn}{p + kn} \geq 0.$$

Therefore,

$$LE(D^k[G] \vee \bar{K}_p)$$
$$= \left|p + kn - \frac{2m'}{n'}\right| + \sum_{i=1}^{n-1} \left|p + k\mu_i - \frac{2m'}{n'}\right| + (k-1) \sum_{i=1}^{n} \left|p + 2d_i - \frac{2m'}{n'}\right|$$
$$+ (p-1)\left|kn - \frac{2m'}{n'}\right| + \left|0 - \frac{2m'}{n'}\right| = 2kn + 2mk^2 + (p - nk)\frac{2m'}{n'}.$$

From this it is clear the Laplacian energy of the graph $(D^k[G] \vee \bar{K_p})$ depends on the parameters p, k, m and n. Therefore all the graphs of the families $(D^k[G_i] \vee \bar{K_p})$ where $i = 1, 2, \ldots$, and $k = 2, 3, \ldots$ having the same parameters p, m, k and n satisfying the conditions of the Theorem, are mutually L-equienergetic. □

Theorem 24 generates families of L-equienergetic graphs in various ways. If we allow p to vary and keep k fixed, we obtain an infinite family of L-equienergetic graphs with same k and if we allow k to vary and keep p fixed, we obtain an infinite family of L-equienergetic graphs with same p and so on.

If $D[G]$ and G^* are respectively the double graph and the extended double cover of the graph G, then the following result gives the construction of L-equienergetic graphs with different number of edges.

Theorem 25 *Let* $G_1(n, m_1)$ *and* $G_2(n, m_2)$ *be two graphs of order* $n \equiv 0 \pmod 4$ *with* $m_2 = m_1 + \frac{n}{4}$. *Then for* $p \geq 4n + k$ *and* $m_2 \leq \frac{n(k-2)}{4} + \frac{k^2}{16}$, $k \geq 4$, *we have*

$$LE(D(G_1^*) \vee \bar{K_p}) = LE(D(G_2)^* \vee \bar{K_p}).$$

Proof. Let μ_i, d_i and μ_i^+ for $i = 1, 2, \ldots, n$ be respectively the L-spectra, degree sequence and Q-spectra of G_1 and let λ_i, d_i' and λ_i^+ be the L-spectra, degree sequence and Q-spectra of the graph G_2. Then by Theorems 11 and 18 and Lemma 3, the L-spectra of the graphs $D(G_1^*) \vee \bar{K_p}$ and $D(G_2)^* \vee \bar{K_p}$ are respectively as $p + 4n$, $p + 2\mu_i$ $(1 \leq i \leq n-1)$, $p + 2\mu_i^+ + 4, p + 2d_i + 2$ (2 times) $(1 \leq i \leq n)$, $4n$ $((p-1)$ times), 0 and $p + 4n$, $p + 2\lambda_i$ $(1 \leq i \leq n-1)$, $p + 2\lambda_i^+ + 4$, $p + 2d_i' + 2$ (2 times) $(1 \leq i \leq n)$, $4n((p-1)$ times), 0, with average vertex degrees

$$\frac{2m_1'}{n'} = \frac{16m_1 + 8n + 8pn}{p + 4n}, \quad \frac{2m_2'}{n'} = \frac{16m_2 + 8n + 8pn}{p + 4n}.$$

Now, if $p \geq 4n + k$ and $m_2 \leq \frac{n(k-2)}{4} + \frac{k^2}{16}$, $k \geq 4$, we have for $i = 1, 2, \ldots, n$

$$p + 2\mu_i - \frac{2m_1'}{n'} = p + 2\mu_i - \frac{16m_1 + 8n + 8pn}{p + 4n}$$

$$= \frac{p(p - 4n) + 2(p + 4n)\mu_i - 16m_1 - 8n - 8pn}{p + 4n}$$

$$\geq \frac{k(4n + k) - 4n(k - 2) - k^2 - 8n}{p + 4n} = 0.$$

Similarly, we can show that

$$p + 2\mu_i^+ + 4 - \frac{2m_1'}{n'} \geq 0, p + 2d_i + 2 - \frac{2m_1'}{n'} \geq 0.$$

Therefore,

$$\begin{aligned} LE(D(G_1^*) \vee \bar{K_p}) &= \left| p + 4n - \frac{2m_1'}{n'} \right| + \sum_{i=1}^{n-1} \left| p + 2\mu_i - \frac{2m_1'}{n'} \right| \\ &+ \sum_{i=1}^{n} \left| p + 2\mu_i^+ + 4 - \frac{2m_1'}{n'} \right| \\ &+ 2\sum_{i=1}^{n} \left| p + 2d_i + 2 - \frac{2m_1'}{n'} \right| \\ &+ (p-1) \left| 4n - \frac{2m_1'}{n'} \right| + \left| 0 - \frac{2m_1'}{n'} \right| \\ &= 16n + 16m_1 + (p - 4n)\frac{2m_1'}{n'}. \end{aligned}$$

Proceeding similarly for the graph $D(G_2)^* \vee \bar{K_p}$ it can be seen that

$$LE(D(G_2)^* \vee \bar{K_p}) = 12n + 16m_2 + (p - 4n)\frac{2m_2'}{n'}.$$

Using the fact $m_2 = m_1 + \frac{n}{4}$, the result follows. □

Let $D[G_1]$ be the double graph of the graph $G_1(n, m_1)$ and let G_2^* be the extended double cover of the graph $G_2(n, m_2)$, then for $p \geq 2n + k$ and $m_1 \leq \frac{k(2n+k)}{8}$, $k \geq 4$, we have from Theorem 23

$$LE(D[G_1] \vee \bar{K_p}) = 4n + 8m_1 + (p - 2n)\frac{2m_1'}{n'}. \tag{3}$$

Also, for $p \geq 2n + k$ and $m_2 \leq \frac{n(k-1)}{2} + \frac{k^2}{4}$, $k \geq 4$, we have by Theorem 20

$$LE(G_2^* \vee \bar{K_p}) = 6n + 4m_2 + (p - 2n)\frac{2m_2'}{n'}. \tag{4}$$

If we suppose that $4m_1 = 2m_2 + n$, then it follows from (3) and 4 that

$$LE(D[G_1] \vee \bar{K_p}) = LE(G_2^* \vee \bar{K_p}).$$

This gives another construction of families of graphs with same Laplacian energy, same number of vertices but different number of edges. Next we give another way of constructing a family of graphs having same number of vertices, same Laplacian energy but different number of edges.

Theorem 26 *Let* $G_1(n, m_1)$ *and* $G_2(n, m_2)$ *be two graphs with* $m_2 = 2m_1$. *Then for* $p \geq 4n + k$ *and* $m_2 \leq \frac{k(4n+k)}{8} - n$, $k \geq 4$, *we have* $LE(D(G_1^*) \vee \bar{K_p}) = LE(G_2^{**} \vee \bar{K_p})$.

Proof. Let μ_i, μ_i^+ and d_i for $i = 1, 2, \ldots, n$ be respectively the L-spectra, Q-spectra and the degree sequence of the graph G_1 and let λ_i and λ_i^+ be the L-spectra and Q-spectra of the graph G_2. Then by Theorems 11 and 18 and Lemma 3, the L-spectra of $D(G_1^*) \vee \bar{K_p}$ is $p + 4n,\ p + 2\mu_i\ \ (1 \leq i \leq n-1),\ p + 2\mu_i^+ + 4,\ p + 2d_i + 2$ (2 times) $(1 \leq i \leq n),\ 4n$ $((p-1)$ times), 0. Also by Theorem 12 and Lemma 3, the L-spectra of the graph $G_2^{**} \vee \bar{K_p}$ is $p + 4n,\ p + \lambda_i\ \ (1 \leq i \leq n-1),\ p + \lambda_i + 2,\ p + \lambda_i^+ + 2,\ p + \lambda_i^+ + 4\ \ (1 \leq i \leq n),\ 4n((p-1)$ times), 0, with average vertex degrees

$$\frac{2m_1'}{n'} = \frac{16m_1 + 8n + 8pn}{p + 4n}, \frac{2m_2'}{n'} = \frac{8m_2 + 8n + 8pn}{p + 4n}.$$

So, if $p \geq 4n + k$ and $m_2 \leq \frac{k(4n+k)}{8} - n$, $k \geq 4$, we have for $i = 1, 2, \ldots, n$

$$p + 2\mu_i - \frac{2m_1'}{n'} = p + 2\mu_i - \frac{16m_1 + 8n + 8pn}{p + 4n}$$

$$= \frac{p(p - 4n) + 2(p + 4n)\mu_i - 16m_1 - 8n - 8pn}{p + 4n}$$

$$= \frac{k(4n + k) - k(4n + k) + 8n - 8n}{p + 4n} = 0.$$

Similarly, we can show

$$p + 2\mu_i^+ + 4 - \frac{2m_1'}{n'} \geq 0, p + 2d_i + 2 - \frac{2m_1'}{n'} \geq 0.$$

Therefore,

$$LE(D(G_1^*) \vee \bar{K_p})$$

$$= |p + 4n - \frac{2m_1'}{n'}| + \sum_{i=1}^{n-1} |p + 2\mu_i - \frac{2m_1'}{n'}| + \sum_{i=1}^{n} |p + 2\mu_i^+ + 4 - \frac{2m_1'}{n'}|$$

$$+2\sum_{i=1}^{n} |p + 2d_i + 2 - \frac{2m_1'}{n'}| + (p-1)|4n - \frac{2m_1'}{n'}| + |0 - \frac{2m_1'}{n'}|$$

$$= 16n + (p - 4n)\frac{2m_1'}{n'} + 16m_1.$$

Proceeding similarly as above for the graph $G_2^{**} \vee \bar{K}_p$, we can see that

$$LE(G_2^{**} \vee \bar{K}_p) = 16n + (p-4n)\frac{2m_2'}{n'} + 8m_2.$$

Using $m_2 = 2m_1$, the result follows. □

Theorem 26 generates L-equienergetic graphs with same number of vertices but different number of edges, infact when one graph contains twice the number of edges as contained in other. Lastly we give the construction of family of graphs with same number of vertices, edges and Laplacian energy by means of Cartesian product and extended double cover.

Theorem 27 *Let* $G_1(n,m)$ *and* $G_2(n,m)$ *be two connected non-bipartite graphs. Then for* $p \geq n+2$, *and* $\min(\mu_n^+, \lambda_n^+) \geq \frac{2m}{n} - 2$ *we have* $LE(G_1^* \times K_p) = LE(G_2^* \times K_p)$ *if and only if* $LE(G_1) = LE(G_2)$.

Proof. Let $0 = \mu_n < \mu_{n-1} \leq \cdots \leq \mu_1$ and $0 < \mu_n^+ < \mu_{n-1}^+ \leq \cdots \leq \mu_1^+$ be respectively the L-spectra and the Q-spectra of the graph G_1 and let $0 = \lambda_n < \lambda_{n-1} \leq \cdots \leq \lambda_1$ and $0 < \lambda_n^+ < \lambda_{n-1}^+ \leq \cdots \leq \lambda_1^+$ be respectively the L-spectra and Q-spectra of the graph G_2. Then by Theorem 11 and Lemma 1, the L-spectra of the graphs $G_1^* \times K_p$ and $G_1^* \times K_p$ are respectively as $\gamma_i + q_j$ and $\theta_i + q_j$, $i = 1, 2, \ldots, 2n$, $\quad j = 1, 2, \ldots, n$, where

$$\gamma_i = \begin{cases} \mu_i, & \text{if } i = 1, 2, \ldots, n \\ \mu_i^+ + 2, & \text{if } i = n+1, n+2, \ldots, 2n, \end{cases}$$

$$\theta_i = \begin{cases} \lambda_i, & \text{if } i = 1, 2, \ldots, n \\ \lambda_i^+ + 2, & \text{if } i = n+1, n+2, \ldots, 2n \end{cases}$$

and $p = q_1 = q_2 = \cdots = q_{p-1}$, $\quad q_p = 0$ with average vertex degree

$$\frac{2m'}{n'} = \frac{2m}{n} + p.$$

Therefore,

$$\begin{aligned} LE(G_1^* \times K_p) &= \sum_{i=1}^{2n} \sum_{j=1}^{n} \left| \gamma_i + q_j - \frac{2m'}{n'} \right| \\ &= (p-1) \sum_{i=1}^{2n} \left| p + \gamma_i - \frac{2m'}{n'} \right| + \sum_{i=1}^{n} \left| \gamma_i - \frac{2m_1'}{n'} \right| \\ &= (p-1)LE(G_1) + 4pn - 4n. \end{aligned}$$

Similarly it can be seen that

$$LE(G_2^* \times K_p) = (p-1)LE(G_2) + 4pn - 4n.$$

It is now clear that $LE(G_1^* \times K_p) = LE(G_2^* \times K_p)$ if and only if $LE(G_1) = LE(G_2)$, therefore the result follows. □

Since G^* is always bipartite, Theorem 27 gives the construction of connected graphs from a given pair of L-equienergetic bipartite graphs having same number of vertices, edges and Laplacian energy. Moreover if t is the first value of p satisfying the conditions in Theorem 27, then every value greater than t also satisfies this condition, therefore we obtain an infinite family of L-equienergetic graph pairs.

References

[1] N. Abreu, D. M. Cardoso, I. Gutman, E. A. Martins, M. Robbiano, Bounds for the signless Laplacian energy, *Linear Algebra Appl.* **435,** 10 (2011) 2365–2374. ⇒90

[2] T. Aleksić, Upper bounds for Laplacian energy of graphs, *MATCH Commun. Math. Comput. Chem.* **60,** 2 (2008) 435–439. ⇒90

[3] R. Balakrishnan, The energy of a graph, *Linear Algebra Appl.* **387** (2004) 287–295. ⇒91

[4] A. S. Bonifacio, C. T. M. Vinagre, N. M. Abreu, Constructing pairs of equienergetic and non-cospectral graphs, *Applied Mathematics Letters* **21,** 4 (2008) 338–341. ⇒97, 105

[5] J. Carmona, I. Gutman, N. J. Tamblay, M. Robbiano, A decreasing sequence of upper bounds for the Laplacian energy of a tree, *Linear Algebra Appl.* **446** (2014) 304–313. ⇒90

[6] Z. Chen, Spectra of extended double cover graphs, *Czechoslovak Math. J.* **54,** 4 (2004) 1077–1082. ⇒91, 92, 97, 101

[7] D. Cvetkovic, M. Doob, H. Sachs, *Spectra of Graphs: Theory and Application*, Academic Press, New York, 1980. ⇒92, 93

[8] D. Cvetkovic, S. K. Simic, Towards a spectral theory of graphs based on signless Laplacian I, *Publ. Inst. Math* (Beograd), **85** (2009) 19–33. ⇒90

[9] K. Ch. Das, I. Gutman, On incidence energy of graphs, *Linear Algebra Appl.* **446** (2014) 329–344. ⇒90

[10] G. H. Fath-Tabar, A. R. Ashrafi, Some remarks on the Laplacian eigenvalues and Laplacian energy of graphs, *Math. Commun.* **15** (2010) 443–451. ⇒90

[11] M. Fiedler, Algebraic connectivity of graphs, *Czechoslovak Math. J.* **23,** 2 (1973) 298–305. ⇒90

[12] I. Gutman, The energy of a graph, *Ber. Math. Statist. Sekt. Forschungszenturm Graz* **103** (1978) 1–22. ⇒90

[13] I. Gutman, On graphs whose energy exceeds the number of vertices, *Linear Algebra Appl.* **429** (2008) 2670–2677. ⇒90
[14] I. Gutman, D. Kiarib, M. Mirzakhahb, On incidence energy of graphs, *MATCH Commun. Math. Comput. Chem.* **62,** 3 (2009) 573–580. ⇒90
[15] I. Gutman, D. Kiarib, M. Mirzakhahb, B. Zhoud, On incidence energy of a graphs, *Linear Algebra Appl.* **431,** 8 (2009) 1223–1233. ⇒90
[16] I. Gutman, O. E. Polansky, *Mathematical Concepts in Organic Chemistry*, Springer Verlag, Berlin 1986. ⇒90
[17] I. Gutman, B. Zhou, Laplacian energy of a graph, *Linear Algebra Appl.* **414** (2006) 29–37. ⇒90
[18] M. R. Jooyandeh, D. Kiani, M. Mirzakhah, Incidence energy of a graph, *MATCH Commun. Math. Comput. Chem.* **62,** 3 (2009) 561–572. ⇒90
[19] X. Li, Y. Shi, I. Gutman, *Graph Energy*, Springer, New York, 2012. ⇒90
[20] J. Liu, B. Liu, E–L equienergetic graphs, *MATCH Commun. Math. Comput. Chem.* **66,** 3 (2011) 971–976. ⇒90
[21] M. S. Marino, N. Z. Salvi, Generalizing double graphs, *Atti dell' Accademia Peloritana dei pericolanti classe di scienze Fisiche, Matematiche e Naturali*, Vol. 85 CIA 0702002 (2007), pp. 1–9. ⇒91, 93, 103
[22] S. Radenković, I. Gutman, Total π-electron energy and Laplacian energy: How far the analog goes?, *J. Serb. Chem. Soc.* **72,** 12 (2007) 1343–1350. ⇒90
[23] H. S. Ramane, H. B. Walikar, S. B. Rao, B. D. Acharya, P. R. Hampiholi, S. R. Jog, I. Gutman, Equienergetic graphs, *Kragujevac J. Math.* **26** (2004) 5–13. ⇒91, 103
[24] L. Shi, H. Wang, The Laplacian incidence energy of graphs, *Linear Algebra Appl.* **439,** 12 (2013) 4056–4062. ⇒90
[25] M. P. Stanić, I. Gutman, On almost equienergetic graphs, *MATCH Commun. Math. Comput. Chem.* **70,** 2 (2013) 681–688. ⇒90
[26] Z. Tang, Y. Hou, On incidence energy of trees, *MATCH Commun. Math. Comput. Chem.* **66,** 3 (2011) 977–984. ⇒90
[27] H. Wang, H. Hua, Note on Laplacian energy of graphs, *MATCH Commun. Math. Comput. Chem.* **59,** 2 (2008) 373–380. ⇒90
[28] F. Zhang, *Matrix Theory, Basic Results and Techniques*, Springer Verlag, Berlin, 1999. ⇒97
[29] B. Zhou, More on energy and Laplacian energy, *MATCH Commun. Math. Comput. Chem.* **64,** 1 (2010) 75–84. ⇒90
[30] B. Zhou, I. Gutman, On Laplacian energy of graphs, *MATCH Commun. Math. Comput. Chem.* **57,** 1 (2007) 211–220. ⇒90

Received: February 3, 2014 • Revised: April 13, 2014

Acta Univ. Sapientiae, Informatica, **6,** 1 (2014) 117–131

Coloring the nodes of a directed graph

Sándor SZABÓ
University of Pécs
Institute of Mathematics and Informatics, Ifjúság u. 6
7624 Pécs, HUNGARY
email: sszabo7@hotmail.com

Bogdán ZAVÁLNIJ
University of Pécs
Institute of Mathematics and Informatics, Ifjúság u. 6
7624 Pécs, HUNGARY
email: bogdan@ttk.pte.hu

Abstract. It is an empirical fact that coloring the nodes of a graph can be used to speed up clique search algorithms. In directed graphs transitive subtournaments can play the role of cliques. In order to speed up algorithms to locate large transitive tournaments we propose a scheme for coloring the nodes of a directed graph. The main result of the paper is that in practically interesting situations determining the optimal number of colors in the proposed coloring is an NP-hard problem. A possible conclusion to draw from this result is that for practical transitive tournament search algorithms we have to develop approximate greedy coloring algorithms.

1 Introduction

Let $G = (V, E)$ be a finite simple graph, that is, G has finitely many nodes and G does not have any loop or double edge. A subgraph D is a clique in G if each two distinct nodes of D are connected in G. If the clique D has k nodes, then we say that D is a k-clique in G. The number of nodes of a clique sometimes referred as the size of the clique. A k-clique in G is a maximum clique if G does not have any $(k+1)$-clique. The graph G may have several maximum

Computing Classification System 1998: G.2.2
Mathematics Subject Classification 2010: 05C5, 03D15
Key words and phrases: simple directed graph, tournament, transitive tournament, maximum transitive tournament, graph coloring, coloring the nodes, NP-hard problems
DOI:10.2478/ausi-2014-0021

cliques but their common size is a well defined number. This number is called the *clique number* of the graph G and it is denoted by $\omega(G)$. The problem of determining the clique number of a given graph is an important problem in many areas of applied discrete mathematics. (For a list of applications see [1].) It is known that the problem is NP-complete. (For proofs see [5] or [12].) The most commonly used clique search algorithms employ coloring of the nodes of a graph to speed up the computations. (See [2, 8, 9, 11, 14].) The coloring of the nodes of the graph G with k colors assigns exactly one color to each node of the graph such that adjacent nodes never receive the same color. This type of coloring of the nodes sometimes referred as legal or well coloring of the nodes. The minimum number of colors with which the nodes of G can be legally colored is a well defined number. It is called the *chromatic number* of G and it is denoted by $\chi(G)$. Determining the chromatic number of a given graph is another important problem in the applied discrete mathematics with many applications. It is known that the problem of deciding if a given graph can be colored with k color is NP-complete for any fixed k, where $k \geq 3$. (See [5] or [12].) The problem for $k = 2$ belongs to the P (polynomial) complexity class.

Let $G = (V, E)$ be a finite simple directed graph. This means that G has finitely many nodes and G does not have any loop or double edges directed in parallel manner. In the particular case when there is exactly one directed edge between any two distinct nodes, then G is called a *tournament*. Tournaments can look back to a venerable history. (See [13, 3, 4, 10].) The directed graph G is transitive if $(x, y), (y, z) \in E$ implies $(x, z) \in E$. Motivated by applications in information theory [7] introduced the problem of determining the size of a maximum subtournament in a given finite simple directed graph. Since coloring proved to be advantageous in improving the efficiency of clique search algorithms we address the problem if coloring can be exploited in the algorithms locating maximum transitive tournaments in a given graph. We propose a type of colorings of the nodes of a finite simple directed graph. We will show that this coloring leads to an NP-hard problem. A practical implication of the result is that for coloring we should rely on approximate greedy coloring procedures instead of trying to compute the optimal number of colors.

Colorings are employed in at least two ways in clique search algorithms. One can color the nodes of a graph G before the clique search starts. In these cases the coloring is used as a possible preprocessing or preconditioning tool. On the other hand if in the course of the clique search algorithm one recolors the nodes of the subgraphs of G under consideration, then we call it an on-line coloring.

It must be ample clear that the requirements for preconditioning or for on-line coloring algorithms are not necessarily the same. In the case of preconditioning the graph coloring is a well separated phase of the computation. We may afford to use more time and memory space. While in the case of on-line coloring we have to trade speed for the quality of coloring.

By the main result of the paper determining the optimal number of colors is not a practically recommended option. In the same time there is a real need to introduce, implement and test various greedy coloring algorithms. We hope that our paper will stimulate this activity.

2 Coloring the nodes

Let T be a tournament whose nodes are $a_1, \ldots, a_k$, where $k \geq 2$.

Lemma 1 *If T is a transitive tournament, then there is a permutation $b_1, \ldots, b_k$ of $a_1, \ldots, a_k$ such that*

(1) $(b_i, b_{i+1}), \ldots, (b_i, b_k)$ are edges of T for each i, $1 \leq i < k$.

(2) The listed $n(n-1)/2$ edges are all the edges of T.

(3) Each subgraph of T is a transitive tournament.

Proof. Statement (1) clearly holds for $k = 2$. We assume that $k \geq 3$ and start an induction on k.

If T has a vertex, say a_1, such that each edge incident to a_1 goes out of a_1, then a_1 can be identified with b_1 and the inductive assumption is applicable to the graph whose nodes are $a_2, \ldots, a_k$.

If T has a vertex, say a_k such that each edge incident to a_k goes into a_k, then a_k can be identified with b_k and the inductive assumption is applicable to the graph whose nodes are $a_1, \ldots, a_{k-1}$.

For the remaining part of the proof of statement (1) we may assume that each vertex of T has an incident edge going in and has an incident edge going out. In this case T contains a directed cycle. On the other hand a transitive tournament cannot contain a directed cycle.

The reason why statement (2) holds is that a transitive tournament has $n(n-1)/2$ directed edges and we listed all of them in statement (1).

Statement (3) follows from the definition of the transitive tournament and statement (1), as any subset of the nodes can be ordered the same way. □

A finite simple directed graph $G = (V, E)$ can be represented by an $|V|$ by $|V|$ adjacency matrix. The rows and the columns are labeled by the nodes of G. If the ordered pair (u, v) is an edge of G, then we put a bullet into the cell at the intersection of row u and column v. The adjacency matrix of course has 0, 1 entries when stored in a computer. The practice of using bullets instead of 1's is taken from [6]. It seems that it has a good visual effect and the computations are less prone to clerical errors when carried out using paper and pencil.

By Lemma 1, a tournament T is a transitive tournament if and only if the rows and the columns of its adjacency matrix can be permuted such that the adjacency matrix became an upper triangular matrix. A simple directed graph H with r nodes has a transitive tournament of r nodes if the adjacency matrix of H can be rearranged such that the upper triangular part is filled with bullets.

Let $G = (V, E)$ be a finite simple directed graph. Let U be a subset of V and let s be an integer such that $U \neq \emptyset$ and $s \geq 3$. The subset U of V is called an s-free subset if U does not contain any transitive tournament with s nodes. A partition of V into the subsets $V_1, \ldots, V_k$ is called an s-free partition of V if V_i is an s-free set for each i, $1 \leq i \leq k$.

A coloring of the nodes of a finite simple directed graph $G = (V, E)$ can be described by means of an onto function $f : V \to \{1, \ldots, k\}$. Here the numbers $1, \ldots, k$ are used as colors and node v receives color $f(v)$. The c-level set V_c of f is defined to be $V_c = \{v : f(v) = c, v \in V\}$.

The coloring $f : V \to \{1, \ldots, k\}$ of the nodes of the finite simple directed graph $G = (V, E)$ is called an s-free coloring if the level sets $V_1, \ldots, V_k$ form an s-free partition of V. The name intends to express the fact that color classes cannot contain any transitive tournament of size s. In other words color classes are free of tournaments of size s.

In the $s = 2$ special case a color class of an s-free coloring cannot contain any edge. ¿From this reason we will mainly deal with the $s \geq 3$ case.

The number of the color classes of an s-free coloring of the graph G can be used to establish an upper estimate of the size of a maximum transitive tournament in G.

Lemma 2 *If the finite simple directed graph G admits an s-free coloring with k colors and G has a transitive tournament of size r, then $r \leq k(s-1)$.*

Proof. Suppose G has a transitive tournament T of size r. Let $V_1, \ldots, V_k$ be the color classes of the s-free coloring and let W be the set of nodes of T. By Lemma 1, the subgraph of T with set of nodes $V_i \cap W$ is a transitive

tournament. Since the coloring is s-free, it follows that $|V_i \cap W| \leq s-1$. We get that

$$r = |W| = |V_1 \cap W| + \cdots + |V_k \cap W| \leq k(s-1),$$

as required. □

From Lemma 2 we can see that the smaller is k, the better is the upper estimate of the size of the maximum transitive tournament in G. The following problem comes to mind naturally.

Problem 3 *Given a finite simple directed graph* $G = (V, E)$*. Further given integers* r*,* s *such that* $r \geq 3$*,* $s \geq 3$*. Decide if* G *has an* s*-free coloring with* r *colors.*

When we deal with coloring the nodes of a graph G we inevitably have to deal with incomplete or partial colorings, where each of the nodes of G receives at most one of the colors $1, \ldots, r$ but some of the nodes of G are left uncolored. Allocating color 0 for the uncolored nodes we can incorporate the incomplete colorings into the family of complete colorings.

3 Two auxiliary graphs

In this section we describe two finite simple directed graphs. They will play the roles of building blocks or switching devices in further constructions. Let r, s be fixed integers such that $r \geq 3$, $s \geq 3$. Set

$$h = (s-1)(r-1) + (s-2) + 2.$$

Let us consider the directed simple graph $H = (V, E)$, where $V = \{1, \ldots, h\}$. Set $W = \{2, \ldots, h-1\}$. We draw directed edges between the nodes in W such that the subgraph of H whose set of nodes is W forms a transitive tournament. From the node 1 we direct edges towards each node of W. Similarly, from the node h we direct edges towards each node of W.

For the sake of the illustration we worked out the special case $r = 3$, $s = 3$ in details. The adjacency matrix of H is in Table 1. A geometric representation of H is depicted in Figure 1.

We spell out the properties of the graph H we will use later as a Lemma.

	1	2	3	4	5	6	7
1		•	•	•	•	•	
2			•	•	•	•	
3				•	•	•	
4					•	•	
5						•	
6							
7		•	•	•	•	•	

Table 1: The adjacency matrix of the graph H in the special case $r = s = 3$.

Lemma 4 *(1) The nodes of the graph* H *have an* s*-free coloring with* r *colors.*

(2) In each s*-free coloring of the nodes of* H *with* r *colors the nodes* 1 *and* h *must receive the same colors.*

(3) Each partial coloring of the nodes of H*, where nodes* 1*,* h *receive the same color (and the remainig nodes of of* H *are left uncolored) can be extended to an* s*-free coloring of the nodes of* H *using* r *colors.*

Proof. In order to prove statement (1) let us consider the subsets $C_1, \dots, C_r$, C_{r+1} of V such that these subsets are pair-wise disjoint and

$$|C_1| = \cdots = |C_{r-1}| = s-1, \quad |C_r| = s-2, \quad C_{r+1} = \{1, h\}.$$

Set $W = \{2, \dots, h-1\}$. Clearly, $C_1, \dots, C_r$ form a partition of W. We use $C_1, \dots, C_r$ as color classes to define a coloring of the subgraph L of H whose set of nodes is W.

By Lemma 1, the subgraph of H whose set of nodes is C_i is a transitive tournament for each i, $1 \le i \le r$. As $|C_i| \le s-1$, it follows that this graph does not contain a transitive tournament with s nodes. Therefore the coloring of L is an s-free coloring. The subgraph of H whose set of nodes is $C_i \cup \{1\}$ is a transitive tournament with s nodes. Consequently, the node 1 cannot receive color i for each i, $1 \le i \le r-1$. On the other hand node 1 can receive color r since the subgraph of H whose set of nodes is $C_r \cup \{1, h\}$ is not an obstruction. Similarly node h may receive color r. This completes the proof of statement (1).

We can use the coloring constructed in the previous part and combine it with the fact that the colors in an s-free coloring of the nodes of H can be permuted among each other freely to settle statement (3).

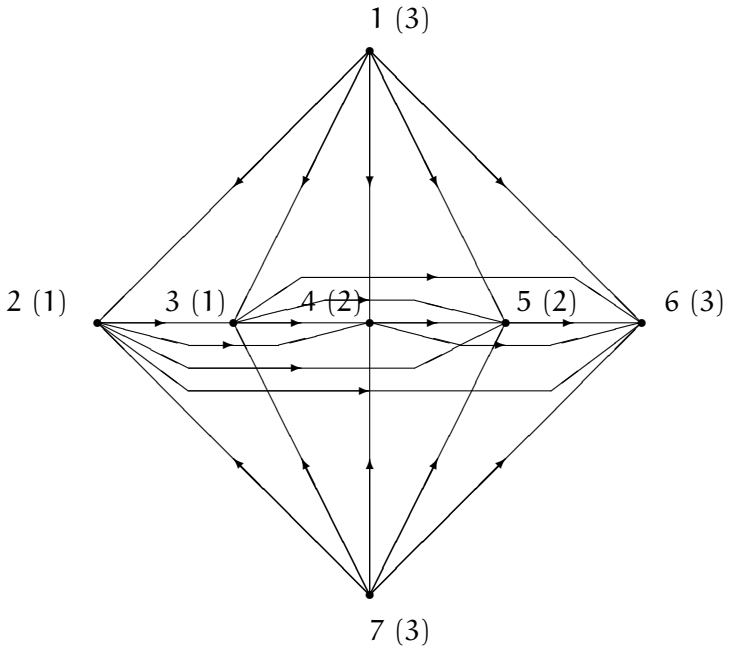


Figure 1: The auxiliary graph H in the special case $r = s = 3$. The numbers in parentheses are the colors of the nodes.

To prove statement (2) let us suppose that $f : V \to \{1, \dots, r\}$ is an s-free coloring of the nodes of H. Let us consider the subgraph L of H whose set of nodes is $W = \{2, \dots, h-1\}$. The restriction of f to W is an s-free coloring of the nodes of L. Let $C_1, \dots, C_r$ be the colors classes of this coloring. We may assume that $|C_1| \geq \cdots \geq |C_r|$ since this is only a matter of exchanging the colors $1, \dots, r$ among each other.

By Lemma 1, the subgraph of H whose set of nodes is C_i is a transitive tournament for each i, $1 \leq i \leq r$. It follows that $|C_i| \leq s-1$. Using

$$|C_1| + \cdots + |C_r| = (s-1)(r-1) + (s-2)$$

we get that $|C_1| = \cdots = |C_{r-1}| = s-1$ and $|C_r| = s-2$. The subgraph of H whose set of nodes is $C_i \cup \{1\}$ is a transitive tournament with s nodes for each i, $1 \leq i \leq r-1$. We get that the node 1 cannot receive color i and so node 1

must receive color r. A similar reasoning gives that node h must receive color r too. This completes the proof of statement (2). □

Let $r, s \geq 3$ be fixed integers. We construct a new auxiliary directed graph K. Let T be a transitive tournament with nodes $1, \dots, s$. We consider $s-1$ isomorphic copies $H_1, \dots, H_{s-1}$ of H. We choose the notation such that $H = (V, E)$ with $V = \{1, \dots, h\}$ and $H_i = (V_i, E_i)$ with $V_i = \{(1,i), \dots, (h,i)\}$. The correspondence

$$1 \longleftrightarrow (1,i), \dots, h \longleftrightarrow (h,i)$$

defines the isomorphism between H and H_i.

Let us consider the nodes $(1,1), \dots, (1,s-1)$ of $H_1, \dots, H_{s-1}$, respectively and solder these nodes together to form a node u of K. Let us consider the nodes $(h,1), \dots, (h,s-1)$ of $H_1, \dots, H_{s-1}$, respectively and solder these nodes together with the nodes $1, \dots, s-1$ of the tournament T, respectively. We rename node s of T to be v.

The graph K has

$$1 + \underbrace{(h-2) + \cdots + (h-2)}_{(s-1) \text{ times}} + s = 1 + (s-1)(h-2) + s$$

nodes. We set $k = 1 + (s-1)(h-2) + s$ and rename the nodes of K by the numbers $1, \dots, k$ such that $1 = u$ and $k = v$. We illustrated the construction in the special cases $s = 3$ and $s = 4$. The geometric versions of K can be seen in Figures 2 and 3.

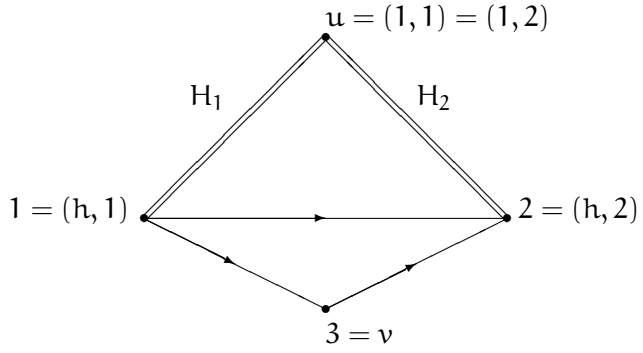


Figure 2: The auxiliary graph K in the special case $s = 3$. The double lines represent isomorphic copies of H.

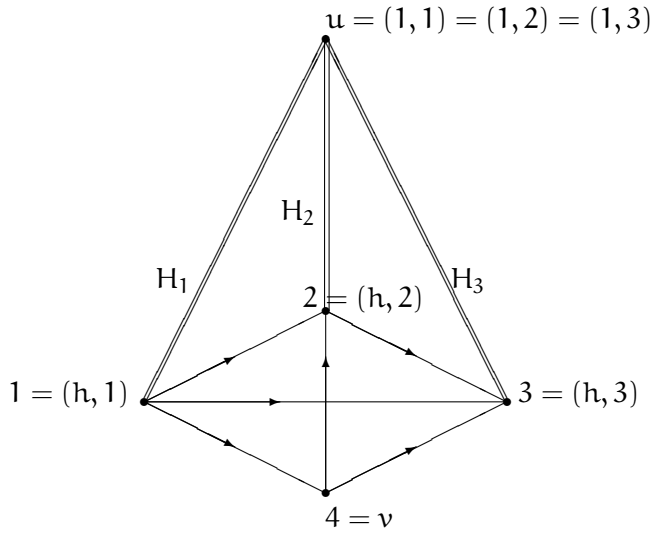


Figure 3: The auxiliary graph K in the special case $s = 4$. The double lines represent isomorphic copies of H.

The next lemma summarizes the essential properties of the graph K what we need later.

Lemma 5 *(1) The nodes of the graph* K *admit an* s*-free coloring with* r *colors.*

(2) In each s*-free coloring of the nodes of the graph* K *with* r *colors the nodes* 1 *and* k *cannot receive the same colors.*

(3) Each partial coloring of the nodes of K*, where the nodes* 1 *and* k *are colored with distinct colors (and the other nodes of* K *are left uncolored) can be extended to an* s*-free coloring of the nodes of* K *using* r *colors.*

Proof. By Lemma 4, the nodes of the graph H admit and s-free coloring with r colors. Consequently, the nodes of each of the graphs $H_1, \ldots, H_{s-1}$ admit an s-free coloring with r colors. These colorings provide the same fixed color for $s - 1$ nodes of the tournament T. The uncolored node v of T can be colored with any of the remaining $r - 1$ colors. This proves statement (1).

By Lemma 4, the node V cannot receive the same color as node u. This settles statement (2).

By Lemma 4, each partial coloring of the nodes of H can be extended to an s-free coloring of the nodes of H using r colors. It follows that each partial coloring of the nodes of H_i can be extended to an s-free coloring of the nodes of H_i using r colors for each i, $1 \leq i \leq s-1$. This provides a partial coloring of the nodes of the tournament T. In this partial coloring of the nodes of T each node except node v receives the same color. Namely the color of node u. The last uncolored node v clearly can be colored with any of the remaining $r-1$ colors. This proves statement (3). □

4 The main result

The main result of this paper is the following theorem

Theorem 6 *Problem 3 is NP-hard for each* $r, s \geq 3$.

Proof. Let $r, s \geq 3$ be fixed integers. Assume on the contrary that Problem 3 is not NP-hard, that is, there is an "efficient" (polynomial running time) algorithm that solves Problem 3. Let $G = (V, E)$ be a finite simple graph with undirected edges. Using G and the auxiliary graphs H, K described in the previous chapter we construct a finite simple directed graph $G' = (V', E')$ such that the following conditions hold.

(1) If the nodes of G' have an s-free coloring with r colors, then the nodes of G have a legal coloring with r colors.

(2) If the nodes of G have a legal coloring with r colors, then the nodes of G' have an s-free coloring with r colors.

(3) The number of nodes G' can be upper bounded by a polynomial of the number of the nodes of G.

Thus for each legal (edge free) coloring problem we can construct a directed s-free coloring problem. If the second can be solved in polynomial time, it means that the first one can be solved in polynomial time as well.

Let $v_1, \ldots, v_n$ be the edges of G. In other words let $V = \{v_1, \ldots, v_n\}$. We consider an isomorphic copy $K_{i,j} = (W_{i,j}, F_{i,j})$ of the auxiliary graph $K = (W, F)$ for each i, j, $1 \leq i < j \leq n$. We recall that the nodes of K are labeled by the numbers $1, \ldots, k$. The nodes of $K_{i,j}$ will be labeled by the ordered triples $(i, j, 1), \ldots, (i, j, k)$. Here the correspondence

$$1 \longleftrightarrow (i,j,1), \ldots, k \longleftrightarrow (i,j,k)$$

defines the isomorphism between K and $K_{i,j}$.

With each of the nodes $v_1, \ldots, v_n$ we associate a node $v'_1, \ldots, v'_n$ of the graph G'. At this moment our only concern is that $v'_1, \ldots, v'_n$ are pair-wise distinct points and they are nodes of G'. But G' may have further nodes.

If the unordered pair $\{v_i, v_j\}$ is an edge of G, then we add additional $k-2$ nodes to G'. We identify the nodes $(i,j,1)$, (i,j,k) of $K_{i,j}$ with the nodes v'_i, v'_j of G', respectively. Next, we add the remaining $k-2$ nodes of $K_{i,j}$ to the nodes of G'. Finally, we add all the edges of $K_{i,j}$ to the edges of G'.

If the unordered pair $\{v_i, v_j\}$ is not an edge of G, then we do not add any nodes and we do not add any edges to G'. Clearly, G' has directed edges and it has $|V|+|E|(k-2)$ nodes. Since r, s are fixed numbers, it follows that $k-2 = c$ is a constant and so $|V'|$ can be upper bounded by $n + cn(n-1)/2$ which is a second degree polynomial in terms of n. This observation shows that condition (3) is satisfied.

In order to show that condition (1) is satisfied let us assume that $f' : V' \to \{1, \ldots, r\}$ is an s-free coloring of the nodes of G'. Using f' we define a coloring $f : V \to \{1, \ldots, r\}$ of the nodes of G. We set $f(v_i)$ to be equal to $f'(v'_i)$.

We claim that $f(v_i) = f(v_j)$ implies that the unordered pair $\{v_i, v_j\}$ is not an edge of G.

To verify the claim we assume on the contrary that $f(v_i) = f(v_j)$ and $\{v_i, v_j\}$ is an edge of G. The restriction of f' to $W_{i,j}$ is an s-free coloring of the nodes of the graph $K_{i,j}$. By Lemma 5, the nodes $(i,j,1)$ and (i,j,k) cannot receive the same color. Using $v'_i = (i,j,1)$, $v'_j = (i,j,k)$ we get the

$$f(v_i) = f'(v'_i) \neq f'(v'_j) = f(v_j)$$

contradiction.

To demonstrate that condition (2) is satisfied let us suppose that $f : V \to \{1, \ldots, r\}$ is a legal coloring of the nodes of G. Using f we define a coloring $f' : V' \to \{1, \ldots, r\}$ of the nodes of G'. We set $f'(v'_i)$ to be equal to $f(v_i)$.

Let us consider two distinct nodes v'_i, v'_j of G'. If the unordered pair $\{v'_i, v'_j\}$ is an edge of G', then by the construction of G' the nodes v'_i, v'_j are identical with the nodes $(i,j,1)$, (i,j,k) of $K_{i,j}$, respectively. Thus $v'_i = (i,j,1)$, $v'_j = (i,j,k)$. Since f is a legal coloring of the nodes of G, it follows that $f(v_i) \neq f(v_j)$ and so by the definition of f', we get that $f'(v'_i) \neq f'(v'_j)$.

For the sake of definiteness let us suppose that $f'(v'_i) = 1$ and $f'(v'_j) = 2$. We have a partial coloring of the nodes of $K_{i,j}$. Namely, the nodes $(i,j,1)$, (i,j,k)

are colored with colors $1, 2$, respectively. Other nodes of $K_{i,j}$ are left uncolored. By Lemma 5, this partial coloring of $K_{i,j}$ can be extended to an s-free coloring of $K_{i,j}$. Since this can be accomplished in connection with each adjacent nodes v'_i, v'_j of G', it follows that the nodes of G' have an s-free coloring with r colors. □

5 A second proof of the main result

In the proof of Theorem 6 we used only the auxiliary graph K. The auxiliary graph H made an appearance only in the proof of Lemma 5 when we established the key properties of the auxiliary graph K. In this section we present an informal new proof where the graph H plays a more direct role. The node edge incidence matrix M of a finite simple graph $G = (V, E)$ is a $|V|$ by $|E|$ matrix. The rows and the columns of M are labeled by the nodes and the edges of G, respectively. If $e = \{u, v\}$ is an edge of G, then we place two bullets into M. We put one bullet into the cell at the intersection of row u and column e. Then we put a bullet into the cell at the intersection of row v and column e.

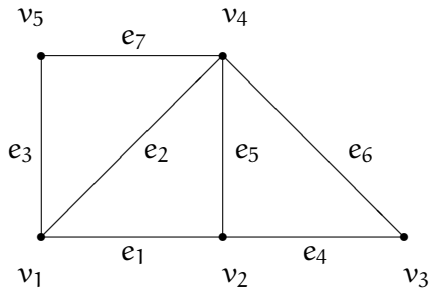


Figure 4: The toy example Γ.

	e_1	e_2	e_3	e_4	e_5	e_6	e_7
v_1	•	•	•				
v_2	•			•	•		
v_3				•		•	
v_4		•			•	•	•
v_5			•				•

Table 2: The node edge incidence matrix of the toy example Γ.

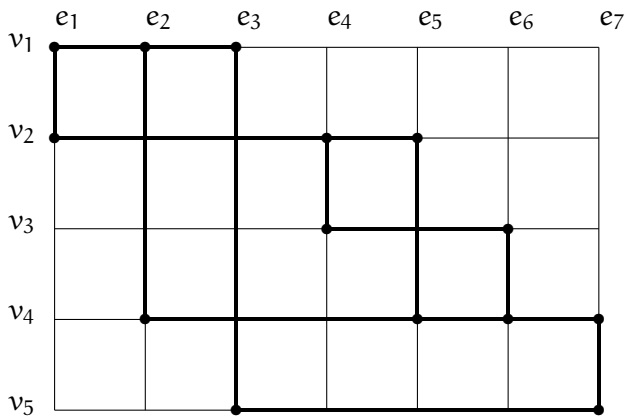


Figure 5: The graph Γ^* associated with the toy example Γ.

For the sake of illustration we included a toy example. The graph Γ can be seen in Figure 4. The node edge incidence matrix of Γ is in Table 2.

Suppose we are given a finite simple undirected graph $G = (V, E)$. Using G we construct a graph $G^* = (V^*, E^*)$. The construction is guided by the node edge incidence matrix of G. Let $V = \{v_1, \ldots, v_n\}$, $E = \{e_1, \ldots, e_m\}$. If $e_t = \{v_i, v_j\}$, then we add the ordered pairs (i, t), (j, t) to the set of nodes of G^*. Thus V^* is a set whose elements are ordered pairs. Clearly $|V^*| = 2|E| = 2m$.

We can form a mesh consisting of n horizontal and m vertical lines. The intersection of the horizontal and vertical lines form $(n)(m)$ mesh points. The nodes of G^* can be identified with some of these mesh points.

Two distinct nodes (i, t) and (j, t) of G^* on a vertical mesh line are connected with a vertical undirected edge in G^*. Two distinct nodes (i, x) and (i, z) of G^* on a horizontal line are connected with a horizontal undirected edge in G^* if there is no node in the form (i, y) such that $x < y < z$. Figure 5 depicts the graph Γ^* associated with the toy example Γ. The mesh lines are represented by thin lines. Bold lines represent the edges of Γ^*.

We replace each horizontal edge of G^* by an isomorphic copy of the auxiliary graph H. Next we replace each vertical edge of G^* by an isomorphic copy of the auxiliary graph K. After all possible replacements we get a finite simple directed graph G'.

Suppose that the nodes of G' have an s-free coloring with r colors. The isomorphic copies of the auxiliary graph H guarantee that the nodes of G' on

a fixed horizontal line all receive the same color. The isomorphic copies of the auxiliary graph K make sure that the two nodes of G' on a fixed vertical line receive distinct colors. In this way we get a legal coloring of the nodes of G with $\mathtt{r}$ colors.

Next suppose that the nodes of G have a legal coloring with $\mathtt{r}$ colors. This coloring will provide partial colorings of the nodes of the isomorphic copies of the graphs H and K. One can extend these partial colorings to a complete $\mathtt{s}$-free coloring of the nodes of G' with $\mathtt{r}$ colors.

Acknowledgements

The research of B. Zaválnij was realized in the frames of TÁMOP 4.2.4. A/2-11-1-2012-0001 "National Excellence Program - Elaborating and operating an inland student and researcher personal support system". The project was subsidized by the European Union and co-financed by the European Social Fund.

References

[1] I. M. Bomze, M. Budinich, P. M. Pardalos, M. Pelillo, The maximum clique problem, in *Handbook of Combinatorial Optimization* Vol. 4, Eds. D.-Z. Du and P. M. Pardalos, Kluwer Academic Publisher, Boston, MA 1999. ⇒118

[2] R. Carraghan, P. M. Pardalos, An exact algorithm for the maximum clique problem, *Operation Research Letters* **9,** 6 (1990), 375–382. ⇒118

[3] P. Erdős, L. Moser, On the representation of directed graphs as unions of orderings. *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **9** (1964), 125–132. ⇒118

[4] P. Erdős, J. W. Moon, On sets of consistent arcs in tournament, *Canad. Math. Bull.* **8** (1965), 269–271. ⇒118

[5] M. R. Garey, D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-completeness*, Freeman, New York, 2003. ⇒118

[6] F. Kárteszi, *Introduction to Finite Geometries*, North-Holland Pub. Co., Amsterdam, 1976. ⇒120

[7] L. Kiviluoto, P. R. J. Östergård, V. P. Vaskelainen, Exact algorithms for finding maximum transitive subtournaments, (manuscript) ⇒118

[8] J. Konc, D. Janežič, An improved branch and bound algorithm for the maximum clique problem, *MATCH Commun. Math. Comput. Chem.* **58,** 3 (2007), 569–590. ⇒118

[9] D. Kumlander, *Some practical algorithms to solve the maximal clique problem*, PhD Thesis, Tallin University of Technology, 2005. ⇒118

[10] J. W. Moon, *Topics on Tournaments*, Holt, Reinhart and Winston, New York, 1968. ⇒118

[11] P. R. J. Östergård, A fast algorithm for the maximum clique problem, *Discrete Appl. Math.* **120,** 1-3 (2002), 197–207. ⇒118

[12] C. H. Papadimitriou, *Computational Complexity*, Addison-Wesley Publishing Company, Inc., Reading, MA, 1994. ⇒118

[13] R. Stearns, The voting problem, *Amer. Math. Monthly* **66** (1959), 761–763. ⇒118

[14] E. Tomita, T. Seki, An efficient branch-and-bound algorithm for finding a maximum clique, Discrete Mathematics and Theoretical Computer Science, *Lecture Notes in Comp. Sci.* **2731** (2003), 278–289. ⇒118

Received: August 25 , 2013 • Revised: April 10, 2014

Acta Univ. Sapientiae, Informatica, **6,** 1 (2014) 132–158

On vertex independence number of uniform hypergraphs

Tariq A. CHISHTI
University of Kashmir
Department of Mathematics
Srinagar, India
email: chishtita@yahoo.co.in

Guofei ZHOU
Nanjing University
Department of Mathematics
Nanjing, China
email: gfzhou@mail.nju.edu.cn

Shariefuddin PIRZADA
University of Kashmir
Department of Mathematics
Srinagar, India
email:
pirzadasd@kashmiruniversity.ac.in

Antal IVÁNYI
Eötvös Loránd University
Faculty of Informatics
Budapest, Hungary
email: tony@inf.elte.hu

Abstract. Let H be an r-uniform hypergraph with $r \geq 2$ and let $\alpha(H)$ be its vertex independence number. In the paper bounds of $\alpha(H)$ are given for different uniform hypergraphs: if H has no isolated vertex, then in terms of the degrees, and for triangle-free linear H in terms of the order and average degree.

1 Introduction to independence in graphs

Let n be a positive integer. A *graph* G on vertex set $V = \{v_1, v_2, \ldots, v_n\}$ is a pair (V, E), where the edge set E is a subset of $V \times V$. n is the *order* of G and $|E|$ is the *size* of G.

Computing Classification System 1998: G.2.2
Mathematics Subject Classification 2010: 05C30, 05C50
Key words and phrases: uniform hypergraph, independence number, lower bound
DOI:10.2478/ausi-2014-0022

Let $v \in V$ and $N(v)$ be the *neighborhood* of v, namely, the set of vertices x so that there is an edge which contains both v and x. Let U be a subset of V, then the *subgraph* of G induced by U is defined as a graph on vertex set U and edge set $E_U = \{(u, v) | u \in u \text{ and } v \in U\}$.

The *degree* $d(v)$ of a vertex $v \in V$ is the number of edges that contains v. Let $d(G)$ be the *average degree* of G, then $nd(G) = \sum_{v \in V} d(v) = 2|E|$ for any graph G. Let $\delta(G)$ be the *minimal degree*, $\Delta(G)$ the *maximal degree* of G A graph G is *regular*, if $\Delta(G) = \delta(G)$, and it is *semi-regular*, if $\Delta(G) - \delta(G) = 1$.

Three vertices v_1, v_2, v_3 form *a triangle* in G if there are distinct verticess $e_1, v_2, v_3 \in F$ such that $\{v_i, v_{i+1}\} \subseteq E$, where the indices are taken mod 3. If G does not contain a triangle, then it is *trianglefree*.

A subset $U \subseteq V$ of vertices in a graph G is called a *vertex independent set* if no two vertices in U are adjacent. The maximum-size vertex independent set is called *maximum vertex independent set*. The size of the maximum vertex independent set is called *vertex independence number* and is denoted by $\alpha(G)$. The problem of finding a vertex maximum independent set and vertex independence number are NP-hard optimization problems [73, 167].

A *maximal vertex independent set* is a vertex independent set such that adding any other vertex to the set forces the set to contain an edge. The problem of finding a maximal vertex independent set can be solved in polynomial time (see e.g. the algorithms due to Tarjan and Trojanowski [155], Karp and Widgerson [101], further the improved algorithms due to Luby [128] and Alon [9].

There are exponential time exact (as Alon [9]) and polynomial time approximate algorithms (as Boppana and Haldórsson [30], Agnarsson, Haldórsson, and Losievskaja [4, 5], Losievskaja [126]) determining $\alpha(G)$. Also there are known algorithms producing the list of all maximum independent sets of graphs (see e.g. Johnson and Yannakakis [93], Lawler, Lenstra, Rinnooy Kan [121]).

An *independent edge set* of a graph G is a subset of the edges such that no two edges in the subset share a vertex of G [166]. An independent edge set of maximum size is called a *maximum independent edge set*, and an independent edge set that cannot be expanded to another independent edge set by addition of any other edge in the graph is called a *maximal independent edge set*. The size of the largest independent edge set (i.e., of any maximum independent edge set) in a graph is known as its *edge independence number* (or *matching number*), and is denoted by $\nu(G)$. The determination of $\nu(G)$ is an easy task for bipartite graphs [49, 50], but it is a polynomially solvable problem for general graphs too [10, 101, 161, 162].

Let $G = (V, E)$ be an n-order graph. The classical Turán theorem [159] gives

a simple lower bound for $\alpha(G)$.

Theorem 1 (Turán [159]) *If* $n \geq 1$ *and* G *is an* n*-order graph, then*

$$\alpha(G) \geq \frac{n}{d(G)+1}. \tag{1}$$

This result was strengthened independently in 1979 by Caro and in 1981 by Wei.

Theorem 2 (Caro [36], Wei, [165]) *If* $G(V, E)$ *is a graph, then*

$$\alpha(G) \geq \sum_{v \in V} \frac{1}{d(v)+1}. \tag{2}$$

Proof. See [36, 165]. □

A nice probabilistic proof of the result can be found in the paper of Alon and Spencer [11]. Since the function $\frac{1}{x+1}$ is convex, $\sum_{v \in V} \frac{1}{d(v)+1} \geq \frac{n}{d+1}$ [170].

Since this bound is the best-possible only for graphs which are unions of cliques, additional structural assumptions excluding these graphs allow improvement of 2 [80, 81]. A natural candidate for such assumptions is connectivity. In 2013 Angel, Campigotto, and Laforest [14] improved (2) for some connected graphs. For locally sparse graphs Ajtai, Erdős, Komlós and Szemerédi improved Turán's bound greatly.

Theorem 3 (Ajtai, Erdős, Komlós and Szemerédi [6, 7, 8]) *If* G *is an* n*-order triangle-free graph with average degree* d*, then*

$$\alpha(G) \geq \frac{cn \ln d}{d+1}. \tag{3}$$

Proof. See [6, 7, 8]. □

They conjectured that $c = 1 - o(1)$ when d tends to ∞. Griggs [72] improved that c can be $\frac{5}{12}$. Shearer [152] finally proved $c = 1 - o(1)$, thus confirming the conjecture. In 1994 Selkow improved the bound due to Caro and Wei supposing that the degrees of the neighbors of the vertices are also known.

Theorem 4 (Selkow [150]) *If* $G(V, E)$ *is a graph, then*

$$\alpha(G) \geq \sum_{v \in V} \frac{1}{d(v)+1} \left(1 + \max\left(0, \frac{d(v)}{d(v)+1} - \sum_{u \in N(v)} \frac{1}{d(u)+1}\right)\right). \tag{4}$$

Proof. See [150]. □

The bound of Selkow is equal to Caro–Wei bound for regular graph and always less then twice the Caro–Wei bound. A recent review on lower bounds for 3-order graphs was published by Henning and Yeo [89].

Let j and k be a positive integers. A subset $I \subseteq V(G)$ is a *vertex-*k*-independent set* of G, if every vertex in I has at most $k-1$ neighbors in I. The *vertex-k-independence number* $\alpha_k(G)$ of G is the cardinality of the largest vertex-k-independent set of G.

A subset $D \subseteq V(G)$ is a *vertex-j-dominating set* of G, if every vertex of D has at least $j-1$ neighbors in D. The *vertex-k-independence number* $\gamma_j(G)$ of G is the cardinality of the largest vertex-j-dominating set of G.

In 1991 Caro and Tuza [38] extended theorem of Turán to the estimation of the maximal size of k-independent sets. Thiele [156] in 1999, Csaba, Pick, and Shokoufandeh [44] in 2012 improved the bound due to Caro and Tuza. In 2008 Favoron, Hansberg and Volkmann [54] analyzed k-domination and minimum degree in graphs. Harant, Rautenbach, and Schiermeier [81, 83, 84, 85] proved different lower bounds on vertex independent number.

In 2012 Chellali and Rad [42] published a paper on k-independence critical graphs. In 2013 Caro and Hansberg [37] proposed a new approach to k-independence of graphs. Recently Chellali, Favaron, Hansberg, and Volkmann [41] published a review on k-independence.

Last year Hansberg and Pepper [79] investigated the connection between $\alpha_k(G)$ and $\gamma_j(G)$. They proved the following theorems.

Theorem 5 (Hansberg, Pepper [79]) *If Let* G *be an* n*-order graph,* j, k *and* m *be positive integers such that* $m = j + k - 1$ *and let* H_m *and* G_m *denote, respectively, the subgraphs induced by the vertices of degree at least* m *and the vertices of degree at least* m*. Then*

$$\alpha_k(H_m) + \gamma_j(G_m) \leq n \tag{5}$$

and

$$\alpha_k(G) + \gamma_j(G) \leq n(G_m). \tag{6}$$

Proof. See [79]. □

Theorem 6 (Hansberg, Pepper [79]) *Let* G *be a connected* n*-order graph with maximum degree* Δ *and minimum degree* $\delta \geq 1$*. Then*

$$\alpha_k(G) + \gamma_j(G) = n(G) \quad \textit{and} \quad \alpha_{k'}(G) + \gamma_{j'}(G) = n(G) \tag{7}$$

for every pair of integers $j,\ k$ *and* $j',\ k'$ *such that* $j+k-1=\delta$ *and* $j'+k'-1=\Delta$ *if and only ig* G *is regular.*

Proof. See [79]. □

Theorem 7 (Hansberg, Pepper [79]) *For any graph* G *the following two statements are equivalent:*

$$\gamma(G)+\alpha_\delta(G)=n(G) \tag{8}$$

and

$$G \quad \textit{is regular or} \quad \gamma(G)+\gamma_2(G)=n(G). \tag{9}$$

Proof. See [79]. □

Spencer [153] also published some extension of Turán theorem.

In 2014 Henning, Löwenstein, Southey and Yeo [87] proved the following theorem, which is an improvement of the result due to Fajtlowicz [53].

Theorem 8 (Henning et al. [87]) *If* G *is a graph of order* n *and* p *is an integer, such that for every clique* X *in* G *there exists a vertex* $x\in X$ *such, that* $d(x)<p-|X|$*, then* $\alpha(G)\geq 2n/p$.

There are results on the independence number of random graphs (e.g. Balogh, Morris, Samotij [18] and Frieze [60], Henning, Löwenstein, Southey and Yeo [87], on the weighted independence number (see e.g. Halldórsson [75], Kako, Ono, Hirata, and Halldórsson [98], further Sakai, Mitsunori, and Yamazaki [149]), and on the enumeration of maximum independent sets (see e.g. Gaspers, Kratsch, and Liedloff [69].

Let $G(n,p)=(V,E)$ the random graph with vertex set $V=\{v_1,\ldots,v_n\}$, p, $\alpha(G_{n,p})$ denote the independence number of $G_{n,p}$. In 1990 Frieze [60] proved, that if $d=np$ and $\epsilon>0$ is fixed, then with probability going to 1 as $n\to\infty$

$$\left|\alpha(G_n,p)-\frac{2n(\ln d-\ln\ln d-\ln 2+1)}{d}\right|\leq\frac{\epsilon n}{d}, \tag{10}$$

provided $d_\epsilon\leq d=o(n)$, where d_ϵ is some fixed constant and p is the join probability for each edge to be included in E.

In 1983 Shearer proved the following lower bound.

Theorem 9 (Shearer [152]) *If* G *is triangle-free, then*

$$\alpha(G)\geq nf(d), \tag{11}$$

where

$$f(x) = \frac{x \ln x - x + 1}{(x-1)^2}\,, \tag{12}$$

$f(0) = 1$ *and* $f(1) = \frac{1}{2}$.

According to the proof of Shearer for $0 < x < \infty$ hold $0 < f(d) < 1$, $f'd) < 0$ and $f''(d) < 0$. Further $f(x)$ satisfies the differential equation

$$(x+1)f(x) = (x+1)d^2f'(x). \tag{13}$$

It is easy to see that

$$\lim_{x\to\infty} \frac{f(x)}{x} = \frac{\ln x}{x}\,. \tag{14}$$

In 1995 Füredi [62] determined the number of different vertex maximal independent set in path graphs.

It is known [22] a minimum covering set of G is also a maximum vertex independent set of G. Therefore we are interested in the results on dominating sets (see e.g. [41, 54, 79, 82, 143].

The structure of the paper is as follows. After this introduction in Section 2 we present a review of results connected with th vertex and edge independence number of hypergraphs, then in Section 3 a lower bound of $\alpha(H)$ is presented for n-order r-uniform hipergraphs with average degree $d(H)$, and finally in Section 4 a similar bound is proved for hypergraphs not containing isolated vertex.

2 Introduction to independence in hypergraphs

Let $n \geq 1$ and $W = \{w_1, w_2, \ldots, w_n\}$ be a finite set called *vertex set*. A *hypergraph* H on vertex set W is a pair (W, F), where the edge set F is a family of the elements of W. We always assume that distinct edges are distinct as subsets. If each edge in F contains exactly $r \geq 2$ vertices, then H is r-*uniform*. So any graph G is a 2-uniform hypergraph.

Let $w \in W$ and $N(w)$ be the *neighborhood* of w, namely, the set of vertices x so that there is an edge which contains both w and x. Let U be a subset of W. The *sub-hypergraph* of H induced by U is defined as a hypergraph on vertex set U with edge set $F_U = \{f \in F : f \subseteq U\}$.

The *degree* $d(w)$ of a vertex $w \in W$ is the number of edges that contain w. Let $d(H) = d$ be the *average degree* of an r-uniform H, then $nd = \sum_{w\in W} d(w) = r|F|$.

For the simplicity we usually omit G and H as arguments of $d(H)$ and similar notations.

A hypergraph H is *linear*, if any two edges of H have at most one vertex in common. Note that a graph G is always linear. Three vertices w_1, w_2, w_3 *form a triangle* in H, if there are distinct edges $f_1, f_2, f_3 \in F$ such that $\{f_i, f_{i+1}\} \subseteq F$, where the indices are taken mod 3.

A subset $U \subseteq W$ of vertices in a hypergraph H is called a *vertex independent set* if no two vertices in U are adjacent. The maximum-size vertex independent set of H is called *maximum vertex independent set*. The size of the maximum vertex independent set is called *vertex independence number* and is denoted by $\alpha(H)$. The problem of finding a maximum vertex -independent set and vertex independence number are NP-hard optimization problems [73, 167].

There are exponential time exact (as Alon [9], Tarjan and Trojanowski [155]) and polynomial time approximate algorithms (as Boppana and Haldórsson [30], Agnarsson, Haldórsson, and Losievskaja [4, 5], Losievskaja [126]). Also there are known algorithms producing the list of all maximum independent sets of graphs (see e.g. Johnson and Yannakakis [93], Lawler, Lenstra, Rinnooy Kan [121]) and hypergraphs (see e.g. Kelsen [107]).

A *maximal vertex independent set* is a vertex independent set such that adding any other vertex to the set forces the set to contain an edge. The problem of finding a maximal vertex independent set can be solved in polynomial time (see e.g. the algorithms due to Tarjan and Trojanowski [155], Karp and Widgerson [101], further the improved algorithms due to Luby [128] and Noga [9]).

In 2012 Dutta, Mubayi, and Subramanian [48] gave new lower bond for the vertex independence number of sparse hypergraphs.

In 2013 Eustis devoted a PhD dissertation to the problems of hypergraph independence numbers [51, 52].

An *independent edge set* of a hypergraph H is a subset of the edges such that no two edges in the subset share a vertex of H [136]. An independent edge set of maximum size is called a *maximum independent edge set*, and an independent edge set that cannot be expanded to another independent edge set by addition of any other edge in the hypergraph is called a *maximal independent edge set*. The size of the largest independent edge set (i.e., of any maximum independent edge set) in a hypergraph is known as its *edge independence number* (or *matching number*), and is denoted by $\nu(H)$. The determination of $\nu(H)$ is an easy task for bipartite graphs [49, 50], but it is a polynomially solvable problem for general graphs too [10].

There are many results on the characterization of hypergraph score se-

quences and on their reconstruction (see e.g. [20, 110, 140, 171, 139, 164, 172]), on the enumeration of different hypergraphs (see e.g. [21, 47, 138, 144, 145]) and directed hypergraphs (see e.g. [15]).

An r-uniform hypergraph with n vertices is called *complete*, if its set of edges has the cardinality $\binom{n}{2}$. The *complement* of an r-uniform hypergraph H is $\overline{H} = (W, \overline{F})$, if $|F \cup \overline{F}| = \binom{n}{2}$ and $|F \cap \overline{F}| = 0$.

A set $P \subseteq W$ is called an *edge cover* of H, if for any non-isolated vertex $x \in W$ there exists an edge $f_i \in P$ that $x \in f_i$. The cardinality of a minimum set which is an edge covering of H is called the *edge covering number* of H, and is denoted by $\nu(H)$.

The following lemma, proved in [97], gives a relation between the edge covering number and the edge independence number in an r-uniform hypergraph H without isolated vertices.

Lemma 10 (Jucovič, Olejník [97]) *For an r-uniform n-order hypergraph H with n without isolated vertices the following inequalities hold:*

$$\alpha(H) \leq n - (kr - 1)\nu(H), \tag{15}$$

$$\alpha(H) + (r - 1)\nu(H) \leq n. \tag{16}$$

$$\nu(H) + (r - 1)r - 1\nu(H) \geq n, \tag{17}$$

Proof. See [97]. □

This lemma generalizes the relations published by Gallai [67] in 1959. In 1991 Tuza [160] extended Gallai's inequalty for uniform hypergraphs.

In 1989 Olejník proved the following three theorems characterizing $\alpha(H)$ and $\nu(H)$.

Theorem 11 (Olejník [136]) *For an r-uniform n-order hypergraph $H = (W, F)$ with n and its complement $\overline{H} = (W, \overline{F})$*

$$\left\lfloor \frac{n}{r} \right\rfloor \leq \nu(H) + \nu(\overline{H}) \leq 2 \left\lfloor \frac{n}{r} \right\rfloor \tag{18}$$

and

$$0 \leq \nu(H)\nu(\overline{H}) \leq \left\lfloor \frac{n}{r} \right\rfloor^2. \tag{19}$$

Proof. See [136]. □

This bounds are direct generalizations of the bounds published by Chartrand and Schuster in 1974 [40].

Theorem 12 (Olejník [136]) *For an* r*-uniform* n*-order hypergraph* $H = (W, F)$ *and its complement* $\overline{H} = (W, \overline{F})$*, where neither* H *nor* $\overline{F}$ *have isolated vertices,*

$$\left\lfloor \frac{n}{r} \right\rfloor \leq \nu(H) + \nu(\overline{H}) \leq 2 \left\lfloor \frac{n}{r} \right\rfloor \tag{20}$$

and

$$0 \leq \nu(H)\nu(\overline{H}) \leq \left\lfloor \frac{n}{r} \right\rfloor^2. \tag{21}$$

Proof. See [136]. □

This result is an extension of the work of R. Laskar and B. Auerbach published in 1978 [120].

Theorem 13 (Olejník [136]) *For an* r*-uniform* n*-order hypergraph* $H = (W, F)$ *and its complement* $\overline{H}, \overline{F}$*, where neither* H *nor* $\overline{H}$ *have isolated vertices and* $n \neq 2r$

$$2 \left\lfloor \frac{n}{r} \right\rfloor \leq \alpha H + \alpha \overline{H} \leq 2n - (r-1) \left\lfloor \frac{n}{r} \right\rfloor - r + 1 \tag{22}$$

and

$$\left\lfloor \frac{n}{r} \right\rfloor^2 \leq \alpha(H)\alpha(\overline{H}) \leq \frac{1}{4} \left(2n - (r-1) \left\lfloor \frac{n}{r} \right\rfloor - k + 1 \right)^2. \tag{23}$$

Proof. See [136]. □

In 1993 Gallo, Longo, Nguyen, and Pallottino [68] studied the applications of directed hypergraphs. In 2004 Vietri [163] wrote on the complexity of the arc-coloring of directed hypergraphs. In 2003 Frank, Király and Király [55] analized the orientation of directed hypergraphs.

Let

$$B(p, q) = \int_0^1 (1-t)^{p-1} t^{q-1} dt \tag{24}$$

denote the beta-function with $p,\ q > 0$. Set constants $0 < a \leq 1$, $0 < b \leq 1$, and $B = B(a, 1-b)$, and let

$$f_r(x) = \frac{1}{B} \int_0^1 \frac{1-t)^a}{(t^b[1 + (x-1)t]} dr. \tag{25}$$

In 2004 Zhou and Li [170] proved the following theorem on sparse hypergraphs.

Theorem 14 (Zhou, Li [170]) *Let* H *be a triangle-free,* r*-uniform* ($r \geq 2$) n*-order linear hypergraph with average degree d. Then its strong vertex independence number* $\alpha_s(G)$ *is at least* $nf_r(d)$*.*

Proof. See [170]. □

In 2004 Greenhill, Ruciński, and Wormald [71] analyzed random hypergraph processes with degree restrictions. In 2008 Plociennik [141] proposed an approximation algorithm for the vertex maximum independence set problem of uniform random hypergraphs. M. Halldórsson, and Losievskaja [4, 5] used semidefinit programming to find maximum vertex independent set of hypergraphs.

Shearer's result ([152], further (11) and (12)) was generalized in [170] with the function $g_r(x)$ satisfying

$$(r-1)^2x(x-1)g_r'(x) + [(r-1)x+1]g_r(x) = 1 \tag{26}$$

for r-uniform, triangle-free linear hypergraphs, with sparse neighborhood and in [125] with the function $g_{r,m}(x)$ satisfying

$$(r-1)^2x(x-m)g_{r,m}'(x) + [(r-1)x+1]g_{r,m}(x) = 1 \tag{27}$$

for r-uniform, triangle-free, and double linear hypergraphs, in which each subhypergraph induced by a neighborhood, has maximum degree less than m. A linear hypergraph is called *double linear* if for any non-adjacent distinct vertices w and z, each edge containing w has at most one neighbor of z. From the uniqueness of solutions of the differential equations, we see that $g_2(x) = g(x)$ and $g_{r,1}(x) = g_r(x)$. It is shown [125] that $g_{2,m}(x) \sim \frac{\log x}{x}$, and for $g_{r,m}(x) \sim \frac{c}{d^{1/(r-1)}}$ for $r \geq 3$, where $c = c(r,m) > 0$ is a constant without knowing exact values.

Independent sets and numbers are studied in many papers (see e.g. the papers of Abraham [1], Alon, Uri and Azar [12], Berger and Ziv [23], Bollobás, Daykin and Erdős [27], Bonato, Brown, Mitsche and Pralat [28, 29], Bordewich, Dyer and Karpiński [31], Boros, Gurvich, Elbassioni, Gurvich and Khachiyan [32, 33], Borowiecki and Michalak [34], Cutler and Radcliffe [45], Greenhill [70], Halldórson and Losievskaja [76], Hofmeister and Lehman [90], Johnson and Yannakakis [93], Khachiyan, Boros, Gurvich, and Elbassioni [108], Lepin [122], Li and Zhang [125], Losievskaja [126], Shachnai and Srinivasan [151], Tarjan and Trojanowski [155], Yuster [168]).

Since independence number and matching number are closely connected, we are interested in the results on maximum matching algorithms too (see e.g. [25, 26, 46, 47, 49, 50, 56, 57, 61, 65, 66, 77, 78, 86, 88, 89, 91, 92, 100, 104, 105, 109, 112, 113, 118, 119, 127, 131, 132, 133, 135, 137, 142, 146, 147, 148, 154, 157, 158, 169]).

Minimum dominating set of H and maximum vertex independent set of H are connected concepts, therefore we are interested in the results on dominating sets of hypergraphs (see e.g. [2, 96]).

Further connected problems are also often analyzed (see e.g. e.g. in the papers of Agnarsson, Egilssson, and Halldórson [3], Alon, Frankl, Huan, Rödl, Ruciński [10], Alon and Yuster [13], Baranyai [19], Balogh, Butterfield, Hu and Lenz [17], Bertram-Kretzberg and Letzman [24], Bujtás and Tuza [35], Cockayne, Hedetniemi, and Laskar [43], Frank, Király and Király [55], Frankl and Rödl [58, 59], Füredi, Ruszinkó, and Selver [63, 64], Hán, Person and Schacht [78], Henning and Yeo [89], Huang, Loh and Sudakov [92], Johnson and Yannakakis [93], Johnston and Lu [94, 95], Jucovič and Olejník [97], Karoński and Luczak [99], Katona [102, 103], Keevash and Sudakov [106], Kelsen [107], Kohayakawa, Rödl, Skokan [111], Krivelevich [115], Kühn and Loose [117], Kostochka, Mubayi, Verstraëte [114], Krivelevich, Nathaniel, and Sudakov [116], Li, Rousseau and Zang [123, 124], Luczak and Szymańska [129, 134], Szymańska [154], Treglown and Zhao [157, 158], Tuza [160], Yuster [169]).

Although hypergraphs are less often used in the practice than the graphs, they also have different applications in the practice.

For example Bailey, Manoukian, Ramamohanaro [16], further Gunopolus, Khardon, Mannila and Toivonen [74] reported on the applications in data mining, Gallo, Longo, Nguyen, and Pallottino [68], further and Maier [130] in relational databases.

In 2000 Carr, Lancia, Istrail, and Genomics [39] reported on Branch-and-Cut algorithms for vertex independent set problem and on their application to solve problems connected with protein structure alignment.

In this paper, we obtain $\alpha(H) \geq \sum_{v \in V} \frac{1-1/r}{d(v)^{1/(r-1)}}$ for any r-uniform hypergraph H without the condition of being triangle-free. The algorithm is naive: it deletes a vertex of maximum degree repeatedly. In order to get a large independent set, a commonly used algorithm is to find a suitable vertex v, then delete v and its neighbors, and then do the iterations. Deleting all neighbors seems to be of no use for hypergraphs as in [125, 170]. After deleting a vertex v, we delete only one vertex other than v from each edge containing v. Our new function $f_r(x)$ satisfies

$$[(r-1)x^2 - x]f_r'(x) + (x+1)f_r(x) = 1. \tag{28}$$

Then $f_r(x) \sim \frac{c}{x^{1/(r-1)}}$ as $x \to \infty$. We do not know the exact value of $c = c(r)$. However, when we run the algorithm, we note that for a vertex v, we delete $1 + d(v)$ vertices instead of deleting $1 + (r-1)d(v)$ vertices as in [125, 170]. So

if c is the constant such that $g_r(x) \sim \frac{c}{x^{1/(r-1)}}$ as $x \to \infty$, then the new constant seems to be $(r-1)c$, namely, $f_r(x) \sim \frac{(r-1)c}{x^{1/(r-1)}}$.

3 Bound for uniform hypergraphs without isolated vertex

The following Theorem 15 is a corollary of Theorem 18, but it has an easy probabilistic proof.

Theorem 15 *Let $H = (V, E)$ be an r-uniform hypergraph of order n and average degree $d \geq 1$, then*

$$\alpha(H) \geq \left(1 - \frac{1}{r}\right) \frac{n}{d^{1/(r-1)}}. \tag{29}$$

Proof. Define a random subset $U \subseteq V$ by $\Pr(v \in U) = p$ for some $0 \leq p \leq 1$ with all these events being mutually independent over $v \in V$.

Let $X(U)$ be the number of vertices in U and let $Y(U)$ be the number of edges in the subgraph induced by U. Note that for one of the edges of H, the probability that all of its vertices belong to U is p^r. By linearity of expectation, we have

$$E(X - Y) = E(X) - E(Y) = np - \frac{nd}{r}p^r. \tag{30}$$

Thus there exists a set U satisfying

$$X(U) - Y(U) \geq E(X) - E(Y). \tag{31}$$

Note that U is not that we require, since the sub-hypergraph of H induced by U may have edges. However, if we delete one vertex from each edge contained in U, then at most $Y(U)$ vertices are deleted, we thus obtain a new set with at least $E(X) - E(Y)$ vertices and whose induced sub-hypergraph has no edges. The desired lower bound follows by taking $p = \frac{1}{d^{1/(r-1)}}$. □

For hypergraphs that are not regular, Theorem 18 is stronger than Theorem 15. We need two lemmas for the proof of Theorem 18.

Lemma 16 *Let $r \geq 2$ be an integer and define*

$$h_r(x) = \begin{cases} 1 - x/r & \text{if } 0 \leq x < 1 \\ \frac{1-1/r}{x^{1/(r-1)}} & \text{if } x \geq 1, \end{cases} \tag{32}$$

then $h_r(x)$ is positive, decreasing and convex. Furthermore, for $x \geq 1$, the function $h_r(x)$ satisfies that $(r-1)x\,h'(x) + h_r(x) = 0$.

Proof. It is easy to see that $h_r(x)$ is positive and

$$h_r'(x) = \begin{cases} -1/r & \text{if } 0 \le x < 1 \\ \frac{-1/r}{x^{r/(r-1)}} & \text{if } x \ge 1. \end{cases} \tag{33}$$

So $h_r'(x)$ is continuous, negative and increasing, thus $h_r(x)$ is decreasing and convex. The fact that $h_r(x)$ satisfies the mentioned differential equation is straightforward. □

Let $\Delta = \Delta(H)$ denote the maximal degree in H and define

$$S(G) = \sum_{x \in V} h(d(x)), \quad S(H) = \sum_{x \in W} h(d(x)). \tag{34}$$

Lemma 17 *If* $\Delta(H) \ge 1$, $w \in W$, $d(w) = \Delta(H)$, *and* $H_1 = H - \{w\}$, *then* $S(H_1) \ge S(G)$.

Proof. For each $x \in V \setminus \{v\}$, denote by n_x the number of edges of H that contain both x and v. Then $n_x = 0$ if x and v are not adjacent, and $n_x \ge 1$ otherwise. It is easy to see

$$\sum_{x \in V \setminus \{v\}} n_x = (r-1)\Delta \tag{35}$$

since H is r-uniform. On the other hand, we have

$$S(H_1) = S(H) - h(\Delta) + \sum_{x \in V \setminus \{v\}} [h(d(x) - n_x) - h(d(x))]. \tag{36}$$

From the fact that $h'(x)$ is negative and increasing, we have

$$h(d(x) - n_x) - h(d(x)) = -h'(\theta_x) n_x \ge -h'(\Delta) n_x, \tag{37}$$

where $\theta_x \in [d(x) - n_x, d(x)]$, thus

$$\begin{aligned} S(H_1) &\ge S(H) - h(\Delta) - h'(\Delta) \sum_{x \in V \setminus \{v\}} n_x \\ &= S(H) - h(\Delta) - (r-1)\Delta h'(\Delta) \\ &= S(H), \end{aligned}$$

proving the claim. □

Theorem 18 *Let* $H = (V, E)$ *be an* r*-uniform hypergraph without isolated vertex, then*

$$\alpha(H) \geq \left(1 - \frac{1}{r}\right) \sum_{v \in V} \frac{1}{d(v)^{1/(r-1)}}. \tag{38}$$

Proof. We write $h_r(x)$ as $h(x)$ for simplicity and define

$$S(H) = \sum_{x \in V} h(d(x)). \tag{39}$$

Repeat the algorithm by deleting the vertex of maximum degree if the degree is at least one, terminate the algorithm if there are no edges. Denote by $H_0 = H, H_1, \ldots, H_\ell$ for the sequence of hypergraphs, where H_ℓ has no edge. We get $S(H_\ell) = n - \ell$ since $h(0) = 1$, where $n - \ell$ is the order of H_ℓ, and $\alpha(H) \geq n - \ell$. So

$$\alpha(H) \geq S(H_\ell) \geq S(H_{\ell-1}) \geq \cdots \geq S(H_0) = S(H), \tag{40}$$

the assertion follows immediately. □

Since the function $\frac{1}{x^{1/(r-1)}}$ is convex, Theorem 15 is truly a corollary of Theorem 18.

Remark. Theorem 18 gives $\alpha(G) \geq \sum_v \frac{1}{2d(v)}$ for a graph G with $\delta(G) \geq 1$, which is weaker than $\alpha(G) \geq \sum_v \frac{1}{d(v)+1}$. However, the later can be proved similarly by replacing the function $h(x)$ with $1/(x+1)$. For details of this algorithm, see Griggs [72].

4 Bound for uniform linear triangle-free hypergraphs

In this section triangle-free hypergraphs are considered. To generalize Shearer's method [152] and to delete less vertices for a hypergraph, we have a definition as follows.

Let $H = (V, E)$ be an r-uniform hypergraph and let v be a vertex of H, denote by $E_v = \{e \in E : v \in e\} = \{e_1, e_2, \ldots, e_{d(v)}\}$ for the set of edges containing v. A *claw* of v is a set of neighbors of v of the form $\{u_1, u_2, \ldots, u_{d(v)}\}$ such that each $u_i \in e_i - v$. For a claw T of v, we write as Q_T, the number of edges that intersect T.

When we run the algorithm in each step, we will delete v and a claw T, so Q_T edges will be deleted. The new function is as follows.

Let $r \geq 2$ be and integer and let $b = \frac{r-2}{r-1}$. Define

$$f_r(x) = \frac{1}{r-1} \int_0^1 \frac{1-t}{t^b[1+((r-1)x-1)t]} dt. \tag{41}$$

Lemma 19 *The function* $f_r(x)$ *satisfies the differential equation*

$$[(r-1)x^2 - x]f_r'(x) + (x+1)f_r(x) = 1, \tag{42}$$

and it is positive, decreasing and convex.

Proof. By differentiating under the integral and then integrating by parts, we have

$$\begin{aligned}
& [(r-1)x^2 - x]f_r'(x) \\
= & -[(r-1)x^2 - x] \int_0^1 \frac{1-t}{t^{1-b}[1+((r-1)x-1)t]^2} dt \\
= & \; x \int_0^1 (1-t)t^{1-b} \frac{d}{dt} \left(\frac{1}{1+[(r-1)x-1]t} \right) \\
= & -x \int_0^1 \frac{1}{1+[(r-1)x-1]t} [(1-t)(1-b)t^{-b} - t^{1-b}] dt \\
= & -(r-1)(1-b)xf_r(x) + x \int_0^1 \frac{t^{1-b}}{1+[(r-1)x-1]t} dt \\
= & -xf_r(x) + \frac{1}{r-1} \int_0^1 \left(\frac{1}{1-t} - \frac{1}{1+[(r-1)x-1]t} \right) (1-t)t^{-b} dt \\
= & -xf_r(x) + 1 - f_r(x) \\
= & \; 1 - (x+1)f_r(x)
\end{aligned}$$

which follows by the differential equation. The monotonicity and convexity of $f_r(x)$ can be seen by repeated differentiation under the integral. □

Theorem 20 *Let* H *be an* r*-uniform* n*-order hypergraph with average degree* d*. If it is triangle-free and linear, then* $\alpha(H) \geq n f_r(d)$.

Proof. We apply induction on $|V|$, the number of vertices of H. The result is trivial for $|V| = 1$, since $f(0) = 1$. Since the case $r = 2$ is exactly what Shearer has given, we suppose that $r \geq 3$.

For each $v \in H$, let $T = \{u_1, u_2, \ldots, u_{d(v)}\}$ be a claw of v. Since H is r-uniform, linear and triangle-free, we have

$$Q_T = d(v) + \sum_{i=1}^{d(v)} (d(u_i) - 1) = \sum_{i=1}^{d(v)} d(u_i). \tag{43}$$

Let $\mathcal{T}_v$ be the set of all claws of v, then $|\mathcal{T}_v| = (r-1)^{d(v)}$. Therefore

$$\sum_{T \in \mathcal{T}_v} Q_T = \sum_{T \in \mathcal{T}_v} \sum_{i=1}^{d(v)} d(u_i) = \sum_{u \in n(v)} (r-1)^{d(v)-1} d(u), \tag{44}$$

and

$$\frac{1}{|\mathcal{T}_v|} \sum_{T \in \mathcal{T}_v} Q_T = \sum_{u \in n(v)} \frac{d(u)}{r-1}. \tag{45}$$

We write $f(x)$ for $f_r(x)$ and set

$$R_T(v) = 1 - (d(v)+1) f(d) + (d d(v) + d - r Q_T) f'(d). \tag{46}$$

Then the average of $R_T(v)$ among $T \in \mathcal{T}_v$ is

$$\frac{1}{|\mathcal{T}_v|} \sum_{T \in \mathcal{T}_v} R_T(v) = 1 - (d(v)+1) f(d) + (d d(v) + d) f'(d) - r \sum_{u \in n(v)} \frac{d(u)}{r-1} f'(d). \tag{47}$$

Note that

$$\frac{1}{n} \sum_{v \in V} \sum_{u \in N(v)} \frac{d(u)}{r-1} = \frac{1}{n} \sum_{v \in V} d^2(v) \geq d^2 \tag{48}$$

as x^2 is a convex function. Since $f'(x) < 0$, we have

$$\frac{1}{n} \sum_{v \in V} \frac{1}{|\mathcal{T}_v|} \sum_{T \in \mathcal{T}_v} R_T(v) \geq 1 - (d+1) f(d) + (d^2 + d - r d^2) f'(d) = 0. \tag{49}$$

Hence there exists a vertex, say v, and a claw of v, say $T = \{u_1, u_2, \ldots, u_{d(v)}\}$, such that $R(v) \geq 0$. Now by deleting v and $u_1, u_2, \ldots, u_{d(v)}$, we obtain a new hypergraph H' with $n - d(v) - 1$ vertices and $\frac{nd}{r} - Q_T$ edges. For an edge e containing v, it contains $r \geq 3$ vertices, and we delete exactly two vertices from e, so H' has some vertices. Note that the average degree $\bar{d}$ of H' is $\frac{nd - rQ_T}{n - d(v) - 1}$. By induction hypothesis, we have

$$\alpha(H) \geq (n - d(v) - 1) f(\bar{d}) = (n - d(v) - 1) f\left(\frac{nd - rQ_T}{n - d(v) - 1}\right). \tag{50}$$

Combining the facts that $\alpha(H) \geq 1 + \alpha(H')$ and $f(x) \geq f(d) + f'(d)(x - d)$ for all $x \geq 0$ as $f(x)$ is convex, we obtain

$$\begin{aligned} \alpha(H) &\geq 1 + (n - d(v) - 1)f\left(\frac{nd - rQ_T}{n - d(v) - 1}\right) \\ &\geq 1 + (n - d(v) - 1)f(d) + (dd(v) + d - rQ_T)f'(d) \\ &= nf(d) + R(v) \geq nf(d) \end{aligned}$$

completing the proof. □

We now get an asymptotic form of $f_r(x)$ as $\frac{c}{x^{1/(r-1)}}$ without knowing exact expression of $c = c(r)$ in hope of improving the old constant based on analysis of the algorithm as mentioned.

Lemma 21 *Let $r \geq 3$ be an integer. Then*

$$\lim x \to \infty f_r(x) = \frac{c}{x^{1/(r-1)}}, \tag{51}$$

where $c = c(r)$ is a positive constant.

Proof. Recall that a first order linear differential equation $\frac{dy}{dx} = p(x)y + q(x)$ has the unique solution of the form

$$y = e^{\phi(x)}\left(y_0 + \int_{x_0}^{x} q(t)e^{-\phi(t)}dt\right) \tag{52}$$

satisfying $y_0 = y(x_0)$, where $\phi(x) = \int_{x_0}^{x} p(t)dt$. From the differential equation that $f_r(x)$ satisfies, we set

$$p(x) = -\frac{x + 1}{(r - 1)x^2 - x}, \quad \text{and} \quad q(x) = \frac{1}{(r - 1)x^2 - x}. \tag{53}$$

For $x_0 = 2$,

$$\phi(x) = -\int_2^x \frac{t + 1}{(r - 1)t^2 - t}dt = \ln \frac{c_1 x}{[(r - 1)x - 1]^{\frac{r}{r-1}}} \tag{54}$$

Hence

$$e^{\phi(x)} = \frac{c_1 x}{[(r - 1)x - 1]^{\frac{r}{r-1}}} \sim \frac{c_2}{x^{1/(r-1)}}. \tag{55}$$

Then we have

$$q(t)e^{-\phi(t)} \sim \frac{1}{c_2(r - 1)}x^{1/(r-1)-2}, \tag{56}$$

implying that $c_3 = \int_2^\infty q(t)e^{-\phi(t)}dt < \infty$, and $\int_2^x q(t)e^{-\phi(t)}dt = c_3 + o(1)$ as $x \to \infty$. Therefore,

$$f_r(x) = e^{\phi(x)}\left(y_0 + c_3 + o(1)\right) \sim \frac{c}{x^{1/(r-1)}}, \tag{57}$$

where $c = c_2(y_0 + c_3)$ and $y_0 = f_r(2)$ are positive constants. □

Acknowledgement The authors thank Attila Kiss (Eötvös Loránd University) and the unknown referee for the useful comments. This research is supported in part by the National Science Foundation of China (No. 11371193).

References

[1] V. Abraham, Independence in hypergraphs. *J. Indian Math. Soc.* (N.S.), **78,** 1-4 (2011) 1–7. ⇒141

[2] B. D. Acharya, Interrelations among the notions of independence, domination and full sets in a hypergraph, *Nat. Acad. Sci. Lett.* **13,** 11 (1990) 421–422. ⇒142

[3] G. Agnarsson, A. Egilsson, M. M. Halldórson, Vertex coloring acyclic digraphs and their corresponding hypergraphs, *Discrete Appl. Math.* **156,** 10 (2008) 1918–1928. ⇒142

[4] G. Agnarsson, M. M. Halldórson, E. Losievskaja, SDP-based algorithms for maximum independent set problems on hypergraphs, in: *Automata, Languages and Programming. Part I*, *Lecture Notes in Comput. Sci.* **5555**, Springer, Berlin, 2009, 12–23. ⇒133, 138, 141

[5] G. Agnarsson, M. M. Halldórson, E. Losievskaja, SDP-based algorithms for maximum independent set problems on hypergraphs, *Theoret. Comp. Sci.* **470** (2013) 1–9. ⇒133, 138, 141

[6] M. Ajtai, P. Erdős, J. Komlós and E. Szemerédi, On Turán theorem for sparse graphs, *Combinatorica* **1,** 3 (1981) 313–317. ⇒134

[7] M. Ajtai, P. Erdős, J. Komlós, E. Szemerédi, A dense infinite Sidon sequence, *European J. Combin.* **2,** 1 (1981) 1–11. ⇒134

[8] M. Ajtai, J. Komlós and E. Szemerédi, A note on Ramsey numbers, *J. Combin. Theory Ser. A* **29** (1980) 354–360. ⇒134

[9] A. Alon, A fast and simple randomized parallel algorithm for the maximal independent set problem, *J. Algorithms* **7,** 4 (1986) 567–583. ⇒133, 138

[10] N. Alon, P. Frankl, H. Huang, V. Rödl, A. Ruciński, Large matchings in uniform hypergraphs and the conjectures of Erdős and Samuels, *J. Combin. Theory Ser. A* **119,** 6 (2012) 1200–1215. ⇒133, 138, 142

[11] N. Alon, J. H. Spencer, *The Probabilistic Method*, Wiley-Interscience, New York, 1992. ⇒134

[12] A. Alon, A. Uri, Y. Azar, Independent sets in hypergraphs with applications to routing via fixed paths. In: *Randomization, approximation, and combinatorial optimization* (Berkeley, CA, 1999), *Lecture Notes in Comput. Sci.* **1671**, Springer, Berlin, 1999, 16–27. ⇒141

[13] N. Alon, R. Yuster, On a hypergraph matching problem, *Graphs Comb.* **21** (2005) 377–384. ⇒142

[14] E. Angel, R. Campigotto, C. Laforest, A new lower bound on the independence number of graphs, *Discrete Appl.* **161,** 6 (2013) 847–852. ⇒134

[15] G. Ausiello, A. D'Atri, D. Saccà, Minimal representation of directed hypergraphs, *SIAM J. Comput.* **15** (1986) 418–431. ⇒139

[16] J. Bailey, T. Manoukian, K. Ramamohanaro, A fast algorithm for computing hypergraph transversals and its application in mining emerging patterns, in: *Proc. of the Third IEEE Int.l Conf. on Data Mining* (*ICDM03*), 2003, 485–488. ⇒142

[17] J. Balogh, J. Butterfield, P. Hu, J. Lenz, D. Mubayi, On the chromatic thresholds of hypergraphs arXiv:1103.1416, 2013, 37 pages. ⇒142

[18] J. Balogh, R. Morris, W. Samotij, Independent sets in hypergraphs, arXiv:1204.6530, 2013, 42 pages. ⇒136

[19] Zs. Baranyai, On the factorization of the complete uniform hypergraph, in: A. Hajnal, R. Rado, V. T. Sós (Eds.), *Infinite and Finite Sets*, Proc. Coll. Keszthely, **1973**, North-Holland, Amsterdam, Netherlands, 1975, pp. 91–108. ⇒142

[20] S. Behrens, C. Erbes, M. Ferrara, S. G. Hartke, B. Reiniger, H. Spinoza, C. Tomlinson, New results on degree sequences of uniform hypergraphs. *Electron. J. Combin.* **20,** 4 (2013) #P14, 18 pages. ⇒139

[21] M. Behrisch, A. Coja-Oghlan, M. Kang, The asymptotic number of connected d-uniform hypergraphs, *Combin. Probab. Comput.* **23,** 3 (2014) 367–385. ⇒139

[22] C. Berge, *The Theory of Graphs and its Applications*, Wiley, London, 1964. ⇒137

[23] E. Berger, R. Ziv, A note on the edge cover number and independence number in hypergraphs, *Discrete Math.* **308,** 12 (2008) 2649–2654. ⇒141

[24] C. Bertram-Kretzberg, H. Lefmann, The algorithmic aspects of uncrowded hypergraphs. *SIAM J. Comput.* **29,** 1 (1999) 201–230. ⇒142

[25] T. Biedl, E. D. Demaine, C. A. Duncan, R. Fleischer, S. G. Kobourov, Tight bounds on maximal and maximum matchings, *Discrete Math.* **285** (2004) 7–15. ⇒141

[26] V. Blinovsky, Fractional matching in hypergraphs, arXiv:1311.2671, 2013, 6 pages. ⇒141

[27] B. Bollobás, D. E. Daykin, P. Erdős, Sets of independent edges of a hypergraph, *Quart. J. Math. Oxford Ser. (2)* **27,** 1 (1976), 25–32. ⇒141

[28] A. Bonato, J. I. Brown, D. Mitsche, P. Pralat, Independence densities of hypergraphs, arXiv:1308.2837, 2013, 16 pages. ⇒141

[29] A. Bonato, J. I. Brown, D. Mitsche, P. Pralat, Independence densities of hypergraphs, *Europ. J. Combin.* **40** (2014) 124–136. ⇒141

[30] R. Boppana, M. M. Halldórson, Approximating maximum independent set by excluding subsets, *BIT*, **32** (1992) 160–196. ⇒133, 138

[31] M. Bordewich, M. Dyer, M. Karpiński, Path coupling using stopping times and counting independent sets and colorings in hypergraphs *Random Structures Alg.* **32,** 3 (2008) 375–399. ⇒141

[32] E. Boros, K. Elbassioni, V. Gurvich, L. Khachiyan,Generating maximal independent sets for hypergraphs with bounded edge-intersections. In: *LATIN 2004: Theoretical informatics*, Lecture Notes in Comput. Sci. **2976**, Springer, Berlin, 2004, 488–498. ⇒141

[33] E. Boros, V. Gurvich, K. Elbassioni, L. Khachiyan, An efficient incremental algorithm for generating all maximal independent sets in hypergraphs of bounded dimension. *Parallel Process. Lett.* **10,** 4 (2000) 253–266. ⇒141

[34] M. Borowiecki, D. Michalak, The independence graphs of hypergraphs and middle graphs, *Discuss. Math.* **7** (1985) 31–37. ⇒141

[35] Cs. Bujtás, Zs. Tuza, Uniform mixed hypergraphs: The possible numbers of colors, *Graphs Combin.*, **24** (2008) 1–12. ⇒142

[36] Y. Caro, New results on the independence number, *Technical Report*, 1979, Tel-Aviv University. ⇒134

[37] Y. Caro, A. Hansberg, New approach to the k-independence number of a graph. *Electron. J. Combin.* **20,** 1 (2013) #P33, 17 pages. ⇒135

[38] Y. Caro, Zs. Tuza, Improved lower bounds on k-independence, *J. Graph Theory* **15** (1991) 99–107. ⇒135

[39] R. D. Carr, G. Lancia, S. Istrail, C. Genomics, Branch-and-Cut algorithms for independent set problems: Integrality gap and an application to protein structure alignment. SAND Report SAND2000-2171, Sandia National Laboratories, 2000. ⇒142

[40] G. Chartrand, S. Schuster, On the independence number of complementary graphs, *Trans. New York Acad. Sci., Series II* **36,** 3 (1974) 247–251. ⇒139

[41] M. Chellali, O. Favaron, A. Hansberg, L. Volkmann, k-domination and k-independence in graphs: a survey, *Graphs Combin.* **28,** 1 (2012) 1–55. ⇒135, 137

[42] M. Chellali, N. J. Rad, On k-independence critical graphs. *Australas. J. Combin.* **53** (2012) 289–298. ⇒135

[43] E. J. Cockayne, S. T. Hedetniemi, R. Laskar, Gallai theorems for graphs, hypergraphs and set systems, *Discrete Math.* **72** (1988) 35–47. ⇒142

[44] B. Csaba, T. A. Pick, A Shokoufandeh, A note on the Caro-Tuza bound on the independence number of uniform hypergraphs, *Australas. J. Combin.* **52** (2012) 235–242. ⇒135

[45] J. Cutler, A. J. Radcliffe, Hypergraph independent sets, *Combin. Probab. Comput.* **22,** 1 (2013) 9–20. ⇒141

[46] B. C. Dean, S. M. Hedetniemi, S. T. Hedetniemii, J. Lewis, A. McRae, Matchability and k-maximal matchings. *Discrete Math.* **159,** 1 (2011) 15–22. ⇒141
[47] A. Dudek, A. Frieze, A. Ruciński, M. Šileikis, Approximate counting of regular hypergraphs, *Inf. Processing Letters*, **113,** 1921 (2013) 785–788. ⇒139, 141
[48] K. Dutta, D. Mubayi, C. R. Subramanian, New lower bounds for the independence number of sparse graphs and hypergraphs. *SIAM J. Discrete Math.* **26,** 3 (2012) 1134–1147. ⇒138
[49] J. Edmonds, Paths, trees, and flowers, *Canadian J. Math.* **17** (1965) 449–467. ⇒133, 138, 141
[50] J. Edmonds, Maximal matching and a polyhedron with 0, 1-vertices, *J. Research Nat. Bureau Standards* **B-69** (1965) 125–130. ⇒133, 138, 141
[51] A. Eustis, *Hypergraph Independence Numbers*, PhD Thesis. University of California, San Diego. 2013. 123 pages. ⇒138
[52] A. Eustis, J. Verstraëte, On the independence number of Steiner systems, *Combinatorics, Prob. Comp.* **22,** 2 (2013) 241–252. ⇒138
[53] S. Fajtlowicz, Independence, clique size and maximum degree, *Combinatorica* **4** (1984), 35–38. ⇒136
[54] O. Favoron, A. Hansberg, L. Volkmann, On k-domination and minimum degree in graphs, *J. Graph Theory* **57,** 1 (2008) 33–40. ⇒135, 137
[55] A. Frank, T. Király, Z. Király, On the orientation of graphs and hypergraphs, *Discrete Appl. Math* **131,** 2 (2003) 385–400. ⇒140, 142
[56] P. Frankl, Improved bounds for Erdős matching conjecture, *J. Combin. Theory Ser. A* **120,** 5 (2013) 1068–1072. ⇒141
[57] P. Frankl, T. Luczak, K.Mieczkowska, On matchings in hypergraphs. *Electron. J. Combin.* **19,** 2 (2012), #P42, 5 pages. ⇒141
[58] P. Frankl, V. Rödl, Some Ramsey-Turn type results for hypergraphs. *Combinatorica* **8,** 4 (1988) 323–332. ⇒142
[59] P. Frankl, V. Rödl, The uniformity lemma for hypergraphs, *Graphs Combin.* **8** (1992) 309–312. ⇒142
[60] M. Frieze, On the independence number of random graphs, *Discrete Math.* **81,** 2 (1990) 171–175. ⇒136
[61] Z. Füredi, Matchings and covers in hypergraphs, *Graphs Combin.* **4** (1988) 115–206. ⇒141
[62] Z. Füredi, The number of maximal independent sets in connected graphs, *J. Graph Theory*, **11,** 4 (1995) 463–470. ⇒137
[63] Z. Füredi, Linear trees in uniform hypergraphs, *Europ. J. Combin.* **35** (2014) 264–272. ⇒142
[64] Z. Füredi, M. Ruszinkó, Uniform hypergraphs containing no grids, *Adv. Math.* **240** (2013) 302–324. ⇒142
[65] H. M. Gabow, An efficient implementation of Edmonds' algorithm for maximum matching on graphs, *J. ACM* **23,** 2 (1976) 221–234. ⇒141

[66] H. M. Gabow, Data structures for weighted matchings and nearest nearest common ancestors with linking, in: *Proc. 1st Annual ACM-SIAM Symp. Discrete Algorithms* 1990, 434–443. ⇒141

[67] T. Gallai, Über extreme Punkt- und Kantenmengen, *Ann. Univ. Sci. Budapest. de Rolando Eötvös Nominatae, Sect. Math.* **2** (1959) 133–138. ⇒139

[68] G. Gallo, G. Longo, S. Nguyen, S. Pallottino, Directed hypergraphs and applications, *Discrete Appl. Math* **42,** 2-3 (1993) 177–201. ⇒140, 142

[69] S.. Gaspers, D. Kratsch, M. Liedloff, Independent sets and bicliques in graphs, in: H. Broersma, T. Erlebach, T. Friedetzky, D. Paulusma (eds.) *Graph-Theoretic Concepts in Computer Science* (Int. Workshop WG2008, Durham, UK, June 30, 2008–July 2, 2008), *Lecture Notes in Comput. Sci.* **5344**, Springer, 2008, 171–182. ⇒136

[70] C. Greenhill, The complexity of counting colourings and independent sets in sparse graphs and hypergraphs. *Comput. Complexity* **9,** 1 (2000) 52–72. ⇒141

[71] C. Greenhill, A. Rucinski, N. C. Wormald, Random hypergraph processes with degree restrictions, *Graphs Combin.* **20** (2004) 319–332. ⇒141

[72] J. Griggs, Lower bounds on the independence number in terms of the degrees, *J. Comb. Theory, Ser. B* **34** (1983) 22–39. ⇒134, 145

[73] J. L. Gross, J. Yellen, P. Zhang, *Handbook of Graph Theory*, CRC Press, Boca Raton, FL, 2014. ⇒133, 138

[74] D. Gunopolus, R. Khardon, H. Mannila, H. Toivonen, Data mining, hypergraph traversals, and machine learning. *Proc. PODS'97*, 1997, pp. 209–2016. ⇒142

[75] M. M. Halldórson, Approximations of weighted independent set and hereditary subset problems, *J. Graph Algorithms Appl.* **4,** 1 (2000) 1–16. ⇒136

[76] M. M. Halldórson, E. Losievskaja, Independent sets in bounded-degree hypergraphs, *Discrete Appl. Math* **157,** 8 (2009) 1773–1786. ⇒141

[77] J. Han, Near perfect matchings in k-uniform hypergraphs, arXiv:1404.1136, 2014, 7 pages. ⇒141

[78] H. Hàn, Y. Person, M. Schaht, On perfect matchings in uniform hypergraphs with large minimum vertex degree, *SIAM J. Discrete Math.* **23,** 2 (2009) 732–748. ⇒141, 142

[79] A. Hansberg, R. Pepper, On k-domination and j-independence in graphs. *Discrete Appl. Math* **161,** 10-11 (2013) 1472–1480. ⇒135, 136, 137

[80] J. Harant, A lower bound on independence number of a graph, *Discrete Math.* **188,** 13 (1998) 239243. ⇒134

[81] J. Harant, A lower bound on independence in terms of degrees, *Discrete Appl. Mathematics* **159,** 10 (2011) 966–970. ⇒134, 135

[82] J. Harant, A. Pruchnewski, M. Voigt, On dominating sets and independent sets of graphs, *Combinatorics, Prob. Comp.* **8,** 6 (1999) 547–553. ⇒137

[83] J. Harant, D. Rautenbach, Independence in connected graphs. *Discrete Appl. Math* **159**, 1 (2011) 79–86. ⇒135

[84] J. Harant, I. Schiermeyer, On the independence number of a graph in terms of order and size, *Discrete Math.*, **232,** 1-3 (2001) 131–138. ⇒135

[85] J. Harant, I. Schiermeyer, A lower bound on the independence number of a graph in terms of degrees. *Discuss. Math. Graph Theory* **26,** 3 (2006), 431–437. ⇒135
[86] M. A. Henning, C. Löwenstein, D. Rautenbach, Independent sets and matchings in subcubic graphs. *Discrete Math.*, **312,** 11 (2012) 1900–1910. ⇒141
[87] M. A. Henning, C. Löwenstein, J. Southey, A. Yeo. A new lower bound on the independence number of a graph and applications, *Electron. J. Comb.* **21,** 1 (2014) #P1.38. ⇒136
[88] M. A. Henning, A. Yeo, Tight lower bounds on the size of a maximum matching in a regular graph. *Graphs Combin.* **3,** 6 (2007) 647–657. ⇒141
[89] M. A. Henning, A. Yeo, Lower bounds on the size of maximum independent sets and matchings in hypergraphs of rank three. *J. Graph Theory*, **72,** 1-2 (2013) 220–245. ⇒135, 141, 142
[90] T. Hofmeister, H. Lefmann, Approximating maximum independent sets in uniform hypergraphs. In: *Mathematical Foundations of Computer Science*, 1998, Brno, *Lecture Notes in Comput. Sci.* **1450**, Springer, Berlin, 1998, 562–570. ⇒ 141
[91] J. E. Hopcroft, R. M. Karp, An $\mathfrak{n}^{5/2}$ algorithm for maximum matchings in bipartite graphs. *SIAM J. Comput.*, **2** (1973) 225–231. ⇒141
[92] H. Huang, P.-S. Loh, B. Sudakov, The size of a hypergraph and its matching number, *Combin. Probab. Comput.* **21,** 3 (2012) 442–450. ⇒141, 142
[93] D. S. Johnson, M. Yannakakis, On generating all maximal independent sets, *Inf. Processing Letters* **27,** 3 (1988) 119–123. ⇒133, 138, 141, 142
[94] T. Johnston, L. Lu, Turn problems on non-uniform hypergraphs, arXiv:1301.1870, 2013. ⇒142
[95] T. Johnston, L. Lu, Strong jumps and Lagrangians of non-uniform hypergraphs, arXiv:1403.1220, 2014. ⇒142
[96] B. K. Jose, Zs. Tuza, Hypergraph domination and strong independence. *Appl. Anal. Discrete Math.* **3,** 2 (2009) 347–358. ⇒142
[97] E. Jucovič, F. Olejník, On chromatic and achromatic numbers of uniform hypergraphs, *Časopis Pěstováni Matematiky*, **99** (1974) 123–130. ⇒139, 142
[98] A. Kako, T. Ono, T. Hirata, M. M. Halldórson, Approximation algorithms for the weighted independent set problem in sparse graphs, *Discrete Appl. Math* **157,** 4 (2009) 617–626. ⇒136
[99] M. Karonski, T. Luczak, The number of connected sparsely edged uniform hypergraphs, *Discrete Math.* **171,** 1-3 (1997) 153–167. ⇒142
[100] R.M. Karp, U.V. Vazirani, V.V. Vazirani, *An optimal online algorithm for optimal bipartite matching*, in: Proceedings of the Twenty-second Annual ACM Symposium on Theory of Computing (STOC '90), ACM, New York, NY, 1990, 352–358. ⇒141
[101] R. M. Karp, A. Widgerson, A fast parallel algorithm for the maximal independent set problem, *J. ACM* **32,** 4 (1985) 762–773. ⇒133, 138

[102] G. O. H. Katona, Continuous versions of some extremal hypergraph problems, in: *Combinatorics* (Keszthely, Hungary, 1976), *Coll. Math. Soc. J. Bolyai* **18** (Math. Soc. J. Bolyai, Budapest, 1978), 653–678. ⇒142

[103] G. O. H. Katona, Continuous versions of some extremal hypergraph problems II, *Acta Math. Acad. Sci. Hungar.* **35** (1980) 67–77. ⇒142

[104] P. Keevash, F. Knox, R. Mycroft, Polynomial-time perfect matchings in dense hypergraphs, arXiv:1307.2608, 2013. 62 pages. ⇒141

[105] P. Keevash, R. Mycroft, A geometric theory for hypergraph matching, arXiv:1108.1757, 2013, 101 pages. ⇒141

[106] P. Keevash, B. Sudakov, On a hypergraph Turán problem of Frankl, *Combinatorica* **25,** 6 (2005) 673–706. ⇒142

[107] P. Kelsen, On the parallel complexity of computing a maximal independent set in a hypergraph. In: *Proc. Twenty-fourth Annual ACM Symp. Theory Comp.* (STOC'92), ACM New York, NY, 1992, 339–350. ⇒138, 142

[108] L. Khachiyan, E. Boros, V. Gurvich, K. Elbassioni, Computing many maximal independent sets for hypergraphs in parallel, *Parallel Process. Lett.* **17,** 2 (2007) 141–152. ⇒141

[109] I. Khan, Perfect matchings in 3-uniform hypergraphs with large vertex degree. *SIAM J. Discrete Math.* **27,** 2 (2013) 1021–1039. ⇒141

[110] M. A. Khan, S. , K. K. Kayibi, Scores, inequalities and regular hypertournaments, *Math. Inequal. Appl.* **15,** 2 (2012), 343–351. ⇒139

[111] Y. Kohayakawa, V. Rödl, J. Skokan, Hypergraphs, quasi-randomness, and conditions for regularity, *J. Combin. Theory Ser. A* **97,** 2 (2002) 307–352. ⇒142

[112] V. Kolmogorov, V. Blossom, A new implementation of a minimum cost perfect matching algorithm, *Math. Programming Pomp.* **1** (2009) 43–67. ⇒141

[113] D. Kőnig, Graphs and matrices (Hungarian), *Matematikai és Fizikai Lapok* **38** (1931) 116–119. ⇒141

[114] A. Kostochka, D. Mubayi, J. Verstraëte, On independent sets in hypergraphs, *Random Structures Algorithms* **44,** 2 (2014) 224–239. ⇒142

[115] M. Krivelevich, Approximate set covering in uniform hypergraphs, *J. Algorithms* **25,** 1 (1997) 118–143. ⇒142

[116] M. Krivelevich, R. Nathaniel, B. Sudakov, Approximating coloring and maximum independent sets in 3-uniform hypergraphs, *J. Algorithms* **41,** 1 (2001) 99–113. ⇒142

[117] D. Kühn, D. Loose, Hamilton cycles in 3-uniform hypergraphs of high minimum degree, *J. Comb. Theory, Ser. B* **96,** 6 (2006) 767–821. ⇒142

[118] D. Kühn, D. Osthus, A. Treglown, Matchings in 3-uniform hypergraphs, *J. Combinatorial Theory Series B* **103** (2013) 291–305. ⇒141

[119] D. Kühn, D. Osthus, T. Townsend, Fractional and integer matchings in uniform hypergraphs, *Europ. J. Combin.* **38** (2014) 83–96. ⇒141

[120] R. Laskar, B. Auerbach, On complementary graphs with no isolated vertices, *Discrete Math.* **24,** 2 (1978) 113–118. ⇒140

[121] E. L. Lawler, J. K. Lenstra, A. H. G. Rinnooy Kan, Generating all maximal independent sets: NP-hardness and polynomial-time algorithms, *SIAM J. Comput.* **9,** 3 (1980), 558–565. ⇒133, 138

[122] V. V. Lepin, An algorithm for finding the independence number of a recursively generated hypergraph (Russian), *Dokl. Nats. Akad. Nauk Belarusi* **49,** 2 (2005), 22–25, 125. ⇒141

[123] Y. Li, C. Rousseau, On book-complete graph Ramsey numbers, *J. Comb. Theory, Ser. B* **68,** 1 (1996) 36–44. ⇒142

[124] Y. Li, C. Rousseau, W. Zang, Asymptotic upper bound for Ramsey functions, *Graphs Combin.*, **17** (2001) 123–128. ⇒142

[125] Y. Li, W. Zang, Differential methods for finding independent sets in hypergraphs, *SIAM J. Discrete Math.* **20** (2006) 96–104. ⇒141, 142

[126] E. Losievskaja, *Approximation Algorithms for Independent Set Problems on Hypergraphs*, PhD Dissertation, School of Computer Science Reykjavik University, Reykyavik, 2009, XV + 80 pages. ⇒133, 138, 141

[127] L. Lovász, M. D. Plummer, *Matching Theory*, North Holland, Amsterdam, 1986. ⇒141

[128] M. Luby, A simple parallel algorithm for the maximal independent set problem, *SIAM J. Comput.* **15,** 1 (1976) 1036–1053. ⇒133, 138

[129] T. Luczak, E. Szymańska, A parallel randomized algorithm for finding a maximal independent set in a linear hypergraph, *J. Algorithms* **25,** 2 (1997) 311–320. ⇒142

[130] D. Maier, Minimum covers in the relational data base model, *J. Assoc. Comp. Mach.* **27** (1980) 664–674. ⇒142

[131] K. Markström, A. Ruciński, Perfect matchings (and Hamilton cycles) in hypergraphs with large degrees, *Europ. J. Combin.* **32,** 5 (2011) 677–687. ⇒141

[132] S. Micali, V.V. Vazirani, *An* $O(\sqrt{V}E)$ *algorithm for finding maximum matching in general graphs,* Proc. 21st Annual IEEE Symposium on Foundations of Computer Science (1980) 17–27. ⇒141

[133] K. Mulmuley, U.V. Vazirani, V.V. Vazirani, *Matching is as easy as matrix inversion*, *Combinatorica* **7,** 1 (1987) 105–114. ⇒141

[134] V. Nikiforov, An analytic theory of extremal hypergraph problems, arXiv:1305.1073, 2013, 31 pages. ⇒142

[135] F. Olejník, On total matching numbers and total covering numbers for k-uniform hypergraphs. *Math. Slovaca* **34,** 3 (1984), 319–328. ⇒141

[136] F. Olejník, On edge independence numbers and edge covering numbers of k-uniform hypergraph, *Math. Slovaca* **39,** 1 (1989) 21–26. ⇒138, 139, 140

[137] L. Özkahiya, M. E. Young, Anti-Ramsey number of matchings in hypergraphs, *Discrete Math.*, **313,** 20 (2013), 2359–2364. ⇒141

[138] S. Pirzada, Hypertournaments-scores, losing scores, total scores and degrees. *J. Combin. Math. Combin. Comput.* **84** (2013) 99–112. ⇒139

[139] S. Pirzada, G. Zhou, On k-hypertournament losing scores, *Acta Univ. Sapientiae, Informatica* **2,** 1 (2010) 5–9. ⇒139

[140] S. Pirzada, G. Zhou, A. Iványi, Score lists in multipartite hypertournaments, *Acta Univ. Sapientiae, Informatica* **2,** 2 (2010) 184–193. ⇒139

[141] K. Plociennik, Approximating independent set and coloring in random uniform hypergraphs. In: *Mathematical Foundations of Computer Science* (2008), Lecture Notes in Comput. Sci. **5162**, Springer, Berlin, 2008, 539–550. ⇒141

[142] M. D. Plummer, Matching theory–a sampler: from Dénes König to the present, *Discrete Math.* **100** (1992) 172–219. ⇒141

[143] A. Pruchnewski, On the domination number of graphs, *Discrete Math.*, **251,** 1-3 (2002) 129–136. ⇒137

[144] J. Qian, Enumeration of unlabeled uniform hypergraphs, Electronic J. Combin **20,** 1 (2013), P46, 10 pages ⇒139

[145] J. Qian, Enumeration of unlabeled uniform hypergraphs. *Discrete Math.* **326** (2014), 66–74. ⇒139

[146] V. Rödl, A. Ruciński, M. Schacht, E. Szemerédi, A note on perfect matchings in uniform hypergraphs with large minimum collective degree, *Comment. Math. Univ. Carolin.* **49,** 4 (2008) 633–636. ⇒141

[147] V. Rödl, A. Ruciński, E. Szemerédi, Perfect matchings in uniform hypergraphs with large minimum degree, *Europ. J. Combin.* **27** (2006) 1333–1349. ⇒141

[148] V. Rödl, A. Ruciński, E. Szemerédi, Perfect matchings in large uniform hypergraphs with large minimum collective degree, J. *Combin. Theory* Ser. A **116,** 3 (2009), 613–636. ⇒141

[149] S. Sakai, M. Mitsunori, K. Yamazaki, A note on greedy algorithms for the maximum weighted independent set problem. *Discrete Appl. Math* **126,** 2-3 (2003) 313–322. ⇒136

[150] S. M. Selkow, A probabilistic lower bound on the independence number of graphs. *Discrete Math.* **132** (1994) 363–365. ⇒134, 135

[151] H. Shachnai, A. Srinivasan, Finding large independent sets in graphs and hypergraphs. *SIAM J. Discrete Math.* **18,** 3 (2004/05) 488–500. ⇒141

[152] J. Shearer, A note on the independence number of triangle-free graphs, *Discrete Math.* **46** (1983) 83–87. ⇒134, 136, 141, 145

[153] J. Spencer, Turán's theorem for k-graphs, *Discrete Math.* **2** (1972) 183–186. ⇒136

[154] E. Szymańska, The complexity of 2-coloring and strong coloring in uniform hypergraphs of high minimum degre, *Discrete Math. Theor. Comput. Sci.* **15,** 3 (2013) 121–137. ⇒141, 142

[155] R. E. Tarjan, A. E. Trojanowski, Finding a maximum independent set, *SIAM J. Comput.* **6,** 3 (1977) 537–546. ⇒133, 138, 141

[156] T. Thiele, A lower bound on the independence number of arbitrary hypergraphs, *J. Graph Theory* **30,** 3 (1999) 213–221. ⇒135

[157] A. Treglown, Y. Zhao, Exact minimum degree thresholds for perfect matchings in uniform hypergraphs, *J. Combin. Theory Ser. A* **119,** 7 (2012) 1500–1522. ⇒141, 142

[158] A. Treglown, Y. Zhao, Exact minimum degree thresholds for perfect matchings in uniform hypergraphs II., *J. Combin. Theory Ser. A* **120,** 7 (2013) 1463–1482. ⇒141, 142

[159] P. Turán, On the theory of graphs, *Colloq. Math.* **3** (1954) 19–30. ⇒133, 134

[160] Zs. Tuza, Extensions of Gallai's graph covering theorems for uniform hypergraphs, *J. Combin. Theory, Series B* **52,** 1 (1991) 92–96. ⇒139, 142

[161] V. V. Vazirani, A theory of alternating paths and blossoms for proving correctness of the $O(\sqrt{V}E)$ maximum matching algorithm, *Combinatorica* **14,** 1 (1994) 71–109. ⇒133

[162] V. V. Vazirani, An improved definition of blossoms and a simpler proof of the *MV matching algorithm*, arXiv:1210.4594, 2013, 32 pages. ⇒133

[163] A. Vietri, The complexity of arc-colorings for directed hypergraphs, *Discrete Appl. Math* **143**,1-3 (2004) 266–271. ⇒140

[164] C. Wang, G. Zhou, Note on the degree sequences of k-hypertournaments, Note on the degree sequences of k-hypertournaments, *Discrete Math.* **308,** 11 (2008) 2292–2296. ⇒139

[165] V. Wei, A lower bound on the stability number of a simple graph, *Bell Laboratories Technical Memorandum*, 1981, No. 81-11217-9. ⇒134

[166] E. W. Weisstein, *Independent Edge Set*. Downloaded 9 May 2014. ⇒133

[167] Wikipedia, *Independent Set*. Downloaded 9 May 2014. ⇒133, 138

[168] R. Yuster, Finding and counting cliques and independent sets in r-uniform hypergraphs. *Inf. Processing Letters* **99,** 4 (2006) 130–134. ⇒141

[169] R. Yuster, Maximum matching in regular and almost regular graphs, *Algorithmica* **66,** 1 (2013) 87–92. ⇒141, 142

[170] G. Zhou, Y. Li, Independence numbers of hypergraphs with sparse neighborhoods, *Europ. J. Combin.* **25,** 3 (2004) 355–362. ⇒134, 140, 141, 142

[171] G. Zhou, S. Pirzada, Degree sequence of oriented k-hypergraphs, *J. Appl. Math. Comput.* **27,** 1-2 (2008) 149158. ⇒139

[172] G. Zhou, T. Yao, K. Zhang, On score sequences of k-hypertournaments, *Europ. J. Combin.* **21,** 8 (2000) 993–1000. ⇒139

Received: January 15, 2014 • Revised: May 11, 2014

Acta Universitatis Sapientiae

The scientific journal of Sapientia Hungarian University of Transylvania publishes original papers and surveys in several areas of sciences written in English.
Information about each series can be found at
http://www.acta.sapientia.ro.

Acta Universitatis Sapientiae, Informatica

Each volume contains two issues.

Sapientia University

Scientia Publishing House

ISSN 1844-6086
http://www.acta.sapientia.ro

Information for authors

Acta Universitatis Sapientiae, Informatica publishes original papers and surveys in various fields of Computer Science. All papers are peer-reviewed.

Papers published in current and previous volumes can be found in Portable Document Format (pdf) form at the address: `http://www.acta.sapientia.ro`.

The submitted papers should not be considered for publication by other journals. The corresponding author is responsible for obtaining the permission of coauthors and of the authorities of institutes, if needed, for publication, the Editorial Board is disclaiming any responsibility.

Submission must be made by email (`acta-inf@acta.sapientia.ro`) only, using the LaTeX style and sample file at the address `http://www.acta.sapientia.ro`. Beside the LaTeX source a pdf format of the paper is necessary too.

Prepare your paper carefully, including keywords, ACM Computing Classification System codes (`http://www.acm.org/about/class/1998`) and AMS Mathematics Subject Classification codes (`http://www.ams.org/msc/`).

References should be listed alphabetically based on the Intructions for Authors given at the address `http://www.acta.sapientia.ro`.

Illustrations should be given in Encapsulated Postscript (eps) format.

One issue is offered each author free of charge. No reprints will be available.

Contact address and subscription:
Acta Universitatis Sapientiae, Informatica
RO 400112 Cluj-Napoca
Str. Matei Corvin nr. 4.
Email: acta-inf@acta.sapientia.ro

Printed by Gloria Printing House
Director: Péter Nagy

ISSN 1844-6086
http://www.acta.sapientia.ro