Volume 87, Number 3,
March 2026

ISSN 0005-1179
CODEN: AURCAT

AUTOMATION AND REMOTE CONTROL

Automation and Remote Control

A JOURNAL OF RUSSIAN ACADEMY OF SCIENCES

Vol. 87, No. 3, March 2026

http://ait.mtas.ru

Automation and Remote Control

ISSN 0005-1179

SCOPE

Automation and Remote Control is one of the first journals on control theory. The scope of the journal is control theory problems and applications. The journal publishes reviews, original articles, and short communications (deterministic, stochastic, adaptive, and robust formulations) and its applications (computer control, components and instruments, process control, social and economy control, etc.).

Automation and Remote Control is abstracted and/or indexed in *ACM Digital Library, BFI List, CLOCKSS, CNKI, CNPIEC Current Contents/Engineering, Computing and Technology, DBLP, Dimensions, EBSCO Academic Search, EBSCO Advanced Placement Source, EBSCO Applied Science & Technology Source, EBSCO Computer Science Index, EBSCO Computers & Applied Sciences Complete, EBSCO Discovery Service, EBSCO Engineering Source, EBSCO STM Source, EI Compendex, Google Scholar, INSPEC, Japanese Science and Technology Agency (JST), Journal Citation Reports/Science Edition, Mathematical Reviews, Naver, OCLC WorldCat Discovery Service, Portico, ProQuest Advanced Technologies & Aerospace Database, ProQuest-ExLibris Primo, ProQuest-ExLibris Summon, SCImago, SCOPUS, Science Citation Index, Science Citation Index Expanded (Sci-Search), TD Net Discovery Service, UGC-CARE List (India), WTI Frankfurt eG, zbMATH.*

Journal website: http://ait.mtas.ru

0005-1179/26. *Automation and Remote Control* (ISSN: 0005-1179 print version, ISSN: 1608-3032 electronic version) is published monthly by Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, 65 Profsoyuznaya street, Moscow 117997, Russia.

Volume 87 (12 issues) is published in 2026.

Publisher: Trapeznikov Institute of Control Sciences, Russian Academy of Sciences.

65 Profsoyuznaya street, Moscow 117997, Russia; e-mail: redacsia@ipu.rssi.ru; http://ait.mtas.ru, http://ait-arc.ru

Contents

Automation and Remote Control
Vol. 87, No. 3, 2026

Stochastic Systems

Robust, Adaptive, and Network Control

Control in Technical Systems

Control in Social Economic Systems

Optimization, System Analysis, and Operations Research

Notes, Chronicles, Information

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2026, Vol. 87, No. 3, pp. 195–208.

Counting Triangles in the Generalized Clustering Attachment Model

N. M. Markovich

Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia
e-mail: nat.markovich@gmail.com

Received February 5, 2025
Revised September 12, 2025
Accepted October 31, 2025

Abstract—The clustering attachment model (CAM) was introduced in Bagrow and Brockmann (2013) as an alternative to preferential attachment models. In Markovich and Vaičiulis (2024), a generalized clustering attachment model (GCAM) is proposed that includes an arbitrary attachment function f. This paper investigates the behavior of the mathematical expectation of the number of triangles under weak constraints on the function f. The logarithmic growth of the mathematical expectation of the number of triangles has been proven. More accurate results, including a central limit theorem for the triangle count, are obtained for a constant f. The expectation of the triangle count in the GCAM with an almost constant function f is considered in our simulation study.

Keywords: generalized clustering attachment model, triangle count, clustering coefficient, node weight, random graph, evolution

DOI: 10.7868/S1608303226030013

1. INTRODUCTION

Let $G_n = (V_n, E_n)$, $n = 1, 2, \ldots$ denote the sequence of random graphs, where V_n is the set of nodes and E_n is the set of edges. The sequence G_n, $n = 1, 2, \ldots$ is formed by the evolution, i.e., the dynamic of the graph in time and space. The preferential attachment model (PAM) for the graph evolution has been first introduced in [1, 2]. The majority of the evolution models of the random graphs concerns to the PAM's, see [3–8] among others due to application to numerous real-world graphs. Let us provide here the definition of linear preferential attachment model (LPAM), considered in [9].

Definition 1. Fix $m \geq 2$ and $\delta > -m$. Then the LPAM is a sequence of random graphs $G_n = (V_n, E_n)$, $n \geq 1$ defined as follows:

(i) the graph G_1 consists of a single node with no edges;

(ii) the graph G_2 consists of two nodes with m edges between them;

(iii) the graph G_{n+1}, $n \geq 2$ is constructed recursively as follows: conditioning on the graph G_n, we add a node $n+1$ to the graph, with m new edges. Newly attached edges connect node $n+1$ and nodes $i_1, \ldots, i_m \in V_n$, which are chosen by using weighted sampling with replacement. All nodes of the graph G_n are equipped with the weights

$$p_{i,n}(\delta) = \frac{D_{i,n} + \delta}{2m(n-1) + \delta n}, \quad i \in V_n, \tag{1}$$

which are recalculated after the choice of each of nodes $i_1, \ldots, i_m \in V_n$. Here and below $D_{i,n}$ denotes the degree of node $i \in V_n$.

In [10], the attachment function f is included in the definition of the PAM. Namely, (1) is replaced by

$$p_{i,n}(f,\delta) = \frac{f(D_{i,n}) + \delta}{\sum_{j \in V_n} (f(D_{i,n}) + \delta)}, \quad i \in V_n, \tag{2}$$

where the preferential attachment function f is deterministic, non-decreasing and non-negative.

The clustering attachment model (CAM) was introduced in [11] as an alternative to the PAM, see [11] for comprehensive argument. This model is less known, even though it is often observed in practice and it can be applicable to local networks, e.g., local social networks, and “systems where nodes are drawn not towards hubs, but towards densely connected groups” [11].

Before providing the definition of so-called generalized CAM, let us recall the definition of the clustering coefficient. The clustering coefficient of node $i \in V_n$ is defined as

$$c_{i,n} = \begin{cases} 0, & D_{i,n} = 0 \text{ or } D_{i,n} = 1, \\ 2\Delta_{i,n}/\left(D_{i,n}\left(D_{i,n} - 1\right)\right), & D_{i,n} \geq 2, \end{cases}$$

where $\Delta_{i,n}$ is the number of triangles of node i, [12].

The clustering coefficient $c_{i,n}$ was proposed first in [13] as “the propensity for two neighbors of the same vertex also to be neighbors of one another, forming a triangle of connections in the network.” It measures a capacity of the ith node to form triangles of the nearest nodes in its neighborhood.

Let $\#A$ denote the cardinality (the number of elements) of arbitrary finite set A. The following definition of the generalized CAM (further the GCAM) is a slight modification of one proposed in [14].

Definition 2. Fix $m \geq 2$ and $\epsilon \geq 0$. Then the GCAM is a sequence of random graphs $G_n = (V_n, E_n)$, $n \geq 1$ defined as follows:

(i) G_1 is a non-random, finite and simple graph, which consists of at least m nodes, i.e.,

$$\#V_1 \geq m. \tag{3}$$

(ii) Each node $i \in V_n$, $n \geq 1$ is equipped with the weight

$$p_{i,n}(f,\epsilon) = \frac{f\left(c_{i,n}\right) + \epsilon}{\sum_{j \in V_n} \left(f\left(c_{j,n}\right) + \epsilon\right)}, \tag{4}$$

where the attachment function f is deterministic, non-decreasing, non-negative on $[0,1]$ and $f(1) < \infty$.

(iii) For $n \geq 1$, G_{n+1} is constructed recursively as follows: conditioning the graph G_n, we add a node $\#V_1 + n$ to the graph, with m new edges. The new edges connect the pairs of nodes

$$\{\#V_1 + n, i_1\}, \{\#V_1 + n, i_2\}, \ldots, \{\#V_1 + n, i_m\},$$

where nodes $i_1, \ldots, i_m \in V_n$ are chosen by applying weighted sampling without replacement.

It is easy to see that the CAM in [11], i.e., GCAM with

$$f(x) = x^\alpha, \quad x \geq 0,\ \alpha > 0 \tag{5}$$

is constructed by replacing $D_{i,n}$ with $c_{i,n}$ in the corresponding PA model. The same noting holds by comparing (2) and (4).

Let us provide several remarks related to comparison of PAM and GCAM.

(1) Cardinalities $\#V_n$ and $\#E_n$ are deterministic for both models. In particular, it follows from Definition 2(iii) that for any $n \geq 1$,

$$\#V_n = \#V_1 + n - 1, \quad \#E_n = \#E_1 + m(n-1). \tag{6}$$

As for LPAM, we have $\#V_n = n$ and $\#E_n = m(n-1)$.

(2) From Definition 2(iii) that for any $n \geq 1$, G_n is a simple graph. The situation is different with LPAM. From Definition 1(iii) it follows that multiple edges allowed in LPAM.

(3) It is known that the LPAM with $m \geq 2$ and $\delta > 0$ produces a power law degree distribution with degree exponent $3 + \delta/m$, see, e.g., [9]. For our best knowledge, the degree distribution is not investigated for the GCAM. Using a heuristic argument it is concluded in [11] that the degree distribution for the CAM belongs to the class of light-tailed distributions. The simulation results in [15] confirm this conclusion.

(4) The main difference between LPAM and GCAM is the following. By considering LPAM one can note that it is likely that the node (say i) with large degree in the graph G_n will be connected with newly introduced node $n+1$. This implies $D_{i,n} < D_{i,n+1}$. This means that LPAM exibits a rich-get-richer phenomenon (Matthew effect). The GCAM has not such property. For example, if $c_{i,n} = 1$ for some node $i \in V_n$ and this node is connected with new node $\#V_1 + n$, then $c_{i,n} > c_{i,n+1}$. Indeed, suppose first that the nodes $i \in V_n$ and $j \in V_n$ selected by weighted random choice with replacement are not connected by an edge. Then $D_{i,n+1} = D_{i,n} + 1$ and $\Delta_{i,n+1} = \Delta_{i,n}$ hold. The last equalities together with the equality $2\Delta_{i,n} = D_{i,n}(D_{i,n} - 1)$ allow us to obtain $c_{i,n+1} = (\Delta_{i,n} - 1)/(\Delta_{i,n} + 1) < 1$. It is easy to verify that when nodes $i \in V_n$ and $j \in V_n$ are connected by an edge, then $c_{i,n+1} = (D_{i,n}(D_{i,n} - 1) + 2)/(\Delta_{i,n}(\Delta_{i,n} + 1)) < 1$ due to $D_{i,n} \geq 2$.

The total number of triangles Δ_n in a random graph G_n is defined by

$$\Delta_n = \frac{1}{3} \sum_{i \in V_n} \Delta_{i,n}.$$

Our objective is to investigate the behavior of the sequence $\mathrm{E}(\Delta_n) - \Delta_1$, $n \geq 2$ for the GCAM with $m = 2$ and $\epsilon > 0$. The GCAM with $m = 2$ can be characterized as a basic. Namely, the clustering coefficient of a newly appended node is equal to 1 when a pair of chosen nodes i_1 and i_2 are connected, or 0 otherwise.

The paper is organized as follows. Related results regarding the expected number of triangles are surveyed in Section 2. Main results including Theorems 1 and 2 are presented in Section 3. The algorithm and simulation results of the sequence $\mathrm{E}\Delta_n - \Delta_1$, $n \geq 2$ are considered in Section 4. Conclusions are given in Section 5. Proofs are provided in Appendix.

2. RELATED RESULTS FOR EXPECTED NUMBER OF TRIANGLES

A triangle of connected nodes is the most studied subgraph and it can be considered as the basic community. Perhaps, it was first proposed for study in [13].

In [16] it is shown that $\mathrm{E}(\Delta_n)$ in the Albert-Barabási model (the LPAM with $\delta = 0$, $m \geq 2$) is of order $\ln^3(n)$. The same order is obtained in [17] for the PAM, where all m edges from the next new vertex are drawn simultaneously and independently, and the probabilities of drawing an edge to vertex i depend linearly on the degree of this vertex in G_n, but are not recalculated after drawing the next edge, as in the definition of 1.

$\mathrm{E}(\Delta_n)$ of LPAM with two parameters δ, m was investigated in [18], see also [19]. In [18] it is shown that for $\delta > 0$ and $m \geq 2$,

$$|\mathrm{E}(\Delta_n) - C_1(\beta, m)\ln(n)| \leq C_2, \quad n \geq 2$$

holds, where C_1 and C_2 are some positive constants. More general, Theorem 2.5 in [9] states that for the LPAM parameters $m \in \mathbb{N} \geq 2$ and $\delta > -m$,

$$\mathrm{E}\left(\Delta_n\right) = C_1(m, \delta)g_1(n) + g_2(n), \quad n \to \infty$$

holds, where the function g_2 is such that $|g_2(n)| \to 0$ as $n \to \infty$, while the function g_1 has the form

$$g_1(n) = \begin{cases} n^{-\delta/(2m+\delta)}, & -m < \delta < 0, \\ \ln^3(n), & \delta = 0, \\ \ln(n), & \delta > 0. \end{cases}$$

The explicit form of the asymptotic constant $C_1(m, \delta)$ is given in Theorem 2.5 of in [9].

Let us list articles containing results on the number of triangles Δ_n. The asymptotic behavior of the number of triangles Δ_n in an inhomogeneous random graph with different connection probabilities to form edges is derived in [20, 21], and a short survey can be found in [22]. It was shown that Δ_n converges in distribution to a Poisson distribution. The asymptotic expected number of Δ_n in power-law uniform random graphs is studied in [23]. The survey of the literature regarding the triangle count Δ_n for homogeneous random graphs is given in [21].

3. MAIN RESULTS

The main result of this work is the following theorem.

Theorem 1. *Let $m = 2$ and $\epsilon > 0$ be parameters of the GCAM. Then for $n \geq 2$,*

$$\frac{1}{\mu^2} + g_1(n) \leq \frac{\mathrm{E}(\Delta_n) - \Delta_1}{4\ln(n)} \leq \mu^2 + g_2(n), \tag{7}$$

where

$$\mu = \frac{\epsilon + f(1)}{\epsilon + f(0)} \tag{8}$$

and $g_i(n) = O\left(1/\ln(n)\right)$, $n \to \infty$, $i = 1, 2$.

Without loss of generality we may assume that $f(1) = 1$ in (4). Indeed, if $f(1) \neq 1$, then $p_{i,n}(f, \epsilon) = p_{i,n}(f/f(1), \epsilon/f(1))$. By considering particular case of the attachment function f we have the following result.

Theorem 2. *Let $m = 2$ and $\epsilon > 0$ be parameters of the GCAM and f is the constant function, namely,*

$$f(x) = \lambda, \quad 0 \leq x \leq 1, \quad \lambda \geq 0. \tag{9}$$

Then

$$\lim_{n\to\infty} \frac{\mathrm{E}(\Delta_n) - \Delta_1}{4\ln(n)} = 1 \tag{10}$$

and

$$\frac{\Delta_n - 4\ln(n)}{2\ln^{1/2}(n)} \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, 1\right), \quad n \to \infty, \tag{11}$$

where $\xrightarrow{\mathrm{d}}$ denotes the convergence in distribution and $\mathcal{N}(0, 1)$ denotes the standard normal distribution.

The statement (11) is the central limit theorem (CLT). The author has not encountered an analogue for a LMAP. A review of the literature on local CLT applied to Δ_n in Erdös-Rényi random graphs is contained in [24].

4. MONTE-CARLO SIMULATIONS

The classical approach to analyze the total triangle count is to simulate a "long enough" sequence of graphs $G_1, \ldots, G_n$ and then build a sample $\Delta_1, \ldots, \Delta_n$. However, this procedure may require a long simulation time. To overcome the problem, we present an approximation for $\mathrm{E}(\Delta_n) - \Delta_1$, $n \geq 2$, based on theoretical calculations presented in the Appendix, to which several references will be made below.

Let $U_k^{(j)}$, $1 \leq j \leq N$, $1 \leq k \leq n-1$ be independent and uniformly distributed random variables (r.v.) on $(0,1)$, and let the random event B_k be defined in (A.4). Then

$$\overline{D_{n,N}} = \frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{n-1} I\left\{U_k^{(j)} < \mathrm{P}(B_k)\right\}, \quad n \geq 2, \tag{12}$$

is an approximation for $\mathrm{E}(\Delta_n) - \Delta_1$. Indeed, using (A.6) and (A.7), we get $\mathrm{E}\left(\overline{D_{n,N}}\right) = \mathrm{E}(\Delta_n) - \Delta_1$. Moreover, for any fixed n the sequence $\overline{D_{n,N}}$, $N = 1, 2, \ldots$ converges almost everywhere to $\mathrm{E}(\Delta_n) - \Delta_1$ as $N \to \infty$. This follows from the strong law of large numbers (see, e.g., Theorem 14 in [28]), since the terms in the sum (12) by j are independent and identically distributed r.v.s with expectation $\mathrm{E}(\Delta_n) - \Delta_1$.

Let $\tilde{n} \geq 2$ denote some natural number. Using (12), one can quickly generate an approximation of the sequence $\mathrm{E}(\Delta_n) - \Delta_1$, $2 \leq n \leq \tilde{n}$. To this end, we need to calculate probabilities $\mathrm{P}(B_k)$, $1 \leq k \leq \tilde{n} - 1$. In the simplest case, when f is a constant function (see, (9)), we have

$$\mathrm{P}(B_k) = \frac{2(\#E_1 + 2(k-1))}{(\#V_1 + k - 1)(\#V_1 + k - 2)}, \quad k \geq 1,$$

see (A.13). Calculating the probabilities $\mathrm{P}(B_k)$, $1 \leq k \leq \tilde{n} - 1$ becomes significantly more complicated when replacing a constant function with an almost everywhere constant (in terms of Lebesgue measure) function

$$f(x) = \begin{cases} 0, & x = 0, \\ 1, & 0 < x \leq 1. \end{cases} \tag{13}$$

For the function f presented in (13), weights (4) have the form

$$p_{i,n}(f, \epsilon) = \frac{1}{(1+\epsilon)V_{n,1} + \epsilon V_{n,2}} \begin{cases} \epsilon, & i \in V_{n,2}, \\ 1+\epsilon, & i \in V_{n,1}, \end{cases} \tag{14}$$

where $V_{n,1} = \{i : c_{i,n} > 0\}$, $V_{n,2} = V_n \setminus V_{n,1}$. Let us also decompose the set of edges E_n into disjoint subsets. Let $E_{n,1} = \{\{i,j\} : \{i,j\} \in E_n,\ i \in V_{n,1},\ j \in V_{n,1}\}$, $E_{n,3} = \{\{i,j\} : \{i,j\} \in E_n,\ i \in V_{n,2},$ $j \in V_{n,2}\}$ and $E_{n,2} = E_n \setminus (E_{n,1} \cup E_{n,3})$ hold. For example, if $G_1 = (V_1, E_1)$ with

$$V_1 = \{1, 2, \ldots, 9\}, \quad E_1 = \{\{1,2\}, \{1,3\}, \{2,3\}, \{2,5\}, \{3,4\}, \{6,7\}\}, \tag{15}$$

then $V_{1,1} = \{1,2,3\}$, $V_{1,2} = \{4, 5, \ldots, 9\}$, $E_{1,1} = \{\{1,2\}, \{1,3\}, \{2,3\}\}$, $E_{1,2} = \{\{2,5\}, \{3,4\}\}$, $E_{1,3} = \{\{6,7\}\}$.

Taking into account the introduced notations (A.7) can be rewritten as follows:

$$\mathrm{P}(B_k) = q_{k,1} + q_{k,2} + q_{k,3}, \quad k \geq 1,$$

where

$$q_{k,\ell} = \sum_{\{i_1,i_2\}\in E_{k,\ell}} \left(\frac{p_{i_1,k}(f,\epsilon)p_{i_2,k}(f,\epsilon)}{1-p_{i_1,k}(f,\epsilon)} + \frac{p_{i_1,k}(f,\epsilon)p_{i_2,k}(f,\epsilon)}{1-p_{i_2,k}(f,\epsilon)} \right), \quad \ell = 1,2,3.$$

Here and further, we assume that the initial graph G_1 and the parameter ϵ are chosen so that $\mu < \#V_1$. Using (14), we find

$$q_{k,1} = \frac{2(1+\epsilon)^2 \#E_{k,1}}{\lambda_k\left(\lambda_k - 1 - \epsilon\right)}, \quad q_{k,2} = \frac{\epsilon(1+\epsilon)\#E_{k,2}}{\lambda_k\left(\lambda_k - 1 - \epsilon\right)} + \frac{\epsilon(1+\epsilon)\#E_{k,2}}{\lambda_k\left(\lambda_k - \epsilon\right)}, \quad q_{k,3} = \frac{2\epsilon^2 \#E_{k,3}}{\lambda_k\left(\lambda_k - \epsilon\right)},$$

where $\lambda_k = (1+\epsilon)\#V_{k,1} + \epsilon\#V_{k,2}$. In addition, we introduce the quantities

$$\begin{aligned}
\bar{q}_{k,1} &= \frac{2(1+\epsilon)^2\left(\#V_{k,1}(\#V_{k,1}-1)/2 - \#E_{k,1}\right)}{\lambda_k\left(\lambda_k - 1 - \epsilon\right)}, \\
\bar{q}_{k,2} &= \frac{\epsilon(1+\epsilon)\left(\#V_{k,1}\#V_{k,2} - \#E_{k,2}\right)}{\lambda_k\left(\lambda_k - 1 - \epsilon\right)} + \frac{\epsilon(1+\epsilon)\left(\#V_{k,1}\#V_{k,2} - \#E_{k,2}\right)}{\lambda_k\left(\lambda_k - \epsilon\right)}, \\
\bar{q}_{k,3} &= \frac{2\epsilon^2\left(\#V_{k,2}(\#V_{k,2}-1)/2 - \#E_{k,3}\right)}{\lambda_k\left(\lambda_k - \epsilon\right)}.
\end{aligned}$$

It is easy to see that $q_{k,1}$ is the probability of choosing an unordered pair of nodes connected by an edge from $E_{k,1}$, and $\bar{q}_{k,1}$ is the probability of choosing an unconnected unordered pair of nodes such that both nodes belong to $V_{k,1}$. The quantities $q_{k,\ell}$ and $\bar{q}_{k,\ell}$, $\ell = 2,3$ can be interpreted similarly.

Let us recall that to form the graph G_{k+1}, a new node $\#V_1 + k$ is added to the graph G_k. Depending on the link between the unordered pair of nodes selected using weighted random selection without deletion, the newly introduced node is added to the set $V_{k+1,1}$ or to the set $V_{k+1,2}$. We also add two edges connected to the node $\#V_1 + k$ to one of the disjoint sets $E_{k+1,1}$, $E_{k+1,2}$, or $E_{k+1,3}$. In particular, the cardinalities of the sets $\#V_{k+1,\ell}$, $\ell = 1,2$, $k \geq 1$ and $\#E_{k+1,\ell}$, $\ell = 1,2,3$, $k \geq 1$ are controlled by the following rules:

1. If the event $\{U_k^{(j)} < q_{k,1}\}$ occurs (i.e. both selected nodes belong to $V_{k,1}$ and they are connected by an edge from $E_{k,1}$), then $\#V_{k+1,1} = \#V_{k,1} + 1$ and $\#E_{k+1,1} = \#E_{k,1} + 2$. If $\{q_{k,1} + q_{k,2} + q_{k,3} \leq U_k^{(j)} < q_{k,1} + q_{k,2} + q_{k,3} + \bar{q}_{k,1}\}$ (i.e., both selected nodes belong to $V_{k,1}$ and are not connected), then $\#V_{k+1,2} = \#V_{k,2} + 1$ and $\#E_{k+1,2} = \#E_{k,2} + 2$.
2. If the event $\{q_{k,1} \leq U_k^{(j)} < q_{k,1} + q_{k,2}\}$ occurs (i.e. one of the selected nodes belongs to $V_{k,1}$, and another one belongs to the set $V_{k,2}$ and the nodes are connected by the edge from $E_{k,2}$), then $\#V_{k+1,1} = \#V_{k,1} + 2$, $\#V_{k+1,2} = \#V_{k,2} - 1$, $\#E_{k+1,1} = \#E_{k,1} + 3$ and $\#E_{k+1,2} = \#E_{k,2} - 1$. If the event $\{q_{k,1} + q_{k,2} + q_{k,3} + \bar{q}_{k,1} \leq U_k^{(j)} < q_{k,1} + q_{k,2} + q_{k,3} + \bar{q}_{k,1} + \bar{q}_{k,2}\}$ occurs (i.e. one of the selected nodes belongs to $V_{k,1}$, and another one belongs to the set $V_{k,2}$, and nodes are not connected), then $\#V_{k+1,2} = \#V_{k,2} + 1$, $\#E_{k+1,2} = \#E_{k,2} + 1$ and $\#E_{k+1,3} = \#E_{k,3} + 1$.
3. If the event $\{q_{k,1} + q_{k,2} \leq U_k^{(j)} < q_{k,1} + q_{k,2} + q_{k,3}\}$ occurs (i.e. both of the selected nodes belong to $V_{k,2}$ and they are connected by the edge from $E_{k,3}$), then $\#V_{k+1,1} = \#V_{k,1} + 3$, $\#V_{k+1,2} = \#V_{k,2} - 2$, $\#E_{k+1,1} = \#E_{k,1} + 3$ and $\#E_{k+1,3} = \#E_{k,3} - 1$. If the event $\{U_k^{(j)} \geq 1 - \bar{q}_{k,3}\}$ occurs (i.e. both of selected nodes belong to $V_{k,2}$ and they are disconnected), then $\#V_{k+1,2} = \#V_{k,2} + 1$ and $\#E_{k+1,3} = \#E_{k,3} + 2$.

The following initial graphs were used in the simulation $G_1 = (V_1, E_1)$:

a) a complete graph $G_1^{(1)}$ with three nodes and three edges;

b) the graph $G_1^{(2)} = (V_1, E_1)$, where V_1 and E_1 are determined in (15);

c) a complete graph $G_1^{(3)}$ with 17 nodes and 136 edges.

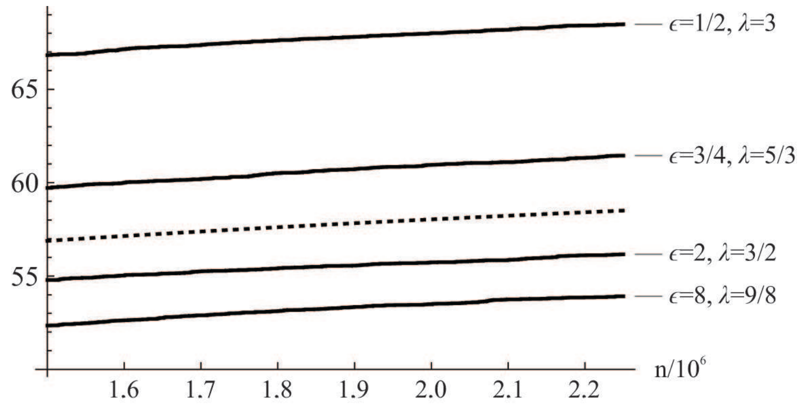

Fig. 1. The graph $g(n) = 4\ln(n)$ (dotted line) and $\overline{D_{n,100}}$, $1.5 \times 10^6 \leq n \leq 2.25 \times 10^6$ for initial graph $G_1^{(1)}$.

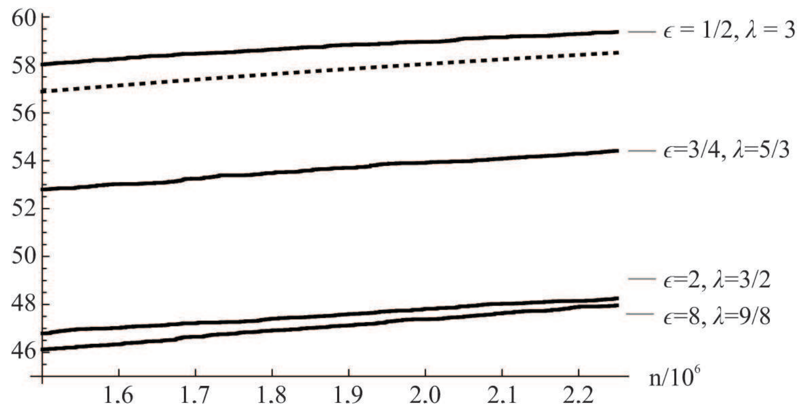

Fig. 2. The graph $g(n) = 4\ln(n)$ (dotted line) and $\overline{D_{n,100}}$, $1.5 \times 10^6 \leq n \leq 2.25 \times 10^6$ for initial graph $G_1^{(2)}$.

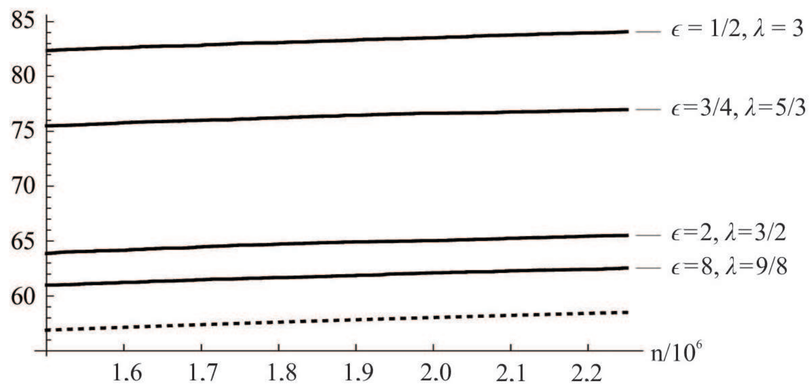

Fig. 3. The graph $g(n) = 4\ln(n)$ (dotted line) and $\overline{D_{n,100}}$, $1.5 \times 10^6 \leq n \leq 2.25 \times 10^6$ for initial graph $G_1^{(3)}$.

In the simulation, the difference $\mathrm{E}(\Delta_n) - \Delta_1$ is approximated by an empirical mean $\overline{D_{N,n}}$. The values $\overline{D_{n,N}}$ are calculated based on (12) using $N = 100$ independent sequences of triangle numbers. The approximation values $\overline{D_{n,100}}$, $2 \leq n \leq 2.25 \times 10^6$ are calculated for $\epsilon = 1/2.3/4.2.8$ and for each initial graph. The simulation results are presented in Figs. 1–3, the smoothness of the curves of which indicates that the chosen value of N is sufficiently large. To show that $\overline{D_{n,100}}$, as a function of n, varies slowly for large n, the simulation results are presented on the interval of natural numbers $[1.5 \times 10^6, 2.25 \times 10^6]$.

Relation (10) states that for $\lambda = 1$, the mathematical expectation of the number of triangles $\mathrm{E}(\Delta_n) - \Delta_1$ is asymptotically equal to $g(n) = 4\ln(n)$ as $n \to \infty$. The graph of the function g is also shown in Figs. 1–3.

We see that the graphs $\overline{D_{n,100}}$ in Figs. 1–3 are approximately parallel to $g(n) = 4\ln(n)$. Therefore, we can assume that for each initial graph G_1 and $\epsilon > 0$, there exists a positive parameter ν, probably depending on G_1 and ϵ, such that the mathematical expectation $\mathrm{E}(\Delta_n) - \Delta_1$ is asymptotically equal to $\nu\ln(n)$ as $n \to \infty$.

The simulation results are as follows:

(1) the graphs $\overline{D_{n,100}}$ in Figs. 1–3 allow us to conclude that ν depends on the parameter ϵ, which is also confirmed by the estimates of the parameter $\hat{\nu}$ obtained by the least squares method and presented in table;

Empirical characteristics $\bar{D} = \overline{D_{2.5\times10^6,100}}$ and $\hat{\nu}$

	Initial graph											
	$G_1^{(1)}$				$G_1^{(2)}$				$G_1^{(3)}$			
ϵ	1/2	3/4	2	8	1/2	3/4	2	8	1/2	3/4	2	8
$\overline{D}$	68	61	56	54	59	54	48	48	84	77	66	62
$\hat{\nu}$	4.69	4.20	3.85	3.69	4.07	3.71	3.29	3.26	5.77	5.29	4.49	4.28

(2) the table shows that $\bar{D} = \overline{D_{2.5\times10^6,100}}$ and ν depend on the assignment of the initial graph;

(3) for all approximations $\overline{D_{2.5\times10^6,100}}$ it holds

$$\lambda^{-2} \leq \frac{\overline{D_{2.5\times10^6,100}}}{4\ln(2.5\times10^6)} \leq \lambda^2, \quad \lambda = \frac{1+\epsilon}{\epsilon},$$

that does not contradict the theoretical result (7);

(4) the approximation (12) can be applied for the next attachment function as well:

$$f(x) = \begin{cases} 0, & 0 \leq x < 1, \\ 1, & x = 1. \end{cases}$$

When designing an algorithm to compute $\mathrm{P}(B_k)$, $1 \leq k \leq n-1$, in this case, it is important to keep in mind that $c_{i,n} = 1$ if and only if i is a vertex of the complete graph.

5. CONCLUSIONS

The expected number of node triangles for the GCAM with $m = 2$, any $\epsilon > 0$ and arbitrary attachment function f satisfying Definition 2 is investigated.

The function f does not have to be continuous in Definition 2(ii), and therefore, in Theorem 1. This function may be discontinuous, with one or several removable discontinuities.

Regarding the initial graph $G_1 = (V_1, E_1)$ from which the evolutions begins, Theorem 1 is valid for any set E_1 including the empty one. The technical assumption (3) with $m = 2$ ensures that each graph G_n, $n \geq 1$ contains at least two nodes, which may be chosen by applying the weighted sampling without replacement. The expected number of triangles is proved to be of the rate $\ln(n)$. It is noted in Section 2, the number of triangles in the linear PAM with $m \geq 2$ and $\delta > 0$ has the same rate $\ln(n)$, see [18], [9].

A fast algorithm to generate a long sequence $\mathrm{E}\Delta_n - \Delta_1$, $n \geq 2$ for the GCAM with the attachment function (13) is provided. Our simulation results validate theoretical result in Theorem 1.

The weights $p_{i,n}(f,\epsilon)$, $i \in V_n$ in (4) are uniformly distributed on V_n if the attachment function is constant. This allowed us to find the asymptotic of $\mathrm{E}\Delta_n - \Delta_1$ and prove the asymptotic normality, see (10) and (11), respectively, in Theorem 2.

The case $m > 2$ requires more calculations due to a complicated combinatorics. As is noted in [11], to investigate the growth of the triangle count by using simulations is a challenge. One can suppose that the expected number of triangles has the rate $\ln(n)$ in the case $m > 2$, too.

Having information about the behavior of the sequence $\mathrm{E}\Delta_n$, $n \geq 2$, one can investigate the behavior of the average clustering coefficient

$$\bar{C}_n = \frac{3\Delta_n}{\text{the number of pairs of connected edges in a graph} G_n}.$$

This is a subject for further research.

FUNDING

This work was supported in part by the Russian Foundation for Basic Research, project no. 24-21-00183.

APPENDIX

We will provide proofs for the statements here.
The digamma function ψ is defined as

$$\psi(x) = \frac{\mathrm{d}}{\mathrm{d}x} \ln\left(\Gamma(x)\right), \quad x > 0, \tag{A.1}$$

where $\Gamma(x) = \int_0^\infty y^{x-1} \mathrm{e}^{-y} \mathrm{d}y$ is the Euler's gamma function, see, e.g., [25]. Let us define

$$\begin{aligned}\Upsilon_n(\mu, \#V_1, \#E_1) = {} & 4\mu^2 \psi(n + \#V_1 - 1 - \mu) \\ & + 2\mu(2\#V_1 - \#E_1)\left(\psi(n + \#V_1 - 1) - \psi(n + \#V_1 - 1 - \mu)\right),\end{aligned}$$

where μ is the same as in (8), initial graph $G_1 = (V_1, E_1)$ satisfies Definition 2(i) and natural n is such that $n \geq n_0$, where natural n_0 satisfies $n_0 + \#V_1 - 1 - \mu > 0$.

Lemma 1. *Let $\#V_k$, $k \geq 1$ and $\#E_k$, $k \geq 1$ be the sequences defined in* (6) *and β be a real positive number, while natural numbers $1 \leq n' < n'' < \infty$ are such that $n' + \#V_1 - 1 - \beta > 0$. Then*

$$2\beta^2 \sum_{k=n'}^{n''-1} \frac{\#E_k}{\#V_k(\#V_k - \beta)} = \Upsilon_{n''}(\beta, \#V_1, \#E_1) - \Upsilon_{n'}(\beta, \#V_1, \#E_1). \tag{A.2}$$

Proof of Lemma 1. The left-hand side of (A.2) is equal to the sum

$$2\beta(\#E_1 - 2\#V_1) \sum_{k=n'}^{n''-1} \frac{1}{k + \#V_1 - 1} + 2\beta\left(2\beta - (\#E_1 - 2\#V_1)\right) \sum_{k=n'}^{n''-1} \frac{1}{k + \#V_1 - 1 - \beta}.$$

The equality (A.2) follows by applying the identity

$$\sum_{k=n'}^{n''-1} \frac{1}{k+\gamma} = \psi(n'' + \gamma) - \psi(n' + \gamma), \quad n' + \gamma > 0. \tag{A.3}$$

To prove (A.3) we decompose the left-hand side of (A.3) as following:

$$\sum_{k=n'}^{\infty}\frac{1}{k+\gamma}-\sum_{k=n''}^{\infty}\frac{1}{k+\gamma} = -\left\{-\gamma_E+\sum_{j=0}^{\infty}\frac{1}{j+1}-\sum_{j=0}^{\infty}\frac{1}{j+n'+\gamma}\right\} + \left\{-\gamma_E+\sum_{j=0}^{\infty}\frac{1}{j+1}-\sum_{j=0}^{\infty}\frac{1}{j+n''+\gamma}\right\},$$

where $\gamma_E \approx 0.5772$ is the Euler–Mascheroni's constant, and use the series expansion

$$\psi(x) = -\gamma_E+\sum_{j=0}^{\infty}\frac{1}{j+1}-\sum_{j=0}^{\infty}\frac{1}{j+x}, \quad x>0,$$

(see, e.g., [25]).

Lemma 2. *Let β be a real positive number, while natural numbers $1 \le n' < n < \infty$ are such that $n' + \#V_1 - 1 - \beta > 0$. Then*

$$\Upsilon_n(\beta, \#V_1, \#E_1) - 4\beta^2 \ln(n) = g(n)$$

holds, where $g(n) = O\left(1/n\right)$, as $n \to \infty$.

Proof of Lemma 2. We can write $\Upsilon_n(\beta, \#V_1, \#E_1) - 4\beta^2 \ln(n) = g(n)$, where

$$\begin{aligned} g(n) &= 2\beta\left(2\beta - (2\#V_1 - \#E_1)\right)\left(\psi(n+\#V_1-1-\beta) - \ln(n+\#V_1-1-\beta)\right) \\ &+ 2\beta(2\#V_1-\#E_1)\left(\psi(n+\#V_1-1) - \ln(n+\#V_1-1)\right) \\ &+ 4\beta^2 \ln\left(1+\frac{\#V_1-1-\beta}{n}\right) + 2\beta(2\#V_1-\#E_1)\ln\left(1+\frac{\beta}{n+\#V_1-1-\beta}\right). \end{aligned}$$

By using inequalities $x/(1+x) \le \ln(1+x) \le x$, $x > -1$ and $|\psi(x) - \ln(x)| < 1/x$, $x > 0$ (see, e.g., [26]), we find that $|g(n)| \le C/n$, where $0 < C < \infty$ is some constant.

Proof of Theorem 1. Obviously, that two random events $\{\Delta_{n+1} = \Delta_n + 1\}$ and

$$B_n = \{\text{an unordered pair of nodes chosen (by weighted sampling without replacement) from } V_n \text{ is connected by an edge from } E_n\} \tag{A.4}$$

are equivalent. Thus, we get

$$\Delta_n - \Delta_1 = \sum_{k=1}^{n-1} I(B_k), \quad n \ge 2, \tag{A.5}$$

where $I(B_k)$ denotes the indicator of the event B_k. Whence, keeping in mind that the graph G_1 is non-random, we get

$$\mathrm{E}\left(\Delta_n\right) - \Delta_1 = \sum_{k=1}^{n-1} \mathrm{P}(B_k), \quad n \ge 2. \tag{A.6}$$

By applying the weighted sampling without replacement, the first node of V_k is randomly selected, using weights $p_{i,k}(f,\epsilon)$, $i \in V_k$, $k \ge 1$. The same node, say i', cannot be selected again. The second node is chosen using renormalized weights $p_{i,k}(f,\epsilon)/\left(1 - p_{i',k}(f,\epsilon)\right)$, $i \in V_k \setminus \{i'\}$. If $\#E_k = 0$, then it immediately follows that $\mathrm{P}(B_k) = 0$. As for the case $\#E_k \ne 0$, we have

$$\mathrm{P}(B_k) = \sum_{\{i_1,i_2\}\in E_k}\left(\frac{p_{i_1,k}(f,\epsilon)p_{i_2,k}(f,\epsilon)}{1-p_{i_1,k}(f,\epsilon)} + \frac{p_{i_1,k}(f,\epsilon)p_{i_2,k}(f,\epsilon)}{1-p_{i_2,k}(f,\epsilon)}\right), \ k \ge 1. \tag{A.7}$$

Here, the sum is taken over all unordered pairs of nodes $\{i_1, i_2\}$ belonging to the set of edges E_k. The assumption $\epsilon > 0$ yields $0 < p_{i,k}(f, \epsilon) < 1$ for any $i \in V_k$, $k \geq 1$. Thus, all summands on the right-hand side of (A.7) are positive.

Under the adopted assumptions (see Definition 2 (ii)) the function f does not decrease on $[0, 1]$, i.e., $f(0) \leq f(x) \leq f(1)$ for all $0 \leq x \leq 1$. Due to gross inequality $0 \leq c_{i,k} \leq 1$, $i \in V_k$, $k \geq 1$, this leads to $f(0) \leq f(c_{i,k}) \leq f(1)$, $i \in V_k$, $k \geq 1$. From here follow the inequalities for the weights (4):

$$\frac{1}{\mu \# V_k} \leq p_{i,k}(f, \epsilon) \leq \frac{\mu}{\# V_k}, \quad i \in V_k,\ k \geq 1. \tag{A.8}$$

Let us consider the upper bound of $\mathrm{E}\,(\Delta_n) - \Delta_1$. Assume first that $\# V_1 - \mu > 0$. By combining (6), (A.7), (A.8) we get

$$\mathrm{P}(B_k) \leq 2\mu^2 \frac{\# E_k}{\# V_k (\# V_k - \mu)} \tag{A.9}$$

for $k \geq 1$. This, together with (A.6) gives

$$\mathrm{E}\,(\Delta_n) - \Delta_1 \leq 2\mu^2 \sum_{k=1}^{n-1} \frac{\# E_k}{\# V_k (\# V_k - \mu)}.$$

By applying Lemma 1 we obtain

$$\begin{aligned} \mathrm{E}\,(\Delta_n) - \Delta_1 &\leq \Upsilon_n(\mu, \# V_1, \# E_1) - \Upsilon_1(\mu, \# V_1, \# E_1) \\ &\leq 4\mu^2 \ln(n) + \left| \Upsilon_n(\mu, \# V_1, \# E_1) - 4\mu^2 \ln(n) \right| + \left| \Upsilon_1(\mu, \# V_1, \# E_1) \right|. \end{aligned}$$

Hence, it follows

$$\frac{\mathrm{E}\,(\Delta_n) - \Delta_1}{4 \ln(n)} \leq \mu^2 + \frac{\left| \Upsilon_n(\mu, \# V_1, \# E_1) - 4\mu^2 \ln(n) \right|}{4 \ln(n)} + \frac{\left| \Upsilon_1(\mu, \# V_1, \# E_1) \right|}{4 \ln(n)}.$$

Now, by Lemma 2, it follows

$$\frac{\mathrm{E}\,(\Delta_n) - \Delta_1}{4 \ln(n)} \leq \mu^2 + g_2(n), \tag{A.10}$$

where $g_2(n) = O\left(1/\ln(n)\right)$, $n \to \infty$.

Assume now $\# V_1 - \mu \leq 0$. Then there exists natural k_0, such that $\# V_k - \mu \leq 0$ for $1 \leq k \leq k_0$ and $\# V_k - \mu > 0$ for $k > k_0$. We decompose the sum in right-hand side of (A.6) as follows:

$$\mathrm{E}\,(\Delta_n) - \Delta_1 = \sum_{k=1}^{k_0} \mathrm{P}(B_k) + \sum_{k=k_0+1}^{n-1} \mathrm{P}(B_k). \tag{A.11}$$

By using a rough bound $\mathrm{P}(B_k) \leq 1$ we get that first sum in right-hand side of (A.11) does not exceed k_0. To find an upper bound of the second sum, we apply the bound (A.9), which holds for $k > k_0$. Thus, we have

$$\mathrm{E}\,(\Delta_n) - \Delta_1 \leq k_0 + \Upsilon_n(\mu, \# V_1, \# E_1) - \Upsilon_{k_0+1}(\mu, \# V_1, \# E_1).$$

It remains to apply Lemma 2 again, to ensure that (A.10) is satisfied.

A similar reasoning can be applied to obtain the lower bound of $(\mathrm{E}\,(\Delta_n) - \Delta_1) / (4 \ln(n))$.

Proof of Theorem 2. By substituting (9) into (4) we get that for any $\epsilon > 0$,

$$p_{i,k}(f, \epsilon) = 1/\#V_k, \quad i \in V_k,\ k \geq 1. \tag{A.12}$$

By combining (A.6), (A.7) and (A.12) we derive

$$\mathrm{E}\left(\Delta_n\right) - \Delta_1 = 2\sum_{k=1}^{n-1} \frac{\#E_k}{\#V_k(\#V_k - 1)},\ n \geq 2, \tag{A.13}$$

and hence,

$$\mathrm{E}\left(\Delta_n\right) - \Delta_1 = \Upsilon_n(1, \#V_1, \#E_1) - \Upsilon_1(1, \#V_1, \#E_1),\ n \geq 2. \tag{A.14}$$

Now (10) follows by applying Lemma 2 with $\beta = 1$.

Let us prove (11). We have

$$\frac{\Delta_n - 4\ln(n)}{2\ln^{1/2}(n)} = \frac{\mathrm{E}(\Delta_n) - \Delta_1 - 4\ln(n)}{2\ln^{1/2}(n)} + \frac{\Delta_1}{2\ln^{1/2}(n)} + \frac{\Delta_n - \mathrm{E}(\Delta_n)}{2\ln^{1/2}(n)}. \tag{A.15}$$

Using (A.14), it is easy to see that the first term on the right-hand side of (A.15) is equal to the ratio

$$\frac{\{\Upsilon_n(1, \#V_1, \#E_1) - 4\ln(n)\} - \Upsilon_1(1, \#V_1, \#E_1)}{2\ln^{1/2}(n)},$$

which by Lemma 2 tends to 0 as $n \to \infty$. Evidently, the second term in the right-hand side of (A.15) tends to 0 as $n \to \infty$, too. It remains to prove that

$$\frac{\Delta_n - \mathrm{E}(\Delta_n)}{2\ln^{1/2}(n)} \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, 1\right), \quad n \to \infty. \tag{A.16}$$

Applying (A.5) and (A.6), we get

$$\Delta_n - \mathrm{E}\left(\Delta_n\right) = \sum_{k=1}^{n-1} \left(I(B_k) - \mathrm{P}\left(B_k\right)\right).$$

The summands in the sum over k are independent since the occurrence of $B_1, \ldots, B_{k-1}$ does not affect the probability $\mathrm{P}(B_k)$. Also, $(I(B_k) - \mathrm{P}\left(B_k\right))$, $k \geq 1$ are non-identically distributed random variables with zero mean.

Let us define

$$\sigma_n^2 = \sum_{k=1}^{n-1} \mathrm{E}\left(I(B_k) - \mathrm{P}\left(B_k\right)\right)^2.$$

By Lyapunov's theorem (see, e.g., Theorem 27.3 in [27]), the relation (A.16) will be proven if we show that it holds

$$\lim_{n\to\infty} \frac{\sigma_n^2}{4\ln(n)} = 1 \tag{A.17}$$

and

$$\lim_{n\to\infty} \frac{1}{\sigma_n^{2+\eta}} \sum_{k=1}^{n-1} \mathrm{E}|I(B_k) - \mathrm{P}\left(B_k\right)|^{2+\eta} = 0 \tag{A.18}$$

for some positive η.

Let us prove (A.17) first. It is easy to find that

$$\sigma_n^2 = \sum_{k=1}^{n-1} \mathrm{P}(B_k) - \sum_{k=1}^{n-1} \mathrm{P}^2(B_k). \tag{A.19}$$

From (A.13) it immediately follows that $\mathrm{P}(B_k) \le C_1/(k + \#V_1 - 2)$, $k \ge 1$, where $C_1 = 4 + 2|\# E_1 - 2\#V_1 + 2|$. This implies

$$\sum_{k=1}^{n-1} \mathrm{P}^2(B_k) \le C_1^2 \sum_{k=1}^{n-1} \frac{1}{(k + \#V_1 - 2)^2} \le C_1^2 \sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{C_1^2 \pi^2}{6}.$$

Whence, it follows that $\ln^{-1}(n) \sum_{k=1}^{n-1} \mathrm{P}^2(B_k) \to 0$, $n \to \infty$. The latter relation with (10), (A.6) and (A.19) yields (A.17).

Lyapunov's condition (A.18) is satisfied with $\eta = 1$. To verify this we write

$$\sum_{k=1}^{n-1} \mathrm{E}|I(B_k) - \mathrm{P}\,(B_k)\,|^3 = \sum_{k=1}^{n-1} \Big\{ \mathrm{P}(B_k) - \mathrm{P}^2(B_k) \Big\} \Big\{ (1 - \mathrm{P}(B_k))^2 + \mathrm{P}^2(B_k) \Big\} \le 2\sigma_n^2.$$

Thus, the left-hand side of (A.18) (with $\eta = 1$) does not exceed $\lim_{n\to\infty} 2/\sigma_n = 0$.

REFERENCES

1. Barabási, A.-L. and Albert, R., Statistical mechanics of complex networks, *Rev. Modern Phys.*, 1999, no. 74, pp. 47–97.
2. Albert, R. and Barabási, A.-L., Emergence of scaling in random networks, *Science*, 2002, no. 286, pp. 509–512.
3. Bollobás, B., Borgs, C., Chayes, J., and Riordan O., Directed Scale-Free Graphs, Society for Industrial and Applied Mathematics, 2003, USA, *Proc. of SODA 2003*, pp. 132–139.
4. Ghoshal, G., Chi, L., and Barabsi, A.L., Uncovering the role of elementary processes in network evolution, *Scientific Reports*, 2013, vol. 3, no. 6, pp. 2920.
5. Markovich, N.M. and Vaičiulis, M., Extreme Value Statistics for Evolving Random Networks, *Mathematics*, 2023, vol. 11, no. 9, pp. 2171.
6. Norros, I. and Reittu, H., On a conditionally poissonian graph process, *Advances in Applied Probability*, 2006, vol. 38, no. 1, pp. 59–75.
7. Wan, P., Wang, T., Davis, R.A., and Resnick, S.I., Are extreme value estimation methods useful for network data? *Extremes*, 2020, no. 23, pp. 171–195.
8. Wang, T. and Resnick, S.I., Poisson Edge Growth and Preferential Attachment Networks, *Methodol. Comput. Appl. Probab.*, 2023, vol. 25, no. 8.
9. Garavaglia, A. and Stegehuis, C., Subgraphs in preferential attachment models, *Advances in Applied Probability*, 2019, no. 51, pp. 898–926.
10. Wang, T. and Resnick, S.I., Consistency of Hill estimators in a linear preferential attachment model, *Extremes*, 2019, no. 22, pp. 1–28.
11. Bagrow, J. P. and Brockmann, D., Natural Emergence of Clusters and Bursts in Network Evolution, *Phys. Rev. X.*, 2013, no. 3, pp. 021016.
12. Newman, M.E.J., *Networks: an introduction*, New York: Oxford, 2010.
13. Newman, M.E.J., *Random Graphs with Clustering*, *Phys. Rev. Lett.*, 2009, vol. 103, no. 5, pp. 058701.

14. Markovich, N.M. and Vaičiulis, M., Investigation of triangle counts in graphs evolved by clustering attachment, *Autom. Remote Control*, 2024, no. 85, pp. 978–989.

15. Markovich, N.M., Ryzhov, M.S., and Kulik, M.R., Investigation of dependencies and distributions in random networks evolved by mixed models of the evolution and node deletion, *J. Large-Scale Syst. Control*, 2024, no. 109, pp. 89–121. (in Russian)

16. Bollobás, B. and Riordan, O.M., *Mathematical Results on Scale-Free Random Graphs.*, Weinheim: Wiley-WCH, 2002.

17. Ryabchenko, A.A. and Samosvat, E.A., On the number of subgraphs of the Barabási–Albert random graph, *Izvestiya: Mathematics*, 2012, vol. 76, no. 3, pp. 607–625.

18. Eggemann, N. and Noble, S.D., The clustering coefficient of a scale-free random graph, *Discrete Appl. Math.*, 2011, no. 159, pp. 953–965.

19. Móri, T.F., On random trees, *Studia Scientiarum Mathematicarum Hungarica*, 2002, vol. 39, no. 1–2, pp. 143–155.

20. Bobkov, S.G., Danshina, M.A., and Ulyanov, V.V., Rate of Convergence to the Poisson Law of the Numbers of Cycles in the Generalized Random Graphs, *Operator Theory and Harmonic Analysis*, Springer: Cham, 2021, pp. 109–133.

21. Liu, Q. and Dong, Z., Limit laws for the number of triangles in the generalized random graphs with random node weights, *Statistics and Probability Letters*, 2020, vol. 161, no. 6, pp. 108733.

22. Michielan, R., Litvak, N., and Stegehuis, C., Detecting hyperbolic geometry in networks: why triangles are not enough, *Phys. Rev. E*, 2022, vol. 106, no. 5, pp. 054303.

23. Gao, P., van der Hofstad, R., Southwell, A., and Stegehuis, C., Counting triangles in power-law uniform random graphs, *Electron. J. Combinat.*, 2020, vol. 27, no. 3, pp. 1–28.

24. Araujo, P. and Mattos, L., Local central limit theorem for triangle counts in sparse random graphs, arXiv:2307.09446v3[math.CO], July 2023.

25. Batir, N., On some properties of digamma and polygamma functions, *J.Math. Anal. Appl.*, 2007, no. 328, pp. 452–465.

26. Alzer, H., On some inequalities for the gamma and psi functions, *Math. Comp.*, 1997, vol. 66, no. 217, pp. 373–389.

27. Bilingsley, P., *Probability and Measures*, 3rd Edition, New York: John Wiley and Sons, 1995.

28. Fristedt, B. and Gray, L., *A Modern Approach to Probability Theory*, Boston: Birkhäuser, 1997.

This paper was recommended for publication by I.V. Rodionov, a member of the Editorial Board

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2026, Vol. 87, No. 3, pp. 209–217.

STOCHASTIC SYSTEMS

Invariance of Stationary Distributions of Exponential Networks with Prohibitions

G. Sh. Tsitsiashvili[*,a] **and M. A. Osipova**[*,**,b]

[*]*Institute for Applied Mathematics, Far Eastern Branch of Russian Academy Sciences, Vladivostok, Russia*
[**]*Far Eastern Federal University, Vladivostok, Russia*
e-mail: [a]*guram@iam.dvo.ru,* [b]*mao1975@list.ru*

Received February 25, 2025
Revised December 11, 2025
Accepted December 24, 2025

Abstract—We consider queuing networks with prohibitions on transitions between network nodes that determine the protocol of their operation. In the graph of transient network intensities, a set of base vertices is allocated (proportional to the number of edges) and the question is raised about deleting some subset of it so that the stationary distribution of the Markov process describing the functioning of the network is preserved. In order for this condition to be fulfilled, it is sufficient that the set of vertices of the graph of transient intensities, after removing a subset of the base vertices, coincide with the set of states of the Markov process and this graph is connected. It is proved that the ratio of the number of remaining base vertices to their total number n converges to $1/2$ for $n \to \infty$.

Keywords: queuing network, transient intensity graph, base vertex of the graph

DOI: 10.7868/S1608303226030023

1. INTRODUCTION

The transition intensity graph of a queuing network determines its operating protocol, and removing its elements reduces the possible transitions between network states. Such a change in the transition intensity graph signifies a change in the queuing network operating protocol and is associated, for example, with solving problems in the design of transportation or computer networks (see [1–6]). One way to change the network operating protocol is to introduce blocking probabilities of transitions between network states [7–9]. However, this method does not imply the introduction of transition restrictions, which is a relevant applied problem and requires the connectivity of the transition intensity graph.

The authors have their own work in this area [10, 11], and this paper is a continuation of that work. These studies consider exponential networks with various operating protocols, including protocols with random variation, and construct a transition intensity graph during their simulation. The stationary distributions of the networks under consideration are calculated under conditions of connectivity of the transition intensity graph.

In this paper, we define conditions under which the removal of some edges of the transition intensity graph that determines the functioning of a queuing network does not change the stationary distribution in it. We are talking about an open Jackson network, including one with a total limited number of requests, and the closed Gordon–Newell network [12, 13]. The search for the required conditions is based on the representation of the transition intensity graph as a union of base graphs. Each base graph consists of a base vertex and is complete; the intersection of the set of edges of any two base graphs is empty. This representation (decomposition) enables to determine a set of base vertices and the graphs corresponding to them, whose removal preserves the total set of graph vertices and the connectivity of the transition intensity graph. Graph decomposition leads to decomposition in the multiplicative theorem.

The main result of the work is the constructive removal of a subset of basic vertices that preserves the stationary distribution of the Markov process describing the network's operation. Removing any vertex from the set of remaining basic vertices violates the condition for preserving this distribution, leading to an optimal solution in a certain sense. The set of vertices in the transition intensity graph after removing the subset of basic vertices coincides with the state set of the Markov process, and this graph is connected. It has been proven that the ratio of the number of remaining basic vertices to the total number of vertices n converges to $1/2$ as $n \to \infty$.

2. DECOMPOSITION IN AN OPEN NETWORK

Consider an open Jackson network G with a Poisson input flow of intensity λ, consisting of m single-channel queuing systems with exponential service times of intensity μ_i, $i = 1, \dots, m$. The dynamics of a request's movement through the network is defined by the routing matrix $\Theta = ||\theta_{ij}| |_{i,j=0}^m$, where θ_{ij} is the transition probability after service from the ith node to the jth node, $\theta_{00} = 0$, and the node number 0 is an external source. It is assumed that the routing matrix is indecomposable:

$$\forall i, j \in \{0, 1, \dots, m\} \ \exists \, i_1, i_2, \dots, i_r \in \{1, \dots, m\} : \ \theta_{ii_1} > 0, \ \theta_{i_1 i_2} > 0, \ \dots, \ \theta_{i_r j} > 0.$$

Then the vector $\Lambda = (\lambda, \lambda_1, \lambda_2, \dots, \lambda_m)$ is the only solution to the system of balance relations

$$\Lambda = \Lambda\Theta. \tag{1}$$

The network functioning (the number of requests in the nodes) is described by a discrete Markov process $y(t)$ with a set of states $\mathbf{N} = \{\mathbf{n} = (n_1, \dots, n_m) : \ n_1, \dots, n_m \geq 0\}$ and transition intensities:

$$\begin{gathered} L(\mathbf{n}, \mathbf{n} + \mathbf{e}_k) = \lambda\theta_{0k}, \quad L(\mathbf{n} + \mathbf{e}_k, \mathbf{n}) = \mu_k \theta_{k0}, \\ L(\mathbf{n} + \mathbf{e}_k, \mathbf{n} + \mathbf{e}_i) = \mu_k \theta_{ki}, \quad 1 \leq k \neq i \leq m, \quad \mathbf{n} \in \mathbf{N}. \end{gathered} \tag{2}$$

Here the element $\mathbf{e}_k \in \mathbf{N}$, has the kth coordinate equal to 1, the rest are 0.

Consider a graph $\Gamma(\mathbf{n})$ with vertex set $V(\mathbf{n}) = \{\mathbf{n}, \ \mathbf{n} + \mathbf{e}_k, \ k = \overline{1, m}\}$ and set of edges connecting them $\{(\mathbf{n}, \mathbf{n} + \mathbf{e}_k), \ (\mathbf{n} + \mathbf{e}_i, \mathbf{n} + \mathbf{e}_j), \ i \neq j, \ 1 \leq k, i, j \leq m\}$, $\mathbf{n} \in \mathbf{N}$ (if $\mathbf{n}_1 \neq \mathbf{n}_2$, then $\Gamma(\mathbf{n}_1, \ \mathbf{n}_2)$ have no common edges). We call the vertex $\mathbf{n}$ defining the graph $\Gamma(\mathbf{n})$ a base vertex, the corresponding graph $\Gamma(\mathbf{n})$ a base graph, and the set of all such vertices $\mathbf{N}_0$ a base set. Note that $\mathbf{N}_0 = \mathbf{N}$ in the case of an open network G.

For some $\mathbf{N}_* \subseteq \mathbf{N}_0$ we set

$$\Gamma(\mathbf{N}_*) = \bigcup_{\mathbf{n} \in \mathbf{N}_*} \Gamma(\mathbf{n}), \quad V(\mathbf{N}_*) = \bigcup_{\mathbf{n} \in \mathbf{N}_*} V(\mathbf{n}).$$

Let us describe the network G_* with restrictions by associating it with the graph $\Gamma(\mathbf{N}_*)$. The absence of an edge $(\mathbf{n}, \mathbf{n} + \mathbf{e}_k)$ in the graph $\Gamma(\mathbf{N}_*)$ means that no requests arrive from outside the node k and do not leave after being serviced there. The absence of an edge $(\mathbf{n} + \mathbf{e}_k, \mathbf{n} + \mathbf{e}_i)$ in the graph $\Gamma(\mathbf{N}_*)$ means that after being serviced at node k, no requests arrive at node i, $k \neq i$. Such permissions and restrictions on transitions between network nodes define its operational protocol.

Theorem 1. *If $\rho_i = \frac{\lambda_i}{\mu_i} < 1$, $i = 1, \dots, m$, $V(\mathbf{N}_*) = \mathbf{N}$, and the graph $\Gamma(\mathbf{N}_*)$ is connected, then the Markov process $y(t)$, describing the functioning of the network G_*, is ergodic, and its stationary distribution $\pi(\mathbf{n})$ where $\mathbf{n} \in \mathbf{N}$, is calculated by the formula*

$$\pi(\mathbf{n}) = C^{-1} \prod_{i=1}^{m} \rho_i^{n_i}, \quad C = \sum_{\mathbf{n} \in N} \prod_{i=1}^{m} \rho_i^{n_i}, \quad \mathbf{n} \in \mathbf{N}. \tag{3}$$

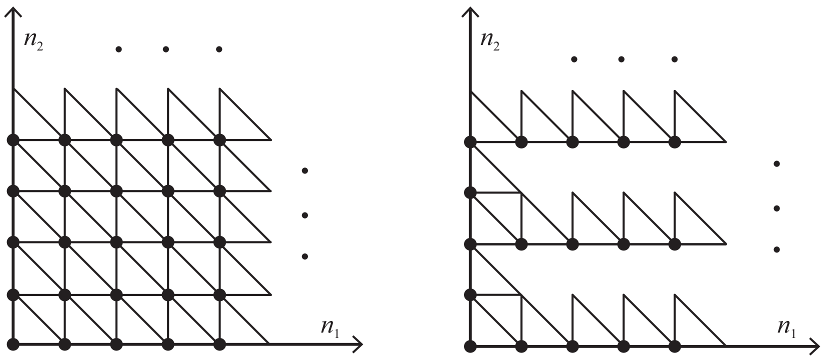

Fig. 1. Graph $\Gamma(\mathbf{N})$ (left) and graph $\Gamma(\mathbf{N}_*)$ (right), sets $\mathbf{N}_0$ (left) and $\mathbf{N}_*$ (right) are highlighted with bold dots.

Connectivity of an undirected graph $\Gamma(\mathbf{N}_*)$ means the existence of a path between any two of its vertices.

Remark 1. The stationary distribution (3) of the Markov process $y(t)$, describing the functioning of network G, remains unchanged for network G_*.

Remark 2. A similar theorem can be formulated and proven for the case of multi-channel nodes.

Remark 3. Theorem 1 can be extended to the case of a finite set of states $\mathbf{N}$ of the Markov process $y(t)$ without the constraints $\rho_i < 1$, $i = 1, \ldots, m$, where $\mathbf{N}_* \subseteq \mathbf{N}_0 \subseteq \mathbf{N}$, $\mathbf{N}_0$ is the set of base nodes that also satisfies the conditions $V(\mathbf{N}_0) = \mathbf{N}$, and the graph $\Gamma(\mathbf{N}_0)$ is connected.

The objective of this work is to construct a subset $\mathbf{N}_*$ of the set of base vertices $\mathbf{N}_0$ in such a way that, on the one hand, the stationary distribution of the Markov process is preserved, and on the other hand, the ratio of the number of vertices of set $\mathbf{N}_*$ to the number of vertices of set $\mathbf{N}_0$ tends to 1/2 with an increase in the number of requests in the network. To preserve the stationary distribution in an open network, it is sufficient to require that the equality $V(\mathbf{N}_*) = \mathbf{N}$ and the graph connectivity $\Gamma(\mathbf{N}_*)$ be satisfied.

Unlimited number of requests. Consider an open network whose operation is described by a discrete $y(t)$ with a set of states $\mathbf{N} = \{\mathbf{n} = (n_1, \ldots, n_m) : \; n_1, \ldots, n_m \geq 0\}$ and transition intensities (2). We construct a network with restrictions. We introduce a subset of the set of base vertices $\mathbf{N}_0 = \mathbf{N}$:

$$\mathbf{N}_* = \bigcup_{k=0}^{\infty} \Big(\mathbf{N}_*(2k) \bigcup (2k+1)\mathbf{e}_m\Big) \subset \mathbf{N}_0, \;\; \mathbf{N}_*(k) = \{\mathbf{n} \in \mathbf{N} : \; n_1, \ldots, n_{m-1} \geq 0, \; n_m = k\}. \tag{4}$$

As an example, the constructed graphs $\Gamma(\mathbf{N})$ and $\Gamma(\mathbf{N}_*)$, for $m = 2$ are shown on Fig. 1. Bold dots indicate the set of base vertices $\mathbf{N}_0$, as well as its subset $\mathbf{N}_*$.

Theorem 2. *For the constructed open network with restrictions* (*the set of base vertices is determined by formula* (4)), *the equality* $V(\mathbf{N}_*) = \mathbf{N}$ *holds, and the graph* $\Gamma(\mathbf{N}_*)$ *is connected. Removing a vertex from the set* $\mathbf{N}_*$ *violates either the equality* $V(\mathbf{N}_*) = \mathbf{N}$, *or the connectivity of the graph* $\Gamma(\mathbf{N}_*)$.

We define the set $\Pi(2k) = \{\mathbf{n} \in \mathbf{N} : \; 0 \leq n_1, \ldots, n_m \leq 2k\}$ and denote by $|D|$ the number of elements of the set D.

Theorem 3. *For the constructed open network with restrictions* (*the set of base vertices is determined by formula* (4)) *the following relation is valid*

$$\frac{|\mathbf{N}_* \bigcap \Pi(2k)|}{|\mathbf{N}_0 \bigcap \Pi(2k)|} \to \frac{1}{2}, \quad k \to \infty, \tag{5}$$

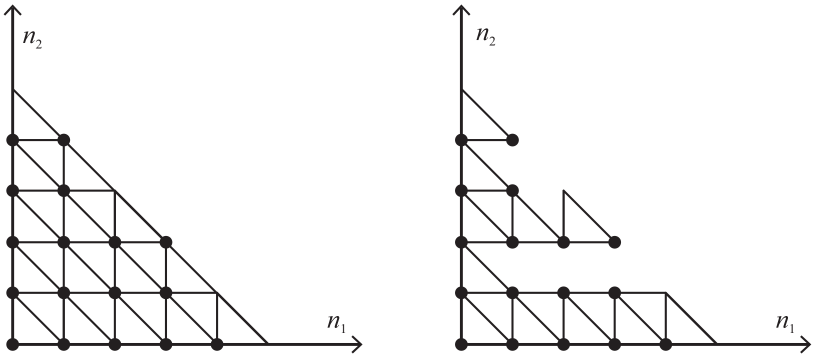

Fig. 2. Graph $\Gamma(\mathbf{N})$ (left) and graph $\Gamma(\mathbf{N}_*)$ (right), sets $\mathbf{N}_0$ (left) and $\mathbf{N}_*$ (right) are highlighted with bold dots.

where by

$$\frac{|\mathbf{N}_* \bigcap \Pi(2k)|}{|\mathbf{N}_0 \bigcap \Pi(2k)|} - \frac{1}{2} = \frac{1}{2(2k+1)} + \frac{k}{(2k+1)^m} \to 0, \quad k \to \infty. \tag{6}$$

The calculation results of $\dfrac{|\mathbf{N}_* \bigcap \Pi(2k)|}{|\mathbf{N}_0 \bigcap \Pi(2k)|} - \dfrac{1}{2}$ for different values of m and k are given in Table.

Estimate of the convergence rate

k	$m = 2$	$m = 3$
5	0.08677686	0.049211119
10	0.046485261	0.024889321
20	0.024092802	0.012485309
50	0.009851975	0.004999025

Limited total number of requests. Consider an open network whose operation is described by a discrete Markov process $y(t)$ with a finite set of states $\mathbf{N} = \{\mathbf{n} = (n_1, \ldots, n_m) : \ n_1, \ldots, n_m \geq 0, \sum_{k=1}^m n_k \leq 2K+1\}$ and transition intensities (2). We construct a network with restrictions by defining a set of base nodes $\mathbf{N}_0 = \{\mathbf{n} \in \mathbf{N} : \ \sum_{k=1}^m n_k \leq 2K\}$. The set $\mathbf{N}_0$ is maximal in the sense that adding a node $\mathbf{n} \notin \mathbf{N}_0$ to it leads to the relation $V(\mathbf{N}_0 \bigcup \mathbf{n}) \neq \mathbf{N}$. We introduce a subset of the set of base nodes

$$\mathbf{N}_* = \bigcup_{k=0}^{K} \mathbf{N}_*(2k) \bigcup \{(2k+1)\mathbf{e}_m, \ k = 0, \ldots, K-1\} \in \mathbf{N}_0, \tag{7}$$

$$\mathbf{N}_*(k) = \left\{ \mathbf{n} \in \mathbf{N} : \ \sum_{k=1}^m n_k \leq 2K - k, \ n_m = k \right\}.$$

As an example, the constructed graphs $\Gamma(\mathbf{N})$ and $\Gamma(\mathbf{N}_*)$ for $m = 2$ are shown on Fig. 2. The set of base vertices is highlighted with bold dots $\mathbf{N}_0$, and as well as its subset $\mathbf{N}_*$.

Theorem 4. *For the constructed open network with restrictions* (*the set of basic vertices is determined by formula* (7)), *the equality* $V(\mathbf{N}_*) = \mathbf{N}$ *holds and the graph* $\Gamma(\mathbf{N}_*)$ *is connected. Removing a vertex from the set* $\mathbf{N}_*$ *violates either the equality* $V(\mathbf{N}_*) = \mathbf{N}$ *or the connectivity of the graph* $\Gamma(\mathbf{N}_*)$.

Theorem 5. *For the constructed open network with restrictions* (*the set of base nodes is determined by formula* (7)) *the following relation is valid:*

$$\frac{|\mathbf{N}_*|}{|\mathbf{N_0}|} \to \frac{1}{2}, \quad K \to \infty. \tag{8}$$

3. DECOMPOSITION IN A CLOSED NETWORK

Consider a closed network (Gordon–Newell network) G with a constant number of requests $2K+1$ and a number of service nodes $m>2$. Request movements in the network are described by the routing matrix $\Theta = ||\theta_{ij}||_{i,j=1}^{m}$, which is assumed to be indecomposable. Then, for any $B>0$, a solution $\Lambda = (\lambda_1, \dots, \lambda_m)$ to the system $\Lambda = \Lambda\Theta$ of balance relations exists and is unique, provided $B = \sum_{i=1}^{m} \lambda_i > 0$. The network's operation (the number of requests at nodes) is described by a discrete Markov process $y(t)$ with a finite set of states

$$\mathbf{N} = \left\{\mathbf{n} = (n_1, \dots, n_m):\ n_1, \dots, n_m \geq 0,\ \sum_{k=1}^{m} n_k = 2K+1\right\}$$

and transition intensities:

$$L(\mathbf{n}+\mathbf{e}_i, \mathbf{n}+\mathbf{e}_k) = \mu_i\theta_{ik},\quad 1 \leq i \neq k \leq m,\quad \mathbf{n}+\mathbf{e}_i, \mathbf{n}+\mathbf{e}_k \in \mathbf{N}. \tag{9}$$

We define the set of all base vertices $\mathbf{N}_0$ by the equality

$$\mathbf{N}_0 = \{\mathbf{n} \in \mathbf{N}:\ n_m > 0\}.$$

We rewrite equalities (9) for $\mathbf{n} \in \mathbf{N}_0$:

$$L(\mathbf{n}, \mathbf{n}-\mathbf{e}_m+\mathbf{e}_i) = \mu_m\theta_{mi},\quad L(\mathbf{n}-\mathbf{e}_m+\mathbf{e}_i, \mathbf{n}-\mathbf{e}_m+\mathbf{e}_k) = \mu_i\theta_{ik},\quad 1 \leq i \neq k < m. \tag{10}$$

Using equalities (10), we define the set

$$V(\mathbf{n}) = \{\mathbf{n}, \mathbf{n}-\mathbf{e}_m+\mathbf{e}_i,\ 1 \leq i < m\}.$$

Consider a graph $\Gamma(\mathbf{n})$, $\mathbf{n} \in \mathbf{N}_0$, with the set of vertices $V(\mathbf{n})$ and the set of edges connecting them ($\Gamma(\mathbf{n}_1)$, $\Gamma(\mathbf{n}_2)$, $\mathbf{n}_1 \neq \mathbf{n}_2$, have no common edges). Obviously, $V(\mathbf{N}_0) = \mathbf{N}$.

Let $\mathbf{N}_* \subset \mathbf{N}_0$. Similar to the open network, we can describe a closed network G with restrictions by associating it with the graph $\Gamma(\mathbf{N}_*)$, and denote it by G^*.

Theorem 6. *If $V(\mathbf{N}_*) = \mathbf{N}$, and the graph $\Gamma(\mathbf{N}_*)$ is connected, then the Markov process $y(t)$ describing the functioning of the network G^* is ergodic and its stationary distribution $\pi(\mathbf{n})$, $\mathbf{n} \in \mathbf{N}$, is calculated by the formula*

$$\pi(\mathbf{n}) = C^{-1}\prod_{i=1}^{m}\left(\frac{\lambda_i}{\mu_i}\right)^{n_i},\ C = \sum_{\mathbf{n}\in\mathbf{N}}\prod_{i=1}^{m}\left(\frac{\lambda_i}{\mu_i}\right)^{n_i},\quad 1 \leq i \leq m.$$

Let us construct $\mathbf{N}_* \subset \mathbf{N}_0$ such that the stationary distribution of the Markov process is preserved and the relation $\frac{|\mathbf{N}_*|}{|\mathbf{N}_0| \to \frac{1}{2}}$ as $K \to \infty$. Due to Theorem 6, to preserve the stationary distribution in a closed network, it suffices to require that the equality $V(\mathbf{N}_*) = \mathbf{N}$ be satisfied and that the graph $\Gamma(\mathbf{N}_*)$ is connected.

To construct the set of base vertices $\mathbf{N}_*$ we associate with each point of the set $\mathbf{n} = (n_1, \dots, n_m) \in \mathbf{N}$ a point $\mathbf{n}' = (n_1, \dots, n_{m-1}, 0)$ and denote the set of such points $\mathbf{N}' = \{\mathbf{n}' :\ \sum_{k=1}^{m-1} n_k \leq 2K+1\}$. Then the set $\mathbf{N}_0$ will correspond to the set $\mathbf{N}'_0 = \{\mathbf{n}' \in \mathbf{N}':\ \sum_{k=1}^{m-1} n_k \leq 2K\}$. We now define

$$\mathbf{N}'_* = \bigcup_{k=0}^{K} \mathbf{N}'_*(2k) \bigcup \{(2k+1)\mathbf{e}_{m-1},\ k = 0, \dots, K-1\},$$

$$\mathbf{N}'_*(k) = \left\{\mathbf{n}' \in \mathbf{N}':\ \sum_{k=1}^{m-1} n_k \leq 2K-k,\ n_{m-1} = k\right\}.$$

Then we construct the set

$$\mathbf{N}_* = \{\mathbf{n} \in \mathbf{N}_0 : \ \mathbf{n}' \in \mathbf{N}'_*\}. \tag{11}$$

Theorem 7. *For the constructed closed network with restrictions (the set of base vertices is determined by formula (11)) the equality $V(\mathbf{N}_*) = \mathbf{N}$ holds and the graph $\Gamma(\mathbf{N}_*)$ is connected. Removing a vertex from the set $\mathbf{N}_*$ violates either the equality $V(\mathbf{N}_*) = \mathbf{N}$ or the connectivity of the graph $\Gamma(\mathbf{N}_*)$.*

Theorem 8. *For the constructed closed network with restrictions (the set of base vertices is determined by formula (11)) the relation*

$$\frac{|\mathbf{N}_*|}{|\mathbf{N}_0|} \to \frac{1}{2}, \quad K \to \infty \tag{12}$$

holds.

4. CONCLUSION

In all the queuing models considered, the set of base nodes $\mathbf{N}_0$, in terms of the number of nodes and, consequently, the number of edges in the transition intensity graph, is approximately half the subset of base nodes $\mathbf{N}_* \subset \mathbf{N}_0$. Exponential queuing systems with restrictions, while maintaining a stationary distribution, were constructed by transitioning to a special type of transition intensity graph. This work began with a special case, namely, with $m = 2$ for an open network and $m = 3$ for a closed one. The variants of the set $\mathbf{N}_*$ proposed in this work are not unique. We plan to continue this work, generalizing the obtained results and extending them to queuing systems whose operation is described by Markov processes.

The authors are currently collaborating with maritime logistics specialists and plan to jointly apply the obtained results to the design of a logistics network for container transportation from one mode of transport to another (road, rail and sea) to ensure uninterrupted delivery of goods.

FUNDING

The research was carried out within the state assignment for IAM FEB RAS no. 075-00460-26-00.

APPENDIX

A.1. PROOF OF THEOREM 1

The proof of formula (3) is based on the ergodicity theorem formulated in [14] for a discrete Markov process. Let us verify its sufficient conditions. 1) To verify that the distribution (3) satisfies the Kolmogorov–Chapman equations

$$\sum_{\mathbf{n}' \in V(\mathbf{N}_0)} \pi(\mathbf{n}) L(\mathbf{n}, \mathbf{n}') = \sum_{\mathbf{n}' \in V(\mathbf{N}_0)} \pi(\mathbf{n}') L(\mathbf{n}', \mathbf{n}), \quad \mathbf{n} \in \mathbf{N}, \tag{A.1}$$

it is enough to verify that it satisfies for $\mathbf{n}_0 \in \mathbf{N}_0$ the equalities

$$\sum_{\mathbf{n}' \in V(\mathbf{n}_0)} \pi(\mathbf{n}) L(\mathbf{n}, \mathbf{n}') = \sum_{\mathbf{n}' \in V(\mathbf{n}_0)} \pi(\mathbf{n}') L(\mathbf{n}', \mathbf{n}), \quad \mathbf{n} \in \mathbf{V}(\mathbf{n_0}). \tag{A.2}$$

Substituting the distribution (3) into formula (A.2) leads to the balance relations (1), with $C < \infty$. 2) The states of the process $y(t)$ are communicating, which follows from the assumption that the graph $\Gamma(\mathbf{N}_*)$ is connected. 3) The regularity condition is also satisfied, since for any $\mathbf{n} \in \mathbf{N}$ the inequality $\sum_{\mathbf{n}' \in \mathbf{N}} L(\mathbf{n}, \mathbf{n}') <$ const. holds. Theorem 1 is proved.

A.2. PROOF OF THEOREM 2

It is obvious that

$$V(\mathbf{N}_*) = \bigcup_{k=0}^{\infty} (\mathbf{N}_*(2k) \cup \mathbf{N}_*(2k+1)) = \mathbf{N}.$$

Since the graph $\Gamma(\mathbf{N}_*(2k))$ is connected and the edge $((2k+1)\mathbf{e}_m, (2k+2)\mathbf{e}_m)$ connects the graphs $\Gamma(\mathbf{N}_*(2k))$, $\Gamma(\mathbf{N}_*(2k+2))$, $k = 0, 1, \ldots$, then the graph $\Gamma(\mathbf{N}_*)$ is also connected.

Let's check the second assertion of the theorem. First, consider the case $m = 2$. The following options are possible.

1. If $\mathbf{n} = (0, 2k)$ and $k > 0$, then the graph $\Gamma(\mathbf{N}_*\backslash\mathbf{n})$ does not contain a path between vertices $\mathbf{n} = (0, 2k)$, $(0, 2k+1)$ and, therefore, is disconnected. If $k = 0$, then $\mathbf{n} \notin V(\mathbf{N}_*\backslash\mathbf{n})$ and, therefore, $V(\mathbf{N}_*\backslash\mathbf{n}) \neq V(\mathbf{N}_*)$.

2. If $\mathbf{n} = (0, 2k+1)$, $k \geq 0$, then in the graph $\Gamma(\mathbf{N}_*\backslash\mathbf{n})$ there is no path between the vertices $(0, 2k+1)$, $(0, 2k+2)$ and, therefore, the graph $\Gamma(\mathbf{N}_*\backslash\mathbf{n})$ is disconnected.

3. If $\mathbf{n} = (n_1, 2k)$, $k \geq 0$, and $n_1 > 1$, then $(n_1, 2k+1) \notin V(\mathbf{N}_*\backslash\mathbf{n})$ and, consequently, $V(\mathbf{N}_*\backslash\mathbf{n}) \neq V(\mathbf{N}_*)$. If $n_1 = 1$, then the vertex $\mathbf{n} = (1, 2k)$ is not connected with the vertex $(2, 2k)$ in the graph $\Gamma(\mathbf{N}_*\backslash\mathbf{n})$ and, therefore, the graph $\Gamma(\mathbf{N}_*\backslash\mathbf{n})$ is disconnected.

Let us now move on to the case $m > 2$. The verification of conditions **1**, **2** when replacing $(0, 2k)$ with $(0, \ldots, 0, 2k)$, and $(0, 2k+1)$ with $(0, \ldots, 0, 2k+1)$ is similar. When checking condition **3**, we set $\mathbf{n} = (n_1, \ldots, n_{m-1}, 2k)$. If $n_1 + \ldots + n_{m-1} > 1$, then it is easy to establish that $(n_1, \ldots, n_{m-1}, 2k+1) \notin V(\mathbf{N}_*\backslash\mathbf{n})$. If $n_1 + \ldots + n_{m-1} = 1$, then in the graph $\Gamma(\mathbf{N}_*\backslash\mathbf{n})$ there is no connection between the vertices $\mathbf{n} = \mathbf{e}_i + 2k\mathbf{e}_m$ and $2\mathbf{e}_i + 2k\mathbf{e}_m$. The Theorem 2 has been proven.

A.3. PROOF OF THEOREM 3

The number of vertices $|\mathbf{N}_0 \bigcap \Pi_m(2k)| = (2k+1)^m$. In turn, the number of vertices $|\mathbf{N}_* \bigcap \Pi_m(2k)| = (2k+1)^{m-1}(k+1) + k$. From here follow formulas (5), (6). Theorem 3 is proved.

A.4. PROOF OF THEOREM 4

The proof of Theorem 4 almost verbatim repeats the proof of Theorem 2.

A.5. PROOF OF THEOREM 5

From [15] it follows that the number of solutions of the Diaphantine equation $n_1 + \ldots + n_m = n$ in non-negative integers is equal to C_{n+m-1}^{n}. Then

$$|\mathbf{N}_*(k)| = \sum_{i=0}^{2K-k} C_{i+m-2}^{m-2} = C_{2K-k+m-1}^{m-1}, \quad |\mathbf{N}_*| = \sum_{k=0}^{K} |\mathbf{N}_*(2k)| + K. \tag{A.3}$$

From (A.3) we obtain the relations $1 = \mathbf{N}_*(2K) < \mathbf{N}_*(2K-1) < \ldots < \mathbf{N}_*(0)$. Consequently, the inequalities hold

$$\sum_{k=0}^{K} |\mathbf{N}_*(2k)| > \sum_{k=1}^{K} |\mathbf{N}_*(2k-1)| > \sum_{k=0}^{K} |\mathbf{N}_*(2k)| - |\mathbf{N}_*(0)|. \tag{A.4}$$

From (A.4) we have

$$2\sum_{k=0}^{K}|\mathbf{N}_*(2k)|-|\mathbf{N}_*(0)|<|\mathbf{N}_0|=\sum_{k=0}^{K}|\mathbf{N}_*(2k)|+\sum_{k=1}^{K}|\mathbf{N}_*(2k-1)|<2\sum_{k=0}^{K}|\mathbf{N}_*(2k)|,$$

$$\frac{|\mathbf{N}_0|}{2}<\sum_{k=0}^{K}|\mathbf{N}_*(2k)|<\frac{|\mathbf{N}_0|+|\mathbf{N}_*(0)|}{2},$$

$$\frac{|\mathbf{N}_0|}{2}+K<|\mathbf{N}_*|<\frac{|\mathbf{N}_0|+|\mathbf{N}_*(0)|}{2}+K. \tag{A.5}$$

In turn, for $K\to\infty$ the asymptotic estimates are valid

$$|\mathbf{N}_0|=\sum_{i=0}^{2K}C_{i+m-1}^{m-1}=C_{2K+m}^{m}\sim\frac{(2K)^m}{m!},\quad |\mathbf{N}_*(0)|=C_{2K+m-1}^{m-1}\sim\frac{(2K)^{m-1}}{(m-1)!},$$

$$\frac{K}{|\mathbf{N}_0|}\to 0,\quad \frac{|\mathbf{N}_*(0)|}{|\mathbf{N}_0|}\to 0.$$

Thus, relation (8) is satisfied by inequality (A.5). Theorem 5 is proven.

A.6. PROOF OF THEOREM 6

The proof of this theorem is similar to the proof of Theorem 1.

A.7. PROOF OF THEOREM 7

By construction, there is a one-to-one correspondence between the sets $\mathbf{N}'_*$ and $\mathbf{N}_*$, and also between the sets $\mathbf{N}'$ and $\mathbf{N}$, so the graphs $\Gamma(\mathbf{N}'_*)$ and $\Gamma(\mathbf{N}_*)$ are isomorphic. In turn, it follows from Theorem 4 that $V(\mathbf{N}'_*)=\mathbf{N}'$ and the graph $\Gamma(\mathbf{N}'_*)$ is connected, and deleting a vertex from the set $\mathbf{N}'_*$ violates at least one of these statements. The statement of the theorem is a consequence of the listed facts. Theorem 7 has been proven.

A.8. PROOF OF THEOREM 8

Since

$$|\mathbf{N}'_*|=|\mathbf{N}_*|,\ |\mathbf{N}'_0|=|\mathbf{N}_0|,$$

then the statement of Theorem 5 implies the statement of Theorem 8.

REFERENCES

1. Kvitko, K.B., Comparative analysis of international transport systems: infrastructure, ratings, transport corridors, *Transport systems and technologies*, 2020, vol. 6, no. 1, pp. 15–29.

2. Komov, M.S., Current development trends and the importance of international transport corridors in the transport services market, *Economy and Entrepreneurship*, 2022, vol. 6, no. 143, pp. 243–246.

3. O'Hara, J., A Toward a Commodity Enterprise Middleware: Can AMQP enable a new era in messaging middleware? A look inside standards-based messaging with AMQP, *Queue*, 2007, vol. 5, no. 4, pp. 48–55.

4. Gudehus, T. and Kotzab, H., *Comprehensive Logistics*, Heidelberg: Springer Science+Business Media, 2012.

5. Konfino, K.V., Topical issues of development of international transport corridors and realization of transport potential of Russia, *Labour and Social Relations Journal*, 2019, vol. 30, no. 5, pp. 101–108.

6. Vinoski, S., Advanced Message Queuing Protocol, *IEEE Internet Computing*, 2006, vol. 10, no. 6, pp. 87–89.
7. Balsamo, S. and De Nitto, V., A survey of Product-form Queueing Networks with Blocking and their Equivalences, *Annals of Operations Research*, 1994, vol. 48, pp. 31–61.
8. Balsamo, S. and Clo, M.C., A Convolution Algorithm for Product-form Queueing Networks with Blocking, *Annals of Operations Research*, 1998, vol. 79, pp. 97–117.
9. Boucherie, R.J. and van Dijk, N.M., On the arrival theorem for product form queueing networks with blocking, *Performance Evaluation*, 1997, vol. 29, pp. 115–176.
10. Tsitsiashvili, G.Sh. and Osipova, M.A., New multiplicative theorems for queueing networks, *Problems of information transmission*, 2005, vol. 41, no. 2, pp. 111–122.
11. Tsitsiashvili, G.Sh. and Osipova, M.A., Probability distribution in queueing networks with variable structure, *Problems of information transmission*, 2006, vol. 42, no. 2, pp. 101–108.
12. Basharin, G.P. and Tolmachev, A.L., *The Theory of Queueing Networks and Its Applications to the Analysis of Information and Computing Systems. Results of Science and Technology, series. "Probability Theory"*, Moscow: VINITI, 1983.
13. Balbo1, G., Marin, A., Olliaro, D. and Sereno, M., Computational Algorithms for the Product Form Solution of Closed Queuing Networks with Finite Buffers and Skip-Over Policy, 2024, arXiv:2409.08075v1 [cs.PF].
14. Ivchenko, G.I., Kashtanov, V.A., and Kovalenko, I.N., *Queueing Theory*, Moscow: Higher School, 1982.
15. Vilenkin, N.Ya., *Combinatorics*, Moscow: Nauka, 1969.

This paper was recommended for publication by M.V. Khlebnikov, a member of the Editorial Board

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2026, Vol. 87, No. 3, pp. 218–226.

ROBUST, ADAPTIVE, AND NETWORK CONTROL

Optimal Spanning Trees in Fault Diagnosis and Design of Fuzzy Devices

D. V. Speranskii

Russian University of Transport, Moscow, Russia
e-mail: Speranskiy.dv@gmail.com

Received September 16, 2023
Revised November 28, 2025
Accepted December 30, 2025

Abstract—The realizability of a fuzzy graph is conceptualized, and two optimization problems are formulated to construct optimal spanning trees for fuzzy graphs. The optimality criteria are graph realizability and tree minimality (by length), with realizability taken into account. Being multicriteria, these problems fundamentally differ from the analogous problem for crisp graphs. An exact method for solving the above problems is described; it has a rather high computational complexity. A simple heuristic algorithm for constructing a minimal spanning tree of a fuzzy graph is proposed; being computationally less intensive, this method, however, does not ensure absolute minimality of the spanning tree. The method is illustrated with an example.

Keywords: fuzzy graph, spanning trees, realizability and tree minimality, exact construction methods, heuristic method

DOI: 10.7868/S1608303226030033

1. INTRODUCTION

When designing electronic devices, one often needs to connect several contacts electrically. It is desirable to obtain a circuit with a minimum number of conductors. This problem can be solved using the model of a connected undirected graph $G(V, E)$, where V and E denote the sets of vertices and possible connections between vertex pairs, respectively. For each edge $(u, v) \in E$, a weight $w(u, v)$ can be assigned, determining the cost of connecting u and v. To minimize the cost of all vertex connections, it is necessary to find an acyclic subset $T \subseteq E$ connecting all vertices with the minimal total weight $w(T) = \sum_{(u,v)\in T} w(u, v)$. Since the set T is acyclic and connects all vertices, it must form a tree, called a minimal spanning tree.

Consider another case: the initial device represents a set of subcircuits, and there is a given set of pairs of subcircuits that can mutually diagnose each other. Assume that these pairs have multichannel links for mutual fault diagnosis. Each multichannel link of a subcircuit pair has a weight equal to its cost. By analogy with the previous problem, it is required to construct an acyclic subset of subcircuits connecting all mutually diagnosable pairs and having minimal weight. Obviously, the solution of this problem also reduces to constructing a minimal spanning tree.

The above problems are of graph theory, and graphs have many optimization-related practical applications in various fields. The extensive literature devoted to their solution largely pertains to classical graph theory, where a graph is deterministic and its description contains no uncertainties. Such graphs are usually called crisp. The numerous results established for crisp graphs, as well as the basic concepts and constructs of graph theory, were presented in the monographs [1, 2]. Therefore, the corresponding references are generally omitted below, assuming the reader's background on the subject.

The problem of constructing spanning trees for classical graphs has been actively studied since the 1920s. For instance, a solution algorithm proposed by O. Borůvka was published in 1926; it was later rediscovered multiple times (K. Florek, S. Perkal, and W. Sołtysiak, occasionally miscalled W. Sollin in the literature). Various approaches were used to solve the problem. For example, based on Borůvka's algorithm, a randomized algorithm with linear (on average) execution time was implemented. E. Minieka, the author of the monograph [3], proposed an original algorithm; this monograph, with Chapter 2 entirely focused on the problem under consideration, also presented the (now widely used) algorithms of R. Prim [4] and J. Kruskal [5]. The algorithms mentioned involve the idea of ordering the graph edges by weight and their incidence.

Following the current proliferation of computer networks and the development of a new mathematical apparatus (e.g., genetic algorithms), the recent publications (for instance, see [6, 7]) have provided new approaches for constructing spanning trees using evolutionary computations.

Unfortunately, the crisp graph model turns out to be inapplicable in many real situations where the operation of an object of study is not precisely described, and such a description cannot be obtained in principle. In this case, new tools are needed to reflect the emerging uncertainties. Such tools were proposed in L. Zadeh's paper [8]: he introduced the concept of a fuzzy set, the foundation of the theory of fuzzy sets, now widely used in applications.

Today, the theory of fuzzy graphs is an actively developing area. The interest of researchers in this theory over the last decade is due to important practical applications. Supplementing the above examples, we also mention the problems of system survivability, assessing the information reliability of complex systems, and modeling heterogeneous systems. The phenomenon of fuzziness arises in many optimization problems of an economic nature as well. For instance, consider the well-known problem of designing a road network for a certain region with minimal construction costs. If the list of regional roads (between cities) and their parameters is precisely known, then the problem has an exact solution using the crisp graph model. Several efficient algorithms have been developed to obtain it [3].

If the reasonability of building roads between particular city pairs is ambiguous, then the resulting uncertainty requires a different model to solve the problem.

To illustrate the method for constructing a spanning tree for a fuzzy graph, we will use the regional road network design problem, owing to its popularity and intuitively clear formulation and interpretation.

In crisp graph theory, the road network design problem reduces to constructing a spanning tree for an original graph. Under uncertainty, this problem can be solved using the fuzzy graph model, and it also reduces to constructing a spanning tree. This paper is devoted to constructing spanning trees for fuzzy graphs that are optimal by criteria different from those adopted for crisp graphs.

2. SOME CONCEPTS AND DEFINITIONS

First, we recall the concepts necessary for further considerations. They are all interpreted as presented in A. Kaufmann's monograph [9]. We begin with the concept of a fuzzy set. Let a set A be a subset of a set E ($A \subset E$). The membership of an element $x \in A$ is written using a function $\mu_A(x)$ that takes any nonnegative value in an ordered set M, particularly in the interval [0, 1]. A mathematical object of the form

$$A = \{(x_1|\mu_A(x_1)), (x_2|\mu_A(x_2)), \ldots, (x_n|\mu_A(x_n))\},$$

where $\mu_A(x_i)$ defines the grade of membership of an element x_i in the set A, is called a fuzzy subset of the set E. Here is the definition of a fuzzy graph [9]. Let E_1, E_2 be two sets, and let $x \in E_1$ and $y \in E_2$. The set of all pairs (x, y) defines the Cartesian product $E_1 \times E_2$. A fuzzy subset G such that

$$\forall (x, y) \in E_1 \times E_2 \, \mu_G(x, y) \in M,$$

where M is the set of membership grades of $E_1 \times E_2$, is called a fuzzy graph. If the pairs (x, y) are considered ordered, then the corresponding graph is usually called a directed fuzzy graph. Fuzzy graphs, like crisp ones, are depicted as vertices and edges (arcs) connecting them.

The terms encountered in this paper generally have the same meaning as in crisp graph theory. If some concepts for fuzzy graphs do not coincide with the analogs for crisp graphs, this will be specified explicitly. Let $a(x, y)$ denote the length of an edge (x, y) in an undirected fuzzy graph. Next, we recall the concept of a spanning tree of a graph: it is an acyclic connected subgraph of a given connected graph that includes all its vertices.

3. PROBLEM FORMULATION IN SUBSTANTIVE TERMS

First, we comment on the problem of constructing a spanning tree for a fuzzy graph. In a region containing several cities, a road construction project is considered to connect each city pair (not necessarily directly; transit connections are possible). There is a given set of city pairs for which the construction of a directly connecting road is possible (allowed). The lengths of roads between city pairs from this set are known. Also, there is an assessment of the reasonability of including each road in the project (e.g., based on the resulting economic benefit and the construction costs). This assessment is defined as a number from some ordered set M (in particular, from the interval [0, 1]), understood as the grade of membership of the road in the network project under design.

As a mathematical model of the road network, we will use a fuzzy graph where the set of regional cities is the set of vertices, and the set of project roads is the set of its edges. In contrast to a crisp graph where the weight of each edge (x, y) is typically a single number interpreted as its length, the weight for a fuzzy graph is an ordered pair of numbers $(a(x, y), \mu_G(x, y))$: the first corresponds to the edge length, and the second to its grade of membership. Obviously, now the optimality of the project can be assessed by two criteria associated with the components of the pair. Therefore, two optimization problems are considered in the paper; see their substantive formulations below. It is required to design a road network project (construct a spanning tree) connecting all cities of the region so that:

1) The maximum probability of tree construction (the maximum sum of the membership grades of its edges) is reached under the minimum possible sum of road lengths included in the project, using the parameters of the fuzzy graph model of the network project.

2) The minimum sum of all road lengths in the network is reached under the maximum possible sum of their membership grades, using the parameters of the fuzzy graph model of the network project.

Hereinafter, the road network project and its mathematical model (the spanning tree of the fuzzy graph) are used as synonyms.

As is known [3], when solving the analog of this problem for a crisp graph, the optimality criterion is the minimum total road length of the network project. In this case, the problem reduces to constructing a minimal spanning tree for the graph where the cities are the vertices, and the lengths of the edges (roads) connecting city pairs are known. The solution of the two problems posed for a fuzzy graph also obviously reduces to constructing some optimal spanning trees, but in a different interpretation (understanding) of optimality. To the best of our knowledge, the problem in the above or similar statements for a general fuzzy graph has not been studied in the literature.

4. REALIZABILITY AND MINIMALITY OF A SPANNING TREE

To clarify the concepts introduced below, we will present the considerations and motivation behind them for fuzzy graphs. If a fuzzy graph $G(X, E)$ is a mathematical model of the network for solving the problem, then the set E consists only of "allowed" edges for which $\mu_G(x, y) > 0$. As mentioned above, the weight of each edge (x, y) is an ordered pair of numbers $a(x, y), \mu_G(x, y)$, where the second number is the grade of membership of (x, y) in the set E. This value can always be transformed to the probability $P(x, y)$ of the belonging of the edge (x, y) to the set E. Clearly, the membership function of a fuzzy set resembles a probability density. The difference is only that the sum of probabilities over all possible values of a random variable always equals 1, while the sum S of the membership function values of all arcs in a fuzzy graph can be any nonnegative number. In this regard, it becomes possible to transform the membership function of a graph edge to a probability distribution function. This is done simply by normalizing the membership function, i.e., dividing its value for each edge of the fuzzy graph by the value $S \neq 0$. Such a method was used in several publications as well.

In fuzzy set theory, an edge's nonzero grade of membership is interpreted only as a potential possibility of including this edge in the graph. Naturally, the greater the grade is, the higher the possibility will be. However, the factual inclusion of an edge (with a nonzero probability estimate) in the graph is not guaranteed. An analogous situation arises in probability theory: an event may have a nonzero probability of occurrence, but this does not mean it will necessarily happen.

By assumption, a given fuzzy graph is connected: otherwise, a spanning tree would not exist for it. However, in reality, some edges with nonzero membership values may be absent, violating the connectivity property of the graph. To avoid this as much as possible, it is necessary to maximize the membership value for each edge (road) included in the spanning tree (i.e., maximize the factual existence probability of this tree).

For the further presentation, we introduce the following notation: $E = \{(a_1, b_1), \ldots, (a_z, b_z)\}$ is the set of edges of a fuzzy graph $G(X, E)$; *Rect* $(G(X, E))$ is the realizability of the graph $G(X, E)$; *Lect* $(G(X, E))$ is the total length of all edges of the graph $G(X, E)$; *Pbe* (x, y) is the probability of the belonging of an edge (x, y) to the set $E = \{(a_1, b_1), \ldots, (a_z, b_z)\}$; Mct $(G(X, E))$ is the minimal spanning tree of the graph $G(X, E)$; and *Lct* $(G(X, E))$ is the length of the spanning tree of the graph $G(X, E)$. Using these designations, we give some definitions.

The realizability of a spanning tree of a fuzzy graph $G(X, E)$ is the sum of the probabilities of the belonging of all edges included in this tree, i.e., *Rect* $(G(X, E)) = \sum_{i=1}^{z} Pbe_G(a_i, b_i)$. Thus, the realizability of a road network project is a probabilistic estimate for the factual existence possibility of a spanning tree for a given fuzzy graph with specified probabilities of the belonging of its edges. Note that the concept of realizability agrees well with common sense: the greater the realizability value is, the higher the chances of constructing a spanning tree will be.

The *length of a spanning tree* of a fuzzy graph is the sum of the lengths of all its edges: *Lect* $(G(X, E)) = \sum_{i=1}^{z} a(a_i, b_i)$. The minimum-length (shortest) spanning tree of a fuzzy graph will be called the *minimal spanning tree* and denoted by Mct $(G(X, E))$.

A *spanning tree* that has maximum realizability and, at the same time, the minimum possible length (*optimality by tree length*) under the specified estimates of the probabilities of the belonging of its edges will be called *optimal by realizability.*

Now we provide a rigorous statement of the first problem posed in the previous section: for a given undirected fuzzy graph, it is required to construct a spanning tree that is *optimal by realizability.* The second problem is now stated as follows: for a given undirected fuzzy graph, it is required to construct a *minimal spanning tree.*

5. A METHOD FOR FINDING OPTIMAL SPANNING TREES

The specification of a fuzzy graph involves a membership function, and the problem of obtaining an appropriate function arises accordingly. Researchers have been studying this problem since the very introduction of the concept of a fuzzy graph. There are many publications on this topic, e.g., [10–12]. Note that calculating the probability of a certain event with a given accuracy requires conducting a significant number of experiments (in principle, tending to infinity). This is not always reasonable due to high complexity, and one often uses some approximate value. We emphasize that methods for determining the values of necessary parameters to obtain results with a given accuracy are now well-developed in sampling theory. For instance, see the monograph [13] for a detailed coverage of these issues. The problem of constructing a membership function goes beyond the scope of this paper; therefore, by assumption, such a function is available together with a fuzzy graph specification.

To construct a spanning tree optimal by realizability for a fuzzy graph, we propose a method based on the results of solving a similar problem (finding a minimal spanning tree) in classical graph theory. For instance, a procedure for constructing all spanning trees of a crisp graph, as well as a proof of its correctness, was presented in [1, Ch. 7, Sec. 2]. Removing the second component (the membership value) from the weight of each edge in a given fuzzy graph, we obtain the corresponding crisp graph $\hat{G}(X, E)$. This graph is a special case of the original fuzzy graph, where all edges are present regardless of their membership values. Using the above procedure, we construct for $\hat{G}(X, E)$ the set of all its spanning trees, denoted by $Ct(\hat{G}(X, E)) = \{G_1, \ldots, G_k\}$. Next, for each edge of the spanning trees in this set, we restore the second component (membership value) to its weight, thereby making these trees fuzzy. Clearly, after this transformation, the crisp spanning trees of the set $Ct(\hat{G}(X, E))$ remain spanning trees and form the set $Ct(G(X, E)) = \{\bar{G}_1, \ldots, \bar{G}_k\}$ of all spanning trees of the given fuzzy graph $G(X, E)$.

For each fuzzy spanning tree from the set $Ct(G(X, E))$, we calculate its realizability $\mathit{Rect}\ (\bar{G}_i)$, $i = 1, \ldots, k$ and find among them the tree (or several trees) with maximum realizability. Next, for each fuzzy spanning tree from the set $Ct(G(X, E))$, we calculate $\mathit{Lect}\ (\bar{G}_i)$, $i = 1, \ldots, k$, its length. For the given graph, the spanning tree optimal by realizability can be found using the data obtained.

Let all spanning trees of the set $Ct(G(X, E))$ be numbered. We form the set *MaxR* of the numbers of all trees with maximum realizability and the set *ML* of the numbers of all trees from the set $Ct(G(X, E))$ with lengths calculated.

Letting $MaxR = \{v_1, \ldots, v_z\}$, we form the set of all pairs (v_i, *the length of tree* v_i), where $i = 1, \ldots, z$. Among all these pairs, we find the one with the minimum tree length; suppose that the corresponding tree has number i_0. Obviously, the spanning tree with number i_0 is optimal by realizability. If the set *MaxR* is a singleton, then the optimal tree has that number. If there are several such trees, we select the tree j_0 with the minimum length among the elements (v_i, *the length of tree* v_i) of the above set. Clearly, the tree with number j_0 will be optimal by realizability.

The second optimization problem (minimizing the spanning tree) is solved by analogy. First, we select the minimal spanning tree in the set of pairs (v_i, *the length of tree* v_i). If this is only one tree with number i_0, then this tree with the realizability value $\mathit{Rect}(G_{i_0})$ is optimal (minimal by length). If there are several such trees (with numbers $i_0^1, i_0^2, \ldots, i_0^r$), then we find among them the tree j_0 with the maximum realizability value. Obviously, this tree will be a minimal spanning tree.

The method proposed rests on the possibility of constructing the set of all spanning trees of a graph. This is needed sometimes when selecting the "best" spanning tree, but the selection criterion itself is very computationally intensive. In such a situation, solving the optimization problem may become practically infeasible. As noted in [1], the number of all spanning trees of a complete connected undirected graph with n vertices is n^{n-2}, the formula derived back in the 19th century. According to the above formula, the number of spanning trees grows very rapidly when increasing n. This growth rate is valid not only for complete but also for all other types of graphs. Therefore, an efficient method is needed to generate an exhaustive set of all trees mentioned, without repetitions. Such methods have been proposed, but unfortunately, they do not always yield the solution in practice due to their rather high complexity.

6. A HEURISTIC ALGORITHM FOR CONSTRUCTING THE MINIMAL SPANNING TREE

Due to the discussion above, for the problem under consideration, it is desirable to have approximate solution methods more acceptable for practical implementation instead of exact ones.

Note that the problems addressed here for a fuzzy graph fundamentally differ from the analogous problem for a crisp graph in the sense of multicriteria optimization [14]. In most real applications, optimization has to be performed by many criteria; in the general case, they can be formulated as follows:

$$\max\{z_1 = f_1(x), \ldots, z_q = f_q(x)\}, \quad g_i(x) \geqslant 0, \quad i = 1, \ldots, m, \quad x \geqslant 0.$$

Here, the search space of solutions is given by

$$S = \{x \in R^n | g_i(x) \geqslant 0, \ i = 1, \ldots, m, \ x \geqslant 0\}.$$

When solving multicriteria problems, one seeks a solution that is better than the others. However, such a best solution by all criteria does not necessarily exist. This occurs due to various conflicts; obviously, a solution may be best by one criterion and, at the same time, worst by others. Therefore, in multicriteria optimization, Pareto optimal solutions are often constructed [15, 16].

In the problems under consideration, we have two criteria, namely, the realizability value of the spanning tree (to be maximized) and the spanning tree length (to be minimized). Multicriteria optimization is a natural extension of ordinary numerical or combinatorial optimization. Therefore, many well-developed methods have been generalized to this class of problems. We emphasize that the central issue of multicriteria optimization is the construction of an objective function. Over the past decades, several approaches to this problem have been developed, including vector evaluation, Pareto ranking, and the use of weighted sums.

Note that the weighted sum approach has been popular in recent years: the new (overall) objective function is constructed from partial objective functions as a weighted sum of the form

$$F(x) = \sum_{i=1}^{k} w_i f_i(x), \quad w_i \in [0, 1], \quad \sum_{i=1}^{k} w_i = 1.$$

Here, each objective function $f_i(x)$ is assigned a particular weight, and the original problem reduces to the scalar case. Different weights w_i yield different Pareto-optimal solutions. Clearly, in view of the diverse nature of partial objective functions, the choice of appropriate weights is a quite complex task.

The approach proposed below will also reduce the original problem to the scalar case, but with a different method.

According to [2, Ch. 23], several algorithms have been developed for finding the minimal spanning tree for a crisp graph. Prim's and Kruskal's algorithms are among the most popular ones. Both are greedy algorithms: their strategy is to choose the best alternative at the current iteration. Generally speaking, this strategy does not ensure the absolute optimum. However, the algorithms have been proven to construct minimal spanning trees. What is also important, for a crisp graph $G(X, E)$, both algorithms run in an estimated time of $O(X \lg E)$ [2], which is a good performance. Based on the above considerations, an acceptable solution would be reducing the original problem for a fuzzy graph to the same problem for a crisp one and applying Prim's or Kruskal's algorithm.

This idea can be implemented by proposing a suitable method for transforming a given fuzzy graph to a crisp one. It seems that such a transformation should preserve the structure of the original graph (the sets of vertices and edges, as well as the connections between them) but suitably replace the weight of each edge (an ordered pair of values) with a single scalar value. This essentially

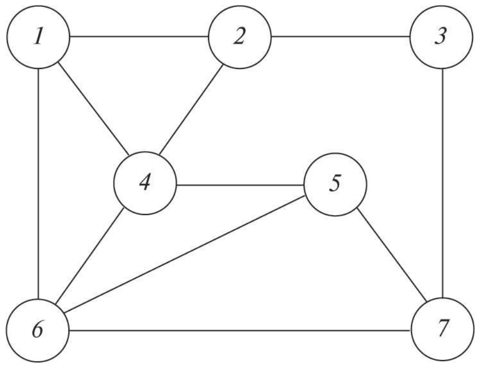

Fig. 1. An example of a fuzzy graph.

means that the scalar value should somehow reflect the information about the edge's probability of belonging and its length.

Clearly, such a transformation should not be computationally-intensive, i.e., should be limited to fairly simple operations over weights. From the standpoint of simplicity, arithmetic operations are suitable as one example. Since these components have different and sometimes incompatible natures, addition, subtraction, and division are often of little sense here; and therefore, multiplication is the most suitable operation.

Below, we propose a heuristic algorithm for constructing the minimal spanning tree of a fuzzy graph based on the idea expressed. In a fuzzy graph $G(X, E)$, each edge (x, y) belongs to it with a given probability $P(x, y)$. This fact can be interpreted as follows: under a sufficiently large number N of trials (simulations of transport flows in the network), the entire road (x, y) will be used not N but only $P(x, y)N$ times. Similarly, in each trial (a simulation of a particular transport route), one can conditionally assume that the road (x, y) of length $a(x, y)$ will be used not entirely but partially, to the length $P(x, y)\ a(x, y)$. When conducting N trials, the total path length on road (x, y) will equal $N(P(x, y)\ a(x, y))$. Then, on average per trial, the path traveled will be $P(x, y)\ a(x, y)$. In this case, the above part can be treated as the "length of edge" (x, y) "weighted" considering its absolute length and probability of belonging. Thus, the weight of an edge in the original fuzzy graph $G(X, E)$, representing an ordered pair, will be replaced by a single number, and the graph becomes crisp. Note that the replacement of the length $a(x, y)$ of an edge with $P(x, y)\ a(x, y)$ cannot, of course, claim to be rigorously justified. It is only an attempt to explain the motivation behind the heuristic change of graph parameters.

After obtaining a crisp graph by the heuristic algorithm, we can apply Prim's and Kruskal's algorithms to this graph and construct a minimal spanning tree. After construction, for each edge in the resulting spanning tree, we restore as its weight the ordered pair corresponding to that edge in the original fuzzy graph. The minimal spanning tree of the fuzzy graph obtained in this way will serve as an approximation to the minimal one.

Obviously, the heuristic algorithm does not surely yield an optimal spanning tree by the adopted minimality criterion. Generally speaking, this algorithm constructs a spanning tree that is minimal (by length), but its realizability value may be non-maximum. As advantages, we note the simplicity and low computational complexity of the algorithm.

To assess the effectiveness of this algorithm, we need sufficient statistical data on the results of its application, which are currently unavailable. According to its limited testing, the heuristic algorithm surely constructs the minimal spanning tree, and the deviations of its realizability values from the maximum are relatively insignificant. Let us illustrate the operation of the algorithm with a small example. Consider a fuzzy graph in Fig. 1; its parameters are combined in the table below.

The parameters of a fuzzy graph

No.	Edge (x, y)	The probability of the belonging of edge, $Pbe(x,y)$	Edge length $a(x,y)$	Replacement length $Pbe(x,y) \times a(x,y)$
1	(1,2)	0.099	70	6.93
2	(1,4)	0.099	60	5.94
3	(1,6)	0.099	50	4.95
4	(2,3)	0.096	30	2.88
5	(2,4)	0.089	30	2.67
6	(3,7)	0.092	50	4.60
7	(4,5)	0.089	20	1.78
8	(4,6)	0.080	40	3.20
9	(5,6)	0.094	40	3.76
10	(5,7)	0.091	60	5.46
11	(6,7)	0.072	90	6.48

A separate column of the table presents the values of $P(x,y)\ a(x,y)$, which replace the edge weights (ordered pairs) of the fuzzy graph with a single number. Using these data, two different spanning trees were constructed by Prim's algorithm. Figure 2 shows the tree constructed starting from vertex 1; Fig. 3, the tree constructed starting from vertex 7. As is easily verified, both trees have the same length, equal to 220, and are minimal for the corresponding crisp graph. A fuzzy spanning tree corresponds to each of these trees. According to the table, we assign the probabilities of belonging to all edges of these graphs; then the corresponding realizability values can be calculated. For the graph in Fig. 2, the realizability value is 0.545; for the graph in Fig. 3, 0.559.

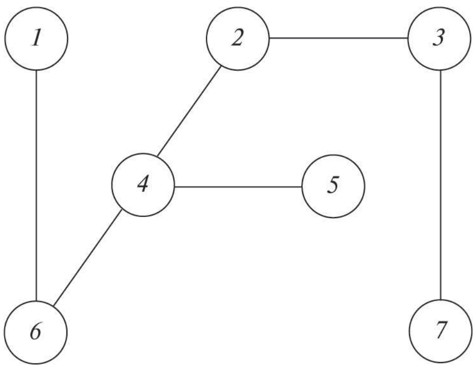

Fig. 2. The first variant of the tree.

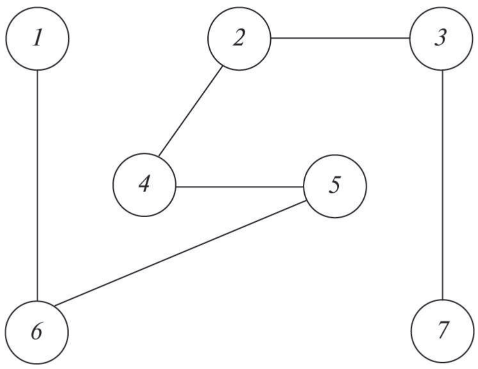

Fig. 3. The second variant of the tree.

Therefore, the heuristic algorithm may yield a spanning tree of non-maximal realizability. Note that the minimal spanning tree of the crisp graph obtained by Prim's algorithm based on the table data has the same structure as the one in Fig. 2 and the same length of 220.

7. CONCLUSIONS

In this paper, we have conceptualized the realizability of a spanning tree of a fuzzy graph and, using this concept, formulated two optimization problems for constructing optimal spanning trees. Previously, analogous problems in such a statement have not been considered in the literature. The optimality criteria used are the maximum value of tree realizability and the minimality of the spanning tree (by its length). An exact solution method for the two problems has been proposed, which suffers from high complexity. In this regard, a heuristic algorithm has been described to construct a minimal spanning tree considering its realizability. The algorithm is fairly simple and has low computational complexity, but does not ensure maximum realizability. This algorithm is based on reducing the original multicriteria problem to a single-criterion one (spanning tree minimality), which is efficiently solved, e.g., by Prim's and Kruskal's algorithms.

REFERENCES

1. Christofides, N., *Graph Theory: An Algorithmic Approach*, New York: Academic Press, 1975.
2. Cormen, T.H., Leiserson, C.E., Rivest, R.L., and Stein, C., *Introduction to Algorithms*, 3rd ed., Cambridge, MA: MIT Press, 2009.
3. Minieka, E., *Optimization Algorithms for Networks and Graphs*, New York: Marcel Dekker, 1978.
4. Prim, R.C., Shortest Connection Network and Some General Realization, *Bell System Technical Journal*, 1957, vol. 36, pp. 1389–1401.
5. Kruskal, J.B., On the Shortest Spanning Subtree of a Graph and Traveling Salesman Problem, *Proc. AMS*, 1956, vol. 7, pp. 48–50.
6. Kuz'min, R.A. and Molchanov, V.A., An Algorithm for Constructing Minimal Spanning Trees in Computer Networks, in *Matematika. Mekhanika* (Mathematics. Mechanics), 2023, no. 25, pp. 46–50.
7. Meshkov, A.N., Construction of a Spanning Tree Using Steiner's Genetic Algorithm, *Trudy V Mezhdunarodnoi nauchno-prakticheskoi konferentsii "Aktual'nye voprosy sovremennykh issledovanii"* (Proceedings of the 5th International Scientific and Practical Conference "Topical Issues in Contemporary Research"), July 5, 2023, Penza: Nauka i Prosveshchenie, 2023, pp. 74–76.
8. Zadeh, L.A., Fuzzy Sets, *Information and Control*, 1965, vol. 8, no. 3, pp. 338–353.
9. Kaufmann, A., *Introduction to the Theory of Fuzzy Subsets. Vol. 1*, New York: Academic Press, 1975.
10. Norwich, A.M. and Turksen, I.B., The Construction of Membership Functions, in *Fuzzy Sets and Their Applications to Cognitive and Decision Processes*, New York: Academic Press, 1975.
11. Qaid, W.A.A., Methods for Constructing Membership Functions of Fuzzy Sets, *Izvestiya SFedU. Engineering Sciences*, 2013, no. 2(139), pp. 144–153.
12. Fedorova, E.V. and Gryaznov, S.M., Features of Forming Membership Functions in the Construction of Fuzzy Control Systems, *Trudy III Vserossiiskoi nauchno-tekhnicheskoi konferentsii "Intellektual'nye sistemy upravleniya i mekhatronika. 2017"* (Proceedings of the 3rd All-Russian Scientific and Technical Conference "Intelligent Control Systems and Mechatronics. 2017"), 2017, pp. 170–174.
13. Cochran, W.G., *Sampling Techniques*, 3rd ed., New York: John Wiley & Sons, 1977.
14. Keeney, R.L. and Raiffa, H., *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*, New York: John Wiley & Sons, 1976.
15. Pareto, V., *Manual di Economica Politica*, Societa Editrice Libraia, Milan, Italia, 1906; translated into English by A.S. Schwier under the title *Manual of Political Economy*, New York: Macmillan, 1971.
16. Gen, M., Cheng, R., and Lin, L., *Network Models and Optimization*, London: Springer-Verlag, 2008.

This paper was recommended for publication by O.N. Granichin, a member of the Editorial Board

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2026, Vol. 87, No. 3, pp. 227–240.

CONTROL IN TECHNICAL SYSTEMS

Modification of the Electrodynamic Three-Axis Attitude Stabilization Method for an Artificial Earth Satellite in the Orbital Frame

M. V. Maksimenko[*,a] **and A. A. Tikhonov**[*,b]

[*] *St. Petersburg State University, St. Petersburg, Russia*
e-mail: [a] *m.v.maksimenko@spbu.ru,* [b] *a.tikhonov@spbu.ru*

Received May 5, 2025
Revised November 5, 2025
Accepted December 1, 2025

Abstract—This paper considers the three-axis attitude stabilization problem of an artificial Earth satellite in an arbitrary (indirect) position in the orbital frame. The problem is solved using an electrodynamic control system that generates the Lorentz torque and the magnetic interaction torque. The control torques contain restoring, dissipative, and compensating components. A new method for forming dissipative torques is proposed, which can be implemented using the available control actuators. The Lyapunov function method is applied to obtain sufficient conditions for the asymptotic stability of the satellite's program motion without imposing constraints on the geomagnetic field model or the satellite's orbit inclination. A combined electrodynamic control is presented to reduce the stabilization time and decrease the amplitude of transient oscillations. The results of numerical experiments are provided to demonstrate the effectiveness and performance of the modified satellite stabilization method.

Keywords: satellite, electrodynamic control, three-axis stabilization, modified control, Lyapunov function, asymptotic stability

DOI: 10.7868/S1608303226030043

1. INTRODUCTION

The interaction of an artificial Earth satellite with the geomagnetic field is a source of torques, which are successfully used to solve a wide range of problems of controlled satellite motion dynamics about its center of mass [1, 2]. This class of problems includes three-axis satellite attitude stabilization in the orbital frame, considered below. The interaction of the geomagnetic field with magnetic coils mounted on the satellite is regarded as a simple, reliable, and efficient way of solving this problem [2–7] and more complex ones requiring other program modes of the satellite's angular motion [8]. Despite the global controllability and observability of magnetic control systems [9], it is necessary to consider their drawback, i.e., the impossibility of generating a control torque in the direction of the geomagnetic field induction vector [4]. The same drawback appears in another well-known control system based on the interaction of an electric charge with the geomagnetic field [10–14]. However, combining the capabilities of these two principles into a unified electrodynamic control system (EDCS) allows achieving better performance compared to each of the two systems used separately [15].

EDCSs were applied to solve several satellite dynamics problems: one-axis and three-axis satellite attitude stabilization in the orbital and Koenig frames, dual-spin satellite stabilization, and others [16]. However, many of the above problems were solved using particular assumptions, e.g., concerning the choice of the geomagnetic field model, the shape and inclination of the satellite's orbit.

In three-axis attitude stabilization problems, the program mode is usually the direct equilibrium position of the satellite in the orbital frame, i.e., the position where its principal central axes of inertia coincide with those of the orbital frame. At the same time, a topical problem is to stabilize a satellite in indirect equilibrium positions, which is fundamentally more complex since in such positions, the significant gravitational torque becomes a disturbance. The problem of stabilizing an Earth-surveying satellite in an oblique position in the orbital frame in a contingency emergency mode was solved in [17]; in particular, the issues of digital control of the magnetic actuator at all transition stages of the satellite attitude control system to the emergency mode and subsequent long-term maintenance of this mode were studied in detail.

Also, there are various known methods for implementing passive or combined satellite attitude stabilization systems without propellant or energy consumption (or with small consumption), as they involve gravitational and aerodynamic torques simultaneously [18–20].

Such a satellite attitude stabilization approach naturally leads to the implementation of indirect equilibrium positions of the satellite in the orbital frame as operation modes. In the case, the problems solved include selecting the most suitable nominal attitude modes, avoiding undesirable resonance effects [21], as well as stabilizing the chosen nominal modes using additional means such as flywheels and active magnetic systems [22, 23]. Consequently, other methods involving the Earth's magnetic field may also be useful. The problem of stabilizing a satellite in indirect equilibrium positions using an EDCS was considered in the paper [24]. The control laws were constructed in accordance with the classical approach, i.e., creating restoring, dissipative, and compensating torques in the control system [25]. Within this approach, the dissipative torque is generated as a linear function of the satellite's relative angular velocity.

In this paper, we propose two modifications of the dissipative torque and their adaptive combination. The high performance of the combined method for forming the dissipative torque is demonstrated. No constraints are imposed on the choice of a particular geomagnetic field model or the satellite's orbit inclination. The study is based on a nonlinear mathematical model, employing the Lyapunov function method to prove the asymptotic stability of the program motion.

2. ELECTRODYNAMIC CONTROL OF SATELLITE'S ANGULAR MOTION

2.1. Satellite's Program Motion

Consider the motion of a satellite as a rigid body about its center of mass. The satellite moves in a Keplerian circular orbit of arbitrary inclination i and radius R in the Earth's magnetic field (EMF). By assumption, the satellite is equipped with a controlled electrostatic charge Q, which is distributed over a certain volume V with a density σ, and a controlled intrinsic magnetic moment $\vec{I}$.

To describe the satellite attitude, we use the following right orthogonal Cartesian frames (Fig. 1).

1) $O_E X^* Y^* Z^*$, the inertial frame where the $O_E X^*$ axis is directed to the ascending node of the orbit and the $O_E Z^*$ axis is along the Earth's diurnal rotation axis.

2) $C\xi\eta\zeta$, the orbital frame with origin at the satellite's center of mass where the $C\xi$ axis (the unit vector $\vec{\xi}_0$) is along the tangent to the orbit in the direction of motion, the $C\eta$ axis (the unit vector $\vec{\eta}_0$) is along the orbit normal, and the $C\zeta$ axis (the unit vector $\vec{\zeta}_0$) is along the local vertical.

3) $Cxyz$, the body-fixed frame whose axes are along the principal central axes of inertia of the satellite.

The relative position of the frames $Cxyz$ and $C\xi\eta\zeta$ is determined by the direction cosine matrix $\mathbf{A} = \{a_{ij}\}_{i,j=1}^{3}$; its elements can be expressed through the "aircraft" angles ψ, θ, and φ according to the following representations of the unit vectors $\vec{\xi}_0$, $\vec{\eta}_0$, and $\vec{\zeta}_0$ of the orbital frame in the

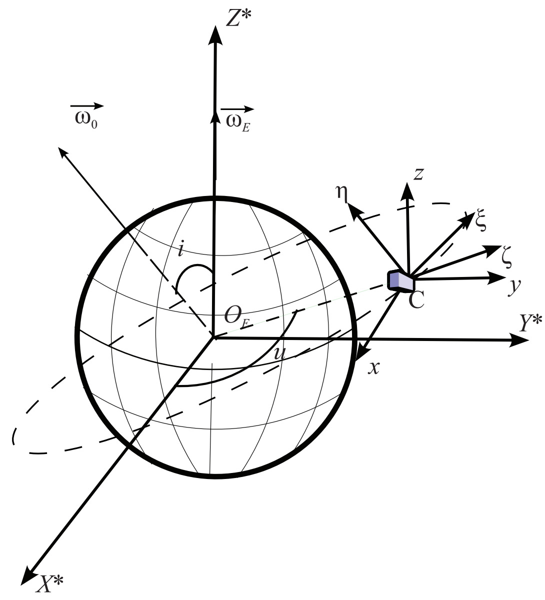

Fig. 1. Frames.

frame $Cxyz$:

$$\vec{\xi}_0 = (\cos\psi\cos\theta,\ \cos\psi\sin\theta\sin\varphi - \sin\psi\cos\varphi,\ \sin\psi\sin\varphi + \cos\psi\sin\theta\cos\varphi)^\top,$$
$$\vec{\eta}_0 = (\sin\psi\cos\theta,\ \cos\psi\cos\varphi + \sin\psi\sin\theta\sin\varphi,\ \sin\psi\sin\theta\cos\varphi - \cos\psi\sin\varphi)^\top,$$
$$\vec{\zeta}_0 = (-\sin\theta,\ \sin\varphi\cos\theta,\ \cos\varphi\cos\theta)^\top.$$

This study takes into account the rotation of the orbital frame relative to the inertial one with the angular velocity $\vec{\omega}_0 = \omega_0\vec{\eta}_0$ and the Earth's rotation with the angular velocity $\vec{\omega}_\mathrm{E}$ (see Fig. 1). Therefore, the absolute angular velocity $\vec{\omega}$ of the satellite can be written as

$$\vec{\omega} = \vec{\omega}' + \vec{\omega}_0, \tag{1}$$

where $\vec{\omega}'$ is its angular velocity relative to the orbital frame. The velocity vector $\vec{v}_C$ of the satellite's center of mass relative to the rotating Earth is expanded in terms of the unit vectors $\vec{\xi}_0$, $\vec{\eta}_0$, and $\vec{\zeta}_0$ as follows:

$$\vec{v}_C = \left(R\left(\omega_0 - \omega_\mathrm{E}\cos i\right),\ R\omega_\mathrm{E}\sin i\,\cos u,\ 0\right)^\top,$$

where $u = \omega_0 t$ is the latitude argument, which will be used below as a dimensionless independent variable.

The mass distribution of the satellite is characterized by the inertia tensor $\mathbf{J} = \mathrm{diag}(A, B, C)$ in the principal central axes of inertia $Cxyz$.

The program attitude of the satellite in the orbital frame is defined by the matrix $\mathbf{A}_0 = \{a_{ij}(\varphi_0, \theta_0, \psi_0)\}_{i,j=1}^3$. Accordingly, it is required to design control torques ensuring the existence of an asymptotically stable program motion of the satellite about its center of mass such that

$$\varphi = \varphi_0, \quad \theta = \theta_0, \quad \psi = \psi_0, \quad \vec{\omega}' = \vec{0}. \tag{2}$$

2.2. Control Torques and Controlled Vectors

When the satellite moves relative to the EMF with magnetic induction $\vec{B}$, the Lorentz torque $\vec{M}_L$ [26] and the magnetic interaction torque $\vec{M}_M$ are excited. They have the form

$$\vec{M}_L = \vec{P} \times \vec{T}, \qquad \vec{M}_M = \vec{I} \times \mathbf{A}^\top \vec{B}, \tag{3}$$

where $\vec{P} = Q\vec{\rho}_0$, $\vec{\rho}_0$ is the radius vector of the satellite's charge center relative to its center of mass, and $\vec{T} = \mathbf{A}^\top(\vec{v}_C \times \vec{B})$. From this point onwards, the controlled vectors $\vec{P}$ and $\vec{I}$ are supposed to be given in the frame $Cxyz$. The torques (3) are the control ones in the electrodynamic control system for the satellite's angular motion.

Note that for a given satellite orbit, the magnetic induction $\vec{B}$ in (3) is a known function of the coordinates of a near-Earth space point. Assume that the EMF is uniform in the part of space occupied by the satellite, and the vector $\vec{B} = B_\xi\vec{\xi}_0 + B_\eta\vec{\eta}_0 + B_\zeta\vec{\zeta}_0$ takes the same value at points of this space as at the satellite's center of mass (point C).

In the program mode (2) of the satellite's rotational motion, the vectors $\vec{T}$ and $\vec{B}$ take the form

$$\vec{T}_0 = \mathbf{A}_0^\top(\vec{v}_C \times \vec{B}), \qquad \vec{B}_0 = \mathbf{A}_0^\top\vec{B}.$$

In the electrodynamic control system, the variation laws of the controlled vectors $\vec{P}$ and $\vec{I}$ must be chosen to implement the program mode of the satellite's rotational motion. Let us design each of them as a sum of the restoring, dissipative, and compensating components:

$$\vec{P} = \vec{P}_{rest} + \vec{P}_{diss} + \vec{P}_{comp}\,, \qquad \vec{I} = \vec{I}_{rest} + \vec{I}_{diss} + \vec{I}_{comp}. \tag{4}$$

In view of (4), the control torques (3) become

$$\begin{aligned} \vec{M}_L &= (\vec{P}_{rest} + \vec{P}_{diss} + \vec{P}_{comp}) \times \vec{T}, \\ \vec{M}_M &= (\vec{I}_{rest} + \vec{I}_{diss} + \vec{I}_{comp}) \times \mathbf{A}^\top\vec{B}. \end{aligned} \tag{5}$$

3. MATHEMATICAL MODEL

3.1. The System of Equations Describing Satellite's Motion

Under the gravitational (disturbing) torque and the control torques (3), the motion of the satellite (a rigid body) about its center of mass obeys the dynamic Euler equations

$$\frac{d}{dt}(\mathbf{J}\vec{\omega}) + \vec{\omega} \times (\mathbf{J}\vec{\omega}) = \vec{M}_{GR} + \vec{M}_L + \vec{M}_M, \tag{6}$$

where $\vec{M}_{GR} = 3\omega_0^2(\mathbf{A}^\top\vec{\zeta}_0) \times (\mathbf{J}\mathbf{A}^\top\vec{\zeta}_0)$ is the gravitational torque [1].

Substituting the absolute angular velocity (1), the gravitational torque, and the control torques into (6) yields

$$\begin{aligned} \frac{d}{dt}\left[\mathbf{J}\left(\vec{\omega}_0\mathbf{A}^\top\vec{\eta}_0 + \vec{\omega}'\right)\right] &+ \left[\vec{\omega}_0\mathbf{A}^\top\vec{\eta}_0 + \vec{\omega}'\right] \times \left[\mathbf{J}\left(\vec{\omega}_0\mathbf{A}^\top\vec{\eta}_0 + \vec{\omega}'\right)\right] \\ &= 3\omega_0^2(\mathbf{A}^\top\vec{\zeta}_0) \times (\mathbf{J}\mathbf{A}^\top\vec{\zeta}_0) + (\vec{P}_{rest} + \vec{P}_{diss} + \vec{P}_{comp}) \times \vec{T} \\ &\quad + (\vec{I}_{rest} + \vec{I}_{diss} + \vec{I}_{comp}) \times \mathbf{A}^\top\vec{B}. \end{aligned} \tag{7}$$

To close the system of differential equations (7), we consider them together with the kinematic Poisson equations

$$\frac{d\vec{\xi}_0}{dt} = \vec{\xi}_0 \times \vec{\omega} - \omega_0\vec{\zeta}_0, \quad \frac{d\vec{\eta}_0}{dt} = \vec{\eta}_0 \times \vec{\omega}, \quad \frac{d\vec{\zeta}_0}{dt} = \vec{\zeta}_0 \times \vec{\omega} + \omega_0\vec{\xi}_0. \tag{8}$$

3.2. Disturbing Torque Compensation

The program mode (2) of the satellite's angular motion is generally not a direct equilibrium position of the satellite in the orbital frame. Therefore, the gravitational torque and the expression $\omega_0^2(\mathbf{A}^\top\vec{\eta}_0) \times (\mathbf{J}\mathbf{A}^\top\vec{\eta}_0)$ from equations (7) do not vanish in the satellite's program motion. To compensate for the total disturbing torque

$$\vec{M}_d = \omega_0^2 \left(3(\mathbf{A}^\top\vec{\zeta}_0) \times (\mathbf{J}\mathbf{A}^\top\vec{\zeta}_0) - (\mathbf{A}^\top\vec{\eta}_0) \times (\mathbf{J}\mathbf{A}^\top\vec{\eta}_0)\right),$$

we will utilize an approach based on creating the corresponding compensating components of the electrodynamic parameters [27] to ensure the equality $\vec{M}_d + \vec{P}_{comp} \times \vec{T} + \vec{I}_{comp} \times (\mathbf{A}^\top\vec{B}) = \vec{0}$. The restoring and compensating components of the controlled vectors $\vec{P}$ and $\vec{I}$ for implementing electrodynamic control to stabilize the program motion of the satellite are given by

$$\vec{P}_{rest} = Qk_L\vec{T}_0, \quad \vec{P}_{comp} = \frac{\vec{M}_d \cdot (\mathbf{A}^\top\vec{B})}{|\vec{B}|\,|\vec{T}|}\left(\mathbf{V}^\top\,(0,0,1)^\top\right),$$

$$\vec{I}_{rest} = k_M\vec{B}_0, \quad \vec{I}_{comp} = \mathbf{V}^\top\left(0, \frac{\vec{M}_d \cdot \left(\left(\mathbf{A}^\top\vec{B}\right) \times \vec{T}\right)}{|\vec{B}|^2\,|\vec{T}|}, -\frac{\vec{M}_d \cdot \vec{T}}{|\vec{B}|\,|\vec{T}|}\right)^\top.$$

Here, k_L and k_M are tunable control parameters (constants or functions of time), and $\mathbf{V} = \{v_{ij}\}_{i,j=1}^3$ is a known matrix with the elements

$$v_{11} = B_x/|\vec{B}|, \quad v_{12} = B_y/|\vec{B}|, \quad v_{13} = B_z/|\vec{B}|,$$

$$v_{21} = T_x/|\vec{T}|, \quad v_{22} = T_y/|\vec{T}|, \quad v_{23} = T_z/|\vec{T}|,$$

$$v_{31} = \frac{B_yT_z - B_zT_y}{|\vec{B}||\vec{T}|}, \quad v_{32} = \frac{B_zT_x - B_xT_z}{|\vec{B}||\vec{T}|}, \quad v_{33} = \frac{B_xT_y - B_yT_x}{|\vec{B}||\vec{T}|}.$$

The choice of the dissipative components $\vec{P}_{diss}$ and $\vec{I}_{diss}$ of the controlled vectors will be justified below.

3.3. Decreasing the Satellite's Relative Angular Velocity

Satellite attitude stabilization implies not only achieving the desired orientation but also ensuring its asymptotic stability. For this purpose, dissipative components are provided in the controlled vectors (4); they can be chosen in various ways. For example, numerous passive devices, including well-established composite schemes in satellites like Vertistat, spherical magnetic dampers (both fluid and eddy-current), and hysteresis rods, generate dissipative torques proportional to the satellite's angular velocity (see [28] and the references therein).

In electrodynamic control systems, dissipative torques are typically created using semi-passive methods. In addition to linear control solutions [15, 29], nonlinear controls with respect to angular velocity are effective in many cases [30].

In control of mechanical systems, as well as in satellite angular motion control, angular momentum-based control laws are widely used [31–34]. In particular, a well-known satellite stabilization solution involves damping proportional to the satellite's angular momentum [6].

In previous works, we applied dissipative components with the control parameters h_L and h_M of the form

$$\vec{P}_{diss} = Qh_L\vec{\omega}' \times \vec{T}, \quad \vec{I}_{diss} = h_M\vec{\omega}' \times (\mathbf{A}^\top\vec{B}). \tag{9}$$

In this study, along with formulas (9), we propose three new variants of forming dissipative components. In the first variant, the relative angular velocity is replaced by $\vec{K} = \mathbf{J}\vec{\omega}'$, the angular momentum of the satellite in relative motion:

$$\vec{P}_{diss} = Qh_L \left(A^{-1}\mathbf{J}\vec{\omega}'\right) \times \vec{T}; \quad \vec{I}_{diss} = h_M \left(A^{-1}\mathbf{J}\vec{\omega}'\right) \times (\mathbf{A}^\top \vec{B}). \tag{10}$$

In the second variant of forming the dissipative component of the controlled vectors, the unit vector of $\vec{K}$ appears instead of $\vec{K}$:

$$\vec{P}_{diss} = \frac{Qh_L}{|\vec{K}|} \left(A^{-1}\mathbf{J}\vec{\omega}'\right) \times \vec{T}; \quad \vec{I}_{diss} = \frac{h_M}{|\vec{K}|} \left(A^{-1}\mathbf{J}\vec{\omega}'\right) \times (\mathbf{A}^\top \vec{B}). \tag{11}$$

Here, stabilization will be achieved formally over an infinite time. Therefore, to implement the corresponding control algorithm, the norm of the vector $\vec{K}$ near zero is set equal to a given small constant Δ, as soon as $|\vec{K}|$ decreases to this constant. In other words, we introduce the modified norm

$$|\vec{K}|_{mod} = \frac{|\Delta - |\vec{K}|| + \Delta + |\vec{K}|}{2}. \tag{12}$$

The third variant is a combination of (10) and (11); it will be justified in Section 5.

Note that compared to (9), the new variants (10) and (11) of forming the dissipative components and their combination do not require additional measurement devices in the control system.

By substituting (5) into (7), we obtain the following system of differential equations, to be considered together with (8):

$$\begin{gathered} \frac{d}{dt}\left[\mathbf{J}\left(\vec{\omega}' + \omega_0\vec{\eta}_0\right)\right] + \left(\vec{\omega}' + \omega_0\vec{\eta}_0\right) \times \left(\mathbf{J}\vec{\omega}'\right) + \omega_0\vec{\omega}' \times \left(\mathbf{J}\vec{\eta}_0\right) \\ = Qk_L\vec{T}_0 \times \vec{T} + \vec{P}_{diss} \times \vec{T} + k_M\vec{B}_0 \times \left(\mathbf{A}^\top\vec{B}\right) + \vec{I}_{diss} \times \left(\mathbf{A}^\top\vec{B}\right). \end{gathered} \tag{13}$$

Thus, the problem is to determine conditions on the parameters k_L, k_M, h_L, and h_M ensuring the asymptotic stability of the program motion (2).

4. MOTION STABILITY ANALYSIS

Let us derive asymptotic stability conditions for the program motion (2) using the variant (10) as more general compared to (9). Assuming the smallness of the angles $\varphi_1 = \varphi - \varphi_0$, $\theta_1 = \theta - \theta_0$, and $\psi_1 = \psi - \psi_0$ and their time derivatives, we expand the control torques on the right-hand side of equation (13) into power series in these small variables, discarding all terms of degree above 1, similar to the approach used in [24, 35]:

$$\begin{aligned} M^{(1)}_{L\nu_m} &= l_{m1}(t)(k_L\varphi_1 + h_L\dot{\varphi}_1) + l_{m2}(t)(k_L\theta_1 + \delta h_L\dot{\theta}_1) + l_{m3}(t)(k_L\psi_1 + \varepsilon h_L\dot{\psi}_1), \\ M^{(1)}_{M\nu_m} &= b_{m1}(t)(k_M\varphi_1 + h_M\dot{\varphi}_1) + b_{m2}(t)(k_M\theta_1 + \delta h_M\dot{\theta}_1) + b_{m3}(t)(k_M\psi_1 + \varepsilon h_M\dot{\psi}_1). \end{aligned}$$

Here $m = 1, 2, 3$, $\nu_1 = x$, $\nu_2 = y$, $\nu_3 = z$, $\delta = B/A$, and $\varepsilon = C/A$. We compile the symmetric matrices $\mathbf{S}_L(t) = \{l_{ij}\}_{i,j=1}^3$ and $\mathbf{S}_M(t) = \{b_{ij}\}_{i,j=1}^3$ with the elements l_{ij} and b_{ij}, respectively (see the Appendix). Then equations (13) turn into the differential equations of the perturbed motion in the three-axis satellite attitude stabilization problem:

$$\mathbf{J}\ddot{\vec{y}} + \mathbf{D}(t)\dot{\vec{y}} + \mathbf{G}\dot{\vec{y}} + \mathbf{H}(t)\vec{y} + \vec{F}(t, \vec{y}, \dot{\vec{y}}) = \vec{0}, \tag{14}$$

where $\vec{y} = \begin{pmatrix} \varphi_1 \\ \theta_1 \\ \psi_1 \end{pmatrix}$, $\mathbf{G} = \omega_0 (A - B + C) \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}$ is a constant matrix, $\mathbf{D}(t) = -A^{-1}\mathbf{J}\left(h_L \mathbf{S}_L(t) + h_M \mathbf{S}_M(t)\right)$ and $\mathbf{H}(t) = -\left(k_L \mathbf{S}_L(t) + k_M \mathbf{S}_M(t)\right)$ are matrices with time-varying coefficients, and the vector $\vec{F}$ nonlinearly depends on $\vec{y}$ and $\dot{\vec{y}}$.

Let us introduce the control parameters k_{L0}, k_{M0}, h_{L0}, and h_{M0} (positive constants) as follows:

$$k_{L0} = k_L Q |\vec{T}(t)|^2, \quad k_{M0} = k_M |\vec{B}(t)|^2, \quad h_{L0} = h_L Q |\vec{T}(t)|^2, \quad h_{M0} = h_M |\vec{B}(t)|^2.$$

Denoting their ratios by $\varepsilon_1 = k_{L0}/k_{M0}$, $\varepsilon_2 = h_{L0}/h_{M0}$, and $\mu = h_{M0}/k_{M0}$, we obtain

$$\mathbf{D}(t) = -\mu k_{M0} \left(A^{-1}\mathbf{J}\mathbf{U}_M(t) + \varepsilon_2 A^{-1}\mathbf{J}\mathbf{U}_L(t) \right),$$
$$\mathbf{H}(t) = -k_{M0} \left(\mathbf{U}_M(t) + \varepsilon_1 \mathbf{U}_L(t) \right),$$

where

$$\mathbf{U}_M(t) = \frac{\mathbf{S}_M(t)}{|\vec{B}(t)|^2}, \quad \mathbf{U}_L(t) = \frac{\mathbf{S}_L(t)}{Q|\vec{T}(t)|^2}.$$

The Lyapunov function can be chosen as a quadratic form with a constant $\chi > 0$:

$$V(t, \vec{y}, \dot{\vec{y}}) = \frac{1}{2}\vec{y}^\top \mathbf{D}(t)\vec{y} + \frac{\chi}{2}\dot{\vec{y}}^\top \mathbf{J}\dot{\vec{y}} + \vec{y}^\top \mathbf{J}\dot{\vec{y}} + \frac{\chi}{2}\vec{y}^\top \mathbf{H}(t)\vec{y}. \tag{15}$$

The time derivative of the function (15) along the trajectories of system (14) is given by

$$\begin{gathered} \dot{V}(t, \vec{y}, \dot{\vec{y}})|_{(14)} = -\chi \dot{\vec{y}}^\top \mathbf{D}(t)\dot{\vec{y}} - \vec{y}^\top \mathbf{G}\dot{\vec{y}} - \vec{y}^\top \mathbf{H}(t)\vec{y} - \chi \dot{\vec{y}}^\top \mathbf{G}\dot{\vec{y}} \\ + \frac{1}{2}\vec{y}^\top \dot{\mathbf{D}}(t)\vec{y} + \dot{\vec{y}}^\top \mathbf{J}\dot{\vec{y}} + \frac{\chi}{2}\vec{y}^\top \dot{\mathbf{H}}(t)\vec{y} - \left(\chi\dot{\vec{y}} + \vec{y}\right)^\top \vec{F}(t, \vec{y}, \dot{\vec{y}}). \end{gathered} \tag{16}$$

For any vectors $\vec{x}, \vec{z} \in \mathbb{R}^3$, the quadratic forms in the expressions (15) and (16) can be estimated as [36, 37]

$$\begin{gathered} \vec{x}^\top \mathbf{D}(t)\vec{x} \geqslant \mu k_{M0} \mathrm{inf}\lambda_{\min}\left(-A^{-1}\mathbf{J}\mathbf{U}_M(t) - \varepsilon_2 A^{-1}\mathbf{J}\mathbf{U}_L(t)\right)\|\vec{x}\|^2, \\ \vec{x}^\top \mathbf{H}(t)\vec{x} \geqslant k_{M0}\mathrm{inf}\lambda_{\min}\left(-\mathbf{U}_M(t) - \varepsilon_1 \mathbf{U}_L(t)\right)\|\vec{x}\|^2, \\ \min\{A, B, C\}\|\vec{x}\|^2 \leqslant \vec{x}^\top \mathbf{J}\vec{x} \leqslant \max\{A, B, C\}\|\vec{x}\|^2, \\ |\vec{x}^\top \mathbf{J}\vec{z}| \leqslant \max\{A, B, C\}\|\vec{x}\|\|\vec{z}\|, \\ \vec{x}^\top \dot{\mathbf{D}}(t)\vec{x} \leqslant \mu k_{M0}\mathrm{sup}\lambda_{\max}\left(-A^{-1}\mathbf{J}\dot{\mathbf{U}}_M(t) - \varepsilon_2 A^{-1}\mathbf{J}\dot{\mathbf{U}}_L(t)\right)\|\vec{x}\|^2, \\ \vec{x}^\top \dot{\mathbf{H}}(t)\vec{x} \leqslant k_{M0}\mathrm{sup}\lambda_{\max}\left(-\dot{\mathbf{U}}_M(t) - \varepsilon_1 \dot{\mathbf{U}}_L(t)\right)\|\vec{x}\|^2, \\ |\vec{x}^\top \mathbf{G}\vec{z}| \leqslant \omega_0 (A - B + C)\|\vec{x}\|\|\vec{z}\|, \quad \vec{x}^\top \mathbf{G}\vec{x} \leqslant \omega_0 (A - B + C)\|\vec{x}\|^2, \end{gathered} \tag{17}$$

where $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$ indicate the largest and smallest eigenvalues, respectively, of a matrix $\mathbf{M}$, and the operations sup and inf are performed for all $t \in \mathbb{R}$. In addition, we denote

$$\begin{gathered} d_1 = \min\{A, B, C\}, \quad d_2 = \max\{A, B, C\}, \quad d_3 = A - B + C, \\ d_{41} = \mathrm{inf}\lambda_{\min}\left[-A^{-1}\mathbf{J}\left(\mathbf{U}_M(t) + \varepsilon_2 \mathbf{U}_L(t)\right)\right], \ d_{42} = \mathrm{inf}\lambda_{\min}\left(-\mathbf{U}_M(t) - \varepsilon_1 \mathbf{U}_L(t)\right), \\ d_{51} = \mathrm{sup}\lambda_{\max}\left[-A^{-1}\mathbf{J}\left(\dot{\mathbf{U}}_M(t) + \varepsilon_2 \dot{\mathbf{U}}_L(t)\right)\right], \ d_{52} = \mathrm{sup}\lambda_{\max}\left(-\dot{\mathbf{U}}_M(t) - \varepsilon_1 \dot{\mathbf{U}}_L(t)\right). \end{gathered} \tag{18}$$

According to (17) and (18), the Lyapunov function satisfies the lower bound

$$V(t,\vec{y},\dot{\vec{y}}) \geqslant \frac{1}{2}\|\vec{y}\|^2\left(\mu k_{M0}d_{41}+\chi k_{M0}d_{42}-d_2a_1\omega_0^2\right)+\frac{1}{2}\|\dot{\vec{y}}\|^2\left(d_1\chi-\frac{d_2}{a_1\omega_0^2}\right),$$

where $a_1 > 0$ is some constant.

Hence, if

$$\chi > \frac{d_2}{d_1a_1\omega_0^2}, \quad \chi > \frac{d_2a_1\omega_0^2-\mu k_{M0}d_{41}}{k_{M0}d_{42}}, \tag{19}$$

the Lyapunov function (15) will be positive definite.

Furthermore, the numbers $c > 0$ and $\delta_0 > 0$ can be assigned so that

$$\begin{aligned}\dot{V}(t,\vec{y},\dot{\vec{y}})|_{(14)} &\leqslant \frac{1}{2}\|\vec{y}\|^2\left(a_2\omega_0^2d_3+\mu k_{M0}d_{51}+\chi k_{M0}d_{52}-2k_{M0}d_{42}\right)\\ &+\|\dot{\vec{y}}\|^2\left(d_2+\frac{d_3}{2a_2}-\chi\mu k_{M0}d_{41}\right)+c\left(\chi\|\dot{\vec{y}}\|+\|\vec{y}\|\right)\left(\|\dot{\vec{y}}\|^2+\|\vec{y}\|^2\right)\end{aligned}$$

for $t \geqslant 0$, $\|\vec{y}\| < \delta_0$, and $\|\dot{\vec{y}}\| < \delta_0$, where $a_2 > 0$ is some constant.

Thus, under the conditions

$$k_{M0} > \frac{a_2\omega_0^2d_3}{2d_{42}-\mu d_{51}-\chi d_{52}}, \quad \chi > \frac{d_2+d_3/(2a_2)}{\mu k_{M0}d_{41}}, \quad \chi < \frac{2d_{42}-\mu d_{51}}{d_{52}}, \tag{20}$$

the derivative of the function (16) along the trajectories of system (14) is negative definite.

Theorem 1. *Assume that for given satellite parameters A, B, C, and Q, orbit parameters R and i, and control parameters k_{L0}, k_{M0}, h_{L0}, and h_{M0}, there exist positive numbers a_1, a_2, and χ satisfying inequalities* (19) *and* (20). *Then the program motion* (2) *of system* (7), (8) *is asymptotically stable.*

The proof of this result is provided in the Appendix.

5. NUMERICAL EXPERIMENTS

To assess the effect of the above modifications of the dissipative torque on the stabilization process of the satellite's angular motion, we conducted a series of numerical experiments for the following parameters: $A = 1000\,\text{kg}\times\text{m}^2$, $B = 1300\,\text{kg}\times\text{m}^2$, $C = 700\,\text{kg}\times\text{m}^2$, and $Q = 15\times10^{-3}\,\text{C}$ (the satellite); $R = 7\times10^6\,\text{m}$ and $i = 30^\circ$ (the orbit); $k_{L0} = 2.5\times10^{-3}\,\text{N}\times\text{m}$, $k_{M0} = 2\times10^{-3}\,\text{N}\times\text{m}$, $h_{L0} = 0.2\,\text{N}\times\text{m}\times\text{s}$, and $h_{M0} = 1.0\,\text{N}\times\text{m}\times\text{s}$ (the control parameters).

The desired satellite attitude was given by the angles $\varphi_0 = 0.2\,\text{rad}$, $\psi_0 = 0.4\,\text{rad}$, and $\theta_0 = 0.6\,\text{rad}$.

In computer simulations, the geomagnetic field was modeled using an octupole approximation, and the Gaussian coefficients were determined according to the IGRF model of the 13th generation [38]. Note that for the electrodynamic control method proposed, the model's order of approximation is not limited: this method can be used with a multipole IGRF model of any order.

For the selected satellite, orbit, and control parameters satisfying the sufficient asymptotic stability conditions, the nonlinear differential equations (7) and (8) were integrated numerically.

Figures 2–5 show the satellite stabilization process for the following initial conditions: $\varphi(0) = -0.2\,\text{rad}$, $\psi(0) = 0.0\,\text{rad}$, $\theta(0) = 0.1\,\text{rad}$, $\Omega_x(0) = \omega_x(0)/\omega_0 = 0.3$, $\Omega_y(0) = \omega_y(0)/\omega_0 = 1.1$, and $\Omega_z(0) = \omega_z(0)/\omega_0 = 0.5$. In all figures, the horizontal axes represent the dimensionless variable u, for which $u = 5$ corresponds to $t = 4638.22\,\text{s}$ or approximately 1 h 17 min. Next, Fig. 2 shows the stabilization process under the modified (10) and original (9) dissipative components of the controlled vectors.

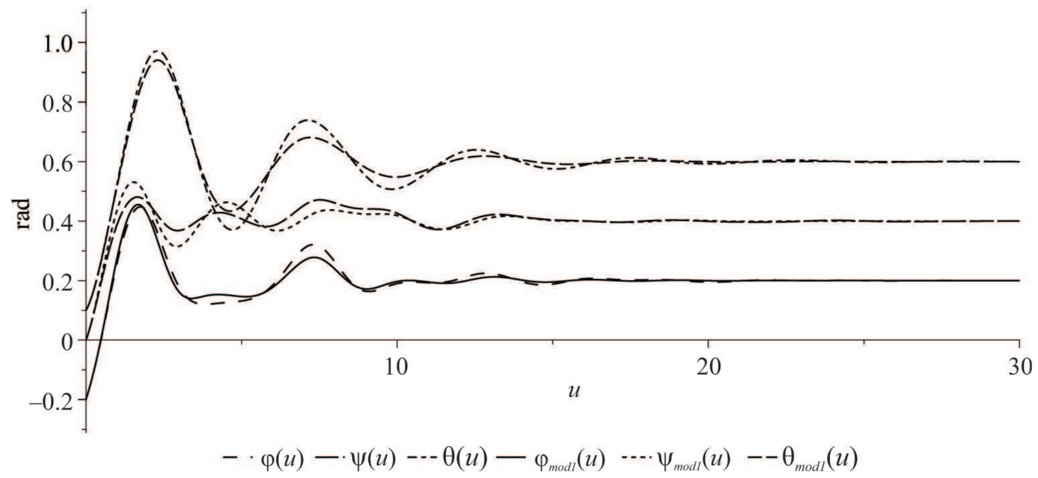

Fig. 2. The dynamics of the satellite attitude angles $\varphi(u)$, $\psi(u)$, and $\theta(u)$ under the variant (9) and $\varphi_{\mathrm{mod1}}(u)$, $\psi_{\mathrm{mod1}}(u)$, and $\theta_{\mathrm{mod1}}(u)$ under the variant (10).

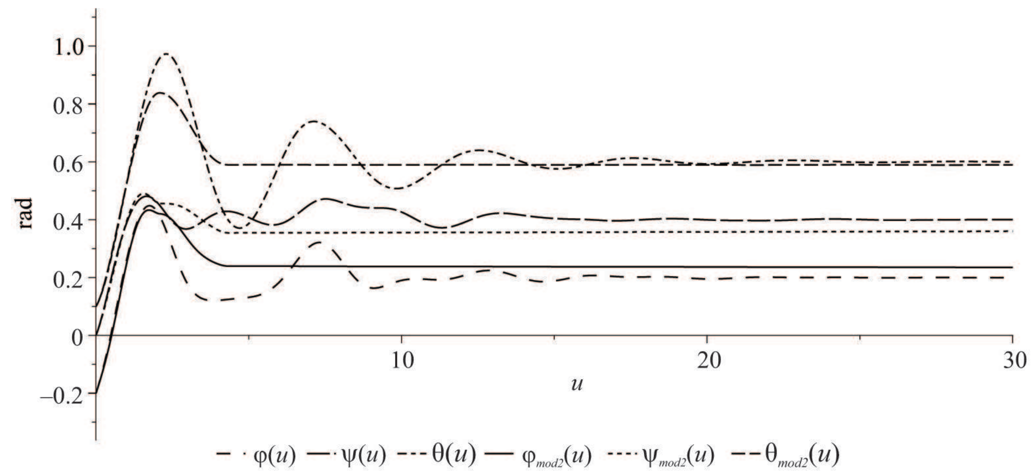

Fig. 3. The dynamics of the satellite attitude angles $\varphi(u)$, $\psi(u)$, and $\theta(u)$ under the variant (9) and $\varphi_{\mathrm{mod2}}(u)$, $\psi_{\mathrm{mod2}}(u)$, and $\theta_{\mathrm{mod2}}(u)$ under the variant (11).

By analogy, Fig. 3 presents the stabilization process under the modified (11) and original (9) dissipative components of the controlled vectors. The value $\Delta = 0.001$ was adopted for the numerical integration of the differential equations in (12).

Note the smoother satellite stabilization transients here. However, the desired attitude was achieved in a rather long time: starting from a certain time instant (Fig. 4), $|\vec{K}|$ became small, leading to a small denominator in formula (11) and a subsequent reduction in control performance.

To overcome this drawback, the approaches (10) and (11) were combined. Figure 5 shows the satellite stabilization process under the combination of the modified dissipative components (10) and (11) of the controlled vectors.

Thus, with combining the variants of forming the dissipative components of the controlled vectors (4), it is possible to stabilize the satellite in the orbital frame in the indirect equilibrium position (2) (in the general case) and, moreover, to obtain relatively smoother and faster-converging transients of the satellite stabilization process using the available control actuators.

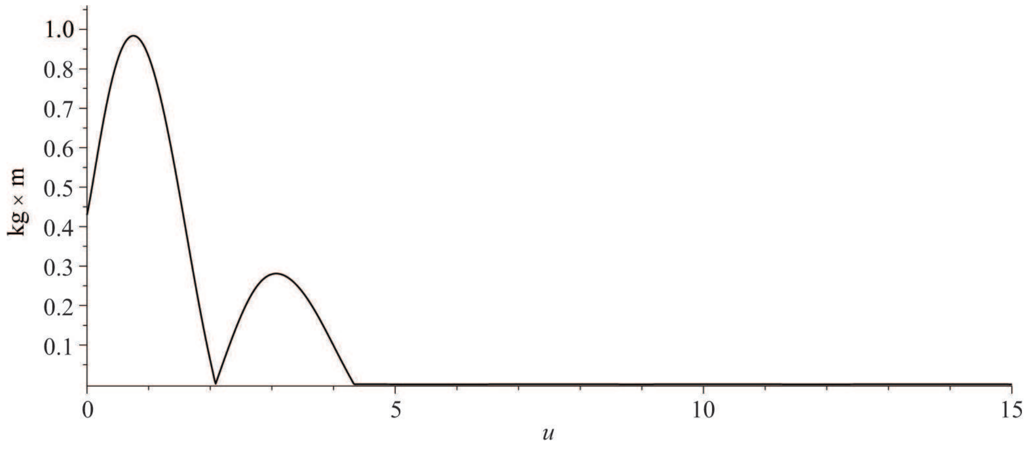

Fig. 4. The magnitude of the vector $\vec{K}$ under the variant (11).

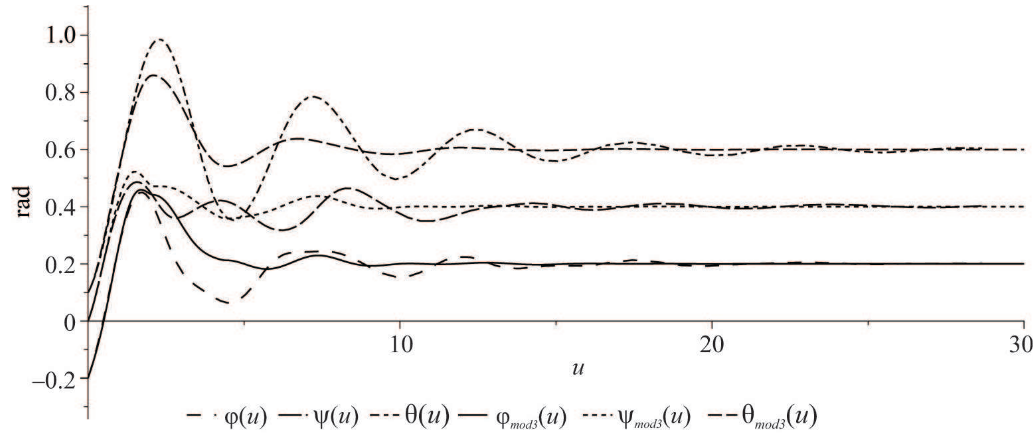

Fig. 5. The dynamics of the satellite attitude angles $\varphi(u)$, $\psi(u)$, and $\theta(u)$ under the variant (9) and $\varphi_{\mathrm{mod3}}(u)$, $\psi_{\mathrm{mod3}}(u)$, and $\theta_{\mathrm{mod3}}(u)$ under the combination of the variants (10) and (11).

During the numerical experiments, we also analyzed the realizability of the control law in terms of the bounded magnitudes of the controlled vectors (4). As established, with this choice of control parameters, the control torques (3) do not exceed the gravitational torque by the order of magnitude.

6. CONCLUSIONS

This paper has addressed the electrodynamic control problem for the angular motion of a charged satellite with an intrinsic magnetic moment, as well as its stabilization problem in the orbital frame in an arbitrary position. Three new variants of forming the dissipative torque have been proposed, which can be implemented using the available control actuators. The first two variants, considered separately, suffer from some drawbacks, which can be eliminated by combining them into a uniform system for forming the dissipative torque (the third variant). Such combined electrodynamic control possesses the advantages of the first and second variants, i.e., reduces the stabilization time and decreases the amplitude of transient oscillations. These performance characteristics are of paramount importance when the satellite carries equipment or devices sensitive to oscillations. The results of numerical experiments have demonstrated the effectiveness and performance of the modified satellite attitude control method.

Using the Lyapunov function method, sufficient stability conditions for the program motion of the satellite have been obtained and expressed explicitly in terms of the satellite, orbit, and control parameters.

FUNDING

This work was supported by the Russian Science Foundation, project no. 24-41-02031, https://rscf.ru/en/project/24-41-02031/.

APPENDIX

The symmetric matrices $\mathbf{S}_L(t) = \{l_{ij}\}_{i,j=1}^3$ and $\mathbf{S}_M(t) = \{b_{ij}\}_{i,j=1}^3$ with the elements l_{ij} and b_{ij} (see Section 4) have the following form:

$$\begin{aligned}
l_{11}(t) &= Q(-B_\zeta(\sin\psi_0 v_{C\xi} - \cos\psi_0 v_{C\eta})(-B_\eta v_{C\xi} + B_\xi v_{C\eta})\sin\theta_0 - B_\zeta^2 v_{C\xi}^2\cos\psi_0 - B_\zeta^2 v_{C\xi}v_{C\eta}\sin\psi_0\\
&\quad - \cos\theta_0(-B_\eta v_{C\xi} + B_\xi v_{C\eta})^2)\cos\varphi_0\\
&\quad - B_\zeta(B_\zeta v_{C\xi}(\sin\psi_0 v_{C\xi} - \cos\psi_0 v_{C\eta})\sin\theta_0\\
&\quad + (B_\xi v_{C\eta} - B_\eta v_{C\xi})(\cos\theta_0 v_{C\xi} - \cos\psi_0 v_{C\xi} - \sin\psi_0 v_{C\eta}))\sin\varphi_0,\\
l_{12}(t) &= l_{21}(t) = Qv_{C\eta}(((B_\eta v_{C\xi} - B_\xi v_{C\eta})\cos\theta_0 + B_\zeta\sin\theta_0(\cos\psi_0 v_{C\eta}\\
&\quad - \sin\psi_0 v_{C\xi}))\sin\varphi_0 - B_\zeta\cos\varphi_0(\cos\psi_0 v_{C\xi} + \sin\psi_0 v_{C\eta}))B_\zeta,\\
l_{13}(t) &= l_{31}(t) = Qv_{C\eta}(((B_\eta v_{C\xi} - B_\xi v_{C\eta})\cos\theta_0 + B_\zeta\sin\theta_0(\cos\psi_0 v_{C\eta}\\
&\quad - \sin\psi_0 v_{C\xi}))\cos\varphi_0 + B_\zeta\sin\varphi_0(\cos\psi_0 v_{C\xi} + \sin\psi_0 v_{C\eta}))B_\zeta;\\
l_{22}(t) &= -Q(-(B_\xi v_{C\eta} - B_\eta v_{C\xi})(B_\zeta\sin\theta_0(-\sin\psi_0 v_{C\xi} + \cos\psi_0 v_{C\eta}) - \cos\theta_0(B_\xi v_{C\eta} - B_\eta v_{C\xi}))\cos\varphi_0\\
&\quad + ((B_\xi v_{C\eta}^2 - B_\eta v_{C\eta}v_{C\xi})\sin\theta_0 + B_\zeta v_{C\eta}(\cos\psi_0 v_{C\eta} - \sin\psi_0 v_{C\xi})\cos\theta_0\\
&\quad - \sin\varphi_0(-B_\eta v_{C\xi} + B_\xi v_{C\eta})(\cos\psi_0 v_{C\xi} + \sin\psi_0 v_{C\eta}))B_\zeta),\\
l_{23}(t) &= l_{32}(t) = -Qv_{C\xi}((((B_\eta v_{C\xi} - B_\xi v_{C\eta})\cos\theta_0\\
&\quad + B_\zeta\sin\theta_0(\cos\psi_0 v_{C\eta} - \sin\psi_0 v_{C\xi}))\cos\varphi_0\\
&\quad + B_\zeta\sin\varphi_0(\cos\psi_0 v_{C\xi} + \sin\psi_0 v_{C\eta}))B_\zeta,\\
l_{33}(t) &= -Q(-v_{C\xi}((B_\eta v_{C\xi} - B_\xi v_{C\eta})\cos\theta_0\\
&\quad + B_\zeta\sin\theta_0(-\sin\psi_0 v_{C\xi} + \cos\psi_0 v_{C\eta}))\sin\varphi_0\\
&\quad + B_\zeta v_{C\eta}(-\sin\psi_0 v_{C\xi} + \cos\psi_0 v_{C\eta})\cos\theta_0 + (-B_\eta v_{C\eta}v_{C\xi} + B_\xi v_{C\eta}^2)\sin\theta_0\\
&\quad + B_\zeta\cos\varphi_0 v_{C\xi}(\cos\psi_0 v_{C\xi} + \sin\psi_0 v_{C\eta}))B_\zeta;\\
b_{11}(t) &= (-B_\zeta(B_\xi\cos\psi_0 + B_\eta\sin\psi_0)\sin\theta_0 + B_\xi B_\eta\sin\psi_0\\
&\quad - B_\eta^2\cos\psi_0 - B_\zeta^2\cos\theta_0)\cos\varphi_0 - ((B_\xi B_\eta\cos\psi_0 + B_\eta^2\sin\psi_0)\sin\theta_0\\
&\quad + B_\zeta(B_\xi\sin\psi_0 - \cos\psi_0 B_\eta + B_\eta\cos\theta_0))\sin\varphi_0,\\
b_{12}(t) &= b_{21}(t) = ((B_\zeta\cos\theta_0 + \sin\theta_0(B_\xi\cos\psi_0 + B_\eta\sin\psi_0))\sin\varphi_0\\
&\quad - \cos\varphi_0(B_\xi\sin\psi_0 - \cos\psi_0 B_\eta))B_\xi;\\
b_{13}(t) &= b_{31}(t) = B_\xi((B_\zeta\cos\theta_0 + \sin\theta_0(B_\xi\cos\psi_0 + B_\eta\sin\psi_0))\cos\varphi_0\\
&\quad + \sin\varphi_0(B_\xi\sin\psi_0 - \cos\psi_0 B_\eta)),\\
b_{22}(t) &= -B_\zeta(B_\zeta\cos\theta_0 + \sin\theta_0(B_\xi\cos\psi_0 + B_\eta\sin\psi_0))\cos\varphi_0 - (B_\xi^2\cos\psi_0\\
&\quad + B_\xi B_\eta\sin\psi_0)\cos\theta_0 + B_\zeta(\sin\theta_0 B_\xi - \sin\varphi_0(B_\xi\sin\psi_0 - \cos\psi_0 B_\eta)),\\
b_{23}(t) &= b_{32}(t) = B_\eta((B_\zeta\cos\theta_0 + \sin\theta_0(B_\xi\cos\psi_0 + B_\eta\sin\psi_0))\cos\varphi_0\\
&\quad + \sin\varphi_0(B_\xi\sin\psi_0 - \cos\psi_0 B_\eta)),\\
b_{33}(t) &= -B_\eta(B_\zeta\cos\theta_0 + \sin\theta_0(B_\xi\cos\psi_0 + B_\eta\sin\psi_0))\sin\varphi_0 - (B_\xi^2\cos\psi_0\\
&\quad + B_\xi B_\eta\sin\psi_0)\cos\theta_0 + \sin\theta_0 B_\xi B_\zeta + \cos\varphi_0 B_\eta(B_\xi\sin\psi_0 - \cos\psi_0 B_\eta).
\end{aligned}$$

Proof of Theorem 1. Resolving inequalities (19) and (20) with respect to χ, we write them as

$$
\begin{aligned}
&1) \quad \chi > \frac{d_2}{d_1 a_1 \omega_0^2},\\
&2) \quad \chi > \frac{d_2 a_1 \omega_0^2 - \mu k_{M0} d_{41}}{k_{M0} d_{42}},\\
&3) \quad \chi < \frac{2k_{M0} d_{42} - \mu k_{M0} d_{51} - a_2 \omega_0^2 d_3}{k_{M0} d_{52}} = \frac{2d_{42} - \mu d_{51}}{d_{52}} - \frac{a_2 \omega_0^2 d_3}{k_{M0} d_{52}},\\
&4) \quad \chi > \frac{d_2 + d_3/(2a_2)}{\mu k_{M0} d_{41}},\\
&5) \quad \chi < \frac{2d_{42} - \mu d_{51}}{d_{52}}.
\end{aligned}
\tag{A.1}
$$

Inequalities 1), 2), and 4) of system (A.1) define a lower bound for the number χ.

Let us denote $\chi_1 = \frac{d_2}{d_1 a_1 \omega_0^2}$, $\chi_2 = \frac{d_2 a_1 \omega_0^2 - \mu k_{M0} d_{41}}{k_{M0} d_{42}}$, and $\chi_4 = \frac{d_2 + d_3/(2a_2)}{\mu k_{M0} d_{41}}$. We emphasize that χ_1 and χ_2 depend on the parameter a_1, and as a_1 increases, $\chi_1(a_1)$ decreases hyperbolically whereas $\chi_2(a_1)$ increases linearly. Thus, to minimize $\max\{\chi_1, \chi_2\}$, the parameter a_1^* is chosen so that $\chi_1(a_1^*) = \chi_2(a_1^*)$.

Inequalities 3) and 5) of system (A.1) define an upper bound for the admissible values of the parameter χ. If $\frac{a_2 \omega_0^2 d_3}{k_{M0} d_{52}} > 0$, then inequality 5) follows from inequality 3). Let the desired bound be denoted by $\chi_3 = \frac{2k_{M0} d_{42} - \mu k_{M0} d_{51} - a_2 \omega_0^2 d_3}{k_{M0} d_{52}}$. Consequently, the system of inequalities (A.1) gets simplified to

$$
\max\{\chi_1(a_1^*), \chi_4\} < \chi_3.
$$

Similarly, we emphasize that χ_3 and χ_4 depend on the parameter a_2. Here, χ_3 decreases linearly whereas χ_4 hyperbolically. To determine χ, it is necessary to find a nonempty interval $\left[a_2^{(1)}, a_2^{(2)}\right]$ where the inequality $\chi_4 < \chi_3$ holds for every point inside it. From the equation $\chi_3(a_2) = \chi_4(a_2)$ we find the roots $a_2^{(1)}$ and $a_2^{(2)}$.

If the length of the desired interval $\left[a_2^{(1)}, a_2^{(2)}\right]$ tends to zero, it shrinks to a point a_2^*, and the value of a_2^* can be determined from the equality of the slopes of the line $\chi_3(a_2)$ and the tangent to the curve $\chi_4(a_2)$. Consequently, $a_2^* = \sqrt{d_{52}/(2\mu \omega_0^2 d_{41})}$. Such a choice of the constant a_2 can be recommended for numerical analysis of the sufficient asymptotic stability conditions. Finally, the system of inequalities (A.1) reduces to

$$
\max\{\chi_1(a_1^*), \chi_4(a_2^*)\} < \chi_3(a_2^*).
\tag{A.2}
$$

If inequality (A.2) fails because $\chi_1(a_1^*) > \chi_3(a_2^*)$, then one can choose an appropriate value of a_2 by reducing it to $a_2^{(1)}$.

The proof of Theorem 1 is complete.

REFERENCES

1. Beletsky, V.V., *Motion of an Artificial Earth Satellite about Its Center of Mass*, Jerusalem: Israel Program for Scientific Translations, 1966.
2. Sarychev, V.A. and Ovchinnikov, M.Yu., Magnetic Attitude Control Systems for Artificial Earth Satellites, in *Itogi nauki i tekhniki. Issledovanie kosmicheskogo prostranstva* (Results of Science and Technology. Space Exploration), vol. 23, Moscow: All-Soviet Institute for Scientific and Technical Information (VINITI), the USSR Academy of Sciences, 1985.

3. Rodriguez-Vazquez, A. and Prats, M.M., Spacecraft Magnetic Attitude Control Using Approximating Sequence of Riccati Equations, *IEEE Transactions on Aerospace and Electronic Systems*, 2015, vol. 51, no. 4, pp. 3374–3385. https://doi.org/10.1109/TAES.2015.130478

4. Ovchinnikov, M.Y., Roldugin, D.S., and Penkov, V.I., Three-Axis Active Magnetic Attitude Control Asymptotical Study, *Acta Astronautica*, 2015, vol. 110, pp. 279–286. https://doi.org/10.1016/j.actaastro.2014.11.030

5. Giri, D., Mukherjee, B., Bidul, T.N., and Sinha, M., Three-Axis Global Magnetic Attitude Control of Earth-Pointing Satellites in Circular Orbit, *Asian Journal of Control*, 2017, vol. 19(3), pp. 2028–2041. https://doi.org/10.1002/asjc.1506

6. Penkov, V.I., Ovchinnikov, M.Yu., and Roldugin, D.S., Magnetic Attitude Control in Orbital Frame. Vector of Finite Rotation Control Method, *Preprints of the Keldysh Institute of Applied Mathematics*, 2018, no. 7. https://doi.org/10.20948/prepr-2018-7

7. Morozov, V.M., and Kalenova, V.I., Controlling the Orientation of a Polar-Orbiting Satellite by Means of Magnetic Moments, *Engineering Journal: Science and Innovation*, 2018, no. 9(81), pp. 8–23. https://doi.org/10.18698/2308-6033-2018-9-1798

8. Ovchinnikov, M. and Roldugin, D., Magnetic Attitude Control and Periodic Motion for the In-Orbit Rotation of a Dual-Spin Satellite, *Acta Astronautica*, 2021, vol. 186, pp. 203–210. https://doi.org/10.1016/j.actaastro.2021.05.035

9. Morozov, V.M. and Kalenova, V.I., Satellite Control Using Magnetic Moments: Controllability and Stabilization Algorithms, *Cosmic Res.*, 2020, vol. 58, no. 3, pp. 158–166. https://doi.org/10.1134/S0010952520030041

10. Tikhonov, A.A., A Method of Semipassive Attitude Stabilization of a Spacecraft in the Geomagnetic Field, *Cosmic Research*, 2003, vol. 41, no. 1, pp. 63–73. https://doi.org/10.1023/A:1022355730291

11. Giri, D.K. and Sinha, M., Magneto-Coulombic Attitude Control of Earth-Pointing Satellites, *J. Guid. Control. Dyn.*, 2014, no. 37(6), pp. 1946–1960. https://doi.org/10.2514/1.G000030

12. Giri, D.K. and Sinha, M., Three-Axis Attitude Control of Earth-Pointing Isoinertial Magneto-Coulombic Satellites, *Int. J. Dyn. Control*, 2017, no. 5(3), pp. 644–652. https://doi.org/10.1007/s40435-015-0206-x

13. Aleksandrov, A.Yu. and Tikhonov, A.A., Averaging Technique in the Problem of Lorentz Attitude Stabilization of an Earth-Pointing Satellite, *Aerospace Science and Technology*, 2020, vol. 104, art. no. 105963. https://doi.org/10.1016/j.ast.2020.105963

14. Prabhat, H., Mukherjee, B.K., Giri, D.K., and Sinha, M., Fault-Tolerant Sliding Mode Satellite Attitude Stabilization Using Magneto-Coulombic Torquers, *Aerosp. Sci. Technol.*, 2022, vol. 121, art. no. 107316. https://doi.org/10.1016/j.ast.2021.107316

15. Antipov, K.A. and Tikhonov, A.A., Parametric Control in the Problem of Spacecraft Stabilization in the Geomagnetic Field, *Autom. Remote Control*, 2007, vol. 68, no. 8, pp. 1333–1344.

16. Tikhonov, A.A., Antipov, K.A., Korytnikov, D.G., and Nikitin, D.Yu., Electrodynamical Compensation of Disturbing Torque and Attitude Stabilization of a Satellite in J2 Perturbed Orbit, *Acta Astronautica*, 2017, vol. 141, pp. 219–227. https://doi.org/10.1016/j.actaastro.2017.10.009

17. Somov, E.I. and Butyrin, S.A., Energy-Saving Digital Control of Magnetic Actuator at an Emergency Mode of a Satellite Orientation, *Izvestiya of Samara Scientific Center of the Russian Academy of Sciences*, 2018, vol. 20, no. 1, pp. 37–44.

18. Beletsky, V.V. and Yanshin, A.M., *Vliyanie aerodinamicheskikh sil na vrashchatel'noe dvizhenie iskusstvennykh sputnikov* (The Influence of Aerodynamic Forces on the Rotational Motion of Artificial Satellites), Kyiv: Naukova Dumka, 1984.

19. Belokonov, I.V., Timbai, I.A., and Barinova, E.V., Design Parameters Selection for CubeSat Nanosatellite with a Passive Stabilization System, *Gyroscopy Navig.*, 2020, vol. 11, pp. 149–161. https://doi.org/10.1134/S2075108720020029

20. Barinova, E.V., Belokonov, I.V., and Timbai, I.A., Technology for Designing the Angular Motion of CubeSat Nanosatellites with a Passive Stabilization System, *Gyroscopy Navig.*, 2024, vol. 15, pp. 346–355. https://doi.org/10.1134/S2075108725700063

21. Barinova, E.V., Belokonov, I.V., Elisov, N.A., et al., Some Features of Dynamics and Attitude Control of Nanosatellites in Low Orbits, *Gyroscopy Navig.*, 2023, vol. 14, pp. 183–204. https://doi.org/10.1134/S2075108723030021
22. Lovera, M. and Astolfi, A., Global Magnetic Attitude Control of Spacecraft in the Presence of Gravity Gradient, *IEEE Transactions on Aerospace and Electronic Systems*, 2006, vol. 42, no. 3, pp. 796–805.
23. Chesi, S., Gong, Q., and Romano, M., Aerodynamic Three-Axis Attitude Stabilization of a Spacecraft by Center-of-Mass Shifting, *Journal of Guidance, Control, and Dynamics*, 2017, vol. 40, no. 7, pp. 1613–1626.
24. Maksimenko, M.V. and Tikhonov, A.A., Electrodynamic Attitude Stabilization of a Satellite in an Indirect Position, *Acta Astronautica*, 2025, vol. 232, pp. 516–524. https://doi.org/10.1016/j.actaastro.2025.03.018
25. Zubov, V.I., *Lektsii po teorii upravleniya* (Lectures on Control Theory), Moscow: Nauka, 1975.
26. Petrov, K.G. and Tikhonov, A.A., Lorentz Torque Acting on a Charged Satellite in the Earth's Magnetic Field. Part 2: Calculation of the Torque and Estimates of Its Components, *Vestnik St. Petersburg Univ.*, 1999, vol. 15, no. 3, pp. 81–91.
27. Tikhonov, A.A., On Electrodynamical Compensation of a Torque Disturbing Satellite Orientation, *Proceedings of 2015 International Conference "Stability and Control Processes" in Memory of V.I. Zubov (SCP)*, St. Petersburg, October 5–9, 2015, pp. 132–134. https://doi.org/10.1109/SCP.2015.7342069
28. Sarychev, V.A., D.E. Okhotsimskii and His Role in the Development of Passive Satellite Orientation Systems, in *Prikladnaya nebesnaya mekhanika i upravleniye dvizheniyem* (Applied Celestial Mechanics and Motion Control), Collection of articles dedicated to the 90th Anniversary of D.E. Okhotsimskii, Akim, E.L., et al., Eds., Moscow: Keldysh Institute of Applied Mathematics, 2010.
29. Tikhonov, A.A., Antipov, K.A., Sablina, M.V., and Spasic, D.T., Optimizing the Electrodynamic Stabilization Method for a Man-Made Earth Satellite, *Autom. Remote Control* 2011, vol. 72, no. 9, pp. 1898–1905.
30. Dosaev, M., Samsonov, V., and Hwang, S.-S., Construction of Control Algorithm in the Problem of the Planar Motion of a Friction-Powered Robot with a Flywheel and an Eccentric Weight, *Applied Mathematical Modelling*, 2021, vol. 89, pp. 1517–1527. https://doi.org/10.1016/j.apm.2020.08.022
31. Smolnikov, B.A., *Problemy mekhaniki i optimizatsii robotov* (Problems of Mechanics and Optimization of Robots), Moscow: Nauka, 1991.
32. Biryukov, V.G. and Chelnokov, Yu.N., Construction of Optimal Laws of Variation of the Angular Momentum Vector of a Rigid Body, *Mech. Solids*, 2014, vol. 49, no. 5, pp. 479–494. https://doi.org/10.3103/S002565441405001X
33. Levskii, M.V., Optimal Control of Kinetic Moment During the Spatial Rotation of a Rigid Body (Spacecraft), *Mech. Solids*, 2019, vol. 54, pp. 92–111. https://doi.org/10.3103/S0025654419010084
34. Smirnov, A.S. and Smolnikov, B.A., Construction and Analysis Modes of Rigid Body Motion Control, *Trudy Mosk. Aviats. Inst.*, 2022, no. 124. https://doi.org/10.34759/trd-2022-124-03
35. Aleksandrov, A.Yu. and Tikhonov, A.A., Asymptotic Stability of a Satellite with Electrodynamic Attitude Control in the Orbital Frame, *Acta Astronautica*, 2017, vol. 139, pp. 122–129. https://doi.org/10.1016/j.actaastro.2017.06.033
36. Kosov, A., The Exponential Stability and Stabilization of Non-Autonomous Mechanical Systems with Non-Conservative Forces, *Journal of Applied Mathematics and Mechanics*, 2007, vol. 71, pp. 371–384. https://doi.org/10.1016/j.jappmathmech.2007.07.011
37. Aleksandrov, A.Y. and Kosov, A.A., Asymptotic Stability of Equilibrium Positions of Mechanical Systems with a Nonstationary Leading Parameter, *J. Comput. Syst. Sci. Int.*, 2008, vol. 47, no. 3, pp. 332–345. https://doi.org/10.1134/S1064230708030027
38. Alken, P., Thebault, E., Beggan, C.D., et al., International Geomagnetic Reference Field: The 13th Generation, *Earth Planets Space*, 2021, vol. 73, art. no. 49. https://doi.org/10.1186/s40623-020-01288-x

This paper was recommended for publication by O.A. Stepanov, a member of the Editorial Board

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2026, Vol. 87, No. 3, pp. 241–249.

CONTROL IN SOCIAL ECONOMIC SYSTEMS

Aggregate Indicators of Oligopoly Game Dynamics

G. I. Algazin[*,a] **and D. G. Algazina**[*,b]

[*] *Altai State University, Barnaul, Russia*
e-mail: [a] *algaz46@yandex.ru,* [b] *darya.algazina@mail.ru*

Received May 24, 2025
Revised June 26, 2025
Accepted June 30, 2025

Abstract—This paper addresses the problem of designing aggregate indicators of reflexive collective decision-making dynamics in oligopoly models. Competitive markets with an arbitrary number of Cournot-reflexive agents are considered. The indicators proposed below are used to estimate the dynamics analytically, establish conditions for its convergence to an equilibrium, and identify the moments of dynamics reversal towards convergence. According to a comparative analysis provided, the numerical simulation results are consistent with the analytical estimation counterparts obtained using the aggregate indicators. The perfect consistency of the results is proved for stabilizing sequences of indicators.

Keywords: Cournot oligopoly, reflexive collective behavior, aggregate description, equilibrium, stabilizing sequences, aggregate indicators

DOI: 10.7868/S1608303226030053

1. INTRODUCTION

This work is devoted to solving the game of agents in an oligopoly market. The state-of-the-art of this problem, as one of the most important in game theory, was reviewed in [1–3].

Under incomplete awareness and the interdependence of behavior, the dynamics of agent's decision-making in a competitive market are built on the basis of reflexive considerations regarding the best individual choice, taking the responses of competitors into account [4–16]. Below, we develop an approach to describing reflexive collective behavior and solving the problem of achieving equilibrium [17].

In contrast to conventional approaches, where the emphasis is on the actions of each agent in each period (time step) of decision-making, the convergence conditions within this approach are formulated for an aggregate description of agent behavior. Aggregation over agent actions and aggregation over time are used. The former aggregation is implemented for each period by summing the residuals of agent actions. And the latter aggregation is implemented over a set of sequential periods by summing the values of dynamics elements [17].

The problem under consideration is to design aggregate indicators that can be used to monitor overall market dynamics and estimate the movement of each agent's dynamics towards an equilibrium analytically. The Cournot oligopoly model is taken as a basic model, where reflexive agents follow a collective behavior procedure under incomplete awareness.

The remainder of this paper is organized as follows. Sections 2 and 3 present the theoretical background of the research under assumptions regarding the awareness of agents that are traditional for indicator-based collective behavior models, as well as a thorough justification of the indicators proposed. For several special cases, the numerical simulation results on the asymptotic convergence of reflexive dynamics are compared with the analytical estimation counterparts obtained using the new aggregate indicators. In Section 4, we consider the applicability of these indicators for the agents themselves. Additional assumptions regarding the awareness of agents are introduced, allowing them to estimate uniformly, via the indicators, the market dynamics independently and on par with the competitors.

2. STARTING POINTS FOR THE RESEARCH

This section presents the theoretical background (previous and new results) underlying the design of aggregate indicators for the convergence of agent dynamics.

The Cournot oligopoly model is a noncooperative game with n players (rational agents) competing in the outputs of a homogeneous product. The classical assumptions of this model (adopted here as well) include a linear inverse demand function $p\left(\sum_{j\in N} q_j\right) = a - b\sum_{j\in N} q_j$ and linear cost functions $\varphi_i(q_i) = c_i q_i + d_i$ of agents, with the following notation: q_i is the output of agent i ($i \in N = \{1, \dots, n\}$); c_i and d_i are the marginal and fixed costs of agent i, respectively; $p\left(\sum_{j\in N} q_j\right)$ is the uniform market price; the parameter a characterizes the maximum possible price of the product under which the demand will tend to zero; and the parameter b characterizes the slope of the demand curve. The payoff (goal) function of agent i depends on his strategy q_i and the set of strategies of all players and has the form [18, 19]

$$\Pi_i\left(q_i, \sum_{j\in N} q_j\right) = \left(a - b\sum_{j\in N} q_j\right) q_i - c_i q_i - d_i \to \max_{q_i}, \quad i \in N. \tag{1}$$

In the case of linear demand and cost functions, a unique solution $q^* = (q_1^*, \dots, q_i^*, \dots, q_n^*)$, $q_i^* > 0\ \forall i \in N$, of the oligopoly game, understood as a static Nash equilibrium in the normal form game [20], exists under the well-known conditions; for example, see [17].

Under game uncertainty (about the actions chosen by the competitive environment) and incomplete knowledge (of the costs, goal functions, etc., of competitors), the game solution is found as the outcome of the reflexive collective behavior dynamics [3, 4, 9, 17, 19, 21]:

$$q_i^{t+1} = q_i^t + \gamma_i^{t+1}(x_i^t - q_i^t), \quad i \in N. \tag{2}$$

Here:

t ($t = 0, 1, 2, \dots$) is the time instant (period) number;

$x_i^t = x_i\left(\sum_{j\in N\setminus\{i\}} q_j^t\right)$ is the current position of the target of agent i (the action ensuring the maximum value of this agent's goal function in period t); to determine it, the agent does not need to know the goal functions of the competitors since the actions of all agents observed by him in period t are supposed to be the same in the current period;

$\gamma_i^{t+1} \in [0, 1]$ is a parameter independently chosen by each agent in period $(t+1)$ (his "step" from the current strategy to the current position of his target);

$q^0 = (q_1^0, \dots, q_i^0, \dots, q_n^0)$ is the vector of initial strategies of all agents. For the linear Cournot model [15, 17, 19], we have

$$x_i^t = \frac{(a - c_i)/b - \sum_{j\in N\setminus\{i\}} q_j^t}{2}. \tag{3}$$

With the change of variables

$$\varepsilon_i^t = q_i^* - q_i^t \quad (i \in N;\ t = 0, 1, 2, \dots), \tag{4}$$

the profit model (1) can be transformed to

$$\Pi_i^t = -\left(\sum_{j\in N} \varepsilon_j^t\right) \varepsilon_i^t \to \max_{\varepsilon_i^t} \quad (i \in N), \tag{5}$$

and the dynamics (2) to

$$\varepsilon_i^{t+1} = \varepsilon_i^t + \gamma_i^{t+1} \left(-\frac{\sum_{j\in N\setminus\{i\}} \varepsilon_j^t}{2} - \varepsilon_i^t\right). \tag{6}$$

In the general case, problem (5) in the static statement may have multiple solutions. For example, for $n = 2$, any vector $\varepsilon = (\varepsilon_1, \varepsilon_2)$ will formally be a solution if $\varepsilon_1 = -\varepsilon_2$, or, more generally, if $\sum_{j\in N} \varepsilon_j = 0$. However, possible nonzero solutions (e.g., $\varepsilon = (5, -5)$ in the case $n = 2$) are not Nash equilibria. Only equilibrium solutions are of interest.

The starting points for this research, explaining the transition to the model with residuals, are presented in the three propositions below.

Proposition 1. *In model* (5)*, there exists the unique static Nash equilibrium* $\varepsilon^* = (\varepsilon_1^*, \ldots, \varepsilon_i^*, \ldots, \varepsilon_n^*)$ *with* $\varepsilon_i^* = 0 \quad \forall i \in N$, *and the convergence of process* (6) *means that* $\varepsilon_i^t \to \varepsilon_i^* = 0$ *as* $t \to \infty$.

The proof of Proposition 1 is given in the Appendix for the first time.

The other two were proved in [17].

Proposition 2. *Process* (6) *converges if and only if process* (2), (3) *converges for model* (1).

Proposition 3. *If* $\sum_{j\in N} \varepsilon_j^t \to 0$ *as* $t \to \infty$, *then* $\varepsilon_i^t \to \varepsilon_i^* = 0\ \forall i \in N$.

The following relations are also important for the subsequent analysis:

$$\sum_{j\in N} \varepsilon_j^{t+1} = \left(1 - 2\tilde{\gamma}^{t+1}\right) \sum_{j\in N} \varepsilon_j^t, \tag{7}$$

$$\left|\sum_{j\in N} \varepsilon_j^{t_0+\tau}\right| \leqslant (1 - 2\bar{\tilde{\gamma}}_+^\tau)^{t_+} \left|1 - 2\bar{\tilde{\gamma}}_-^\tau\right|^{t_-} \left|\sum_{j\in N} \varepsilon_j^{t_0}\right| \quad (\tau = t_+ + t_-). \tag{8}$$

(Here, the expression $\frac{\tilde{\gamma}^t(1+n)}{4}$, used in [17], is redenoted by $\tilde{\gamma}^t$.)

In (8) and the Appendix, the superscripts t_+ and t_- are the powers, whereas the superscripts $(t_0 + \tau)$ and t_0 are not.

In (8), we also employ the following designations: t_0 is an arbitrary period and $\tau > 0$; $t_+(t_-)$ is the number of periods in the set T_+ (T_-, respectively), $T_+ = \{t \in T | 1 - 2\tilde{\gamma}^t > 0\}$, $T_- = \{t \in T | 1 - 2\tilde{\gamma}^t < 0\}$, $T = \{t_0 + 1, \ldots, t_0 + \tau\}$, and $\bar{\tilde{\gamma}}_+^\tau = \frac{1}{t_+} \sum_{t\in T_+} \tilde{\gamma}^t$ and $\bar{\tilde{\gamma}}_-^\tau = \frac{1}{t_-} \sum_{t\in T_-} \tilde{\gamma}^t$ are the average values of the parameter

$$\tilde{\gamma}^t = \left(\sum_{j\in N} \gamma_j^t + \sum_{j\in N} \gamma_j^t \varepsilon_j^{t-1} \Big/ \sum_{j\in N} \varepsilon_j^{t-1}\right) \Big/ 4 \tag{9}$$

over τ periods and the sets T_+ and T_-, respectively.

In the special case where $\tilde{\gamma}^t = \frac{1}{2}$ (and therefore, $\sum_{j\in N} \varepsilon_j^t = 0$), the market dynamics improve in the sense that the residuals of each agent in period $(t + 1)$ are smaller, by absolute value, than his residuals in period t [17].

In dynamics series, the ratio of two sequential levels of the series is called the chain growth rate. In the case under consideration, for the series of aggregate residuals, this rate is expressed via the aggregate indicator $\tilde{\gamma}^t$ of agent actions as $\frac{\sum_{j\in N} \varepsilon_j^t}{\sum_{j\in N} \varepsilon_j^{t-1}} = 1 - 2\tilde{\gamma}^t$, see (7).

3. AGGREGATE INDICATORS IN NEW STUDIES OF DYNAMICS CONVERGENCE

First, let us dwell on an important special case of the sequence $\{\tilde{\gamma}^t\}$, as it has a quite natural explanation in terms of the steps γ_i^t chosen by each agent i and can be simply verified experimentally; for example, see [4, 15, 19]). For instance, an agent does not change his steps over time or repeats them in a known way.

Assume that the sequence $\{\tilde{\gamma}^t\}$ stabilizes to $\tilde{\gamma}^{t_s}$ starting from period t_s if t_s is the smallest natural number such that $\tilde{\gamma}^t = \tilde{\gamma}^{t_s}$ $\forall t \geqslant t_s$ [22].

Proposition 4. *Let the sequence* $\{\tilde{\gamma}^t\}$ *stabilize to* $\tilde{\gamma}^{t_s}$. *The dynamics* (6) *converge if and only if* $0 < \tilde{\gamma}^{t_s} < 1$.

The proof of this proposition is given in the Appendix.

Example 1. Consider $n = 4$ agents in a market and compile the vector of agents' steps $\gamma^t = (\gamma_1^t, \gamma_2^t, \gamma_3^t, \gamma_4^t)$. First, we analyze the case where $\gamma^t = (0.2; 1; 1; 1)$ in all periods, i.e., the agents do not change their steps. According to a numerical experiment, the sequence $\{\tilde{\gamma}^t\}$ stabilizes, within the third decimal place, to the level $\tilde{\gamma}^{t_s} = 1.039$ starting from period $t_s = 35$, and the dynamics (6) diverge. By Proposition 4, the dynamics (6) also diverge since, in this case, $\tilde{\gamma}^{t_s} > 1$. Now let $\gamma^1 = \gamma^5 = \gamma^9 = \ldots = (0.2; 1; 1; 1)$, $\gamma^2 = \gamma^6 = \gamma^{10} = \ldots = (1; 0.2; 1; 1)$, $\gamma^3 = \gamma^7 = \gamma^{11} = \ldots = (1; 1; 0.2; 1)$, and $\gamma^4 = \gamma^8 = \gamma^{12} = \ldots = (1; 1; 1; 0.2)$. In this case, the sets of parameters repeat every 3 periods. According to a numerical experiment, starting from period $t_s = 21$, the sequence $\{\tilde{\gamma}^t\}$ stabilizes to $\tilde{\gamma}^{t_s} = 0.955$, and the dynamics (6) converge. By Proposition 4, the dynamics (6) also converge since $\tilde{\gamma}^{t_s} < 1$. Note that in both cases described, three agents take full steps (equal to one) and one takes a partial step (equal to 0.2), the average values of the steps coincide (equal to 0.8), and the sequences $\{\tilde{\gamma}^t\}$ stabilize. However, the results turn out to be directly opposite: in the first case, the dynamics (6) diverge, converging in the second, which may be somewhat non-obvious and unexpected.

For stabilizing sequences, the numerical simulation results are always perfectly consistent with the analytical estimation counterparts obtained using $\tilde{\gamma}^t$; see Example 1 above. Also, this example illustrates situations where the aggregate description makes it possible to formulate the convergence conditions of decision-making dynamics quite simply and uniformly, which may be not the case with a detailed description.

Let us proceed to general cases of the dynamics (6), where the sequences $\{\tilde{\gamma}^t\}$ may be non-stabilizing.

We introduce the following aggregate indicator for process (6):

$$z_{t_0,t_1} = \sum_{t=t_0}^{t_1} z^t \quad (z_{t_0,t_0} = z^{t_0}), \tag{10}$$

where

$$z^t = \begin{cases} \tilde{\gamma}^t, & \tilde{\gamma}^t < \dfrac{1}{2}(t \in T_+), \\ 1 - \tilde{\gamma}^t, & \tilde{\gamma}^t > \dfrac{1}{2}(t \in T_-), \\ \dfrac{1}{2}, & \tilde{\gamma}^t = \dfrac{1}{2}. \end{cases} \tag{11}$$

The properties of $z_{t_0,t}$:

(1) For $t_0 < t_1 < t_2$, $z_{t_0,t_2} = z_{t_0,t_1} + z_{t_1+1,t_2}$.

(2) For $t^* \in M(\tau) = \{t^* | z_{1,t^*} = \min_{t=1,\ldots,\tau} z_{1,t}\}$, $z_{t^*+1,t} \geqslant 0 \quad \forall t = t^*+1, \ldots, \tau$.
(3) For $t^* = \max_{t \in M(\tau)} t$, $z_{t^*+1,t} > 0 \quad \forall t = t^*+1, \ldots, \tau$.

Proposition 5. *Let the parameters* $\tilde{\gamma}^t, z^t$, *and* $z_{1,t}$ *of the Cournot oligopoly be given by* (9), (10), *and* (11), *respectively. If* $M(\tau) = \{t^* | z_{1,t^*} = \min_{t=1,\ldots,\tau} z_{1,t}\} \neq \oslash$ *as* $\tau \to \infty$, *then process* (6) *converges to an equilibrium.*

The proof of this (main) result of the paper is provided in the Appendix.

By the proposition, if the minimum of $z_{1,t}$ with respect to t exists (i.e., $z_{1,t}$ does not go to minus infinity) and t^* is the point of minimum, then $z_{t^*+1,t} \geqslant 0$ ($\forall t \geqslant t^*+1$) and the dynamics become convergent.

In the context of Proposition 5, the formation of $z_{1,t}$ plays a key role. In this regard, we discuss the possible changes of its value in the (next) period $(t+1)$.

The case $\tilde{\gamma}^{t+1} < \frac{1}{2}$ corresponds to the fact that, due to (10), $z_{1,t+1} = z_{1,t} + \tilde{\gamma}^{t+1}$. If $0 < \tilde{\gamma}^{t+1}$, then $z_{1,t+1} > z_{1,t}$. The value of z increases, which favorably affects the convergence of the process. If $\tilde{\gamma}^{t+1} < 0$, then $z_{1,t+1} < z_{1,t}$. The value of z decreases, which may deteriorate the convergence of the process. If $\tilde{\gamma}^{t+1} = 0$, then $z_{1,t+1} = z_{1,t}$. The value of z remains the same.

The case $\frac{1}{2} < \tilde{\gamma}^{t+1}$ corresponds to the fact that, due to (10) and (11), $z_{1,t+1} = z_{1,t} + 1 - \tilde{\gamma}^{t+1}$. If $\tilde{\gamma}^{t+1} < 1$, then $z_{1,t+1} > z_{1,t}$. The value of z increases, which favorably affects the convergence of the process. If $1 < \tilde{\gamma}^{t+1}$, then $z_{1,t+1} < z_{1,t}$. The value of z decreases, which may deteriorate the convergence of the process. If $\tilde{\gamma}^{t+1} = 1$, then $z_{1,t+1} = z_{1,t}$. The value of z remains the same.

The case $\tilde{\gamma}^t = \frac{1}{2}$ favors the convergence of the process and $z_{1,t+1} = z_{1,t} + \frac{1}{2}$.

Proposition 5 and the above discussion give grounds for treating $z_{t_0,t}$ as an aggregate indicator for estimating the convergence of the dynamics (6). A positive current value of the indicator points to an improvement in the overall market dynamics relative to the reference period (in the sense of (8)), which (see Proposition 3) favors the convergence of each agent's dynamics. A negative current value of the aggregate indicator is insufficient for asserting a relative improvement in the dynamics. Generally speaking, a positive contribution to the indicator and to the convergence of the process is made by periods where $0 < \tilde{\gamma}^t < 1$.

In dynamics series, the base growth rate is used as an intensity indicator; in the case under consideration, for the series of aggregate residuals and the time interval $[t_0, t_0+\tau]$, it has the form $\frac{\sum_{j \in N} \varepsilon_j^{t_0+\tau}}{\sum_{j \in N} \varepsilon_j^{t_0}}$. The indicator $z_{t_0,t_0+\tau}$, aggregated over the agents' actions and the interval, estimates the dynamics of this rate and points the periods of its reversal towards convergence to zero.

Example 2. A numerical calculation fragment. This example illustrates Proposition 5. As before, consider $n = 4$ agents. Figure 1 shows the dynamics of the indicator $\tilde{\gamma}^t$, which is calculated by (9) based on some initial conditions ε_i^0 and agents' steps γ_i^t $(i = \overline{1,4})$. On an interval of thirty periods, the dynamics of $\tilde{\gamma}^t$ do not stabilize. This figure also shows the corresponding dynamics of the indicator $z_{1,t}$, which is calculated by (10), (11). The influence of the dynamics of the indicator $z_{1,t}$ on the dynamics of the residuals ε_i^t $(i = \overline{1,4})$ can be traced in Fig. 2. Up to the twentieth period, no convergence trend of the residuals of all agents is observed since on the interval of thirty periods, the indicator $z_{1,t}$ achieves minimum only in the twentieth period (see Fig. 1). Then the indicator $z_{21,t}$, which can be calculated by the formula $z_{21,t} = z_{1,t} - z_{1,20}$ (see Property 1) above), has a positive sign for $t \geqslant 21$ and increases up to $t = 29$ inclusive. Therefore, the dynamics of agents' residuals move towards the theoretical static Cournot–Nash equilibrium $\varepsilon_i^* = 0$ $(i = \overline{1,4})$.

Let us return to Example 1 with stabilizing indicator sequences $\{\tilde{\gamma}^t\}$ to demonstrate their application. For the case of convergent dynamics of ε_i^t, in view of (11), the aggregate indicator starting from $t_s = 21$ will be calculated by the formula $z_{t_s,\tau} = \sum_{t=t_s}^{\tau} (1 - \tilde{\gamma}^t) = (\tau - t_s + 1)(1 - \tilde{\gamma}^{t_s}) = (\tau - t_s + 1)\, 0.445$. Its value will be bounded from below as $\tau \to \infty$; therefore, based on Proposi-

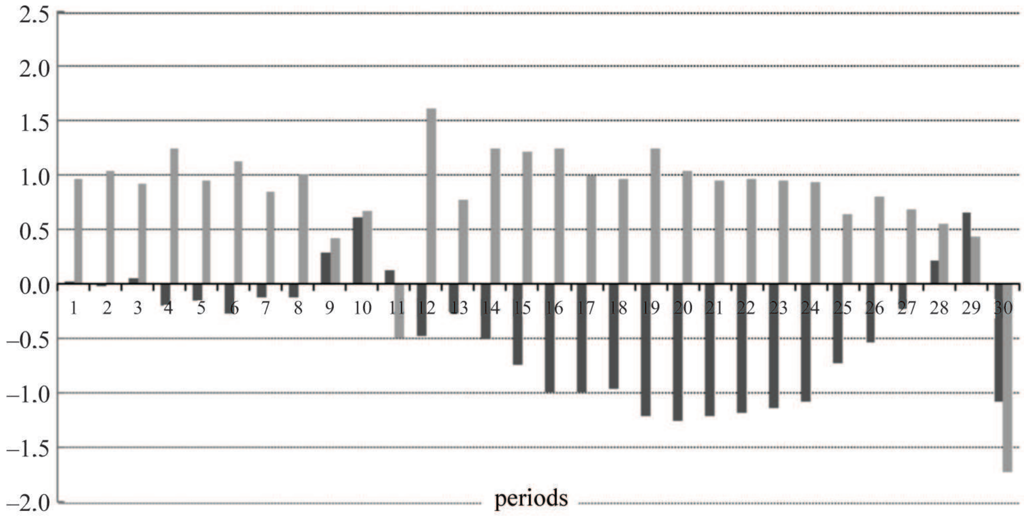

Fig. 1. The dynamics of the indicators $\tilde{\gamma}^t$ (light bars) and $z_{1,t}$ (dark bars).

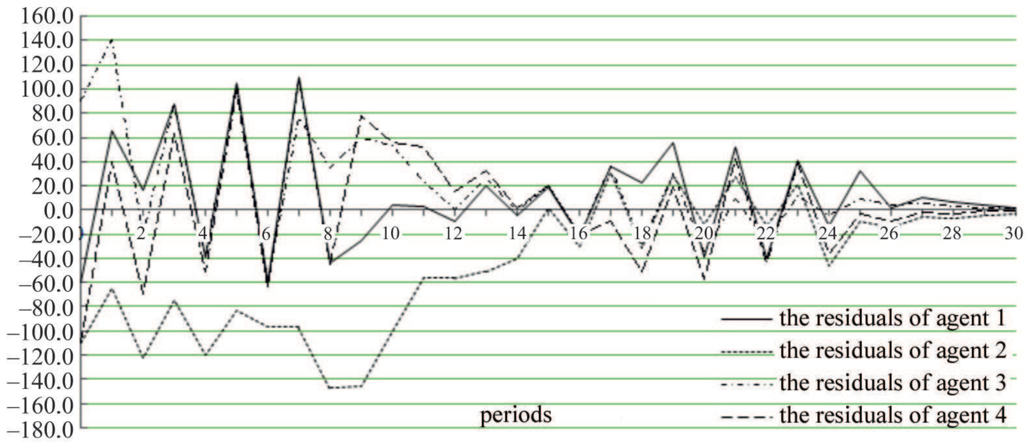

Fig. 2. The dynamics of the residuals of individual agent actions.

tion 5, the convergence of (6) according to this aggregate indicator is confirmed. For the case of divergent dynamics of ε_i^t, the aggregate indicator starting from $t_s = 35$ will be calculated by the same formula $z_{t_s,\tau} = (\tau - t_s + 1)(1 - \tilde{\gamma}^{t_s}) = (\tau - t_s + 1)(-0.039)$. Its value will not be bounded from below as $\tau \to \infty$. Hence, $M(\tau) = \varnothing$ as $\tau \to \infty$, and by Proposition 5, the convergence of (6) according to the z-indicator is not confirmed.

4. APPLICATION OF INDICATORS BY AGENTS

The above indicators may be of not only theoretical but also practical interest. Here is a possible scenario in which agents apply aggregate indicators in order to estimate their dynamics and the overall market dynamics as well.

Traditionally, in the collective behavior procedure (2), an agent decides on his current output without knowledge of the equilibrium output and without the ability to judge how his previous actions and the actions of other agents have contributed to achieving the equilibrium. Obviously, such an ability could be useful for an agent to make his current choice.

An agent will acquire this ability by using the aggregate indicators $\tilde{\gamma}^t$ and $z_{t_0,t}$.

If $0 < \tilde{\gamma}^t < 1$, the steps chosen in the previous period t by all agents have positively contributed to the convergence of the process. And if $z_{t_0,t} > 0$, the steps chosen on the time interval $t_0, t_0+1, \ldots, t$ have improved the overall market dynamics in period t relative to period t_0.

The main challenge regarding the applicability of the indicators for the agents themselves is that the collective behavior procedures (2) and (6) are described in different "coordinate frames": in one frame, the current outputs of agents are calculated by (2); in the other, the residuals of current outputs are calculated by (6), and the indicators are formed by (9)–(10).

To link these coordinate frames, we present the following considerations. With $Q^t = \sum_{j\in N} q_j^t$, from (4) and $\sum_{j\in N} \varepsilon_j^{t+1} = (1-2\tilde{\gamma}^{t+1})\sum_{j\in N}\varepsilon_j^t$ it follows that $\sum_{j\in N}\varepsilon_j^t = \frac{Q^{t+1}-Q^t}{2\tilde{\gamma}^{t+1}}$. Then $\sum_{j\in N}\varepsilon_j^t - \sum_{j\in N}\varepsilon_j^{t-1} = \frac{Q^{t+1}-Q^t}{2\tilde{\gamma}^{t+1}} - \frac{Q^t-Q^{t-1}}{2\tilde{\gamma}^t} = -(Q^t - Q^{t-1})$, and we arrive at the recurrence relation

$$\tilde{\gamma}^{t+1} = \frac{Q^{t+1}-Q^t}{Q^t-Q^{t-1}} \cdot \frac{\tilde{\gamma}^t}{1-2\tilde{\gamma}^t}. \tag{12}$$

Let us adopt several assumptions so that each agent, while staying within the coordinate frame of process (2), will objectively estimate the market dynamics on par with other competitors.

(1) All agents are "intellectual" to the level of knowing formula (12) for calculating the indicator $\tilde{\gamma}^t$ recursively and formulas (10), (11) for calculating the indicator $z_{t_0,t}$.

(2) In a certain period t_0, all agents choose the same step γ^{t_0}. (Starting from t_0, the beliefs of all agents about the process will be synchronized. The agents choose it by agreement. According to (9), $\tilde{\gamma}^{t_0} = \gamma^{t_0}(1+n)/4$. The value $\tilde{\gamma}^{t_0}$ is supposed to be known to all agents.)

The first assumption is quite natural for an agent using indicators. It means that in period $(t+1)$, each agent will know the calculated indicators $\tilde{\gamma}^t$ and the aggregate outputs Q^t $(t = 1, 2, \ldots)$ of the previous periods. The indicators $\tilde{\gamma}^t$ are calculated by the agents themselves using the recurrence formula (12), and the agents know the aggregate outputs based on the conditions of the collective behavior procedure (2), (3). In turn, in period $(t+1)$, each agent will also know the calculated indicators $z_{t_0,t}$, which (see Property 1) above) are calculated by the agent using the recurrence formula $z_{t_0,t} = z_{t_0,t-1} + z^t$, with z^t given by (11).

The second assumption is intended to eliminate the influence of equilibrium outputs on the indicator $\tilde{\gamma}^t$ and process (6).

Under these assumptions, by observing the previous aggregate outputs Q^t according to model (2), each agent can obtain, independently of the others, the uniform estimates $\tilde{\gamma}^t$ and $z_{t_0,t}$ of the market dynamics synchronized with the estimates of his competitors starting from period t_0.

5. CONCLUSIONS

The practical and scientific significance, as well as the prospects for future studies of collective behavior dynamics, is determined by an analytical confirmation of the intuitive idea that their aggregate description over significant time intervals can be no less accurate than a detailed one. The application of aggregate indicators expands our understanding of the set of situations that, under incomplete awareness and the absence of common knowledge, can lead oligopoly models to reflexive equilibria (stable outcomes of agent interaction). It seems promising to design and apply aggregate indicators for analytically solving reflexive decision-making problems within other oligopoly and behavior models, different from those considered in this work, e.g., with the Stackelberg responses of agents, as well as taking into account the mental and behavioral components of agent activity [9].

APPENDIX

Proof of Assertion 1. Based on (5), the optimal strategies of agents can be determined by solving the system of n homogeneous linear equations $\partial\Pi_i/\partial\varepsilon_i = -\left(\sum_{j\in N}\varepsilon_j\right) - \varepsilon_i = 0$ $(i \in N)$ with n unknowns. The determinant of the coefficient matrix of the unknowns is nonzero. Therefore, the system has only the trivial solution $\varepsilon_i^* = 0$ $\forall i \in N$. This solution will be a Nash equilibrium. Suppose that only agent i has a finite deviation $\Delta\varepsilon_i \neq 0$ from the trivial solution. Due to (5), we have $\Pi_i(\varepsilon^*) = 0 > -(\Delta\varepsilon_i)^2$: by deviating, the agent reduces his profit. In other words, the trivial solution is an equilibrium.

The proof of Proposition 1 is complete.

Proof of Assertion 4. Let $0 < \tilde{\gamma}^{t_s} < 1$. According to (7), we have $\sum_{j\in N}\varepsilon_j^{t_0+\tau} = (1-2\tilde{\gamma}^{t_s})^\tau \sum_{j\in N}\varepsilon_j^{t_0}$. Hence, $\sum_{j\in N}\varepsilon_j^{t_0+\tau} \to 0$ as $\tau\to\infty$ and, by Proposition 3, $\varepsilon_i^{t_0+\tau} \to \varepsilon_i^* = 0$ $\forall i \in N$.

If $\tilde{\gamma}^{t_s} < 0$ or $1 < \tilde{\gamma}^{t_s}$, then $\left|\sum_{j\in N}\varepsilon_j^{t_0+\tau}\right| \to \infty$, and therefore the dynamics cannot converge.

The proof of Proposition 4 is complete.

Proof of Assertion 5. Let $t^* = \max_{t\in M(\tau)} t$ as $\tau \to \infty$. Then, by Property 3) for $z_{t^*+1,t}$, we obtain $z_{t^*+1,t^*+\tau} = \sum_{t\in T_+}\tilde{\gamma}^t + t_- - \sum_{t\in T_-}\tilde{\gamma}^t > 0$ with $\tau \geqslant 1$. In view of the the average value expressions $\bar{\tilde{\gamma}}_+^t = \frac{1}{t_+}\sum_{t\in T_+}\tilde{\gamma}^t$ and $\bar{\tilde{\gamma}}_-^t = \frac{1}{t_-}\sum_{t\in T_-}\tilde{\gamma}^t$, it follows that $t_+\bar{\tilde{\gamma}}_+^\tau - t_-(\bar{\tilde{\gamma}}_-^\tau - 1) > 0$. This inequality implies $\frac{t_+(1-2\bar{\tilde{\gamma}}_+^\tau)+t_-(2\bar{\tilde{\gamma}}_-^\tau-1)}{t_++t_-} < 1$.

Due to inequality (8), we have $\left|\sum_{j\in N}\varepsilon_j^{t^*+\tau}\right| \leqslant (1-2\bar{\tilde{\gamma}}_+^\tau)^{t_+}|1-2\bar{\tilde{\gamma}}_-^\tau|^{t_-}\left|\sum_{j\in N}\varepsilon_j^{t^*}\right|$ $(\tau = t_+ + t_-)$.

The expression $(1-2\bar{\tilde{\gamma}}_+^\tau)^{t_+}|1-2\bar{\tilde{\gamma}}_-^\tau|^{t_-}$ on the right-hand side of this inequality can be estimated by using a well-known result: for $x_1, x_2 \geqslant 0$ and $1 > \alpha > 0$, $\alpha x_1 + (1-\alpha)x_2 \geqslant x_1^\alpha x_2^\alpha$, and the strict equality holds only if $x_1 = x_2$.

Consequently,

$$(1-2\bar{\tilde{\gamma}}_+^\tau)^{t_+}|1-2\bar{\tilde{\gamma}}_-^\tau|^{t_-} = (1-2\bar{\tilde{\gamma}}_+^\tau)^{t_+}(2\bar{\tilde{\gamma}}_-^\tau-1)^{t_-} \leqslant \left[\frac{t_+(1-2\bar{\tilde{\gamma}}_+^\tau)+t_-(2\bar{\tilde{\gamma}}_-^\tau-1)}{t_++t_-}\right]^{t_++t_-}.$$

It has been shown that the expression in square brackets is below 1. Therefore, $\left[\frac{t_+(1-2\bar{\tilde{\gamma}}_+^\tau)+t_-(2\bar{\tilde{\gamma}}_-^\tau-1)}{t_++t_-}\right]^{t_++t_-} \to 0$ as $\tau \to \infty$, and based on Proposition 3, process (6) converges to an equilibrium.

The proof of Proposition 5 is complete.

REFERENCES

1. *Handbook of Game Theory and Industrial Organization. Vol. I: Theory*, Corchon, L.C. and Marini, M.A., Eds., Edward Elgar Publishing, UK, USA, 2018.

2. *Handbook of Game Theory and Industrial Organization. Vol. II: Applications*, Corchon, L.C. and Marini, M.A., Eds., Edward Elgar Publishing, UK, USA, 2018.

3. Geraskin, M.I., A Survey of the Latest Advances in Oligopoly Games, *Autom. Remote Control*, 2023, vol. 84, no. 6, pp. 565–578.

4. Novikov, D.A., and Chkhartishvili, A.G., *Reflexion and Control: Mathematical Models*, Leiden: CRC Press, 2014.

5. Ueda, M., Effect of Information Asymmetry in Cournot Duopoly Game with Bounded Rationality, *Appl. Math. Comput.*, 2019, vol. 362, art. no. 124535. https://doi.org/10.1016/j.amc.2019.06.049.124535

6. Peng, Y., Xiao, Y., Lu, Q., Wu, X., and Zhao, Y., Chaotic Dynamics in Cournot Duopoly Model with Bounded Rationality Based on Relative Profit Delegation Maximization, *Physica A: Statist. Mech. Appl.*, 2020, vol. 560, art. no. 125174.

7. Koriyama, Y. and Ozkes, A.I., Inclusive Cognitive Hierarchy, *J. Econom. Behavior Organiz.*, 2021, vol. 186, pp. 458–480.

8. Novikov, D.A., Bounded Rationality and Control, *Autom. Remote Control*, 2022, vol. 83, no. 7, pp. 990–1009.

9. Novikov, D.A., Dynamics Models of Mental and Behavioral Components of Activity in Collective Decision-Making, *Large-Scale Systems Control*, 2020, no. 85, pp. 206–237.

10. Skarzhinskaya, E. and Tsurikov, V., On the Possibility of Successive Approximation towards an Equilibrium in a Coalition Game with Reiterating Collective Action, *Economics and Mathematical Methods*, 2020, vol. 56, no. 4, pp. 103–115.

11. Askar, S., On Complex Dynamics of Cournot–Bertrand Game with Asymmetric Market Information, *Appl. Math. Comput.*, 2021, vol. 393, no. 3, art. no. 125823. https://doi.org/10.1016/j.amc.2020.125823

12. Alcantara-Jimenez, G. and Clempner, J.B., Repeated Stackelberg Security Games: Learning with Incomplete State Information, *Reliability Engineer. Syst. Safety*, 2020, vol. 195, art. no. 106695. https://doi.org/10.1016/j.ress.2019.106695

13. Long, J. and Zhao, H., Stability of Equilibrium Prices in a Dynamic Duopoly Bertrand Game with Asymmetric Information and Cluster Spillovers, *Int. J. Bifurcat. Chaos Appl. Sci. Engineer.*, 2021, vol. 31, no. 16, pp. 215–240.

14. Xiao, Y., Zhang, S., and Peng, Y., Dynamic Investigations in a Stackelberg Model with Differentiated Products and Bounded Rationality, *J. Comput. Appl Math.*, 2022, vol. 414, no. 3, art. no. 114409. https://doi.org/10.1016/j.cam.2022.114409

15. Korepanov, V.O., *Modeli refleksivnogo gruppovogo povedeniya i upravleniya* (Models of Reflexive Group Behavior and Control), Moscow: Trapeznikov Institute of Control Sciences RAS, 2011.

16. Ougolnitsky, G.A. and Usov, A.B., The Interaction of Economic Agents in Cournot Duopoly Models under Ecological Conditions: A Comparison of Organizational Modes, *Autom. Remote Control*, 2023, vol. 84, no. 2, pp. 175–189.

17. Algazin, G.I. and Algazina, D.G., Aggregate Estimates of Reflexive Collective Behavior Dynamics in a Cournot Oligopoly Model, *Autom. Remote Control*, 2024, vol. 85, no. 9, pp. 821–831.

18. Cournot, A., *Researches into the Mathematical Principles of the Theory of Wealth*, London: Hafner, 1960. (Original 1838.)

19. Algazin, G.I. and Algazina, Yu.G., Reflexive Dynamics in the Cournot Oligopoly under Uncertainty, *Autom. Remote Control*, 2020, vol. 81, no. 2, pp. 345–359.

20. Nash, J., Non-Cooperative Games, *Ann. Math.*, 1951, no. 54, pp. 286–295.

21. Opoitsev, V.I., *Ravnovesie i ustoichivost' v modelyakh kollektivnogo povedeniya* (Equilibrium and Stability in Collective Behavior Models), Moscow: Nauka, 1977.

22. Yakovlev, M.K., Construction of the Real Field and the Theory of Limits of Sequences Based upon the Notion of the Stabilizer of Sequences of Non-Terminating Decimals, *Matematika v Vysshem Obrazovanii*, 2010, no. 8, pp. 41–52.

This paper was recommended for publication by A.G. Chkhartishvili, a member of the Editorial Board

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2026, Vol. 87, No. 3, pp. 250–269.

CONTROL IN SOCIAL ECONOMIC SYSTEMS

Market Design, Pricing, and the Distribution of Market Power in Finite Markets

R. K. Shavshin

National Research University Higher School of Economics, Moscow, Russia
e-mail: ruslanshavshin@gmail.com

Received April 28, 2025
Revised October 8, 2025
Accepted October 11, 2025

Abstract—This paper analyzes the impact of market design on equilibrium characteristics in finite markets. We study a directed search model in the labor market under two variations: firm-led and worker-led contract terms. The results demonstrate that the preferences of the market sides (employees, firms, and society) regarding market design depend on market size and may either align or diverge. Furthermore, the analysis reveals that for certain market-side ratios, the expected number of concluded contracts increases by more than 10% upon switching to a more efficient design.

Keywords: market design, directed search, finite markets, labour market

DOI: 10.7868/S1608303226030063

1. INTRODUCTION

The concept of directed search, which underpins a broad class of research, offers a distinct perspective on market structure and interaction, providing an explanation for several key empirical phenomena. One such phenomenon is the coexistence of agents on opposite sides of the market who could potentially trade but remain unmatched due to market frictions. Examples include the simultaneous existence of vacancies and unemployment in the labor market, unmatched individuals in the marriage market, and unserved buyers alongside sellers with unsold inventories in the goods market.

The core idea defining the directed search framework is that the matching process between agents from opposite sides of the market is not random; rather, it emerges as a result of strategic interaction.

(a) In the first stage, agents on one side of the market (*Proposers*) simultaneously and independently publicly announce the terms under which they are willing to contract with any agent from the opposite side of the market (*Respondents*).

(b) In the second stage, agents on the *Respondents* side, *having observed all proposed contracts*, simultaneously and independently decide which agent on the *Proposers* side they wish to approach. It is at this point that "directed search" occurs, in the sense that the *Respondents*' decisions are made strategically, based on the terms offered by the *Proposers* and the anticipated behavior of other *Respondents*.

(c) In the third stage, interaction occurs locally within the resulting subgroups (or submarkets), each consisting of a *Proposer* and all *Respondents* who chose to apply to that specific agent. A matching mechanism determines which pair, if any, forms a contract within each subgroup. While such mechanisms may involve complex formats like various types of auctions, interest often lies in a non-strategic random matching rule. This corresponds, for example, to a scenario where a seller posts a price and subsequently selects a buyer at random from the pool of applicants.

The foundational work most closely related to this research is [1], which examines a product market with homogeneous firms and buyers, where each firm holds a single unit of inventory and matching occurs via a random mechanism between a firm and one of its potential customers. Subsequently, numerous advancements and extensions have been proposed, a comprehensive review of which is provided in [2]. However, generalizations that preserve the possibility of obtaining closed-form solutions are virtually non-existent. One such exception is [3], where the equilibrium is derived analytically for a setting in which firms are permitted to hold an arbitrary, positive volume of inventory.

Similar motifs arise in the stable matching problem. It is well known that the Gale–Shapley algorithm ([4]) identifies two solutions that constitute the upper and lower elements of the lattice of stable matchings. This implies that the version of the algorithm where firms act as the proposing side always yields a stable matching that is firm-optimal and worker-pessimal among all possible stable matchings. The converse holds when workers are the proposing side. In contrast, in the model considered here, depending on the market size, firms — as well as workers — may find it beneficial to forgo the proposer role. Furthermore, in a balanced market with an equal number of agents on both sides, the possibility of either side gaining an advantage by changing the design is absent.

Despite these considerations, existing literature has yet to address the question of how equilibrium characteristics depend on a key element of the basic market structure: namely, the assignment of "buyer" and "seller" roles. However, this question plays a crucial role in policy determination and the consequences of its implementation. For instance, an online platform connecting students and tutors can, by altering its architecture and user interaction design, structure its operations either through tutors posting resumes and students responding to them, or through students creating public requests and interested tutors responding. Depending on the estimated number of users on each side of the market, the platform — which typically earns a commission for creating a successful match — may implement specific functionalities to maximize the expected number of stable pairs formed and, ultimately, the expected commission. A similar question is relevant for service platforms (e.g., *profi.ru*), job search sites (e.g., *hh.ru*), and many others. It is important to note that, in practice, platforms often implement a combination of these options rather than a single pure variant. Nevertheless, understanding how equilibrium characteristics relate to one another in these two limiting cases within a stylized model can shed light on the structure of optimal platform design in more realistic settings.

As part of the analysis in this work, the problem is formulated in terms of a labor market with homogeneous firms, each possessing a single vacancy and willing to pay a maximum wage $wage_{\max}$, and homogeneous employees seeking employment provided a minimum wage $0 \leqslant wage_{\min} < wage_{\max}$ is guaranteed. A comparative analysis of equilibrium characteristics was conducted as a function of market size — i.e., the number of firms and employees — for cases where contract terms, in the form of wages, are proposed by employees through resume publication versus cases where wages are announced by firms via job postings.

The primary contribution of this study consists in characterizing the preference structures of various market participants — employees, firms, and society at large — as a function of market size. It is demonstrated that:

1) From a social welfare perspective, defined as the expected number of market transactions, it is preferable for the side with the smaller population to publish contract terms in the first stage, as this configuration yields a higher objective value than the alternative.

2) From the perspective of individual agents, when the numerical disparity between the two sides is relatively small, each side maximizes its own surplus by delegating the obligation of publishing contract terms to the opposing side, thereby assuming the role of *Respondents* in the second stage. Conversely, in scenarios where the number of firms and employees differs significantly, it becomes more advantageous for each side to act as the *Proposer* and publish the contract terms themselves.

3) Due to the inherent asymmetry and the fact that the threshold for "significant disparity" varies for each side based on their respective sizes, the set of all admissible market sizes is partitioned into distinct regions based on the various combinations of agent preferences regarding the *Proposer* role.

Furthermore, the extent of market inefficiency is quantified, demonstrating that at specific market scales, a mere adjustment of the market design can increase the expected number of executed contracts by more than 10%.

Another finding of the analysis pertains to the heterogeneity of preference motivations: a scenario is identified in which one market side may experience a decrease in its relative share of the aggregate surplus but nonetheless achieve a net gain in monetary terms. This occurs because improved coordination among market agents increases the aggregate surplus to such an extent that both sides realize a monetary benefit.

Consequently, these results provide a framework for determining the optimal design of online platforms or other market environments, contingent upon the objective function of the decision-maker.

2. MODEL

2.1. Market Description

Consider a labor market consisting of $n_f \geqslant 2$ homogeneous firms, each possessing exactly one job vacancy, and $n_w \geqslant 2$ job-seeking employees. Let the set of market sides be denoted by $S = \{\text{Firms, Workers}\}$. We assume that the minimum wage at which an employee is willing to work is $wage_{\min} \geqslant 0$, and the maximum wage a firm is willing to pay is $wage_{\max} > wage_{\min}$. Let the difference $wage_{\max} - wage_{\min}$ be denoted as $\Delta wage$. The market is structured as follows.

In the first stage, agents on one side of the market, the *Proposers* $\in S$, simultaneously and independently set a wage vector. For instance, if firms are the *Proposers*, each firm publishes a job description specifying a fixed wage. If workers are the *Proposers*, each worker publishes a resume specifying their desired wage. In either case, the announced wage cannot be subsequently altered and becomes a binding part of the resulting contract upon successful negotiations.

In the second stage, after observing the contract terms offered by the *Proposers*, the agents on the other side of the market (*Respondents* $\in S$) simultaneously and independently decide which single counterparty from the *Proposers*' side they wish to attempt to form a contract. If firms published job descriptions in the first stage, potential employees respond to them in the second stage, with each employee being able to apply to exactly one vacancy. Conversely, if employees initially published resumes, firms invite them for employment in the second stage, with each firm being able to approach exactly one employee.

In the third stage, the *Proposers* who initially posted the contract terms review the responses received. If an agent on this side receives at least one response, they select a counterparty for the transaction uniformly at random from among the applicants and successfully execute the contract. *Proposers* who receive no responses, as well as *Respondents* whose application was not accepted, remain unmatched and realize zero utility / profit.

All agents act rationally and strategically. Each agent maximizes their payoff, which is defined as $wage - wage_{\min}$ for employees and $wage_{\max} - wage$ for firms, in the event of a successful contract.

We denote the model where firms post vacancies in the first stage as FP (Firms Posting). In this model, firms are the *Proposers* and workers are the *Respondents*. We denote the model where employees post resumes in the first stage as WP (Workers Posting). In this case, workers are the *Proposers* and firms are the *Respondents*. It is worth noting that each version of the model independently represents an adaptation of the canonical directed search model [1] within the labor market context. In the following section, we derive the symmetric subgame perfect Nash equilibrium for each model and compare their equilibrium characteristics.

2.2. Equilibrium

In this section, we move away from the "firm/worker" terminology and reformulate the problem in terms of the *Proposers* and *Respondents* sides introduced in the introduction. Let the number of agents on the *Proposers* side be $n_p \geqslant 2$, and on the *Respondents* side be $n_r \geqslant 2$. Let the utility of each agent from concluding a contract (before accounting for costs or the price) be $val_p \geqslant 0$ for *Proposers* and $val_r \geqslant 0$ for *Respondents*. Depending on the structure of the utility functions, the payoffs for *Proposers* and *Respondents* take one of two forms:

$$\begin{cases} gain_p(val) = val - val_p, \\ gain_r(val) = val_r - val, \end{cases}$$

if, upon contract conclusion, a monetary transfer val flows in the direction of *Respondents* $\Longrightarrow$ *Proposers* (implying $val_r > val_p$), and

$$\begin{cases} gain_p(val) = val_p - val, \\ gain_r(val) = val - val_r \end{cases}$$

in the reverse case.

Theorem 1. *In the model under consideration, there exists a unique uncoordinated subgame perfect symmetric equilibrium in which each Proposer offers a contract price in the first stage equal to*

$$val = \frac{val_p \left(\frac{n_r}{n_p}\right)\left(1 - \frac{1}{n_p}\right)^{n_r} + val_r \left(1 - \left(1 - \frac{1}{n_p}\right)^{n_r - 1} \frac{n_r}{n_p} - \left(1 - \frac{1}{n_p}\right)^{n_r}\right)}{\frac{n_r}{n_p}\left(1 - \frac{1}{n_p}\right)^{n_r} + 1 - \left(1 - \frac{1}{n_p}\right)^{n_r - 1} \frac{n_r}{n_p} - \left(1 - \frac{1}{n_p}\right)^{n_r}},$$

and each Respondent in the second stage chooses a Proposer to approach uniformly at random with probability $\gamma = \frac{1}{n_p}$.

Proof. The proof follows the logic presented in [2] for Proposition 5, subject to differences in notation. Its key steps are provided below.

The pure strategy of each agent j among the *Proposers* is

$$val_j \in \left[\min(val_r, val_p), \max(val_r, val_p)\right].$$

Let $\mathbf{val} = (val_1, \ldots, val_{n_p})$ denote the strategy profile of the *Proposers*.

The search strategy for agent i among the *Respondents* is the vector

$$\gamma_{i*} = (\gamma_{i1}, \ldots, \gamma_{in_p}),$$

where γ_{ij} represents the probability that *Respondent* i chooses to attempt a contract with *Proposer* j. Let

$$\underset{n_r \times n_p}{\Gamma} = (\gamma_{1*}, \ldots, \gamma_{n_r *})^T = \begin{pmatrix} \gamma_{11} & \cdots & \gamma_{1n_p} \\ \vdots & \ddots & \vdots \\ \gamma_{n_r 1} & \cdots & \gamma_{n_r n_p} \end{pmatrix}$$

denote the strategy profile of the *Respondents*.

An equilibrium is a pair $(\mathbf{val}, \Gamma)$ such that no agent has an incentive to deviate; specifically, a symmetric equilibrium is one where $val_j = val$ and $\gamma_{ij} = \frac{1}{n_p}$. To verify whether a given pair $(\mathbf{val}, \Gamma)$ constitutes an equilibrium, we must examine the outcome following a unilateral deviation by one of the *Proposers*. We begin with the symmetric vector $\mathbf{val} = (val, \ldots, val)$ and, without loss of generality, assume that the first *Proposer* deviates from strategy val to strategy val^{dev}. Thus, the profile becomes $\mathbf{val} = (val^{dev}, val, \ldots, val)$. We must determine whether this deviation is profitable, assuming a symmetric equilibrium in the subgame following the deviation.

Suppose that for any agent i among the *Respondents*, the probability of attempting a contract with the deviating *Proposer* is $\gamma^{dev} = \gamma_{i1}(val^{dev}, val)$. The symmetric subgame perfect equilibrium is characterized by val and $\gamma^{dev}(val^{dev}, val)$ satisfying the following conditions:

(1) $val^{dev} = val$ maximizes the utility function of the potentially deviating *Proposer*, given no deviations by other agents;

(2) $(\gamma^{dev})^*(val^{dev}, val)$ constitutes a subgame equilibrium for any val^{dev} and val;

(3) on the equilibrium path, we have $\gamma_{ij} = \frac{1}{n_p}$, while after a deviation, *Respondents* visit the deviating *Proposer* with probability $\gamma^{dev} = \gamma^{dev}(val^{dev}, val)$, and all other *Proposers* uniformly at random with probability $\gamma^{nondev} = \frac{1-\gamma^{dev}}{n_p-1}$.

The probability that at least one *Respondent* attempts to contract with the deviating *Proposer* is equal to

$$\alpha_p^{dev} = 1 - \left(1 - \gamma^{dev}\right)^{n_r}.$$

Let the probability that a *Respondent* successfully concludes a contract, given they approach the deviating *Proposer*, be $\alpha_r^{dev} = \alpha_r^{dev}(val^{dev}, val)$. Note that $n_r\gamma^{dev}\alpha_r^{dev} = \alpha_p^{dev}$, as the left-hand side represents the expected number of *Respondents* with whom the deviating *Proposer* concludes a contract, while the right-hand side represents the expected number of successful contracts formed by the deviator. Thus,

$$\alpha_r^{dev} = \frac{1 - \left(1 - \gamma^{dev}\right)^{n_r}}{n_r\gamma^{dev}}.$$

The expected utility of the deviating *Proposer* is equal to

$$V_p^{dev}(val^{dev}, val) = \alpha_p^{dev} gain_p(val^{dev}) = \left(1 - \left(1 - \gamma^{dev}\right)^{n_r}\right) gain_p(val^{dev}).$$

The expected utility of a *Respondent* attempting to contract with the deviating *Proposer* is equal to

$$V_r^{dev}(val^{dev}, val) = \alpha_r^{dev} gain_r(val^{dev}) = \frac{1 - \left(1 - \gamma^{dev}\right)^{n_r}}{n_r\gamma^{dev}} gain_r(val^{dev}).$$

The expected utility of a *Respondent* attempting to contract with any of the non-deviating *Proposers* is equal to

$$V_r^{nondev}(val^{dev}, val) = \frac{1 - \left(1 - \gamma^{nondev}\right)^{n_r}}{n_r\gamma^{nondev}} gain_r(val).$$

Given the above, the first-order condition (FOC) for maximizing the utility $V_p^{dev}(val^{dev}, val)$ of the potentially deviating *Proposer* (assuming no deviations by others) is:

$$\begin{aligned} 0 = \frac{\partial V_p^{dev}}{\partial val^{dev}} &= sgn(val_r - val_p)\alpha_p^{dev} + gain_p(val^{dev})\frac{\partial \alpha_p^{dev}}{\partial val^{dev}} \\ &= sgn(val_r - val_p)\alpha_p^{dev} + n_r(1 - \gamma^{dev})^{n_r-1} gain_p(val^{dev})\frac{\partial \gamma^{dev}}{\partial val^{dev}}. \end{aligned}$$

In a symmetric mixed-strategy subgame equilibrium, *Respondents* must be indifferent between attempting a contract with the deviating *Proposer* or any of the non-deviating ones. Thus, γ^{dev} satisfies the indifference condition:

$$\frac{1 - (1 - \gamma^{dev})^{n_r}}{n_r\gamma^{dev}} gain_r(val^{dev}) = \frac{1 - (1 - \gamma^{nondev})^{n_r}}{n_r\gamma^{nondev}} gain_r(val).$$

For any $\gamma^{dev} \in (0, 1)$, one can obtain the implicit derivative $\frac{\partial \gamma^{dev}}{\partial val^{dev}}$ and substitute it into the FOC derived above for the deviating *Proposer*'s expected utility. Simplifying the expression by substituting $\gamma^{dev} = \frac{1}{n_p}$ and $val^{dev} = val$ proves that the deviation is not profitable if and only if val takes the value defined in the statement of this theorem.

The theorem is proven.

We now present the formulas for the key equilibrium characteristics that will be used in the subsequent comparative analysis of market designs.

Corollary 1. *In equilibrium, the unconditional probability of matching for each agent on the Respondents side is given by*

$$\nu(\gamma, n_r) = \sum_{i=1}^{n_p} \gamma\mu(\gamma, n_r) = \mu(\gamma, n_r) = \frac{1 - \left(1 - \frac{1}{n_p}\right)^{n_r}}{\frac{n_r}{n_p}},$$

where $\mu(\gamma, n_r)$ is the probability that a Respondent successfully forms a contract with their chosen Proposer, provided that this Respondent selects the specified Proposer with probability one, while the other $n_r - 1$ Respondents select that same Proposer with the equilibrium probability γ.

The equilibrium expected utility for each agent on the Respondents side is given by

$$\mathbf{E}\left[u_r\right] = gain_r \nu(\gamma, n_r) = \frac{|val_r - val_p| \left(1 - \frac{1}{n_p}\right)^{n_r} \left(1 - \left(1 - \frac{1}{n_p}\right)^{n_r}\right)}{\frac{n_r}{n_p}\left(1 - \frac{1}{n_p}\right)^{n_r} + 1 - \left(1 - \frac{1}{n_p}\right)^{n_r - 1} \frac{n_r}{n_p} - \left(1 - \frac{1}{n_p}\right)^{n_r}}$$

and is bounded as follows:

$$\mathbf{E}\left[u_r\right] \in |val_r - val_p| \left(\left(1 - \frac{1}{n_p}\right)^{n_r}, \left(1 - \frac{1}{n_p}\right)^{n_r - 1}\right).$$

The equilibrium expected Respondents Surplus (RS) is given by

$$\mathbf{E}\left[RS\right] = n_r \mathbf{E}\left[u_r\right].$$

The equilibrium expected utility for each agent on the Proposers side is given by

$$\mathbf{E}\left[u_p\right] = gain_p n_r \gamma \mu(\gamma, n_r)$$

$$= \frac{|val_r - val_p| \left(1 - \left(1 - \frac{1}{n_p}\right)^{n_r}\right) \left(1 - \left(1 - \frac{1}{n_p}\right)^{n_r - 1} \frac{n_r}{n_p} - \left(1 - \frac{1}{n_p}\right)^{n_r}\right)}{\frac{n_r}{n_p}\left(1 - \frac{1}{n_p}\right)^{n_r} + 1 - \left(1 - \frac{1}{n_p}\right)^{n_r - 1} \frac{n_r}{n_p} - \left(1 - \frac{1}{n_p}\right)^{n_r}}$$

and is bounded as follows:

$$\mathbf{E}\left[u_p\right] \in |val_r - val_p| \left(1 - \left(1 - \frac{1}{n_p}\right)^{n_r - 1} \frac{n_r}{n_p} - \left(1 - \frac{1}{n_p}\right)^{n_r}, \right.$$
$$\left. \left(1 - \left(1 - \frac{1}{n_p}\right)^{n_r - 1} \frac{n_r}{n_p} - \left(1 - \frac{1}{n_p}\right)^{n_r}\right)\left(1 + \frac{1}{n_p - 1}\right)\right).$$

The equilibrium expected Proposers Surplus (PS) is given by

$$\mathbf{E}\left[PS\right] = n_p \mathbf{E}\left[u_p\right].$$

The equilibrium expected Total Surplus (TS) is given by

$$\mathbf{E}\left[TS\right] = \mathbf{E}\left[PS\right] + \mathbf{E}\left[RS\right] = n_p \mathbf{E}\left[u_p\right] + n_r \mathbf{E}\left[u_r\right]$$
$$= |val_r - val_p|\, n_r \nu(\gamma, n_r) = |val_r - val_p|\, n_p \left(1 - \left(1 - \frac{1}{n_p}\right)^{n_r}\right).$$

Several points should be noted here.

First, using the aforementioned upper bound of the Respondent's expected utility in conjunction with the lower bound of the Proposer's expected utility yields the exact expression for the expected total surplus.

Second, under the assumption that the arguments n_p and n_r of the expected total surplus function are real-valued, the first partial derivative with respect to each argument equals the lower bound of the expected utility function for Proposers and Respondents, respectively.

The equilibrium expected Number of Deals (ND) is given by

$$\mathbf{E}\left[ND\right] = \frac{1}{|val_r - val_p|}\mathbf{E}\left[TS\right].$$

The expected Share of Deals (SD), representing the proportion of realized contracts relative to the maximum possible number, is given by

$$\mathbf{E}\left[SD\right] = \max\left(\frac{1}{n_p}, \frac{1}{n_r}\right)\mathbf{E}\left[ND\right].$$

The share of the expected Respondents Surplus within the expected Total Surplus is bounded as

$$\frac{\mathbf{E}\left[RS\right]}{\mathbf{E}\left[TS\right]} \in \left(\frac{n_r\left(1-\frac{1}{n_p}\right)^{n_r}}{n_p\left(1-\left(1-\frac{1}{n_p}\right)^{n_r}\right)}, \frac{n_r\left(1-\frac{1}{n_p}\right)^{n_r-1}}{n_p\left(1-\left(1-\frac{1}{n_p}\right)^{n_r}\right)}\right).$$

The equilibrium contract price can be expressed as

$$val = var_r + \left(val_p - val_r\right)\frac{\mathbf{E}\left[RS\right]}{\mathbf{E}\left[TS\right]}.$$

The proof is based on substituting the equilibrium price into the corresponding expressions followed by algebraic manipulations.

3. MARKET DESIGN ANALYSIS

Having described the equilibrium market characteristics in the general case in the previous section, we now return to the firm-employee model to examine how market size influences the preferences of each side regarding its design.

3.1. Market Efficiency and Social Preferences

We begin our analysis with one of the key aspects: the efficiency of the market mechanism. The following metrics can be used interchangeably as parameters characterizing efficiency: the expected number of contracts in equilibrium, the expected share of contracts, or the expected total surplus. These quantities were described individually in Corollary 1. Here, our primary interest lies in the differences in these values as a function of the market design.

Definition 1. The difference in the expected number of contracts between the FP and WP models as a function of market sizes is given by

$$\begin{aligned}\Delta_{ND}\left(n_f, n_w\right) &= \mathbf{E}\left[ND\middle|Proposers = Firms\right] - \mathbf{E}\left[ND\middle|Proposers = Workers\right] \\ &= n_f\left(1-\left(1-\frac{1}{n_f}\right)^{n_w}\right) - n_w\left(1-\left(1-\frac{1}{n_w}\right)^{n_f}\right).\end{aligned}$$

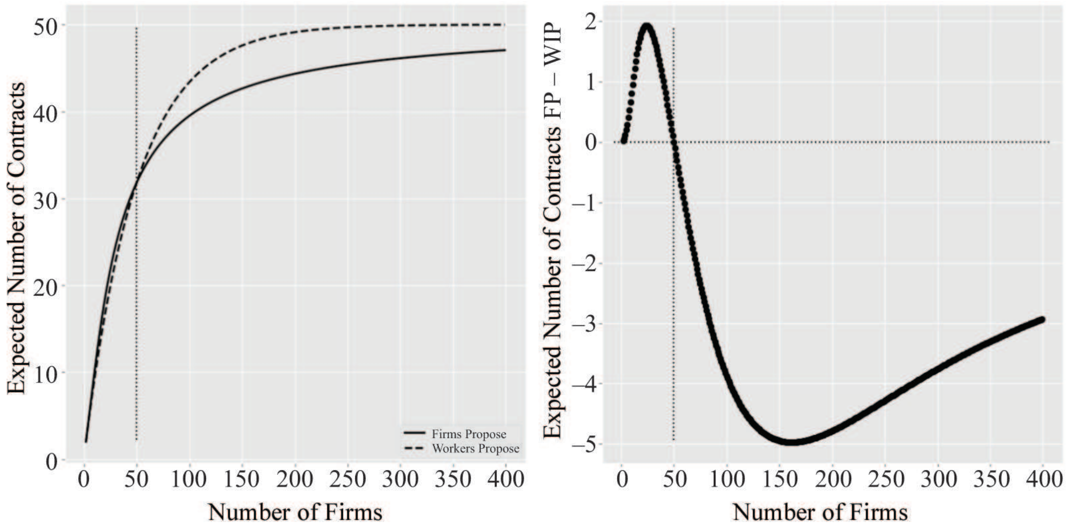

Fig. 1. Expected equilibrium number of contracts in the FP and WP models as a function of the number of firms for $n_w = 50$, $wage_{\min} = 100$, $wage_{\max} = 200$.

Definition 2. The difference in the expected share of concluded contracts relative to the maximum possible number between the FP and WP models, as a function of market sizes, is given by

$$\Delta_{SD}(n_f, n_w) = \mathbf{E}\left[SD \middle| Proposers = Firms\right] - \mathbf{E}\left[SD \middle| Proposers = Workers\right]$$
$$= \max\left(\frac{1}{n_f}, \frac{1}{n_w}\right)\Delta_{ND}(n_f, n_w).$$

Definition 3. The difference in the expected total surplus between the FP and WP models as a function of market sizes is given by

$$\Delta_{TS}(n_f, n_w) = \mathbf{E}\left[TS \middle| Proposers = Firms\right] - \mathbf{E}\left[TS \middle| Proposers = Workers\right]$$
$$= (wage_{\max} - wage_{\min})\,\Delta_{ND}(n_f, n_w).$$

It is worth noting that all three difference functions described above have the same sign for identical values of their arguments. For clarity and ease of interpretation, we refer to the graphical illustration in Fig. 1 and then formulate the analytical properties of these functions.

Theorem 2. *Consider the function $\Delta_{ND}(n_f, n_w)$ of the difference in the expected number of contracts between the FP and WP models. This function has the following properties:*

1) $\Delta_{ND}(n_f, n_w) \geqslant 0 \iff n_f \leqslant n_w$, *with* $\Delta_{ND}(n_f, n_w) = 0 \iff n_f = n_w$,

2) $\Delta_{ND}(n_f; n_w)$ *as a function of* n_f *for a fixed* n_w

(a) *is increasing on the interval* $[2, \theta_1)$, *where* $2 < \theta_1 < n_w$, *for* $n_w \geqslant 5$ (*for* $2 \leqslant n_w \leqslant 4$, *we define* θ_1 *to be* 2),

(b) *is decreasing on the interval* $[\theta_1, \theta_2)$, *where* $n_w < \theta_2 < \infty$,

(c) *is increasing on the interval* $[\theta_2, \infty)$.

The approximate values of θ_1 *and* θ_2 *are given by*

$$\begin{cases} \theta_1 \simeq \left(\dfrac{1}{k_2}\right) n_w \simeq 0.4747 n_w, \\ \theta_2 \simeq \left(\dfrac{1}{k_1}\right) n_w \simeq 3.2565 n_w, \end{cases}$$

where k_1 *and* k_2 *are the smaller and larger positive roots of the equation*

$$1-(1+k)\exp(-k)-\exp\left(-\frac{1}{k}\right)=0. \tag{1}$$

3) *In the limit as either argument tends to infinity, the function vanishes for any fixed value of the other argument.*

Proof. First and foremost, observe that

$$\Delta_{ND}(n_f=2;n_w)=2\left(1-\left(1-\frac{1}{2}\right)^{n_w}\right)-n_w\left(1-\left(1-\frac{1}{n_w}\right)^2\right)$$
$$=2\left(1-\frac{1}{2^{n_w}}\right)-n_w\left(1-\left(1-\frac{2}{n_w}+\frac{1}{n_w^2}\right)\right)=2-\frac{1}{2^{n_w-1}}-n_w\left(\frac{2}{n_w}-\frac{1}{n_w^2}\right)$$
$$=2-\frac{1}{2^{n_w-1}}-2+\frac{1}{n_w}=\frac{1}{n_w}-\frac{1}{2^{n_w-1}}>0,\quad \forall n_w\geqslant 3$$

since the exponential function grows faster than the linear one. Also, note that $\Delta_{ND}(n_w;n_w)=0$ due to symmetry.

We extend the definition of $\Delta_{ND}(n_f;n_w)$ to be a function of a real argument n_f for a given parameter n_w.

Calculating the first derivative yields

$$\Delta'_{ND}(n_f;n_w)=1-\left(1-\frac{1}{n_f}\right)^{n_w}-\frac{n_w}{n_f}\left(1-\frac{1}{n_f}\right)^{n_w-1}+n_w\ln\left(1-\frac{1}{n_w}\right)\left(1-\frac{1}{n_w}\right)^{n_f}.$$

Using the following approximations

$$\ln\left(1-\frac{1}{a}\right)\simeq-\frac{1}{a},$$
$$\left(1-\frac{1}{a}\right)^b\simeq\exp\left(-\frac{b}{a}\right),$$

we can simplify the expression for the first derivative as follows:

$$\Delta'_{ND}(n_f;n_w)\simeq 1-\left(1-\frac{1}{n_f}\right)^{n_w}-\frac{n_w}{n_f}\left(1-\frac{1}{n_f}\right)^{n_w-1}-\left(1-\frac{1}{n_w}\right)^{n_f},$$
$$\Delta'_{ND}(n_f;n_w)\simeq 1-\exp\left(-\frac{n_w-1}{n_f}\right)\left(1+\frac{n_w-1}{n_f}\right)-\exp\left(-\frac{n_f}{n_w}\right).$$

Applying the approximation $\frac{n_w-1}{n_f}\simeq\frac{n_w}{n_f}$, we introduce the substitution $\frac{n_w}{n_f}=k>0$.

$$\Delta'_{ND}(n_f;n_w)\simeq 1-(1+k)\exp(-k)-\exp\left(-\frac{1}{k}\right).$$

The first-order conditions defining the extrema of $\Delta_{ND}(n_f;n_w)$ are given by

$$\Delta'_{ND}(n_f;n_w)=0 \iff 1-(1+k)\exp(-k)-\exp\left(-\frac{1}{k}\right)=0.$$

The specified equation has exactly two solutions satisfying the condition $k>0$ (verified via WolframAlpha):

$$\begin{cases} k_1^*\simeq 0.307074,\\ k_2^*\simeq 2.106541.\end{cases}$$

Thus, the extrema expressed as functions of the parameter n_w take the form

$$\begin{cases} n^*_{f_1} \simeq 3.25654 n_w, \\ n^*_{f_2} \simeq 0.47471 n_w. \end{cases}$$

It should be noted that numerical analysis reveals that the correction $n^*_{f_1} \simeq 3.25654 n_w - 1$ provides a slightly better approximation of the minimum when Δ_{ND} is treated as a function of an integer argument. However, for the sake of consistency, the uncorrected version will be used throughout the main text.

The second derivative is given by

$$\Delta''_{ND}(n_f; n_w) = -\frac{n_w(n_w-1)}{n_f^3}\left(1-\frac{1}{n_f}\right)^{n_w-2} + n_w \ln^2\left(1-\frac{1}{n_w}\right)\left(1-\frac{1}{n_w}\right)^{n_f}.$$

Simplifying this expression via the logarithm approximation and comparing it with zero yields:

$$\Delta''_{ND}(n_f; n_w) < 0 \iff \frac{1}{n_w}\left(1-\frac{1}{n_w}\right)^{n_f} < \frac{n_w(n_w-1)}{n_f^3}\left(1-\frac{1}{n_f}\right)^{n_w-2},$$

$$\Delta''_{ND}(n_f; n_w) < 0 \iff \left(1-\frac{1}{n_f}\right)^{n_w-2}\frac{1}{n_f^3} > \left(1-\frac{1}{n_w}\right)^{n_f-1}\frac{1}{n_w^3}.$$

Taking the natural logarithm of both sides, applying the approximation once more, and performing algebraic simplifications, we obtain:

$$\Delta''_{ND}(n_f; n_w) < 0 \iff 3\ln(n_f) + \frac{n_w-2}{n_f} - \frac{n_f-1}{n_w} - 3\ln(n_w) < 0.$$

Substituting the parameterized expression for the extrema into the second-derivative inequality, we obtain the following:

$$\Delta''_{ND}\left(n^*_{f_2}(n_w); n_w\right) < 0 \iff 3\ln(0.475 n_w) + \frac{n_w-2}{0.475 n_w} - \frac{0.475 n_w - 1}{n_w} - 3\ln(n_w) < 0,$$

$$\Delta''_{ND}\left(n^*_{f_2}(n_w); n_w\right) < 0 \iff -0.6034 - \frac{-1.5253}{n_w} - \frac{-2.1065}{n_w-1} < 0 \ \forall n_w \geqslant 2.$$

Thus, the point $n^*_{f_2}$ corresponds to a maximum of the function.

Performing a similar exercise for the point $n^*_{f_1}$, we obtain

$$\Delta''_{ND}\left(n^*_{f_1}(n_w); n_w\right) > 0 \iff 3\ln(3.2565 n_w) + \frac{n_w-2}{3.2565 n_w} - \frac{3.2565 n_w - 1}{n_w} - 3\ln(n_w) < 0,$$

$$\Delta''_{ND}\left(n^*_{f_1}(n_w); n_w\right) > 0 \iff 7.1056 > \frac{7.2565}{n_w} + \frac{0.3071}{n_w-1} \ \forall n_w \geqslant 2.$$

Thus, the point $n^*_{f_1}$ corresponds to a minimum of the function.

It should be noted that

$$\begin{cases} n^*_{f_1} \simeq 3.25654 n_w > n_w, \\ n^*_{f_2} \simeq 0.47471 n_w < n_w, \end{cases}$$

$\forall n_w \geqslant 2$, i.e., the maximum of $\Delta_{ND}(n_f, n_w)$ is located to the left of n_w, while the minimum is to its right.

Let us now consider the asymptotic behavior.

$$\lim_{n_f\to\infty}\Delta_{ND}\left(n_f,n_w\right)=\lim_{n_f\to\infty}n_f\left(1-\left(1-\frac{1}{n_f}\right)^{n_w}\right)-\lim_{n_f\to\infty}n_w\left(1-\left(1-\frac{1}{n_w}\right)^{n_f}\right)=0,$$

$$\lim_{n_f\to\infty}n_w\left(1-\left(1-\frac{1}{n_w}\right)^{n_f}\right)=n_w,$$

$$\lim_{n_f\to\infty}n_f\left(1-\left(1-\frac{1}{n_f}\right)^{n_w}\right)=n_w.$$

In conclusion, a few points deserve mention.

First, since the original argument n_f of the function $\Delta_{ND}\left(n_w;n_w\right)$ is an integer rather than a real number, the identified minimum and maximum points require an adjustment to the nearest integer that yields the minimum or maximum value of Δ_{ND}.

Second, although approximations were used in the theorem's formulation and its proof, numerical analysis shows that the difference between the true integer extremum and the rounded approximate value does not exceed 1 for any value of n_w.

This completes the proof of the theorem.

Corollary 2. *The function $\Delta_{TS}\left(n_f,n_w\right)$, representing the difference in the expected total surplus between the FP and WP models, shares the same properties as described in Theorem 2 for the expected contract number difference $\Delta_{ND}\left(n_f,n_w\right)$.*

Proof. By definition,

$$\Delta_{TS}\left(n_f,n_w\right)=\left(wage_{\max}-wage_{\min}\right)\Delta_{ND}\left(n_f,n_w\right).$$

Thus, with respect to the properties in Theorem 2, these functions are equivalent as they differ only by a positive constant factor.

Corollary 3. *The function $\Delta_{SD}\left(n_f,n_w\right)$, representing the difference in the expected share of concluded contracts relative to the maximum possible number between the FP and WP models, exhibits the same properties as described in Theorem 2 for the expected number of contracts, subject to an adjustment of the maximum point coordinate λ_1. The approximate value of the point λ_1 is given by*

$$\lambda_1\simeq k_1n_w\simeq 0.3071n_w,$$

where k_1 is the smaller of the two positive roots of equation 1.

Proof. In the case $n_f\geqslant n_w$, by the definition of the function $\Delta_{SD}\left(n_f,n_w\right)$, we have:

$$\Delta_{SD}\left(n_f,n_w\right)=\frac{1}{n_w}\Delta_{ND}\left(n_f,n_w\right),$$

i.e., scaling by $\frac{1}{n_w}$ affects neither the zeros of Δ_{ND} nor its behavior as a function of n_f for a fixed n_w.

In the case $n_f<n_w$, the argument follows the same logic as in the proof of Theorem 2.

The above discussion leads to the following formulation.

Proposition 1. *From the perspective of social welfare and market efficiency, measured by any of the methods given in definitions 1, 2, and 3, it is preferable that the contract terms in the first stage be published by the side of the market with fewer participants.*

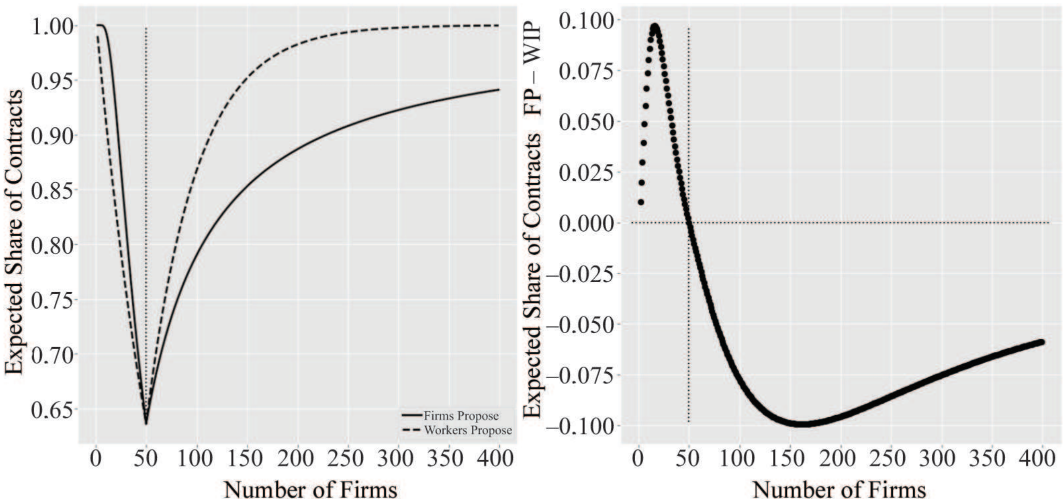

Fig. 2. Expected equilibrium share of concluded contracts in the FP and WP models as a function of the number of firms for $n_w = 50$, $wage_{\min} = 100$, $wage_{\max} = 200$.

At the same time, if we move beyond a binary comparison of relative efficiency between FP and WP for fixed market sizes and instead examine how the superiority of one design over another evolves as market sizes change, it is more appropriate to consider relative metrics, such as the expected share of concluded contracts in equilibrium, a graphical illustration of which is presented in Fig. 2.

It is worth noting that market sizes $n_f = n_w$ are the least efficient in terms of the expected share of concluded contracts in both models (FP and WP); furthermore, due to this extreme inefficiency, society cannot gain any advantage by switching the market design. Conversely, for any other market size, one of the designs is strictly preferable from a social standpoint. It should be noted, however, that in the case of a correlated equilibrium, a market with an equal number of agents on both sides is the most efficient, as it ensures not only a 100% share of concluded contracts but also the absence of unmatched agents on either side.

This situation is analogous to using the ROC-AUC (Area Under the Receiver Operating Characteristic Curve) metric for evaluating the performance of binary classification models in machine learning. A model with an ROC-AUC of 0.5 represents the worst possible case, as it corresponds to random guessing and cannot be improved. Conversely, a model with a very low ROC-AUC, such as 0.01, which implies an inverse ranking of classes, can be significantly improved to an ROC-AUC of $(1 - 0.01)$ by simply flipping the class labels.

Let us now consider the potential gains in market efficiency achievable through a change in its design.

Corollary 4. *The maximum and minimum values of the function $\Delta_{SD}(n_f; n_w)$, representing the difference in the expected share of concluded contracts relative to the maximum possible number between the FP and WP models as a function of n_f for a fixed n_w, are approximately equal in absolute value to*

$$\Delta_{SD}(k_1 n_w; n_w) \simeq -\Delta_{SD}\left(\frac{n_w}{k_1}; n_w\right) \simeq 1 - \frac{1}{k_1} + \frac{1}{k_1}\exp(-k_1) - \exp\left(-\frac{1}{k_1}\right) \simeq 10.04\%$$

and are attained at the points

$$\begin{cases} \lambda_1 \simeq k_1 n_w \simeq 0.3071 n_w, \\ \theta_2 \simeq \left(\dfrac{1}{k_1}\right) n_w \simeq 3.2565 n_w, \end{cases}$$

where k_1 is the smaller of the two positive roots of equation 1.

Thus, for certain market sizes, the gain in efficiency solely due to a change in design can exceed 10% in terms of the expected share of concluded contracts.

Remark 1. It is important to note that, due to the approximations used, the approximate maximum and minimum values of the function $\Delta_{SD}(n_f; n_w)$ are independent of the parameter n_w. The true maximum and minimum values do not possess this property. However, the absolute discrepancy does not exceed 0.4% for $n_w \geqslant 50$ and diminishes as n_w increases further.

3.2. Firms' Preferences

Let us now examine the preferences of firms regarding the market design as a function of its size. The characteristics analyzed include the expected profit of an individual firm and the expected aggregate surplus of all firms, as well as the dependence of these variables on the market mechanism design.

Definition 4. The difference in a firm's expected profit between the FP and WP models, as a function of market sizes, is given in approximate form by

$$\Delta_{u_F}(n_f, n_w) = \mathbf{E}\left[u_p \middle| Proposers = Firms\right] - \mathbf{E}\left[u_r \middle| Proposers = Workers\right]$$
$$\simeq \Delta wage\left(1 - \left(1 - \frac{1}{n_f}\right)^{n_w - 1}\frac{n_w}{n_f} - \left(1 - \frac{1}{n_f}\right)^{n_w} - \left(1 - \frac{1}{n_w}\right)^{n_f}\right).$$

Remark 2. In the FP model, firms act as *Proposers*, whereas in the WP model, they act as *Respondents*; therefore, to obtain the exact difference function, one must use the corresponding values from Corollary 1. However, a full-scale analysis of the exact function is not feasible. Consequently, we employ the aforementioned approximation, which lies between the lower and upper bounds of the exact function derived using the corresponding bounds from Corollary 1.

Definition 5. The difference in the expected total surplus of firms between the FP and WP models, as a function of market sizes, is given in approximate form by

$$\Delta_{FS}(n_f, n_w) = \mathbf{E}\left[PS \middle| Proposers = Firms\right] - \mathbf{E}\left[RS \middle| Proposers = Workers\right] = n_f \Delta_{u_F}(n_f, n_w).$$

A graphical illustration for a particular case is depicted in Fig. 3 for the expected firms' profits and in Fig. 4 for the expected firms' aggregate surplus. Let us now formulate the analytical properties of these functions.

Theorem 3. *Consider the function $\Delta_{u_F}(n_f, n_w)$, representing the difference in a firm's expected profit between the FP and WP models. This function has the following properties.*

1) *$\Delta_{u_F}(n_f; n_w)$, as a function of n_f for a fixed n_w, has exactly two roots, θ_1 and θ_2, the approximate values of which are given by*

$$\begin{cases} \theta_1 \simeq \left(\dfrac{1}{k_2}\right) n_w \simeq 0.4747 n_w, \\ \theta_2 \simeq \left(\dfrac{1}{k_1}\right) n_w \simeq 3.2565 n_w, \end{cases}$$

where k_1 and k_2 are the smaller and larger positive roots of equation 1, respectively. Furthermore,

$$\begin{cases} \Delta_{u_F}(n_f; n_w) > 0 \iff n_f < \theta_1 \vee n_f > \theta_2, \\ \Delta_{u_F}(n_f; n_w) < 0 \iff \theta_1 < n_f < \theta_2. \end{cases}$$

2) *$\Delta_{u_F}(n_f; n_w)$, as a function of n_f for a fixed n_w,*

(a) *is increasing on the interval $[2, \xi_1)$, where $2 < \xi_1 < \theta_1$, for $n_w \geqslant 9$ (for $2 \leqslant n_w \leqslant 8$ we define ξ_1 to be 2),*

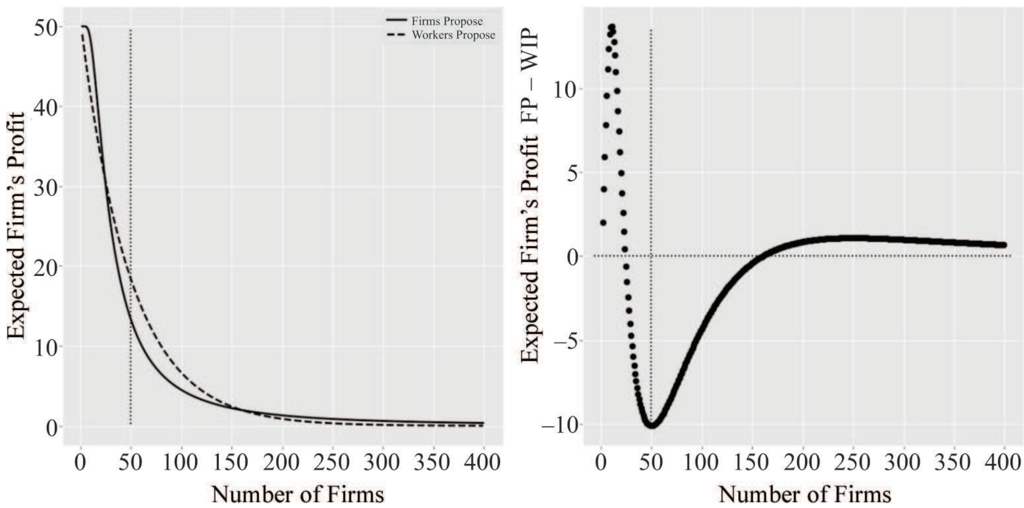

Fig. 3. Expected equilibrium firm's profit in the FP and WP models as a function of the number of firms for $n_w = 50$, $wage_{\min} = 100$, $wage_{\max} = 200$.

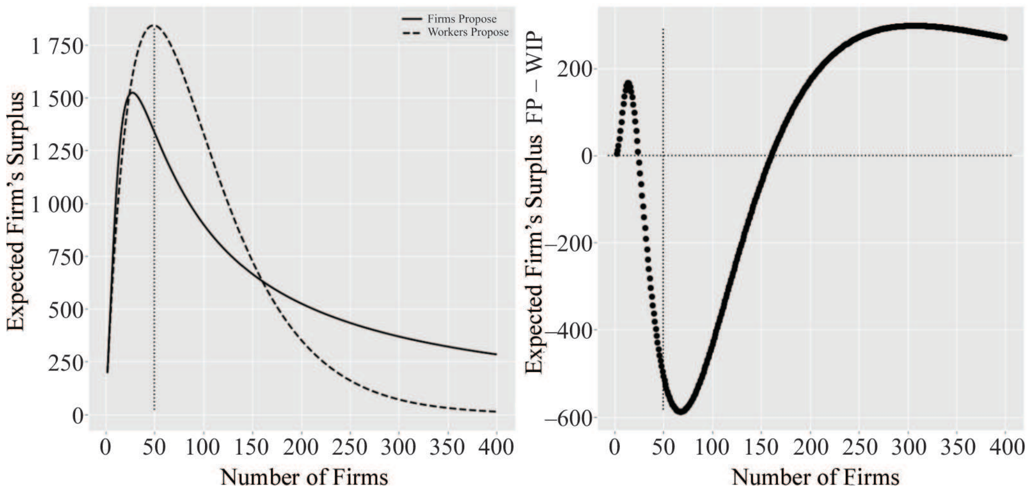

Fig. 4. Expected equilibrium firms' surplus in the FP and WP models as a function of the number of firms for $n_w = 50$, $wage_{\min} = 100$, $wage_{\max} = 200$.

(b) *is decreasing on the interval* $[\xi_1, \xi_2)$, *where* $\xi_2 = n_w$,
(c) *is increasing on the interval* $[\xi_2, \xi_3)$, *where* $\theta_2 < \xi_3 < \infty$,
(d) *is decreasing on the interval* $[\xi_3, \infty)$.

The approximate values of ξ_1, ξ_2, *and* ξ_3 *are given by*

$$\begin{cases} \xi_1 \simeq p_1 n_w \simeq 0.1975 n_w, \\ \xi_2 = p_2 n_w = n_w, \\ \xi_3 \simeq p_3 n_w \simeq 5.0639 n_w, \end{cases}$$

where p_1, p_2, *and* p_3 *are roots of the equation*

$$3\ln(p) + \frac{1}{p} - p = 0 \tag{2}$$

in increasing order. It is worth noting that $p_1 = \frac{1}{p_3}$.

3) *In the limit as either argument tends to infinity, the function vanishes for any fixed value of the other argument.*

Proof. We extend the definition of $\Delta_{u_F}(n_f; n_w)$ to treat n_f as a real-valued argument for a given parameter n_w.

The proof of the first part of this theorem is covered by the proof of Theorem 2. This follows from the fact that the approximation of the function $\Delta_{u_F}(n_f; n_w)$ given in Definition 4 is identically equal to the approximation of the first derivative of $\Delta_{ND}(n_f, n_w)$ used in the proof of Theorem 2. Consequently, the zeros of the first derivative of $\Delta_{ND}(n_f; n_w)$ and the zeros of $\Delta_{u_F}(n_f; n_w)$ coincide.

Subject to the approximations used, the first derivative of the function $\Delta_{u_F}(n_f; n_w)$ is equal to the second derivative of the function $\Delta_{ND}(n_f, n_w)$, the expression for the zeros of which has already been derived in the proof of Theorem 2.

$$\Delta'_{u_F}(n_f; n_w) = 0 \iff 3\ln\left(\frac{n_f}{n_w}\right) + \frac{n_w}{n_f} - \frac{n_f}{n_w} - \frac{2}{n_f} + \frac{1}{n_w} = 0.$$

Within the framework of the approximation, neglecting the term $-\frac{2}{n_f}+\frac{1}{n_w}$ and substituting $\frac{n_f}{n_w} = p$, we obtain the equation

$$\Delta'_{u_F}(n_f; n_w) = 0 \iff 3\ln(p) + \frac{1}{p} - p = 0.$$

Note that $p = 1$ is a root. Furthermore, if some p^* is a root, then $\frac{1}{p^*}$ is also a root. The approximate value of the root other than unity (obtained via WolframAlpha) is 0.1975. Using the interval method, it can be shown that ξ_1 and ξ_3 are local maxima, whereas ξ_2 is a local minimum.

This completes the proof of the theorem.

Corollary 5. *The function $\Delta_{FS}(n_f, n_w)$, representing the difference in the expected aggregate surplus of firms between the FP and WP models as a function of market sizes, exhibits the same properties as described in Theorem 3 for the function $\Delta_{u_F}(n_f, n_w)$, adjusted for the coordinates of the extrema ϕ_1, ϕ_2, and ϕ_3. The approximate values of these extrema are given by*

$$\begin{cases} \phi_1 \simeq \left(\dfrac{1}{r_3}\right) n_w \simeq \frac{1}{3.7779} n_w \simeq 0.2647 n_w, \\ \phi_2 \simeq \left(\dfrac{1}{r_2}\right) n_w \simeq \frac{1}{0.7388} n_w \simeq 1.3536 n_w, \\ \phi_3 \simeq \left(\dfrac{1}{r_1}\right) n_w \simeq \frac{1}{0.1608} n_w \simeq 6.2189 n_w, \end{cases}$$

where r_1, r_2, and r_3 are the roots of the equation

$$1 - \left(1 + r + r^2\right)\exp(-r) - \left(1 - \frac{1}{r}\right)\exp\left(-\frac{1}{r}\right) = 0$$

given in increasing order.

Proof. The proof is similar to that of Theorem 3.

The economic interpretation of the results obtained above can be stated as follows.

Proposition 2. *In a market where the sizes of the two sides do not differ significantly:*

$$\frac{1}{k_2} \simeq 0.4747 \leqslant \frac{n_f}{n_w} \leqslant \frac{1}{k_1} \simeq 3.2565,$$

firms prefer to take the side of Respondents, leaving the responsibility of publishing the contract terms to the workers. Conversely, if the market is significantly unbalanced in either direction, it is more preferable for the firms to publish the terms themselves as Proposers.

3.3. Workers' Preferences

We now turn our attention to the analysis of workers' preferences regarding the market design. In this regard, both the expected utility of an individual worker and the expected aggregate surplus of all workers are of particular interest; however, the primary focus lies in the differences in these values depending on the market design.

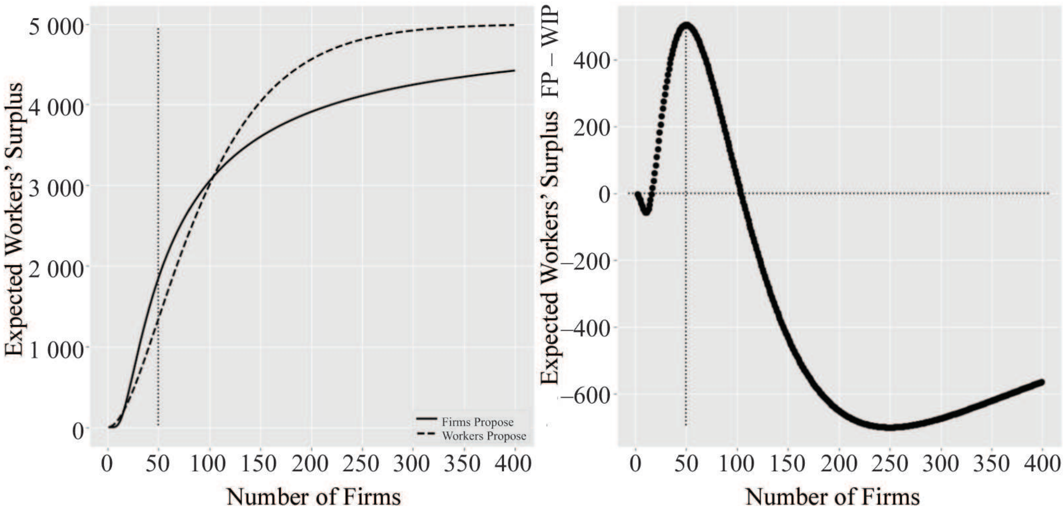

Fig. 5. Expected equilibrium workers' surplus in the FP and WP models as a function of the number of firms for $n_w = 50$, $wage_{\min} = 100$, $wage_{\max} = 200$.

Definition 6. The difference in a worker's expected utility between the FP and WP models, as a function of market sizes, is given in approximate form by

$$\Delta_{u_W}(n_f, n_w) = \mathbf{E}\left[u_r \middle| Proposers = Firms\right] - \mathbf{E}\left[u_p \middle| Proposers = Workers\right]$$
$$\simeq \Delta wage\left(\left(1 - \frac{1}{n_f}\right)^{n_w} + \left(1 - \frac{1}{n_w}\right)^{n_f - 1}\frac{n_f}{n_w} + \left(1 - \frac{1}{n_w}\right)^{n_f} - 1\right).$$

Definition 7. The difference in the expected total surplus of workers between the FP and WP models, as a function of market sizes, is given in approximate form by

$$\Delta_{WS}(n_f, n_w) = \mathbf{E}\left[RS \middle| Proposers = Firms\right] - \mathbf{E}\left[PS \middle| Proposers = Workers\right] \simeq n_w \Delta_{u_W}(n_f, n_w).$$

A graphical illustration for a particular case is presented in Fig. 5. The functions introduced above possess the following analytical properties.

Theorem 4. *Consider the function $\Delta_{u_W}(n_f, n_w)$, representing the difference in a worker's expected utility between the FP and WP models. This function has the following properties.*

1) *$\Delta_{u_W}(n_f; n_w)$, as a function of n_f for a fixed n_w, has exactly two roots, λ_1 and λ_2, the approximate values of which are given by*

$$\begin{cases} \lambda_1 \simeq k_1 n_w \simeq 0.3071 n_w, \\ \lambda_2 \simeq k_2 n_w \simeq 2.1065 n_w, \end{cases}$$

where k_1 and k_2 are the smaller and larger positive roots of equation 1. Furthermore,

$$\begin{cases} \Delta_{u_W}(n_f; n_w) < 0 \iff n_f < \lambda_1 \vee n_f > \lambda_2, \\ \Delta_{u_W}(n_f; n_w) > 0 \iff \lambda_1 < n_f < \lambda_2. \end{cases}$$

2) *$\Delta_{u_W}(n_f; n_w)$, as a function of n_f for a fixed n_w,*

(a) *is decreasing on the interval $[2, \xi_1)$, where $2 < \xi_1 < \lambda_1$, for $n_w \geqslant 9$ (for $2 \leqslant n_w \leqslant 8$ we define ξ_1 to be 2),*

(b) *is increasing on the interval* $[\xi_1, \xi_2)$*, where* $\xi_2 = n_w$*,*
(c) *is decreasing on the interval* $[\xi_2, \xi_3)$*, where* $\lambda_2 < \xi_3 < \infty$*,*
(d) *is increasing on the interval* $[\xi_3, \infty)$*.*
The approximate values of ξ_1*,* ξ_2*, and* ξ_3 *are given by*

$$\begin{cases} \xi_1 \simeq p_1 n_w \simeq 0.1975 n_w, \\ \xi_2 = p_2 n_w = n_w, \\ \xi_3 \simeq p_3 n_w \simeq 5.0639 n_w, \end{cases}$$

where p_1*,* p_2*, and* p_3 *are roots of the equation 2 in increasing order.*
3) *In the limit as either argument tends to infinity, the function vanishes for any fixed value of the other argument.*

The proof is similar to that of Theorem 3.

Corollary 6. *The function* $\Delta_{WS}(n_f, n_w)$*, representing the difference in the expected aggregate surplus of workers between the FP and WP models as a function of market sizes, exhibits the same properties as described in Theorem 4 for the function* $\Delta_{u_W}(n_f, n_w)$*.*

The economic interpretation of the results obtained above can be stated as follows.

Proposition 3. *In a market where the sizes of the two sides do not differ significantly:*

$$k_1 \simeq 0.3071 \leqslant \frac{n_f}{n_w} \leqslant k_2 \simeq 2.1065,$$

workers prefer to take the side of Respondents, leaving the responsibility of publishing the contract terms to the firms. Conversely, if the market is significantly unbalanced in either direction, it is more preferable for the workers to publish the terms themselves as Proposers.

3.4. Market Power and Equilibrium Wage

Another important perspective illustrating the comparative advantage of firms over workers (or vice versa) and reflecting their market power is the expected share of one side's surplus in the expected aggregate surplus. The higher this share, the more favorable the position of one market side relative to the other. It should be noted, however, that this analysis focuses solely on the division of the "pie" rather than its absolute size.

It should be noted that the economic significance of the degree of market power is rooted in the nature of the equilibrium market price. As significance in Corollary 1, the equilibrium wage, being a weighted average of $wage_{\min}$ and $wage_{\max}$, is expressed as a strictly monotonic linear function of the share of one side's expected surplus in the total expected surplus. For convenience, and without loss of generality, the following analysis is conducted using the normalized version.

Definition 8. The difference in the share of the workers' expected surplus in the total expected surplus between the FP and WP models, as a function of market sizes, is approximately given by

$$\Delta_{\frac{WS}{TS}}(n_f, n_w) \simeq \frac{\mathbf{E}\left[RS \middle| Proposers = Firms\right]}{\mathbf{E}\left[TS \middle| Proposers = Firms\right]} - \frac{\mathbf{E}\left[PS \middle| Proposers = Workers\right]}{\mathbf{E}\left[TS \middle| Proposers = Workers\right]}$$

$$\simeq \frac{n_w\left(1 - \frac{1}{n_f}\right)^{n_w}}{n_f\left(1 - \left(1 - \frac{1}{n_f}\right)^{n_w}\right)} + \frac{n_f\left(1 - \frac{1}{n_w}\right)^{n_f}}{n_w\left(1 - \left(1 - \frac{1}{n_w}\right)^{n_f}\right)} - 1.$$

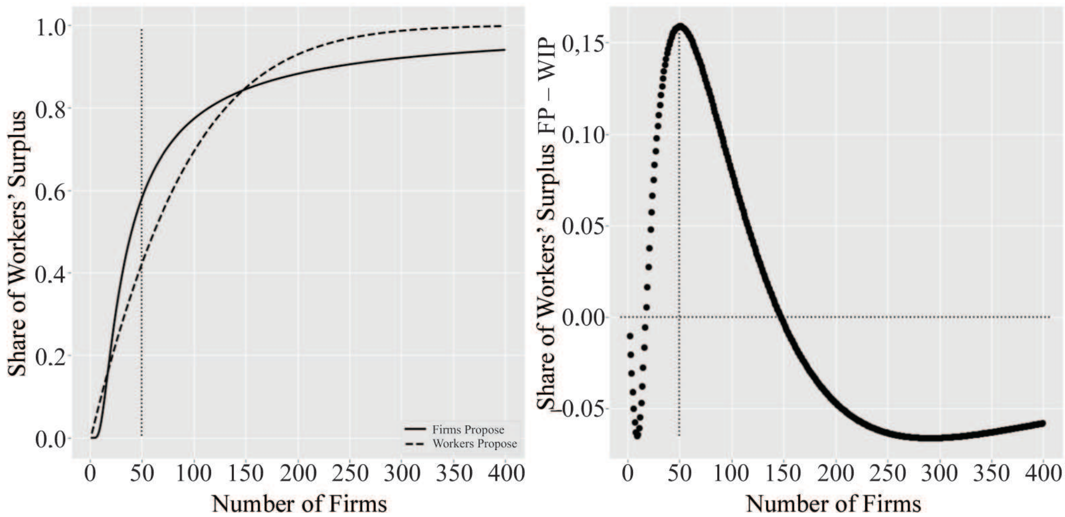

Fig. 6. The share of the workers' expected surplus in the total expected surplus in the FP and WP models as a function of the number of firms for $n_w = 50$, $wage_{\min} = 100$, $wage_{\max} = 200$.

By employing an additional approximation, it may be rewritten as:

$$\Delta_{\frac{WS}{TS}}(x) \simeq \frac{1}{x}\left(\frac{\exp\left(-\frac{1}{x}\right)}{1-\exp\left(-\frac{1}{x}\right)}\right) + x\left(\frac{\exp(-x)}{1-\exp(-x)}\right) - 1,$$

where $x = \frac{n_f}{n_w}$.

A graphical illustration of the function under discussion for a particular case is presented in Fig. 6. The introduced function possesses the following analytical properties.

Theorem 5. *Consider the function $\Delta_{\frac{WS}{TS}}(n_f, n_w)$, representing the difference in the share of the workers' expected surplus in the total expected surplus between the FP and WP model. This function has the following properties.*

1) $\Delta_{\frac{WS}{TS}}(n_f; n_w)$, *as a function of n_f for a fixed n_w, has exactly two roots, α_1 and α_2, the approximate values of which are given by*

$$\begin{cases} \alpha_1 \simeq s_1 n_w \simeq 0.3337 n_w, \\ \alpha_2 \simeq s_2 n_w \simeq 2.9963 n_w, \end{cases}$$

where s_1 and s_2 are the smaller and larger positive roots of the equation

$$\frac{1}{s}\left(\frac{\exp\left(-\frac{1}{s}\right)}{1-\exp\left(-\frac{1}{s}\right)}\right) + s\left(\frac{\exp(-s)}{1-\exp(-s)}\right) - 1 = 0.$$

Furthermore,

$$\begin{cases} \Delta_{\frac{WS}{TS}}(n_f; n_w) < 0 \iff n_f < \alpha_1 \vee n_f > \alpha_2, \\ \Delta_{\frac{WS}{TS}}(n_f; n_w) > 0 \iff \alpha_1 < n_f < \alpha_2. \end{cases}$$

2) $\Delta_{\frac{WS}{TS}}(n_f; n_w)$, *as a function of n_f for a fixed n_w,*

(a) *is decreasing on the interval $[2, \omega_1)$, where $2 < \omega_1 < \alpha_1$, for $n_w \geqslant 11$ (for $2 \leqslant n_w \leqslant 10$ we define ω_1 to be 2),*

(b) *is increasing on the interval $[\omega_1, \omega_2)$, where $\omega_2 = n_w$,*

(c) *is decreasing on the interval $[\omega_2, \omega_3)$, where $\alpha_2 < \omega_3 < \infty$,*

(d) *is increasing on the interval $[\omega_3, \infty)$.*

The approximate values of ω_1, ω_2, *and* ω_3 *are given by*

$$\begin{cases} \omega_1 \simeq t_1 n_w \simeq 0.1697 n_w, \\ \omega_2 = t_2 n_w = n_w, \\ \omega_3 \simeq t_3 n_w \simeq 5.8925 n_w, \end{cases}$$

where t_1, t_2, *and* t_3 *are positive roots of the equation*

$$\frac{t + \exp\left(-\frac{1}{t}\right) - t\exp\left(-\frac{1}{t}\right)}{\left(\exp\left(-\frac{1}{t}\right) - 1\right)^2 t^3} + \frac{1-t}{\exp(t) - 1} - \frac{t}{(\exp(t) - 1)^2} = 0$$

in increasing order.

3) *In the limit as either argument tends to infinity, the function vanishes for any fixed value of the other argument.*

The proof is similar to that of Theorem 3.

Remark 3. Since the shares of the workers' and firms' expected surpluses in the total expected surplus sum to 1, the difference function for the firms' surplus share is given by $1 - \Delta_{\frac{WS}{TS}}(n_f, n_w)$. This function possesses corresponding properties, and its graph is a reflection of the graph of $\Delta_{\frac{WS}{TS}}(n_f; n_w)$ across the horizontal axis.

The economic interpretation of the results obtained above can be stated as follows.

Proposition 4. *In a market where the sizes of the two sides do not differ significantly:*

$$s_1 \simeq 0.3337 \leqslant \frac{n_f}{n_w} \leqslant s_2 \simeq 2.9963,$$

workers, from the perspective of maximizing their share of the expected total surplus, prefer to take the side of Respondents, leaving the responsibility of publishing the contract terms to the firms. Conversely, if the market sizes differ significantly in either direction, workers find it more preferable to take the side of Proposers to maximize their share.

The preferences of firms regarding the maximization of their share of the expected total surplus are, in this case, exactly the opposite.

Remark 4. It is important to highlight the key distinction between Propositions 3 and 4. In Proposition 3, workers are concerned with maximizing their utility or surplus in monetary terms. In contrast, in Proposition 4, workers choose the market design to maximize their share of the expected total surplus for given market sizes. A consequence of the different thresholds for these two objectives is the emergence of market size intervals where these goals come into conflict with each other.

3.5. Comparative Analysis of the Market Participants' Preferences

Having examined the preferences of each market side individually (workers, firms, and society as a whole), we now consolidate the results of this analysis.

Figure 7 graphically illustrates the combined results of Propositions 1–4 regarding the preferences of the market sides. Based on this, the following conclusions can be drawn.

First, if the parties' preferences are considered strictly in monetary terms, the $\frac{n_f}{n_w}$ axis is divided into six intervals. Within two of these, specifically $\frac{n_f}{n_w} \in (0.3071, 0.4747)$ and $\frac{n_f}{n_w} \in (2.1065, 3.2565)$, the preferences of all parties coincide; in the remaining intervals, they come into conflict.

Second, incorporating the parties' preferences regarding the maximization of their own share of total surplus reveals a significant effect. For $\frac{n_f}{n_w} \in (0.3071, 0.3337)$, from the perspective of maximizing monetary value, it is more beneficial for society, firms, and workers alike if firms take

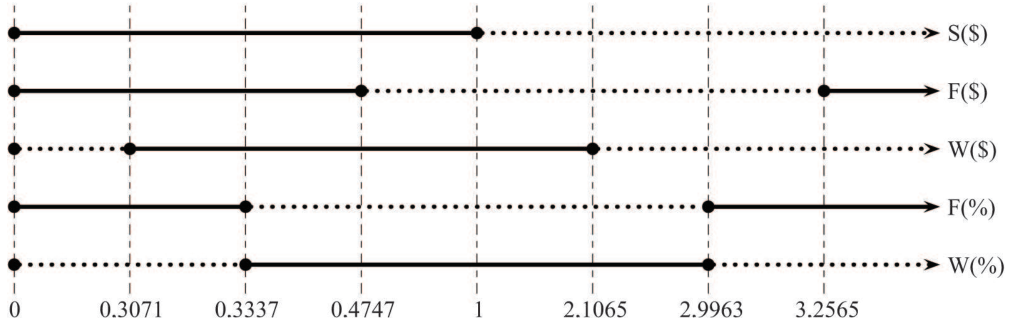

Fig. 7. Market design preferences of firms, workers, and society as a function of the side ratio $\frac{n_f}{n_w}$. (The solid line represents the preference for firms to act as *Proposers*, while the dashed line represents the preference for workers to act as *Proposers*.)

the role of *Proposers*. However, from the standpoint of maximizing their share of the total surplus, workers would prefer to be the proposing side themselves. This implies a situation where one party's pursuit of a larger "slice of the pie" undermines the monetary interests of all market participants, including that very party. A similar effect is observed for firms when $\frac{n_f}{n_w} \in (2.9963, 3.2565)$.

4. CONCLUSION

The conducted analysis describes the structure of preferences of various market participants regarding market design as a function of market size. It was demonstrated that this structure is non-homogeneous: depending on the firm-to-worker ratio, the preferences of the parties may shift several times, occasionally coinciding with one another. The system is most sensitive to changes in design at the ratios $\frac{n_f}{n_w} \simeq 0.3071$ and $\frac{n_f}{n_w} \simeq \frac{1}{0.3071}$, where the gain in the expected share of concluded contracts can exceed 10%. Another key result is the demonstration of the heterogeneity in the choice of objective variables when determining agent preferences. It was shown that a market side may lose a portion of its relative share of the aggregate surplus due to a change in design, yet still gain in monetary terms due to an increase in the total surplus resulting from more efficient agent coordination.

Overall, the findings underscore the importance of market design as a fundamental element of market structure. These results can be applied, for instance, by various online platforms to optimize their objective functions.

FUNDING

This work was supported by HSE University (Basic Research Program).

REFERENCES

1. Burdett, K., Shi, S., and Wright, R., Pricing and Matching with Frictions, *Journal of Political Economy*, 2001, vol. 109, no. 5, pp. 1060–1085.
2. Wright, R., Kircher, P., Julien, B., and Guerrieri, V., Directed Search and Competitive Search Equilibrium: A Guided Tour, *Journal of Economic Literature*, 2021, vol. 59, no. 1, pp. 90–148.
3. Sandomirskaia, M. and Shavshin, R., Pricing, Market Power, and Friction in a Finite Market: The Role of Capacities, *Higher School of Economics Research Paper*, 2024, vol. 267.
4. Gale, D. and Shapley, L., College Admissions and The Stability of Marriage, *The American Mathematical Monthly*, 1962, vol. 69, no. 1, pp. 9–15.

This paper was recommended for publication by F.T. Aleskerov, a member of the Editorial Board

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2026, Vol. 87, No. 3, pp. 270–286.

Optimization Statements of Signal Detection Problems

A. A. Galyaev[*,a], **L. M. Berlin**[*,b], **P. V. Lysenko**[*,c], **and V. G. Babikov**[*,d]

[*] *Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia*
e-mail: [a]*galaev@ipu.ru,* [b]*berlin.lm@phystech.edu,* [c]*pavellysen@ipu.ru,* [d]*babikov@ipu.ru*

Received July 7, 2025
Revised September 1, 2025
Accepted September 24, 2025

Abstract—The article develops the authors' previous work and examines the sequence of problems that form an approach to solving the problem of detecting a signal in noise, including cases with a low signal-to-noise ratio. The sequence of mathematical problems needed to be solved is determined. For the case of proximity of two hypotheses, analytical constructions are used to obtain test statistics, observational statistics, and decision rules based on frequency, time, and time-frequency distributions. In this paper, the possibility of increasing the signal-to-noise ratio is established when distinguishing between two hypotheses for d-signals. The theoretical results are established by the proof of the corresponding lemmas.

Keywords: detection criteria, signal processing, Fourier transform, signal detection in noise

DOI: 10.7868/S1608303226030073

1. INTRODUCTION

At first glance, the problem of detecting a deterministic signal in the presence of environmental noise appears to be a binary classification problem. However, the deeper a researcher dives into formalizing the problem, the more questions arise regarding what information they possess about the properties of both the signal and the noise and about how to correctly interpret the solution to this problem. In the general case, the detection problem is a binary classification problem that requires formulating a decision rule regarding the presence or absence of a useful signal in a signal-noise mixture. In this context, features derived from the input time series may belong to a set of different classes which include both the useful signal and the noise. The more classes available for decision-making, the more precisely one needs to know the frequency-time properties of the signal and noise [1, 2].

In binary classification, the discrimination between two hypotheses relies on the discrimination function [3], which indicates the limit of the methods ability to detect a deterministic signal in noise. This can be viewed as an analogue of the uncertainty principle in quantum mechanics or digital signal processing (the Kotelnikov theorem) [2]: even by increasing the number of features, it is impossible to solve the problem of signal detection in noise at a low signal-to-noise ratio. It follows that the key physical characteristics of a signal-noise mixture that are known or can be measured are the type and length of the time window, the sampling rate, the signal-to-noise ratio, the statistical properties of the noise, and the number of harmonics [4]. The remaining characteristics can be classified as local or integral information characteristics, which are functions of the input time series [5–8]. Therefore, it is necessary to formulate optimal rules for synthesizing an information system that solves the detection problem based on limited prior information [9, 10].

A distinctive feature of such optimal processing is the instability to variations in the statistical properties of noise and the prior uncertainty regarding the signal properties [11, 12]. It turns out that when solving the problem of detecting a deterministic signal under a low signal-to-interference ratio, the discriminatory power of the hypotheses serves as additional information [3, 13]. Overall, practically relevant problems include the detection of deterministic signals under a low signal-to-noise ratio and the discrimination between chaotic and random signals. Mathematical formulations of such problems reasonably rest on on additional assumptions concerning the properties of the signal and noise. In practice, however, many mathematical constructs prove to be either unimplementable or implementable only with significant errors.

The present work is devoted to establishing the properties of the detection problem and developing an approach to improve the quality of the solution. Several successive stages of formalizing and solving the detection problem will be considered, building upon known formulations of the binary hypothesis testing problem and the Neyman–Pearson lemma. Subsequently, it is necessary to identify the criteria, properties, and features of a deterministic signal, to apply optimal signal processing methods with the computation of probabilistic time, frequency, and time-frequency distributions, and, at the final stage, to construct statistics for model experiments.

2. NEUMANN-PEARSON LEMMAS, OPTIMAL TESTS, AND THE FORM OF THE ERROR FUNCTION

2.1. Detection Problem

The problem of detecting the signal $s(n)$ is traditionally reduced to the problem of distinguishing between two hypotheses

$$\begin{cases} \Gamma_0 : x(n) = w(n), \\ \Gamma_1 : x(n) = s(n) + w(n), \ n = 1, \dots, N. \end{cases}$$

Hypothesis Γ_0 corresponds to the decision that only noise is received, while hypothesis Γ_1 corresponds to the reception of a mixture of the signal of interest and noise, where the sequences $x(n)$, $n = 1, \dots, N$ represent a time series of the received data, $s(n)$ is the signal of interest, $w(n)$ is additive random noise, and N is the length of the data time series.

The random variables of the time series $(x(1),...,x(n),...,x(N))$ take values $(x_1,...,x_n,...,x_N)\in\mathbb{R}^N$.

2.2. Classical Neyman–Pearson Lemma

To obtain an analytical expression for estimating the probability of hypothesis discrimination error, one can apply a variant of the Neyman–Pearson lemma, which serves as a condition for optimality [14, 15].

Lemma 1 [Neumann–Pearson (classical)]. *Let there be an arbitrary measurable function of several variables $(x_1, \dots, x_N) \in \mathbb{R}^N$, called a decision rule or test, such that*

$$\varphi(x_1, \dots, x_N) = \begin{cases} 1, \ \textit{hypothesis } \Gamma_0 \textit{ is true}, \\ 0, \ \textit{hypothesis } \Gamma_1 \textit{ is true}, \end{cases}$$

by which we find

$$\begin{aligned} \alpha(\varphi) &= \textit{probability } (\textit{of accepting } \Gamma_1 | \Gamma_0 \textit{ is true}), \\ \beta(\varphi) &= \textit{probability } (\textit{of accepting } \Gamma_0 | \Gamma_1 \textit{ is true}). \end{aligned}$$

Then the decision rule φ^* is optimal if

$$1 - \beta(\varphi^*) = \sup_{\varphi}[1 - \beta(\varphi)] \ \textit{if} \ \alpha(\varphi^*) = \alpha_0, \tag{1}$$

where the supremum is taken over all possible tests.

Here, $\alpha(\cdot)$ is the false alarm probability, and $\beta(\cdot)$ is the probability of missing a useful signal.

2.3. Unconditional Neyman–Pearson Lemma

Lemma 2 [Neumann–Pearson (unconditional)]. *Let there be an arbitrary measurable function of several variables* $(x_1, \ldots, x_N) \in \mathbb{R}^N$*, called a decision rule or test, such that*

$$\varphi(x_1, \ldots, x_N) = \begin{cases} 1, & \text{hypothesis } \Gamma_0 \text{ is true}, \\ 0, & \text{hypothesis } \Gamma_1 \text{ is true}, \end{cases}$$

from which we find

$$\alpha(\varphi) = \text{probability (of accepting } \Gamma_1 | \Gamma_0 \text{ is true)},$$
$$\beta(\varphi) = \text{probability (of accepting } \Gamma_0 | \Gamma_1 \text{ is true)}.$$

Then the decision rule φ^* *is optimal if*

$$\alpha(\varphi^*) + \beta(\varphi^*) = \inf_{\varphi}[\alpha(\varphi) + \beta(\varphi)] = \mathcal{E}r(N; \Gamma_0, \Gamma_1) \text{ — error function}, \tag{2}$$

where the infimum is taken over all possible tests.

The peculiarity of applying this variant of the Neyman–Pearson lemma lies in the requirement to avoid errors when accepting one hypothesis in place of the other.

The following exact formula holds for the error function:

$$\mathcal{E}r(N; \Gamma_0, \Gamma_1) = 1 - \frac{1}{2}\|P_0^{(N)} - P_1^{(N)}\| = 1 - TV(P_0, P_1), \tag{3}$$

where $P_0^{(N)}$ is the multivariate distribution function of the observation statistics under hypothesis Γ_0, $P_1^{(N)}$ is the multivariate distribution function of the observation statistics under hypothesis Γ_1, and $TV(P_0, P_1)$ is the total variation of the signed measure. Here, $\|Q\| = 2\sup_A |Q(A)|$. Thus, if the support sets of the measures P_0 and P_1 do not intersect, then the hypotheses can be distinguished without error. If, however, the measures $P_0^{(N)}$ and $P_1^{(N)}$ are close, then $\|P_0^{(N)} - P_1^{(N)}\| \approx 0$, and then $\mathcal{E}r(N; \Gamma_0, \Gamma_1) \approx 1$.

Remark 1. For the problem of detecting a deterministic useful signal, for example, under low signal-to-noise ratio conditions, the case where $\|P_0^{(N)} - P_1^{(N)}\| = 2TV(P_0, P_1) \approx 0$ is of interest, as is the possibility of reasonably estimating this quantity. Therefore, when the probability of a total error in distinguishing between two hypotheses is close to one, it becomes possible to use the analytical expression $TV(P_0, P_1)$ to construct a criterion for the problem of detecting a useful signal in a signal-noise mixture.

Let us emphasize once again that in the classical Neyman–Pearson lemma, the probability of signal detection is maximized for a given false alarm probability, whereas in the unconditional Neyman–Pearson lemma, the total error of false alarms and missed signals is minimized. Due to the analytical representation of the error function, the unconditional lemma is preferable when working with low signal-to-noise ratios and unknown noise conditions. The following problem arises.

Problem 1. Determine the rules for selecting a criterion and constructing distributions P_0, P_1 that minimize the hypothesis discrimination error in a signal detection problem under low signal-to-noise ratio.

Here and in the following, we consider discrete probability distributions $P = \{p = (p_1, \ldots, p_i, \ldots, p_N)\}$, which, by definition, possess the following properties:

$$\forall\ p_i \in [0, 1], \quad \sum_{i=1}^{N} p_i = 1. \tag{4}$$

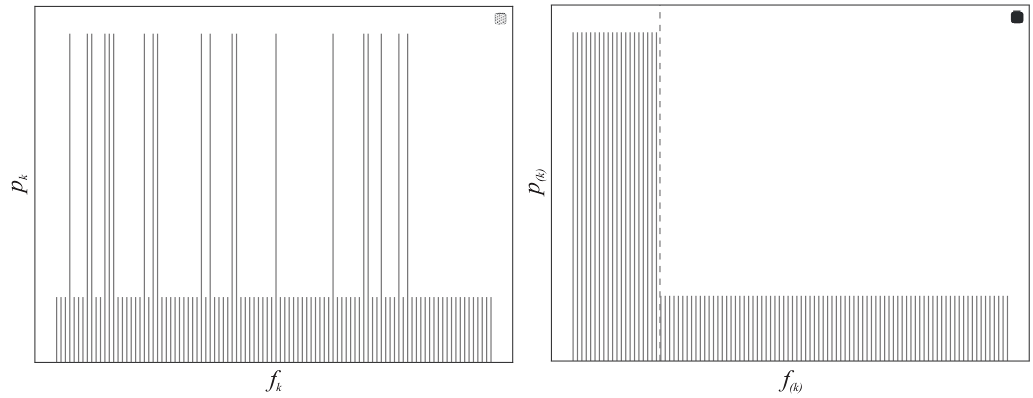

Fig. 1. The energy disordered and ordered spectra of a d-signal in uniform noise ($d = 20, N = 100$).

We now give the definition of a d-signal, which will be used in this work. To this end, we consider a distribution of the following form, where each probability p_k can take one of two distinct values:

$$\begin{cases} p_k = \dfrac{1 - p_{\max}}{N - d}, & \forall\, k = 1, \ldots, N - d, \\ p_k = p_N = \dfrac{p_{\max}}{d}, & \forall\, k = N - d + 1, \ldots, N. \end{cases} \tag{5}$$

Definition 1. A signal $s(t)$ for which the time, spectral, or frequency-time representations can be expressed as a distribution (5) is called a d-signal.

In [16], the following definition of a d-signal is given.

Definition 2. A signal consisting of d sinusoidal signals of equal amplitude will be referred to as a d-signal, and the individual frequencies of this signal will be referred to as harmonics. Then the spectrum of this signal will take the form

$$A_d = \sum_{i=1}^{d} A_0 \exp(-j2\pi f_i \Delta T), \tag{6}$$

where A_0 is the amplitude of the unit spectrum sample, f_i is the frequency of the ith sample (Hz), ΔT is the signal time interval in seconds.

An example of the spectrum and the ordered spectrum of a d-signal is shown in Fig. 1.

Remark 2. The frequencies of the d-harmonics f_i and the window size ΔT, where a window is defined as the time interval of the signal over which the discrete Fourier transform (DFT) is performed, are such that the following condition holds: $\forall f_i : f_i\Delta T = r_i$, $r_i \in \mathbb{N}$, $r_i < N/2$.

The meaning of the expression $r_i \in \mathbb{N}$ is that the integer number of periods r_i of the corresponding ith d-harmonic spectrum must fit within the time window ΔT of the signal-noise mixture.

Remark 3. Depending on the values of $p_{\max}$ and d, different d-signals are obtained.

3. OPTIMAL DISCRIMINATION OF TWO HYPOTHESES

It follows from the classical Neyman–Pearson lemma that the best method for distinguishing hypothesis is based on the likelihood ratio for discrete time of the following form:

$$l_N(x(t)) = \ln \frac{P_0(x(1), \ldots, x(N))}{P_1(x(1), \ldots, x(N))}, \tag{7}$$

where P_0 and P_1 are the joint distributions of the random process ξ_t, which takes values $x(1), \ldots, x(N)$ at times $t = \Delta t, \ldots, t = N\Delta t$. For a given time window width $T = N\Delta t$, there exists a strict functional relationship between α (false alarm probability) and β (probability of missing a useful signal) under optimal reception conditions

$$\alpha = \alpha(\beta, T) \text{ or } \beta = \beta(\alpha, T) \text{ or } T = T(\alpha, \beta). \tag{8}$$

It follows that $T(\alpha, \beta)$ represents the minimum observation time required to distinguish two hypotheses with given false alarm and missed detection rates [3].

Remark 4. Since the number of samples N can be increased in the discrete Fourier transform (DFT), each observation can be considered sufficient for investigating patterns

$$\alpha = \alpha(\beta) \text{ or } \beta = \beta(\alpha). \tag{9}$$

Definition 3. Let us define the mutual entropy of hypothesis Γ_0 relative to Γ_1 as the Kullback–Leibler divergence

$$H_{0/1}(P_0, P_1) = D_{KL}(P_0, P_1). \tag{10}$$

Proposition 1. *If the probability of error β is not too small, specifically $-\ln\beta \ll N$, then*

$$\beta(\alpha) = \exp(-H_{0/1}(P_0, P_1))(1 - \alpha) + o\left(\frac{\ln\beta}{N}\right). \tag{11}$$

Definition 4. We will define the discriminatory ability of hypotheses as the symmetrized Kullback–Leibler distance, also known as the information distance,

$$I_D = D_{KL}(P_0, P_1) + D_{KL}(P_1, P_0). \tag{12}$$

Discriminatory ability is used when the results of experiments under hypotheses Γ_0 and Γ_1 are close to each other, i.e., when the task is to detect a weak signal. In general, discriminatory ability depends on the signal-to-noise ratio. The following examples illustrate this feature.

Example 1. Let there be two hypotheses Γ_0 and Γ_1, and let the false alarm probability α and the missed detection probability β be calculated; then, provided the conditions of Proposition 1 are satisfied, it follows that

$$I_D = -\ln\frac{\beta}{(1-\alpha)} - \ln\frac{\alpha}{(1-\beta)} = \ln\frac{(1-\alpha)(1-\beta)}{\alpha\beta} \approx -(\ln\alpha + \ln\beta). \tag{13}$$

Example 2. Let there be two hypotheses Γ_0 and Γ_1 for distinguishing a deterministic signal from Gaussian noise with spectral density σ^2, and let us represent the deterministic signal as

$$s(t) = s_0 + \sum_{i=1}^{\infty} a_i \cos(\omega_i t + \varphi_i), \tag{14}$$

where $\{\omega_i\}$ and $\{\varphi_i\}$ are constants. Then

$$I_D = \frac{\int_0^T s^2(t)dt}{\sigma^2} = \frac{s_0^2}{2\pi\sigma^2} + \frac{1}{4\pi}\sum_{i=1}^{\infty}\frac{a_i^2}{\sigma^2}. \tag{15}$$

This means that in separating signal and noise, the discriminatory ability depends on the signal-to-noise ratio [3].

Example 3. Let there be two hypotheses Γ_0 and Γ_1 for distinguishing a deterministic d-signal from Gaussian noise with spectral density σ^2, where $d \ll N$, and the spectrum of this signal takes the form

$$A_d = \sum_{i=1}^{d} A_0 \exp(-j\omega_i + \varphi_i), \tag{16}$$

where A_0 is the amplitude of the unit spectral bin, ω_i is the frequency of the ith bin, and φ_i is the phase of the ith bin. Then, using spectral complexity and order statistics to construct the ordered spectral density of the given signal-to-noise ratio [17], the following is obtained

$$I_D \approx \ln\left(\frac{N}{d}\right)\left(\frac{\int_0^T s^2(t)dt}{\sigma^2}\right). \tag{17}$$

This means that the discriminatory ability depends on the number of signal harmonics and the signal-to-noise ratio.

The examples demonstrate the limits of hypothesis differentiation under given approaches to their comparison. Examples 1 and 2 show that the statistical and energy-based approaches yield the same results under otherwise identical conditions. Example 3 illustrates the difference in the ability to distinguish between hypotheses using the energy-based approach, provided that the spectral bandwidth of the signal is known at a given sampling rate. Example 1, adapted to the conditions of Example 3, establishes that under its conditions, the probability of signal rejection decreases for a given false alarm rate.

The functions of information divergences presented above can be unified under the general concept of *f-divergence* [18]:

$$D_f(p||q) = \sum_{x\in\mathbb{R}^N} q(x) f\left(\frac{p(x)}{q(x)}\right). \tag{18}$$

The choice of the function f gives rise to a whole family of different divergences:

- The Kullback–Leibler divergence $D_{KL}(p,q)$ is obtained from (18) by choosing $f(x) = x\log_2(x)$, $x > 0$.
- The Jensen–Shannon divergence is obtained from (18) by choosing

$$f(x) = x\log_2\frac{2x}{x+1} + \log_2\frac{2}{x+1},\ x\ >\ 0. \tag{19}$$

- The total variation is obtained for $f(x) = \frac{1}{2}|1-x|$:

$$TV(p,q) = \frac{1}{2}\sum_{x\in\mathbb{R}^N} |p(x) - q(x)|. \tag{20}$$

Remark 5. For discrete distributions of the form (4), the expression for the total variation of a measure denoted by TV, written as

$$TV(P_0, P_1) = \frac{1}{2}\sum_{i=1}^{N} |p_i^0 - p_i^1|. \tag{21}$$

Furthermore, $TV(p,q)$ is also a metric on the space of probability distributions and provides an upper bound on the Jensen–Shannon divergence:

$$JSD(p||q) \leq TV(p,q). \tag{22}$$

Each of the functions discussed in this section can be used to construct observation statistics during the secondary processing of a sequence of measurement sets, each containing $i = 1, \ldots, N$ samples.

To formalize criteria that take into account not only the random component of signals but also the deterministic component, let us introduce the concepts of the disequilibrium function D and the statistical complexity C of a distribution

$$C(p) = H(p)D(p,q),$$

where $H(p)$ is the entropy of the distribution p, and the disequilibrium $D(p,q)$ defines the distance between the signal and noise distributions. The simplest example of disequilibrium is the Euclidean distance in the space of discrete probability distributions [19], which is convenient to use when evaluating and comparing signals with a spectral distribution q that is close to uniform.

4. CONSTRUCTING THE DISTRIBUTIONS P_0 AND P_1

Let us describe the approaches to constructing the distributions P_0 and P_1 for a single observation. Three methods for constructing these distributions are proposed: based on the time series $\{x_i\}$, based on its discrete Fourier transform, and based on a frequency-time transform, specifically the classical or robust spectrogram.

4.1. Optimization of Frequency-Time Distributions and Types of Spectrograms

It turns out that the short-time Fourier transform can be obtained as the solution to an optimization problem [20].

Problem 2. It is necessary to find

$$F_s(t,\omega) = m^* = \arg\min_m I(t,\omega,m), \tag{23}$$

where

$$I(t,\omega,m) = \sum_{n=-N/2}^{N/2-1} \mathrm{w}(n\Delta t)\mathbf{F}(e(t,\omega,n,m)). \tag{24}$$

Here, $\mathbf{F}(e)$ is the loss function, $\mathrm{w}(n\Delta t)$ is the window function, Δt is the sampling interval, and the error function takes the form

$$e(t,\omega,n,m) = s(t+n\Delta t)\exp(-i\omega n\Delta t) - m.$$

The error function determines the residual value that expresses the similarity between the signal and the harmonic $\exp(i\omega n\Delta t)$. Different types of spectrograms can be obtained using various functions $\mathbf{F}(e)$.

Remark 6. If $\mathbf{F}(e) = |e|^2$, hen the necessary extremum conditions for the Problem 2

$$\frac{\partial I(t,\omega,m)}{\partial m} = 0$$

yield its solution in the form

$$F_s(t,\omega) = \frac{1}{a_{\mathrm{w}}} \sum_{n=-N/2}^{N/2-1} \mathrm{w}(n\Delta t) s(t+n\Delta t) \exp\left(-i\omega n\Delta t\right), \quad a_{\mathrm{w}} = \sum_{n=-N/2}^{N/2-1} \mathrm{w}(n\Delta t). \tag{25}$$

It follows that the short-time Fourier transform $S_t(\omega)$ of the signal $s(t)$ is equal to

$$S_t(\omega) = F_s(t,\omega),$$

and the corresponding spectrogram is given by

$$P_S(t,\omega) = |F_s(t,\omega)|^2. \tag{26}$$

Remark 7. Since (t,ω) are discrete variables in discrete transforms, we can define

$$P_0 = \frac{P_S(t,\omega)}{\sum_{t,\omega} P_S(t,\omega)}.$$

Remark 8. The maximum likelihood estimation method is applicable when the noise distribution density $p(e)$ is known. MLE assumes a loss function $\mathbf{F}(e) = -\ln p(e)$. Thus, the MLE estimate of signal spectra distorted by Gaussian noise $p(e) \sim \exp(-|e|^2)$, with the corresponding loss function $\mathbf{F}(e) = |e|^2$, is the short-time Fourier transform. Furthermore, in many cases, MLE estimates are highly sensitive to deviations from the parametric model and the assumed distribution. Even a small deviation from the assumption can lead to a significant deterioration of such an estimate.

Following the Remarks 6 and 8, a minimax robust approach [8] was developed in statistics as an alternative to the traditional MLE to reduce the sensitivity of MLE estimate. The loss function in this case takes the form $\mathbf{F}(e) = |e|$. The robust minimax short-time Fourier transform (M-STFT) is based on the solution of the nonlinear system of equations (27)–(29)

$$F_s(t,\omega) = \frac{1}{a_{\mathrm{w}}(t,\omega)} \sum_{n=-N/2}^{N/2-1} d(t,\omega,n) s(t+n\Delta t) \exp\left(-i\omega n\Delta t\right), \tag{27}$$

$$d(t,\omega,n) = \frac{\mathrm{w}(n\Delta t)}{|s(t+n\Delta t)\exp\left(-i\omega n\Delta t\right) - F_s(t,\omega)|}, \tag{28}$$

$$a_{\mathrm{w}}(t,\omega) = \sum_{n=-N/2}^{N/2-1} d(t,\omega,n). \tag{29}$$

Definition 5. Let $\mathbf{F}(e) = |e|$ and suppose that (27)–(29) are given in Problem 2; then the robust spectrogram is defined as

$$P_s(t,\omega) = I(t,\omega,0) - I(t,\omega,F_s(t,\omega)). \tag{30}$$

Note that for a quadratic loss function, the expressions (30) and (25) are identical.

Another way to construct a robust spectrogram is to construct a windowed discrete Wigner–Ville distribution. The distribution is defined as

$$W_s(t,\omega) = \frac{1}{a_{\mathrm{w}}} \sum_{n=-N/2}^{N/2} \mathrm{w}(n\Delta t) s(t+n\Delta t) s^*(t-n\Delta t) \exp\left(-i2\omega n\Delta t\right), \tag{31}$$

where

$$a_{\mathrm{w}} = \sum_{n=-N/2}^{N/2-1} \mathrm{w}(n\Delta t). \tag{32}$$

It can be found by solving the following optimization problem.

Problem 3. It is necessary to find

$$W_s(t,\omega) = m^* = \arg\min_m J(t,\omega,m), \tag{33}$$

where

$$J(t,\omega,m) = \sum_{n=-N/2}^{N/2-1} \mathrm{w}(n\Delta t)\mathbf{F}(|e|). \tag{34}$$

The error function is defined as follows:

$$e(t,\omega,n,m) = s(t+n\Delta t)s^*(t-n\Delta t)\exp\left(-i2\omega n\Delta t\right) - m,$$

and $\mathbf{F}(|e|) = |e|^2$.

4.2. Discrete Fourier Transform

Problem 4. Let $\{x_1,\ldots,x_{2N+2}\}$ be a realization of the sequence of independent random variables $\{\xi_1,\ldots,\xi_{2N+2}\}$ with zero expected value, to which the DFT is applied

$$X_k = \sum_{n=1}^{2N+2} x_n e^{-2\mathbf{i}\pi k(n-1)/(2N+2)}, \tag{35}$$

which defines the random variable

$$\Xi_k = \sum_{n=1}^{2N+2} \xi_n e^{-2\mathbf{i}\pi k(n-1)/(2N+2)}, \tag{36}$$

where $k = 0,\ldots,N$, since, by the symmetry property of the DFT of a real signal, the second half of the $N+1,\ldots,2N+1$ complex amplitudes of the spectral samples is complex conjugate to the first.

It is necessary to find the discrete probability function of the normalized ordered spectral distribution $n_k(N)$ as the normalized mean for each kth value of the random variable

$$\eta_k(N) = \frac{(\mathbf{T}I)_k}{E_X}, \tag{37}$$

where $I_k = \Xi_k\Xi_k^*$ (the square of the amplitude modulus or the energy of the spectral bin), E_X is half the signal energy, and $\mathbf{T}$ is the operator that orders the series in non-increasing order.

The solution of Problem 4 in the case of observing independent normally distributed random variables ξ_n, $n = 1,\ldots,2N+2$ with mean zero and variance σ_0^2 is given in [17] and [13]. The estimates of the probability function $\widetilde{n}_k(N)$ of the normalised ordered discrete spectrum, generated by the corresponding exponential random variable I_k, are as follows:

$$\widetilde{n}_k(N) = \frac{1}{N}\left(\sum_{i=1}^{N}\frac{1}{i} - \sum_{i=1}^{k-1}\frac{1}{i}\right) = \frac{1}{N}\sum_{i=k}^{N}\frac{1}{i}. \tag{38}$$

To obtain an approximate value of $n_k(N)$, let us choose the following formula

$$n_k(N) = -\frac{1}{NK_N}\ln\frac{2k-1}{2N+1}, \quad K_N = -\frac{1}{N}\sum_{k=1}^{N}\ln\frac{2k-1}{2N+1}. \tag{39}$$

Remark 9. Assuming that the normalised ordered statistics $(\eta_1,\ldots,\eta_N)$ take the values $P_0 = \{p = (\widetilde{n}_1,\ldots,\widetilde{n}_N)\}$ or $P_0 = \{p = (n_1,\ldots,n_N)\}$, and the observation statistics $P_1 = \{p = (p_1,\ldots,p_N)\}$, the spectral complexity can also be defined on the basis of the distribution (38) using the formula

$$C_{SS}(p) = -\frac{1}{4\log_2 N}\left(\sum_{k=1}^{N} p_k\log_2 p_k\right)\left(\sum_{k=1}^{N}|p_k - \widetilde{n}_k(N)|\right)^2 \tag{40}$$

or its approximate equivalent (39) using the formula

$$C_S(p) = -\frac{1}{4\log_2 N}\left(\sum_{k=1}^{N} p_k\log_2 p_k\right)\left(\sum_{k=1}^{N}|p_k - n_k(N)|\right)^2. \tag{41}$$

Now let us assume that $N = N_1N_2$, where N_1 is the number of samples in a single window and N_2 is the number of windows. Since the entropy in a single window does not depend on the order of the spectral samples, it can be defined as

$$H(n_k(N_1)) = -\frac{1}{\log_2 N_1}\sum_{k=1}^{N_1} n_k(N_1)\log_2 n_k(N_1).$$

The entropy in N_2 statistically identical windows is equal to

$$H(P_S(N_2,N_1)) = -\frac{1}{\log_2 N_1 + \log_2 N_2}\left(\log_2 N_2 + \sum_{k=1}^{N_1} p_k(N_1)\log_2 p_k(N_1)\right).$$

5. OPTIMISATION OF INFORMATION CRITERIA AND d-SIGNAL DETECTION

Now let us turn to some well-known information criteria for discriminating between hypotheses and establish their properties. Since Lemma 2 establishes that the discrimination error function for two hypotheses depends on the total variance $TV(p,q)$, one more concept of disequilibrium and statistical complexity based on it is introduced [10]. It is also shown there that such a criterion is one of the best for solving the problem of distinguishing between two hypotheses and indicating the presence of a deterministic component of a weak useful signal in white noise:

$$C_{TV}(p) = H(p)D_{TV}(p), \tag{42}$$

where

$$D_{TV}(p) = \frac{1}{4}\left(\sum_{i=1}^{N}\left|p_i - \frac{1}{N}\right|\right)^2, \tag{43}$$

$$H(p) = \frac{1}{\log_2 N}\left(-\sum_{i=1}^{N} p_i\log_2 p_i\right) \tag{44}$$

— normalised Shannon entropy [21]. When calculating the sum in (44), it is assumed that $\frac{0}{\log_2 0} = 0$ by continuity. It was also shown there that the maximum value of C_{TV} is attained on a distribution consisting of only two values, with an optimal number of samples for each of these values. The following holds

Lemma 3. *The maximum statistical complexity* (42) *is attained on the class of distributions* (5). It is necessary to determine the optimal values for d and $p_{\max}$. To do this, let us compute the value of the disequilibrium D_{TV} on the distribution (5):

$$D_{TV}(\omega, p_{\max}) = (p_{\max} - \omega)^2, \quad \omega = \frac{d}{N}. \tag{45}$$

In turn, the entropy is given by

$$H(\omega, p_{\max}) = 1 - \frac{1}{\log_2 N}\left((1-p_{\max})\log_2\frac{1-p_{\max}}{1-\omega} + p_{\max}\log_2\frac{p_{\max}}{\omega}\right). \tag{46}$$

The table shows the optimal values of the parameters for the formulas (45) and (46), which yield the maximum statistical complexity $C_{TV}(\omega, p_{\max})$.

Optimal parameters $C_{TV}(\omega, p_{\max})$ for various values of N

N	$C_{TV}(\omega^*, p^*_{\max})$	$p^*_{\max}$	$1-\omega^*$	d^*
3	0.1289	0.8241	0.6751	1 or 2
256	0.4789	0.9976	0.8752	32
512	0.5120	0.9991	0.8901	56
1024	0.5410	0.9997	0.9022	100
2048	0.5667	0.9999	0.9122	180

From Lemma 3 and the table, it is clear that d-signals are important for solving the detection problem. This leads to the following

Problem 5. Detection of a d-signal in a signal-noise mixture of length N, $N > d$.

To solve Problem 5 on d-signal detection, we apply the Renyi entropy H_γ with parameter γ ($\gamma > 0$, $\gamma \neq 1$), the detailed properties of which are presented in [22].

$$H_\gamma = \frac{1}{1-\gamma}\left(\log_2\sum_{k=1}^{N} p_k^\gamma\right). \tag{47}$$

It turns out that the Renyi entropy for a d-signal is calculated as

$$H_\gamma^{(d)} = \frac{1}{1-\gamma}\log_2\left(d\left(\frac{p_{\max}}{d}\right)^\gamma + (N-d)\left(\frac{1-p_{\max}}{N-d}\right)^\gamma\right). \tag{48}$$

For large N, it can be estimated as follows:

$$H_\gamma^{(d)} \approx \widetilde{H}_\gamma^{(d)} = \begin{cases} \dfrac{1}{1-\gamma}\log_2(d^{1-\gamma}p_{\max}^\gamma) = \log_2 d + \dfrac{\gamma}{1-\gamma}\log_2(p_{\max}), & \gamma > 1, \\ \log_2 N + \dfrac{\gamma}{1-\gamma}\log_2(1-p_{\max}), & 0 < \gamma < 1, \end{cases} \tag{49}$$

which shows an approximate linear relationship between the logarithms of of the number of harmonics in the signal, $p_{\max}$ and the Renyi entropy.

From the table and (49), for $\gamma > 1$ the following equality holds:

$$\widetilde{H}_\gamma^{(d^*)} = \log_2 d^*. \tag{50}$$

Similar relations are known for the frequency-time processing of known signals [20]. The Renyi entropy on white noise takes its maximum value $\widetilde{H}_\gamma^{(N)} = \log_2 N$.

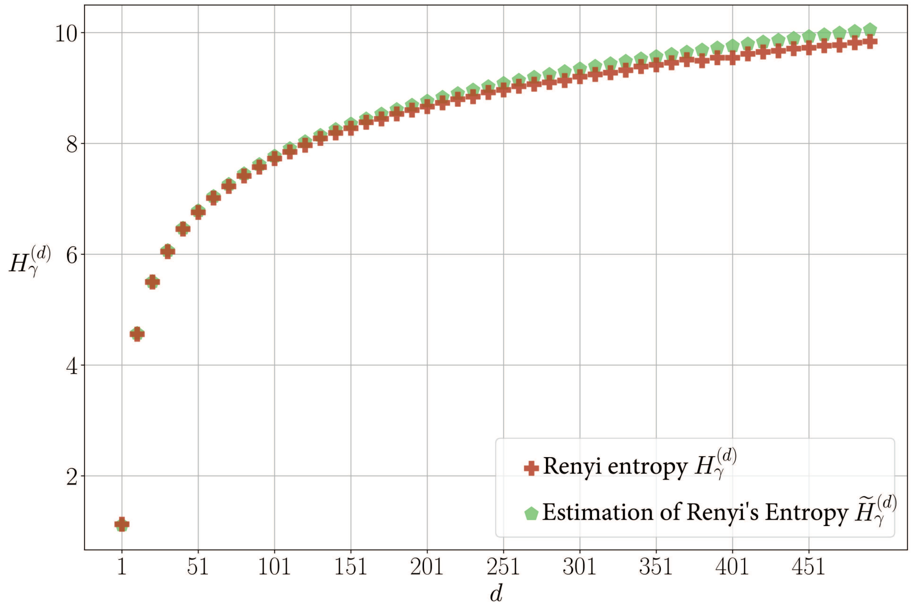

Fig. 2. A comparison of the Renyi entropy ($\gamma = 10$), calculated using formula (47), and its estimate (53).

Definition 6. Let us define the signal-to-noise ratio for a d-signal in the case where the system (5) corresponds to the power spectrum of a signal-noise mixture as:

$$SNR = 10\log_{10}\delta = 10\log_{10}\left(d\left(\frac{p_{\max}}{d}\right)\Big/(N-d)\frac{(1-p_{\max})}{(N-d)}\right) = 10\log_{10}\left(\frac{p_{\max}}{1-p_{\max}}\right). \quad (51)$$

In this case, $p_{\max}$ can be expressed in terms of SNR in order to be used in the calculation of Renyi entropy

$$p_{\max} = \frac{10^{\frac{SNR}{10}}}{1+10^{\frac{SNR}{10}}}. \quad (52)$$

Then $\widetilde{H}_\gamma^{(d)}$ can be written for $\gamma > 1$ according to (49) in the following form:

$$\widetilde{H}_\gamma^{(d)} = \log_2 d + \frac{\gamma}{1-\gamma}\log_2\left(\frac{10^{\frac{SNR}{10}}}{1+10^{\frac{SNR}{10}}}\right) = \log_2 d + \frac{\gamma}{1-\gamma}\left[\frac{SNR}{10}\log_2 10 - \log_2\left(1+10^{\frac{SNR}{10}}\right)\right]. \quad (53)$$

Figure 2 shows a comparison of the Renyi entropy, calculated using formula (47), and its estimate (53) as a function of the number d of harmonics in the generated signal, mixed with white noise, at $SNR = 0\ dB$. It can be seen that the estimate is close to the calculated values.

Corollary 1. *With a low signal-to-noise ratio, the Renyi entropy takes the form*

$$\widetilde{H}_\gamma^{(d)} \approx \log_2 d + \frac{\gamma\log_2 10}{10(1-\gamma)}SNR. \quad (54)$$

Thus, when the maximum Renyi entropy is known, the minimum SNR can be estimated as

$$SNR_{\min} = \frac{10(1-\gamma)}{\gamma\log_2 10}\left(\log_2 N - \log_2 d\right). \quad (55)$$

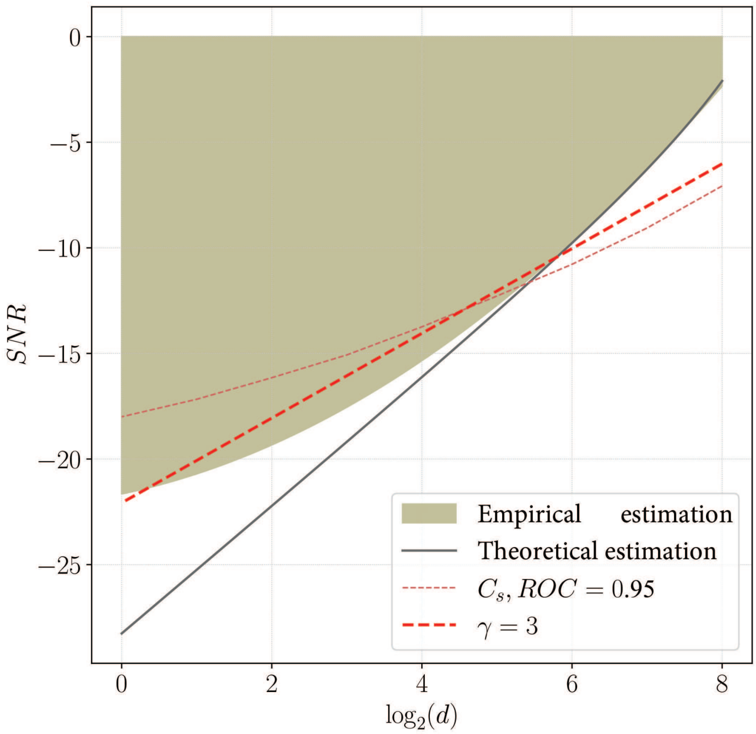

Fig. 3. Limit values of SNR for signal detection in white noise as a function of the number d ($N = 2048$).

According to [4], the one-to-one correspondence between the set H (entropy), C_{TV} (statistical complexity) and the set SNR (signal-to-noise ratio), $\log d$ (discreteness) makes it possible to use metric (analytical)[1] grids in signal detection and classification tasks in noise.

A comparative analysis of SNR estimates as a function of the number d of harmonics in the spectrum is shown in Fig. 3 ($N = 2048$). The shaded area illustrates the results of an empirical study of the limits of signal classification in white noise using a statistical complexity diagram. The solid black line represents the theoretical estimate obtained using the formula from [16]:

$$SNR > 10 \lg \left(\frac{3d}{N - 3d} \right). \tag{56}$$

The thin dotted line in Fig. 3 shows the limiting capabilities of signal detection using the C_s criterion (spectral complexity). The result of the dependence of the limiting SNR on the number of harmonics in the spectrum was obtained through a numerical experiment (ROC $= 0.95$)[2]. The thick dotted line represents the estimate according to formula (55) for $\gamma = 3$.

The formula (53) obtained for $\widetilde{H}_{\gamma}^{(d)}$ allows us to formulate the following lemma for the estimate $\hat{d}$ of the number of components of a d-signal.

[1] The concept of a metric grid is explained in detail in [4]. In the context of the complexity diagram, a metric grid is an approximate one-to-one mapping between a set of information characteristics of a signal-noise mixture, such as information entropy and statistical (spectral) complexity, and a set of characteristics such as discreteness (the number of harmonics in the d-signal spectrum) and the signal-to-noise ratio.

[2] The area under the ROC (Receiver Operating Characteristic) curve is the AUC. In practice, for machine learning model performance metrics, an ROC-AUC value of 0.5 indicates an inability to detect a signal in the noise based on a given criterion; a value between 0.7 and 0.8 is considered acceptable; between 0.8 and 0.9 is considered high; and a value above 0.9 is considered excellent.

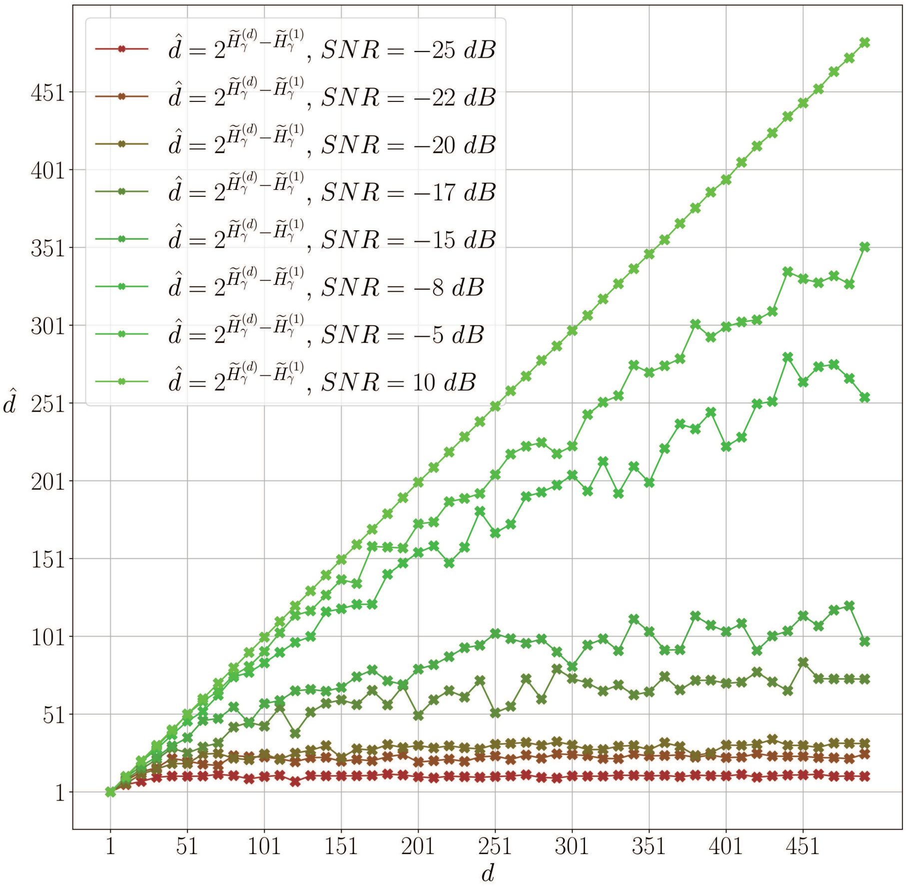

Fig. 4. Estimation of the number of components in a d-signal of length $N = 65536$ in white noise using Renyi entropy ($\gamma = 10$) for different SNR values.

Remark 10. The estimate $\hat{d}$ for the number of d-signal harmonics, using Renyi entropy, is written as follows:

$$\hat{d} = 2^{\widetilde{H}_\gamma^{(\hat{d})} - \widetilde{H}_\gamma^{(1)}}. \tag{57}$$

Proof. Let us write out equation (53) for $d = \hat{d}$ and $d = 1$:

$$\begin{cases} \log_2 \hat{d} = \widetilde{H}_\gamma^{(\hat{d})} - \dfrac{\gamma}{1-\gamma}\left[\dfrac{SNR}{10}\log_2 10 - \log_2\left(1 + 10^{\frac{SNR}{10}}\right)\right], \\ \log_2 1 = 0 = \widetilde{H}_\gamma^{(1)} - \dfrac{\gamma}{1-\gamma}\left[\dfrac{SNR}{10}\log_2 10 - \log_2\left(1 + 10^{\frac{SNR}{10}}\right)\right]. \end{cases} \tag{58}$$

By expressing in this way the term dependent on SNR, which is identical in both cases, in the system (58) via $\widetilde{H}_\gamma^{(1)}$, we obtain the statement of the lemma.

Figure 4 presents a numerical validation of the remark 10 in the form of a dependence of the component number estimate, calculated using formula (57), for generated signals consisting of d harmonic components mixed with white noise at various SNR values. It can be seen from Fig. 4 that the estimate is reasonably stable for up to 100 components, provided that the white noise level is up to $-10\ dB$.

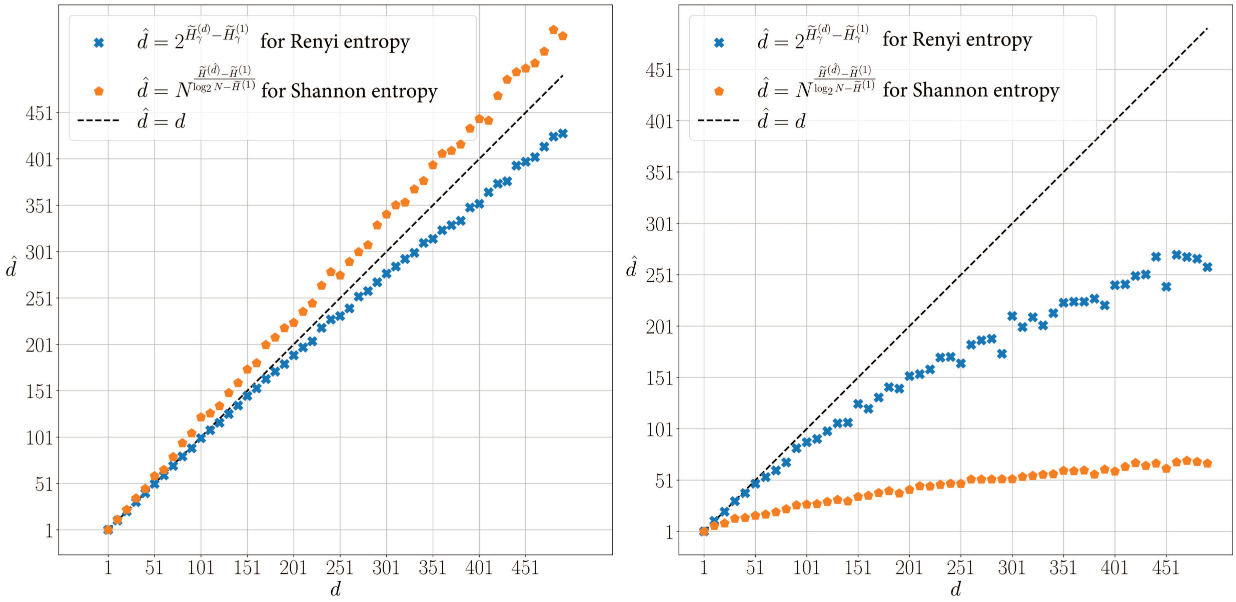

Fig. 5. Estimation of the number of components in a d-signal of length $N = 65536$ in white noise using Renyi entropy ($\gamma = 10$) and Shannon entropy for $SNR = 0\ dB$ and $SNR = -8\ dB$.

Further, let us estimate the number of components using unnormalised Shannon entropy. To do this, let us consider the expression $\widetilde{H}^{(d)}$, which estimates the Shannon entropy for a d-signal when N is large.

$$\begin{aligned} H^{(d)} &= -\left(d\frac{p_{\max}}{d}\log_2\frac{p_{\max}}{d} + (N-d)\frac{1-p_{\max}}{N-d}\log_2\frac{1-p_{\max}}{N-d}\right) \\ &= -\left(p_{\max}\log_2\frac{p_{\max}}{d} + (1-p_{\max})\log_2\frac{1-p_{\max}}{N-d}\right) \\ &= -\left(p_{\max}\log_2 p_{\max} + (1-p_{\max})\log_2(1-p_{\max})\right) + \left(p_{\max}\log_2 d + (1-p_{\max})\log_2(N-d)\right). \end{aligned} \tag{59}$$

The terms in the left-hand side are constrained by the constraint $0 < p_{\max} < 1$, so they will not appear in the expression for $\widetilde{H}^{(d)}$.

$$H^{(d)} \approx \widetilde{H}^{(d)} = p_{\max}\log_2 d + (1-p_{\max})\log_2 N. \tag{60}$$

To obtain the estimate $\hat{d}$, by analogy with the proof of remark 10, let us consider Shannon entropy for $d = \hat{d}$ and $d = 1$.

$$\widetilde{H}^{(1)} = (1-p_{\max})\log_2 N \Longrightarrow p_{\max} = 1 - \frac{\widetilde{H}^{(1)}}{\log_2 N}. \tag{61}$$

Let us use the expression $p_{\max}$ obtained for $\widetilde{H}^{(\hat{d})}$.

$$\widetilde{H}^{(\hat{d})} = \left(1 - \frac{\widetilde{H}^{(1)}}{\log_2 N}\right)\log_2 d + \log_2 N\frac{\widetilde{H}^{(1)}}{\log_2 N}. \tag{62}$$

Hence

$$\log_2\hat{d} = \frac{(\widetilde{H}^{(\hat{d})} - \widetilde{H}^{(1)})\log_2 N}{(\log_2 N - \widetilde{H}^{(1)})} \Longrightarrow \hat{d} = N^{\dfrac{\widetilde{H}^{(\hat{d})} - \widetilde{H}^{(1)}}{\log_2 N - \widetilde{H}^{(1)}}}. \tag{63}$$

Consequently, the following remark is valid.

Remark 11. The estimate $\hat{d}$ of the component number of a d-signal, using Shannon entropy, is expressed as follows:

$$\hat{d} = N^{\dfrac{\widetilde{H}^{(\hat{d})} - \widetilde{H}^{(1)}}{\log_2 N - \widetilde{H}^{(1)}}}. \tag{64}$$

Figure 5 shows a comparison of the performance of component number estimates based on the formulas (57) and (64), which are based on the Renyi and Shannon entropies respectively. It can be seen that the Renyi entropy provides a more robust estimate.

6. CONCLUSION

This paper establishes a sequence of optimisation sub-problems that need to be solved in order to detect a signal in noise, including cases with a low signal-to-noise ratio. In the general case, it is necessary to determine the optimal statistical test for the problem of distinguishing between two hypotheses, to establish optimal rules for assigning test statistic observations based on discrete spectral or frequency-time distributions, and to specify the form of the optimal discrimination criterion. Analytical constructions of criteria, the concept of d-signals, and information-based detection criteria form the basis for solving this binary classification problem: whether or not a signal is present. Further improvement in the quality of the solution to this problem is possible by refining the statistical properties of the noise and the signal in the noise.

FUNDING

This work was supported by the Russian Scientific Foundation, project no. 23-19-00134.

REFERENCES

1. Rihaczek, A., Signal energy distribution in time and frequency, *IEEE Transactions on Information Theory*, 1968, vol. 14, no. 3, pp. 369–374.
2. Cohen, L., *Time-Frequency Analysis*, New Jersey: Prentice Hall, 1995.
3. Dobrushin, R., Pinsker, M., and Shiryaev, A., An application of the concept of entropy to problems of detecting a signal in the noise, *Lithuanian Mathematical Journal*, 1963, vol. 3, no. 1, pp. 107–122.
4. Babikov, V. and Galyaev, A., Analytical representation of complexity diagrams, *Problems of Information Transmission*, 2025, vol. 61, no. 1, pp. 31–46.
5. Horie, T. et al., Sample entropy in electrocardiogram during atrial fibrillation, *Yonago Acta Med.*, 2018, vol. 61, no. 1, pp. 49–57.
6. Ribeiro, M. et al., The entropy universe, *Entropy*, 2021, vol. 23, no. 2, pp. 1–35.
7. Zunino, L., Soriano, M., and Rosso, O., Distinguishing chaotic and stochastic dynamics from time series by using a multiscale symbolic approach, *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, 2012, vol. 86, no. 4, pp. 1–10.
8. Shen, J., Hung, J., and Lee, L., Robust entropy-based endpoint detection for speech recognition in noisy environments, in *5th International Conference on Spoken Language Processing (ICSLP 1998)*, ISCA: ISCA, 1998, pp. 1–4.
9. Allen, R. and Mills, D., *Signal Analysis: Time, Frequency, Scale, and Structure*, New Jersey: John Wiley & Sons, 2004.
10. Berlin, L., Galyaev, A., and Lysenko, P., Comparison of information criteria for detection of useful signals in noisy environments, *Sensors*, 2023, vol. 23, no. 4, pp. 1–17.

11. Ramirez, J. et al., A new Kullback–Leibler VAD for speech recognition in noise, *IEEE Signal Process. Lett.*, 2004, vol. 11, no. 2, pp. 266–269.

12. Mehrotra, K., Mohan, C., and Huang, H., *Anomaly detection principles and algorithms*, Terrorism, Security, and Computation, Cham, Switzerland: Springer International Publishing, 2018.

13. Galyaev, A. et al., Order statistics of the normalized spectral distribution for detecting weak signals in white noise, *Autom. Remote Control.*, 2024, vol. 12, no. 1, pp. 1–16.

14. Shiryaev, A., *Veroyatnostno-statisticheskie metody v teorii prinyatiya reshenij*, (Probabilistic and Statistical Methods in Decision-Making Theory), M.: MCNMO: NMU, 2020.

15. Johnson, P., Moriarty, J., and Peskir, G., Detecting changes in real-time data: a user's guide to optimal detection, *Philos. Trans. A Math. Phys. Eng. Sci.*, 2017, vol. 375, no. 2100, pp. 1–16.

16. Babikov, V. and Galyaev, A., Information diagrams and their capabilities for classifying weak signals, *Problems of Information Transmission*, 2024, vol. 60, no. 2, pp. 127–140.

17. Galyaev, A. et al., A new spectral measure of complexity and its capabilities for detecting signals in noise, *Dokl. Math.*, 2024, vol. 110, no. 1, pp. 361–368.

18. Sason, I., On f-divergences: Integral representations, local behavior, and inequalities, *Entropy*, 2018, vol. 20, no. 5, pp. 1–32.

19. López-Ruiz, R., Shannon information, LMC complexity and rényi entropies: a straightforward approach, *Biophys. Chem.*, 2005, vol. 115, no. 2–3, pp. 215–218.

20. Boashash, B., *Time-Frequency Signal Analysis and Processing. A Comprehensive Reference*, London: Academic Press, 2016.

21. Shannon, C., A mathematical theory of communication, *Bell Syst. Tech. J.*, 1948, vol. 27, pp. 379–423.

22. Rényi, A., On measures of entropy and information, in *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, vol. 4, University of California Press, 1961, pp. 547–562.

This paper was recommended for publication by A.V. Nazin, a member of the Editorial Board

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2026, Vol. 87, No. 3, pp. 287–291.

15th IFAC Workshop on Adaptive and Learning Control Systems

DOI: 10.7868/S1608303226030083

From July 2 to 4, 2025, 15th IFAC Workshop on Adaptive and Learning Control Systems (ALCOS 2025) took place in Mexico City, Mexico. It was held under the auspices of Technical Committee 1.2 on Adaptive and Learning Systems of the International Federation of Automatic Control (IFAC) and was organized by the Faculty of Electrical Engineering and Electronics of the Autonomous Technological Institute of Mexico (Instituto Tecnologico Autonomo de Mexico, ITAM). The Mexican Association of Automatic Control (Asociacion de Mexico de Control Automatico, AMCA) was the technical sponsor of the event.

ALCOS is one of IFAC regular workshops, which aim is to exchange ideas and results in the field of adaptive and learning control systems that are used under conditions of parametric and non-parametric uncertainty, nonlinear dynamics, and external disturbances. The event was first held in San Francisco in 1983, and then in Lund, Sweden (1986), Glasgow, Scotland (1989), Grenoble, France (1992), Budapest, Hungary (1995), Glasgow, Scotland (1998), Cernobbio-Como, Italy (2001), Katayama, Japan (2004), St. Petersburg, Russia (2007) [1], Antalya, Turkey (2010), Caen, France (2013) [2], Eindhoven, Netherlands (2016), Southampton, United Kingdom (2019), Casablanca, Morocco (2022).

It should be noted that until 2019 the name of the event was the Adaptation and Learning in Control and Signal Processing (ALCOSP) workshop. Moreover, until 1998, the event was called a symposium (according to IFAC rules, an event is a symposium if the number of participants consistently exceeds 100). However, starting from 1998, the number of participants fell below 100. In 2001, Professor S. Bittanti, who organized ALCOSP in Italy, proposed to broaden its scope and hold the Periodic Control Systems (PSYCO 2001) workshop simultaneously with ALCOSP-2001. This indeed attracted new participants interested in the control of oscillations, including chaotic ones, in mechanics, biology, and other fields, and the total number of participants exceeded 100. Although in subsequent years the number of participants did not reach 100, it was convenient to hold these two events together, and the tradition continued. At the same time, combining two differently focused events at the same time and place sometimes caused confusion, and in early 2017, during an online meeting of Technical Committee 1.2, the issue of changing the event format and name was raised. In fact, there were practically no papers on signal processing at ALCOSP, as signal processing is a part of interest of another technical committee: Committee 1.1 "Identification and Signal Processing." After a brief discussion in January 2017, it was decided to change the seminar name to the current ALCOS. The field of oscillation control was decided to be implicitly included in the scope of the new workshop.

In its current format, the event was held in Southampton in 2019, in Casablanca in 2022, and in Mexico City in 2025.

Scientists from Russia took an active part in organizing and conducting ALCOS 2025. In particular, A.A. Bobtsov (ITMO University, St. Petersburg) was the co-chair of the conference, and the members of the international program committee were A.L. Fradkov (IPME RAS and SPbU, St. Petersburg), A.A. Pyrkin (ITMO University, St. Petersburg), A.I. Glushchenko (ICS RAS, Moscow), and six participants from Russia (A.L. Fradkov IPME RAS and SPbU, O.N. Granichin SPbU, A.A. Pyrkin, ITMO University, A.A. Vedyakov, ITMO University, A.A. Peregudin, ITMO University, A.I. Glushchenko, ICS RAS) were chairs or co-chairs of seven sessions.

Of the 53 manuscripts submitted to ALCOS in 2025, the international program committee selected 51 papers that were included in the final workshop program. Twenty-one associate editors and 236 independent reviewers were involved in the review process. Each submitted paper was evaluated by at least two reviewers. The final program included 3 plenary sessions, 1 invited session, and 10 regular sessions. Two parallel sessions were held daily.

About 60 participants from 13 countries (Russia, Mexico, France, India, USA, Brazil, Great Britain, Colombia, China, Morocco, Norway, Canada, Germany) took part in the sessions. The largest number of reports (20) were presented by scientists from Russia. Next in terms of the number of presentations were Mexico (9), India (5), and France (4). Russia was represented by scientists from St. Petersburg, Moscow, Krasnodar and Krasnogorsk.

The relatively small number of presentations is largely due to the particular topic of the workshop scope, in contrast to conferences on control theory in general, such as the European Control Conference (ECC), American Control Conference (ACC), Conference on Decision and Control (CDC), etc. At the same time, such narrow scope makes all sessions being interesting to attend, and only two parallel sections provide the opportunity to attend almost half of all presentations. Looking back over the past 20 years, we note that in 2010, 2016, and 2019, the number of presentations at ALCOS (ALCOSP) also did not exceed 60, in 2007 90 presentations were accepted, and only in 2013 and 2022 this number was above 100 presentations.

The workshop program included three plenary talks.

Professor A.L. Fradkov, representing the Institute for Problems in Mechanical Engineering of Russian Academy of Sciences and Saint Petersburg State University, devoted his presentation to the application of the speed gradient method for modeling, control, adaptation, and learning. He described the current state of the speed gradient method, developed mainly in the 1970s and 1980s, and provided brief information about speed gradient algorithms and their applicability, optimality, and passivity conditions. Applications of the speed gradient method to adaptive control and identification, nonlinear control, network and distributed system control were demonstrated, and examples of the method application in engineering, physical, biological, and environmental systems were provided. Recent modifications and extensions of the method were also mentioned, including non-Euclidean speed gradient algorithms based on Lyapunov–Bregman functions. The concluding part of the report demonstrated the applicability of the speed gradient method to develop models of natural phenomena. The report triggered a lively discussion, during which both classical questions about the connection between the speed gradient method and the second Lyapunov method for the design of adaptive controllers were discussed, as well as important points related to the future development of this method. For example, the possibility of its application to solve control problems in the presence of constraints. Such an extension would be important, in particular, for branches of control theory, in which the task of control design is formulated as an optimization problem.

Professor Joao Hespanha, representing the University of California, Santa Barbara, presented a report on reinforcement learning for large-scale games. The presentation discussed the use of reinforcement learning in zero-sum Markov games for two players with a finite but large state space, the goal of which is to find minimax strategies with "moderate" computations, *i.e.*, the task is to prove the optimality of strategies without exploring the entire state space of the game. An overview was given of the Q-learning method, which was proposed in the late 1980s as an alternative to the principle of dynamic programming for single-player game and zero-sum games for two players. The approach is based on reinforcement learning and does not require a known Markov model of the system. The report went on to show that most of the studies devoted to proving the correctness of Q-learning are based on establishing that each iteration converges to a single fixed point of a Bellman-type equation, which usually requires exploring the entire state space. The main part of the report demonstrated that for zero-sum games, mathematically correct optimal strategies can be designed using algorithms inspired by Q-learning, without the need for the Q-function to converge across the entire state space. In fact, the samples used to update the Q-function may not even explore the entire set of reachable states, and for certain classes of games, the proportion of explored states becomes smaller and smaller as the size of the state space increases.

Professor Zhong-Ping Jiang from New York University presented a report on the application of learning-based approaches for the design of adaptive optimal controllers. Learning-based control is a direct adaptive control method aimed at adaptation of optimal controllers based on real-time data. The report first demonstrated well-known developments in learning-based control for linear and nonlinear systems with unknown dynamics in continuous time, and then presented the latest results of stability analysis of continuous-time controllers designed on the basis of learning. Considering linear systems, it is possible to use parameterizations in the form of linear regression equations for the synthesis of an adaptive LQ controller, on the basis of which the laws of adaptation of the parameters of such a controller can be obtained, whereas, talking about nonlinear systems, the design of adaptive optimal controllers requires application of universal approximators, in particular — two neural networks in the controller structure (one — to approximate the value of the objective function, the other — to approximate the ideal controller). This leads to the locality of the obtained result in terms of stability, and in the case of disturbances — to the need to assume their smallness. In conclusion, the effectiveness of learning-based control was demonstrated.

The topics of ALCOS 2025 sessions were focused on: adaptive control and identification tasks for linear and nonlinear systems, adaptive observers, network systems control, data-driven control methods, and reinforcement learning. Particular attention was paid to bridge the theoretical developments with applications in such areas as robotics, aerospace, energy systems, and autonomous vehicles. As it was noted at the end of one of the sessions, all presentations were followed by very fruitful discussions, since almost everyone involved in the above-mentioned topics gathered at one place — at ALCOS sessions.

ALCOS 2025 program included only one special session, which was organized by professor O.N. Granicin and E.Yu. Tarasova (Saint Petersburg State University) and was devoted to problem of multi-agent systems control based on the application of the so-called intelligent approaches, including cases of dynamic network topology, noise and external disturbances, etc.

The following should be noted. Recently, not only at ALCOS, but also at ACC, ECC, and CDC, there has been a significant increase in the number of reports on data-driven control and reinforcement learning. Such reports always attract a lot of attention from participants, especially among young scientists. Perhaps this can be explained by the fact that data analysis, machine learning, deep neural networks, and other data processing tools are currently quite popular direction of students teaching. And, for graduates of such specialties, it is the data-driven control and reinforcement learning that seem to be the simplest and most understandable way to enter the control theory. However, everything has its price. For data-driven control, for linear systems, the requirement for a full-rank matrix composed of collected measurements actually means that the parameters of such a system can be estimated, which reduces the solution of the problem to the application of well-known methods of adaptive and robust control. For reinforcement learning, this price is the strict assumptions to prove the stability, and the locality of the result obtained. This should be kept in mind. This paragraph is debatable, reflects the authors' subjective opinion, and the authors are open to discuss the above by receiving feedback from readers.

We would like to highlight the presentation by professor Miroslav Krstic from the University of California, San Diego, USA. In his report, he proposed an alternative to the well-known sigma modification used in the adaptive laws to ensure: 1) bounded parametric error in the presence of external disturbances and non-parametric uncertainty, 2) convergence of the tracking error, albeit to a bounded set, but with a non-adjustable bound that is proportional to the value of the unknown ideal controller parameters, non-parametric uncertainty, and disturbances. Unlike known solutions, the method described in the report ensures required tracking error in the presence of arbitrary external disturbances and non-parametric uncertainty through the simultaneous application of dead zone, nonlinear damping, and parametric adaptation.

In the opinion of the authors of this note, the workshop venue was extremely well chosen. Mexico City is a very large metropolis with beautiful parks and fascinating architecture. The population of the Mexico City metropolitan area exceeds 21 million people. Thanks to the high-altitude subtropical climate, the air temperature remained moderate throughout the workshop, at around +22°C, which was conducive to calm discussion of scientific topics and communication between scientists. ITAM University, founded in 1946, where all the sessions were held, is a small (about 3 500 students) private university in Mexico City, which campus is very well organized and creates a cozy atmosphere. Of course, the seminar participants also had time to visit numerous attractions in Mexico City, such as Teotihuacan, Chapultepec Castle, Constitution Square, National Museum of Anthropology, Palace of Fine Arts, and many others.

We would like to express our gratitude to the workshop organizers, particularly professor Romeo Ortega, Chair of the International Program Committee, and members of the National Organizing Committee, professor Jose Guadalupe Romero, professor Rafael Cisneros, and professor Hugo Rodriguez. Our colleagues not only organized ALCOS 2025 at a very high level, but also held a seminar the day before its start entitled "Commemorating forty years of scientific collaboration between Russia and Mexico in the field of Control Theory," which participants, in addition to the colleagues already mentioned in this note, included professors L.M. Fridman (National Autonomous University of Mexico, UNAM), A.S. Poznyak (Center for Research and Advanced Studies of the National Polytechnic Institute, CINVESTAV-IPN), J. Moreno (National Autonomous University of Mexico, UNAM), I. Chairez (Monterrey Institute of Technology and Higher Education, ITESM), L. Aguilar (Instituto Politecnico Nacional, IPN), J. Davila (Instituto Politecnico Nacional, IPN), V.O. Nikiforov (ITMO University), and others. The presentations made by the seminar attendees focused both on scientific results and familiarizing the audience with the history of Russia-Mexico scientific cooperation in the field of control theory. Such seminars are very valuable for new generations of researchers, as knowledge of history allows them to set meaningful goals and choose scope for further research.

The venue for the next workshop will be chosen in 2026 at IFAC World Congress, which will be held in Busan, South Korea.

More detailed information can be found on the official website https://alcos2025.itam.mx/

REFERENCES

1. Andrievskii, B.R. and Fradkov, A.L., The 9th IFAC Workshop "Adaptation and Learning in Control and Signal Processing" (ALCOSP 2007) and 3rd IFAC Workshop "Periodic Control Systems" (PSYCO 2007), *Automation and Remote Control*, 2008, vol. 69, no. 4, pp. 733–736.

2. Bobtsov, A.A., Fradkov, A.L., and Zhiri, F., The 11th IFAC Workshop on Adaptation and Learning in Control and Signal Processing (ALCOSP 2013) and the 5th IFAC Workshop on Periodic Control Systems (PSYCO 2013), *Avtomatika i Telemekhanika*, 2014, no. 10, pp. 153–157. (In Russian)

A. A. Bobtsov (bobtsov@itmo.ru)
ITMO University, Saint Petersburg, Russia
A. I. Glushchenko (aiglush@ipu.ru)
Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia
K. A. Lastochkin (lastconst@ipu.ru)
Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia
R. Ortega (romeo.ortega@itam.mx)
Autonomous Technological Institute of Mexico, Mexico City, Mexico
A. L. Fradkov (fradkov@mail.ru)
Institute for Problems in Mechanical Engineering of Russian Academy of Sciences,
Saint Petersburg State University, Saint Petersburg, Russia