


Springer Proceedings in Mathematics & Statistics

Ilya Bychkov  
Valery A. Kalyagin  
Panos M. Pardalos  
Oleg Prokopyev *Editors*

# Network Algorithms, Data Mining, and Applications

NET, Moscow, Russia, May 2018

 Springer

**Springer Proceedings in Mathematics &  
Statistics**

Volume 315

## **Springer Proceedings in Mathematics & Statistics**

This book series features volumes composed of selected contributions from workshops and conferences in all areas of current research in mathematics and statistics, including operation research and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

More information about this series at <http://www.springer.com/series/10533>

Ilya Bychkov · Valery A. Kalyagin ·  
Panos M. Pardalos · Oleg Prokopyev  
Editors

# Network Algorithms, Data Mining, and Applications


NET, Moscow, Russia, May 2018

 Springer

*Editors*

Ilya Bychkov  
Higher School of Economics  
National Research University  
Nizhny Novgorod, Russia

Valery A. Kalyagin  
Higher School of Economics  
National Research University  
Nizhny Novgorod, Russia

Panos M. Pardalos   
Department of Industrial and Systems  
Engineering  
University of Florida  
Gainesville, FL, USA

Oleg Prokopyev  
Department of Industrial Engineering  
University of Pittsburgh  
Pittsburgh, PA, USA

ISSN 2194-1009                      ISSN 2194-1017 (electronic)  
Springer Proceedings in Mathematics & Statistics  
ISBN 978-3-030-37156-2              ISBN 978-3-030-37157-9 (eBook)  
<https://doi.org/10.1007/978-3-030-37157-9>

Mathematics Subject Classification (2010): 05C82, 90B10, 90B15, 90-02, 90C31, 90C27, 90C09, 90C10, 90C11, 90C35, 90B06, 90B18, 90B40, 68R01

© Springer Nature Switzerland AG 2020, corrected publication 2020

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Preface

This volume is based on the papers presented at the 8th International Conference on Network Analysis held in Moscow, Yandex office, Russia, May 18–19, 2018. The main focus of the conference and this volume is centered around the development of new network algorithms as well as underlying analysis and optimization of network structures generated by complex networks. Various applications to network data mining and social networks are also considered. The previous books based on the papers presented at the 1st–7th International Conferences on Network Analysis can be found in [1–7]. The current volume consists of three major parts, namely, Network Algorithms, Network Data Mining, and Network Applications, which we briefly overview next.

The first part of the book is focused on network algorithms. The chapter “[Fairness in Resource Allocation: Foundation and Applications](#)” presents a comprehensive review of fairness in resource allocation and its foundation including a complex network analysis. Fairness is applied when the resources divided on multiple demands are limited. Implementing fairness in resource allocation is a challenging task since fairness and efficiency are contradicting objectives. Hence, a variety of approaches from the literature are discussed in this paper.

In the chapter “[Mixed Integer Programming for Searching Maximum Quasi-Bicliques](#)”, the problem of finding the maximal quasi-bicliques in a bipartite graph (bigraph) is considered. Several models of mixed-integer programming (MIP) to search for a quasi-biclique are constructed and tested for working efficiency.

In the chapter “[Graph Clustering Via Intra-Cluster Density Maximization](#)”, the clustering problem is formulated as a combinatorial optimization problem. The main contribution is a novel problem formulation that maximizes intra-cluster density, a statistically meaningful quantity, which is designed to prevent common degeneracies, like “mega clusters”. Some numerical solution techniques are presented.

In the chapter “[Computational Complexity of SRIC and LRIC Indices](#)”, the computational complexity of short-range and long-range interaction centrality (SRIC and LRIC) is investigated. Several modes are proposed to decrease the

computational complexity of the indices. The runtime comparison of the sequential and parallel computation of the proposed models is also given.

The chapter “[A Survey on Variable Neighborhood Search Methods for Supply Network Inventory](#)” focuses on reverse logistics and closed-loop supply chain networks that have gained substantial interest in business and academia. The dynamic lot-sizing problem with product returns and recovery are reviewed and recent successful applications of Variable Neighborhood Search (VNS) for the efficient solution of such problems are presented.

The second part of the book presents several network data mining techniques. Chapter “[GSM: Inductive Learning on Dynamic Graph Embeddings](#)” studies the problem of learning graph embeddings for dynamic networks and the ability to generalize to unseen nodes called inductive learning. An improved model GSM based on GraphSAGE algorithm is introduced and the experiments on datasets CORA, Reddit, and HSEcite are conducted. The results show a good performance of the new model.

In the chapter “[Collaborator Recommender System](#)”, a recommender system for the scientists from the National Research University Higher School of Economics to help them find coauthors for their future work is designed and investigated.

The chapter “[User Preference Prediction in a Set of Photos Based on Neural Aggregation Network](#)” focuses on the problem of user interests’ classification in visual product recommender systems. A new two-stage procedure is proposed. It is shown that this procedure can capture the relationships between the product images purchased by the same user. Experiments on the Amazon product dataset confirm a good performance of the procedure.

In the chapter “[Network Structure and Scheme Analysis of the Russian Language Segment of Wikipedia](#)”, a network of the Russian-language segment of Wikipedia is created and an analysis of its structure is conducted.

In the chapter “[Indirect Influence Assessment in the Context of Retail Food Network](#)”, an application of long-range interaction centrality (LRIC) to the problem of the influence assessment in the global retail food network is considered where node-to-node influence is transformed into the influence index. The model is applied to the food trade network based on the World International Trade Solution database.

The chapter “[Facial Clustering in Video Data Using Deep Convolutional Neural Networks](#)” presents an automatic system that structures information in video surveillance systems based on the analysis of facial images. The cluster analysis in video data using face detection in each video frame and feature extraction with pre-trained deep convolutional neural networks is suggested. Different aggregation techniques to combine frame features into a single video descriptor are implemented to organize video data based on clustering techniques. An experimental study with YouTube Faces dataset shows high efficiency of the model.

The third part of the book is on applications of network analysis. In the chapter “[The Existence and Uniqueness Theorem for Initial-Boundary Value Problem of the Same Class of Integro-Differential PDEs](#)”, the second initial-boundary value problem for a class of nonlinear PDEs of the second order and an integral operator of a given form is considered. The existence and uniqueness theorem of the corresponding initial-boundary value problem is proved.

In the chapter “[Mapping of Politically Active Groups on Social Networks of Russian Regions \(On the Example of Karachay-Cherkessia Republic\)](#)”, social and political activity in online social networks are investigated. Clusters of political activity in social networks of some Russian regions are obtained by the author’s method of seed clustering, each cluster being analyzed by network methods.

In the chapter “[Social Mechanisms of the Subject Area Formation. The Case of “Digital Economy”](#)”, a wide range of texts about digital economy was analyzed, making it possible to show the thematic structure of this subject area. Central and peripheral concepts were identified to characterize theoretical core concepts and related topics clarifying the application of digital economy.

In the chapter “[Methodology for Measuring Polarization of Political Discourse: Case of Comparing Oppositional and Patriotic Discourse in Online Social Networks](#)”, speech markers and semantic concepts typical for patriotic and oppositional discourse in social networks are analyzed. An alternative method to tf-idf metric for specific text markers identification is proposed. The features of oppositional discourse in comparison with the patriotic discourse were formulated.

In the chapter “[Network Analysis Methodology of Policy Actors Identification and Power Evaluation \(The Case of the Unified State Exam Introduction in Russia\)](#)”, a new methodology for identifying policy actors for policy fields is proposed and investigated. The presented methodology is based on text parsing and mining and producing networks with analysis of the text processing results. The example of the Russian Unified State Exam is developed as the real case of policy formulation and implementation. The methodology was shown to have great potential for verifying the theories of policy studies and for a broader application in the areas where analysis of policy actors and their power, influence, and impact is needed.

We would like to take this opportunity to thank all the authors and referees for their efforts. This work is supported by the Laboratory of Algorithms and Technologies for Network Analysis (LATNA) of the National Research University Higher School of Economics.

Nizhny Novgorod, Russia  
Nizhny Novgorod, Russia  
Gainesville, FL, USA  
Pittsburgh, PA, USA

Ilya Bychkov  
Valery A. Kalyagin  
Panos M. Pardalos  
Oleg Prokopyev



## References

1. Goldengorin, B.I., Kalyagin, V.A., Pardalos, P.M. (eds.): Models, algorithms and technologies for network analysis. In: Proceedings of the First International Conference on Network Analysis. Springer Proceedings in Mathematics and Statistics, vol. 32. Springer, Cham (2013)
2. Goldengorin, B.I., Kalyagin, V.A., Pardalos, P.M. (eds.): Models, algorithms and technologies for network analysis. In: Proceedings of the Second International Conference on Network Analysis. Springer Proceedings in Mathematics and Statistics, vol. 59. Springer, Cham (2013)
3. Batsyn, M.V., Kalyagin, V.A., Pardalos, P.M. (eds.): Models, algorithms and technologies for network analysis. In: Proceedings of Third International Conference on Network Analysis. Springer Proceedings in Mathematics and Statistics, vol. 104. Springer, Cham (2014)
4. Kalyagin, V.A., Pardalos, P.M., Rassias, T.M. (eds.): Network models in economics and finance. In: Springer Optimization and Its Applications, vol. 100. Springer, Cham (2014)
5. Kalyagin, V.A., Koldanov, P. A., Pardalos, P.M. (eds.): Models, algorithms and technologies for network analysis. In: NET 2014, Nizhny Novgorod, Russia, May 2014. Springer Proceedings in Mathematics and Statistics, vol. 156. Springer, Cham (2016)
6. Kalyagin, V.A., Nikolaev, A.I., Pardalos, P.M., Prokopyev, O.A. (eds.): Models, algorithms and technologies for network analysis. In: NET 2016, Nizhny Novgorod, Russia, May 2016. Springer Proceedings in Mathematics and Statistics, vol. 197. Springer, Cham (2017)
7. Kalyagin V.A., Pardalos, P.M., Prokopyev, O.A., Utkina I.E. (eds.): Computational aspects and applications in large-scale networks. In: Springer Proceedings in Mathematics & Statistics, vol. 247. Springer International Publishing AG, part of Springer Nature (2018)

# Contents

## Network Algorithms

<b>Fairness in Resource Allocation: Foundation and Applications</b> . . . . .	3
Hamoud S. Bin-Obaid and Theodore B. Trafalis	
<b>Mixed Integer Programming for Searching Maximum Quasi-Bicliques</b> . . . . .	19
Dmitry I. Ignatov, Polina Ivanova and Albina Zamaletdinova	
<b>Graph Clustering Via Intra-Cluster Density Maximization</b> . . . . .	37
Pierre Miasnikof, Leonidas Pitsoulis, Anthony J. Bonner, Yuri Lawryshyn and Panos M. Pardalos	
<b>Computational Complexity of SRIC and LRIC Indices</b> . . . . .	49
Sergey Shvydun	
<b>A Survey on Variable Neighborhood Search Methods for Supply Network Inventory</b> . . . . .	71
Angelo Sifaleras and Ioannis Konstantaras	

## Network Data Mining

<b>GSM: Inductive Learning on Dynamic Graph Embeddings</b> . . . . .	85
Marina Ananyeva, Ilya Makarov and Mikhail Pendiukhov	
<b>Collaborator Recommender System</b> . . . . .	101
Anna Averchenkova, Alina Akhmetzyanova, Konstantin Sudarikov, Stanislav Petrov, Ilya Makarov, Mikhail Pendiukhov and Leonid E. Zhukov	
<b>User Preference Prediction in a Set of Photos Based on Neural Aggregation Network</b> . . . . .	121
Kirill V. Demochkin and Andrey V. Savchenko	

<b>Network Structure and Scheme Analysis of the Russian Language Segment of Wikipedia</b> .....	129
Sergey Makrushin	
<b>Indirect Influence Assessment in the Context of Retail Food Network</b> .....	143
Fuad Aleskerov, Natalia Meshcheryakova and Sergey Shvydun	
<b>Facial Clustering in Video Data Using Deep Convolutional Neural Networks</b> .....	161
Anastasiia D. Sokolova and Andrey V. Savchenko	
<b>Network Applications</b>	
<b>The Existence and Uniqueness Theorem for Initial-Boundary Value Problem of the Same Class of Integro-Differential PDEs</b> .....	173
A. I. Egamov	
<b>Mapping of Politically Active Groups on Social Networks of Russian Regions (On the Example of Karachay-Cherkessia Republic)</b> .....	187
Galina Gradoselskaya, Ilia Karpov and Tamara Shcheglova	
<b>Social Mechanisms of the Subject Area Formation. The Case of “Digital Economy”</b> .....	201
Oxana Mikhailova, Galina Gradoselskaya and Alexander Kharlamov	
<b>Methodology for Measuring Polarization of Political Discourse: Case of Comparing Oppositional and Patriotic Discourse in Online Social Networks</b> .....	219
Tamara Shcheglova, Galina Gradoselskaya and Ilia Karpov	
<b>Network Analysis Methodology of Policy Actors Identification and Power Evaluation (The Case of the Unified State Exam Introduction in Russia)</b> .....	231
Dmitry Zaytsev, Gregory Khvatsky, Nikita Talovsky and Valentina Kuskova	
<b>Correction to: User Preference Prediction in a Set of Photos Based on Neural Aggregation Network</b> .....	C1
Kirill V. Demochkin and Andrey V. Savchenko	

# Contributors

**Alina Akhmetzyanova** National Research University Higher School of Economics, Moscow, Russian Federation

**Fuad Aleskerov** National Research University Higher School of Economics, Moscow, Russian Federation;  
V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, Moscow, Russian Federation

**Marina Ananyeva** National Research University Higher School of Economics, Moscow, Russian Federation

**Anna Averchenkova** National Research University Higher School of Economics, Moscow, Russian Federation

**Hamoud S. Bin-Obaid** University of Oklahoma, Norman, OK, USA;  
King Saud University, Riyadh, Saudi Arabia

**Anthony J. Bonner** University of Toronto, Toronto, ON, Canada

**Kirill V. Demochkin** Laboratory of Algorithms and Technologies for Network Analysis, National Research University Higher School of Economics, Nizhny Novgorod, Russian Federation

**A. I. Egamov** Lobachevsky State University of Nizhny Novgorod, Nizhny Novgorod, Russian Federation

**Galina Gradoselskaya** National Research University Higher School of Economics, Moscow, Russian Federation;  
Higher School of Economics, Moscow, Russian Federation

**Dmitry I. Ignatov** National Research University Higher School of Economics, Moscow, Russian Federation;  
St. Petersburg Department of Steklov Mathematical Institute of Russian Academy of Sciences, Saint Petersburg, Russian Federation

**Polina Ivanova** National Research University Higher School of Economics, Moscow, Russian Federation

**Ilia Karpov** National Research University Higher School of Economics, Moscow, Russian Federation

**Alexander Kharlamov** Higher School of Economics, Moscow, Russian Federation;  
Institute of Higher Nervous Activity and Neurophysiology of RAS, Moscow, Russian Federation;  
Moscow State Linguistic University, Moscow, Russian Federation

**Gregory Khvatsky** National Research University Higher School of Economics, Moscow, Russian Federation

**Ioannis Konstantaras** Department of Business Administration, School of Business Administration, University of Macedonia, Thessaloniki, Greece

**Valentina Kuskova** National Research University Higher School of Economics, Moscow, Russian Federation

**Yuri Lawryshyn** University of Toronto, Toronto, ON, Canada

**Ilya Makarov** National Research University Higher School of Economics, Moscow, Russian Federation;  
University of Ljubljana, Faculty of Computer and Information Science, Ljubljana, Slovenia

**Sergey Makrushin** Financial University under the Government of the Russian Federation, Moscow, Russian Federation

**Natalia Meshcheryakova** National Research University Higher School of Economics, Moscow, Russian Federation;  
V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, Moscow, Russian Federation

**Pierre Miasnikof** University of Toronto, Toronto, ON, Canada

**Oxana Mikhailova** Higher School of Economics, Moscow, Russian Federation

**Panos M. Pardalos** University of Florida, Gainesville, FL, USA;  
National Research University HSE, Nizhny Novgorod, Russian Federation

**Mikhail Pendiukhov** Analytical Software Solutions LLC, Moscow, Russian Federation

**Stanislav Petrov** National Research University Higher School of Economics, Moscow, Russian Federation

**Leonidas Pitsoulis** Aristotle University of Thessaloniki, Thessaloniki, Greece

**Andrey V. Savchenko** Laboratory of Algorithms and Technologies for Network Analysis, National Research University Higher School of Economics, Nizhny Novgorod, Russian Federation

**Tamara Shcheglova** National Research University Higher School of Economics, Moscow, Russian Federation

**Sergey Shvydun** National Research University Higher School of Economics, Moscow, Russian Federation;  
V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, Moscow, Russian Federation

**Angelo Sifaleras** Department of Applied Informatics, School of Information Sciences, University of Macedonia, Thessaloniki, Greece

**Anastasiia D. Sokolova** Laboratory of Algorithms and Technologies for Network Analysis, National Research University Higher School of Economics, Nizhny Novgorod, Russian Federation

**Konstantin Sudarikov** National Research University Higher School of Economics, Moscow, Russian Federation

**Nikita Talovsky** National Research University Higher School of Economics, Moscow, Russian Federation

**Theodore B. Trafalis** University of Oklahoma, Norman, OK, USA

**Albina Zamaletdinova** National Research University Higher School of Economics, Moscow, Russian Federation

**Dmitry Zaytsev** National Research University Higher School of Economics, Moscow, Russian Federation

**Leonid E. Zhukov** National Research University Higher School of Economics, Moscow, Russian Federation

# Network Algorithms

# Fairness in Resource Allocation: Foundation and Applications



Hamoud S. Bin-Obaid and Theodore B. Trafalis

**Abstract** This paper presents a comprehensive review of fairness in resource allocation and its foundation. Fairness is applied when the resources divided on multiple demands are limited. Implementing fairness in resource allocation is a challenging task since fairness and efficiency are contradicting objectives. A variety of approaches to find fair resource allocation from the literature are discussed such as max-min fairness, lexicographic ordering, proportional fairness in addition to some fairness measures. Both strength points and drawbacks for each approach are illustrated, and some connections among the approaches are elaborated. Examples of applications where fairness is applied are reviewed.

## 1 Introduction

Fairness has been gaining great interest in the past few decades. The decision-maker (DM) objective is to maximize efficiency, but if fairness is not considered, the service receivers are not satisfied and claim that the distribution of resources is unfair or unjust. If the objective is to distribute resources fairly, the DM is not satisfied if the resources are not utilized. A balanced solution between fairness and efficiency is the goal of fair resource allocation. Fairness in its early development was applied in the microeconomics of social welfare. Every individual or demand is assigned a utility function  $u$  based on the preference of the individual assuming that the DM is aware of the preferences of each individual. The set of resources  $X$  is to be distributed among individuals. Given that  $x \in X$  is a feasible allocation of resources among the

---

H. S. Bin-Obaid (✉) · T. B. Trafalis  
University of Oklahoma, 202 W. Boyd St., Lab 28, Norman, OK 73071, USA  
e-mail: [hsbinobaid@ou.edu](mailto:hsbinobaid@ou.edu)

T. B. Trafalis  
e-mail: [ttrafal@ou.edu](mailto:ttrafal@ou.edu)

H. S. Bin-Obaid  
King Saud University, 800, Riyadh 11421, Saudi Arabia

© Springer Nature Switzerland AG 2020  
I. Bychkov et al. (eds.), *Network Algorithms, Data Mining,  
and Applications*, Springer Proceedings in Mathematics & Statistics 315,  
[https://doi.org/10.1007/978-3-030-37157-9\\_1](https://doi.org/10.1007/978-3-030-37157-9_1)



individuals are chosen by the DM, then  $f_i(x)$  is the utility of individual  $i$  for every  $i = 1, \dots, n$ . This leads to the utility set  $U$  for all individuals:

$$U = \{u_i = f_i(x), \forall i = 1, \dots, n\} \quad (1)$$

Fair distribution of resources was initiated by observing and measuring the differences in the level of income of individuals in a society or a country by the statistician. Gini [26] developed the measure Gini index or Gini coefficient to measure income inequality, then he discussed the relationship between the Gini index and the Lorenz curve in [19]. Later in time, fair division of resources was introduced by Steinhaus [40] through the Cake Cutting Problem, where resources are distributed fairly or what is called an envy-free division. Fairness has gained great interest, and many approaches have been developed to distribute resources fairly. The scope of this paper is to review fair resource allocation approaches in general and more specifically in networks.

The paper is organized as follows. In Sect. 1.1, fairness in communication networks is discussed. In Sect. 1.2, fairness in facility location is reviewed. In Sect. 1.3, fairness in evacuation and traffic management is discussed. In Sect. 1.4, fairness in air traffic control is reviewed.

In Sect. 2, the early development of fairness in resource distribution is discussed. In Sect. 3, some measures of fairness are explored. Max-min fairness and lexicographic ordering are discussed in Sects. 4 and 5, respectively, elaborating on the relationship between them. Then, proportional fairness and  $\alpha$ -fairness are presented in Sects. 6 and 7, respectively, discussing the connection between them. In Sect. 8, the price of fairness is reviewed. Finally, the paper is summed up with concluding remarks in Sect. 9.

## 1.1 Fairness in Communication Networks

Recently, more than 90% of the literature of fair resource allocation has been extensively applied specifically in communication networks and in networks in general. Megiddo [31] has introduced fairness to networks. He finds the optimal solution to the fair and maximum flows using the lexicographical ordering of the individual flows for sources and sinks. Nace and Pióro [33] discuss the max-min fairness approach and its variations to fairly distribute bandwidth among a set of demands in communication networks. Max-min fair bandwidth allocation is studied in different network structures such as multi-channel wireless mesh networks, wireless multihop networks, cellular networks, and packet switches and routers by Tang et al. [41], Thulasiraman et al. [43], Boche et al. [8], and Pan and Yang [34], respectively. When the bandwidth demand is very high, the network or routers become congested. Mahajan et al. [29] and Siu and [38] explore congestion in routers and asynchronous transfer mode (ATM) networks, respectively.

Another application is in communication networks, where fairness in resource allocation is applied in the area of wireless sensor networks. Wireless sensors are placed in remote and hard to reach areas to sense the environment such as temperature, wind speed, humidity, etc. Then, the data are sent to the server through the shortest path to optimize the energy. The source of energy for these sensors is from solar panels or wind propellers. Some sensors sense and send the information while other sensors sense information, receive information from other sensors, and send the information to other sensors or the server, and these processes require energy. The objective is to maximize the utilization and fairly sense the information from all the sensors in the network. Sridharan and Krishnamachari [39] use max-min fairness to maximize the utilization and fairly collect information from all sensors in the network. Hsu et al. [22] claim that it is inefficient to use max-min fairness for underwater sensor networks (UWSN) and propose mixed-integer linear programming (MILP) model to find a max-min fair solution.

## ***1.2 Fairness in Facility Location***

Fairness in resource allocation is applied to other applications in networks. Fairness is applied in facility location and location allocation problems. When placing public facilities such as schools, libraries, or outpatient clinics, minimizing the distance from the facility to the service receivers is the main objective. However, minimizing the total distances may result in placing the facility very far from some service receivers and very close to others which may drive some service receivers to claim that it is unfair or unjust. Beheshtifar and Alimoahmadi [3] develop a model with multiple objectives to determine optimal sites for new clinics. Two of the objectives are minimizing the total distance to reach the clinics and minimizing the inequity to access the clinics. Buzna et al. [11] propose an approach to solve the facility location problem using a lexicographic minimax objective to find an equitable and efficient solution.

However, when placing fire stations, police departments, or ambulance stations, modeling the problem can be slightly different. The problem becomes a set covering problem since the response time is a vital factor in the rescue process. Goldberg [20] develops a mixed integer programming (MIP) model that determines the optimal location and dispatching process of ambulances. The model maximizes the number of potential patient-covered locations (within nine minutes distance) and fairly allocates the patients to the ambulance locations by distributing the demand fairly among the ambulance locations. This leads us to conclude that fairness can be applied not only to service receivers but also to service providers. Another example of locating facilities fairly is by Erkut et al. [15]. They introduced a multicriteria facility location model for solid waste, and one of the objectives is to locate the facilities fairly.

### ***1.3 Fairness in Evacuation and Traffic Management***

Fairness has been applied to evacuation processes and traffic management. The objectives of fair evacuation models are to minimize the total evacuation time and fairly allocate evacuees to the shortest route with minimum travel time. The delay experienced by the evacuees due to congestion is a main factor in the evacuation process. A very common approach to model an evacuation model is the cell transmission model (CTM) developed by Daganzo [14]. The objective in CTM is to minimize the network clearance time (NCT), or system optimal (SO), neglecting fairness in routing the evacuees, which may cause high congestion in some parts of the network and lead to further losses in lives and properties. User equilibrium (UE) or Wardrop's rule by [45] is the optimal fair solution that schedules travelers on the shortest routes such that no traveler can improve their travel time by changing routes resulting in a fair distribution of resources and minimum congestion over the network.

Other objectives are tested by Bish et al. [7] in the evacuation model. Minimizing the average travel time (ATT) and the average evacuation time (AET) of evacuees results in different optimal solutions. Minimizing the ATT is ideal in normal traffic congestion control while minimizing the AET is suitable for emergency evacuation. Moreover, Bayram et al. [2] introduce a model to compromise system efficiency and users' interest in shelter location during evacuation which is an application considering evacuation and facility location.

### ***1.4 Fairness in Air Traffic Control (ATC)***

The demand for air transportation has been increasing significantly leading to air traffic congestion and more challenging scheduling tasks. The scheduling process starts with strategic planning where flights are assigned takeoff and landing slots. Before execution, tactical planning takes place due to uncontrolled delays caused by inclement weather or technical issues. These delays incur a cost in billions of dollars. Decision-making tools are being developed to improve the scheduling process. The majority of the optimization models fail in considering the distribution of delay equally among the airline carriers. As a result, minimizing the total delay minutes (or total cost) is an efficient solution but it remains unimplemented due to the lack of fairness. Jonker et al. [25] have proposed a MIP model imposing fairness using real-world data spanning across 6 days. Since fairness and efficiency are contradicting objectives, they achieved fairness compromising less than 10% of the total cost. See Jonker et al. [25] for more details.

## 1.5 Fairness in Job Scheduling

Job scheduling on parallel processors has been extensively studied in the past few decades. Job scheduling can become very challenging as the number of jobs increases since they are scheduled over time and space. In addition, jobs are scheduled on multiple threads which lead to an extremely large number of combinations. As a result, the model complexity is NP-complete. Hence, approximation models and heuristics have been introduced in the literature. See Feitelson and Rudolph [16] and Schwiegelshohn and Yahyapour [37] for more details. Due to the high demand of different sizes of job processing, the fairness issue has arisen. Algorithms and heuristics are introduced for fair job scheduling. See Wang et al. [44] and Zaharia et al. [46]. Another application of fair scheduling is applied in heterogeneous vehicular networks by Zhang et al. [47]. They used a max-min fair scheduling approach to maximize the mobile service amount. See Thawari et al. [42] and Zhang [47] for more details.

## 2 Fairness Early Development

The classical fair division of an object between two partners is by letting one partner halve the object and the other to choose its half. The first partner is satisfied by being allowed to split the object into two halves, and the other is pleased by being given the freedom to choose one of the two halves. This problem is known by the cake cutting problem or envy-free division, and an example of a divisible object is a land (pastures or fields). Using this method becomes complicated if the number of partners is more than three. Steinhaus [40] has proposed an approach to solve this problem for  $n$  partners proposed by Knaster and Banach. Every partner has the right to cut a part of the cake until the last one cuts the last piece he touched, then he is eliminated. The remaining  $n - 1$  repeat the same procedure until two partners are left, then the classical rule is applied. Examples of divisible resources are salary, bonus, performance incentives, and severance pay.

Steinhaus then proposed an approach to fairly divide indivisible objects such as houses, animals, cars, etc., among  $n$  partners. Every partner estimates the value of each object, then every object value is decided by the highest estimated value and attributed to the partner who estimated it. Then, partners who received the higher value objects compensate the partners with the lower value objects to reduce the differences to zero. See Brams and Taylor [10] and Golovin [21] for more details.

Fairness has been applied in the microeconomic theory of social welfare [30]. Every consumer is assigned a utility function based on their preferences of the desired commodities, then all utility functions of all consumers are aggregated into a social welfare function. The objective is to maximize the social welfare function with respect to fairness and justice. A way to fairly maximize the utility functions of all consumers is to maximize the worst-off utility function or to rank alternatives based

on their worst outcomes given their probabilities. This approach is called maximin or Rawlsian social welfare function by Rawls [36], but it is not Pareto optimal in most cases. Another approach is the leximin ordering of social welfare function by using lexicographical ordering by Chen [12], and it provides a Pareto optimal solution for the optimal leximin social welfare function. Lexicographical ordering will be discussed in detail in a later section. Finally, max-min fairness is the most widely accepted approach since it fairly distributes resources and maximizes the utilization of the system. The relationship between leximin maximum and max-min fair solution will be discussed in detail in Sect. 4.

### 3 Fairness Measures

In this section, some measures are used to determine the variability in the distributed resources among the set of demands. However, some of these measures can be employed as objectives.

#### 3.1 Basic Fairness Measures

To measure equity or fairness, simple statistical measures such as the mean, standard deviation, and variance can give an excellent indication of how the resource is distributed on individual demands. The resource can be bandwidth in communication networks, distance in facility location, or processing time in job scheduling. In addition, these measures can be set as objectives in mathematical programming. Some of the basic statistical measures discussed in Leclerc et al. [28] are the maximum of absolute deviations (MAD), sum-of-squared deviations (SSD), and the sum of the absolute deviations (SAD). Note that  $u_i$  is the resource assigned to demand  $i$  given that the number of demands is  $n$ , and  $\bar{u}$  is the average resource of all the  $n$  demands.

$$MAD = \max_{i=1,2,\dots,n} |u_i - \bar{u}| \quad (2)$$

$$SSD = \sum_{i=1}^n (u_i - \bar{u})^2 \quad (3)$$

$$SAD = \sum_{i=1}^n |u_i - \bar{u}| \quad (4)$$

These measures perform well as objectives in mathematical programming. However, there are drawbacks associated with these measures if they are employed as objectives. These measures can produce excellent results with low total deviation but

with poor efficiency. All demands can be assigned zero resources and still provide excellent results. Another drawback is that some demands are assigned zero resource at the cost of improving other demands with more resources. These measures quantify the differences on average and not considering every demand independently.

### 3.2 Gini Index

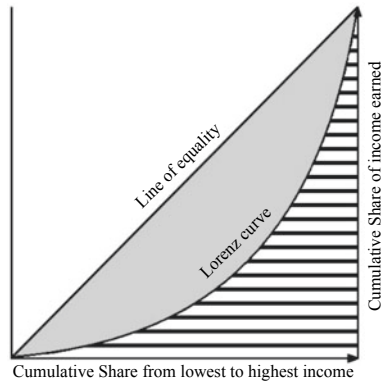
One of the measures to measure the income inequality in a society developed by Gini [18] is the Gini index or Gini coefficient as shown in (5). We denote by  $u_i$  the level of income of individual  $i$ . If the income of all individuals are equal, the absolute difference among incomes is zero indicating perfect equality. However, if one individual receives all the income and all others receive zero income, the Gini index equals one as seen in (6). Values greater than one are possible if the utility function of an individual is negative.

$$G = \frac{\sum_{j=1}^n \sum_{i=1}^n |u_i - u_j|}{2n \sum_{i=1}^n u_i} \tag{5}$$

$$0 \leq G \leq 1 \tag{6}$$

To find the Gini index value from the Lorenz curve shown in Fig. 1, the shaded area is divided by both the shaded and the striped areas under the curve. If Lorenz curve falls on the line of equality, then the shaded area becomes zero resulting in Gini index value of zero. See Gastwirth [17] for more details.

**Fig. 1** The cumulative share of people from lowest to highest income is plotted against the cumulative share of income. As the area between the Lorenz curve and the line of equity decreases (the gray area), a more equitable share of income is obtained resulting in a fairer distribution resources



### 3.3 Jain's Index

Jain's index, illustrated in (7), is derived from the coefficient of variation (COV). The variability of a series of numbers is measured independently by the COV. See Abdi [1] for more illustration. Jain's index is equal to  $1/(1+\text{COV}^2)$  indicating negative correlation with the COV. Jain's index is bounded by  $1/n$  and 1 and the higher the index the fairer the solution is as shown in (8). If  $k$  out of  $n$  demands are allocated resources fairly, Jain's index is equal to  $k/n$  meaning that it is very intuitive. See Jain et al. [24] for more details

$$\mathcal{J} = \frac{(\sum_{i=1}^n u_i)^2}{n \sum_{i=1}^n u_i^2} \quad (7)$$

$$1/n \leq \mathcal{J} \leq 1 \quad (8)$$

### 3.4 Unfairness

Unfairness is a measure used in traffic congestion control and traffic flow problems. The majority of traffic congestion control models are based on traffic assignment models. The traffic assignment models are based on different objectives. These models are System Optimal (SO), User Equilibrium (UE), Constrained System Optimal (CSO), and Nearest Allocation (NA). In SO, the sum of latencies of all entities in the network is minimized to minimize the network clearance time, which is suitable for emergency evacuation processes. SO is equivalent to minimizing the sum of utilities of all individuals for an efficient solution. UE is the state when all travel times for all entities are minimum and no entity can improve its travel time by changing routes. Hence, UE is applicable in nonemergency traffic control to minimize the overall congestion in the network. UE is the fairest solution in congestion control and equivalent to max-min fairness, which will be discussed in a later section. In CSO, a set of constraints are added to the SO to give a solution between SO and UE. The evacuees are assigned to the nearest shelter in NA in emergency evacuations.

The measure that provides an insight into the maximum difference between the highest value and lowest value of the assigned resource in the form of a fraction is the unfairness  $u(x)$ . See Correa et al. [13] for more details.  $v_i(x)$  is the cost of the resource  $x_i \forall i \in D$  given that  $D$  is the set of demands as shown in (9). The lower bound of the fraction is one indicating perfect fairness as seen in (10).

$$U(x) = \max \left\{ \frac{v_i(x)}{v_j(x)} : i, j \in D, x_i, x_j > 0 \right\} \quad (9)$$

$$u(x) \geq 1 \quad (10)$$

There are four variations of the unfairness measure. These fairness measures are loaded unfairness, normal unfairness, user equilibrium unfairness, and free-flow unfairness. In loaded unfairness, the ratio is the entity travel time to the fastest traveler on the same source–destination pair. Normal unfairness delivers the ratio of the length of the entity path to the shortest path of the same source–destination pair. User equilibrium unfairness provides the ratio of the entity travel time to the travel time on the same source–destination pair under user equilibrium. Finally, the ratio of the entity travel time to the length of the fastest path for the same source–destination pair is found by the free-flow unfairness. For more details see [23].

## 4 Lexicographic Ordering

To find an equitable solution using lexicographic ordering, the following lexicographic minimax model is applied.

$$\min_x \left\{ \max_{j=1, \dots, n} f_j(x), x \in X \right\} \quad (11)$$

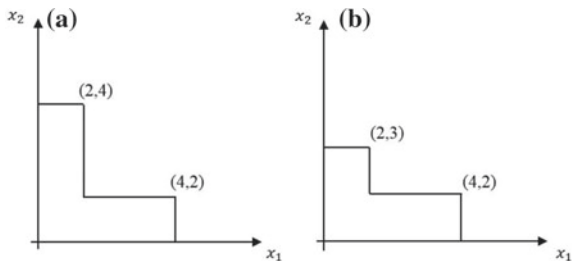
where  $X$  is the set of feasible solutions and  $f_j$ ,  $j = 1, \dots, n$  are the objective functions. The model (11) is solved to find the lexicographic minimax vector of optimal solutions. A vector is leximin larger than another vector if the nondecreasing ordered vector is lexicographically larger than the other nondecreasing ordered vector. The vector  $x$  is said to be lexicographically greater than the vector  $y$  ( $x > y$ ) if  $x_i = y_i$  and  $x_j > y_j$ ,  $\forall j > i$ . The vector  $x$  is said to be lexicographically greater or equal than the vector  $y$  ( $x \succeq y$ ) if  $x > y$  or  $x = y$ . The relationship between MMF vector and leximin maximal vector can be described by the following statement; if a vector  $x \in X$  is MMF over the set  $X$ , then the vector is leximin maximal over the set  $X$ . A vector  $x$  is leximin maximal over the set  $X$  if for all  $y \in X$ ,  $x \succeq y$ . However, the opposite relationship is not always true. The leximin maximal vector  $x$  is not necessarily MMF vector over the set  $X$  since leximin maximal vector  $x$  is not necessarily unique. For example, Fig. 2a does not have an MMF solution since the two points (2,4) and (4,2) contradict the MMF definition as we will discuss in the next section. However, these two points are leximin maximal solutions. Although Fig. 2b has a unique leximin maximal point (4,2), it is not MMF since  $x_2$  can be increased by decreasing  $x_1$ . See Radunovi and Boudec [35] for more details.

## 5 Max-Min Fairness

Max-Min Fairness (MMF) as defined by Bertsekas and Gallager [4] is a state, where all resources are utilized with the fairest possible allocation. A vector  $\gamma \in \Gamma$  is said to be max-min fair if it is feasible, and if there exists  $\gamma'_s > \gamma_s \forall s \in 1, \dots, n$ ,



**Fig. 2** An example of leximin maximum solutions that are not MMF solution to show that MMF is leximin max solution and unique but not the opposite



then  $\gamma'_t < \gamma_t \leq \gamma_s$  for  $t \in 1, \dots, n$ . Let us assume that the vector  $\gamma$  is MMF on the set  $\Gamma$ . A component  $\gamma_s$  in the vector  $\gamma$  cannot be increased without worsening another component  $\gamma_t$  that is less than or equal to  $\gamma_s$  on the same set. Water filling or progressive filling algorithm is an algorithm that gives an MMF solution. See Nace [33] for more illustration. MMF solution is unique and provides the fairest solution with the highest efficiency. As a result, it is a widely accepted approach to find fair and efficient resource allocation solutions. The algorithm is called water filling or progressive filling since it starts filling the lowest capacity components with the available resource until the component is filled, then it proceeds with the other components until the resource is consumed or all components are filled. Suppose there is a set of demands  $D$ , and the goal is to supply these demands with the resource  $\alpha$ . At iteration  $k = 0$ , the set of blocking demands  $L_k$  is an empty set meaning that the resource in all demands in constraint (13) is maximized when maximizing  $\alpha$  in the objective function (12). The LP problem  $P_k$  is described as follows:

$$\max \alpha, \quad (12)$$

s.t.

$$f_d(x) \geq \alpha, \quad \forall d \in DL_k, \quad (13)$$

$$f_d(x) \geq \lambda_d, \quad \forall d \in L_k, \quad (14)$$

$$x \in X. \quad (15)$$

The progressive filling algorithm to find the max-min fair solution is defined as follows:

1. Set  $k := 0$  and  $L_0 = \phi$ ;
2. While  $L_k \neq D$  do:

- a. Set  $k = k + 1$ . Solve the LP problem  $P_k$ , compute  $\alpha$
- b. Identify the set  $D_k$  of saturated demands, set  $\lambda_d = \alpha \forall d \in D_k$  and  $L_k = L_{k-1} \cup D_k$ .

The flow vector obtained at the last step is leximin maximal, and the obtained multi-commodity flow is thus a max-min fair one.

The iterative approach to find MMF flow in networks is very common, but there are some challenges related to it. The first challenge with this approach is solving the LP model a number of times that can be computationally demanding depending on the size of the network. Another challenge is in identifying the blocking constraints. The method to identify the blocking constraints in MMF is to identify the binding constraints through the dual variables. The corresponding dual variables to the binding constraints are positive according to the strict complementary slackness theorem, but the complementary slackness condition is not necessary. As a result, only a subset of the binding constraints is identified in each iteration leading to degeneracy since  $\alpha_k = \alpha_{k+1} = \alpha_{k+2} = \dots$  in subsequent iterations  $k$ . This eventually would lead to higher computational time, but convergence is guaranteed. In addition, water filling has some limitations depending on the structure of the problem. The feasible space has to be convex and the utility function is concave for the water filling algorithm to work. Bin Obaid and Trafalis [6] introduced an approximation model to find the max-min fair solution for bandwidth allocation in communication networks. The model works on convex and nonconvex problems. However, the max-min fair solution is not guaranteed based on the structure of the problem. Next, we discuss the concept of proportional fairness as an alternative fairness concept.

## 6 Proportional Fairness

Given a set of users  $R$ , a vector of rates  $x$  with components  $x_i, i \in R$  is proportionally fair if it is feasible and if for any other vector of rates  $y$  with components  $y_i, i \in R$ , the aggregate proportional changes is less than or equal to zero:

$$\sum_{i \in R} \frac{y_i - x_i}{x_i} \leq 0 \quad (16)$$

Proportional fairness is applied when the service received is proportional to the amount paid by the service receiver. Suppose bandwidth is sent to a set of demands  $N$  through a network with a set of capacitated resources  $J$  of capacity  $C$ , and a set  $A$  given  $A_{ji} = 1$  if demand  $i$  uses resource  $j$ . The objective (17) is to maximize the utility function for all demands subject to capacity constraints (18) and non-negativity constraints (19).

$$\max \sum_{i \in N} U_i(x_i), \quad (17)$$

s.t

$$Ax \leq C, \quad (18)$$

$$x \geq 0. \quad (19)$$

Assume that the amount paid  $w_r$  per unit time is decided by user  $r$  and receives flow  $x_r$  proportional to  $m_r$  given that  $x_r = m_r/\lambda_r$ , where  $\lambda_r$  is the cost per unit of time. Now the objective (20) is to maximize the utility of user  $i$  and minimize the amount paid by every user given that the amount paid is nonnegative (21).

$$\max U_i \left( \frac{m_i}{\lambda_i} \right) \quad (20)$$

s.t

$$m_i \geq 0 \quad (21)$$

Suppose that the vector  $m$  is known. The optimal solution of the objective (22) subject to the constraints (23) and (24) is proportionally fair. See Bonald et al. [9] for more details. If  $m_i = 1 \forall i \in R$ , the vector of rates is a Nash bargaining solution. See Kelly [26] and Kelly et al. [27] for more details.

$$\max \sum_{r \in R} m_i \log x_i, \quad (22)$$

s.t

$$Ax \leq C, \quad (23)$$

$$x \geq 0. \quad (24)$$

## 7 $(p, \alpha)$ - Proportional Fairness

Let  $p = (p_1, p_2, \dots, p_{|R|})$  where  $p_i, i = 1 \dots R$ , and  $\alpha$  are positive numbers. A vector of rates  $x$  with components  $x_i, i \in R$  is  $(p, \alpha)$ —proportionally fair if it is feasible and for any other vector  $y$  with components  $y_i, i \in R$  the aggregate proportional changes is less than or equal to zero:

$$\sum_{i \in R} p_i \frac{y_i - x_i}{x_i^\alpha} \quad (25)$$

Although max-min fairness is a widely accepted approach for fair and efficient resource allocation, some researchers claim that max-min fairness gives absolute

priority to fairness.  $(p, \alpha)$ —proportional fairness is a generalization of max-min fairness and proportional fairness. Since fairness and utilization are contradicting objectives,  $(p, \alpha)$ —proportional fairness compromises between resource utilization, in objective (17), and proportional fair solution in objective (22). Note that  $(p, \alpha)$ —proportional fairness, as shown in (26), is a generalization to proportional fairness [32]. For  $\alpha = 0$ , the maximum utilization is found  $U(0) = \sum_{r \in R} x_r$  and for  $\alpha = 1$ , the result is the proportional fair solution  $U(1) = \sum_{r \in R} \log(x_r)$ .

$$U_i(x_i, \alpha) = \begin{cases} \frac{x_i^{1-\alpha}}{1-\alpha} & \text{for } \alpha \geq 0, \alpha \neq 0 \\ \log(x_i) & \text{for } \alpha = 1 \end{cases} \quad (26)$$

As  $\alpha$  goes to  $\infty$ , the solution converges to max-min fairness. See Mo and Walrand [32] for more details.

## 8 Price of Fairness

The price of fairness is a measure of the percentage of utilization lost when fairness is incorporated in the model, Bertsimas and Farias [5]. To find an efficient solution and its optimal allocation, the model (27)–(28) is solved, and the optimal value is denoted with  $SYSTEM(U)$ .

$$\max e^T u, \quad (27)$$

s.t.

$$u \in U. \quad (28)$$

If the DM objective is to fairly distribute the resources among demands, the sum of utilities is denoted by  $FAIR(U; \phi) = e^T \phi(U)$ . Since efficiency and fairness are contradicting objectives, the sum of utilities will mostly decrease resulting in a loss in efficiency. The percentage of loss can be measured by the price of fairness  $POF(U; \phi)$  in (29). The closer the POF to 0, the lower efficiency is compromised.

$$POF(U, \phi) = \frac{SYSTEM(U) - FAIR(U; \phi)}{SYSTEM(U)} \quad (29)$$

The upper bounds of the POF when MMF and PF are the fairness schemes considered as seen in (30) and (31). See Bertsimas and Farias [5] for more details.

$$POF(U, \phi^{PF}) \leq 1 - \frac{2\sqrt{n} - 1}{n} \quad (30)$$

$$POF(U, \phi^{MMF}) \leq 1 - \frac{4n}{(n+1)^2} \quad (31)$$

where  $\phi^{PF}$  is the proportional fairness scheme, and  $\phi^{MMF}$  is the max-min fairness scheme.

## 9 Concluding Remarks

In this paper, we provided a comprehensive review of fairness in resource allocation. Fairness is applicable in applications with multiple demands. The set of demands are assigned utility functions based on their preferences, then the utilities are maximized fairly for a fair efficient solution given that fairness is more important than efficiency in some applications. There is no single method to find a fair and efficient resource allocation. Every approach has its strength points and weaknesses. Using basic statistical measures may not give the fairest or the most efficient solution since the average or total deviation is minimized disregarding the individual allocations. Gini index, Jain's index, and unfairness are nonlinear functions and can lead to more computational complexity. Lexicographic ordering and max-min fairness can be computationally demanding with the problem size if the progressive filling algorithm is used. Finding a proportional fair or  $(p, \alpha)$ —proportional fair solution can be challenging. Lagrangean duality is one of the approaches to find a proportionally fair solution.

**Acknowledgements** Dr. Theodore Trafalis was supported by RSF Grant 14-41-00039, and he conducted research at National Research University Higher School of Economics.

## References

1. Abdi, H.: Coefficient of variation. *Encycl. Res. Des.* **1**, 169–171 (2010)
2. Bayram, V., Tansel, B.Ç., Yaman, H.: Compromising system and user interests in shelter location and evacuation planning. *Transp. Res. Part B Methodol* **72**, 146–163 (2015)
3. Beheshtifar, S., Alimoahmadi, A.: A multiobjective optimization approach for location?allocation of clinics. *Int. Trans. Oper. Res.* **22**(2), 313–328 (2015)
4. Bertsekas, D., Gallager, R.: *Data networks*. Prentice-Hall (1987)
5. Bertsimas, D., Farias, V.F., Trichakis, N.: The price of fairness. *Oper. Res.* **59**(1), 17–31 (2011)
6. Bin Obaid, H., Trafalis, T.B.: Linear Max-Min Fairness in Multi-commodity Flow Networks. In: *International Conference on Network Analysis*, pp. 3–10. Springer, Cham (2016)
7. Bish, D.R., Sherali, H.D., Hobeika, A.G.: Optimal evacuation planning using staging and routing. *J. Oper. Res. Soc.* **65**(1), 124–140 (2014)
8. Boche, H., Wicznanowski, M., Stanczak, S.: Unifying view on min-max fairness, max-min fairness, and utility optimization in cellular networks. *EURASIP J. Wireless Commun. Netw.* **2007**(1), 034869 (2007)
9. Bonald, T., Massoulié, L., Proutiere, A., Virtamo, J.: A queueing analysis of max-min fairness, proportional fairness and balanced fairness. *Queue. Syst.* **53**(1–2), 65–84 (2006)

10. Brams, S.J., Taylor, A.D.: Fair division: from cake-cutting to dispute resolution (1996)
11. Buzna, L., Koháni, M., Janáček, J.: An approximation algorithm for the facility location problem with lexicographic minimax objective. *J. Appl. Math.* (2014)
12. Chen, M.A.: Individual monotonicity and the leximin solution. *Econ. Theory* **15**(2), 353–365 (2000)
13. Correa, J.R., Schulz, A.S., Stier-Moses, N.E.: Fast, fair, and efficient flows in networks. *Oper. Res.* **55**(2), 215–225 (2007)
14. Daganzo, C.: The cell transmission model Part I: a simple dynamic representation of highway traffic. *PATH Rep.* **93-0409**, 3 (1993)
15. Erkut, E., Karagiannidis, A., Perkoulidis, G., Tjandra, S.A.: A multicriteria facility location model for municipal solid waste management in North Greece. *Eur. J. Oper. Res.* **187**(3), 1402–1421 (2008)
16. Feitelson, D.G., Rudolph, L.: Parallel job scheduling: issues and approaches. In: *Workshop on Job Scheduling Strategies for Parallel Processing*, pp. 1–18. Springer, Berlin, Heidelberg (1995)
17. Gastwirth, J.L.: The estimation of the Lorenz curve and Gini index. *Rev. Econ. Stat.* 306–316 (1972)
18. Gini, C.: Variabilità e mutabilità. In: Pizetti, E., Salvemini, T. (eds) *Reprinted in Memorie di metodologica statistica*. Rome: Libreria Eredi Virgilio Veschi
19. Gini, C.: Sulla misura della concentrazione e della variabilità dei caratteri. *Atti del Reale Istituto veneto di scienze, lettere ed arti* **73**, 1203–1248 (1914)
20. Goldberg, J.B.: Operations research models for the deployment of emergency services vehicles. *EMS Manage. J.* **1**(1), 20–39 (2004)
21. Golovin, D.: Max-min fair allocation of indivisible goods (2005)
22. Hsu, C.C., Lin, K.C.J., Lai, Y.R., Chou, C.F.: On exploiting spatial-temporal uncertainty in max-min fairness in underwater sensor networks. *IEEE Commun. Lett.* **14**(12), 1098–1100 (2010)
23. Jahn, O., Möhring, R.H., Schulz, A.S., Stier-Moses, N.E.: System-optimal routing of traffic flows with user constraints in networks with congestion. *Oper. Res.* **53**(4), 600–616 (2005)
24. Jain, R., Duresi, A., Babić, G.: Throughput fairness index: an explanation, p. 99. Department of CIS, The Ohio State University, Tech. rep. (1999)
25. Jonker, G.M., Meyer, J.J., Dignum, F.P.M.: Efficiency and fairness in air traffic control. In: *Proceedings 7th Belgium-Netherlands conference on artificial intelligence*, pp. 151–157. KVAB (2005)
26. Kelly, F.: Charging and rate control for elastic traffic. *Eur. Trans. Telecommun.* **8**(1), 33–37 (1997)
27. Kelly, F.P., Maulloo, A.K., Tan, D.K.: Rate control for communication networks: shadow prices, proportional fairness and stability. *J. Oper. Res. Soc.* **49**(3), 237–252 (1998)
28. Leclerc, P.D., McLay, L.A., Mayorga, M.E.: Modeling equity for allocating public resources. In: *Community-based operations research*, pp. 97–118. Springer, New York (2012)
29. Mahajan, R., Floyd, S., Wetherall, D.: Controlling high-bandwidth flows at the congested router. In: *Network Protocols, 2001. Ninth International Conference on*, pp. 192–201. IEEE (2001)
30. Mas-Colell, A.: Microeconomic theory/Andreu Mas-Colell. Michael D. Whinston and Jerry R. Green (1995)
31. Megiddo, N.: Optimal flows in networks with multiple sources and sinks. *Math. Program.* **7**(1), 97–107 (1974)
32. Mo, J., Walrand, J.: Fair end-to-end window-based congestion control. *IEEE/ACM Trans. Netw.* **8**(5), 556–567 (2000)
33. Nace, D., Pióro, M.: Max-min fairness and its applications to routing and load-balancing in communication networks: a tutorial. *IEEE Commun. Surv. Tutorials* **10**(4), (2008)
34. Pan, D., Yang, Y.: Max-min fair bandwidth allocation algorithms for packet switches. In: *2007 IEEE international parallel and distributed processing symposium*, p. 52. IEEE (2007)
35. Radunović, B., Boudec, J.Y.L.: A unified framework for max-min and min-max fairness with applications. *IEEE/ACM Trans. Netw. (TON)* **15**(5), 1073–1083 (2007)

36. Rawls, J.: A theory of justice. Harvard University Press (1971)
37. Schwiegelshohn, U., Yahyapour, R.: Analysis of first-come-first-serve parallel job scheduling. *SODA* **98**, 629-638 (1998)
38. Siu, K.Y., Tzeng, H.Y.: Congestion control for multicast service in ATM networks. In: Global Telecommunications Conference, 1995. GLOBECOM'95., IEEE, (Vol. 1, pp. 310–314). IEEE (1995)
39. Sridharan, A., Krishnamachari, B.: Maximizing network utilization with max?min fairness in wireless sensor networks. *Wireless Netw.* **15**(5), 585–600 (2009)
40. Steinhaus, H.: The problem of fair division. *Econometrica* **16**(1), (1948)
41. Tang, J., Xue, G., Zhang, W.: Maximum throughput and fair bandwidth allocation in multi-channel wireless mesh networks. In: INFOCOM 2006. 25th IEEE International Conference on Computer Communications. Proceedings, pp. 1–10. IEEE (2006)
42. Thawari, V.W., Babar, S.D., Dhawas, N.A.: An efficient data locality driven task scheduling algorithm for cloud computing. *Int. J. Multi. Acad. Res. (SSIJMAR)* **1**(3), (2012)
43. Thulasiraman, P., Chen, J., Shen, X.: Multipath routing and max-min fair QoS provisioning under interference constraints in wireless multihop networks. *IEEE Trans. Parallel Distrib. Syst.* **5**, 716–728 (2011)
44. Wang, Y., Tan, J., Yu, W., Zhang, L., Meng, X., Li, X.: Preemptive reduceTask scheduling for fair and fast job completion. In: ICAC, pp. 279–289 (2013)
45. Wardrop, J.G.: Some theoretical aspects of road traffic research. In: Inst Civil Engineers Proc London/UK/ (1952)
46. Zaharia, M., Borthakur, D., Sen Sarma, J., Elmeleegy, K., Shenker, S., Stoica, I.: Delay scheduling: a simple technique for achieving locality and fairness in cluster scheduling. In: Proceedings of the 5th European conference on Computer systems, pp. 265–278. ACM (2010)
47. Zhang, Y., Xiong, K., An, F., Di, X., Su, J.: Mobile-service based max-min fairness resource scheduling for heterogeneous vehicular networks (2016). [arXiv:1603.03645](https://arxiv.org/abs/1603.03645)

# Mixed Integer Programming for Searching Maximum Quasi-Bicliques



Dmitry I. Ignatov , Polina Ivanova  and Albina Zamaletdinova

**Abstract** This paper is related to the problem of finding the maximal quasi-bicliques in a bipartite graph (bigraph). A quasi-biclique in the bigraph is its “almost” complete subgraph. The relaxation of completeness can be understood variously; here, we assume that the subgraph is a  $\gamma$ -quasi-biclique if it lacks a certain number of edges to form a biclique such that its density is at least  $\gamma \in (0, 1]$ . For a bigraph and fixed  $\gamma$ , the problem of searching for the maximal quasi-biclique consists of finding a subset of vertices of the bigraph such that the induced subgraph is a quasi-biclique and its size is maximal for a given graph. Several models based on Mixed Integer Programming (MIP) to search for a quasi-biclique are proposed and tested for working efficiency. An alternative model inspired by biclustering is formulated and tested; this model simultaneously maximises both the size of the quasi-biclique and its density, using the least-square criterion similar to the one exploited by triclustering TriBOX.

**Keywords** Quasi-biclique · Maximal quasi-biclique · Mixed integer programming · Biclustering · Triclustering

---

D. I. Ignatov (✉) · P. Ivanova · A. Zamaletdinova  
National Research University Higher School of Economics, Moscow, Russian Federation  
e-mail: [dignatov@hse.ru](mailto:dignatov@hse.ru)

P. Ivanova  
e-mail: [ivanova.p.m@gmail.com](mailto:ivanova.p.m@gmail.com)

A. Zamaletdinova  
e-mail: [aazamaletdinova\\_1@edu.hse.ru](mailto:aazamaletdinova_1@edu.hse.ru)

D. I. Ignatov  
St. Petersburg Department of Steklov Mathematical Institute of Russian Academy of Sciences,  
Saint Petersburg, Russian Federation

© Springer Nature Switzerland AG 2020  
I. Bychkov et al. (eds.), *Network Algorithms, Data Mining, and Applications*, Springer Proceedings in Mathematics & Statistics 315,  
[https://doi.org/10.1007/978-3-030-37157-9\\_2](https://doi.org/10.1007/978-3-030-37157-9_2)



## 1 Introduction

There are many data sources that can be represented as a bipartite graph; for example, in recommender systems and web stores, users can interact with different items like movies, books, clothes, and other products. The most commonly studied data usually has a structure of a bipartite graph, whose vertices form two disjoint sets. For example, social network data, where a binary relation between two sets show interactions between people and communities, advertisement data with a set of consumers and a corresponding set of products, and so on.

In this study, we are interested in the analysis of such bipartite data and search for dense communities, where almost all elements are connected. A situation where all elements of a community are involved can be described by a concept of a biclique or a complete subgraph of a bipartite graph.

Unfortunately, the community completeness requirement excludes almost complete communities frequently met in real-world data. Due to this reason, we allow some edges to be absent and introduce the concept of quasi-biclique. In order to bound the size of quasi-biclique, we can use the subgraph minimal density or the maximum number of absent edges needed to complete a subgraph.

The problem of searching for maximal quasi-clique is NP-hard [15] as well as the problem of searching for maximal quasi-biclique [12]; the maximum edge biclique problem is known to be NP-complete [16]. Many algorithms that solve those problems are being developed [1, 11, 17, 19]. For instance, Veremyev et al. [18] offered an exact Mixed Integer Programming model for searching for maximal quasi-clique but the case of bipartite graphs for quasi-bicliques was not yet considered within the MIP framework.

The aim of this paper is to propose a Mixed Integer Programming models for finding a maximum quasi-bicliques in a bipartite graph and compare the results obtained by those models with those of existing algorithms.

The paper is organised as follows. Section 2 introduces several basic definitions, namely biclique, quasi-bicliques, and its density and provides a short overview of related work along with important propositions on algorithmic aspects. Section 3 proposes two Mixed Integer Programming models for quasi-biclique search. In Sect. 4, the chosen datasets are described. Section 5 summarises the experimental results. Section 6 concludes the paper.

## 2 Maximum Quasi-Cliques and Quasi-Bicliques

### 2.1 Basic Definitions

Let us introduce several basic notions.

**Definition 1** In a graph  $G = (V, E)$  a subgraph  $G' = (V', E')$ , where  $V' \subseteq V$ ,  $E' \subseteq E$ , is called a vertex-induced subgraph. Let us denote such graph as  $G[V']$ .

**Definition 2** A complete subgraph of a graph  $(V, E)$  is called a clique.

**Definition 3** A complete bipartite subgraph in a bipartite graph  $(U, V, E \subseteq U \times V)$  is called a biclique.

**Definition 4** The density of an arbitrary graph is the ratio of the number of edges to the maximum possible number of edges.

The density of a bipartite graph  $G = (V, U, E)$  is  $\rho = \frac{|E|}{|V||U|}$ .

**Definition 5** A subgraph  $G' = (V', E')$  of a given graph  $G = (V, E)$  is called  $f(k)$ -dense, if  $G'$  is a subgraph induced by a vertex subset  $V' \subseteq V$ ,  $|V'| = k$  and  $|E'| \geq f(k)$ , where  $f : Z_+ \rightarrow R_+$  is a chosen function.

**Definition 6** A  $\gamma$ -quasi-biclique in a bipartite graph  $G = (U, V, E)$  is its induced bipartite subgraph  $G' = (V', U', E' \subseteq V' \times U')$  with the density at least  $\gamma \in (0, 1]$ .

## 2.2 Maximum Quasi-Cliques

Let us consider properties and searching algorithms of cliques in a graph  $G = (V, E)$ .

For a graph  $G = (V, E)$  and a fixed  $\gamma \in (0, 1]$ , we need to find a  $V' \subseteq V$  such that  $G[V']$  is a  $\gamma$ -quasi-clique and  $|V'|$  is maximal.

Problem of searching for maximum quasi-clique as well as the problem of searching for maximum clique is NP-hard [12, 15]. In addition to that, the assumption of graph incompleteness leads to the loss of useful properties of a clique. For instance, inheritance property which is used in most maximum clique searching algorithms does not hold. Namely, if  $G[V]$  is a clique, then  $G[V']$  is a clique as well, where  $V'$  is a subset of  $V$ . This property does not hold for  $\gamma$ -quasi-cliques: i.e., a subset of a  $\gamma$ -quasi-clique is not necessarily a  $\gamma$ -quasi-clique.

However, for quasi-cliques we can define the property of quasi-inheritance:  $\gamma$ -quasi-clique with  $|V| > 1$  is a strict superset to a  $\gamma$ -quasi-clique with  $|V| - 1$  vertices [15].

## 2.3 Maximum Quasi-Bicliques

The problem of maximum quasi-biclique in a bipartite graph  $G = (U, V, E)$  with fixed  $\gamma \in (0, 1]$  is to find  $U' \subseteq U$  and  $V' \subseteq V$  such that vertex-induced subgraph  $G[U', V']$  is a  $\gamma$ -quasi-biclique of size  $|U'| + |V'|$ , maximum for this graph. Let us denote a maximum  $\gamma$ -quasi-biclique in the graph  $G$  by  $\omega_\gamma(G)$ .

Let us consider several commonly met definitions of quasi-biclique. In Liu et al. [12], we can find the following definition.

**Definition 7** A induced subgraph  $G'[U', V']$  is called a  $\delta$ -quasi-biclique ( $0 \leq \delta \leq 0.5$ ) in a bipartite graph  $G = (U, V, E)$  if:

1.  $\forall u \in U' : d(u, V') = |\{v \in V' | (u, v) \in E\}| \geq (1 - \delta) \cdot |V'|$ ,
2.  $\forall v \in V' : d(v, U') = |\{u \in U' | (u, v) \in E\}| \geq (1 - \delta) \cdot |U'|$ .

In order to consider the third definition of quasi-biclique by Sim et al. [17], let us introduce some useful notations. The neighbourhood of a vertex  $v \in V$  in a graph  $G = (V, E)$  is a set of vertices  $\Gamma(v) = \{u \in V | (u, v) \in E\}$ .

For a vertex set  $V' \subseteq V$  and a vertex  $v \in V \setminus V'$ , let us denote a set of vertices from  $V'$  adjacent to  $v$  as  $\Gamma_{V'}(v) = \{u | (u, v) \in E \& u \in V'\}$ . By a set of vertices  $\Gamma(V') = \cup_{v \in V'} \Gamma(v)$ , we denote a loose neighbourhood of subset  $V'$ .

**Definition 8** A subgraph  $G'[U', V']$  of a bipartite graph  $G(U, V, E)$  is called an  $\epsilon$ -quasi-biclique, if for some small positive integer  $\epsilon$ :

1.  $\forall u \in U' |V'| - |\Gamma_{V'}(u)| \leq \epsilon$ ,
2.  $\forall v \in V' |U'| - |\Gamma_{U'}(v)| \leq \epsilon$ .

*Remark* Obviously, Definitions 7 and 8 of quasi-bicliques can be reduced to the definition of  $\gamma$ -quasi-biclique.

1. In Definition 7, let us sum the first condition over all vertices from  $U'$ . We get that  $\sum_{u \in U'} d(u, V') \geq (1 - \delta) \cdot |V'| |U'|$ , where  $\sum_{u \in U'} d(u, V')$  is a number of edges in a  $\delta$ -quasi-biclique,  $|V'| |U'|$  is the maximum possible number of edges in a bipartite graph. Thus a  $\delta$ -quasi-biclique is a  $\gamma$ -quasi-biclique with  $\gamma = 1 - \delta$ . Both definitions of quasi-biclique are equivalent if  $\gamma \in [0.5, 1]$ .
2. By summing both conditions over sets  $U'$  and  $V'$ , respectively, in Definition 8 we get

$$\frac{\sum_{u \in U'} |\Gamma_{V'}(u)|}{|U'| |V'|} \geq 1 - \frac{\epsilon}{|V'|}, \quad \frac{\sum_{v \in V'} |\Gamma_{U'}(v)|}{|U'| |V'|} \geq 1 - \frac{\epsilon}{|U'|}.$$

Since  $\sum_{u \in U'} |\Gamma_{V'}(u)| = \sum_{v \in V'} |\Gamma_{U'}(v)|$  is a number of edges in an  $\epsilon$ -quasi-biclique  $G[U', V']$ , then the density of  $G[U', V']$  is

$$\rho(G[U', V']) \geq 1 - \frac{\epsilon}{\min(|U'|, |V'|)}.$$

Bounding the size of a quasi-clique vertex sets from below  $\omega_l^{(1)} \leq |U'|$  and  $\omega_l^{(2)} \leq |V'|$ , we can establish a connection between these definitions. If we let

$$\gamma = 1 - \frac{\epsilon}{\min(\omega_l^{(1)}, \omega_l^{(2)})},$$

we obtain that  $G[U', V']$  is a  $\gamma$ -quasi-clique under condition  $\epsilon \in [0, \min(\omega_l^{(1)}, \omega_l^{(2)})]$ .

Most properties of quasi-cliques naturally fulfil for quasi-bicliques as well. However, since the density definition of quasi-biclique differs from the case of quasi-clique and the maximum number of edges is a function of two variables with no convex properties, most algorithms searching for maximum quasi-clique are not directly applicable to search for maximum quasi-biclique.

Pattillo et al. [15] established inequalities for upper bounds for the size of maximum quasi-clique shown below.

**Proposition 1** *In a graph  $G = (V, E)$  with  $|V| = n$  and  $|E| = m$  the maximum size of a quasi-clique  $\omega_\gamma(G)$  satisfies the following inequality:*

$$\omega_\gamma(G) \leq \frac{\gamma + \sqrt{\gamma + 8\gamma m}}{2\gamma}. \quad (1)$$

In order to obtain similar bound for a quasi-biclique, we need to allow the following conditions on quasi-biclique.

**Proposition 2** *In a bipartite graph  $G(U, V, E)$ , with  $|U| = n_U$ ,  $|V| = n_V$  and  $|E| = m$ , the maximum size of a quasi-biclique  $\omega_\gamma(G)$  satisfies the following inequalities:*

1.  $\omega_\gamma(G) \leq \sqrt{\frac{4m}{\gamma}}$ , for balanced quasi-biclique (the sizes of two vertex sets  $U$  and  $V$  are equal),
2.  $\omega_\gamma(G) \leq \min \left\{ (2 + \theta) \cdot \sqrt{\frac{m}{\gamma(1 - \theta)}}, \left(1 + \frac{1}{1 - \theta}\right) \cdot \sqrt{\frac{m(1 + \theta)}{\gamma}} \right\}$ , if  $\theta \in (0, 1)$  and sizes of vertex sets differ from each other by no more than in  $\theta$ .

**Proof** Let  $U'$  and  $V'$  be vertex sets of a maximum  $\gamma$ -quasi-biclique and let  $n_{U'}$  and  $n_{V'}$  be their cardinalities, respectively.

1. For balanced quasi-clique  $n_{U'} = n_{V'}$ , hence  $\omega_\gamma(G) = 2 \cdot n_{U'}$ . Obviously, the maximum possible number of edges in a quasi-biclique is less than the total number of graph edges. Then

$$\gamma \cdot n_{U'}^2 = \gamma \cdot \left( \frac{\omega_\gamma(G)}{2} \right)^2 \leq m,$$

2.  $\gamma$ -quasi-biclique is ‘‘almost’’ balanced when  $(1 - \theta) n_{V'} \leq n_{U'} \leq (1 + \theta) n_{V'}$ . Thus,

$$\begin{aligned} \omega_\gamma(G) = n_{U'} + n_{V'} &\leq (2 + \theta)n_{V'} \Rightarrow \\ m \geq \gamma \cdot n_{U'} \cdot n_{V'} &\geq \gamma \cdot (1 - \theta) \cdot n_{V'}^2 \geq \gamma \cdot (1 - \theta) \cdot \left( \frac{\omega_\gamma(G)}{2 + \theta} \right)^2 \Rightarrow \\ &\Rightarrow \omega_\gamma(G) \leq \sqrt{\frac{m(2 + \theta)^2}{\gamma(1 - \theta)}}. \end{aligned}$$

Analogously,

$$\begin{aligned} \omega_\gamma(G) &\leq \left(1 + \frac{1}{1-\theta}\right) n_{U'}, \quad \gamma \cdot n_{U'} \cdot n_{V'} \geq \gamma \cdot \frac{1}{(1+\theta)} n_{U'}^2 \Rightarrow \\ &\Rightarrow \omega_\gamma(G) \leq \left(1 + \frac{1}{1-\theta}\right) \sqrt{\frac{m(1+\theta)}{\gamma}}. \end{aligned}$$

Now let us discuss a few chosen algorithms that implement maximum quasi-biclique search.

A greedy algorithm for searching maximum quasi-bicliques according to Definition 7 is discussed in detail by Liu et al. [12]. The algorithm uses two parameters: (1)  $\delta$  to control the size of the quasi-biclique ( $\delta = 1 - \gamma$ ) and (2)  $\tau$  to control the smallest possible number of vertices that belong to one of the partitions of a quasi-biclique. Let us denote by  $U'$  and  $V'$  vertex sets of quasi-biclique of the graph  $G(U, V, E)$ . At the beginning of the algorithm, we set  $U' = \emptyset$  and  $V' = V$ . From the vertex set  $U \setminus U'$ , we choose such vertex  $u$  that its degree is maximum and delete from  $V'$  all vertices for which  $d(v, U') < (1 - \delta) \cdot |U'|$ . This process continues as long as the size of  $U' < \tau$ . However, this algorithm can miss possible vertex candidates, thus the authors introduce the second step of the algorithm: if there is a vertex  $u$  outside of the current vertex set  $U'$  such that its degree is maximum in  $U \setminus U'$  and  $U' \cup \{u\}$  remains a quasi-biclique, then it can be added to  $U'$ . The same applies to  $V'$  as long as there is a vertex to add.

### 3 Quasi-biclique Searching Models

#### 3.1 Model 1

In this section, we will show how to adapt the model **F3** from Veremyev et al. [18] for searching for maximum quasi-bicliques. Let us consider disjoint sets  $U' \cup V'$ ,  $U' \cap V' = \emptyset$  that form a quasi-biclique of a bipartite graph  $G = (U, V, E)$ . Using similar techniques as in Veremyev et al. [18], we introduce the following variables:

$$\begin{aligned} u_i &= 1 \Leftrightarrow i \in U', \\ v_j &= 1 \Leftrightarrow j \in V', \\ y_{ij} &= 1 \Leftrightarrow \exists(i, j) \in E \cap (U' \times V') \\ z_k^{(1)} &= 1 \Leftrightarrow |U'| = k, \quad z_k^{(2)} = |V'|, \end{aligned}$$

$\omega_l^{(1)}, \omega_u^{(1)}$  are the lower and upper bounds, respectively, for the vertex set  $U'$ ,  
 $\omega_l^{(2)}, \omega_u^{(2)}$  are the lower and upper bounds, respectively, for the vertex set  $V'$ .

We can refine the sizes of vertex sets of a quasi-biclique using Proposition 2. Then we build Model 1:

**Model 1**

$$\omega_\gamma(G) = \max_{u,v,y,z} \left[ \sum_{i \in U} u_i + \sum_{j \in V} v_j \right], \quad (2)$$

$$\text{under conditions } \sum_{(i,j) \in E} y_{ij} \geq \gamma \sum_{n=\omega_l^{(1)}}^{\omega_u^{(1)}} \sum_{m=\omega_l^{(2)}}^{\omega_u^{(2)}} n \cdot m \cdot z_n^{(1)} \cdot z_m^{(2)}, \quad (3)$$

$$\forall i \in U, \forall j \in V : y_{ij} \leq u_i, y_{ij} \leq v_j, y_{ij} \geq u_i + v_j - 1, \quad (4)$$

$$\sum_{i \in U} u_i = \sum_{n=\omega_l^{(1)}}^{\omega_u^{(1)}} n z_n^{(1)}, \sum_{j \in V} v_j = \sum_{m=\omega_l^{(2)}}^{\omega_u^{(2)}} m z_m^{(2)}, \quad (5)$$

$$\sum_{n=\omega_l^{(1)}}^{\omega_u^{(1)}} z_n^{(1)} = 1, \sum_{m=\omega_l^{(2)}}^{\omega_u^{(2)}} z_m^{(2)} = 1, \quad (6)$$

$$\forall i \in U, \forall j \in V : u_i \in \{0, 1\}, v_j \in \{0, 1\}, \forall i < j, (i, j) \in E : y_{ij} \in \{0, 1\}, \quad (7)$$

$$\forall n \in \{\omega_l^{(1)} : z_n^{(1)} \geq 0, \dots, \omega_u^{(1)}\}, \forall m \in \{\omega_l^{(2)} : z_m^{(2)} \geq 0, \dots, \omega_u^{(2)}\} : z_m^{(2)} \geq 0. \quad (8)$$

As in the model **F3**, we can bound  $z_k^{(1)}$  and  $z_k^{(2)}$  and recast them from binary into continuous variables. Suppose there exists an optimal solution  $(u^*, v^*, y^*, \overline{z}^{(1)}, \overline{z}^{(2)})$  of Model 1, where vectors  $\overline{z}^{(1)}$  and  $\overline{z}^{(2)}$  are not binary ( $\overline{z}_n^{(1)} \geq 0, \overline{z}_n^{(2)} \geq 0$ ). Let  $\widehat{z}^{(1)}$  be a binary vector with  $\widehat{z}_k^{(1)} = 1 \Leftrightarrow |U'| = k$  and  $\widehat{z}_k^{(1)} = 0$  otherwise, where  $k \in \{\omega_l^{(1)}, \dots, \omega_u^{(1)}\}$ ; analogously, vector  $\widehat{z}^{(2)} : \widehat{z}_k^{(2)} = 1 \Leftrightarrow |V'| = k$  and 0 otherwise. Hence, it is obvious that vectors  $\widehat{z}^{(1)}$  and  $\widehat{z}^{(2)}$  satisfy constraint (6). Constraints (3) and (5) can be rewritten as follows:

$$\begin{aligned} \sum_{i \in U} u_i^* &= \sum_{n=\omega_l^{(1)}}^{\omega_u^{(1)}} n \widehat{z}_n^{(1)}, \sum_{j \in V} v_j^* = \sum_{m=\omega_l^{(2)}}^{\omega_u^{(2)}} m \widehat{z}_m^{(2)} \text{ (by definition),} \\ \sum_{(i,j) \in E} y_{ij}^* &\geq \gamma \sum_{n=\omega_l^{(1)}}^{\omega_u^{(1)}} \sum_{m=\omega_l^{(2)}}^{\omega_u^{(2)}} n \cdot m \cdot \overline{z}_n^{(1)} \cdot \overline{z}_m^{(2)} = \\ &= \gamma \left( \sum_{n=\omega_l^{(1)}}^{\omega_u^{(1)}} n \cdot \overline{z}_n^{(1)} \right) \left( \sum_{m=\omega_l^{(2)}}^{\omega_u^{(2)}} m \cdot \overline{z}_m^{(2)} \right) = \gamma \left( \sum_{i \in U} u_i^* \right) \left( \sum_{m=\omega_l^{(2)}}^{\omega_u^{(2)}} \sum_{j \in V} v_j^* \right) \\ &= \gamma \left( \sum_{n=\omega_l^{(1)}}^{\omega_u^{(1)}} n \cdot \widehat{z}_n^{(1)} \right) \left( \sum_{m=\omega_l^{(2)}}^{\omega_u^{(2)}} m \cdot \widehat{z}_m^{(2)} \right). \end{aligned}$$

This means that  $(u^*, v^*, y^*, \widehat{z}^{(1)}, \widehat{z}^{(2)})$  is also an optimal solution of the problem and usage of continuous variables  $z_n^{(1)}$  and  $z_m^{(2)}$  in Model 1 is proved.

In the worst case, when  $\omega_l^{(1)} = \omega_l^{(2)} = 1$ ,  $\omega_u^{(1)} = |U|$ ,  $\omega_u^{(2)} = |V|$ , the model has  $|U| + |V|$  binary variables and  $|E| + |U| + |V|$  continuous.

*Remark* In Model 1, condition (3) is not linear, so we can linearise it. Let us introduce a new variable  $z_{n,m} = z_n^{(1)} \cdot z_m^{(2)}$ . Then left side of the inequality (3) is

$$\sum_{n=\omega_l^{(1)}}^{\omega_u^{(1)}} \sum_{m=\omega_l^{(2)}}^{\omega_u^{(2)}} (n \cdot m) \cdot z_{n,m}.$$

Conditions (5) are changed as follows:

$$\sum_{i \in U} u_i = \sum_{n=\omega_l^{(1)}}^{\omega_u^{(1)}} \sum_{m=\omega_l^{(2)}}^{\omega_u^{(2)}} n z_{n,m}, \quad \sum_{j \in V} v_j = \sum_{n=\omega_l^{(1)}}^{\omega_u^{(1)}} \sum_{m=\omega_l^{(2)}}^{\omega_u^{(2)}} m z_{n,m},$$

where  $c_{n,m}^{(1)} = n$  and  $c_{n,m}^{(2)} = m$ .

Using this substitution for variables  $z_n^{(1)}$  and  $z_m^{(2)}$ , the model becomes a linear integer programming model. In the worst-case scenario, for dense graph there are  $|U| + |V|$  binary variables and  $|E| + |U| \cdot |V|$  continuous variables to be optimised.

### 3.2 Model 2

Let us look at different maximising criteria for related Mixed Integer Programming models. In papers [9, 14] which are dedicated to triclustering generation,  $K = (G, M, B, I)$  is a triadic context with  $G$ , the set of objects,  $M$ , the set of attributes,  $B$ , the set of conditions, and  $I \subseteq G \times M \times B$ , the ternary relation. The proposed triclustering algorithm searches for clusters that maximise the following criteria:

$$f_3(T) = \rho^2(T) |X| |Y| |Z|. \quad (9)$$

By narrowing this criteria for binary contexts, it is possible to obtain another maximising criteria for Model 7 **GF3**( $f$ ) from Veremyev et al. [18], p. 191.

For a bipartite graph  $G = (U, V, E)$  and its induced subgraph  $G[C_1, C_2]$ , function  $f$  is maximised over the density and size of biclique.

$$f(C_1, C_2) = \rho^2(G[C_1, C_2]) \cdot |C_1| \cdot |C_2| = \frac{(|\{(i, j) : i \in C_1, j \in C_2, (i, j) \in E\}|)^2}{|C_1| \cdot |C_2|}. \quad (10)$$

Using variables definitions from the previous model, we can rewrite function  $f$ :

$$f(C) = \frac{\left(\sum_{(i,j) \in E} y_{ij}\right)^2}{\left(\sum_{i \in U} u_i\right) \cdot \left(\sum_{j \in V} v_j\right)}$$

Since function  $f$  is multiplicative, the direct way to transform it to an additive function is logarithmisation:

$$\begin{aligned} f_{\log}(C) &= 2 \cdot \log |\{(i, j) : i \in C_1, j \in C_2, (i, j) \in E\}| - \log |C_1| - \log |C_2| = \\ &= 2 \cdot \log \left( \sum_{(i,j) \in E} y_{ij} \right) - \log \left( \sum_{i \in U} u_i \right) - \log \left( \sum_{j \in V} v_j \right). \end{aligned} \quad (11)$$

As in Model 1,

$$\sum_{i \in U} u_i = \sum_{n=\omega_i^{(1)}}^{\omega_u^{(1)}} n z_n^{(1)}, \quad \sum_{j \in V} v_j = \sum_{m=\omega_i^{(2)}}^{\omega_u^{(2)}} m z_m^{(2)}.$$

Now we introduce a new variable:  $w_k = 1 \Leftrightarrow |\{(i, j) : i \in C_1, j \in C_2, (i, j) \in E\}| = k$ , then  $\sum_{(i,j) \in E} y_{ij} = \sum_{(i,j) \in E} k w_k$ .

$$\begin{aligned} f_{\log}(C) &= 2 \cdot \log \left( \sum_{(i,j) \in E} k w_k \right) - \log \left( \sum_{n=\omega_i^{(1)}}^{\omega_u^{(1)}} n z_n^{(1)} \right) - \log \left( \sum_{m=\omega_i^{(2)}}^{\omega_u^{(2)}} m z_m^{(2)} \right) = \\ &= 2 \cdot \sum_{(i,j) \in E} \log(k) w_k - \sum_{n=\omega_i^{(1)}}^{\omega_u^{(1)}} \log(n) z_n^{(1)} - \sum_{m=\omega_i^{(2)}}^{\omega_u^{(2)}} \log(m) z_m^{(2)}. \end{aligned} \quad (12)$$

Obviously, the equality  $\log \left( \sum_{(i,j) \in E} k w_k \right) = \sum_{(i,j) \in E} \log(k) w_k$  because  $w_k$  is binary variable and  $\sum_{(i,j) \in E} w_k = 1$ . Thus there exists a unique number  $k^*$  such that  $w_{k^*} = 1$ .

It follows that  $\log \left( \sum_{(i,j) \in E} k w_k \right) = \log(k^*) = w_{k^*} \cdot \log(k^*) = \sum_{(i,j) \in E} \log(k) w_k$ . A

similar statement is true for  $\log \left( \sum_{n=\omega_i^{(1)}}^{\omega_u^{(1)}} n z_n^{(1)} \right)$  and  $\log \left( \sum_{m=\omega_i^{(2)}}^{\omega_u^{(2)}} m z_m^{(2)} \right)$ .

Without extra conditions on the sizes of vertex sets of a quasi-biclique and its minimum number of edges, the model has  $2 \cdot (|U| + |V|) + 2 \cdot |E|$  variables.



## Model 2

$$\begin{aligned}
& 2 \sum_{k=1}^{|E|} \log(k) \cdot w_k - \sum_{n=1}^{|U|} \log(n) z_n^{(1)} - \sum_{m=1}^{|V|} \log(m) z_m^{(2)} \xrightarrow{w, z^{(1)}, z^{(2)}} \max, \\
& \text{under conditions } \sum_{k=1}^{|E|} w_k \geq \gamma \sum_{n=\omega_l^{(1)}}^{\omega_u^{(1)}} \sum_{m=\omega_l^{(2)}}^{\omega_u^{(2)}} n \cdot m \cdot z_n^{(1)} \cdot z_m^{(2)}, \\
& \sum_{(i,j) \in E} y_{ij} = \sum_{k=1}^{|E|} k \cdot w_k, \\
& \sum_{i \in U} u_i = \sum_{n=\omega_l^{(1)}}^{\omega_u^{(1)}} n z_n^{(1)}, \sum_{j \in V} v_j = \sum_{m=\omega_l^{(2)}}^{\omega_u^{(2)}} m z_m^{(2)}, \\
& \sum_{k=1}^{|E|} w_k = 1, \sum_{n=1}^{|U|} z_n^{(1)} = 1, \sum_{m=1}^{|V|} z_m^{(2)} = 1, \\
& \forall i \in U : u_i \in \{0, 1\} \forall j \in V : v_j \in \{0, 1\} \\
& \forall i < j, (i, j) \in E : y_{ij} \in \{0, 1\}, \forall k \in \{1, \dots, |E|\} : w_k \in \{0, 1\}, \\
& \forall n \in \{\omega_l^{(1)}, \dots, \omega_u^{(2)}\} : z_n^{(1)} \geq 0, \forall m \in \{\omega_l^{(2)}, \dots, \omega_u^{(2)}\} : z_m^{(2)} \geq 0.
\end{aligned}$$

*Remark* In order to simplify the model, we can add extra constraints for variables  $w_k$ ,  $k \in \{1, \dots, |E|\}$ . Let  $k$  be a possible number of edges in a quasi-biclique, then

1.  $k \leq \omega_u^{(1)} \cdot \omega_u^{(2)}$ .
2. If  $\gamma \cdot \omega_l^{(1)} \cdot \omega_l^{(2)} \leq |E| \Rightarrow k \geq \gamma \cdot \omega_l^{(1)} \cdot \omega_l^{(2)}$ .
3. Let us consider  $U'$  such that  $|U'| = \omega_l^{(1)}$  and  $\forall u \in U' \deg(u) \leq \min_{x \in U \setminus U'} \{\deg(x)\}$ . That is,  $U'$  is a subset of  $U$  with the minimum possible size and with all smallest degree vertices with respect to  $U$ . Then  $k \geq \gamma \sum_{u \in U'} \deg(u)$ .
4. Similarly, for  $V' \subseteq V$ :  $|V'| = \omega_l^{(2)}$  and  $\forall v \in V' \deg(v) \leq \min_{x \in V \setminus V'} \{\deg(x)\}$ , then  $k \geq \gamma \sum_{v \in V'} \deg(v)$ .

## 4 Datasets

Datasets for testing the performance of the algorithms are mainly taken from [2, 5].

1. Southern Women:  $|U| = 18$ ,  $|V| = 14$ ,  $|E| = 89$  edges, a classic ethnographic dataset with a bipartite graph of 18 women, which met in a series of 14 informal social events [6].

2. Divorce in the US:  $|U| = 9$ ,  $|V| = 50$  vertices, and  $|E| = 225$  edges. This graph describes the particular causes of divorce in the United States.
3. Dutch Elite:  $|U| = 3810$ ,  $|V| = 937$  vertices, and  $|E| = 5221$  edges. This graph describes the connections between people and the most important for the Netherlands government administrative authorities.
4. Dutch Elite (TOP-200):  $|U| = 200$ ,  $|V| = 395$  vertices, and  $|E| = 877$  edges. The list of people in the first partition of the graph consists of the most influential persons regarding their membership in administrative authorities.
5. Movie-Lens (ml-latest-small):  $|U| = 99125$ ,  $|V| = 50$  vertices, and  $|E| = 20340$  edges, a bipartite graph of “movie-genre” relation from Movie-Lens project [7].

## 5 Experimental Verification

### 5.1 Implementation Description

The greedy algorithm of searching for maximal  $\gamma$ -quasi-biclique in a bipartite graph was implemented in Python 2.7. The MIP models were implemented with the optimisation package CPLEX, created by IBM. All computations were carried out on a laptop with macOS operating system, 2.7 GHz Intel Core i5 processor, and RAM 8 GB 1867 MHz.

The search for solutions in the CPLEX package was performed by means of the branch-and-cut method, which is similar to the branch-and-bound algorithmic approach. The method uses a search tree, where each node represents a subproblem that needs to be solved and possibly analysed further.

The BRANCH procedure creates two new nodes from the active parent node. Generally, at this point, the boundaries of one variable are applied and stored for the current node and all its child nodes. In its turn, the CUT procedure adds a new constraint to the model. As a result of any cut, the solution space for the subproblems, which are presented in the nodes, is reduced, and the number of branches needed to process decreases. CPLEX processes active nodes in the tree until no more active nodes are available or a certain limit is reached.<sup>1</sup>

The standard solution with the CPLEX software package assumes only one of the optimal solutions as the answer. However, in CPLEX it is possible to obtain a set of optimal solutions using the *solution pool* method, which allows one to find and store several solutions of MIP models.

The generation of multiple solutions works in two steps. The first step is identical to the usual solution search using the CPLEX software package. At this step, the algorithm finds the only optimal solution to the integer programming problem. It also saves nodes in the search tree that could potentially be useful; for example, if

---

<sup>1</sup>CPLEX user manual: [https://www.ibm.com/support/knowledgecenter/SSSA5P\\_12.8.0/ilog.odms.cplex.help/CPLEX/homepages/usrmanplex.html](https://www.ibm.com/support/knowledgecenter/SSSA5P_12.8.0/ilog.odms.cplex.help/CPLEX/homepages/usrmanplex.html).

not all the variable constraints are taken into account or if all the nodes contain a suitable value, but the target function is not optimal.

In the second step, using previously calculated and stored information in the first stage, several solutions are generated, and the tree is traversed again, in particular within the branches rooted from the additional nodes stored in the first stage.

### 5.2 Illustrative Examples

On a toy example of a graph with 12 vertices, we consider the search results for maximal  $\gamma$ -quasi-bicliques,  $\gamma = 0.8$ , using Models 1 and 2, respectively (Fig. 1).

The results for both models are the same (with respect to the solutions' output order). Even for this small-sized problem, the time is tangible: the computations with Model 1 took 2.16 s, and for Model 2, it was 2.94 s. A comparison of the executed models and the greedy algorithm in terms of computational time is given below for the selected bipartite graphs.

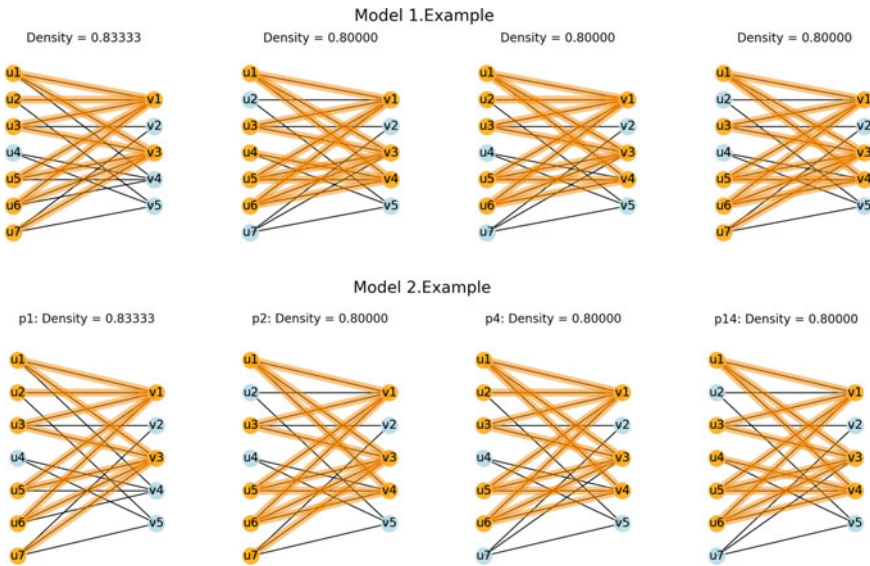


Fig. 1 The results of search for quasi-bicliques using Models 1 and 2

### 5.3 Comparison of Algorithms

The algorithm of searching for the maximal quasi-biclique using the CPLEX software package was implemented for Models 1 and 2 (see Sect. 3) and compared with the GREEDY algorithm from [12] (let us denote it as Greedy Algorithm).

There are no comparison results presented for the model **F3** [18]: despite its fast work, the algorithm based on this model chose quasi-bicliques of very small size and maximum density (i.e., bicliques). This phenomenon is rather expected since the model **F3** implies a completely different function of the density of the subgraph. Therefore, the comparison, in this case, is irrelevant. The description of Complete QB in [17] lacks of important implementation details.

The weakness of the constructed MIP models was identified during the finding solution. Since the problem of enumerating all maximal quasi-bicliques in practice requires considerable time, the software package can discard some solutions, if it has found quite a few optimal ones already. First of all, the search is carried out among unbalanced quasi-bicliques (no constraints on the approximately equal size of the quasi-clique partitions have been given). For large graphs, this means that the number of vertices in one of the parts of the found optimal solution may exceed the number of vertices in the second part by hundreds of times or more.

This issue can be addressed in two ways. Firstly, one can set roughly equal limits on the size of the partitions. Secondly, it is possible to adapt the model for finding an almost balanced quasi-bicliques, which means that sizes of partitions of a quasi-biclique differ by  $\theta$ . To do this, the following conditions should be added to Model 1 or Model 2:

$$\sum_{n=\omega_l^{(1)}}^{\omega_u^{(1)}} z_n^{(1)} \geq (1 - \theta) \sum_{m=\omega_l^{(2)}}^{\omega_u^{(2)}} z_m^{(2)}, \quad (13)$$

$$\sum_{n=\omega_l^{(1)}}^{\omega_u^{(1)}} z_n^{(1)} \leq (1 + \theta) \sum_{m=\omega_l^{(2)}}^{\omega_u^{(2)}} z_m^{(2)}. \quad (14)$$

Models with additional conditions (13) and (14) have not been tested.

It has also been noted that small-sized quasi-bicliques can be useless in practice, but their recalculation is costly. Therefore, for each dataset, we can establish minimum bounds on the size of a quasi-biclique (of the order of the smallest vertex degree with respect to the graph partitions).

The results of the algorithms execution are presented in Table 1 for  $\gamma = 0.6$ , Table 2 for  $\gamma = 0.7$ , and in Table 3 for  $\gamma = 0.8$ .<sup>2</sup> For each algorithm its main parameters are indicated: the algorithm running time (time), the number of found maximum quasi-bicliques (count), and the maximum size of the found solution.

---

<sup>2</sup>The size column in Table 3 shown as the result of summation  $|U'|$  and  $|V'|$ .

**Table 1** Results of maximum  $\gamma$ -quasi-biclique search. Parameters:  $\gamma = 0.6$ 

Data	Model 1			Model 2			Greedy algorithm		
	Time	Count	Size	Time	Count	Size	Time	Count	Size
Southern women	678 ms	4	(18, 4)	801 ms	2	(18, 4)	234 ms	4	(17, 5)
Divorce in US	1.23 s	1	(4, 50)	3.38 s	1	(4, 50)	360 ms	1	(2, 46)
Dutch Elite (top200)	7602 s	2	(26, 1)	181 s	1	(11, 3)	3 s	1	(10, 3)
Dutch Elite	–	–	–	6968 s	1	(45, 2)	1954 s	1	(40, 2)
Movie-Lens (small)	28068 s	2	(692, 2)	13851 s	5	(900, 3)	5976 s	2	(754, 2)

**Table 2** The results of maximum  $\gamma$ -quasi-biclique search for  $\gamma = 0.7$ 

Data	Model 1			Model 2			Greedy algorithm		
	Time	Count	Size	Time	Count	Size	Time	Count	Size
Southern women	1.29 s	1	(16, 3)	1.11 s	1	(10, 6)	309 ms	1	(16, 2)
Divorce in US	1.56 s	1	(2, 45)	2.66 s	3	(5, 36)	320 ms	1	(2, 28)
Dutch Elite (top200)	8497 s	1	(23,1)	1668 s	3	(10,3)	1.63 s	1	(10, 3)
Dutch Elite	–	–	–	6166 s	1	(20, 2)	1511 s	1	(20, 1)
Movie-Lens (small)	–	–	–	10719 s	6	(800, 3)	–	–	–

**Table 3** The results of maximum  $\gamma$ -quasi-biclique search for  $\gamma = 0.8$ 

Data	Model 1			Model 2			Greedy algorithm		
	Time	Count	Size	Time	Count	Size	Time	Count	Size
Divorce in US	8.53 s	1	38	1.7 s	2	33	313 ms	1	25
Dutch Elite (top200)	–	–	–	4834 s	2	13	2.5 s	1	13
Dutch Elite	–	–	–	7129 s	1	47	1718	1	21
Movie-Lens (small)	–	–	–	9046 s	2	445	–	–	–

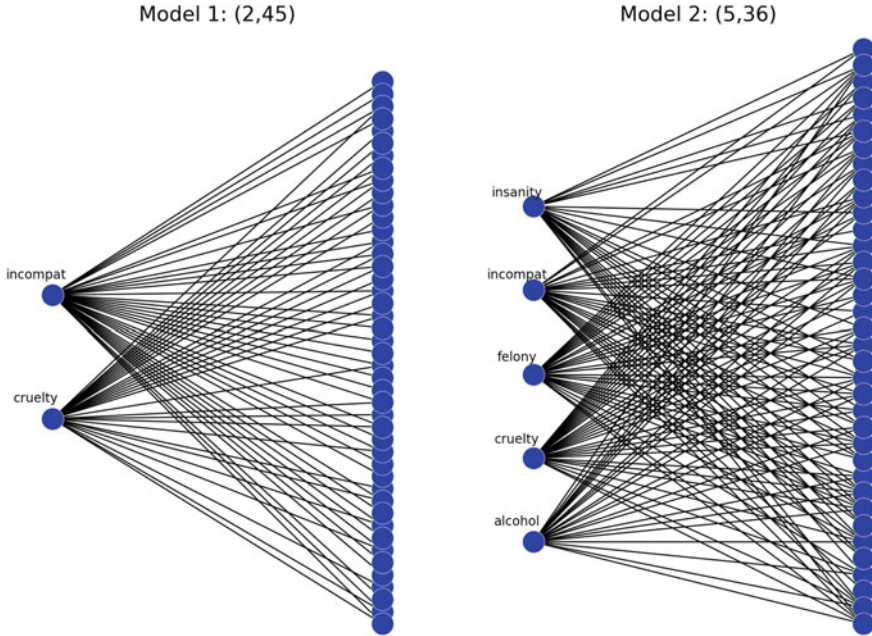
Two found  $\gamma$ -quasi-bicliques for the dataset Divorce in US are shown in Fig. 2.

Dashes (“–”) in the following tables mean that the algorithm worked 10 hours and did not find a solution. If one of the partitions of the maximal quasi-biclique has a unit size, this is marked in the table as  $(U', V')$ , where  $U'$  and  $V'$  are the sizes of the partitions.

## 6 Results and Conclusions

One can note, that mixed linear programming models work an order of magnitude slower than the greedy algorithm by Liu et al. [12], but they find more quasi-cliques and generally each of them has a larger size.

For small graphs, the time for finding the solution by the considered models is acceptable. Model 1 contains a fewer number of variables that must be optimised,



**Fig. 2** Quasi-bicliques obtained by the studied MIP models for the dataset Divorce in US with  $\gamma = 0.7$

but its maximisation criterion is costly for large graphs. Thus, on large-sized graphs Model 1 works too long (more than 10 hours), especially for high  $\gamma$  density thresholds. The dependence of the speed and quality of processing on  $\gamma$  is also apparent for two other algorithms: for high thresholds on density, those methods work longer since the number of possible optimal solutions to the problem is reduced. Model 2 on similar graphs showed better results, but the processing time is still quite large. For *DutchElite* data with a large number of vertices and a small number of edges, MIP-based algorithms work much longer than on more dense graphs.

If we consider the results, not in terms of speed, but terms of quality, then Model 2 was the best one. This model produced more unique and larger quasi-bicliques than other algorithms.

The following ways of future work seems to be relevant: (1) further improvements of the proposed models by establishing tighter bounds for different constraints and using optimisation tricks; (2) exploration of new optimisation criteria; (3) comparison of different MIP solvers with the state-of-the-art approaches of searching for quasi-bicliques in a larger set of experiments.

Another interesting avenue for research could be a study on connection between various approximations of formal concepts (fault-tolerant concepts [4] and object-attribute biclusters [8, 10]), Boolean matrix factorization [3, 13], and quasi-bicliques.

**Acknowledgements** The work of Dmitry I. Ignatov shown in all the sections, except 5 and 6, has been supported by the Russian Science Foundation grant no. 17-11-01276 and performed at St. Petersburg Department of Steklov Mathematical Institute of Russian Academy of Sciences, Russia. The authors would like to thank Boris Mirkin, Vladimir Kalyagin, Panos Pardalos, and Oleg Prokopyev for their piece of advice and inspirational discussions. Last but not least, we are thankful to anonymous reviewers for their useful feedback.

## References

1. Abello, J., Resende, M.G.C., Sudarsky, S.: Massive quasi-clique detection. In: Rajsbaum, S. (ed.) *LATIN 2002: Theoretical Informatics*, pp. 598–612. Springer, Berlin (2002)
2. Batagelj, V., Mrvar, A.: Pajek. In: *Encyclopedia of Social Network Analysis and Mining*, pp. 1245–1256 (2014). [https://doi.org/10.1007/978-1-4614-6170-8\\_310](https://doi.org/10.1007/978-1-4614-6170-8_310)
3. Belohlávek, R., Outrata, J., Trnečka, M.: Factorizing boolean matrices using formal concepts and iterative usage of essential entries. *Inf. Sci.* **489**, 37–49 (2019). <https://doi.org/10.1016/j.ins.2019.03.001>
4. Besson, J., Robardet, C., Boulicaut, J.: Mining a new fault-tolerant pattern type as an alternative to formal concept discovery. In: *Conceptual Structures: Inspiration and Application*, 14th International Conference on Conceptual Structures, ICCS 2006, Aalborg, Denmark, July 16–21, 2006, Proceedings, pp. 144–157 (2006). [https://doi.org/10.1007/11787181\\_11](https://doi.org/10.1007/11787181_11)
5. Borgatti, S.P., Everett, M.G., Freeman, L.C.: UCINET. In: *Encyclopedia of Social Network Analysis and Mining*, pp. 2261–2267 (2014). [https://doi.org/10.1007/978-1-4614-6170-8\\_316](https://doi.org/10.1007/978-1-4614-6170-8_316)
6. Freeman, L.C.: Finding social groups: A meta-analysis of the southern women data. In: Breiger, R., Carley, K., Pattison, P. (eds.) *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*, National Academies Press (2003)
7. Harper, F.M., Konstan, J.A.: The movielens datasets: history and context. *ACM Trans. Interact. Intell. Syst.* **5**(4):19:1–19:19 (2015). <https://doi.org/10.1145/2827872>
8. Ignatov, D.I., Kuznetsov, S.O., Poelmans, J.: Concept-based biclustering for internet advertisement. In: 12th IEEE International Conference on Data Mining Workshops, ICDM Workshops, Brussels, Belgium, 10 Dec 2012, pp. 123–130 (2012). <https://doi.org/10.1109/ICDMW.2012.100>
9. Ignatov, D.I., Gnatyshak, D.V., Kuznetsov, S.O., Mirkin, B.G.: Triadic formal concept analysis and triclustering: searching for optimal patterns. *Mach. Learn.* **101**(1), 271–302 (2015). <https://doi.org/10.1007/s10994-015-5487-y>
10. Ignatov, D.I., Semenov, A., Komissarova, D., Gnatyshak, D.V.: Multimodal clustering for community detection. In: *Formal Concept Analysis of Social Networks*, pp. 59–96 (2017). [https://doi.org/10.1007/978-3-319-64167-6\\_4](https://doi.org/10.1007/978-3-319-64167-6_4)
11. Liu, H.B., Liu, J., Wang, L.: Searching maximum quasi-bicliques from protein-protein interaction network. *J. Biomed. Sci. Eng.* **1**(03), 200 (2008a)
12. Liu, X., Li, J., Wang, L.: Quasi-bicliques: complexity and binding pairs. In: Hu, X., Wang, J. (eds.) *Computing and Combinatorics*, pp. 255–264. Springer, Berlin Heidelberg, Berlin, Heidelberg (2008b)
13. Miettinen, P.: Fully dynamic quasi-biclique edge covers via boolean matrix factorizations. In: *Proceedings of the Workshop on Dynamic Networks Management and Mining*, pp. 17–24. ACM, New York, NY, USA, *DyNetMM '13* (2013). <https://doi.org/10.1145/2489247.2489250>
14. Mirkin, B.G., Kramarenko, A.V.: Approximate bicluster and tricluster boxes in the analysis of binary data. In: *Rough Sets, Fuzzy Sets, Data Mining and Granular Computing - 13th International Conference, RSFDGrC 2011*, pp. 248–256. Moscow, Russia, 25–27 June 2011. Proceedings (2011). [https://doi.org/10.1007/978-3-642-21881-1\\_40](https://doi.org/10.1007/978-3-642-21881-1_40)
15. Pattillo, J., Veremyev, A., Butenko, S., Boginski, V.: On the maximum quasi-clique problem. *Discret. Appl. Math.* **161**(1):244–257 (2013). <https://doi.org/10.1016/j.dam.2012.07.019>

16. Peeters, R.: The maximum edge biclique problem is np-complete. *Discret. Appl. Math.* **131**(3), 651–654 (2003). [https://doi.org/10.1016/S0166-218X\(03\)00333-0](https://doi.org/10.1016/S0166-218X(03)00333-0)
17. Sim, K., Li, J., Gopalkrishnan, V., Liu, G.: Mining maximal quasi-bicliques to co-cluster stocks and financial ratios for value investment. In: Sixth International Conference on Data Mining (ICDM'06), pp. 1059–1063 (2006). <https://doi.org/10.1109/ICDM.2006.111>
18. Veremyev, A., Prokopyev, O.A., Butenko, S., Pasiliao, E.L.: Exact mip-based approaches for finding maximum quasi-cliques and dense subgraphs. *Comp. Opt. Appl.* **64**(1), 177–214 (2016). <https://doi.org/10.1007/s10589-015-9804-y>
19. Wang, L.: Near optimal solutions for maximum quasi-bicliques. *J. Comb. Optim.* **25**(3), 481–497 (2013). <https://doi.org/10.1007/s10878-011-9392-4>



# Graph Clustering Via Intra-Cluster Density Maximization



Pierre Miasnikof, Leonidas Pitsoulis, Anthony J. Bonner, Yuri Lawryshyn  
and Panos M. Pardalos

**Abstract** Graph clustering, also often referred to as network community detection, is the process of assigning common labels to vertices that are densely connected to each other but sparsely connected to the rest of the graph. There are many different approaches to clustering in the literature. However, in this article, we formulate the clustering problem as a combinatorial optimization problem. Our main contribution is a novel problem formulation that maximizes intra-cluster density, a statistically meaningful quantity. It requires the number of clusters, a softbound on cluster size and a penalty coefficient as parameter inputs. More importantly, it is designed to prevent common degeneracies, like the so-called “mega-clusters”. We end with some suggestions on numerical solution techniques and note that an ensemble-like optimization routine seems promising.

## 1 Introduction

Graph clustering is the process of grouping of vertices into densely connected subsets of vertices that have sparse connections to other subsets of vertices. These subsets are referred to as clusters. The process of assigning cluster labels to vertices, grouping them into clusters, is referred to as graph clustering (or network community detection).

---

Note on vocabulary: Although there are subtle differences between the concepts of graph clustering and network community detection, in this document we use the two interchangeably.

---

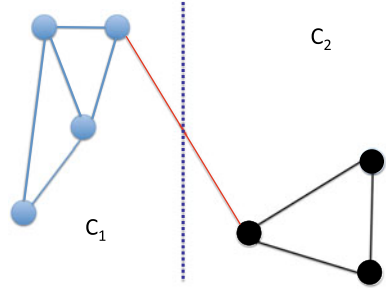
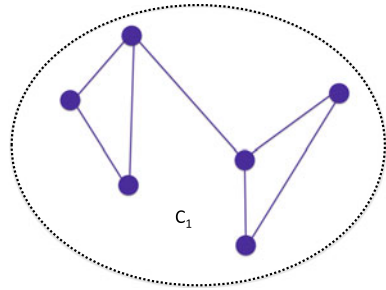
P. Miasnikof (✉) · A. J. Bonner · Y. Lawryshyn  
University of Toronto, Toronto, ON, Canada  
e-mail: [p.miasnikof@mail.utoronto.ca](mailto:p.miasnikof@mail.utoronto.ca)

L. Pitsoulis  
Aristotle University of Thessaloniki, Thessaloniki, Greece

P. M. Pardalos  
University of Florida, Gainesville, FL, USA

National Research University HSE, Nizhny Novgorod, Russian Federation

© Springer Nature Switzerland AG 2020  
I. Bychkov et al. (eds.), *Network Algorithms, Data Mining,  
and Applications*, Springer Proceedings in Mathematics & Statistics 315,  
[https://doi.org/10.1007/978-3-030-37157-9\\_3](https://doi.org/10.1007/978-3-030-37157-9_3)

**Fig. 1** Good clustering**Fig. 2** Bad clustering

The definition of graph clusters (or network communities) remains a matter of debate in the literature (e.g., [10, 25]). However, most authors agree that a cluster can be described as a dense subgraph within a sparse graph (e.g., [8, 22, 23, 28], we quote these authors, but their definition is very common throughout the literature).

In Fig. 1, we see properly labeled (clustered) vertices. On either side of the dotted line, we observe densely connected vertices and very sparse (only one edge) connections between each cluster ( $C_1$  and  $C_2$ ). On the other hand, in Fig. 2, we observe two cliques connected to each other by only a single edge being labeled as all belonging to the same cluster. Clearly, in this example, labeling each triangle as belonging to a separate cluster would be more reasonable.

It is also important to draw a distinction between clusters and cliques (e.g., “the clique problem” or “maximal clique problem” [27]), since clusters may or may not be cliques. In fact, according to Fortunato and Hric [10], clusters typically are not cliques.

In this article, we formulate the clustering problem as a mean intra-cluster density maximization problem. While most authors who have used optimization-based approaches maximize modularity, we maximize a statistically meaningful and robust quantity. Modularity is known to be fragile and problematic in many ways and has been shown to be less responsive to graph structure than mean intra-cluster density (e.g., [9, 17]). To the best of our knowledge, this formulation is novel and has not been used in the past.

## 2 Previous Work

A complete review of the very rich graph clustering literature is beyond the scope of this article. However, we note many authors have approached graph clustering using various techniques. There exist many competing formulations and solution techniques in the literature. The main ones are

- Spectral (e.g., [16]),
- Markov (e.g., [5]), and
- Optimization.
  - Modularity maximization (e.g., [1, 10, 18]),
  - Other objective functions (e.g., [6, 7, 15]).

While there are many competing approaches to clustering, Fortunato and Hric claim that determining which is the best clustering technique under all circumstances is not a clear-cut case [10] and that most algorithms cannot adapt to every dataset and consistently provide superior clusterings.

It is also important to note that spectral methods are very costly and do not scale well at all. Spectral clustering methods require eigendecomposition of the graph's Laplacian, a projection into a metric space and the application of a k-means algorithm. This difficulty with scaling has also been noted by Schaeffer, in her review [25]. While there are techniques that allow us to take advantage of sparsity and the partial computation of eigenvalues, the combined costs of such an approach make it prohibitive for large graphs. Even in more recent work, where authors claim their algorithms are “faster and more accurate” than legacy techniques in this area, they still require onerous computations (e.g., [13]). Finally, it should also be noted that Fortunato and Hric describe spectral methods as inaccurate in the case of sparse graphs [10] and that clusterable graphs are typically sparse.

As for Markov-based techniques, they require simulating a random walk over the graph (i.e., matrix multiplications), as well as multiple element-wise and row operations. While Markov clustering does not require the number of clusters as input, it relies on costly operations. Also, as highlighted by Fortunato and Hric, algorithms that do not require the number of clusters as an input parameter have been found to be less accurate than those that do require it [10].

On the other hand, optimization-based approaches lend themselves very well to approximate solution techniques, which carry a lower computational cost. Indeed, because of the NP-hardness of the problem [8, 25], solving these and other types of combinatorial optimization problems is often successfully done via (meta-)heuristic solution techniques (e.g., [21]), which explore subsets of the solution space. In the specific case of graph clustering, many authors have made use of meta-heuristic optimization techniques (e.g., [1, 14, 18, 26]), in order to find approximate solutions and overcome the NP-hard nature of the problem. Additionally, meta-heuristic optimization techniques are easily parallelizable and well suited to implementation on high-performance computing platforms.

### 3 Problem Formulation

Using the almost universally accepted definition that a cluster is a dense subgraph within a sparser graph, we formulate an optimization problem that assigns cluster labels to nodes in a manner that yields a high mean intra-cluster density. Our goal is to assign cluster labels so that the mean intra-cluster density is higher than the graph's global density.

Recall, (sub)graph density is defined as the ratio of the total number of edges or sum of edge weights ( $|E_i|$ ) over the maximum possible number of edges in a given (sub)graph, given the number of vertices ( $n_i$ ):

$$\kappa_i = \frac{|E_i|}{0.5 \times n_i(n_i - 1)}$$

In the case of an unweighted graph, the quantity  $\kappa_i$  represents the empirical estimate of the probability that two vertices are connected. In the case of a weighted graph, it represents the mean edge weight. This definition of density and its interpretations apply to graphs without multiple edges or self-loops.

We assert that a graph whose vertices have been well clustered (labeled) will display a mean intra-cluster density, a quantity we call  $\bar{K}_{\text{intra}}$  [17], that is higher than the graph's global density. This assertion was initially presented and detailed in Miasnikof et al. [17]. Mathematically, we assert that, for a properly clustered graph with  $|C|$  clusters, the following inequality should hold:

$$\bar{K}_{\text{intra}} = \frac{1}{|C|} \sum_{i=1}^{|C|} \kappa_i > K = \frac{|E|}{0.5 \times N(N - 1)}$$

where

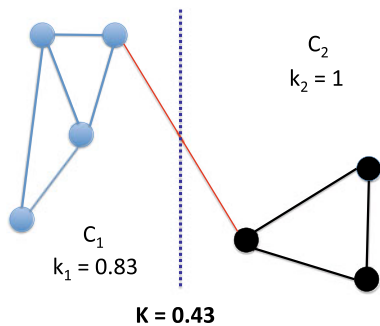
$$N = \sum_{i=1}^{|C|} n_i$$

$$\kappa_i = \frac{|E_{ii}|}{0.5 \times n_i(n_i - 1)}$$

In the equations above,  $|E|$  is the total number (weight) of edges on the graph and  $N$  is the total number of vertices. Similarly, we compute  $\kappa_i$  for each cluster, where  $|E_{ii}|$  is the total number (weight) of edges connecting two vertices in cluster “ $i$ ” and  $n_i$  is the number of vertices in that same cluster.

For example, in Fig. 3, we show an example of a well-labeled graph. The graph's global density is  $K = 0.43$ , while the mean intra-cluster density (assuming only two clusters) is  $\bar{K}_{\text{intra}} = \frac{1}{2}(\kappa_1 + \kappa_2) = \frac{1}{2}(0.83 + 1) = 0.915$ . Here, our earlier assertion regarding the inequality  $K < \bar{K}_{\text{intra}}$  clearly holds and can also be easily visualized in this small example.

**Fig. 3** Global and intra-cluster densities



### 3.1 Optimization Problem: Mean Intra-Cluster Density Maximization

We opt for an objective function inspired by the quadratic maximization formulation presented by Fan and Pardalos [6], which we modify slightly. These authors maximize the following objective function:

$$\begin{aligned} & \underset{x_{i,k}, x_{j,k}}{\text{maximize}} \left\{ z = \sum_{i,j,k} w_{i,j} x_{i,k} x_{j,k} \right\} & (1) \\ & (x_{i,k} \in \{0, 1\}, w_{i,j} \in \mathbb{R}_+, \forall i, j \in V, \forall k \in C) \end{aligned}$$

In this function,  $x_{i,k}$  is an indicator variable that is equal to 1 if vertex  $i$  is assigned to cluster “ $k$ ” and  $w_{i,j}$  is the weight of the edge connecting vertices  $i$  and  $j$ . The summation iterates over every vertex( $i$ )-vertex( $j$ )-cluster( $k$ ) triplet, given a vertex set  $V$  and a set of clusters  $C$ .

In our modified formulation, we maximize  $\bar{K}_{\text{intra}}$  and also add a penalty function,  $P(M)$ , that penalizes very large or very small clusters and prevents degenerate solutions:

$$\underset{x_{i,k}, x_{j,k}}{\text{maximize}} \left\{ \sum_{k=1}^{|C|} \frac{|E_{k,k}|}{0.5 \times n_k(n_k - 1)} - \lambda P_k(M) \right\} \quad (2)$$

where

$$n_k = \sum_j x_{j,k} \quad (\text{num vertices in cluster ‘}k\text{’}) \quad (3)$$

$$|E_{k,k}| = \sum_{i,j} w_{i,j} x_{i,k} x_{j,k} \quad (\text{sum edges in cluster ‘}k\text{’}) \quad (4)$$

$$P_k(M) = \max\{0, n_k - M\} \quad (5)$$

OR,

$$P_k(M) = (n_k - M)^2 \quad (6)$$

$M$  is a parameter input specifying a softbound for reasonable cluster sizes. It is determined by judgement and domain expertise. Similarly,  $\lambda$  is a penalty coefficient which is also determined through domain expertise. Our penalty function is a type of ( $L1$  or  $L2$ ) regularization [11]. It favors clusters with fewer than  $M$  vertices in the first case (Eq. 5) and clusters with roughly  $M$  vertices in the second case (Eq. 6), but keeps larger and smaller sized clusters within the feasible set.

Putting it all together,

$$\underset{x_{i,k}, x_{j,k}}{\text{maximize}} \left\{ \sum_{k=1}^{|C|} \left[ \sum_{i,j} \left( \frac{w_{i,j} x_{i,k} x_{j,k}}{0.5 \times n_k (n_k - 1)} - \lambda P_k(M) \right) \right] \right\} \quad (7)$$

Our model maximizes the mean probability of connection/edge weight within clusters (mean density). In the case of unweighted graphs, it maximizes the similarity between our clusters and cliques, on average (and cliques are arguably very strong clusters). It also avoids mega-clusters, the tendency displayed by many clustering techniques to create extremely large uninformative clusters and degenerately small clusters as well. The numerator in the fractions corresponds to the quantity “ $z$ ” in Eq. 1. Dividing our numerator by the denominator yields a measure of connection density within each cluster. The denominator also acts as a natural form of cardinality constraint, limiting the number of vertices assigned to each cluster. The penalty function also enforces a soft upper bound on cluster cardinality, in the case of Eq. 5 or a soft cardinality upper and lower bound in the case of Eq. 6. Naturally, we also add constraints forcing each vertex to belong to exactly one cluster.

As mentioned previously, our formulation does not require cardinality constraints for the number of vertices in each cluster, as in the Fan and Pardalos formulation [6]. Together the denominators and penalty function in the inner summation ensure our cluster sizes remain reasonable and faithful to the graph’s structure. We do, however, impose a set of constraints that ensure all vertices are assigned to exactly one cluster:

$$\sum_{k=1}^{|C|} x_{i,k} = 1 \quad \forall i \in V \quad (V \text{ is the set of all vertices}) \quad (8)$$

### 3.2 Overcoming Common Degeneracies

Our formulation is designed to avoid two common degeneracies observed in graph clustering, the appearance of uninformative mega-clusters and “garbage collector clusters”. These degeneracies are especially frequent and exacerbated when clustering is conducted using optimization-based approaches that maximize a sum of non-negative terms.

### 3.2.1 Mega-Clusters

When attempting to cluster vertices, it is not uncommon for an algorithm to group all nodes together into only one or a few very large clusters, leaving the vast majority of clusters very sparsely populated. This is particularly true in the case of an optimization-based approach that maximizes modularity, a widely used clustering quality measure originally introduced by Newman and Girvan in 2004 [19] (Modularity maximization is by far the most common optimization-based approach to clustering.)

This grouping of vertices into a few very large clusters is due to a well-known and documented degeneracy of modularity, known as resolution limit. Fortunato and Bathélemy [9] describe how any clustering quality function that is defined as a sum of quality measures (scores) of individual clusters, as is the case with modularity, suffers from this limit. The authors describe how terms from smaller clusters are dominated by terms from larger clusters. Because the smaller clusters' contribution to the sum is dominated by the larger clusters, the final result is also dominated, which leads to the resolution limit.

Our formulation has two features that prevent this domination of large clusters. First, the  $\bar{K}_{\text{intra}}$  portion of our objective function contains a denominator which is proportional to the number of vertices. The resulting ratio ensures independence between the score of each cluster and its number of vertices. Second, our formulation imposes a penalty to the scores of very small or very large clusters, which may even assign a negative score to such clusters.

### 3.2.2 Garbage Collector Clusters

The garbage collector cluster is a special case of a mega-cluster. When maximizing a sum of unweighted cluster-level quality measurements (nonnegative numbers), an algorithm can return a solution which contains a few very small but very strong very dense clusters with very-high-quality scores and include all remaining vertices in one or a few very large cluster with a very low score. The risk of this pitfall is exacerbated in the case of  $\bar{K}_{\text{intra}}$ , since all cluster scores are weighted equally, regardless of size. For example, in a graph with three clusters, a clustering algorithm could return two clusters each containing only two connected vertices and lump all remaining vertices in one large poorly connected cluster. In such a case, the aggregate score would be high, because two of the three clusters would have a very high score.

To prevent this situation, many formulations impose cardinality constraints on the clusters (e.g., [6]). We impose a penalty which can potentially assign a negative score to an unreasonably small or large cluster with low density.

## 4 Numerical Experiments

Numerical implementation is still work in progress and is beyond the scope of this article. We are still in the process of implementing the methods discussed here. We have, however, very early-stage experimental results and find our experiences with an ensemble-like routine to be of interest.

As mentioned previously, the clustering problem is NP-hard. Solving it for a graph of even just moderate size is impossible to do exactly. For this reason, we employ meta-heuristic optimization techniques. When it is hard or impossible to obtain analytical solutions to an optimization problem, there exists a vast array of global optimization search techniques to approximate a globally optimal solution. While each of these techniques has its own specificities, they all explore the feasible set of a problem in some systematic manner.

In our experiments, we use a greedy algorithm and simulated annealing, both separately and in combination, to optimize the  $L1$  regularized version of our objective function. We experiment with a greedy algorithm alone, simulated annealing alone, and an ensemble-like routine which uses the best solution yielded by the greedy algorithm as a starting point for simulated annealing.

We deliberately keep graph sizes small to obtain a proof of concept, but do explore sensitivity to graph size (number of edges and vertices). We are still pursuing our experiments on larger graphs, refining our algorithms, and implementing them on high-performance parallel computing platforms, but find the results yielded by our ensemble-like routine worthy of mention.

### 4.1 *Meta-Heuristic Techniques*

As mentioned earlier, we use a greedy algorithm, simulated annealing and an ensemble-like routine which combines both techniques. Details about each technique are provided in the following sections.

#### 4.1.1 **Simulated Annealing**

Simulated annealing [2, 3, 24] is a well-known meta-heuristic optimization technique. It does not systematically apply a greedy logic to determine a move from the current solution to the next one. In simulated annealing, we search the feasible set by moving from a current solution to a new one according to the following set of search rules. If the new solution in our search improves the objective function, then the move is automatically accepted. In the case where the new solution does not improve the objective function, it is accepted if a random draw is lower than an evolving probability of acceptance. This acceptance procedure allows simulated annealing to break out of local optima.



In our implementation, we begin with a simulated annealing algorithm with a random starting point, as is common practice. We also use the best solution returned by our greedy algorithm as a starting point. Finally, using both starting points, we experiment with 1 million and 5 million runs, to examine sensitivity to the number of search iterations.

### 4.1.2 Greedy Algorithm

We apply a greedy algorithm to vertices sorted in depth-first order. Our decision to sort vertices in a depth-first order is motivated by the work of Creusefond et al. [4] who used a lexicographic depth-first ordering (LexDFS) to detect clusters. While we did not use LexDFS but used a plain depth-first traversal to order vertices instead, we did notice better results than when we greedily assigned vertices in random order. The steps in our algorithm are detailed below:

- **INPUT:**

- Graph: “ $G$ ”,
- Number of clusters: “ $k$ ”,
- Percent of nodes to be randomly assigned: “ $pRand$ ”, and
- Number of repeated runs: “ $R$ ”.

- **OUTPUT:**

- Vertex-cluster assignments
- **Steps:**
  - Randomly assign  $\lfloor pRand \times \text{num vertices} \rfloor$  uniformly to each cluster;
  - Sort remaining vertices using depth-first ordering with a random starting point;
  - While remaining vertices list is not empty:
- Assign vertex to cluster where objective function improves the most;
- **Steps:**
  - Record objective function value;
  - Repeat ‘ $R$ ’ times and select best run;
  - **Return** best vertex-cluster assignments.

### 4.1.3 Ensemble-Like Learning Routine

Here, we are guided by the work of Ovelgönne and Geyer-Schulz [20], who combined weak clustering algorithms to obtain better results. Our routine consists of using the best solution obtained by a greedy algorithm that was run for 50,000 iterations as a starting point for our simulated annealing algorithm. We use this seeded starting point, instead of using a random starting point.

## 4.2 Preliminary Results

We conduct numerical experiments on two different synthetic graphs of varying sizes. These graphs were generated using the stochastic block model [12]. Each graph’s characteristics are reported in Table 1. We record the best objective function values (best solutions) returned by each algorithm in (Table 2). We then compare them among themselves, to the synthetic graphs’ known features, and the graphs’ global density.

While our initial results are still a distance away from the synthetic graph’s figures, the clusters identified by the combined greedy-simulated annealing routine fit our definition of a good cluster, since the mean intra-cluster density is greater than the graph’s global density, in both experiments ( $0.46 > 0.38$  and  $0.59 > 0.34$ ). Our initial results do indicate that the ensemble-like optimization routine which combines a greedy step and simulated annealing yields better results than simulated annealing alone. We find that beginning with a seeded starting point consisting of the solution obtained with just 50,000 iterations of a greedy algorithm significantly improves the results returned by simulated annealing alone.

Our ensemble approach yields better results than running simulated annealing for millions of additional iterations. Increasing the number of iterations of the simulated annealing algorithm does not improve the outcome, while using a seeded starting point does. These results suggest the choice of starting point is critical and that our objective function seems to have multiple local optima. This lack of improvement indicates the presence of a local optimum and outlines the need for an algorithm that will explore the solution set more widely.

**Table 1** Test graph characteristics

	Graph 1	Graph 2
Number of vertices ( $ V $ )	38	104
Number of edges ( $ E $ )	236	2,036
K (global graph wide density)	0.34	0.38
$\bar{K}_{\text{intra}}$	0.87	0.88
Number of clusters ( $ C $ )	4	4
Size of solution set	$\sim 10^{22}$	$\sim 10^{62}$

**Table 2** Best objective function values

	Graph 1	Graph 2
Simulated annealing ( $10^6$ runs)	0.28	0.30
Simulated annealing ( $5 \times 10^6$ runs)	0.28	0.30
Greedy (50,000 runs)	0.52	0.37
Simulated annealing w/ Greedy ( $10^6$ runs)	0.59	0.46
Simulated annealing w/ Greedy ( $5 \times 10^6$ runs)	0.59	0.46

## 5 Conclusion and Future Work

We have formulated the graph clustering problem as a combinatorial optimization problem, using a novel formulation that maximizes mean intra-cluster density. Our objective function penalizes unreasonably sized clusters, which eliminates the need for cluster-size constraints used in other formulations in the literature. While numerical implementation remains work in progress, we do note that an ensemble-like approach which combines a greedy first pass with simulated annealing yields far better results than simulated annealing alone, regardless of the number of iterations.

Future work will involve testing on a wider array of graphs, exploring the use of other meta-heuristic solution techniques and implementation on parallelized high-performance computing platforms. Specific focus will be placed on starting point and escaping local optima.

## References

1. Aloise, D., Caporossi, G., Hansen, P., Liberti, L., Perron, S., Ruiz, M.: Modularity maximization in networks by variable neighborhood search. In: Bader, D.A., Meyerhenke, H., Sanders, P., Wagner, D. (eds.) *Graph Partitioning and Graph Clustering*, 10th DIMACS Implementation Challenge Workshop, Georgia Institute of Technology, Atlanta, GA, USA, 13–14 Feb 2012, pp. 113–128 (2012). <http://www.ams.org/books/conm/588/11705>
2. Bertsimas, D., Tsitsiklis, J.: Simulated annealing. *Stat. Sci.* **8**(1), 10–15 (1993)
3. Brownlee, J.: *Clever Algorithms: Nature-Inspired Programming Recipes*, 1st edn. Lulu.com (2011)
4. Creusefond, J., Largillier, T., Peyronnet, S.: Finding compact communities in large graphs. In: *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, ASONAM '15*, pp. 1457–1464. ACM, New York, NY, USA (2015). <https://doi.org/10.1145/2808797.2808868>
5. van Dongen, S.: *Graph clustering by flow simulation*. Ph.D. thesis, Faculteit Wiskunde en Informatica, Universiteit Utrecht (2000)
6. Fan, N., Pardalos, P.M.: Linear and quadratic programming approaches for the general graph partitioning problem. *J. Global Optim.* **48**(1), 57–71 (2010). <https://doi.org/10.1007/s10898-009-9520-1>
7. Fan, N., Pardalos, P.M.: Robust optimization of graph partitioning and critical node detection in analyzing networks. In: *Proceedings of the 4th International Conference on Combinatorial Optimization and Applications—Volume Part I, COCOA'10*, pp. 170–183. Springer, Berlin, Heidelberg (2010). <http://dl.acm.org/citation.cfm?id=1940390.1940405>
8. Fortunato, S.: Community detection in graphs. *Phys. Rep.* **486**, 75–174 (2010). <https://doi.org/10.1016/j.physrep.2009.11.002>
9. Fortunato, S., Barthélemy, M.: Resolution limit in community detection. *Proc. Natl. Acad. Sci.* **104**(1), 36–41 (2007). <http://www.pnas.org/content/104/1/36.abstract>
10. Fortunato, S., Hric, D.: Community detection in networks: a user guide. ArXiv e-prints (2016)
11. Hastie, T., Tibshirani, R., Friedman, J.: *The Elements of Statistical Learning*, Second Edition: Data Mining, Inference, and Prediction, 2nd ed. Springer Series in Statistics. Springer (2009)
12. Holland, P.W., Laskey, K.B., Leinhardt, S.: Stochastic blockmodels: First steps. *Soc. Netw.* **5**(2), 109–137 (1983). [https://doi.org/10.1016/0378-8733\(83\)90021-7](https://doi.org/10.1016/0378-8733(83)90021-7). <http://www.sciencedirect.com/science/article/pii/0378873383900217>
13. Jin, J.: Fast community detection by score. *Ann. Stat.* **43** (2015)

14. Kazakovtsev, L., Antamoshkin, A.: Genetic algorithm with fast greedy heuristic for clustering and location problems. *Informatica (Slovenia)* **38**(3) (2014). <http://www.informatica.si/index.php/informatica/article/view/704>
15. Lancichinetti, A., Radicchi, F., Ramasco, J.J., Fortunato, S.: Finding statistically significant communities in networks. *PLoS ONE* **6**, e18,961 (2011). <https://doi.org/10.1371/journal.pone.0018961>
16. von Luxburg, U.: A Tutorial on Spectral Clustering. *CoRR* **abs/0711.0189** (2007). <http://arxiv.org/abs/0711.0189>
17. Miasnikof, P., Shestopaloff, A., Bonner, A., Lawryshyn, Y.: A statistical performance analysis of graph clustering algorithms. In: *Lecture Notes in Computer Science*. Springer (2018)
18. Nascimento, M., Pitsoulis, L.: Community detection by modularity maximization using GRASP with path relinking. *Comput. Oper. Res.* **40**, 3121–3131 (2013)
19. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. *Phys. Rev. E, Stat. Nonlinear, Soft Matter Phys.* **69**, 026,113 (2004)
20. Ovelgönne, M., Geyer-Schulz, A.: An ensemble learning strategy for graph clustering. In: Bader, D.A., Meyerhenke, H., Sanders, P., Wagner, D. (eds.) *Graph Partitioning and Graph Clustering, 10th DIMACS Implementation Challenge Workshop, Georgia Institute of Technology, Atlanta, GA, USA, 13–14 Feb 2012*, pp. 113–128 (2012). <http://www.ams.org/books/comm/588/11705>
21. Papadimitriou, C., Steiglitz, K.: *Combinatorial Optimization: Algorithms and Complexity*. Dover Books on Computer Science. Dover Publications (1998). <https://books.google.ca/books?id=u1RmDoJqkF4C>
22. Prokhorenkova, L.O., Prałat, P., Raigorodskii, A.: Modularity of complex networks models. In: Bonato, A., Graham, F., Prałat, P. (eds.) *Algorithms and Models for the Web Graph*, pp. 115–126. Springer International Publishing, Cham (2016)
23. Prokhorenkova, L.O., Prałat, P., Raigorodskii, A.: Modularity in several random graph models. In: *The European Conference on Combinatorics, Graph Theory and Applications (EUROCOMB'17)*, *Electronic Notes in Discrete Mathematics* **61**, 947–953 (2017). <https://doi.org/10.1016/j.endm.2017.07.058>. <http://www.sciencedirect.com/science/article/pii/S1571065317302238>
24. Russell, S., Norvig, P.: *Artificial Intelligence: A Modern Approach*, pp. 92–115. Prentice-Hall Inc. (1995)
25. Schaeffer, S.E.: Survey: graph clustering. *Comput. Sci. Rev.* **1**(1), 27–64 (2007). <https://doi.org/10.1016/j.cosrev.2007.05.001>
26. Tasgin, M., Herdagdelen, A., Bingol, H.: *Community Detection in Complex Networks Using Genetic Algorithms*. ArXiv e-prints (2007)
27. Weisstein, E.: *Clique*. MathWorld—A Wolfram Web Resource (2018). <http://mathworld.wolfram.com/Clique.html>
28. Yang, J., Leskovec, J.: Defining and Evaluating Network Communities Based on Ground-truth. *CoRR* **abs/1205.6233** (2012). <http://arxiv.org/abs/1205.6233>

# Computational Complexity of SRIC and LRIC Indices



Sergey Shvydun

**Abstract** Over the past years, there is a deep interest in the analysis of different communities and complex networks. Identification of the most important elements in such networks is one of the main areas of research. However, the heterogeneity of real networks makes the problem both important and problematic. The application of SRIC and LRIC indices can be used to solve the problem since they take into account the individual properties of nodes, the possibility of their group influence, and topological structure of the whole network. However, the computational complexity of such indices needs further consideration. Our main focus is on the performance of SRIC and LRIC indices. We propose several modes on how to decrease the computational complexity of these indices. The runtime comparison of the sequential and parallel computation of the proposed models is also given.

**Keywords** Influence in networks · Short-range interaction centrality · Long-range interaction centrality · Computational complexity · Group influence · Simple paths

## 1 Introduction

One of the key directions of network analysis is the detection of topologically central nodes in the network. There have been developed many indices that measure the centrality level of each node. Some of them are based on the number of links to the other nodes with (or without) respect to the importance of adjacent nodes (degree measures, eigenvector centrality, PageRank, etc.) [1–4]. Other techniques consider how close each node is located to the other nodes of the network in terms of the distance

---

S. Shvydun (✉)

National Research University Higher School of Economics,  
20 Myasnitskaya Str., Moscow, 101000, Russian Federation

e-mail: [shvydun@hse.ru](mailto:shvydun@hse.ru)

URL: <https://www.hse.ru/en/staff/Shv>

V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, 65  
Profsoyuznaya Str., Moscow 117997, Russian Federation

© Springer Nature Switzerland AG 2020

I. Bychkov et al. (eds.), *Network Algorithms, Data Mining, and Applications*, Springer Proceedings in Mathematics & Statistics 315,  
[https://doi.org/10.1007/978-3-030-37157-9\\_4](https://doi.org/10.1007/978-3-030-37157-9_4)

(e.g., the closeness centrality) or how many times it is on the shortest paths connecting any given node-pairs (e.g., the betweenness centrality) [5–8]. All measures consider different ideas on which properties central elements should satisfy.

However, most real networks are very complex. Nodes of the network are heterogeneous and may have various attributes, which characterize their size, importance, level of influence, etc. In many cases, the influence on two distinct nodes may be calculated differently depending on their attributes. Connections between nodes may also be different and describe different types of relations. Some nodes may coalesce in a group and perform the same behavior to affect a particular node of a network. All these aspects lead to the fact that initial connections do not represent the actual picture of nodes' influence. Hence, classical centrality measures cannot be straightforwardly applied to most real networks. Moreover, they do not give any information about the most interdependent pairs of nodes.

Since most classical centrality measures do not take into account individual attributes of nodes, the possibility of group influence, or long-range interactions, there is a need for new models that considers all these features.

In [9] there was proposed a new model of influence in a network based on short-range interactions. In the original paper, it was applied to the loan market and, consequently, called a key borrower index (KBI). In this paper, we use another name Short-Range Interaction Centrality (SRIC). A distinct feature of the model is that it allows to consider individual characteristics of nodes and their group influence. However, the proposed model was quite limited since the main intuition was that a node could be affected only by adjacent nodes both directly or indirectly. In the next Section, we will discuss the SRIC index in more detail.

In [10–12] there were proposed several models called Long-Range Interaction Centralities (LRIC) which were aimed to remove the drawbacks of the SRIC index. The models take into account individual attributes of nodes and a possibility of a group influence but, in contrast to the SRIC index, they also allow the influence between nodes which are connected via  $k$  intermediate nodes through considering different paths or chain reactions between them. However, the computational complexity of such indices requires a further consideration.

Since real networks are large and complex, centrality measures with a high computational complexity cannot be applied to them. Thus, there is a need to evaluate the computational complexity of SRIC and LRIC indices in order to determine whether they can be applied to complex networks.

In this paper, we focus on the computational complexity of SRIC and LRIC indices. As a result, we propose a novel approach on how to decrease the complexity of the LRIC index. We also generate various graphs to evaluate the runtime and compare it to the previous approaches.

The paper is organized as follows. In Sect. 2 we give some definitions and describe SRIC and LRIC indices. In Sect. 3 there is a discussion of their computational complexity. In Sect. 4 we consider some properties of LRIC index and propose several models that may decrease its computational complexity. In Sect. 5 we perform some experimental results and define when these indices can be applied to real networks. Finally, we conclude the main results of the paper.

## 2 Methodology

### 2.1 Mathematic Notions

Consider a graph  $G = (V, E, W)$ , where  $V = \{1, \dots, n\}$  is a set of nodes,  $E \subseteq V \times V$  is a set of edges (edge  $(i, j) \in E$ ),  $W = \{|w_{ij}|\}$  is a set of weights associated with edge  $(i, j) \in E$ . We consider directed graphs, i.e., a graph where the existence of edge  $(i, j)$  does not imply the existence of the edge  $(j, i)$ . Additionally, each node  $i$  may have a set of attributes  $w_i^k$ , where  $k$  is an attribute,  $k \in K$ , and some predefined threshold of influence  $q_i$ .

**Definition 1** ([11]) A critical group for node  $j$  is a subset of nodes whose total influence on node  $j$  exceeds this quota  $q_j$ . More formally,  $\Omega(j) \subseteq V$  is a critical group for node  $j$  if

$$\sum_{i \in \Omega(j)} w_{ij} \geq q_j. \quad (1)$$

**Definition 2** ([11]) A node  $l$  is pivotal for some group if its exclusion from this group makes the group noncritical. Formally speaking,  $\Omega^p(j) \subseteq \Omega(j)$  is a subset of pivotal nodes of group  $\Omega(j)$  if

$$\sum_{i \in \Omega(j) \setminus \{k\}} w_{ij} < q_j \quad \forall k \in \Omega^p(j). \quad (2)$$

**Definition 3** ([11]) A simple path between nodes  $i$  and  $j$  in graph  $G$  is a sequence of edges that connects them and contains distinct nodes. More precisely,

$$(i, k_1), (k_1, k_2), (k_2, k_3), \dots, (k_{s-1}, j), i \neq k_1 \neq \dots \neq k_{s-1} \neq j$$

is a simple path of length  $s$  denoted as  $P_{i-j}^t(s)$ , where  $t$  is the number of the path.

The problem lies in the identification of the most influential elements in the network taking into account quotas and other individual attributes of nodes. In [9–12], there were proposed SRIC and LRIC indices which we describe below.

### 2.2 Short-Range Interaction Centrality (SRIC)

The SRIC index was first introduced in the voting theory [13] and then adapted to the estimation of the influence of agents in the loan market network [9]. The main idea of the index is analyze the influence on each individual node in order to reconstruct initial connections in a network.

The SRIC index can be divided into several stages. Since connections between nodes do not represent the real influence, the first stage evaluates normalized inten-

sities of direct influence of node  $i$  to node  $j$  denoted by  $p_{ij}$  and calculated as

$$p_{ij} = \frac{w_{ij}}{\sum_k w_{kj}}, \quad (3)$$

as well as their indirect influence through some node  $h$  denoted by  $p_{ihj}$  and calculated as

$$p_{ihj} = \left\{ \begin{array}{l} \frac{w_{ih}}{\sum_k w_{kj}}, \text{ if } w_{ij} > 0, w_{hj} > 0, w_{hj} \geq w_{ih}, i \neq h \neq j, \\ \frac{w_{hj}}{\sum_k w_{kj}}, \text{ if } w_{ij} > 0, w_{hj} > 0, w_{hj} < w_{ih}, i \neq h \neq j, \\ 0, \text{ otherwise.} \end{array} \right\} \quad (4)$$

Obviously, only adjacent nodes of node  $j$  can be pivotal since they determine whether a group is critical or not. Hence, knowing critical group  $\Omega_k(j)$  where node  $i$  is pivotal and normalized intensities of connections, the second stage calculates the intensity of its influence denoted by  $f(i, \Omega_k(j))$

$$f(i, \Omega_k(j)) = \frac{p_{ij} + \sum_{h \in \Omega_k(j)} p_{ihj}}{|\Omega_k(j)|}, \quad (5)$$

The intensity of influence within every critical group can be calculated of each pivotal node by formula (5). Note that the size of a group is taken into account so the bigger a group is, the less contribution each pivotal node has.

Thus, the main idea of the SRIC index is to find all critical groups where node  $i$  is pivotal for node  $j$  and find the total influence  $\chi_i(j)$ , i.e.,

$$\chi_i(j) = \sum_k f(i, \Omega_k(j)), \quad (6)$$

and then normalize the value taking into account information about the influence of other nodes on node  $j$

$$\hat{\chi}_i(j) = \frac{\chi_i(j)}{\sum_k \chi_k(j)}, \quad (7)$$

The normalized influence  $\hat{\chi}_i(j)$  is calculated for all distinct nodes  $i$  and  $j$  and then aggregated into a single vector  $\alpha_{SRIC}(i)$  based on attributes of nodes which are influenced by node  $i$ .

The SRIC index was the first attempt to evaluate the influence of nodes based on nodes attributes and the possibility of group influence but it has several drawbacks. First, it considers the influence only between adjacent nodes while in many networks indirect nodes may influence each other. Second, critical group  $\Omega_k(j)$  may include nodes which have no connection to node  $j$ . It leads to the fact that the total influence  $\chi_i(j)$  between two adjacent nodes  $i$  and  $j$  actually depends on nodes that have no relation to  $i$  and  $j$ . Moreover, if we add some isolated node  $k$  into the graph, the influence between two nodes  $i$  and  $j$  will be increased while it should intuitively remain constant. Thus, it is more reasonable that only adjacent nodes may form crit-



ical groups. Third, the use of normalization is also under discussion. Although  $\chi_i(j)$  measure provides information on which node more influences node  $j$ , it is impossible to compare  $\chi_i(j)$  and  $\chi_i(k)$  values and determine which node is more influenced by node  $i$ . Furthermore, a node that influences some other node individually and a node that influences the other node only in a group may have the same normalized values. Finally, the algorithm requires the enumeration of all possible subsets of nodes set which is a computationally complex task.

### 2.3 Long-Range Interaction Centrality (LRIC)

The LRIC index was proposed in [10–12] to remove the drawbacks of the SRIC index. A distinct feature of the LRIC index is that non-adjacent nodes may influence each other. In this section, we do not consider the LRIC index based on simulations which were proposed in [12] because of its low computational complexity.

Contrary to the SRIC index, the LRIC index provides a completely different idea on how to measure the influence between nodes. While the SRIC index requires the evaluation of normalized intensities of direct and indirect influence and the detection of all possible critical groups, the pairwise influence in the LRIC index is based on another idea. According to it, the direct influence of node  $i$  on node  $j$  denoted by  $c_{ij}$  is equal to its maximal possible share of the total influence in a group where node  $i$  is pivotal, i.e.,

$$c_{ij} = \begin{cases} \max_{i \in \Omega_k^p(j)} \frac{w_{ij}}{\sum_{h \in \Omega_k(j)} w_{hj}}, & i \in \Omega_k^p(j) \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

Thus, the influence in the LRIC index is based on the intuition that only adjacent nodes may affect some other nodes and no indirect or isolated nodes can change the direct influence. Moreover, the LRIC index also satisfies the rationality principle where each individual node aims to affect some other nodes in a group where its influence is maximal. If node  $i$  influences node  $j$  individually, the total influence is equal to 1. On the contrary, if node  $i$  is pivotal in none of the critical groups, the total influence is equal to 0.

The indirect influence of nodes in the LRIC index is based on information about direct influence. First, it assumes that nodes may affect other nodes through no more than  $s$  intermediate nodes, where value  $s$  is a parameter that depends on the problem. The second assumption is that all these intermediate nodes should be distinct, which also seems reasonable in many cases. Since the direct influence varies from 0 to 1, there were proposed several ways how to calculate the indirect influence through some  $t$ th path  $P_{i-j}^t(s)$

$$f_{\text{multi}}(P_{i-j}^t(s)) = c_{ik_1(t)} \times c_{k_1(t)k_2(t)} \times \cdots \times c_{k_{s-1}(t)j}, \quad (9)$$

or

$$f_{\min}(P_{i-j}^t(s)) = \min \{c_{ik_1(t)}, c_{k_1(t)k_2(t)}, \dots, c_{k_{s-1}(t)j}\}, \quad (10)$$

Formula (9) might be interpreted as a joined probability while formula (10) is thought of as minimal flow capacity in the path.

Since there can be multiple paths between to nodes, there is a need to aggregate all of them. In [10, 11], there were proposed several ideas on how to aggregate the influence through each path. The first idea is that total influence between nodes  $c_{ij}^*(s)$  is equal to the sum of influences of all possible simple paths between them, i.e.,

$$c_{ij}^*(s) = \min \left\{ \sum_{k:|P_{i-j}^k| \leq s} f(P_{i-j}^k), 1 \right\} \quad (11)$$

The second idea of the total influence lies on the detection of the strongest, in some sense, path between two nodes. For instance, the strongest path is a path with the maximal influence, i.e.,

$$c_{ij}^*(s) = \max_{k:|P_{i-j}^k| \leq s} f(P_{i-j}^k) \quad (12)$$

or it is a path which is the best one by the aggregation rule proposed in [14], i.e.,

$$c_{ij}^*(s) = f(P_{i-j}^z) \quad (13)$$

where

$$z = \operatorname{argmin}_{k:n(k) \leq s} v(P_{i-j}^k) \quad (14)$$

and

$$v(P_{i-j}^k) = \sum_{l=1}^{n(k)-1} (s+1)^{m-v_l(P_{i-j}^k)} \quad (15)$$

$n(k)$  is a length of path  $P_{i-j}^k$ ,  $v_l(P_{i-j}^k)$  is a grade that was assigned to edge between  $l$ -th and  $(l+1)$ -th nodes in a path  $P_{i-j}^k$ ,  $m$  is a maximal possible grade in a network.

As a result, there were proposed 5 different indices that evaluate the pairwise influence of nodes (see Table 1).

**Table 1** Models of indirect influence assessment for the LRIC index

	Paths aggregation		
	Sum of paths	Maximal path	Threshold rule
Multiplication of direct influence	LRIC (Sum)	LRIC (Max)	LRIC (MultT)
Minimal direct influence	–	LRIC (MaxMin)	LRIC (MaxT)

Similarly to the SRIC index, the LRIC index can be obtained by aggregating information about the pairwise influence and attributes of nodes. It should be noted here that some other versions of the LRIC index can be obtained by applying classical centrality measures to the network of direct influence.

### 3 Computational Complexity of SRIC and LRIC Indices

#### 3.1 Computational Complexity of the SRIC Index

In the previous section, we have discussed the SRIC index and its main drawbacks and emphasized that computational complexity is a weak point. In order to evaluate the influence of node  $i$  on node  $j$  we need to consider all possible critical groups of nodes where node  $i$  is pivotal. In general, node  $i$  can be pivotal in all possible critical groups and the total number of which is equal to  $2^{n-1} - 1$ . Since the direct influence is calculated for each individual node, the total number of examined combinations is equal to  $2^n - 1$ . Thus, the SRIC index is applicable only to small networks.

In [11], it was mentioned that to minimize the total number of studied critical groups we can additionally limit the size of critical groups. Indeed, in many applications coordination costs are so high that only small groups of nodes may coalesce to influence the other node. If we introduce this additional parameter  $n_0$ , the total number of considered groups for each node in the worst case will be equal to  $\sum_{l=1}^{n_0} (l \cdot C_{n-1}^l)$ . Here we should note that since a group that contains only nonadjacent nodes is not critical for a particular node, the upper bound of considered groups can be decreased. However, in the case of a large number of nodes in a network the computational complexity is relatively high.

Additionally, it was discussed in [11] that we can introduce an additional parameter of a node  $q_w$  and do not consider its adjacent nodes with a weight less than this parameter. Unfortunately, the accuracy of the SRIC index will be decreased in that case. Moreover, the definition of parameter  $q_w$  is arguable.

We can formulate two theorems about critical and pivotal nodes.

**Theorem 1** *If node  $i$  with weight  $w_{ij}$  is pivotal for node  $j$ , any node  $k$  which is connected to  $j$  is pivotal if  $w_{kj} \geq w_{ij}$ .*

**Proof** Since node  $i$  is pivotal, there exists a critical group  $\Omega(j)$  for  $j$  where node  $i$  is pivotal and the following conditions hold:

1.  $\exists \Omega(j) : i \in \Omega^p(j)$ ;
2.  $\sum_{l \in \Omega(j)} w_{lj} \geq q_j$ ;
3.  $\sum_{l \in \Omega(j) \setminus \{i\}} w_{lj} < q_j$ .

There are two possible situations: node  $k$  is also in critical group  $\Omega(j)$  or it is not in a group. Let us consider the first case. If node  $k$  is in critical group  $\Omega(j)$  and  $w_{kj} \geq w_{ij}$ , it is true that

$$\sum_{l \in \Omega(j) \setminus \{k\}} w_{lj} \geq \sum_{l \in \Omega(j) \setminus \{i\}} w_{lj}.$$

Thus, node  $k$  is also pivotal by condition 3, i.e.,  $\sum_{l \in \Omega(j) \setminus \{k\}} w_{lj} \leq q_j$ .

Consider the second case where node  $k$  is not in critical group  $\Omega(j)$ . It is true that group  $\Omega(j)$  is not critical without node  $i$  by condition 3. Examine another group  $\Omega'(j) = \{k\} \cup \Omega(j) \setminus \{i\}$ . Since we know that  $w_{kj} \geq w_{ij}$  and by condition 2

$$\sum_{l \in \Omega(j) \setminus \{i\}} w_{lj} + w_{ij} \geq q_j.$$

it is obvious that

$$\sum_{l \in \Omega(j) \setminus \{i\}} w_{lj} + w_{kj} \geq q_j.$$

Thus, node  $k$  is pivotal in group  $\Omega'(j)$ . We considered all possible cases, consequently, any node  $k$  such that  $w_{kj} \geq w_{ij}$  is always pivotal for node  $j$  if node  $i$  is pivotal.

**Theorem 2** *If node  $i$  with weight  $w_{ij}$  is pivotal for node  $j$ , any node  $k$  where  $w_{kj} \leq w_{ij}$  is not pivotal for node  $j$ .*

**Proof** Let us prove the theorem by making a contradiction. Assume there exists node  $k$  where  $w_{kj} \leq w_{ij}$  and this node is pivotal for node  $j$ . It leads to the fact that there exists a critical group for  $j$  where node  $k$  is pivotal. In other words, the following conditions hold:

1.  $\exists \Omega(j) : k \in \Omega^p(j)$ ;
2.  $\sum_{l \in \Omega(j)} w_{lj} \geq q_j$ ;
3.  $\sum_{l \in \Omega(j) \setminus \{k\}} w_{lj} < q_j$ .

There are two possible situations: node  $i$  is also in critical group  $\Omega(j)$  or it is not in a group. If node  $i$  is in critical group  $\Omega(j)$  and  $w_{kj} \leq w_{ij}$  it is true that

$$\sum_{l \in \Omega(j) \setminus \{i\}} w_{lj} \geq \sum_{l \in \Omega(j) \setminus \{k\}} w_{lj}.$$

Thus, node  $i$  is also pivotal by condition 3, i.e.,  $\sum_{l \in \Omega(j) \setminus \{i\}} w_{lj} \leq q_j$ .

Suppose node  $i$  is not in critical group  $\Omega(j)$ . It is true that group  $\Omega(j)$  is not critical without node  $k$  by condition 3. Examine another group  $\Omega'(j) = \{i\} \cup \Omega(j) \setminus \{k\}$ . Since we know that  $w_{kj} \leq w_{ij}$  and by condition 2

$$\sum_{l \in \Omega(j) \setminus \{k\}} w_{lj} + w_{kj} \geq q_j.$$

it is obvious that

$$\sum_{l \in \Omega(j) \setminus \{k\}} w_{lj} + w_{ij} \geq q_j.$$

Thus, node  $i$  is pivotal in group  $\Omega'(j)$ . Again we got a contradiction that node  $i$  is not pivotal. Hence, node  $k$  where  $w_{kj} < w_{ij}$  is not pivotal for node  $j$  if node  $i$  is also not pivotal.

Theorems 1, 2 play a significant role for calculation of the SRIC index. The SRIC evaluates the pairwise influence  $\chi_i(j)$ , consequently, if  $\chi_i(j) = 0$  then for any node  $k$  such that  $w_{kj} < w_{ij}$  there will not be any influence on node  $j$ , i.e.,  $\chi_k(j) = 0$ . These results allow to adapt the binary search algorithm to evaluate the pairwise influence between nodes and, hence, reduce the computational complexity. However, if node  $i$  is pivotal for node  $j$  the SRIC index still requires to consider all possible critical groups. Nevertheless, Theorems 1, 2 can be used to reduce the total number of possible critical groups under consideration.

### 3.2 Computational Complexity of the LRIC Index

The computational complexity is a weak point of the LRIC index. Similarly to the SRIC index, it requires to consider different groups of nodes to evaluate the pairwise influence. Moreover, indirect influence between nodes in the LRIC index is based on information about all possible simple paths which is also a combinatorially difficult task. Thus, we divide the problem of the LRIC calculation into two separate parts: how to calculate direct and indirect influence.

#### (1) Direct influence calculation of the LRIC index

Contrary to the SRIC index, the LRIC index considers groups which contain only adjacent nodes. Thus, if each node  $j$  has a relatively small number of neighbors denoted by  $N(j)$ , the computational complexity of pairwise influence evaluation is low. In the previous section, we have discussed that the total number of considered groups for each node is equal to  $2^{N(j)} - 1$  in general and  $\sum_{l=1}^{n_0} (l \cdot C_{N(j)}^l)$  if we limit the maximal group size. Unfortunately, for complete graphs we have the same problem as it was for the SRIC index and, consequently, the LRIC index cannot be applied.

We have also discussed in Theorems 1, 2 that in many cases we do not need to evaluate the influence of each node  $i$  to node  $j$ . However, contrary to the SRIC index, the direct influence  $c_{ij}$  does not depend on all possible critical groups of nodes. The LRIC index requires a detection of only one minimal critical group, where node  $i$  is pivotal for node  $j$ . Since we do not need to consider all critical groups, the computational complexity can be decreased. Moreover, we can adapt Theorem 2 and remove all non-pivotal nodes from further consideration as any group  $\Omega(j)$  that contains non-pivotal nodes will not be minimal.

Consider the direct influence on node  $j$ . The problem lies in the evaluation of direct influence for each adjacent node  $i$ . In other words, we need to find a critical group where node  $i$  is pivotal and has a maximal possible share of the total group influence. Since node  $i$  is pivotal, it should satisfy the same conditions which are discussed in Theorem 1.

We can observe that there is a need to find group  $\Omega'(j) \subseteq N(j) \setminus \{i\}$  with a minimal total influence where  $\sum_{l \in \Omega'(j)} w_{lj} \in [q_j - w_{ij}, q_j)$ . The upper bound of the total influence of some critical group we denote by  $Q_{ij}$  (initially,  $Q_{ij} = q_j$ ).

The problem can be formulated as an integer linear programming problem. Let  $x_l = \{0, 1\}$  be the value that shows if node  $l$  is in group  $\Omega'(j)$ . Since group  $\Omega'(j) \cup \{i\}$  should be critical for node  $j$ , we add the following constraints:

$$q_j - w_{ij} \leq \sum_l w_{lj} \cdot x_l < Q_{ij}. \quad (16)$$

Subject to (16), the goal is to find values  $x_{ij}$  that minimizes the total group influence of group  $\Omega'(j)$ , i.e.,

$$\sum_l w_{lj} \cdot x_l \rightarrow \min.$$

Additionally, we can add the constraint on the maximal critical group size  $s$

$$\sum_l x_l < s. \quad (17)$$

If we solve this integer linear programming problem, we will get the critical group with a minimal where node  $i$  is pivotal. Additionally, if the solution exists, we can use Theorem 2 and update the upper bound  $Q_{kj}$  for other nodes  $k$  such that  $w_{ij} \leq w_{kj}$  to solve the same problem for these nodes with an updated bound. On the other hand, if no solution is found, we do not need to solve the same problem for nodes  $k$  which are adjacent to  $j$  and  $w_{kj} \leq w_{ij}$  by Theorem 1.

Thus, we transformed the problem of critical group detection to integer linear programming problems, thus, reducing the computational complexity. We can also use Theorems 1, 2 and adapt the binary search algorithm to reduce the total number of integer linear programming problems.

## (2) Indirect influence calculation of the LRIC index

In [10, 11] it was proposed to evaluate the indirect influence between nodes through considering all possible simple paths. As a result, there were proposed 5 indices that aggregate information about direct and indirect influence.

### (a) LRIC (Sum) index

The LRIC (Sum) index considers all possible paths between nodes. If we know all paths as well as their influence, we can define the total pairwise influence. Again,

the problem of all simple paths detection has high computational complexity. In the worst case, there can be  $2^{n-2} - 1$  simple paths between two nodes and, since self-influence is not considered, there are  $\frac{n \cdot (n-1)}{2} \cdot (2^{n-2} - 1)$  paths in total.

In many applications, long connections between nodes do not play a significant role or are not representative in terms of indirect influence. Hence, we can introduce parameter  $s$  that defines the lengths of connections that we take into account. In that case, the total number of possible paths between two nodes will be equal to  $\sum_{l=1}^{\min(s, n-2)} \frac{(n-2)!}{(n-2-l)!}$ . For instance, there are  $n - 2$  simple paths of length 2,  $n^2 - 5n + 6$  simple paths of length 3, and  $n^3 - 9n^2 + 26n - 24$  between pair of nodes. Obviously, the total number of paths between the two nodes grows exponentially.

To decrease the computational complexity of the problem, we can adapt the well-known depth-first search (DFS) technique [15] that is used for scanning and traversing finite graphs. A distinct feature of the algorithms is that it finds all nodes reachable from some other node and runs in linear time  $O(|V| + |E|)$ . Unfortunately, we did not find any rational adaptation of the DFS algorithm since for our problem we need to keep in memory/cache enormous information about visited paths.

To evaluate the total influence of all possible simple paths between a pair of nodes, let us consider the problem of the detection of simple paths of length  $k$ . If we find the solution to the problem, we can extend it to the problem of all simple paths detection.

It is well-known that the total number of paths of length  $k$  can be obtained by multiplication of adjacency matrix to the power of  $k$ . In other words, if  $A = [a_{ij}]_{n \times n}$  is an adjacent matrix such that  $a_{ij} = 1$  if  $(i, j) \in E$  and  $a_{ij} = 0$ , otherwise, the total number of paths of length  $k$  from node  $i$  to node  $j$  will be equal to  $[A^k]_{ij}$ .

Since the LRIC (Sum) index is based on the idea of accumulation of paths influence where the influence of each path is equal to multiplication of edge values, we can observe that the total influence of node  $i$  to node  $j$  through all paths of length  $k$  is equal to  $[C^k]_{ij}$ , where  $C = [c_{ij}]_{n \times n}$  is a matrix of direct influence according to formula (8) and  $C^k$  is the  $k$ th power of matrix  $C$ . Naturally, if only direct influence is allowed, the total influence is equal to the direct influence.

However, the result of matrix multiplication  $[C^k]_{ij}$  includes not only simple paths but also paths with cycles, i.e., paths where a particular node appears several times in a path. Thus, we need to find all paths with cycles and remove them from the total value. For instance, for  $k = 2$  there is only one possible path with cycle (see Fig. 1).

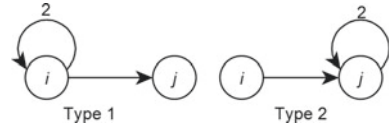
Note that self-influence is not allowed in the LRIC index, i.e.,  $\forall i \in N c_{ii} = 0$ . Hence, paths of length 2 with cycles are possible only between the same nodes. Thus, the total indirect influence  $c_{ij}(k)$  for  $k = 2$  can be calculated as

$$c_{ij}(2) = \begin{cases} [C^2]_{ij}, & \text{if } i \neq j, \\ 0, & \text{otherwise.} \end{cases} \tag{18}$$

**Fig. 1** Path with cycles of length 2



**Fig. 2** Path with cycles of length 3



**Fig. 3** Path with cycles that appears in Type 1 and Type 2



Let us consider the paths of length 3, i.e.,  $k = 3$ . In fact, we can divide all paths with cycles for node  $i$  to node  $j$  in 2 types (see Fig. 2).

The total influence of each path with cycle provided in Fig. 3 can be calculated as

1. Type 1:  $[C^2]_{ii} \cdot c_{ij}$ ;
2. Type 2:  $c_{ij} \cdot [C^2]_{jj}$ ;

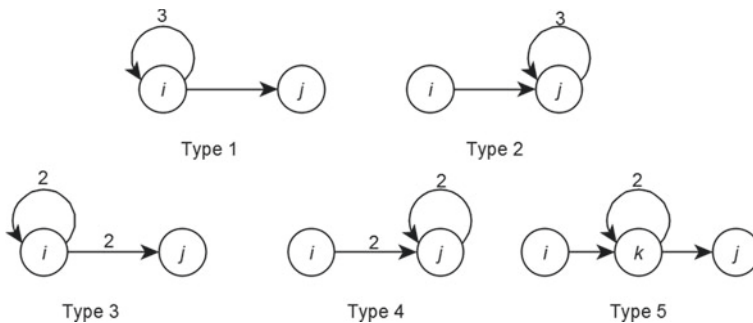
Although we presented different patterns of paths of length 3, it is unclear where these patterns intersect. Thus, we have a problem of path detection that exists in several patterns simultaneously which we will discuss below. In our case, we have only 1 path that appears in Type 1 and Type 2 (see Fig. 3).

Thus, the total influence of nodes through all simple paths of length 3 can be evaluated as

$$c_{ij}(3) = \begin{cases} [C^3]_{ij} - ([C^2]_{ii} + [C^2]_{jj}) \cdot c_{ij} + (c_{ij})^2 \cdot (c_{ji})^2, & \text{if } i \neq j, \\ 0, & \text{otherwise.} \end{cases} \quad (19)$$

Similarly, we can evaluate the paths of length 4. All paths with cycles are divided in 5 types (see Fig. 4).

The total influence of paths can be computed using the matrix of direct influence  $C$  (see Table 2). Again, we have the problem that there are some paths which appear in several path types simultaneously. If we consider types' intersections, we will obtain 4 more path types (see Fig. 5).

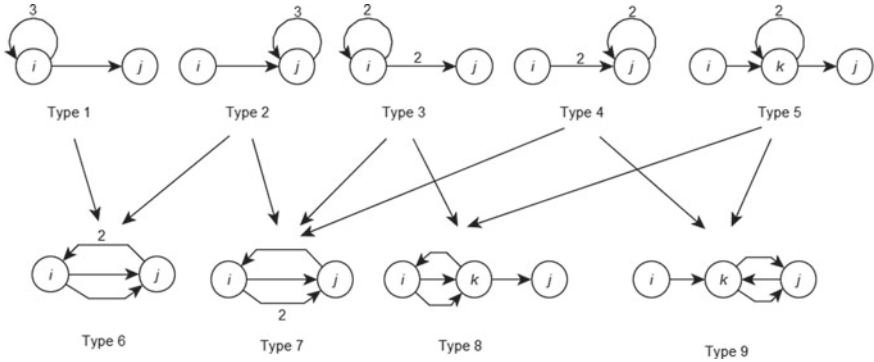


**Fig. 4** Path with cycles of length 4



**Table 2** Calculation of paths with cycles ( $k = 4$ )

Type of cycle	Formula for nodes $i$ and $j$
1	$[C^3]_{ii} \cdot c_{ij}$
2	$[C^3]_{jj} \cdot c_{ij}$
3	$[C^2]_{ii} \cdot [C^2]_{ij}$
4	$[C^2]_{ij} \cdot [C^2]_{jj}$
5	$\sum_k c_{ik} \cdot [C^2]_{kk} \cdot c_{kj}$



**Fig. 5** Intersection of path types ( $k = 4$ )

Paths of type 6 appear in types 1 and 2, consequently, since we subtract types 1–2 from the total influence, we should add type 6 to the total influence. Similarly, we can do the same procedure for types 8 and 9. As for paths of type 7, they appear in types 2–4, thus, we need to add the influence of these paths 2 times to the total influence.

Thus, the total influence of nodes through all simple paths of length 4 can be evaluated as

$$c_{ij}(4) = \begin{cases} [C^4]_{ij} - ([C^3]_{ii} + [C^3]_{jj} - c_{ij}[C^2]_{ji})c_{ij} - ([C^2]_{ii} + [C^2]_{jj} - \\ - 2c_{ij}c_{ji})[C^2]_{ij} + \sum_k c_{ik}(c_{ki}c_{ik} + c_{kj}c_{jk} - [C^2]_{kk})c_{kj}, & \text{if } i \neq j, \\ 0, & \text{otherwise.} \end{cases} \quad (20)$$

Note that types 6–9 do not intersect with each other which is not true in general. For instance, if we apply the same methodology for  $k > 4$ , there will be some intersections. Hence, let us generalize the idea of paths intersection. The problem of intersection of path types has is rather complex. One of the ways to solve this issue in general is to consider all path types as a finite sets  $B_1, B_2, \dots, B_r$  and use the Inclusion–Exclusion principle [16] which is usually attributed to Abraham de Moivre; it is sometimes also named for Daniel da Silva, Joseph Sylvester, or Henri Poincare (see formula 19)

$$|\bigcup_{l=1}^r B_l| = \sum_{X \subseteq \{1, \dots, r\}} (-1)^{|X|+1} \cdot |\bigcup_{l \in X} B_l| \tag{21}$$

The Inclusion–Exclusion principle considers all possible combination of sets  $B_1, B_2, \dots, B_r$ . In general, the total number of such combinations is equal to  $2^r - 1$ . In other words, the Inclusion–Exclusion principle cannot be used for large values  $r$ . In our case, value  $r$  is the number of path types which depends on the path length (see Table 3).

According to Table 3, even paths of length 6 require to consider more than 500,000 comparisons. Thus, we cannot apply the Inclusion–Exclusion principle in our case.

We can also observe that most path types do not intersect with each other. Thus, based on the idea which is provided in Fig. 6, we propose another model on how to find the intersections of path types that requires less comparisons.

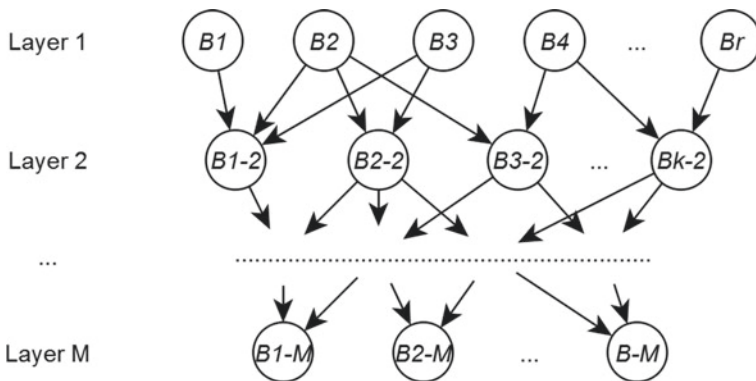
The illustration of the model is presented in Fig. 6. Consider a hierarchical structure that contains on each level  $t$  some set of elements  $B^t = \{B_1^t, B_2^t, \dots, B_{r(t)}^t\}$ , where each element  $B_i^t$  is a set of paths of level  $t$ . Suppose, the next level  $t + 1$  in the hierarchy is constructed by the following rule:

$$x \in B^{t+1} \Leftrightarrow \exists i, j \in \{1, \dots, r(t)\} : i \neq j \ \& \ x = B_i^t \cap B_{ij}^t \tag{22}$$

We can also define a binary relation  $P$  over two adjacent layers by the following rule:

**Table 3** Total paths types depending on the path length

Path length	Number of path types
1, 2	0
3	2
4	5
5	10
6	19
7	35
8	63
9	112
10	198
11	239
12	614
13	1079
14	1895
15	3327



**Fig. 6** Paths types intersection model

$$(B_i^t, B_j^{t+1}) \in P \Leftrightarrow \exists l \in \{1, \dots, r(t)\} : i \neq l \ \& \ B_j^{t+1} = B_i^t \cap B_l^t$$

Additionally, each element  $B_i^t$  may have some weight  $v_i^t$  which can be calculated as

$$v_i^{t+1} = 1 - \sum_{l:j:(j,i) \in P^t} v_j^t.$$

In other words, the weight  $v_i^t$  is equal to the difference of 1 and the sum of weights of all its predecessors (note that  $v_i^1 = 1$ ). Since the set of elements of the first level is finite and we consider different intersections, it is clear that the total number of layers constructed by formula (22) is finite, i.e.,  $\exists t_0 : |B^{t_0}| = 0$ .

Let the first level ( $t = 1$ ) of the hierarchy contains different path types  $B_1^1, B_2^1, \dots, B_{r(1)}^1$  and weight  $v_i^1$  for each type  $i$  is equal to 1, i.e.,  $\forall i \in \{1, \dots, r(1)\} v_i^1 = 1$ . Then, based on pairwise comparisons we can define various path types of all next levels and evaluate the value  $r_i^t$ . Since, the total influence of each path type  $f(B_i^t)$  can be evaluated, we propose the following formula:

$$f(B_1^1 \cup B_2^1 \cup \dots \cup B_{r(1)}^1) = \sum_{t=1}^{t_0} \sum_{i=1}^{r(t)} f(B_i^t) \cdot v_i^t. \tag{23}$$

We provided another model on how to evaluate the total number of distinct elements in different sets which requires fewer comparisons. Thus, we can use formula (23) to evaluate the total influence of all paths with cycles.

Thus, we proposed a generalized model that evaluates the total influence through all simple paths of length  $k$ . We also developed an algorithm that performs the calculation of the LRIC (Sum) index. The model is based on the matrix multiplication

and it removes all paths with cycles. The LRIC (Sum) matrix of pairwise influence can be calculated as

$$c_{ij}^*(s) = \sum_{k=1}^s c_{ij}(k).$$

Matrix operations play a key role to reduce the computational complexity of considering all simple paths. The matrix addition simply adds the corresponding entries together. As for matrix multiplication, it is the most expensive operation of our model. Although the naive algorithm of matrix product has  $O(n^3)$  complexity, there have been developed various faster algorithms. In 1969, there was proposed a well-known Strassen algorithm [17] that has  $O(n^{2.81})$  complexity. The fastest known algorithm, proposed in 1987 by Don Coppersmith and Winograd [18], runs in  $O(n^{2.38})$  time. It is further believed that some algorithms with almost  $O(n^2)$  time exist. Thus, the computation of the LRIC (Sum) index depends on the issue whether there exists more optimal algorithms for matrix multiplication.

The LRIC (Sum) index calculation was implemented in the programming environment R and Python. Section 4 provides the runtime of the proposed algorithm.

#### **(b) LRIC (Max), LRIC (MaxMin), LRIC (MultT), and LRIC (MaxT)**

Contrary to LRIC (Sum) index, other versions of LRIC index do not require consideration of all possible simple paths. Indeed, all these versions are based on the detection of the strongest path according to some criteria. In other words, there exist some algorithms that calculate these indices without considering all paths between nodes.

Since all indices are based on the same approach but use different methods of path strength evaluation, we can propose the same intuition for these measures. Actually, we can transform the considered problem to the shortest path problem with different criteria.

As we have discussed, in terms of the shortest path problem the distance of each path  $d(P_{i \rightarrow j}^k)$  is calculated in LRIC (Max) index by formula (9), in LRIC (MaxMin) by formula (10) while in LRIC (MultT) and LRIC (MaxT) it is calculated by (15). The problem lies in the detection of the path with the shortest distance.

To find all shortest paths in a graph, there are well-known Floyd–Warshall algorithm [19, 20] and adapted Dijkstra algorithm [21]. The Floyd–Warshall algorithm finds all shortest paths in  $O(n^3)$  time while adapted Dijkstra algorithms runs in  $O(n \cdot |E| + n^2 \log(n))$ . There is also Johnson’s algorithm that uses both Dijkstra and Bellman–Ford as subroutines. However, since we consider the model where node  $i$  may influence other node  $j$  through no more than  $s$  intermediate nodes, we cannot apply these algorithms as they do not guarantee the shortest paths of maximal length  $s$ . Note that all shortest path problem can be used as other versions of indirect influence calculation for the LRIC index.

Thus, we present another model that evaluates all shortest paths of maximal length  $s$  in a graph. Based on formulae (9), (10), (15) consider the matrix of direct influence  $C = [c_{ij}]_{n \times n}$  and the matrix of total influence  $C^* = [c_{ij}^*]_{n \times n}$  where  $\forall i, j \in N$   $c_{ij}^* = c_{ij}$

for LRIC (Max) and LRIC (MaxMin) and  $c_{ij}^* = -(s+1)^{(m-v_l(c_{ij}))}$  if  $c_{ij} > 0$  and  $c_{ij}^* = -\infty$  otherwise for LRIC (MaxT) and LRIC (MultT).

We propose an operation  $*$  of raising matrix  $C^*$  to a power of 2 that gives a matrix  $C^{*2} = [c_{ij}^{*2}]_{n \times n}$  such that

$$c_{ij}^{*2} = \max_k (c_{ij}^*, f(c_{ik}^*, c_{kj})), \quad (24)$$

where  $f(c_{ik}^*, c_{kj})$  is an operator that works as

1.  $f(c_{ik}^*, c_{kj}) = c_{ik}^* \cdot c_{kj}$  for LRIC (Max);
2.  $f(c_{ik}^*, c_{kj}) = \min(c_{ik}^*, c_{kj})$  for LRIC (MaxMin);
3.  $f(c_{ik}^*, c_{kj}) = c_{ik}^* - (s+1)^{m-v_l(c_{kj})}$  for LRIC (MaxT) and LRIC (MultT).

Matrix  $C_{ij}^{*s}$  actually represents information about the strongest path of maximal length  $s$  according to indices under consideration. We have converted the problem of LRIC indices calculation to the problem of simple mathematical operations. Since  $c_{ij}$  values vary from 0 to 1, any path with a cycle will not get a greater value than the same path without cycles by formula (24). Thus, we do not have the problem of simple paths detection here while our complexity is equal to the complexity of raising a matrix to the power of  $s$ .

The LRIC (Max), LRIC (MaxMin), LRIC (MaxT), LRIC (MultT) indices calculation was implemented in the programming environment R and Python.

## 4 Experimental Results

To compare the proposed models with previous analogues, we generated various networks and compared their runtime. We focused on the LRIC indices calculation as there is no significant improvement of SRIC calculation. We generated different graphs of the following types:

- Complete graphs;
- Graphs with an exponential degree distribution.

The comparison of the runtime on complete graphs allows us to evaluate the efficiency of the algorithm on dense graphs when each node has a lot of neighbors. Similarly, we tested our algorithms on graphs with exponential degree distribution to examine the computational complexity on sparse graphs.

Since problems of direct and indirect influence detection are separate tasks we compared 3 implementations of the LRIC index:

1.  $LRIC_{naive}$ : Naive LRIC index (existing solution);
2.  $LRIC_{direct}$ : optimized direct influence calculation;
3.  $LRIC_{opt}$ : optimized direct and indirect influence calculation.

The comparison of implementations 1 and 2 shows the improvement of direct influence calculation while the comparison of implementations 2 and 3 shows the improvement of indirect influence calculation.

#### 4.1 Experimental Results on Complete Graphs

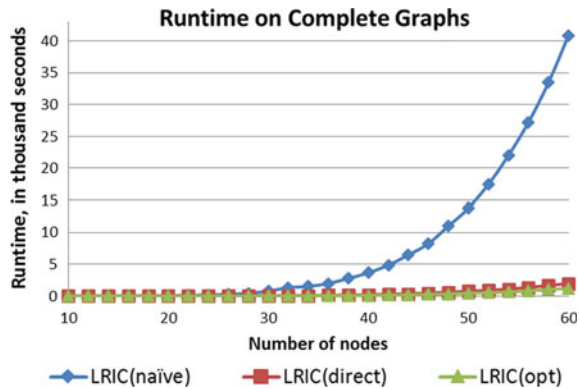
We generated weighted graphs with the total number of nodes from 10 to 60. Since in many cases coordination costs are high, suppose that only small groups of nodes may collaborate to influence other nodes. Thus, we considered all critical groups of maximal size 5, so their total number is  $\sum_{l=1}^5 (l \cdot C_{n-1}^l)$ . We also limit indirect influence through no more than 1 intermediate node.

In Fig. 7, we provide the comparison of the considered algorithms. The main difficulty of the LRIC index calculation is on the evaluation of direct influence which grows exponentially. We can observe that if we consider the problem of direct influence as an integer linear problem and adapt ideas of binary search, we will get 14 times improvement of runtime on average. As for the  $LRIC_{opt}$  algorithm, it results in a 100% improvement of indirect calculation in average.

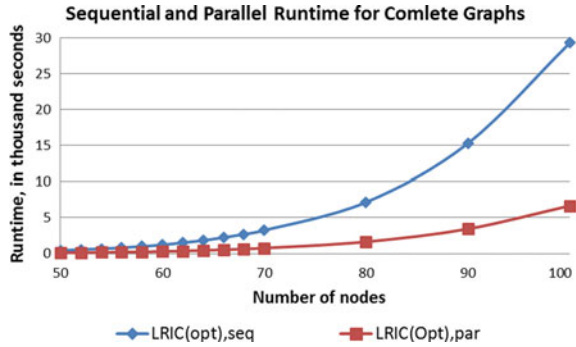
For instance,  $LRIC_{naive}$  algorithm requires 1 min to calculate the LRIC index on complete graphs with 20 nodes while  $LRIC_{direct}$  and  $LRIC_{opt}$  algorithms run in 9 and 4 seconds correspondingly. For complete graphs with 40 nodes, the  $LRIC_{naive}$  algorithms require 1 h to calculate the index while  $LRIC_{direct}$  and  $LRIC_{opt}$  algorithms run in 3,8, and 2 min. Finally, the LRIC index calculation requires more than 11 h in calculation while the proposed algorithm runs in 20 min.

It is clear that the  $LRIC_{naive}$  algorithm cannot be applied in case of a large number of nodes neighbors. Obviously, it is impractical to use this index on networks with hundreds of nodes because of the large number of possible critical groups. However, the runtime efficiency of the proposed algorithms allows to apply the LRIC index on these graphs. Thus, the  $LRIC_{opt}$  algorithm is much more efficient than its previous

Fig. 7 Runtime of LRIC index on complete graphs



**Fig. 8** Sequential and Parallel Runtime of LRICopt algorithm on complete graphs



analogues. We have also compared the runtime of  $LRIC_{opt}$  algorithm in case of the sequential and parallel computation for complete graphs with 50–100 nodes. (see Fig. 8).

According to Fig. 8, the parallel implementation of  $LRIC_{opt}$  algorithm runs 3 times faster than its sequential analogues. We can observe that it allows to find the most influential nodes for graphs with 60 nodes in 4.5 min, 80 nodes in 27 min and 100 nodes in less than 2 h. Thus, we can apply the parallel implementation of the LRIC algorithm to graphs, where each node has a lot of strongly connected neighbors.

Finally, we should note that for complete graphs we considered quota where most neighbors of a node will be pivotal in coalitions. In many networks it is not the case, so the runtime on these networks should be even lower.

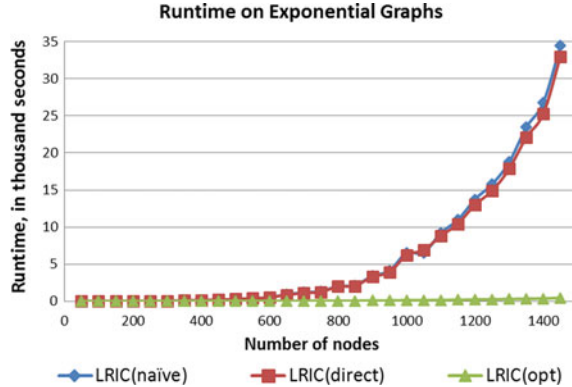
## 4.2 Experimental Results on Complete Graphs

Since most real graphs have an exponential degree distribution, we examined the computational complexity of the LRIC index on these types of graphs. We generated weighted graphs with the total number of nodes from 50 to 1500. Similarly to complete graphs, we considered all critical groups of maximal size 5 while the indirect influence was allowed through no more than 2 intermediate nodes. In Fig. 9, we provide the comparison of the considered algorithms.

According to Fig. 9, contrary to complete graphs, the main difficulty of the LRIC index calculation is on consideration of all possible simple paths between nodes. We can observe that the proposed solution of direct influence calculation provides 5% improvement of runtime in general. However, the proposed approach of simple paths evaluation gives 40–60 times improvement of runtime on average.

Let us describe  $LRIC_{naive}$ ,  $LRIC_{direct}$  and  $LRIC_{opt}$  algorithms.  $LRIC_{naive}$  and  $LRIC_{direct}$  implementation provides similar results in terms of the runtime. For instance, graphs with less than 300 nodes require less than a minute, graphs with 500 nodes runs in 5 min. Unfortunately, these algorithms cannot be applied to graphs with

**Fig. 9** Sequential and Parallel Runtime of LRICopt algorithm on complete graphs



thousands of nodes (graph with 1000 nodes runs in 2.5 h while graphs with 1500 nodes require almost 10 h).

Contrary to other implementations,  $LRIC_{opt}$  algorithm allows to make computations on graphs with thousands of nodes. Graphs with less than 900 nodes run in less than a minute while graphs with 1000 and 1500 require 2 and 7 min correspondingly. Thus, we obtained incredibly large improvement by the proposed models of indirect influence evaluation.

Thus, it is clear that the  $LRIC_{naïve}$  algorithm cannot be applied in case if we consider long connections between nodes. On the other hand, the runtime efficiency of the proposed algorithms allows to apply the LRIC index on large graphs with long connections. Thus, the  $LRIC_{opt}$  algorithm is much more efficient than its previous analogues.

We have also compared the runtime of  $LRIC_{opt}$  algorithm in case of the sequential and parallel computation for exponential graphs with 50–1500 nodes. However, the runtime of the parallel implementation of the algorithm showed 10 times worse performance due to the costs of parallelization. Thus, it is better to apply the sequential implementation of the LRIC algorithm to large graphs.

*Availability and implementation.* The SRIC and LRIC library can be downloaded from this site: <https://github.com/SergSHV/slric>. All versions of SRIC and LRIC indices were implemented in the programming environment R and Python.

## 5 Conclusion

Over the past decades, there has been a huge attention attracted to the evaluation of the most important participants in networks. Since real networks are heterogeneous, nodes may have additional attributes and thresholds of influence and also affect other nodes in groups, classical centrality measures cannot be applied to these networks.



SRIC and LRIC indices relate to a class of indices that take into account these features of networks, however, they have a high computational complexity.

In our work, we focused on SRIC and LRIC indices and proposed several algorithms with a low computational complexity. The problem of directed influence evaluation was reformulated as an integer linear programming problem. Two theorems about properties of pivotal nodes were given. We also transformed the problem of indirect influence evaluation into the problem of matrix multiplication.

To prove the efficiency of proposed modes, we generated various complete and exponential graphs. The algorithm performance showed that we can apply the proposed models to networks with large number of nodes, its neighbors, and long connection between them.

**Acknowledgements** The paper was prepared within the framework of the HSE University Basic Research Program and funded by the Russian Academic Excellence Project '5-100'. This work is also supported by the Russian Foundation for Basic Research under grant -18-01-00804a Power of countries in the food security problem.

## References

1. Freeman, L.C.: Centrality in social networks: conceptual clarification. *Soc Netw* **1**, 215–239 (1979)
2. Bonacich, P.: Technique for analyzing overlapping memberships. *Sociol. Methodol.* **4**, 176–185 (1972)
3. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. *Comput. Netw.* **30**, 107–117 (1998)
4. Katz, L.: A new status index derived from sociometric index. *Psychometrika* 39–43 (1953)
5. Freeman, L.C.: A set of measures of centrality based upon betweenness. *Sociometry* **40**, 35–41 (1977)
6. Freeman, L.C., Borgatti, S.P., White, D.R.: Centrality in valued graphs: a measure of betweenness based on network flow. *Soc. Netw.* **13**, 141–154 (1991)
7. Newman, M.E.J.: A measure of betweenness centrality based on random walks. *Soc. Netw.* **27**, 39–54 (2005)
8. Rochat, Y.: Closeness centrality extended to unconnected graphs: the harmonic centrality index. ASNA (2009)
9. Aleskerov, F.T., Andrievskaya, I.K.: Permjakova Key borrowers detected by the intensities of their short-range interactions. Working papers by NRU Higher School of Economics. Series FE "Financial Economics". No. WP BRP 33/FE/2014 (2014)
10. Aleskerov, F.T., Meshcheryakova, N.G., Shvydun, S.V.: Power in network structures. In: Kalyagin, V.A., Nikolaev, A.I., Pardalos, P.M., Prokopyev, O. (eds.) *Models, Algorithms, and Technologies for Network Analysis*. Springer Proceedings in Mathematics and Statistics, vol. 197, pp. 79–85. Springer International Publishing, Berlin (2017)
11. Aleskerov, F., Meshcheryakova, N., Shvydun, S.: Centrality measures in networks based on nodes attributes, long-range interactions and group influence, arXiv preprint [arXiv:1610.05892](https://arxiv.org/abs/1610.05892)
12. Meshcheryakova, N., Shvydun, S.: Power in network structures based on simulations. In: Cherifi, C., Cherifi, H., Karsai, M., Musolesi, M. (eds.) *Complex Networks and Their Applications VI. Complex Networks: Studies in Computational Intelligence*, vol. 689. Springer, Cham (2018)
13. Aleskerov, F.T.: Power indices taking into account agents preferences. In: Simeone, B., Pukelsheim, F. (eds.) *Mathematics and Democracy*, p. 118. Springer, Berlin (2006)

14. Aleskerov, F.T., Yuzbashev, D.V., Yakuba, V.I.: Threshold aggregation for three graded rankings. *Autom. Remote Control (in Russian)* **1**, 147–152 (2007)
15. Lucas, E.: *Recreations Mathematiques*, Paris (1882)
16. Van Lint, J., Wilson, R.: *A Course in Combinatorics*. Cambridge University Press, Cambridge (2001)
17. Strassen, V.: Gaussian elimination is not optimal. *Numer. Math.* **13**, 354–356 (1969)
18. Coppersmith, D., Winograd, S.: Matrix multiplication via arithmetic progressions. In: *STOC 87: Proceedings of the nineteenth annual ACM symposium on Theory of Computing* (1987)
19. Floyd, R.W.: Algorithm 97: shortest path. *Commun. ACM* **5**(6), 345 (1962)
20. Warshall, S.: A theorem on Boolean matrices. *J. ACM* **9**(1), 11–12 (1962)
21. Dijkstra, E.W.: A note on two problems in connexion with graphs. *Numerische Mathematik* **1**, 269–271 (1959)
22. Johnson, D.B.: Efficient algorithms for shortest paths in sparse networks. *J. ACM* **24**(1), 1–13 (1977)

# A Survey on Variable Neighborhood Search Methods for Supply Network Inventory



Angelo Sifaleras  and Ioannis Konstantaras 

**Abstract** Over the past two decades, reverse logistics and closed-loop supply chain networks have gained substantial interest in business and academia. The dynamic lot-sizing problem with product returns and recovery is one of the most extensively researched topics in inventory control literature. Several interesting generalizations of this optimization problem have lately emerged that include the multiproduct case, the case with capacity constraints, and others. In this chapter, we present recent successful applications of variable neighborhood search (VNS) for the efficient solution of such problems, review the state-of-the-art solution methods, and also discuss some future research directions.

## 1 Introduction

One of the most important issues that should be tackled by a company is the determination of the replenishment policy of the different goods (spare parts, raw material, components, or finished goods) involved in the supply chain network. Manufacturers have started to integrate remanufacturing facilities into the regular production environment. Due to the increasing attention on the sustainability in industry, inventory control problems have been studied in the field of closed-loop systems. This problem becomes more complex when the demand varies with time through a finite planning horizon with  $T$  periods. Several interesting generalizations of this optimization problem have lately emerged.

---

A. Sifaleras (✉)

Department of Applied Informatics, School of Information Sciences,  
University of Macedonia, 156 Egnatia Str., 54636 Thessaloniki, Greece  
e-mail: [sifalera@uom.gr](mailto:sifalera@uom.gr)

I. Konstantaras

Department of Business Administration, School of Business Administration,  
University of Macedonia, 156 Egnatia Str., 54636 Thessaloniki, Greece  
e-mail: [ikonst@uom.gr](mailto:ikonst@uom.gr)

© Springer Nature Switzerland AG 2020

I. Bychkov et al. (eds.), *Network Algorithms, Data Mining, and Applications*, Springer Proceedings in Mathematics & Statistics 315,  
[https://doi.org/10.1007/978-3-030-37157-9\\_5](https://doi.org/10.1007/978-3-030-37157-9_5)

Inventory theory constitutes a large part of supply chain optimization. Advances in inventory theory include, among others, advances in lot-sizing methods [30], forecasting methods for product demands, mathematical models for optimal inventory levels, and others. Inventory optimization problems [6] arise in several cases as a part of modern supply chain networks. Inventory optimization problems can be classified into deterministic or stochastic (e.g., if the demand is assumed to be time varying and not constant). A deterministic optimization problem consists of minimizing a real-valued objective function  $f(x)$ , where  $x$  is a feasible solution belonging in the feasible set  $X$  which is part of the solution space  $S$ . Depending on whether  $S$  is a finite but large set or equal to  $R^n$ , we define either a discrete or a continuous optimization problem, respectively. The optimality of a solution  $x^*$  belonging to  $X$  follows when  $f(x^*)$  is no larger than  $f(x)$  for all feasible solutions  $x$ .

The majority of real-world inventory optimization problems are  $\mathcal{NP}$ -hard problems and are computationally difficult to solve. This is the reason why, several researchers use heuristic and/or metaheuristic solution methods in order to compute optimal or suboptimal solutions in a short amount of time. Metaheuristics (based on local search procedures) try to continue the search by other means after finding the first local minimum.

This survey paper differentiates itself from prior research works in that, to date, the majority of VNS survey papers [4, 9, 10] have mainly focused on the methodological aspects. However, the interest of the researchers working with the VNS methodology has significantly grown [27] and the number of published works on modern supply chain network problems using VNS, has also significantly increased in the literature. Due to this reason, this work aims to present recent advances of VNS applications to various inventory optimization problems such as the economic lot-sizing problem with product returns and recovery (ELSR), the multiproduct dynamic lot-sizing problem with remanufacturing (MDLSRP), the uncapacitated multilevel lot-sizing problems (MLLSs), the capacitated lot-sizing problems (CLSPs), the inventory routing problems (IRPs), and the location inventory routing problems (LIRPs).

This work is structured as follows. Section 2 present a short overview of the VNS methodology. Section 3 presents recent VNS contributions in supply network inventory. Finally, Sects. 4 and 5 draw up concluding remarks and describe some future research directions, respectively.

## 2 Variable Neighborhood Search

Variable neighborhood search (VNS) is a metaheuristic based on systematic changes in the neighborhood structure. In spite of its simplicity, it has provided a huge number of successful applications in a wide range of areas. The success of VNS is based on three simple facts: (i) A local minimum with respect to one neighborhood structure is not necessarily so for another, (ii) a global minimum is a local minimum with respect to all possible neighborhood structures, and (iii) for many problems, local minima with respect to one or several  $N_k$  are relatively close to each other. Let  $N_k$ , ( $k = 1$ ,

$\dots, k_{max}$ ) be a finite set of preselected neighborhood structures. Let  $N_k(x)$  be the set of solutions in the  $k$ th neighborhood of  $x$ . Most local search heuristics use only one neighborhood structure, i.e.,  $k_{max} = 1$ . An optimal solution  $x_{opt}$  (or global minimum) is a feasible solution where a minimum is reached. We call  $x'$  a local minimum with respect to  $N_k$  (w.r.t.  $N_k$  for short), if there is no feasible solution  $x \in N_k(x')$  such that  $f(x) < f(x')$ .

VNS has shown to be very successful for solving hard either combinatorial or global optimization problems. Moreover, a large number of VNS variants have been already proposed in the literature, e.g., reduced VNS (RVNS), basic VNS (BVNS), variable neighborhood descent (VND), general VNS (GVNS), skewed VNS (SVNS), nested VNS (NVNS), variable neighborhood decomposition search (VNDS). All these VNS variants exhibit differences either in the intensification and/or in the diversification phase. Generally speaking, a VNS method alternately executes the improvement phase (intensification phase) and the shaking procedure (diversification phase) together with the neighborhood change step until some stopping criterion is fulfilled.

### 3 VNS Contributions to Supply Chain Optimization Problems

#### 3.1 VNS for the ELSR

The economic lot-sizing problem with product returns and recovery has the following problem parameters: a number of periods ( $T$ ), setup/holding costs ( $k_M, k_R, h_M, h_R$ ), the demand for each time period ( $D(t)$ ) and also, the number of returned items per period and product ( $R(t)$ ) that can be completely remanufactured and sold as new. The ELSR aims to compute the number of new  $x_M(t)$  and/or remanufactured products  $x_R(t)$  and also, the inventory level of serviceable items  $y_M(t)$  and/or items that can be remanufactured  $y_R(t)$  per period. The objective of the problem is to minimize the total cost due to manufacturing and/or remanufacturing setup cost, and also, holding cost for the serviceable items and/or recoverable items per unit time.

In [26], several new neighborhoods for this combinatorial optimization problem were presented and an efficient local search method for exploring them was described. The computational results obtained on an established set of benchmark problems with 6,480 instances showed that the VNS metaheuristic algorithm outperformed previous state-of-the-art heuristic methods from the literature.

### 3.2 VNS for the MDLSRP

An interesting generalization occurs by considering multiple products and thus now there is a number of periods ( $T$ ), a number of different products ( $K$ ), setup/holding costs ( $k_M(k)$ ,  $k_R(k)$ ,  $h_M(k)$ ,  $h_R(k)$ ), the demand for each time period and product ( $D(k, t)$ ) and also the number of returned items per period and product ( $R(k, t)$ ) that can be completely remanufactured and sold as new. The MDLSRP aims to compute the number of new  $x_M(k, t)$  and/or remanufactured products  $x_R(k, t)$  and also, the inventory level of serviceable items  $y_M(k, t)$  and/or items that can be remanufactured  $y_R(k, t)$  per period. The objective of the problem is again to minimize the total cost due to manufacturing and/or remanufacturing setup cost, and also, holding cost for the serviceable items and/or recoverable items per unit time.

In [24, 25], a subset of the most efficient neighborhoods for the ELSR was appropriately generalized for the MDLSRP, since the local search part now requires more time. The authors proposed and solved a new, larger, benchmark set with 300 products, 52 periods (publicly available from: <http://users.uom.gr/~sifalera/benchmarks.html>). Furthermore, the VNS approach outperformed the state-of-the-art Gurobi optimizer. Also, the authors have implemented a simple constructive Heuristic for the MDLSRP, where the total demand is fulfilled by a single lot (without remanufacturing units) for each product in the first period, in order to find an initial solution to the MDLSRP. This approach did not depend on any other, commercial or not, solver for either computing a starting solution or an intermediate computation. Thus, it was a self-contained solver without any link to other callable library API.

### 3.3 VNS for MLLSs

Xiao et al. [32] in 2011 presented a reduced variable neighborhood search algorithm for solving uncapacitated multilevel lot-sizing (MLLS) problems. Such uncapacitated MLLS problems constitute a basic form of several other extended versions of the MLLS problem under various constraints. The authors used the modified randomized cumulative Wagner–Whitin (RCWW) method for computing initial solutions and also applied six other techniques in order to enhance the efficiency of the searching process. The same authors [33] in 2012, investigated the use of three indexes, i.e., distance, changing range, and changing level, for characterizing the neighborhood of the incumbent solution in three aspects, and improving the efficacy of a VNS algorithm.

Seeaner et al. [23] in 2013 combined the variable neighborhood decomposition search method and the MIP-based fix & optimize heuristic for solving problems that typically arise in the consumer goods industry. Such problems require simultaneously determining lot-sizes and sequences for parallel lines and multiple production stages. The authors defined different neighborhood structures based on product, (micro) period, production line decompositions, and bill-of-materials.

Furthermore, Xiao et al. [34] in 2014, developed the Ancestors Depth-first Traversal Search (ADTS), which is an effective local search method and can be embedded in the VNS framework for significant improvements of the solution quality. Additionally, the authors proposed a new formulation for the MLLS problem which allowed them to devise a Recursive Accumulation of the Cost Reduction (RACR) method for fast updating of the objective function after a change of an incumbent solution.

### 3.4 VNS for CLSPs

Almada-Loboa and James [1] in 2010, addressed the multi-item capacitated lot-sizing and scheduling problem with sequence-dependent setup times and costs, which is an extension of the Wagner–Whitin model taking also into account capacity constraints. The authors applied a construction heuristic that contains five, forward and backward, steps that are able to efficiently compute feasible solutions and then employed a VNS method for the improvement of this solution for this  $\mathcal{NP}$ -hard problem. Three different types of moves (neighborhoods) were used in the improvement phase of the VNS: (i) an insertion move which takes a job and inserts it before another job in the sequence, (ii) a fractional insertion move which furthermore allows the option of splitting a job into two jobs, and (iii) a swap move which takes two jobs in the sequence and swaps their location in the sequence. According to the authors, if sufficient time was available, they found that the proposed VNS method outperformed the tabu search method for the same problem.

Chen [5] in 2015, proposed a new fix-and-optimize (FO) approach for two dynamic multilevel capacitated lot-sizing problems (MLCLSP). Based on their FO method, the authors also develop a VNS approach for the MLCLSP without setup carryover, which can further improve the solution obtained by the FO, by diversifying the search space. Also, numerical experiments on benchmark instances showed that, their VNS approach outperforms other previous FO approaches.

Recently, Liuxi et al. [17] in 2017, motivated from a real-world steel enterprise problem, proposed a piecewise linear approximation method for solving two nonlinear mathematical models for stochastic multi-item capacitated lot-sizing problems with and without setup carryovers. Moreover, the authors presented an integrative approach combining a FO method and VNS (FO-VNS), for the efficient solution of these approximation models.

### 3.5 VNS for IRPs

Lejeune [16] in 2006, considered a three-stage supply chain which combines an inventory–production–distribution plan over a multi-period horizon. The authors proposed a VNDS method, where the decomposition scheme is related to the relaxation of the integrality constraints on the integer decision variables. The authors

validated their VNDS method using a real-industrial life problem faced by a large North American chemical company and outperformed other commercial MIP solvers.

Liu and Lee [19] in 2011, applied a hybrid method that integrates VNS and Tabu Search, in order to address the inventory routing problem with time windows. Later in 2012, Liu and Chen [18] proposed a VNS for simultaneously solving an integrated model for the inventory routing and scheduling problem (IRSP) rather than sequentially and separately taking the inventory, routing, and scheduling decisions, based on the minimal cost criterion. Also in 2012, Popović et al. [22] applied a Randomized Variable Neighborhood Descent (RVND) method for minimizing the total cost of a multiproduct multi-period Inventory Routing Problem (IRP) in fuel delivery with multi-compartment homogeneous vehicles, with a deterministic consumption that varies with each petrol station and each fuel type. Thus, this IRP resulted from the combination of a Vehicle Routing Problem (VRP) in fuel delivery, including petrol stations inventory management. Their proposed stochastic RVND method was based on random changes of the neighborhoods rather than changing of neighborhoods in a deterministic way as for example in the VND method.

Mjirda et al. [20] in 2014 presented a two-phase VNS metaheuristic method for solving a multiproduct inventory routing problem. The authors applied a VNS method in the first phase, for solving a capacitated VRP at each period and finding an initial solution without taking into account the inventory. Afterward, an iterative improvement of the initial solution while minimizing both the transportation and inventory costs is taking place in the second phase. The proposed VNS used four neighborhood structures while the shaking part alternated between only two neighborhood structures.

Hasni et al. [11] in 2017 proposed a GVNS method for minimizing both the transportation and inventory holding costs of a multiproduct deterministic version of the IRP, with a fleet of heterogeneous vehicles and known demands. Their GVNS approach utilized six different neighborhoods, i.e., insertion moves in the same route or between two routes from the same period, or between two routes in different periods, and exchange moves within the same route, or between two routes from the same period, or between two routes from different periods. The authors demonstrated the efficiency of their approach against a branch & cut algorithm, using benchmark instances ranging from 10 to 100 customers, and with a maximum of seven periods, five vehicles, and five products.

More recently Gruler et al. [8] in 2018 presented a simulation–optimization (simheuristic) approach integrating Monte Carlo simulation (MCS) within a VNS framework to solve the multi-period IRP with stochastic customer demands. The authors aimed at computing the optimal refill policies for each customer–period combination. Thus, a constructive procedure using MCS techniques was applied for generating an initial solution, and afterward it was iteratively improved by the VNS framework.



### **3.6 VNS for LIRPs**

Turan et al. [29] in 2017 tackled a two-echelon inventory routing problem with uncertain demand, where perishable products are distributed from one central warehouse to multiple retail stores. The authors proposed a VNS procedure, which was extended with a dynamic programming subroutine to handle inventory allocation. The intensification (local search) phase consisted of randomly selecting one of the following operators (neighborhoods): Intra-route Two-Opt, Intra-route Swap, and Intra-route Relocate. The diversification (shaking) phase, was similarly based on a random selection of one of the following operators: Inter-route Two-opt, Inter-route Swap, and Inter-route Relocate.

Recently, Karakostas et al. [14] in 2018 proposed two BVNS metaheuristic approaches for the solution of a more realistic LIRP version, which integrates economic and environmental decisions. This new variant is denoted as the Pollution LIRP (PLIRP). The same authors [15] proposed in 2019 a GVNS metaheuristic method for the solution of the Location Inventory Routing with Distribution Outsourcing (LIRPDO). This new problem variant represents a more realistic situation, in which a company is required to outsource its distribution operation, (e.g., in cases where it cannot afford vehicles acquisition). Therefore, more decisions, such as the most efficient allocation of the company's opened depots to the selected providers and the selection of the proper vehicles providers needs to be made.

### **3.7 VNS for Supplier Selection and Performance Evaluation in Inventory Control**

Since traditional methods often fail to identify the factors affecting supplier selection and evaluation, Vahdani et al. [3] in 2017 proposed a novel hybrid approach based on continuous GVNS and least square-support vector machine (LS-SVM). The authors applied GVNS to improve the generalization performance in searching the support vector machine model, which is used to build the nonlinear relationship between selection, evaluation criteria, and performance rating of suppliers. Additionally, a real-world case study of a supplier selection and evaluation problem from the cosmetics industry is shown for demonstrating the performance of the proposed integrated model.

### **3.8 VNS for Sustainable Order Allocation (inventory) and Sustainable Supply Chain**

Govindan et al. [7] in 2015 applied a novel multi-objective hybrid optimization approach in order to simultaneously minimize the total costs and environ-

mental effects, occurred by strategically designing a sustainable forward supply chain network (SCN) with stochastic demand. The authors combined an adaptive multi-objective variable neighborhood search (AMOVNS) method as a local search together with an adapted multi-objective electromagnetism mechanism algorithm. Their proposed supply chain network was composed of five echelons including suppliers classified in different classes, plants, distribution centers, and cross-docks. Furthermore, a real-world case study from an automobile industry was used to demonstrate the applicability of this approach.

### ***3.9 VNS for EOQ-Based Inventory Problems***

Recently, Đorđević et al. [21] in 2017 presented a BVNS method for tackling the static time-continuous multiproduct economic order quantity (EOQ) based inventory management problem with the storage space constraints. Also, the authors examined the efficiency of their approach using a preliminary computational analysis with randomly generated instances with the same number of products.

### ***3.10 VNS for Joint Replenishment Problems***

Wang et al. [31] in 2018 approximately obtained lower and upper bounds of joint replenishment and delivery problem (JRD) problem using a bounding procedure. Afterward, they applied a VNS method utilizing several new neighborhood structures with the bounding method to solve the JRD problem. This hybrid VNS was shown to perform better than the best known heuristic and metaheuristic approaches for JRD in terms of accuracy, using randomly generated examples.

### ***3.11 VNS for Vendor Managed Inventory Problems***

Hemmelmayr et al. [12] in 2010, applied a VNS approach for optimal planning delivery routes for the supply of blood products to hospitals by a blood bank. Their method allowed for low-cost and robust plans that hedge against the natural uncertainty associated with blood product usage at hospitals. The authors used 15 neighborhoods that were defined from combinations of three operators, i.e., move, cross-exchange, and change-combination.

### 3.12 VNS for Multi-criteria Inventory Classification

Kaabi et al. [13] in 2015 proposed an automatic learning method (ALM) that combined the benefits of artificial intelligence (AI) and multi-criteria decision-making (MCDM) techniques. The authors applied a continuous VNS to infer the criteria weights of the ABC analysis used in inventory management, in order to produce a classification of items that minimizes the inventory cost. Furthermore, the technique for order preference by similarity to ideal solution (TOPSIS) method was used to compute the items score, using an aggregation function combining the item evaluations on the different criteria and their weights. Also, the performance of the proposed ALM with respect to some other ABC inventory classification models was evaluated based on a benchmark data set of 47 items from an Hospital Respiratory Therapy Unit.

## 4 Conclusions

This chapter presented recent successful applications of VNS for the efficient solution of inventory optimization problems that often arise in modern closed-loop supply chain networks. A thorough literature review on several state-of-the-art VNS solution methods is intended as a reference for both scientists and practitioners in these disciplines. Table 1 summarizes several research works and provides a taxonomy based on different problem variants and their characteristics.

## 5 Future Research Guidelines

Skouri et al. [28] in 2018 recently presented a survey of open problems in green supply chain modeling and optimization with carbon emission targets. Nowadays, research on pollutant emissions management is very important and a large number of, computationally difficult, inventory lot-sizing optimization problems arise that could improve the effectiveness of carbon management in the supply chain.

In order to tackle such hard optimization problems, a future research direction consists of exploiting parallelism in multicore CPU+GPU systems. Antoniadis and Sifaleras [2] in 2017 proposed a hybrid parallel VNS method using a CPU-GPU system via OpenMP and OpenACC on a set of recent benchmark problem instances for the multiproduct dynamic lot-sizing problem with product returns and recovery, which appears in reverse logistics and is known to be  $\mathcal{NP}$ -hard.

**Table 1** Key literature VNS contributions on supply network inventory problems

Reference	Problem type	Demand type	Commodity type	Level type	Capacitated version	Closed-loop supply chain	VNS variant
[26]	Inventory	Deterministic	Single	Single	-	✓	GVNS
[24]	Inventory	Deterministic	Multiple	Single	-	✓	VND
[25]	Inventory	Deterministic	Multiple	Single	-	✓	GVNS
[32]	Inventory	Deterministic	Multiple	Multilevel	-	-	RVNS
[23]	Inventory	Deterministic	Multiple	Multilevel	Product lines	-	VNDS
[33]	Inventory	Deterministic	Multiple	Multilevel	-	-	Iterated VNS
[34]	Inventory	Deterministic	Multiple	Multilevel	-	-	VNS modification
[1]	Inventory	Deterministic	Multiple	Single	Product lines	-	GVNS
[5]	Inventory	Deterministic	Multiple	Multilevel	Product lines	-	Hybrid VNS with fix-and-optimize heuristic
[17]	Inventory	Stochastic	Multiple	Single	Product lines	-	Hybrid VNS with fix-and-optimize heuristic
[18]	Inventory/Routing/Scheduling	Deterministic	Single	Single	Vehicles	-	VNS modification
[19]	Inventory Routing	Deterministic	Single	Single	Vehicles	-	Hybrid VNS with Tabu Search
[16]	Inventory Routing	Deterministic	Single	Single	Vehicles	-	VNDS
[22]	Inventory/Routing	Deterministic	Multiple	Single	Reservoir for fuel	-	RVND
[20]	Inventory Routing	Deterministic	Multiple	Single	Vehicles	-	VND, VNS
[11]	Inventory Routing	Deterministic	Multiple	Single	Customers, Vehicles	-	GVNS
[8]	Inventory Routing	Stochastic	Multiple	Single	Retail centers, Vehicles	-	Simheuristic VNS with Monte Carlo
[29]	Location/Inventory/Routing	Stochastic	Multiple	Single	-	-	Hybrid VNS with dynamic programming
[14]	Location/Inventory/Routing	Deterministic	Single	Single	Depots, Vehicles	-	BYNS
[15]	Location/Inventory/Routing	Deterministic	Single	Single	Depots, Vehicles	-	GVNS

## References

1. Almada-Lobo, B., James, R.J.: Neighbourhood search meta-heuristics for capacitated lot-sizing with sequence-dependent setups. *Int. J. Prod. Res.* **48**(3), 861–878 (2010)
2. Antoniadis, N., Sifaleras, A.: A hybrid CPU-GPU parallelization scheme of variable neighborhood search for inventory optimization problems. *Electron. Notes Discret. Math.* **58**, 47–54 (2017)
3. Branch, S.T.: A new enhanced support vector model based on general variable neighborhood search algorithm for supplier performance evaluation: a case study. *Int. J. Comput. Intell. Syst.* **10**, 293–311 (2017)
4. Caporossi, G., Hansen, P., Mladenović, N.: Variable neighborhood search. In: *Metaheuristics*, pp. 77–98. Springer, Berlin (2016)
5. Chen, H.: Fix-and-optimize and variable neighborhood search approaches for multi-level capacitated lot sizing problems. *Omega* **56**, 25–36 (2015)
6. Cunha, J.O., Konstantaras, I., Melo, R.A., Sifaleras, A.: On multi-item economic lot-sizing with remanufacturing and uncapacitated production. *Appl. Math. Model.* **43**, 678–686 (2017)
7. Govindan, K., Jafarian, A., Nourbakhsh, V.: Bi-objective integrating sustainable order allocation and sustainable supply chain network strategic design with stochastic demand using a novel robust hybrid multi-objective metaheuristic. *Comput. Oper. Res.* **62**, 112–130 (2015)
8. Gruler, A., Panadero, J., de Armas, J., Pérez, J.A.M., Juan, A.A.: A variable neighborhood search simheuristic for the multiperiod inventory routing problem with stochastic demands. *Int. Trans. Oper. Res.* (2018)
9. Hansen, P., Mladenović, N., Brimberg, J., Pérez, J.A.M.: Variable neighborhood search. In: M. Gendreau, J.Y. Potvin (eds.) *Handbook of Metaheuristics*, pp. 57–97. Springer, Berlin (2019)
10. Hansen, P., Mladenović, N., Todosijević, R., Hanafi, S.: Variable neighborhood search: basics and variants. *EURO J. Comput. Optim.* **5**(3), 423–454 (2017)
11. Hasni, S., Toumi, S., Jarboui, B., Mjirda, A.: GVNS based heuristic for solving the multi-product multi-vehicle inventory routing problem. *Electron. Notes Discret. Math.* **58**, 71–78 (2017)
12. Hemmelmayr, V., Doerner, K.F., Hartl, R.F., Savelsbergh, M.W.: Vendor managed inventory for environments with stochastic product usage. *Eur. J. Oper. Res.* **202**(3), 686–695 (2010)
13. Kaabi, H., Jabeur, K., Enneifar, L.: Learning criteria weights with TOPSIS method and continuous VNS for multi-criteria inventory classification. *Electron. Notes Discret. Math.* **47**, 197–204 (2015)
14. Karakostas, P., Sifaleras, A., Georgiadis, M.C.: Basic VNS algorithms for solving the pollution location inventory routing problem. In: *Variable Neighborhood Search. ICVNS 2018. Lecture Notes in Computer Science*, vol. 11328, pp. 64–76. Springer, Cham, Sifonia, Halkidiki, Greece (2019)
15. Karakostas, P., Sifaleras, A., Georgiadis, M.C.: A general variable neighborhood search-based solution approach for the location-inventory-routing problem with distribution outsourcing. *Comput. Chem. Eng.* **126**, 263–279 (2019)
16. Lejeune, M.A.: A variable neighborhood decomposition search method for supply chain management planning problems. *Eur. J. Oper. Res.* **175**(2), 959–976 (2006)
17. Li, L., Song, S., Wu, C., Wang, R.: Fix-and-optimize and variable neighborhood search approaches for stochastic multi-item capacitated lot-sizing problems. *Math. Probl. Eng.* **2017**, (2017)
18. Liu, S.C., Chen, A.Z.: Variable neighborhood search for the inventory routing and scheduling problem in a supply chain. *Expert. Syst. Appl.* **39**(4), 4149–4159 (2012)
19. Liu, S.C., Lee, W.T.: A heuristic method for the inventory routing problem with time windows. *Expert. Syst. Appl.* **38**(10), 13223–13231 (2011)
20. Mjirda, A., Jarboui, B., Macedo, R., Hanafi, S., Mladenović, N.: A two phase variable neighborhood search for the multi-product inventory routing problem. *Comput. Oper. Res.* **52**, 291–299 (2014)

21. Dorđević, L., Antić, S., Čangalović, M., Lisec, A.: A metaheuristic approach to solving a multiproduct EOQ-based inventory problem with storage space constraints. *Optim. Lett.* **11**(6), 1137–1154 (2017)
22. Popović, D., Vidović, M., Radivojević, G.: Variable neighborhood search heuristic for the inventory routing problem in fuel delivery. *Expert. Syst. Appl.* **39**(18), 13390–13398 (2012)
23. Seeanner, F., Almada-Lobo, B., Meyr, H.: Combining the principles of variable neighborhood decomposition search and the fix & optimize heuristic to solve multi-level lot-sizing and scheduling problems. *Comput. Oper. Res.* **40**(1), 303–317 (2013)
24. Sifaleras, A., Konstantaras, I.: General variable neighborhood search for the multi-product dynamic lot sizing problem in closed-loop supply chain. *Electron. Notes Discret. Math.* **47**, 69–76 (2015)
25. Sifaleras, A., Konstantaras, I.: Variable neighborhood descent heuristic for solving reverse logistics multi-item dynamic lot-sizing problems. *Comput. Oper. Res.* **78**, 385–392 (2017)
26. Sifaleras, A., Konstantaras, I., Mladenović, N.: Variable neighborhood search for the economic lot sizing problem with product returns and recovery. *Int. J. Prod. Econ.* **160**, 133–143 (2015)
27. Sifaleras, A., Salhi, S., Brimberg, J. (eds.): Variable Neighborhood Search - 6th International Conference, ICVNS 2018, Sithonia, Greece, October 4–7, 2018, Revised Selected Papers. *Lecture Notes in Computer Science*, vol. 11328. Springer, Cham (2019)
28. Skouri, K., Sifaleras, A., Konstantaras, I.: Open problems in green supply chain modeling and optimization with carbon emission targets. In: P.M. Pardalos, A. Migdalas (eds.) *Open Problems in Optimization and Data Analysis*, vol. 141. Springer Optimization and Its Applications (2018)
29. Turan, B., Minner, S., Hartl, R.F.: A VNS approach to multi-location inventory redistribution with vehicle routing. *Comput. Oper. Res.* **78**, 526–536 (2017)
30. Wagner, H., Whitin, T.: Dynamic version of the economic lot size model. *Manag. Sci.* **5**, 89–96 (1958)
31. Wang, L., Liu, R., Liu, S.: Variable neighborhood search incorporating a new bounding procedure for joint replenishment and delivery problem. *J. Oper. Res. Soc.* **69**(2), 201–219 (2018)
32. Xiao, Y., Kaku, I., Zhao, Q., Zhang, R.: A reduced variable neighborhood search algorithm for uncapacitated multilevel lot-sizing problems. *Eur. J. Oper. Res.* **214**(2), 223–231 (2011)
33. Xiao, Y., Kaku, I., Zhao, Q., Zhang, R.: Neighborhood search techniques for solving uncapacitated multilevel lot-sizing problems. *Comput. Oper. Res.* **39**(3), 647–658 (2012)
34. Xiao, Y., Zhang, R., Zhao, Q., Kaku, I., Xu, Y.: A variable neighborhood search with an effective local search for uncapacitated multilevel lot-sizing problems. *Eur. J. Oper. Res.* **235**(1), 102–114 (2014)

# **Network Data Mining**

# GSM: Inductive Learning on Dynamic Graph Embeddings



Marina Ananyeva, Ilya Makarov and Mikhail Pendiukhov

**Abstract** In this paper, we study the problem of learning graph embeddings for dynamic networks and the ability to generalize to unseen nodes called inductive learning. Firstly, we overview the state-of-the-art methods and techniques for constructing graph embeddings and learning algorithms for both transductive and inductive approaches. Secondly, we propose an improved GSM based on GraphSAGE algorithm and set up the experiments on datasets CORA, Reddit, and HSEcite, which is collected from Scopus citation database across the authors with affiliation to NRU HSE in 2011–2017. The results show that our three-layer model with attention-based aggregation function, added normalization layers, regularization (dropout) outperforms suggested by the respective authors' GraphSAGE models with mean, LSTM, and pool aggregation functions, thus giving more insight into possible ways to improve inducting learning model based on GraphSAGE model.

**Keywords** Graph embeddings · Dynamic graphs · Inductive learning approach

---

M. Ananyeva · I. Makarov (✉)  
National Research University Higher School of Economics,  
3 Kochnovskiy Proezd, 107207 Moscow, Russian Federation  
e-mail: [iamakarov@hse.ru](mailto:iamakarov@hse.ru)

M. Ananyeva  
e-mail: [ananyeva.me@gmail.com](mailto:ananyeva.me@gmail.com)

I. Makarov  
University of Ljubljana, Faculty of Computer and Information Science,  
Vecna pot 113, 1000 Ljubljana, Slovenia

M. Pendiukhov  
Analytical Software Solutions LLC, 3 proezd Marinoy Roshchi 40, bld. 1,  
127018 Moscow, Russian Federation  
e-mail: [dx@apsolutions.ru](mailto:dx@apsolutions.ru)

© Springer Nature Switzerland AG 2020  
I. Bychkov et al. (eds.), *Network Algorithms, Data Mining,  
and Applications*, Springer Proceedings in Mathematics & Statistics 315,  
[https://doi.org/10.1007/978-3-030-37157-9\\_6](https://doi.org/10.1007/978-3-030-37157-9_6)



# 1 Introduction

Real-world networks are full of useful information that can be extracted to help solve complex problems in various application fields. To make this goal feasible, networks have to be transformed into simplified representations called graphs. Graph-based models are used in a wide range of application areas: social network analysis [1–3], trade and financial transactions [4], co-authorship networks [5–7], neural connections in the human brain [8], biochemical protein–protein interactions [9]. There are several conventional ways to operate with graph models, which are based either on the basic data representation structures of the original graph (e.g., adjacency list, incidence or adjacency matrices) or on the graph representation in vector space over real numbers. In this work, we are going to focus on the latter approach.

Graph embedding is an effective method to convert the initial graph into a space of lower dimension. In particular, it learns a mapping from a given graph to a vector space and optimizes it by finding the best possible option to preserve the network properties and graph structural information. The problem of graph embeddings lies between machine learning task and representation learning. The problem setting can be broadly divided into four categories: (1) node classification, (2) link prediction, (3) visualization, and (4) clustering [10, 11]. In this paper, we set the node classification task. The second research problem aims to obtain the best possible data representations, and it covers supervised, semi-supervised, and unsupervised tasks for learning embeddings. Vector representations are more efficient and simple to work within data science tasks in comparison to raw graphs, which are limited to a certain subset of machine learning and statistical models [12].

Most of the existing embedding approaches are mainly designed for static graphs. In fact, many real-world networks have dynamic nature, e.g., financial transactions flows, reposts graphs on Facebook, citations in arXiv. In these cases, it may seem more reasonable to model dynamic graphs considering its changes over time. These graphs can be called dynamic in terms of either node or edge temporal data or graph structure. In our case, we use the second definition. It means that instead of considering the dynamics of the internal changing information of nodes (or edges), we consider that the graph structure can evolve over time periods by acquiring, preserving, or losing its nodes and edges. In order to extend the static embedding algorithms, dynamic graphs are often divided into several snapshots that represent the state of the graph at some time point. Transductive algorithms are applied to each snapshot independently, however the final results still remain unsatisfactory and unfold a promising direction of further scientific studies. Poor performance is mostly explained by the challenges needed to overcome while working with dynamic graphs: scalability, efficiency, stability, and flexibility. The following research questions stay open for dynamic case. What are the effective ways to use only a local structure for node embeddings instead of the whole graph? How to chose the node’s neighborhood? Which orders of proximity should be considered? What are the best practices for aggregating information from neighbors?

In this research, we propose an improved model *GSM* (*GraphSAGE attention-Modified*), one of the most effective models with inductive extension, and test it on the open-source datasets (CORa, Reddit), including our own data collected from Scopus that we called HSEcite. We demonstrate its performance on the multilabel classification task. The results of applying our model reveal the significance of a deliberate choice of aggregation functions, normalization, and regularization for GraphSAGE-based algorithms. We put the experiments for different types of aggregation functions (mean, max pooling, LSTM, attention-based), apply normalization and dropout (regularization) for layers.

In Sect. 2, we give the basic definitions and formalize the problem. Section 3 provides an overview of transductive graph embedding methods. The CNN, GAE, and most notable algorithms for inductive learning embeddings on dynamic graphs are listed in Sect. 4. Finally, Sect. 5 contains the set up of experiments and discussion of the results.

## 2 Notation and Basic Definitions

First, we denote the basic definitions. Let us assume the input of representation learning algorithm being an undirected graph denoted as follows:

**Definition 1** *Graph* is an unordered pair  $G = (V, E)$ , where  $V$  is a set of vertices  $v \in V$ , and  $E$  is a set of edges  $e \in E$ .

Matrix  $A$  will be associated with binary adjacency matrix of size  $n \times n$ . We also make an assumption that the algorithms can take an input of a real-valued matrix of vertex attributes which represent its metadata  $X \in \mathbb{R}^{k \times |V|}$ , where  $k$  is a number of attributes. Hence, the goal of embedding is to leverage the information from matrices  $A$  and  $X$  to map each vertex into a vector  $\mathbf{v} \in \mathbb{R}^d$ ,  $d \ll |V|$ . We denote the first-order and second-order proximities the same way as Cai et al. [12]:

**Definition 2** The *first-order proximity*  $s_{ij}^{(1)}$  between vertices  $i$  and  $j$  is called the weight of the edge  $e_{ij}$ , which in simple case equals  $A_{i,j}$ .

Hence, let  $s_i^{(1)} = [s_{i,1}^{(1)}, s_{i,2}^{(1)}, \dots, s_{i,|V|}^{(1)}]$  define the first-order proximity between  $v_i$  and other vertices. The more two connected vertices are similar, the greater is the weight of their common edge.

**Definition 3** The *second-order proximity*  $s_{ij}^{(2)}$  between vertices  $i$  and  $j$  is a similarity between  $s_i^{(1)}$  and  $s_j^{(1)}$ ,  $i$ 's and  $j$ 's neighborhoods.

For comparison of nodes' neighbourhoods  $s_i^{(1)}$  and  $s_j^{(1)}$  cosine similarity, Jaccard index or any other applicable measure can be considered. The proximity between two nodes is equal to zero in case they do not share common neighbors. The abovementioned definitions of proximities will help us to express the meaning of embedding.

**Definition 4** *Graph embedding* learns a mapping  $f : i \rightarrow \mathbf{i} \in \mathbb{R}^d$ , where  $d \ll |V|$ . The function’s objective is to produce the similarity between  $\mathbf{i}$  and  $\mathbf{j}$  preserving the first- and second-order proximities of nodes  $i$  and  $j$  as much as possible.

Formally, Leskovec and Zitnik [13] defined the following steps:

1. Denote an encoder (such function  $\phi(u)$  that maps a node  $u$  into a  $d$ -dimensional vector  $\mathbf{u}$ ).
2. Define a similarity function for nodes of initial graph.
3. Optimize parameters of the encoder:  $\text{similarity}(u, v) \approx \phi_u^T \phi_v$ .

We expect the vertices which are close to each other in the original graph get a close representation in the vector space. Generally, embeddings methods use various notions of closeness between nodes: connectivity, common neighbors, similarity of local structure. For instance, if we assume the weight of an edge as a good measure of proximity, it approximately equals the scalar product of nodes’ embeddings. Hence, the angle between the vectors of close nodes should be minimal. The loss function will take the following form:

$$L = \sum_{(u,v) \in V \times V} \|\phi_u^T \phi_v - A_{u,v}\|^2.$$

Alternatively, we can define the node’s higher order neighborhoods ( $k$ -hop neighbors). We discuss these differences in detail in subsection dedicated to random walk based methods. We also have to provide the essential definitions for generalization to the dynamic case.

**Definition 5** *Dynamic graph* is a sequence of static graphs over discrete time steps called snapshots, i.e.,  $G = \{G_1, \dots, G_T\}$ , where  $G_t = (V_t, E_t)$ ,  $T$ —the number of snapshots.

We assume the setup with expanding graphs, allowing new vertices to join the dynamic graph  $V_t \subseteq V_{t+1}$  and to add edges to existing vertices  $E_t \subseteq E_{t+1}$ . The disappeared vertices and edges can remain as part of the graph with zero weight to others, but we use the datasets with evolving structures with all previous connections included. By considering embeddings on a dynamic graph, we extend the notion for dynamic graph embedding.

**Definition 6** *Dynamic graph embedding* is a time series of mapping functions  $F = \{f_1, \dots, f_T\}$  on dynamic graph  $G = \{G_1, \dots, G_T\}$ , such that  $f_t$  corresponds to a graph embedding for  $G_t$  and all mappings preserve the proximity measure, respectively, to their graphs.

Most of the graph-based methods optimize the mappings as *unsupervised* task, using only the information from matrices  $A$  and  $X$  and generating scores based on sampled paths or node neighbors. Other models use node labels for optimization of embeddings during *supervised* representation learning. In *semi-supervised* tasks, the

data contains labeled and mostly unlabeled instances. Let  $L$  and  $L^*$  be the numbers of labeled and unlabeled objects. Then, we define  $[v_1 : v_L] = [v_1, v_2, \dots, v_L]$  and  $[v_{L+1} : v_{L+L^*}]$  as feature vectors of labeled and unlabeled objects. The labels are  $[y_1 : y_L]$ . Based on the given sample, the goal is to learn a classifier  $f : v \rightarrow v$  and to use the labeled objects to improve the overall performance based only on a small labeled set. The hypothesis in graph-based semi-supervised learning implies that nearby nodes tend to have the same labels. Transductive learning applies this classifier  $f$  on unlabeled instances observed at training time. Inductive learning tries to learn a parameterized classifier, which can be generalized on unseen instances. In this research, we are interested in a more complicated inductive case.

Considering the evolving network, a sample above includes only observed objects called out-of-sample nodes, for which we want to infer the learned embeddings. For instance, Hamilton et al. [14] suggested inductive node classification for texts and protein–protein interactions.

The above basic concepts and definitions will be useful for a better understanding of the descriptions of the algorithms in the following sections.

### 3 Related Research

Transductive methods imply that nodes or edges can be predicted only for ones observed during training time and do not naturally generalize to unseen instances. The disadvantages of such kind of task generalization are computational inefficiency, especially for large graphs. A brief review of existing methods divided into matrix factorization-based methods, random walk based algorithms, and deep learning architectures for graphs is presented based on [10–12].

#### 3.1 Matrix Factorization Based Methods

Embedding based on the matrix factorization approach uses the properties of the graph represented in a matrix form, e.g., node pairwise similarity, and is aimed at the decomposition of this matrix into the product of others to get node embedding. In most cases, the algorithm’s input has to be a graph from nonrelational data features in high-dimensional space. One can simply use a column vector or row vector of adjacency matrix as the vector representation of nodes, but the representation space will be  $N$ -dimensional, where  $N$  is the number of nodes in a graph. Therefore, the goal is to form and learn low-rank vector space for the initial matrix preserving the network’s properties. The most common matrix factorization-based methods for embeddings are *Singular Value Decomposition* [15, 16], *Non-negative matrix factorization* [17], *Locally Linear Embedding* [18]. *Laplacian Eigenmaps* approach tries to preserve local distances and learn manifold structure. The disadvantage of such methods is that it cannot be applied to large graphs, because they operate on

dense matrices. The generated network representations are obtained through factorizing the Laplacian matrix of the adjacency matrix, therefore it exploits only the first-order proximity and demonstrates the importance of second-order proximity, which help preserving network structure. **HOPE** algorithm preserves high-order proximity and enables to transform the original SVD problem to a generalized one, but it also requires the whole graph matrix [16]. **GraRep** is another factorization method taking into account local and global structural information [19]. It uses SVD and transformed the transition probability matrix of DeepWalk, concatenating the final representations. However, it is not scalable and also requires the whole graph matrix. **LINE** [20] preserves the first- and second-order proximities separately and trains the embeddings by negative sampling, concatenating the obtained representations, which is a suboptimal solution. Nevertheless, LINE adopts a shallow structure, for which it is difficult to capture the highly nonlinear graph structure in the network.

### 3.2 *Random Walk Based Methods*

The idea of random walk based methods is to define a similarity based on stochastically denoted higher order neighborhoods of nodes. For unsupervised feature learning tasks, we learn node embeddings such that nodes nearby are close to each other, preserving similarity from an initial graph in  $d$ -dimensional vector space. The advantages of random walk implementation are the efficiency and flexibility, because we take only a part of node pairs with probabilities of cooccurrences for training set and provide a stochastic definition of similarity [13]. Given any node  $u$ , we learn its feature representation  $\phi(u)$  predicted by closest nodes from  $N(u)$ —target node’s  $k$ -hop neighborhood. Hence, the goal is to find the embedding of  $\phi_u$  such that it predicts close nodes from neighborhood obtained via random walk simulation.

Among the most frequently used algorithms is **DeepWalk** [21], which holds an idea to use unbiased random walks of fixed-length starting from each node and creates a matrix of  $d$ -dimensional node embeddings using the SkipGram algorithm. SkipGram is applied to the set of random walks maximizing the probability of the node’s neighborhood conditioned by the node’s embedding. In this way, nodes with similar neighborhoods (having large second-order proximity values) share similar embedding. For input it takes  $G(V, E)$ , window size  $w$ , embedding size  $d$ , the number of walk for each node  $\gamma$ , and walk length  $t$ . As a result, we obtain a matrix of vertex representations in  $R^{|V|*d}$ . Although it shows a good performance on different network datasets, it does not clarify the definition of the objective function for preserving graph structural information and is prone to keep only the second-order node proximity. Recently, there were suggested modifications of DeepWalk, e.g., *Max-Margin DeepWalk* [22]—a semi-supervised model that jointly optimizes the max-margin classifier and the social representation learning, also holding discriminative characteristics. Another well-known approach is **Node2Vec** introduced by Grover and Leskovec [23]. The key point was to use flexible and biased random walks, searching for a trade-off between exploration of global and local network

properties. Thus, the authors suggested to define two parameters:  $p$ —return back to the previous node,  $q$ —moving outwards or inwards, the strategies of DFS and BFS. For the input of learning features algorithm, we initialize a graph  $G = (V, E, W)$ ,  $d$  dimensions, walk length  $l$ ,  $r$  walks per node, context size  $k$ , and probabilities of return ( $p$ ), and in-out ( $q$ ). As DeepWalk, this model is scalable and local (does not require the entire graph), nevertheless, it is a hyperparameter-supervised approach, extending DeepWalk by introducing two parameters to control random walk sampling,  $p$  and  $q$ . Given  $p = 1, q = 1$ , we get back to DeepWalk. In addition, DeepWalk uses hierarchical softmax, while Node2Vec is based on negative sampling.

As an alternative approach to random walks, the diffusion simulations were suggested for graphs. Random walks tend to suffer from producing extra information by revisiting the same node several times, slowly spread across the network, and inefficient generation of proximity statistics [24]. Instead, we apply diffusion-like process to extract a subgraph of the node's neighbors—diffusion graph. Then, we find Euler tour to use it as sequence with more complete information on local  $k$ -hop neighborhood and all adjacencies in the graph in comparison to random walk methods. This approach was named **Diff2Vec** [24]. Fast Sequence Based Embedding is a further extension of Diff2Vec—it uses the sequences from Diff2Vec and uses it in the input of single-layer neural network with  $d$  neurons for learning  $d$ -dimensional embedding.

### 3.3 Graph Convolutional Networks and GAE

Convolutional networks and its modifications for the graphs have been widely adopted for learning graph embeddings. In general, the differences in approaches lie in the ways they formulate a similar to image convolution operation for working on graphs: to define the convolution in the spectral domain or treat it as neighborhood matching in the spatial set. Concerning the autoencoders—it is unsupervised models which are composed of two parts, i.e., the encoder and decoder. The autoencoders aim to minimize the loss function as a difference between input and output representation while intermediately reducing the data dimension. In terms of graph embeddings, adopting autoencoder models means their usage for proximity matrix factorization, for e.g., adjacency matrix factorization for the reconstruction process. Such an autoencoder will also make the nodes with similar neighborhood sets have similar embeddings. The following deep learning models are popular for graph embeddings learning: **EmbedNN** [25], **SDNE** [26].

## 4 Inductive Learning Embeddings on Dynamic Graphs

Most part of the existing methods requires the whole graph with all nodes for learning the embeddings, because otherwise they cannot generalize on to unseen instances. Inductive methods try to overcome this problem, which is especially relevant for large

networks evolving in time. The following techniques are mentioned and used in the scientific research papers: (1) application of static embedding algorithms to each snapshot of the dynamic graph and rotational alignment of the resulting embeddings across all time steps [14]; (2) explicit imposing of temporal regularizer in order to ensure temporal smoothness across embeddings from time snapshots; (3) information propagation [27]; (4) loss optimization, which encourages smooth changes between vertices with edges; (5) learning a mapping from node’s features by imposing a manifold regularizer obtained from the graph [28]. We are going to list the most effective methods for inductive learning embeddings problem: **Planetoid**, **DynGEM** [29], **Graph2Gauss**, **DepthLGP** [30].

**GraphSAGE** [14] is a method that generates node embeddings by sampling node neighborhood and aggregating attributes from its neighbors while providing an inductive framework supporting the node feature usage to efficiently generate graph embeddings for previously unseen data.

For each vertex  $v \in V$ , we aggregate the information from its neighbors  $u \in N(v)$  and itself:

$$h_v^k = \sigma([W_k * AGG(h_u^{k-1}, \forall v \in N(v)), B_k h_v^{k-1}]),$$

where  $AGG$  is a generalized differentiable aggregation function. The details of this algorithm will be regarded in the next sections.

## 5 Experimental Setup

In the experiments, we tested the performance of GraphSAGE model with different aggregation function against our model GSM in three tasks: (1) the classification of academic articles of HSE into 25 scientific research areas using Scopus citation database; (2) classification of research papers on Machine Learning topics into 7 subcategories (CORA dataset); (3) classification of posts placed on Reddit into 41 different communities. For all experiments, we make the predictions for unseen nodes that were not used during the model’s supervised training.

### 5.1 Baseline

We selected three models as baselines for the empirical results of inductive benchmark: GraphSAGE with mean, pooling, and LSTM aggregation functions. In GraphSAGE-modified model, we used attention-based aggregation function, added normalization of the two last layers, dropout for the network’s regularization and Adam optimization method instead of stochastic gradient descent (SGD). The cross-entropy was used as the loss function for supervised training. The authors of Graph-

SAGE used rectified linear units (ReLU) as the non-linearity function,  $K = 2$  and neighborhood's samples with sizes 25 and 10. In our experiments, the best results were obtained for LeakyReLU non-linearity function, implemented in PyTorch python module,  $K = 3$  and neighborhood's samples with sizes 15, 10, and 5. As shown in the table of results, the increase of the depth by one unit did not significantly affect the results for CORA and Reddit datasets, but it worked for HSEcite data. The use of one more “aggregator-encoder” bunch was helpful for the case, when the feature matrix is too sparse and the number of features is significantly more than objects. For providing a fair comparison of results, all models shares had the same loss function, the way to sample the neighborhood, and the number of minibatch iterations. The final set of hyperparameters' values was formed on early stage through validation tests on the subsets of CORA, HSEcite, and Reddit data that were discarded from the further analysis.

## 5.2 Datasets

The experiments were conducted on three evolving graphs, which represent citation and social networks.

**HSEcite.** We collected our own dataset from the Scopus citation database, gathering all papers, where the author's affiliation organization was National Research University Higher School of Economics for the time period 2011–2017. The dataset consists of 6279 unique articles and 21601 edges, which indicate that one paper cited another. It forms undirected and unweighted graph, containing 27836 features for each paper. First, we formed the dictionary of all keywords mentioned in the articles, and then transformed it to a binarized vector for each paper, where zero value means the absence of keyword in the article. In total, there are 25 scientific fields, which were coded from categorical to numerical values and are used as labels in this dataset. We want to predict paper subject categories in a multilabel classification task.

**Cora.** The dataset represents a citation network of Machine Learning papers. It includes the labels of seven categories: Genetic Algorithms, Theory, Case Based, Neural Networks, Probabilistic Methods, Reinforcement Learning, and Rule Learning. In total, the dataset is the corpus of 2,708 labeled papers and 5,429 directed links, where each article cited or was cited by at least one other article. Besides seven classes, it contains 1,433 features—the number of words used in paper abstracts and stored in a dictionary. Each paper is described by a 0/1-valued word vector. The first file contains the paper's id, word attributes, and class label, while the second one: unique id of cited paper and id of citing paper.

**Reddit.** The multilabel dataset contains 232 965 nodes, which represent users posts in online forum Reddit, and 5 376 619 edges. The labels are the communities, so-called “subreddits”, to which the post belong to. This social network can be constructed as a post-to-post graph, where the nodes are connected to each other if the same user left his comments on both posts. For building a graph, they were sampled 50 out of the largest communities. The first 20 days were used in training subset, 30%—



**Table 1** The datasets description

Dataset	Network	V	E	Features	Labels
HSEcite	Citation	6279	21601	27836	25
CORA	Citation	2708	5429	1433	7
Reddit	Social	232965	5376619	602	41

for validation, and the rest—for testing. The features are also transformed into word vectors containing the mean embedding of the post’s title, the mean embedding of all its comments, scores, and the number of comments of the post (Table 1).

### 5.3 Model Framework

In this subsection, we describe the generation of embeddings by the modified GraphSAGE algorithm (*Algorithm 1*). Assuming the model was trained and its parameters are fixed, we apply forward propagation algorithm. We suppose to learn the parameters of  $K$  aggregation functions ( $K = 3$  in our case, while in the architecture of GraphSAGE the two depth layers were used) and the weight matrices  $W^k$  for all  $k$  for further propagation between model’s layers. For each depth layer  $k \in K$ , the function incrementally aggregates information across neighboring nodes, which is transferred into encoder function. The input of the algorithm takes the whole graph  $G(V, E)$ , all features  $x_v, \forall v \in V$ . So,  $k$  defines the step from the outer loop,  $h^k$  stands for a representation of node at the current step. Each node  $v$  aggregates the neighbors and the representations of nodes  $h_u^{k-1}, \forall u \in N(v)$  into a vector  $h_{N(v)}^{k-1}$ . By introducing the subset  $V_0 \subset V$ , we imply that each vertex can be dropped with probability (e.g.,  $p = 0.2$ ). In fact, in the experiments we used the implementation of the dropout procedure in PyTorch and Tensorflow. The aggregation depends on the previous iteration  $k - 1$  and  $k = 0$  by the representations. After completing this step, the algorithm concatenates the aggregated neighbor’s vector  $h_{N(v)}^{k-1}$  with the last obtained representation of node  $h_v^{k-1}$ . This vector is transferred into a fully connected layer with the use of any nonlinear activation function (e.g., we used LeakyReLU). The output of this algorithm is denoted through  $z_v$ , the final representation.

The following aggregation functions were proposed and used by the authors of GraphSAGE algorithm for  $k$ th layer:

1. Averaging (the weighted average of the neighbors, a linear approximation of a localized spectral convolution):

$$AGG = \sum_{u \in N(v)} \frac{h_u^{k-1}}{|N(v)|}.$$

**Algorithm 1:** GSM embedding generation algorithm (forward propagation)

---

**Input:** Graph  $G(V, E)$ ; input features  $\{x_v, \forall v \in V\}$ ; depth  $K$ ; dropout probability  $p$ ; weight matrices  $W^k, \forall k \in \{1, \dots, K\}$ ; non-linearity  $\sigma$ , attention aggregation function  $AGG_k, \forall k \in \{1, \dots, K\}$ , neighborhood function  $N : v \rightarrow 2^V$

**Output:** Vector representations  $z_v$  for all  $v \in V$

```

1  $h_v^0 \leftarrow x_v, \forall v \in V$ ;
2 for  $k = 1 \dots K$  do
3     if  $k \neq K$  do
4          $\forall v \in V : P(v \in V_0) = p \rightarrow V_0 \subseteq V$  do
5              $V = V_0$ 
6         for  $v \in V_0$  do
7              $h_{N(v)}^0 \leftarrow AGG_k(\{h_u^{k-1}, \forall u \in N(v)\})$ ;
8              $h_v^k \leftarrow \sigma(W^k \cdot \text{CONCAT}(h_v^{k-1}, h_{N(v)}^k))$ ;
9         end
10     $h_v^k \leftarrow h_v^k / \|h_v^k\|_2, \forall v \in V$ 
11 end
12  $z_v \leftarrow h_v^k \forall v \in V$ 

```

---

2. Pooling (average or maximum value by element, there is no significant difference between mean- and max-pooling):

$$AGG = \gamma(Qh_u^{k-1}, \forall u \in N(v)).$$

3. LSTM (which is not symmetric, but holds larger extensive power):

$$AGG = LSTM([h_u^{k-1}, \forall u \in \pi(N(v))]).$$

We suggest to use attention-based aggregation function instead. Among the benefits of attention is that it gives the opportunity to work with inputs of variable size, focusing on the most important parts of the input. For computing a new representation of a single sequence, it is referred to self-attention [31]. The main idea is to compute the representations of each node following the self-attention and neighbors attending strategy. Firstly, this operation is efficient, because it can be parallelized across pairs of neighbors. Secondly, it is not limited by the node's degree, as far as we can adjust the arbitrary weights of neighbors. Thirdly, it is applicable to inductive learning tasks.

In the experiments, the attention component is a single (feedforward) neural network layer with weight vector  $\mathbf{a} \in \mathbb{R}^{2F}$  and  $\sigma$  nonlinear activation function (e.g., LeakyReLU). The coefficients can be denoted as

$$\alpha_{i,j} = \frac{\exp(\sigma(\mathbf{a}^T \cdot [W\mathbf{h}_i || W\mathbf{h}_j]))}{\sum_{k \in N_i} \exp(\sigma(\mathbf{a}^T \cdot [W\mathbf{h}_i || W\mathbf{h}_k]))},$$

where  $|\cdot|$  is the concatenation.

The first step is a linear transformation applied to every node, which is parametrized by a weight matrix  $W \in R^{F' \cdot F}$ , where  $F$  and  $F'$  are the numbers of features for each node in initial (input) and new(output) feature sets. Next, we compute self-attention on the nodes by calculating attention coefficients, that indicate the so-called importance of node  $j - s$  features to another node  $i$ :  $e_{ij} = a(W\mathbf{h}_i, W\mathbf{h}_j)$ , where  $a : R^F \cdot R^{F'} \rightarrow R$ .

## 5.4 Evaluation Metrics

To test the performance of different embedding methods on multilabel classification task, we report Micro- $F_1$  and Macro- $F_1$  scores. Micro- $F_1$  calculate metrics globally by counting the total true positives, false positives, and false negatives, while Macro- $F_1$  score calculate metrics for each label, and find their unweighted mean. 80/20% train/test split was used. XGBoost classifier was used in the experiments.

## 5.5 Discussion

In fact, the GSM model with attention-based aggregator function performed the best scores F1-micro and F1-macro on all three datasets (CORA, HSEcite, and Reddit), showing better results in comparison to mean, pool, and LSTM aggregators. One can observe a greater boost in accuracy for HSEcite dataset, which shows that attention-based aggregator better captures the information from neighboring nodes for this type of graph structure. All models, GraphSAGE and GSM variants, worked fast on CORA dataset with  $K = 2$  and  $K = 3$  depths due to the sizes of the feature matrix and the dataset. The runtime for the rest datasets, especially with  $K = 3$  depth, was increased with a factor of 5–80x that depends on the neighborhood size sampling. In fact, the GSM model with attention-based aggregator function performed the best scores F1-micro and F1-macro (0.881–0.873 on CORA, 0.694–0.691 on HSEcite, 0.918–0.911 on Reddit), showing better results in comparison to mean, pool, and LSTM aggregators (Table 2).

**Table 2** The results of experiments for multilabel classification task

Algorithm	CORA		HSEcite		Reddit	
	F1-micro	F1-macro	F1-micro	F1-macro	F1-micro	F1-macro
GraphSAGE-mean	0.866	0.864	0.583	0.581	0.893	0.889
GraphSAGE-pool	0.871	0.869	0.587	0.579	0.911	0.904
GraphSAGE-LSTM	0.873	0.871	0.522	0.519	0.915	0.898
GSM-attention	0.881	0.873	0.694	0.691	0.918	0.911

## 6 Conclusion

In this research, we investigated the problem of learning graph embeddings with respect to its applications in dynamic networks and inductive formalization in order to generalize to unseen nodes. We presented an overview and the classification of the modern methods for graph embeddings and learning algorithms for both transductive and inductive approaches. Selecting GraphSAGE algorithm as one of the state-of-the-art approaches for inductive learning, we proposed an improved model based on GraphSAGE algorithm and set up the experiments on datasets (CORa, Reddit, HSEcite), including our own data. The results evaluated by F1-micro and -macro metrics show that our model outperforms simple GraphSAGE models with mean, LSTM, and pool aggregation functions, which were taken as baselines. The key advantages of GSM model: (i) GraphSAGE-based model with attention-based aggregation function has the better ability to generalize not only to unseen nodes, but also to unseen graphs; (ii) the results were significantly improved in comparison to GraphSAGE on the Reddit dataset, used in the original paper, CORa and on our own data HSEcite; (iii) we showed the significance of deliberate choice of aggregation function, optimization method, more normalization layers and regularization (e.g., dropout) for a certain type of neural network architecture.

**Acknowledgements** Sections 2–5 were prepared under the support by the Russian Science Foundation under grant 17-11-01294, performed at National Research University Higher School of Economics, Russia. Section 1 was prepared under support by RFBR grant 16-29-09583 “Development of methodology, methods and tools for identifying and countering the proliferation of malicious information campaigns in the Internet”.

## References

1. Mislove, A., Marcon, M., Gummadi, K.P., Druschel, P., Bhattacharjee, B.: Measurement and analysis of online social networks. In: Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, pp. 29–42. ACM (2007)
2. Lapsuev, R., Ananyeva, M., Meinster, D., Makarov, I., Karpov, I., Zhukov, L.: Information propagation strategies in online social networks. In: Large Scale Networks - Computational Aspects and Applications - Computational Aspects and Applications, pp. 1–8 (2018)
3. Khayrullin, R.M., Makarov, I., Zhukov, L.E.: Predicting psychology attributes of a social network user. In Proceedings of EEML Workshop. Ceur WP, pp. 1–10 (2017)
4. Kyriakopoulos, F., Thurner, S., Pühr, C., Schmitz, S.W.: Network and eigenvalue analysis of financial transaction networks. *Eur. Phys. J. B* **71**(4), 523 (2009)
5. Makarov, I., Gerasimova, O., Sulimov, P., Zhukov, L.E.: Recommending co-authorship via network embeddings and feature engineering: the case of national research university higher school of economics. In: Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, ser. JCDL '18, pp. 365–366. New York, NY, USA, ACM, (2018). <https://doi.org/10.1145/3197026.3203911>
6. Makarov, I., Bulanov, O., Gerasimova, O., Meshcheryakova, N., Karpov, I., Zhukov, L.E.: Scientific matchmaker: collaborator recommender system. In: van der Aalst, W.M., Ignatov, D.I., Khachay, M., Kuznetsov, S.O., Lempitsky, V., Lomazova, I.A., Loukachevitch, N., Napoli,

- A., Panchenko, A., Pardalos, P.M., Savchenko, A.V., Wasserman, S. (eds.) *Analysis of Images, Social Networks and Texts*, pp. 404–410. Springer International Publishing, Cham (2018)
7. Makarov, I., Bulanov, O., Zhukov, L.E.: Co-author recommender system. In: Kalyagin, V.A., Nikolaev, A.I., Pardalos, P.M., Prokopyev, O.A. (eds.) *Models, Algorithms, and Technologies for Network Analysis*, pp. 251–257. Springer International Publishing, Cham (2017)
  8. Kurmukov, A., Ananyeva, M., Dodonova, Y., Gutman, B., Faskowitz, J., Jahanshad, N., Thompson, P., Zhukov, L.: Classifying phenotypes based on the community structure of human brain networks. In: *Graphs in Biomedical Image Analysis, Computational Anatomy and Imaging Genetics*, pp. 3–11. Springer, Berlin (2017)
  9. Brohee, S., Van Helden, J.: Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinform.* **7**(1), 488 (2006)
  10. Goyal, P., Ferrara, E.: Graph embedding techniques, applications, and performance: a survey. *Knowl.-Based Syst.* **151**, 78–94 (2018)
  11. Cui, P., Wang, X., Pei, J., Zhu, W.: A survey on network embedding. [arXiv:1711.08752](https://arxiv.org/abs/1711.08752) (2017)
  12. Cai, H., Zheng, V.W., Chang, K.: A comprehensive survey of graph embedding: problems, techniques and applications. *IEEE Trans. Knowl. Data Eng.* (2018)
  13. Leskovec, J.: Deep learning for network biology. part 1: network propagation and node embeddings. *Tutorial* (2018)
  14. Hamilton, W., Ying, Z., Leskovec, J.: Inductive representation learning on large graphs. In: *Advances in Neural Information Processing Systems*, pp. 1024–1034 (2017)
  15. Mees, A., Rapp, P., Jennings, L.: Singular-value decomposition and embedding dimension. *Phys. Rev. A* **36**(1), 340 (1987)
  16. Ou, M., Cui, P., Pei, J., Zhang, Z., Zhu, W.: Asymmetric transitivity preserving graph embedding. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1105–1114. ACM (2016)
  17. Wang, X., Cui, P., Wang, J., Pei, J., Zhu, W., Yang, S.: Community preserving network embedding. In: *AAAI*, pp. 203–209 (2017)
  18. Wang, J.: Locally linear embedding. In: *Geometric Structure of High-Dimensional Data and Dimensionality Reduction*, pp. 203–220. Springer, Berlin (2012)
  19. Cao, S., Lu, W., Xu, Q.: Grarep: learning graph representations with global structural information. In: *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, pp. 891–900. ACM (2015)
  20. Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., Mei, Q.: Line: large-scale information network embedding. In: *Proceedings of the 24th International Conference on World Wide Web*, pp. 1067–1077. International World Wide Web Conferences Steering Committee (2015)
  21. Perozzi, B., Al-Rfou, R., Skiena, S.: Deepwalk: online learning of social representations. In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 701–710. ACM (2014)
  22. Tu, C., Zhang, W., Liu, Z., Sun, M., et al.: Max-margin deepwalk: discriminative learning of network representation. In: *IJCAI*, pp. 3889–3895 (2016)
  23. Grover, A., Leskovec, J.: Node2vec: scalable feature learning for networks. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 855–864. ACM (2016)
  24. Rozemberczki, B., Sarkar, R.: Fast sequence-based embedding with diffusion graphs. In: *International Workshop on Complex Networks*, pp. 99–107. Springer, Berlin (2018)
  25. Weston, J., Ratle, F., Mobahi, H., Collobert, R.: Deep learning via semi-supervised embedding. In: *Neural Networks: Tricks of the Trade*, pp. 639–655 (2012)
  26. Wang, D., Cui, P., Zhu, W.: Structural deep network embedding. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1225–1234. ACM (2016)
  27. Zhu, X., Ghahramani, Z., Lafferty, J.D.: Semi-supervised learning using gaussian fields and harmonic functions. In: *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pp. 912–919 (2003)

28. Belkin, M., Niyogi, P., Sindhvani, V.: Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.* **7**, 2399–2434 (2006)
29. Ma, J., Cui, P., Zhu, W.: Depthlgp: learning embeddings of out-of-sample nodes in dynamic networks. In: *AAAI* (2018)
30. Bojchevski, A., Günnemann, S.: Deep gaussian embedding of attributed graphs: unsupervised inductive learning via ranking. [arXiv:1707.03815](https://arxiv.org/abs/1707.03815) (2017)
31. Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks, **1**(2). [arXiv:1710.10903](https://arxiv.org/abs/1710.10903) (2017)

# Collaborator Recommender System



Anna Averchenkova, Alina Akhmetzyanova, Konstantin Sudarikov,  
Stanislav Petrov, Ilya Makarov, Mikhail Pendiukhov and Leonid E. Zhukov

**Abstract** Nowadays, a lot of scientists' works aim to improve the quality of people's life but it could be quite complicated without building a successful collaboration. Productive partnerships can increase research efficiency in many cases and make a huge impact on society. However, today there is no clear way to find such collaborators. In this paper, we propose a recommender system for the scientists from the Higher School of Economics university to help them find co-authors for their prospective studies.

**Keywords** Recommender systems · Collaboration network · Link prediction

---

A. Averchenkova · A. Akhmetzyanova · K. Sudarikov · S. Petrov · I. Makarov (✉) · L. E. Zhukov  
National Research University Higher School of Economics, Moscow, Russian Federation  
e-mail: [iamakarov@hse.ru](mailto:iamakarov@hse.ru)

A. Averchenkova  
e-mail: [aaverchenkova@yandex.ru](mailto:aaverchenkova@yandex.ru)

A. Akhmetzyanova  
e-mail: [lady\\_ahmetzyanova@mail.ru](mailto:lady_ahmetzyanova@mail.ru)

K. Sudarikov  
e-mail: [kon.sudarikov@yandex.ru](mailto:kon.sudarikov@yandex.ru)

S. Petrov  
e-mail: [stasdp@mail.ru](mailto:stasdp@mail.ru)

L. E. Zhukov  
e-mail: [dx@apsolutions.ru](mailto:dx@apsolutions.ru)

I. Makarov  
Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia

M. Pendiukhov  
Analytical Software Solutions LLC, Moscow, Russian Federation  
e-mail: [lzhukov@hse.ru](mailto:lzhukov@hse.ru)

© Springer Nature Switzerland AG 2020  
I. Bychkov et al. (eds.), *Network Algorithms, Data Mining, and Applications*, Springer Proceedings in Mathematics & Statistics 315,  
[https://doi.org/10.1007/978-3-030-37157-9\\_7](https://doi.org/10.1007/978-3-030-37157-9_7)

## 1 Introduction and Related Work

Collaboration networks could be found in many fields of social science. Particularly, scientific publications networks play an important role these days, when the annual number of scientific articles is rapidly growing. The most challenging problem related to this kind of network is the search of the most appropriate collaboration teams of authors for the particular person. Main methods used for making such recommendations are based either on link analysis or content analysis. The purpose of this work is to propose a new approach to the problem by combining some of the existing methods with the novel techniques.

One of the directions of the collaboration recommender systems studies is focused on finding the most important nodes in the collaboration network. From different points of view, the concept of the most important vertex in the network differs. The significant problem in that field is finding the list of the most appropriate collaborators for new authors, who have a lack of co-authorship information and few publications.

For instance, CollabSeer [2] is one such systems that recommends potential authors to the scientists based on the co-author network structure and on the scientist's research interests. Another example of the considerable research related to this area is the publication of the Tang et al. [5]. They discovered the problem of inter-domain collaboration recommendation and proposed methods for recommending possible future co-authors by modeling cross-domain topics.

However, most studies consider co-author link prediction that is unsuitable for the isolated researchers because of the lack in co-authorship information. Moreover, they do not access the quality of collaborations. Another approach that overcomes this is proposed by Huynh et al. [3] The key idea of this study is adding additional information as new features in order to predict the best co-authorships. The researchers describe these possible features as the following metrics:

- the content similarity (using TF-IDF metric);
- organization similarity (computing the strength of the relationship between organizations represented by researchers);
- the researcher importance rating (using citation network);
- the researcher activity (given some time period).

They also access collaboration quality given the assumption that the more publication this collaboration can generate in the future, the better is this collaboration. The proposed methods were tested by the experiment on the data crawled from the Microsoft Academic Search using Support Vector Machines. By measuring the quality of the predicted collaborations, it was revealed that some of the proposed metrics increase the quality of recommended collaborators for the isolated researchers. Overall, the proposed algorithm allows finding the list of the best possible co-authors based on the given data and the metadata (as citation networks and time of the publication).

In [6], Zhang et al produce a dynamic collaboration recommendation method. The main idea is to take into account the experience that scholars gain over time.



Thus, for one source author, the model gives a ranking list of ten target authors in chronological order.

To choose the best candidate at each step, two factors are considered, namely the similarity and the status. The following variants of similarity calculation are present: the Jaccard similarity based on the authors' publication venues and keywords, and the Cosine similarity based on the titles and abstracts. For each scholar its popularity is calculated based on the sum of similarities with all the other authors. Then the collaborator selection is optimized by Gradient Value Iteration algorithm, which is built based on Value-Iteration algorithm. It provides the opportunity to consider the target author's status in the network. For the case when multiple authors pretend to collaborate with one target author, a competition function is proposed. It gives the result based on the social proximity between the competing authors.

The model was applied to the topic machine learning of the ACM dataset and compared with 5 baseline models. According to the results, the multiagent reinforcement learning method outperforms other methods in such evaluation metrics as MRR, P@3, P@5.

Today, in [4], the system for collaboration recommendation uses a variety of features to support decision-making. To generate a list of prospective co-authors, the system proceeds through four stages:

1. filtering data from ACM DL [1] offline dataset
2. extending the knowledge graph by adding information from ACM DL online dataset
3. supplementing the results
4. generate the relevant author list

In the first stage, the system gets a list of keywords from the user and fetches basic details from ACM offline data source considering relevance by time—they filter out all publications that are more than 3 years old. Also system considers user history data to enhance the accuracy of the search. On the next step, they extend knowledge graph by querying online ACM and getting publication and authors details such as:

- affiliation,
- average citations per article,
- citation count,
- publication count,
- number of downloads,
- average download per article.
- etc.

The third step is designed to extend author evaluation and to supplement the author information. Extending author evaluation is done by using Scopus API to access publications, citations, H-index, co-authors and area of interest. Then, based on previous data, the system builds co-author's graph and processes other information from Scopus. On the final step, it aggregates results from all steps and generates a list of relevant authors. The procedure of aggregating includes the following:

- computing recency;
- computing relevance by taking into account:
  - publication keywords,
  - keyword frequency,
  - order of the authors in authors list,
  - author similarity based on cosine distance between their publications.
- computing impact by taking into account:
  - citation count,
  - downloads,
  - the quality of collaboration colleagues.
- considering user expertise by utilizing TF/IDF method to discover the keywords in user publications
- computing quality by taking into account:
  - citation literature,
  - rank of journal.

## 2 Problem Statement

The scientists inside the publication network face a choice of future collaborations in order to do researches and to write articles together. Particularly, researchers from the Higher School of Economics university form a network inside which possible future collaborations could be found. Therefore, one of the major problems in this field is to find a collaborator for a chosen author from the publication network of the Higher School of Economics in 2018.

In this paper, we propose such a recommender system based only on the outside information of the network (as chosen author's works explain) or information combined with the inside structure of the network (neighbourhood of the authors within the network, etc.).

## 3 Data Description

The dataset provides information about **all published papers** that are affiliated with the Higher School of Economics university. By affiliation with the Higher School of Economics, we refer to all such publication where **at least one of the authors** is affiliated with the Higher School of Economics University (so, some of the researchers could be actually not from the university directly). Main information about all such papers was collected from the Scopus website (<https://www.scopus.com/home>) and is represented in Fig. 1. The data totals 9476 articles and contains such fields as

Авторы	Название	Год	Название источника
Shevgunov, T., Efimov, E.	Artificial neural networks implementing maxii	2019	Advances in Intelligent Systems and Computing
Budnitskii, O.	A Harvard project in reverse: Materials of the c	2018	Kritika
Demishev, S.V., Gilmanov, M.I., Samarin, A.N., S	Magnetic resonance probing of ground state ir	2018	Scientific Reports
Castro Santa, J., Exadaktylos, F., Soto-Faraco, S.	Beliefs about others' intentions determine wh	2018	Scientific Reports
Kaznadzey, A., Shelyakin, P., Belousova, E., Eren	The genes of the sulphoquinovose catabolism	2018	Scientific Reports
Tereshina, I.S., Kostyuchenko, N.V., Tereshina-C	The Mn<math>^{12}</math>-type phases for magnets w	2018	Scientific Reports
Nesterov, Y., Shikhman, V.	Dual subgradient method with averaging for o	2018	European Journal of Operational Research
Savchenko, A.V., Belova, N.S.	Unconstrained face identification using maxin	2018	Expert Systems with Applications
Proskuryakova, L.	Updating energy security and environmental p	2018	Journal of Environmental Management
Yakushkina, T., Saakian, D.B.	New versions of evolutionary models with letl	2018	Physica A: Statistical Mechanics and its Applications
Bespalov, P.A., Mizonova, V.G., Savina, O.N.	Reflection from and transmission through the	2018	Journal of Atmospheric and Solar-Terrestrial Physics

Fig. 1 Head of the dataset table

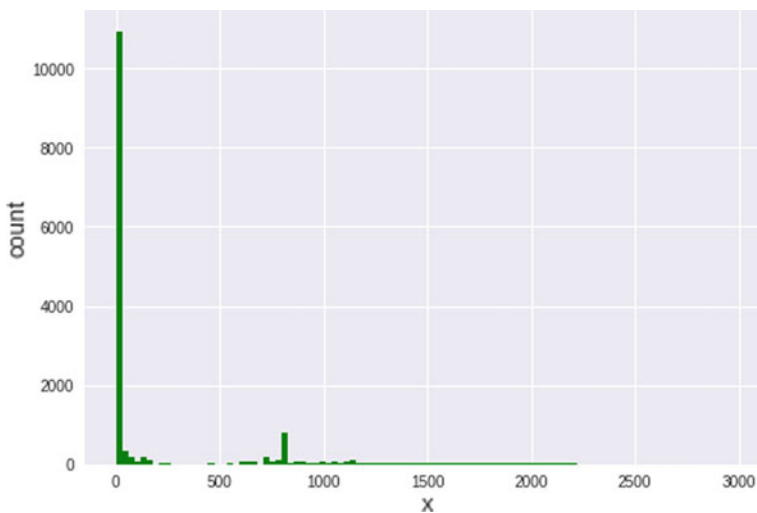
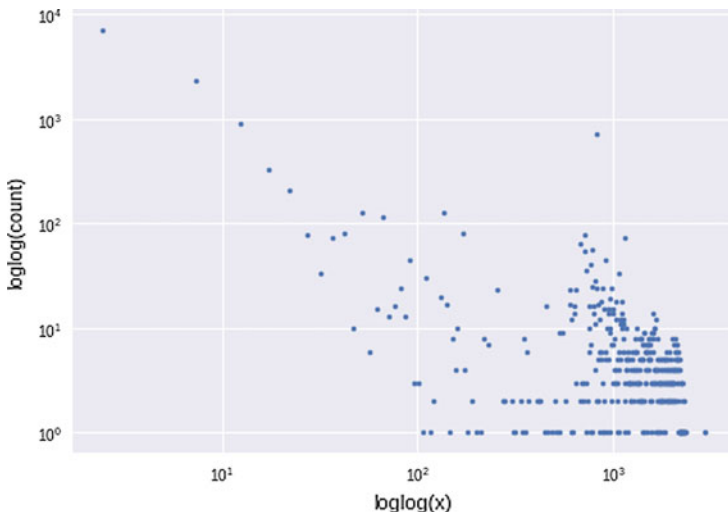


Fig. 2 Degree Distribution

- list of authors,
- title,
- year of publication,
- name of source (journal) where it was published,
- DOI,
- citations,
- type of document(e.g. Conference paper, article, review),
- etc.

From that data, a graph was built by taking into account co-authors relation. In other words, it was built in such a way that the set of vertices  $V$  of this graph is the set of all authors from the dataset and the set of edges  $E$  is all the pairs of authors that had at least one common publication.

The obtained graph has degree distribution shown in Fig. 2 and in log–log scale in Fig. 3. It is easily seen that Power Law holds for the current network.



**Fig. 3** Degree Distribution in log–log scale

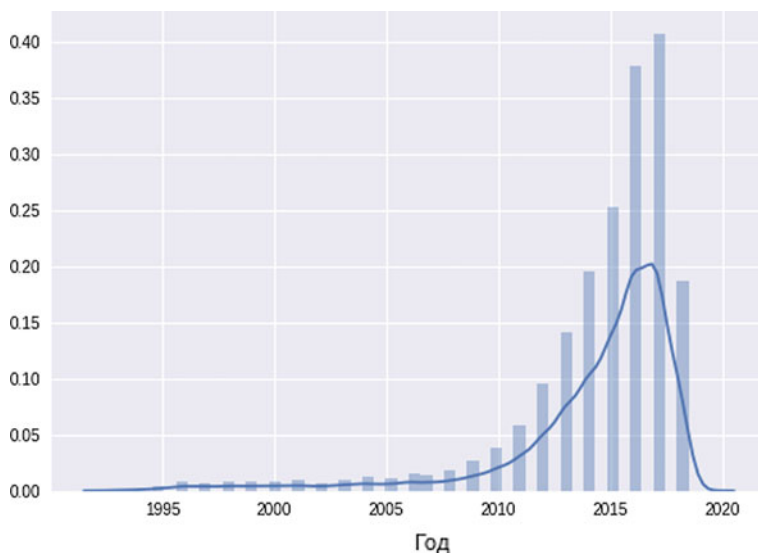
Another data insight can be obtained by plotting the distribution of papers by years in Fig. 4. It is obvious that the number of papers grows every year. However, it can be seen that the last two columns on this plot are smaller than the previous ones. This happens because the dataset contains only the articles until June 2018 (inclusive—i.e. the freshest articles) and does not contain any of the articles published after this month except 2 preprints from the year 2019.

In addition, we provide the basic statistics of the data:

- The total number of papers in the dataset = 9476;
- The total number of authors (vertices in the graph)  $|V| = 14794$ ;
- The total number of edges in the obtained graph  $|E| = 1669338$ .

And lastly for understanding which journals are the most popular in our dataset, the top 10 journals by papers published is given below:

#	Name of journal	Count of papers
1	Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)	232
2	CEUR Workshop Proceedings	193
3	Mathematical Notes	178
4	Voprosy Obrazovaniya	169
5	Russian Education and Society	104
6	Doklady Mathematics	96
7	Sotsiologicheskie Issledovaniya	88
8	Journal of Physics: Conference Series	86
9	Automation and Remote Control	82
10	Springer Proceedings in Mathematics and Statistics	82



**Fig. 4** Degree Distribution of papers by year

Moreover, for accurate and precise results, we also need to utilize the information about the importance and citation count of the journals where the given author is published.

For this purpose, the data from the SJR (ScimaJoR) website\* (<https://www.scimagojr.com/journalrank.php>) was collected and integrated with the initial dataset in such a way that for every article the following fields were added:

- quartile\* of the article (see the next section for the precise definition);
- Hirsch index of the article (see the next section for the explanation);
- main topics of the journal (those that pretend for any quartile).

\*Here as quartile there was taken the best of all the quartiles given to the chosen journal (every journal gets a quartile for each of the domains separately, see Fig. 5).

\*\*The list of all 2017 SJR grades was grabbed from the site.

## 4 Model Description

In this section, we propose our models. Let us begin with the baseline model which does not consider the graph structure.

For it, we used 8 various characteristics of each author  $i$  as

- $pt_i$ —the total number of papers.

We assign as  $pt_i$  the number of all papers from the dataset where the given person

Title	SJR Best CH index	Categories
CA - A Cancer Journal for Clinicians	Q1	137 Hematology (Q1); Oncology (Q1)
Nature Reviews Genetics	Q1	307 Genetics (Q1); Genetics (clinical) (Q1); Molecular Biology (Q1)
MMWR. Recommendations and reports : Morbidity and mortality weekly reports	Q1	125 Epidemiology (Q1); Health Information Management (Q1); Health (social)
National vital statistics reports : from the CDC	Q1	85 Life-span and Life-course Studies (Q1)
Nature Reviews Molecular Cell Biology	Q1	372 Cell Biology (Q1); Molecular Biology (Q1)
Quarterly Journal of Economics	Q1	219 Economics and Econometrics (Q1)
Nature Reviews Immunology	Q1	332 Immunology (Q1); Medicine (miscellaneous) (Q1)
Nature Reviews Materials	Q1	33 Biomaterials (Q1); Electronic, Optical and Magnetic Materials (Q1); Energy
Cell	Q1	682 Biochemistry, Genetics and Molecular Biology (miscellaneous) (Q1)
Handbook of International Economics	Q1	9 Economics and Econometrics (Q1); Economics, Econometrics and Finance
Nature Reviews Cancer	Q1	373 Cancer Research (Q1); Oncology (Q1)
Chemical Reviews	Q1	581 Chemistry (miscellaneous) (Q1)

Fig. 5 Head of the journal information table

$i$  stays as one of the authors of the paper. This amount reflects the publication productivity of the person.

**Limitations:** an integer number  $\geq 1$ .

- $s_i$ —the total number of co-authors.

We denote by  $s_i$  the number of all researchers that have at least one paper written together with the person  $i$  with other persons as collaborators or within them. It is relevant to calculate  $s_i$  score since it is the communication score of the person  $i$ , i.e., a valuable metric in a scientific community

**Limitations:** an integer number  $\geq 0$ .

- $c_i$  — the total number of citations

The number of all citations of all papers where the given person  $i$  stays as the author. This value is widely used by many resources such as Google Scholar. It represents the impact of the person to the scientific society.

**Limitations:** an integer number  $\geq 0$ .

- $pf_i$ —the number of those papers where the given author stays as the first author.

The first place in the authors order of the article shows the importance of the person inside the investigation and writing of the article. Moreover, the first author is responsible for the quality of the paper most. Thus, the number of all such papers is the index of the author's responsibility in his (her) researches.

**Limitations:** an integer number  $\geq 0$ .

- $q_i$ —the average quartile of the journal.

The  $q_i$  score shows the average significance of the authors paper measured in quartiles of journals where they are published. There exist four categories of quartiles: **Q1**, **Q2**, **Q3** and **Q4**. **Q1** reflects the best category, while the **Q4** score means that the journal is in the less powerful class of all journals. The average of all categories reflects well the average quality of the  $i$ th author list of papers.

**Limitations:** a decimal number  $\geq 1$ . For every paper, a quartile is an integer number from 1 to 4. For all papers that do not appear in the 2017th SJR list, we assume that the quartile = 5. The explanation is that often the journals are not sufficiently good to be included in this list, so it is reasonable to give them a quartile score slightly higher than the largest possible.

- $h_i$ —the largest value of the Hirsch Index of all journals.

The Hirsch index of the journal is very similar to the Hirsch index for the particular

author. So, the Hirsch index  $k$  of the journal means that this journal has at least  $k$  papers with at least  $k$  citations each of them. Overall, by taking the largest value of the Hirsch Index across all papers of the author  $i$ , we obtain a score that shows the best journal where the given person  $i$  was published ever.

**Limitations:** an integer number  $\geq 0$  and  $\leq 1053$ . All papers that do not appear in SJR list are assumed to have 0-score since the journal that they have published in is not sufficiently good (even if it has citations are not enough influential).

- $y_i$ —the latest year when the given author published a paper.  
By  $y_i$ , we denote the last year of all those when the given person  $i$  has published. So, this score reflects well the activity of the person inside the network across the timeline.

**Limitations:** an integer number  $\geq 1993$  and  $\leq 2019$

- $th_i$ —the number of topics that the author ever wrote on.  
The  $th_i$  value represents the number of themes that were ever covered by the particular author  $i$ . The theme here is given from the list of the main topics of the journal (got from the SJR data). The number of all the topics of the author  $i$  reflects the variety of topics that the  $i$ th author is interested in. Thus, it shows the variety and the coverage of domains where this author is a specialist.

**Limitations:** an integer number  $\geq 0$ . It equals to 0 since no one paper of the author  $i$  was published in a journal that was mentioned in the SJR-2017 list.

All these variables are considered as the set of attributes to the  $i$ th vertex in the graph which means the  $i$ th publication. The distributions of these variables throw all the vertices and the comments are given in the Appendix to the article.

To find new collaborations, we introduce a metric that helps to compare the vectors of characteristics and find the closest ones not considering the graph structure. The first step is to scale characteristics somehow. The way we decided to do it is to perform MinMaxScaling: for each of our 8 characteristics  $v_i$  we first compute maximum  $max_j v_{i,j}$  and minimum  $min_j v_{i,j}$  values, where  $i$  is the number of characteristic and  $j$  is the identifier of object. Then for each  $v_{i,j}$  we calculate new feature value using the formula:

$$v_{i,j}^{new} = \frac{v_{i,j} - min_j v_{i,j}}{max_j v_{i,j} - min_j v_{i,j}}$$

After this transformation, all the values will be in the range from 0 to 1.

\*Note that this technique is different from the other widely used that assigns new weights according to the formula:

$$v_{i,j}^{new} = \frac{v_{i,j} - mean_j v_{i,j}}{\sqrt{var_j v_{i,j}}}$$

We do not use this technique, because we would like to use variables with a different kind of nature that would be normalized from 0 to 1.

## Model 1

For each scientist, we get the ordered list of the “closest” candidates to cooperate with. In order to do this, we compare the vectors of current scientist with all others’ vectors according to Euclidean (also known as L2) distance  $d$ :

$$d^1(i, j) = \sqrt{(v_{1,i} - v_{1,j})^2 + (v_{2,i} - v_{2,j})^2 + \dots + (v_{n,i} - v_{n,j})^2}$$

where  $n = 8$  and  $(v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8) = (pt_i, s_i, c_i, pf_i, q_i, h_i, y_i, th_i)$  and then choose top-5 of them.

**Modification:** Let us complicate the model. Now for each author, we will consider only those potential colleagues with which the number of common topics exceeds the given threshold. Indeed, let us add to a journal (vertex) attributes the union of topics for each journal, obtained from the SJR list (see the previous section). So for each vertex in the graph, we get the union of topics of all journals where this publication was published. In our system, the scientist can choose the number of such topics that must appear together in both lists of topic by this scientist and the possible collaborator. That modification allows searching this author for the collaboration from the close or the same field of study. On the other hand, if the author will set up the value of the threshold to 0, then the unmodified model will be used and the choice will be made of all the authors within the network.

Here and after, the modified versions of the models will be implemented.

## Model 2

Let us further improve the model by adding the second part to the metric. The additional part will include only the inner structure of the network. For this purpose, we need to consider the two following pairwise metrics of the authors  $i$  and  $j$ :

- $cp_{ij}$ —the number of common papers.  
The  $cp_{ij}$  value represents the number of papers where  $i$  and  $j$  are co-authors (not necessary the only ones). This metric reflects the first approximation network connection of two authors.

**Limitations:** an integer number  $\geq 0$ .

- $ca_{ij}$ —the number of common authors.  
The  $ca_{ij}$  value reflects the number of common neighbours of the vertices  $i$  and  $j$  in the publication network. In other words, it equals to the number of authors that have at least one publication with  $i$  and at least one with  $j$  (no necessary the same). In other words, the metric reflects the second approximation network connection of two authors.

**Limitations:** an integer number  $\geq 0$ .

More formally,

$$ca_{ij} = \text{common\_authors}(a_i, a_j) = |a : a \in \text{neighbours}(a_i), a \in \text{neighbours}(a_j)|$$

$$cp_{ij} = \text{common\_papers}(a_i, a_j) = |p : p \in \text{papers}(a_i), p \in \text{papers}(a_j)|$$



For example, if two authors write all their papers together, *common\_papers* equals 1. These features are normalized in the following way:

$$\overline{ca_{ij}} = \text{common\_authors}(a_i, a_j) = \frac{|a : a \in \text{neighbours}(a_i), a \in \text{neighbours}(a_j)|}{\max(|\text{neighbours}(a_i)|, |\text{neighbours}(a_j)|)}$$

$$\overline{cp_{ij}} = \text{common\_papers}(a_i, a_j) = \frac{|p : p \in \text{papers}(a_i), p \in \text{papers}(a_j)|}{\max(|\text{papers}(a_i)|, |\text{papers}(a_j)|)}$$

The idea of making normalization is to make the model smoother. Normalizing by the maximum of sets lengths was done in order to provide the normalization itself and to prevent deleting by zero.

Now, our metric will look the following way:

$$d^2(i, j) = \sqrt{(v_{1,i} - v_{1,j})^2 + (v_{2,i} - v_{2,j})^2 + \dots + (v_{n,i} - v_{n,j})^2} - \sqrt{(\overline{ca_{i,j}})^2 + (\overline{cp_{i,j}})^2}$$

where  $n = 8$  and  $(v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8) = (pt_i, s_i, c_i, pf_i, q_i, h_i, y_i, th_i)$ .

### Model 3

A rough modification of the previous model is to consider only the fact that authors have at least one common paper or/and common neighbour. This means that they have already collaborated with each other in case of common papers, and that they have worked with the same co-author. So the metric will slightly change:

$$d^3(i, j) = \sqrt{(v_{1,i} - v_{1,j})^2 + (v_{2,i} - v_{2,j})^2 + \dots + (v_{n,i} - v_{n,j})^2} - \sqrt{I[ca_{i,j} > 0] + I[cp_{i,j} > 0]}$$

where  $n = 8$  and  $(v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8) = (pt_i, s_i, c_i, pf_i, q_i, h_i, y_i, th_i)$   
where  $I[k > 0]$  is an indicator function which is determined as follows:

$$I[k > 0] = \begin{cases} 1, & \text{if } k > 0; \\ 0, & \text{otherwise.} \end{cases}$$

This model seems slightly less smooth than the previous one since we consider only the fact if there exists at least one common neighbour/paper and does not consider the true amount or a share of them.

## 5 Results

We have constructed a recommendation system based on one of the three models. Now we will demonstrate the strength of the performance of our system for a particular author and compare it.

For implementation, the Python 3.5 code was used with such libraries as pandas and network. Its working time is less than 1 min that opens wide perspectives in the future.

It should also be mentioned that the simple parsing was conducted for the names of all authors. However, in the initial data there were names of authors with patronymic in some places and in some places without them. There appeared some difficulties in matching those scientists, because some persons can have the same patronymic and thus they cannot be simply matched to the name without patronymic but with the same first and second name. However, all other possible difficulties were overcome.

Let us look at the top-ranking results of the algorithm for the author ‘Kuznetsov, S.O.’.

Model 1, 0-threshold	Model 2, 0-threshold
– Pardalos, P.M.	– Ignatov, D.I.
– Saenko, V.S.	– Napoli, A.
– Ignatov, D.I.	– Pardalos, P.M.
– Bondarenko, G.G.	– Saenko, V.S.
– Babkin, E. ,	– Bondarenko, G.G.
Model 1, 3-threshold	Model 3, 0-threshold
– Pardalos, P.M.	– Pardalos, P.M.
– Ignatov, D.I.	– Ignatov, D.I.
– Savchenko, A.V.	– Savchenko, A.V.
– Napoli, A.	– Napoli, A.
– Aleskerov, F.	– Aleskerov, F.

We can see from here that the Model 3 gives the result which coincides with Model 1 for 0-threshold. Furthermore, several persons are top-3 in all rankings (‘Pardalos, P.M.’, ‘Ignatov, D.I.’) that reflect the sustainability of the models between each other. Moreover, these persons could be especially recommended for the author ‘Kuznetsov S.O.’ because of this reason. We also can notice that both Pardalos and Ignatov researchers have similar to Kuznetsov publication activity and close to Kuznetsov domains of interests, so these particular results are confirmed.

Now it is time to look at the author ‘Makarov, I.’.

Model 1, 0-threshold	Model 2, 0-threshold	Model 3, 0-threshold
– Neznanov, A.A.	– Karpov, I.	– Karpov, I.
– Musabirov, I.	– Zhukov, L.E.	– Zhukov, L.E.
– Galitsky, B.	– Gerasimova, O.	– Gerasimova, O.
– Kozyrev, O.	– Polyakov, P.	– Polyakov, P.
– Bernstein, A. ,	– Zyuzin, P.	– Zyuzin, P.

Model 1, 3-threshold	Model 2, 3-threshold	Model 3, 3-threshold
- Kuznetsov, S.O.	- Gerasimova, O.	- Karpov, I.
- Karpov, I.	- Polyakov, P.	- Gerasimova, O.
- Nikolaev, D.	- Zyuzin, P.	- Polyakov, P.
- Napoli, A.	- Martynov, M.	- Zyuzin, P.
- Yakovlev, A.A.	- Tokmakov, M.	- Tokmakov, M.

First of all, we can see that Model 1 with the 0-threshold gives the results different from other models. It can be explained by the fact that here only the outside information of graph is used and no topic intersection is done. Thus given all other characteristics there are possible collaborators.

As for other models, Model 1 with 3-threshold gives the results that still do not consider the internal structure of the graph, but take into account only those authors that have at least 3 common topics with ‘Makarov, I.’. Furthermore, ‘Karpov, I.’ that appears here also appears in other models in the top which says that it is an extremely good choice of collaborator.

Models 2 and 3 both have persons ‘Gerasimova, O.’ as well as ‘Polyakov, P.’ and ‘Zyuzin, P.’. These results are sufficiently more reliable since they consider both internal structure of the graph and nodes attributes. So, these could also be highly recommended possible collaborators since they have a similar profile to profile of ‘Makarov, I.’.

In addition, if to consider only the 0-threshold (which means any count of common topics), Models 2 and 3 suggest ‘Zhukov, L.E.’ as a good collaborator, which makes sense.

Let us consider the results for the author ‘Ignatov, D.I.’. They are the following (M stands for the Model index, T is a Threshold value)

M	T	Top-1	Top-2	Top-3	Top-4	Top-5
1	0	Babkin, E.	Aleskerov, F.	Napoli, A.	Petrosyants, K.O.	Mirkin, B.
2	0	Kuznetsov, S.O.	Poelmans, J.	Babkin, E.	Napoli, A.	Savchenko, A.V.
3	0	Aleskerov, F.	Napoli, A.	Savchenko, A.V.	Lomazova, I.A.	Kalyagin, V.A.

Let us increase the threshold:

M	T	Top-1	Top-2	Top-3	Top-4	Top-5
1	7	Babkin, E.	Napoli, A.	Petrosyants, K.O.	Mirkin, B.	Buzmakov, A.
2	7	Kuznetsov, S.O.	Poelmans, J.	Babkin, E.	Napoli, A.	Savchenko, A.V.
3	7	Napoli, A.	Savchenko, A.V.	Lomazova, I.A.	Kalyagin, V.A.	Nikolenko, S.

It can be seen that with threshold 7, which means that the proposed co-authors must have at least 7 common topics with Ignatov D.I., Model 2 gives the same

result as for threshold 0. The explanation of such results is that Model 2 promotes the colleagues with maximum number of common authors and papers. Overlapping of mutual topics is a consequence of a significant number of common authors and papers.

Model 1 results contain Aleskerov, F. without threshold and does not contain this candidate with threshold 7. Indeed, we can see that Aleskerov, F. and Ignatov D.I. have only 6 common topics of their research, while with Buzmakov, A. they have 13 common topics. For the same reason Aleskerov, F. is absent in Model 3 results for threshold 7.

## 6 Conclusion

As a result of the work, the recommendation system for the user from the existing publication network was created. Three models of collaborative recommendations were constructed. Every model considers several metrics of the vertices, some of them are individual and other reflects the inside structure of the publication network.

Model 1 is based on the vector description closeness, which is measured by the standard Euclidean norm. Model 1 modification filters the results, so that the number of common topics of the proposed co-author should exceed the selected threshold.

Model 2 additionally takes into account such inside metrics of the graph as the number of authors which are common for two selected authors and the number of papers written collaboratively. The more co-authors or common papers the authors have, the better for them to cooperate next time.

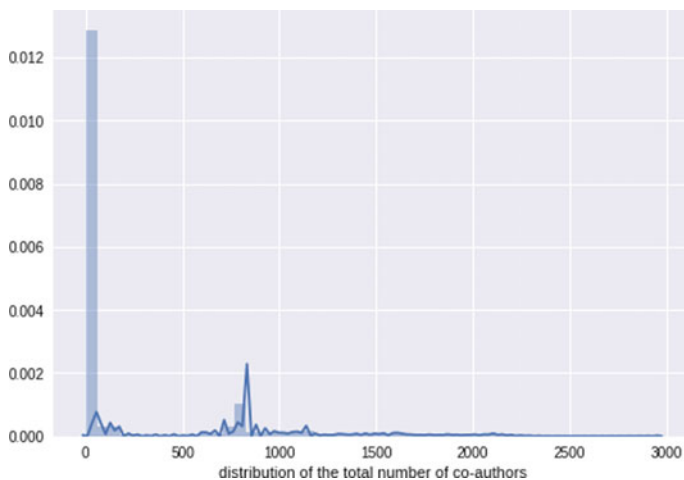
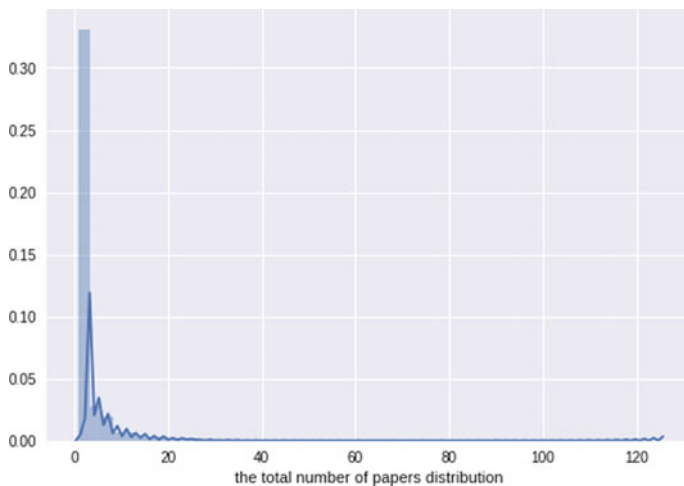
Model 3 treats mutual papers and ‘neighbours’ another way in comparison with Model 2. Only the fact that authors have worked with each other or have worked with at least one common person is considered. So, it equalizes the chances to work together for persons who have already done it once or much more times.

All proposed models were derived and tested for the Higher School of Economics publications network. This network was obtained by combining databases of Scopus and SJR. It consists of all authors affiliated with this university or having publication with such authors. Quite productive and well-behaved results were produced. Moreover, the program that analyzes the data works very fast, and therefore could be implemented for more complex analysis, social networks analysis or for a financial product in the future.

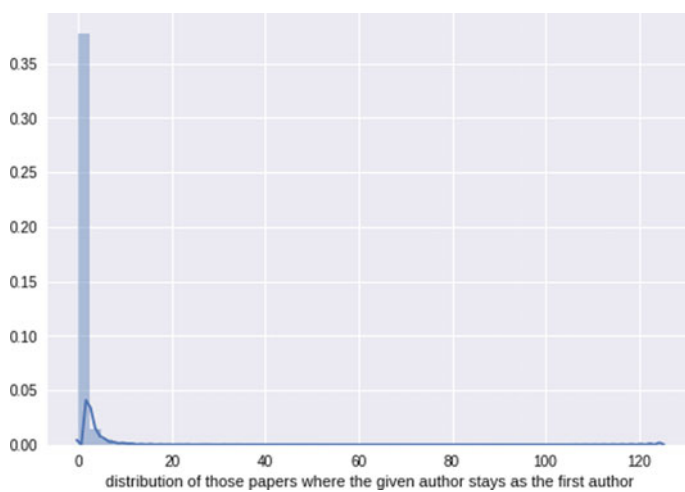
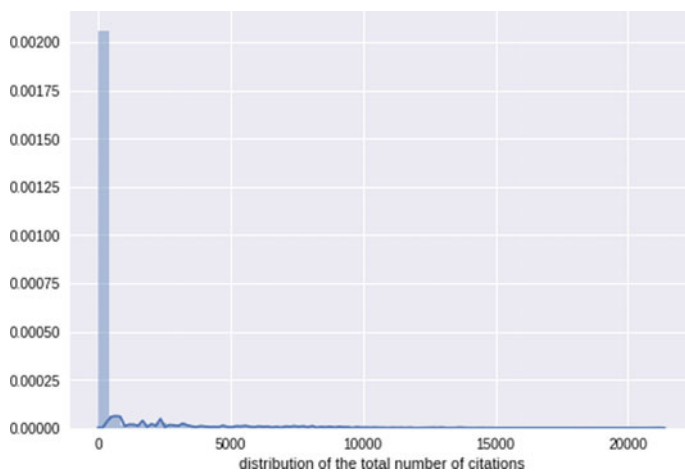
**Acknowledgements** Sections 2–5 were prepared under the support by the Russian Science Foundation under grant 17-11-01294, performed at National Research University Higher School of Economics, Russia. Section 1 was prepared under support by RFBR grant 16-29-09583 ‘Development of methodology, methods and tools for identifying and countering the proliferation of malicious information campaigns in the Internet’.

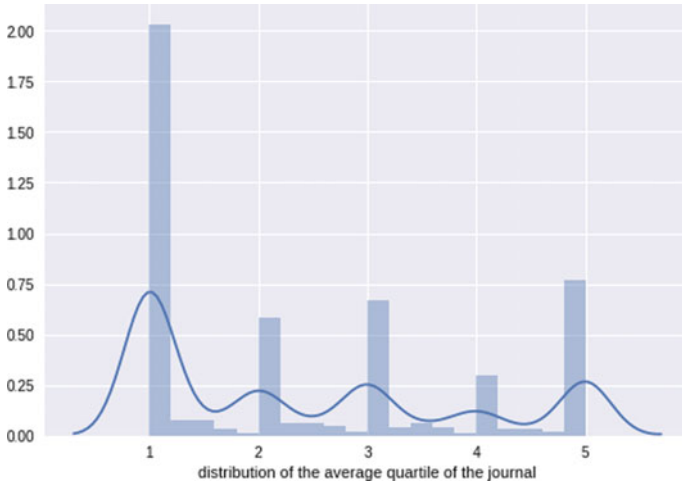
## 7 Appendix

In this section, the distribution of parameters from the models are given.

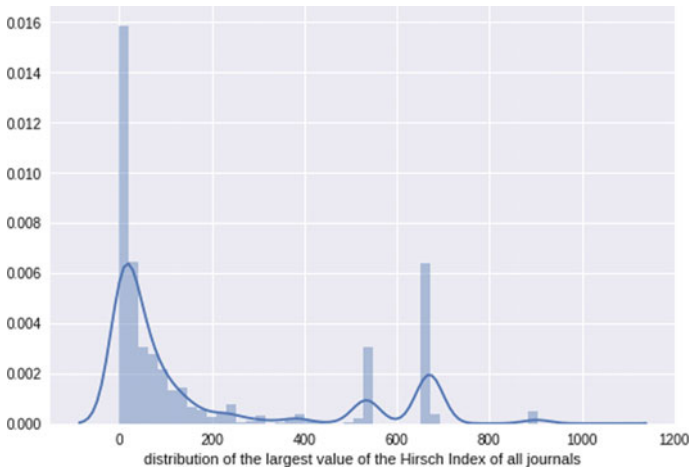


It is interesting that here exist a hump in the region of the number 800.

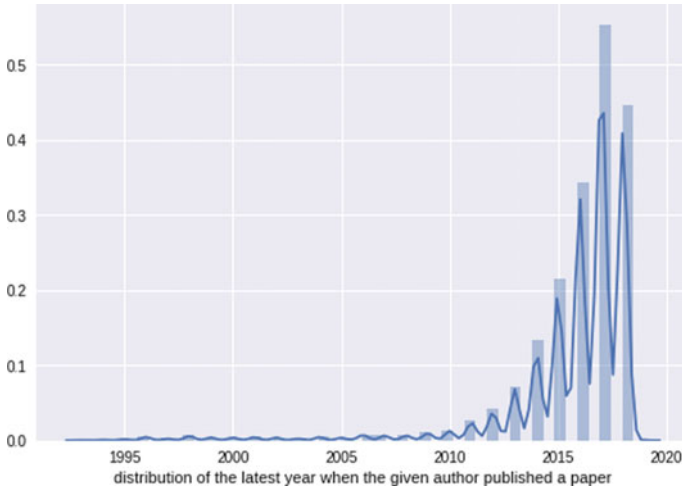




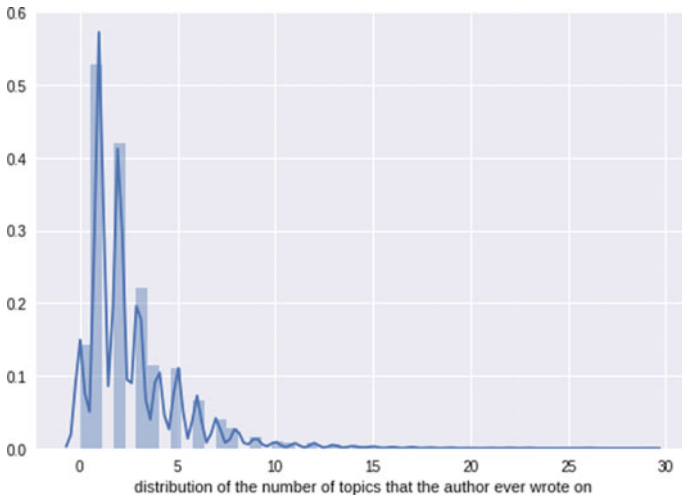
As it was expected, the highest hills are seen in the regions of the integer values of the variable. Moreover, there is a plenty of observations having exactly a score of 5 , which means that the journals where there were published are not good enough nowadays.



An interesting plot could be seen below. Of course, it could be expected that there is a large number of authors that have their journals only in low-citing journals. However, there are two bars not at the ends of the graph, which is quite interesting.



It can be seen from the graph that this distribution correlates with the year distribution shown in the section of the dataset description.



This graph shows that there is only a small amount of authors that have only papers with no topic association, which is very good. Moreover, as most of the previous graphs it looks as a power-law graph.



## References

1. ACM: ACM digital library <https://dl.acm.org/>
2. Chen, H.H., Gou, L., Zhang, X., Giles, C.L.: Collabseer: a search engine for collaboration discovery. In: Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries. pp. 231–240. ACM (2011)
3. Huynh, T., Takasu, A., Masada, T., Hoang, K.: Collaborator recommendation for isolated researchers. In: 2014 28th International Conference on Advanced Information Networking and Applications Workshops (WAINA), pp. 639–644. IEEE (2014)
4. Alinani, K., Wang, G., Alinani, A., Narejo, D.H.: Aggregating author pro-files from multiple publisher networks to build a list of potential collaborators. *IEEE Access* **6**, 20298–20308 (2018)
5. Lopes, G.R., Moro, M.M., Wives, L.K., De Oliveira, J.P.M.: Collaboration recommendation on academic social networks. In: International Conference on Conceptual Modeling. pp. 190–199. Springer (2010)
6. Zhang, Y., Zhang, C., Liu, X.: Dynamic scholarly collaborator recommendation via competitive multi-agent reinforcement learning. In: RecSys'17 Eleventh ACM Conference on Recommender Systems, pp. 331–335. ACM (2017)

# User Preference Prediction in a Set of Photos Based on Neural Aggregation Network



Kirill V. Demochkin and Andrey V. Savchenko

**Abstract** In this paper, we focus on the problem of user interests' classification in visual product recommender systems. We propose a two-stage procedure. At first, the image features are learned by fine-tuning the convolutional neural network, e.g., MobileNet. In the second stage, we use such learnable pooling techniques as neural aggregation network and context gating in order to compute a weighted average of image features. As a result, we can capture the relationships between the products' images purchased by the same user. We provide an experimental study with the Amazon product dataset. It was shown that our approach achieves an F1-measure of 0.90 for 15 recommendations, which is much higher when compared to 0.66 F1-measure classifications of traditional averaging of the feature vector.

**Keywords** Visual recommender system · Deep convolutional neural networks · Learnable pooling · Neural aggregation network · Context gating

## 1 Introduction

The relevance of user preference prediction task is explained by a recent surge in the amount of user's data being available to businesses based on his/her purchases of products. E-commerce shops would greatly benefit from a visual recommender system [1–4] that allows being able to quickly and reliably assess the potential categories of products relevant to the specific user based on the information gathered from, e.g., a smartphone app. Such a system will be able to make a prediction about the categories of interest for that user based on a gallery of photos on the mobile

---

The original version of this chapter was revised: The title of this Chapter has been renamed. The correction to this chapter is available at [https://doi.org/10.1007/978-3-030-37157-9\\_17](https://doi.org/10.1007/978-3-030-37157-9_17)

---

K. V. Demochkin (✉) · A. V. Savchenko  
Laboratory of Algorithms and Technologies for Network Analysis, National  
Research University Higher School of Economics, Nizhny Novgorod, Russian Federation

A. V. Savchenko  
e-mail: [avsavchenko@hse.ru](mailto:avsavchenko@hse.ru)

© Springer Nature Switzerland AG 2020, corrected publication 2020  
I. Bychkov et al. (eds.), *Network Algorithms, Data Mining,  
and Applications*, Springer Proceedings in Mathematics & Statistics 315,  
[https://doi.org/10.1007/978-3-030-37157-9\\_8](https://doi.org/10.1007/978-3-030-37157-9_8)

device. There is an additional challenge that the proposed system needs to be run on smartphones. This imposes constraints such as being able to fit the entire model in memory and not relying on computationally expensive algorithms that would be slow to provide inference on a user device.

The categories of products that a user is interested in are often correlated. Hence, in this paper, we propose to use modern learnable pooling techniques to capture the interdependencies between images of the same user. Among such techniques, the most successful are the neural aggregation model [5], which has been applied to the video-based face recognition, and the context gating [6], which won the prestigious YouTube 8M Large-Scale Video Understanding challenge.

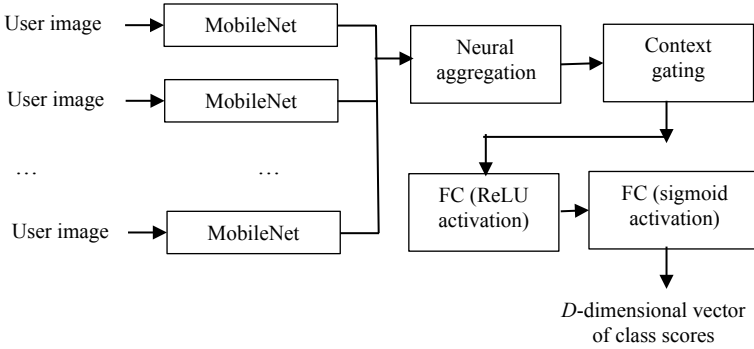
The rest of the paper is organized as follows: in Sect. 2 we formulate the proposed approach for the classification of user preferences based on the aggregated features extracted from images of products. In Sect. 3, we conduct an experimental study of the implemented methods and compare our results with the conventional averaging. In Sect. 4, we present the findings and give concluding comments.

## 2 Decision-Making Using Neural Aggregation of Visual Data

Let there be given  $N$  collections of images of products from  $D$  unique categories that are associated with  $N$  users. Each  $n$ th collection corresponding to the  $n$ th user contains  $M_n$  images of products  $\{X_n(m)\}$ ,  $m = 1, 2, \dots, M_n$ , that this user has purchased or interacted with. We assume that each image contains only a single product that can belong to one or more of  $D$  categories. Each image is labeled with a binary vector  $\mathbf{y}$  of length  $D$ , where the  $d$ th entry of  $\mathbf{y}$  is set to 1 if the item on the image belongs to the  $i$ th category and 0 otherwise. The task is to predict the relevant classes of products to a user based on a collection of images of products that the user has interacted with, i.e., generate a  $D$ -dimensional vector of scores (estimates of posterior probabilities) that the corresponding category is relevant to the user.

Nowadays, it is common practice to build an image classifier on top of a deep convolutional neural network based model trained on a large dataset such as ImageNet [7]. Therefore, we propose the following two-stage decision-making procedure. At first, we split the  $N$  collections of images into two disjoint sets with size  $N_1$  and  $N_2$ . The first set is used to learn the relevant features. In particular, we first apply transfer learning [8] to the preliminary trained MobileNet model [9] using only images from the first subset of  $N_1$  users. This fine-tuned model is used to extract the  $K$ -dimensional feature vector  $\mathbf{x}_n(m)$  for each image from the second subset.

At the second stage, we learn the aggregation (pooling) of these features in order to produce the final  $K$ -dimensional feature vector  $\mathbf{x}_n$ , which describes the  $n$ th user as a weighting sum:



**Fig. 1** Proposed neural network architecture

$$\mathbf{x}_n = \sum_{m=1}^{M_n} w(\mathbf{x}_n(m)) \mathbf{x}_n(m), \tag{1}$$

where the weights  $w$  depend on the features  $\mathbf{x}_n(m)$ . If all weights are equal, conventional averaging is implemented. However, in this paper, we decided to use the neural aggregation module based on the attention mechanism originally used in video-based face recognition [5]. Moreover, we additionally use the context gating [6], which applies a scaling mask to the resulting aggregated vector (1). The intuition behind context gating is that the products from certain categories are likely to appear together. Hence, the weights for images of those products should be scaled up if they are present in a single collection. The opposite is also true: for items that are not likely bought together, we would like their weights to be scaled down. Finally, the aggregated vectors (1) are fed into a fully connected layer with dropout. Since the target vectors are not one-hot encoded as a product can correspond to many categories, we used the sigmoid activation on the prediction layer. This layer finally produces  $D$  predicted scores or probabilities that the  $d$ th product category is relevant to the particular user. The overall neural architecture is shown in Fig. 1.

### 3 Experimental Results

In the experimental study, the “Home and Kitchen” 5-core subset of the Amazon Product Data dataset [10] was used (Fig. 2), meaning only the items that have at least 5 unique users interacted with and only users that have interacted with at least 5 unique items are kept. Such subset contains 547,700 entries of  $N = 66519$  unique users interacting with 28,237 unique items from  $D = 1000$  categories. The list of categories includes “Cookware”, “Storage and Organization”, “Coffee”, etc. For each user, there is data available on the items that the user has bought. The number of items per user  $M_n$  varies from 5 to 40; the average user has interacted with 8 unique



**Fig. 2** Examples of images from the “Home and Kitchen” subset

items. Each user was assigned a  $D$ -dimensional vector  $y$  where the  $d$ th element is 0 if the user has not bought any items from category  $d$ , and 1 otherwise if the user has bought at least a single product from category  $d$ . All experiments were conducted on a.

As part of the preprocessing stage, each image was resized to  $224 \times 224$  pixels and RGB values were normalized to the range  $[-1; 1]$  to conform to the input format required by the MobileNet v1 pretrained with ImageNet weights. At first, 70% of all images ( $N_1 = 0.7$ ,  $N = 46563$ ) selected by random split were used to fine-tune the MobileNet model. It should be noted that since each item belongs to only a few of the categories the resulting target vectors are sparse. To reduce the class imbalance, we implemented the weighted binary cross-entropy objective function. Several values were tested for the positive class weight  $\{10, 36, 72, 140\}$  and the best quality was achieved for positive class weight equal to 36. We also found that the hidden layer with 2048 neurons and 50% dropout probability worked best in our experiment. The model was first trained with 22 frozen deepest layers in batches of 64 samples using the ADAM optimizer with a learning rate of 0.001,  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$  for 10 epochs. Then it was trained with all layers unfrozen for 20 more epochs with only the learning rate changed to 0.0001.

To speed up the experiments, we pre-extracted the  $K$ -dimensional feature vectors for the remaining 30% of images, grouped them by the user and saved the resulting array. Next, we simply label each user with a  $D$ -dimensional vector where the  $d$ th component is equal to 0 if the user has not interacted with an item from the  $d$ th category and 1 if the user has interacted with at least one item of the  $d$ th class.

As for the aggregation step, we tested three approaches (Average pooling, Neural Aggregation, Neural Aggregation + Context Gating). We used 70% of our data to learn the pooling weights. The algorithms were tested on the other 30% of users. In the case of average pooling, we just take the mean of all feature vectors for a user. For the second approach, we have two stacked attention blocks as in the original paper [5]. For the final approach, we used two stacked attention blocks, followed by a context gating unit proposed [6], which takes in a vector (1) and dynamically reweights its entries.

After the features are pooled into a single vector, it is passed through a single dense hidden layer of size 2048 with a linear layer that predicts the final relevance of categories. We utilized the same custom weighted binary cross-entropy function with positive weight 36 as the objective function. The model was trained with the ADAM optimizer with learning rate = 0.001,  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ . The dependence of the F1-measure on the number of recommendations  $k$  is shown in Fig. 3. Details about precision @  $k$  and recall @  $k$  [11] are presented in Table 1. Here the best results are marked in bold.

Here one can notice that the highest F-measure is achieved by combined neural aggregation and context gating 12–35% higher when compared to simple averaging of visual features. The addition of Context Gating [6] to the neural aggregation network makes it possible to improve the decision-making quality in 5–14%.

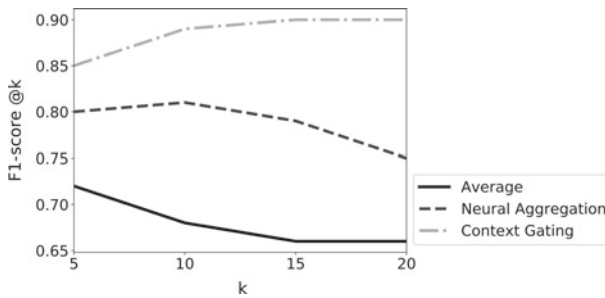


Fig. 3 Dependence of F1-measure on the number of recommendations  $k$

Table 1 Recall @  $k$  and Precision @  $k$  for different aggregation strategies

$k$	Aggregation	Recall @ $k$	Precision @ $k$
5	Average	0.704867	0.749925
	Neural aggregation	0.772574	0.839458
	Neural aggregation + context gating	<b>0.792203</b>	<b>0.922438</b>
10	Average	0.797340	0.595867
	Neural aggregation	0.901716	0.710123
	Neural aggregation + context gating	<b>0.91846</b>	<b>0.881151</b>
15	Average	0.815469	0.561431
	Neural aggregation	0.932418	0.710123
	Neural aggregation + context gating	<b>0.942565</b>	<b>0.868210</b>
20	Average	0.820141	0.553453
	Neural aggregation	0.943513	0.636783
	Neural aggregation + context gating	<b>0.947498</b>	<b>0.864384</b>

## 4 Conclusion and Future Work

In this paper, we discussed the possibility to use learnable pooling techniques to predict user preferences based on images of items from their previous interactions. These techniques were previously successfully implemented in various video recognition tasks. It was experimentally shown that the neural aggregation [5] with context gating outperforms [6] the naive averaging method by up to 34% (Fig. 3).

The main direction for further research is to expand the proposed approach into a complete mobile recommender system that would suggest the items from relevant categories to that user based only on the images on the user's device. Additionally, we are interested in comparing the performance of our approach with traditional recommender system, e.g., collaborative filtering or factorization machines [12]. Finally, it is important to work with other open datasets, e.g., the Behance dataset that features interactions of users with graphic designs created by other users.

**Acknowledgements** The article was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE University) in 2019 (Grant No. 19-04-004) and within the framework of the Russian Academic Excellence Project "5-100".

## References

1. Hidasi, B., Quadrana, M., Karatzoglou, A., Tikk, D.: Parallel recurrent neural network architectures for feature-rich session-based recommendations. In: Proceedings of the 10th ACM Conference on Recommender Systems, pp. 241–248 (2016)
2. Shankar, D., Narumanchi, S., Ananya, H.A., Kompalli, P., Chaudhury, K.: Deep learning based large scale visual recommendation and search for e-commerce (2017). arXiv preprint [arXiv:1703.02344](https://arxiv.org/abs/1703.02344)
3. Andreeva, E., Ignatov, D.I., Grachev, A., Savchenko, A.V.: Extraction of visual features for recommendation of products via deep learning. In: van der Aalst W. et al. (eds.) Analysis of Images, Social Networks and Texts. AIST 2018. Lecture Notes in Computer Science, vol. 11179, pp. 201–210. Springer, Cham (2018)
4. Zhai, A., Kislyuk, D., Jing, Y., Feng, M., Tzeng, E., Donahue, J., Du, Y.L., Darrell, T.: Visual discovery at pinterest. In: Proceedings of the 26th International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee, pp. 515–524 (2017)
5. Yang, J., Ren, P., Zhang, D., Chen, D., Wen, F., Li, H., Hua, G.: Neural aggregation network for video face recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR), pp. 4362–4371 (2017)
6. Miech, A., Laptev, I., Sivic, J.: Learnable pooling with Context Gating for video classification (2017). arXiv preprint [arXiv:1706.06905](https://arxiv.org/abs/1706.06905)
7. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 1097–1105 (2012)
8. Pan, S.J., Qiang, Y.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**(10), 1345–1359 (2010)

9. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Adam, H.: Mobilenets: efficient convolutional neural networks for mobile vision applications (2017). arXiv preprint [arXiv:1704.04861](https://arxiv.org/abs/1704.04861)
10. McAuley, J., Targett, C., Shi, Q., Van Den Hengel, A.: Image-based recommendations on styles and substitutes. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 43–52 (2015)
11. Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst. (TOIS)* **22**(1), 5–53 (2004)
12. Zhou, Y., Wilkinson, D., Schreiber, R., Pan, R.: Large-scale parallel collaborative filtering for the Netflix prize. In: International Conference on Algorithmic Applications in Management, pp. 337–348 (2008)



# Network Structure and Scheme Analysis of the Russian Language Segment of Wikipedia



Sergey Makrushin

**Abstract** Nowadays Wikipedia is much more than the most popular online encyclopedia: it is a unique source of semi-structured knowledge about the world. Data from Wikipedia is actively used in different approaches to create universal semantic networks. In the current research we created a network of the Russian language segment of Wikipedia and conducted an analysis of its structure. In our network Wikipedia articles and categories are regarded as nodes and references are regarded as links. The emphasis of the paper is on studying nodes degree distribution. In the work it is suggested to use the adaptive kernel density estimation method for clearing diffusion of tails of nodes degree distributions. Using this method, we found that nodes degree distribution for out-links of articles doesn't fit to a power law and has extensive artifacts: a large amount of outlier points which deviate very much from the trend. We found that these artifacts were induced by using navigation templates in Wikipedia articles. For eliminating these artifacts, we include a new type of nodes and links, representing Wikipedia templates and their references, into the Wikipedia network scheme. It enables separating article-to-article links, generated by navigation templates, from regular links and eliminating these artifacts.

## 1 Introduction

Wikipedia [1] represents a unique combination of well-structured knowledge and a wealth of information: at the moment all language segments of Wikipedia have more than 47 million of articles, and more than one million articles belong to 14 language segments.

Information in Wikipedia is organized as a body of hypertext documents, but as compared to World Wide Web as a whole, pages in Wikipedia are much more strictly organized. Layout of all Wikipedia pages is defined by Media Wiki wiki-engine [2].

---

S. Makrushin (✉)

Financial University under the Government of the Russian Federation,  
49 Leningradsky Prospekt, GSP-3, 125993 Moscow, Russian Federation  
e-mail: [SVMakrushin@fa.ru](mailto:SVMakrushin@fa.ru)

© Springer Nature Switzerland AG 2020

I. Bychkov et al. (eds.), *Network Algorithms, Data Mining, and Applications*, Springer Proceedings in Mathematics & Statistics 315,  
[https://doi.org/10.1007/978-3-030-37157-9\\_9](https://doi.org/10.1007/978-3-030-37157-9_9)

Every page in Media Wiki belongs to a particular type and there are strict rules (defined by Media Wiki and by Wikipedia community) of structuring and formatting for all of these page types. Rules are thoroughly developed for articles - it is the main Wikipedia page type. These rules make information in Wikipedia much more well-structured comparing to WWW which has pages with structure constrained only by HTML format.

The fact that Wikipedia is relatively well structured of Wikipedia helps to use it not only in regular way but also for automatic information retrieval. For example, data from Wikipedia is actively used in different approaches to create universal semantic networks—for instance, DBPedia [3]. Wikipedia itself can also be used as a weakly structured semantic network. From this point of view, the structure of its network is of great interest. In our research we created a network of the Russian language segment and made basic analysis of its network structure using methods of complex network theory.

## 2 Related Work

In the current research we created a network of the Russian language segment of Wikipedia and conducted an analysis of its structure. The emphasis of the paper is on studying nodes degree distribution. Analysis of the network of Wikipedia is quite popular in the literature. Particularly because Wikipedia is a huge semi-structured knowledge base, its network is widely used for knowledge and semantic extraction [4, 5].

Node degree distribution analysis is a basic method of network analysis and it is widely applied to the network of Wikipedia. For example, in [6] the node degree distribution is used for comparison of linkstructures in the network of Wikipedia and in the subgraph of Web graph. The authors found that Wikipedia link structure is similar to the Web. Main finding of the article is that both out-links and in-links are good indicators of relevance of a Wikipedia article, whereas on the Web the in-links are more important.

In [7] analysis of node degree distribution is used for identification of the statistical properties and the growth model of Wikipedia. The main feature of this article is the emphasis on the study of the Wikipedia network dynamics (growth) and its growth mechanism. In [7] as in [6] the authors found out that the Wikipedia network is scale free. The authors of [7] found a simple statistical model which could reproduce main features of the Wikipedia network by local rules such as the preferential attachment mechanism.

It is important to notice that research of growth mechanisms for real world networks (such as the Wikipedia network), which could lead to node degree distributions similar to scale free distribution, is quite popular in the complex networks' science. Since the well-known publication of Barabasi and Albert [8], there have been a lot of publications (see [9, 10]) about development of growth mechanisms for real world networks, which are more appropriate for networks similar to Web graphs.

A common logic of this publications is that a deeper understanding of specific logic of a network could lead to more precise models of the network growth. In the current paper we go more deeply in the specific logic of a certain network to clean and prepare data for using more common models, which could describe such networks properties as node degree distribution.

### 3 Wikipedia Network Creation

To define the scheme of our Wikipedia network, we need to identify essential Wikipedia structure which forms its semantic framework. As was already mentioned, all Wikipedia pages belong to various types defined by page namespaces. A Wikipedia namespace is a named set of Wikipedia pages. Each Wikipedia page belongs to one namespace. Pages from each namespace are used for special goals and have strict rules for their layout and content formation. In fact, a namespace defines the type of a page. A full name of each page (except pages from the main namespace) begins with a namespace followed by a colon, e.g. the name of Wikipedia page which describes the Wikipedia namespace concept is “Wikipedia:Namespace”.

Wikipedia main namespace is used for Wikipedia articles—encyclopedia articles, lists, disambiguation pages and encyclopedia redirects. There is one exception from the rule of full Wikipedia page name creation for the pages from the main namespace: the prefix for them is omitted. In the Table 1 you can see the definition

**Table 1** Basic Wikipedia namespaces

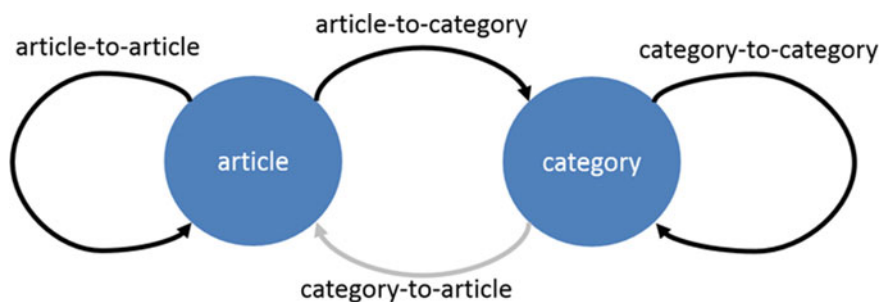
Namespace id / Talk namespace page id	Namespace name (Russian Wikipedia Namespace name)	Namespace description (namespace classification)
0 / 1	Main/Article [no prefix in the page name]	encyclopedia articles, lists, disambiguation pages and encyclopedia redirects
2 / 3	User (Участник)	user pages and pages created by users for personal use (support)
4 / 5	Wikipedia (Википедия)	pages connected with the Wikipedia project itself: information, policy, discussion, etc. (support)
6 / 7	File (Файл)	file description pages for image, videos or audio files (structure)
8 / 9	MediaWiki (MediaWiki)	interface texts, e.g. messages for automatically generated pages (support)
10 / 11	Template (Шаблон)	pages that are intended to be included into other pages in order to insert standard text or boxes such as info boxes
12 / 13	Help (Справка)	pages which provide help in using Wikipedia and its software (support)
14 / 15	Category (Категория)	category pages, which display a list of pages and subcategories that have been added to a particular category (structure)
100 / 101	Portal (Портал)	portals that help readers to find articles related to a specific topic (structure)

of major Wikipedia namespaces. For each namespace there is a supporting “talk” namespace which is used for creation of discussion pages for pages from this namespace. Namespaces can be classified into two main classes: namespaces responsible for Wikipedia infrastructure and community support (marked with “support” class name in the table) and namespaces responsible for Wikipedia information structure and meta-information (marked with “structure” class name in the table). Since our task is the definition of Wikipedia network scheme, the point of our interest are the namespaces from the last class.

Pages from the Category namespace are crucial for the formation of semantic relations between Wikipedia articles. Categories are used in Wikipedia to link articles under a common topic and allow readers to navigate through Wikipedia and find related articles. A page in the main namespace must have a link to at least one category (these categories can be found at the bottom of the article page). Affiliation of terms, which are described in articles, to the same topic is very important semantic information and must be presented in the created Wikipedia network.

As well as an article, a category in Wikipedia must have at least one link to a parent category. There is one exception: the root category “Category:Wikipedia\_categories” (in the Russian language segment it is called “Kategoriya:Vse”) does not have a parent category. As in the case of a usual reference from article to article, a reference to a parent category can be represented as a directed link from a child article or category to a parent category. Categories in Wikipedia are hierarchical, which means that the links between categories can’t form a directed cycle. As a result of this rule, categories in Wikipedia form a directed acyclic graph.

Based on the analysis given above, the following scheme of Wikipedia network is proposed (see Fig. 1). Essential semantic information from Wikipedia could be represented as a network of Wikipedia articles and categories. Articles and categories are considered as network nodes, which have the following attributes: name, type (article or category) and some additional attributes (e.g. length of page text). Full Wikipedia page text is not stored in nodes. References between Wikipedia pages are represented in the network as directed links between nodes. A link between two nodes in the Wikipedia network could be interpreted as the existence of semantic relation between terms. There are three main types of links: usual links from article



**Fig. 1** Wikipedia network scheme with two types of nodes

to article; links from article to category (they mean that an article belongs to some category), links from category to category (they form hierarchy of categories). Links from categories to articles are very rare and could be omitted. By analogy with the webgraph—a directed graph, whose vertices correspond to the pages of the WWW, and directed edges describe hyperlinks between pages, our Wikipedia network could be considered as its special case—wikigraph.

For creation of the Wikipedia network for the Russian-language segment of Wikipedia we gathered information about 5.3 million of Wikipedia pages and more than 150 million of links between them. As a data source, the official dump of Wikipedia database distributed by Wikimedia Foundation was used. MySQL RDBMS and SQL queries were used for data preparation - for instance, for removing technical Wikipedia pages, such as talk articles. Here are the main steps of our technology of Wikipedia network creation:

1. Official SQL dump files of WikiMedia (it is wiki engine of Wikipedia) database for the Russian-language segment of Wikipedia are downloaded.
2. Tables from dump, which describe article categories and references between them, are deployed in MySQL database.
3. Information, which is important for our Wikipedia network, is downloaded to CSV files using SQL queries.
4. CSV files are cleaned and normalized.
5. Information from CSV files is imported into Neo4j graph database [11].
6. Further analysis of the network is made in Neo4j database through queries in Cypher query language.

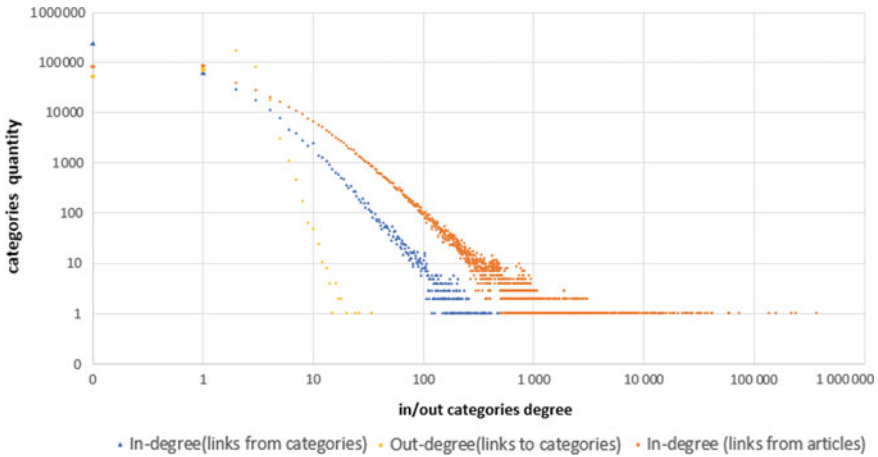
## 4 Analysis of Nodes degree Distribution for Article Nodes

As the result of our technology of Wikipedia network creation, the network of Russian language segment of Wikipedia was built. It consists of 3,305,000 articles and 405,000 categories. Basic statistics of the network is reported in Table 2.

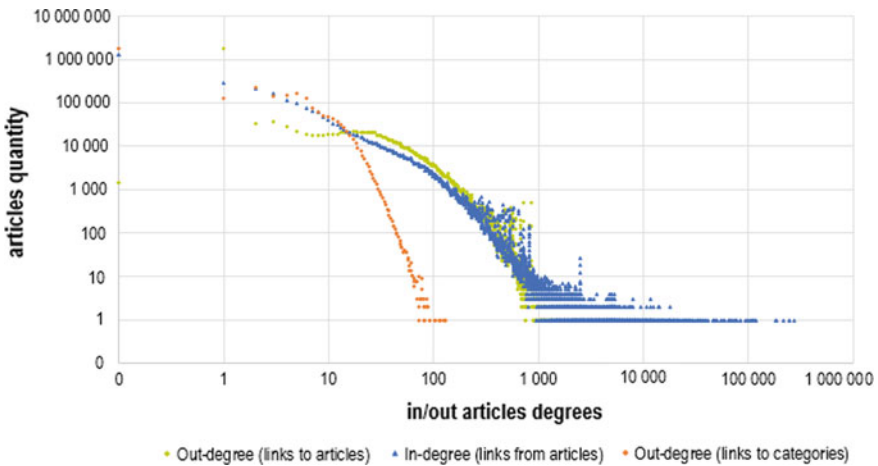
The next step was the analysis of nodes degree distributions. In Fig. 2 there are node degree distributions’ plots for links between categories and ingoing article-to-category links. In Fig. 3 node out-degree distribution for article-to-category links is presented. Tails of all these distributions are quite similar to linear relationship in log-log plot. This means that it is possible to try to fit power law distribution to

**Table 2** Basic statistics of the network built for the Russian-language segment of Wikipedia

Node types	Nodes’ quantity (in thousands)	Links’ quantity (in thousands)	
		To articles	To categories
Articles	3,305	92,167	9,187
Categories	405	–	770
Total	3,710	92,167	9,958



**Fig. 2** Log-log plot of nodes degree distributions for category nodes of the Russian-language Wikipedia



**Fig. 3** Log-log plot of nodes degree distribution for article nodes of the Russian-language Wikipedia

tails . However, even without strict statistical testing we can see that artifacts in the ends of tails of category nodes in-degree distributions will sufficiently alter estimated parameters and it will lead to a rejection of the statistical hypothesis of acceptance of any simple distribution law. Hence firstly we need to research the causes of these deviations, and only after that we could try to find certain models which explain these distributions.

In Fig. 3 there are node degree distributions' plots for links between articles and for outgoing article-to-category links. Nodes degree distribution for in-degree

(article-to-article links) has two linear relationship ranges with different inclination. This distribution also has a lot of artifacts for nodes degrees' values more than 200. Artifacts here is a large amount of outlier points which deviate very much from the distribution trend.

There is quite a different situation for nodes degree distribution for out-degrees of article-to-article links. This distribution doesn't comply with the power law because its tail is concave in log-log plot (see Figs. 2 and 3). Also, this node degrees distribution has a lot of artifacts in degrees' range from 300 to 1000.

To describe nodes degree distributions for article-to-article links, especially for out-degree case, we need some special model of network growth sufficiently different from the classical preferential attachment model. For these purposes, we propose to adapt Barabási-Albert model. Adaptation must consider multi-step process of Wikipedia articles formation instead of single-step process of adding a node to a network which is considered in the basis of Barabási-Albert model [8]. A growing network model could incorporate two types of growth processes: node creation with immediate attachment to a existing nodes, and link creation between preexisting nodes. Moreover, the model could include non-trivial rules for selection of the created links direction. There are examples of such adaptations in [7, 9, 12, 13].

The behavior of ends of tails of these distributions is not clear because of significant diffusion of values. Let's consider the cause of this diffusion. A nodes degree distribution is a distribution of a discrete random variable which takes non-negative integer values—node degrees (it is denoted as  $k$ ). Probability of the random node has a degree  $k'$  is  $p_{k'} = P(X = k')$  which is a real number in range from 0 to 1. But a nodes degree distribution for an empirical network is an empirical distribution function which estimates a probability density function of a random variable by consideration of a sample containing  $n$  values. For an empirical network with set of nodes  $N$ , which contain  $n$  nodes, we can get only empirical probability that a random network node has a degree  $k'$ :

$$\hat{p}_{k'} = \frac{|\{n|k(n) = k', n \in N\}|}{n}$$

Since the sample size  $n$  is finite, values of  $\hat{p}_{k'}$  are discrete and the lower positive value of  $\hat{p}_{k'}$  is  $1/n$ . It means that for a certain sample size (in our case a certain size of an empirical network) some  $p_{k'}$  will be relatively low and  $np_{k'}$ —expected value of  $|\{n|k(n) = k', n \in N\}|$  (quantity of nodes with degree  $k'$  in this network)—will be near to 1 or lower than 1. Hence in a range of  $k'$  values, for which  $p_{k'}$  are relatively low, values of  $\hat{p}_{k'}$  will randomly take values between several discreet levels. In particular, if  $k$  in certain range has expected value  $np_k \ll 1$ , then  $\hat{p}_k$  in this range will take one of the two values 0 or  $1/n$  by random. But the real share of 0 and 1 values will be dependent of a certain  $p$  value. In our case this mechanism of appearance of several discreet levels of nodes degree distribution for relatively low  $p_k$  explains diffusion of tails of nodes degree distribution in plots in Figs. 2 and 3.

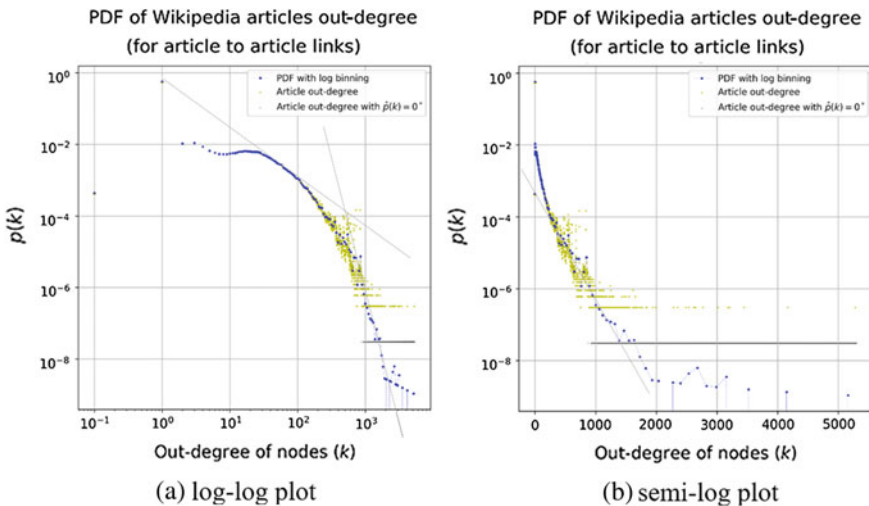
For reconstruction of low  $p_{k'}$  value we could use information about several  $\hat{p}_k$  values in the neighborhood of the considered value of  $k'$ . Estimation of  $p_{k'}$  based

on several neighbor  $\hat{p}_k$  values relies on the hypothesis  $p_k$  that values in considered range of degrees  $k$  are sufficiently close.

### 5 Methods for Clearing Diffusion of Tails of Nodes Degree Distributions

There are several methods for clearing diffusion of tails of nodes degree distributions. One of the simplest is using log bin estimation (e. g. see [14]) of node degree distribution. We used this method for increasing quality of nodes degree distribution analysis for out-degree of article-to-article links (see Fig. 4). In this method empirical distribution function is replaced with normalized frequency histogram. Instead of a constant bin width in a log bin estimation, each next bar bin width is  $q$  times wider than the previous one. This method actively uses aggregation of information about empirical probability for neighboring  $k$  values. In a plot with a logarithmic scale of a horizontal axis it looks like a graph with constant horizontal step (width) of bars. In Fig. 4 log bin estimation is shown for visual clarity in a form of a scatter plot instead of a histogram.

More precise analysis of this node’s degree distribution could be done with adaptive kernel density estimation method [15, 16]. This method also uses aggregation of information about empirical probability for neighboring  $k$  values but doing it in more precise and adaptive way. Adaptive kernel density estimation method selects



**Fig. 4** Plot of nodes degree distribution for out-degrees of article-to-article links of the Russian language Wikipedia and its log bin estimation (\* because of the impossibility to show values with  $\hat{p}(k) = 0$  on a plot with logarithmic y axis, these values are shown with gray dots and an artificial value, which is set an order of magnitude lower then minimum positive  $\hat{p}(k)$ )

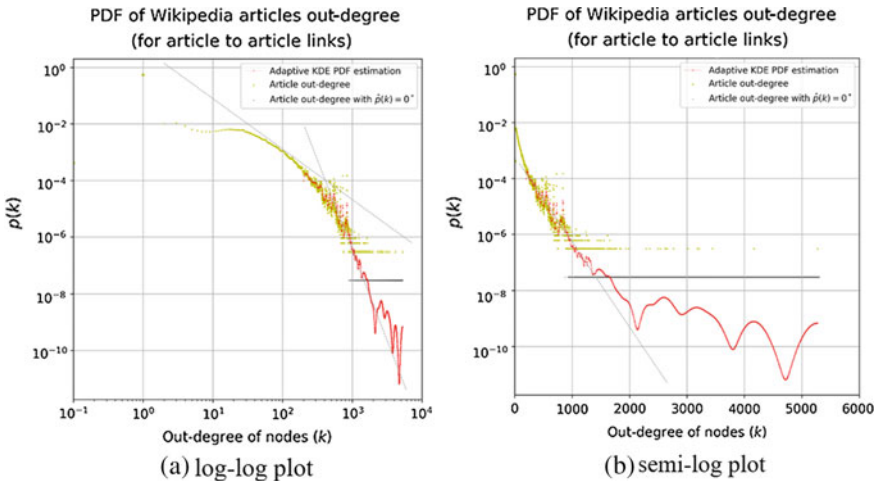


the width of kernel to minimize mean integrated error between the estimated rate and unknown underlying rate.

Like a log bin estimation of nodes degree distribution, the adaptive kernel density estimation method allows to clear diffusion of tails of nodes degree distributions. But this method is not constrained by exponentially growing step along horizontal axis, it allows to calculate probability estimation virtually for any point of distribution.

The usage of gaussian kernel or other kernels with similar shape allows to get a more precise estimation of probability for certain value of degree  $k$  in case of sharp changing of probability density. The reason is that this method can adapt the width of kernel to actual distribution probability. In particular, it can use a narrower kernel width, if in some range of nodes degrees probability  $\hat{p}_k$  is increased with increasing  $k$ . This adaptivity for bin width is impossible in log bin method. In [16] it is shown that adaptive kernel density estimation method, compared to other methods, gives lower integrated square error in case of estimation of probability density with sawtooth shape. It is similar to our case: in nodes degree ranges for artifacts, difference in probability of neighbor points could be more than one order of magnitude.

In our case adaptive kernel density estimation method has demonstrated its superiority over log bin estimation of nodes degree distribution in degree ranges with intensive artifacts. As can be seen in Fig. 4, log bin estimation of nodes degree distribution definitely indicates only one artifact in degree range from 600 to 700. A more precise analysis of nodes out-degree distribution for article-to-article links with adaptive kernel density estimation method sufficiently better shows (see Fig. 5) deviations of the density function around the trend line. It indicates 6 different artifacts of nodes degree distribution density in nodes degree range from 300 to 700.



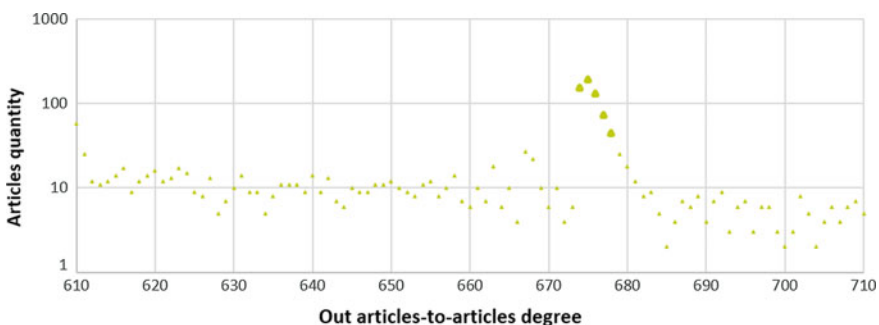
**Fig. 5** Plot of nodes degree distribution for out-degrees of article-to-article links of the Russian-language Wikipedia and its adaptive kernel density estimation (\* because of the impossibility to show values with  $\hat{p}(k) = 0$  on a plot with logarithmic y axis, these values are shown with an artificial value, which is set an order of magnitude lower than minimum positive  $\hat{p}(k)$ )

Adaptive kernel density estimation method has made it possible to understand that there are several systematically repeated abnormal peaks of nodes degree distribution density and find certain values of these peaks. It helps to do empirical research of certain Wikipedia pages from the different distribution density peaks and to formulate a hypothesis about the cause of these artifacts.

## 6 Elimination of the Source of Deviation for Out-Degree Probability Density Function

For identification of the cause of deviations of the nodes out-degree distribution around the trend line we made an analysis of outlier points in the distribution. In particular, we reviewed Wikipedia articles with out-degree in range from 674 to 678. Pages with out-degrees from this range occur an order of magnitude more often than pages with out-degrees near to the range (see Fig. 6). As the result of reviewing of Wikipedia articles with out-degree in the considered range, we have found that there are too many articles about French geographical names (more precisely, communes—units of administrative division in the French Republic). We reviewed Wikipedia articles of these communes (e.g. see article <https://ru.wikipedia.org/wiki/%D0%9B%D0%B5-%D0%91%D1%8E%D1%80%D0%B3%D0%BE> Le Burgaud commune) and found usual pages with about 10 regular outgoing links to other articles. But in the bottom of every of these pages there is a navigation block, which includes several hundred references to all communes of one of the French departments. By default, navigations blocks in Wikipedia pages are minimized and typical users usually doesn't see and notice all these links.

Navigation blocks are added to articles pages by using Wikipedia templates - pages from special namespace which are intended to be included into other pages to insert standard text. That is why all communes of the department have the same navigation block with the same quantity of links in it. Any navigation block which



**Fig. 6** Log-log plot for out-degrees of article-to-article links of the Russian-language Wikipedia (out-degree range which includes outliers in range from 674 to 678)

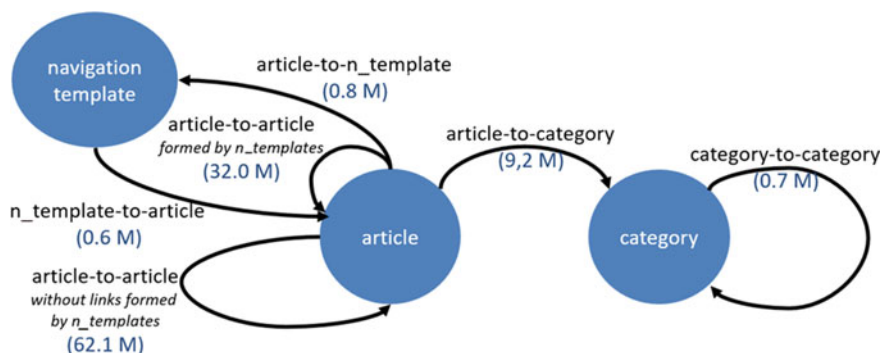
has references to  $N$  pages and is included in all  $N$  pages, which are referenced in this block, adds to Wikipedia  $N^2$  links.

As typically, pages which are included in the same navigation block have quite similar structure and size, they have about the same amount of regular outgoing links. Let  $L$  be an average quantity of regular links for pages from a navigation block. Then,  $N$  links from the navigation block shift out-degree for all its pages to a neighborhood of out-degree value  $L + N$ . If  $N$  is quite big, then addition of  $N$  new pages to the neighborhood of out-degree value  $L + N$  adds much more pages per one out-degree value than typical frequency of pages for these out-degree values. That is why navigation templates become the cause of deviations from trendline of out-degree distribution. An example of this process for the case of communes in the French department is shown in the Fig. 6. If we separate all article-to-article links into regular and formed by navigation templates, we could test if the presence of navigation blocks is the reason for all major outlier points in the nodes out-degree distribution.

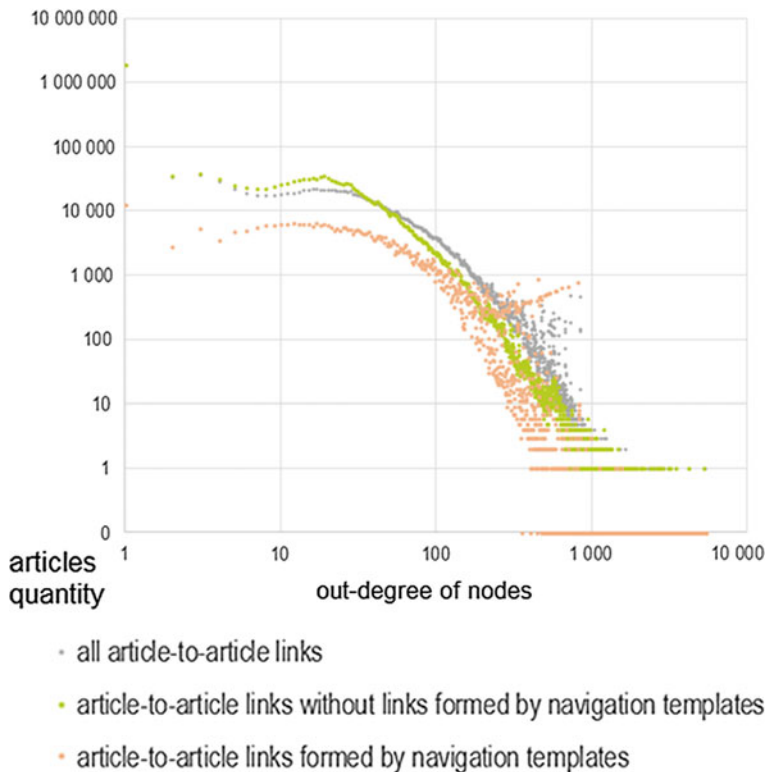
In general, pages which use the same navigation template (or other types of templates) form could be considered as a special class of pages. That is, templates could be regarded as a sort of categories. It is especially appropriate for navigation templates, because pages included in one navigation block usually describe objects from one class. We can add a new type of nodes—the navigation template type—to the scheme of Wikipedia network. The changed Wikipedia network scheme is shown in Fig. 7.

With the new type of nodes, we also need to add new types of links: article to navigation template links (article-to-n\_template); navigation template to article links (n\_template-to-article). Using information from article-to-n\_template links and n\_template-to-article links we could separate all article-to-article links into two categories: links, formed by navigation templates, and regular links. As a result, we can eliminate outliers from the out-degree distribution for regular links.

32 million links from 94.1 million of article-to-article links in Russian language segment of Wikipedia are formed due to using navigation templates. It means that



**Fig. 7** Wikipedia network scheme with three types of nodes and quantity of different types of links (in million of pcs)



**Fig. 8** Log-log plot of PDF of Wikipedia articles out-degree (for article to article links)

34% of all article-to-article links are formed due to affiliation of articles to a special type of categories implemented in the form of navigation templates. These links have other semantic information about relation between articles and their separation from regular links could improve quality of estimation of semantic relation between terms based on Wikipedia links.

Separation of article-to-article links formed by navigation templates indeed eliminates all major outlier points in out-degree probability density function for regular out-links (see Fig. 8). By contrast, out-degree probability density function for out-links, generated by navigation templates, has a lot of points which have a several order of magnitude larger quantity of articles than the trendline for density function. Furthermore, outlier points for distribution for all article-to-article out-links are situated in the same range approximately from degree 200 to degree 1000—the same range where most of outlier points are situated for out-links generated by navigation templates. Moreover, the tail of out-degree distribution for regular article-to-article links is sufficiently less concave than for the non-separated distribution. After this correction the tail of this distribution can be explained by models which are alike to the preferential attachment model [8]; in particular based on an adaptation of this model with multi-step growth process [7].

## 7 Conclusion

In the current research, essential analysis of scheme and network structure of Russian language segment of Wikipedia has been done. Within the scope of this work, technology of Wikipedia network creation and its analysis with the use of graph database were developed. Essential network scheme with two type of nodes (for Wikipedia articles and categories respectively) and three type of links were developed. It has been found that tails of nodes degree distributions for categories out- and in- links could have the power law distribution and could be explained by models based on the preferential attachment model. Nodes degree distribution for articles in- links has two linear ranges with different inclination and has a lot of outlier points which must be explained.

It was found that nodes degree distribution for articles out- links doesn't fit the power law and has extensive artifacts: large amount of outlier points which deviate very much from the trend. This artifacts and end of tail of distribution have been analyzed using the log bin estimation method and the adaptive kernel density estimation method. Analysis of certain artifacts has shown that they were induced by using navigation templates in Wikipedia articles. For eliminating these artifacts, the Wikipedia network scheme was modified: a new type of nodes and links representing Wikipedia templates and references linked to them was added. It was made possible to separate article-to-article links, generated by navigation templates, from regular links. As a result of separation, it was found that 34% of all article-to-article links in the Russian-language segment of Wikipedia are formed due to using navigation templates. These links have other semantic information about relations between articles and their separation from regular links could improve the quality of evaluation of semantic relations between terms based on Wikipedia links. As expected, this separation eliminated all major outlier points in out-degree probability density function for regular out-links.

## References

1. Wikipedia, the free encyclopedia. <http://wikipedia.org>. Accessed in 2018
2. MediaWiki, free software open source wiki package. <https://www.mediawiki.org/wiki/MediaWiki>. Accessed in 2018
3. DBpedia, Large, multilingual, semantic knowledge graph. <http://dbpedia.org>. Accessed in 2018
4. Zhou, B., et al.: Wikipedia-graph based key concept extraction towards news analysis. In: CEC '09 Proceedings of the 2009 IEEE Conference on Commerce and Enterprise Computing, pp. 285–292 (2009). <https://doi.org/10.1109/CEC.2009.54>
5. Chernov, S., et al.: Extracting semantics relationships between Wikipedia categories. SemWiki 206 (2006)
6. Kamps, J., Koolen, M.: Is Wikipedia link structure different? In: WSDM '09 Proceedings of the Second ACM International Conference on Web Search and Data Mining, pp. 232–241 (2009). <https://doi.org/10.1145/1498759.1498831>
7. Capocci, A., et al.: Preferential attachment in the growth of social networks: the internet encyclopedia Wikipedia. Phys. Rev. E **74**(3), 036116 (2006). <https://doi.org/10.1103/PhysRevE.74.036116>

8. Barabasi, A.-L., Albert, R.: Emergence of scaling in random networks. *Science* **286**(5439), 509–512 (1999)
9. Krapivsky, P.L., et al.: Degree distributions of growing networks. *Phys. Rev. Lett.* **86**, 5401 (2001). <https://doi.org/10.1103/PhysRevLett.86.5401>
10. Caldarelli, G., et al.: Scale-free networks from varying vertex intrinsic fitness. *Phys. Rev. Lett.* **89**(25), 258702 (2002). <https://doi.org/10.1103/PhysRevLett.89.258702>
11. Neo4j, graph database management system. <https://neo4j.com>. Accessed in 2018
12. Tadic, B.: Dynamics of directed graphs: the world-wide Web. *Physica A* **293**, (2001)
13. Barrat, A., Pastor-Satorras, R.: Rate equation approach for correlations in growing network models. *Phys. Rev. E* **71**, 036127 (2005)
14. Gravino, P., Servedio, V., Barrat, A., Loreto, V.: Complex structures and semantics in free word association. *Adv. Complex Syst.* **15**(34), (2012)
15. Parzen, E.: Estimation of a probability density-function and mode. *Ann. Math. Stat.* **27**(3), 832–837 (1956)
16. Shimazaki, H., Shimomoto, S.: Kernel bandwidth optimization in spike rate estimation. *J. Comput. Neurosci.* **29**(1-2), 171–182 (2010)

# Indirect Influence Assessment in the Context of Retail Food Network



Fuad Aleskerov, Natalia Meshcheryakova and Sergey Shvydun

**Abstract** We consider an application of long-range interaction centrality (LRIC) to the problem of the influence assessment in the global retail food network. Firstly, we reconstruct an initial graph into the graph of directed intensities based on individual node's characteristics and possibility of the group influence. Secondly, we apply different models of the indirect influence estimation based on simple paths and random walks. This approach can help us to estimate node-to-node influence in networks. Finally, we aggregate node-to-node influence into the influence index. The model is applied to the food trade network based on the World International Trade Solution database. The results obtained for the global trade by different product commodities are compared with classical centrality measures.

**Keywords** Influence estimation · Network analysis · Food trade network

## 1 Introduction

There is no human being who can survive without adequate nutrition in our world. There are a lot of factors in the context of nutrition that influence the quality of life and the health of people. It includes ground productivity of the area where people live at, the availability of products in markets, the opportunity of inhabitants to buy food there, trade relations with other countries, etc. A huge number of organizations (FAO [1], WHO [2], etc.) were funded in order to govern trade systems and keep track of the nutrition situation in the world. These organizations, along with coordinating issues, study statistical indicators of countries concerning diet quality, quantity, diversity, conduct public surveys among people and households, etc. However, not so many

---

F. Aleskerov · N. Meshcheryakova (✉) · S. Shvydun  
National Research University Higher School of Economics,  
20 Myasnitskaya Str., Moscow 101000, Russian Federation  
e-mail: [natamesc@gmail.com](mailto:natamesc@gmail.com)

V.A. Trapeznikov Institute of Control Sciences of Russian Academy  
of Sciences, 65 Profsoyuznaya Str., Moscow 117997, Russian Federation

© Springer Nature Switzerland AG 2020  
I. Bychkov et al. (eds.), *Network Algorithms, Data Mining,  
and Applications*, Springer Proceedings in Mathematics & Statistics 315,  
[https://doi.org/10.1007/978-3-030-37157-9\\_10](https://doi.org/10.1007/978-3-030-37157-9_10)

studies and measures are dedicated to interrelations between countries with the aim to reveal some new patterns and knowledge about the worldwide system. The study of this problem from the global point of view can give us a fundamental understanding of the structure and organization of these relations in the whole world.

In this work, we join both statistical measures and global retail food network in order to study trading relations of countries. Undoubtedly, this type of relations directly impacts the inner situation with nutrition in each country. Our main goal is to analyze how countries influence each other through trade relations with the aim of detecting the most influential participants as well as the most dependent ones. We assume that countries can influence each other even if they are weakly or completely not bargaining with each other directly. Long-distant connections play a significant role in this type of network and countries may have an impact on each other through intermediate participants of these long trading chains.

We compare the results of measures that take into account long-range connections with classical centrality indices that elucidate important nodes in network structures and with some statistical indicators of countries related to food context. We analyze both total trade and different product commodities in order to reveal key elements on each product level.

This paper is organized as follows. First, we describe some measures that rank countries in terms of their nutrition state and measures that detect key elements in the network. Second, we analyze and compare different approaches that consider long-range connections between elements. Third, we describe the database that provide data concerning bilateral trade statistics. Finally, we provide results and make a conclusion.

## 2 Literature Review

In this Section, we describe measures that rank countries with respect to their food balance and nutrition as well as some network measures that help us to detect key elements in networks.

We address to FAOSTAT database [3] in order to present different food measures. Food balance sheet (FBS) includes several estimations as Domestic Supply Quantity (DSQ), Food, Losses, Production, Food Supply Quantity (FSQ), etc.

DSQ is a measure that includes such components as production, total imports and exports values and changes in stocks. These components form supply for domestic utilization. Food data include the total amount of a commodity available for the period in question for human dietary needs. Losses are calculated as the amount of a particular product lost through wastage during different stages of processing. FSQ indicates the amount of food supply in kilograms per capita per year.

All these measures are estimated at a very detailed level for different types of products, for example, different race of potatoes, various citrus fruits, etc. We analyze aggregated commodities and select cereals, meat, and vegetables in order to rank countries. The results by different measures for 2010 are provided in Table 1. We



**Table 1** Top-5 countries by different food balance sheet indicators in 2010

DSQ			Food		
Cereals	Meat	Vegetables	Cereals	Meat	Vegetables
China	China	China	China	China	China
USA	USA	India	India	USA	India
India	Brazil	USA	Indonesia	Brazil	USA
Brazil	Russia	Turkey	USA	Russia	Egypt
Indonesia	Germany	Egypt	Bangladesh	Germany	Iran
Losses			Production		
Cereals	Meat	Vegetables	Cereals	Meat	Vegetables
China	Peru	China	China	China	China
India	Canada	India	USA	USA	India
Brazil	Argentina	Turkey	India	Brazil	USA
Indonesia	Japan	USA	Brazil	Germany	Turkey
Mexico	Myanmar	Iran	France	Russia	Iran
FSQ					
Cereals		Meat	Vegetables		
Morocco		HK	China		
Egypt		USA	Armenia		
Lesotho		Australia	Tunisia		
Azerbaijan		Bahamas	Montenegro		
Mali		Macao	Albania		

can see that China and USA are leading almost for all indicators by all types of commodities. India is also very representative for cereals and vegetables while Brazil is among leaders for meat.

However, these measures indicated in the food balance sheet do not consider relationships between countries, which is vitally important for the understanding of global processes.

When we talk about relations between elements, we appeal to social network analysis that helps us represent different types of connections. Export/import relations are a vitally important part of the analysis of quality and safety of nutrition of countries. In this work, we analyze countries in terms of the influence and dependency of each other and products they purchase.

One of the ways to detect the most important elements in a network is to apply classical centrality measures. There exist various well-known measures that address the problem of key nodes detection. The simplest centrality measure refers to *degrees of nodes* [4]. In the context of food export/import network, *weighted in-degree measure* indicates total import volumes, *weighted out-degree measure* indicates total export volumes and *unweighted* versions of these measures indicate the total number of

trading partners. Generally, this measure does not analyze long-range connections in a network but only considers local environment of nodes.

In order to take into account distant connections in a network, we can apply measures that are based on eigenvector calculation. This group of measures includes *eigenvector centrality*, *PageRank*,  $\alpha$ -*centrality*, *Katz centrality*, *Hubs and Authorities centralities*, etc. [4–6]. All these measures estimate the importance of nodes considering the importance of their neighbors, neighbors of neighbors, etc.

Two the most popular measures that are based on the shortest paths in a network are called *betweenness* and *closeness* centralities [4]. *Betweenness centrality* estimates the number of times a particular node is met on the shortest paths between pairs of nodes while *closeness centrality* takes into account the distance from a particular node to other nodes in a graph. Similar ideas (*decay centrality*, *stress centrality* and *percolation centrality*) are proposed in [7–9]. These centralities also consider the length of paths in a graph.

The main disadvantage of classical centrality measures in the context of retail food network is that they do not account for the size of nodes. When we talk about the size of nodes we mean some individual attributes that determine the importance and the power of a node. As a consequence, the same amount of a flow can be significant for one “small” node and this value of a flow can be negligible for some “big” node. This is why it is critical to normalize flows to the size of nodes with respect to their individual attributes.

The approach that takes into account individual parameters of nodes is described in [10]. *Long-range interaction centrality index* (LRIC) has various applications for the influence estimation in different network structures. This index detects both powerful nodes and dependent nodes. Next, we give a brief description of this index and determine some new ways of long-range interactions estimation.

### 3 Long-Range Interaction Centrality (LRIC)

In this Section, we describe LRIC index in the context of a food trade network and show the importance of all stages. The first stage of LRIC calculation involves attributes analysis. Each country has some related attributes as production level, consumption level, total export, total import, population, standard rate of consumption, etc. Hence, when we assess the influence of one country on another country through the trade flow between them we need to consider these additional factors. Roughly speaking, for each country its production minus export plus import is the total amount of some good that can be devoted to domestic needs. Imagine that some country stopped exporting to this country, which means that the supply level decreased for some percent. If the remaining resource is enough for dairy needs of a country population, then the exporting country evidently is not influential in terms of food trade. Hence, we can introduce the notion of a quota for each country. Quota is some number homogeneous to a flow when this country becomes affected by other countries, in other words, quota is a critical level of the amount of a good, at which

country incurs a deficit of this good. Quota can be calculated as some aggregated value of different parameters of a country.

When we define quotas for each country, next we start analyzing direct flows between elements of a network. LRIC takes into account direct influence of nodes on each other individually as well as in groups. More precisely, if a country cannot overcome a quota of its trading partner on its own it can affect this partner if it unites with some other countries. Hence, we will call group  $\Omega(j)$  of node  $j$  as the influential one on node  $j$  if it satisfies the following condition:

$$\sum_{i \in \Omega(j)} w_{ij} \geq q_j, \quad (1)$$

where  $w_{ij}$  is a flow from country  $i$  to country  $j$  and  $q_j$  is an individual quota of node  $j$ . We will say that node  $k \in \Omega(j)$  is a key or pivotal node of group  $\Omega(j)$  if without this node group  $\Omega(j) \setminus \{k\}$  is not influential anymore, i.e.,  $k$  is a key node if

$$\sum_{i \in \Omega(j)} w_{ij} \geq q_j \quad \text{and} \quad \sum_{i \in \Omega(j) \setminus \{k\}} w_{ij} < q_j. \quad (2)$$

Finally, we will denote a minimal influential group of node  $j$  by  $\Omega_p(j)$ , i.e., this is an influential group where all nodes are pivotal:

$$\sum_{i \in \Omega_p(j)} w_{ij} \geq q_j \quad \text{and} \quad \forall k \in \Omega_p(j) \quad \sum_{i \in \Omega_p(j) \setminus \{k\}} w_{ij} < q_j. \quad (3)$$

LRIC defines an individual influence of each node considering influential groups and the fact that nodes can be pivotal in these groups in the following way:

$$c_{ij} = \begin{cases} \max_{\substack{\Omega_p(j): \\ i \in \Omega_p(j)}} \frac{w_{ij}}{\sum_{k \in \Omega_p(j)} w_{kj}}, \\ 0, \quad \text{if node } i \text{ is not pivotal in } \forall \Omega_p(j). \end{cases} \quad (4)$$

Formula (4) defines the maximal possible intensity of the direct influence of node  $i$  on its adjacent node  $j$  denoted by  $c_{ij}$ . The intensity varies between 0 and 1, where  $c_{ij} = 0$  is the absence of the influence and  $c_{ij} = 1$  is the maximal influence. Evidently, if node  $i$  exceeds the quota of node  $j$  alone, i.e.,  $w_{ij} \geq q_j$ , then the influence of node  $i$  on node  $j$  should be maximal and equal to 1. In other cases, the intensity of the direct influence of node  $i$  on node  $j$  is defined as the maximal possible intensity with respect to all  $\Omega_p(j)$  where node  $i$  is pivotal. If nodes are not connected or node  $i$  is not pivotal in any minimal influential group of node  $j$ , then the intensity of the direct influence is equal to zero and we eliminate such edges from the network. Additionally, we can introduce the parameter of the maximal group size denoted by  $m$ . As a rule, this parameter is not very large as in many real-life situations participants do not unite in large groups in order to influence other participants.

As the result, this stage helps us normalize initial values of flows with respect to individual parameters of countries. But this reconstruction of a network to the network of direct intensities does not consider long-range connections yet. In real processes, some actors may influence other actors through intermediate participants. The same intuition can be applied to the influence in networks when some nodes influence other nodes through intermediate nodes. This assumption is analyzed during the next stage.

LRIC considers long-range connections through all simple paths in a network. A simple path in a graph is a chain of edges that connect adjacent distinct nodes. One node can influence other node and as the result the target node becomes affected. This leads to the fact that this node can influence its neighbors with some power and this domino effect can spread deeply. In terms of a trade network, it can be the result of the fact that some country stops exporting to one of its neighbors and this neighbor has to decrease its export to its partner proportionally to the deficit of the product, etc. Basically, there can be several paths between a pair of nodes through which one node can influence the other one.

Denote by  $P_k(i, j)$  the  $k$ -th simple path between nodes  $i$  and  $j$ . In other words,  $P_k(i, j)$  is a chain on edges  $(i, h_1), (h_1, h_2), \dots, (h_{l-1}, j)$  and  $l_k(i, j)$  is the length of this path. LRIC suggests various techniques how to estimate the influence of node  $i$  on node  $j$  through path  $P_k(i, j)$ . For instance, we can multiply the intensities of the direct influences on the edges of this path, which is thought of as joint probability of the influence denoted by  $\Pi(P_k(i, j))$ :

$$\Pi(P_k(i, j)) = c_{ih_1} \times c_{h_1h_2} \times \dots \times c_{h_{l-1}j}. \quad (5)$$

More precisely, node  $i$  influences node  $h_1$  with intensity  $c_{ih_1}$  and node  $h_1$  in its turn influences node  $h_2$  with intensity  $c_{h_1h_2}$ , which means that  $i$  influences  $h_2$  through  $h_1$  with the intensity of  $c_{ih_1} \times c_{h_1h_2}$  and so on until we reach node  $j$ . It can be seen that the more distant nodes are, the less impact the source node has on the target node.

Another way of the influence estimation of node  $i$  on node  $j$  through path  $P_k(i, j)$  proposed in [10] is to take the minimal value of the intensity on this path denoted by  $M(P_k(i, j))$ :

$$M(P_k(i, j)) = \min(c_{ih_1}, c_{h_1h_2}, \dots, c_{h_{l-1}j}). \quad (6)$$

The idea of this approach is to take the upper bound estimate of the influence as one node cannot influence another node through the path between them more than the minimal value on this path. This idea is applied in various problem areas and is known as a channel capacity problem. Again, the more remote a node is, the less impact we have on this node. Hence, it is reasonable to introduce the parameter of the maximal length of the path denoted by  $l_{max}$  and to consider paths with  $l_k(i, j) \leq l_{max}$ .

As there can be several paths between two nodes, we need to aggregate the influence of one node on another node over all paths between them. In [10] several ways of aggregation are proposed, for instance, we can sum up over all values of  $\Pi(P_k(i, j))$  as

$$c_{ij}^{sum\Pi} = \min \left( \sum_k \Pi(P_k(i, j)), 1 \right). \quad (7)$$

As we consider the influences in the interval from 0 to 1 and the summation over  $\Pi(P_k(i, j))$  can be more than 1 we limit this value with 1.

Another way of the paths aggregation is to take the maximal of the influences over all paths

$$c_{ij}^{max\Pi} = \max_k (\Pi(P_k(i, j))), \quad (8)$$

$$c_{ij}^{maxM} = \max_k (M(P_k(i, j))). \quad (9)$$

As the result, there exist several approaches of long-range interactions assessment that help us to obtain total node-to-node influence in a network. The approaches described in [10] are based on all simple paths enumeration which is a complex issue.

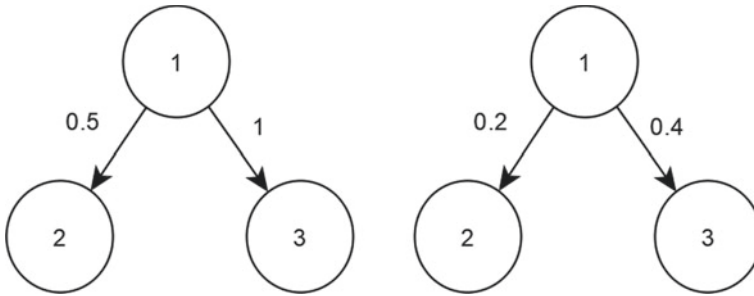
Another way to assess the indirect influence between nodes through long-range interactions is to calculate classical centrality measures for the graph of direct influences. If we consider the problem of ranking and do not estimate node-to-node influence, we can calculate eigenvector, Katz, PageRank and other centralities based on walks in a graph. However, if we want to assess pairwise influence we cannot use classical centrality measures in this case.

One of the ways to estimate node-to-node influence is to apply personalized PageRank centrality [11]. The most central nodes according to classical PageRank centrality are those nodes that can be reached with higher probability while random walking on a graph. More precisely, the value of node  $i$  is calculated as

$$PR(i) = d \left( \sum_j \frac{PR(j)}{\sum_k w_{jk}} \right) + (1 - d)E, \quad (10)$$

where  $d$  is a damping factor and  $E = [E_i]$  is a vector of probabilities,  $E_i = \frac{1}{n} \forall i$ ,  $n$  is the total number of nodes.

As we can see from formula (10) the structure of the classical PageRank combines both random walks and jumps to arbitrary nodes which we start walking again with. According to this formula, we jump to each node with equal probability as the second term of the formula is the same for each node in a graph. Personalized PageRank allows to introduce the probabilities of jumps for each node individually, i.e., vector  $E$  in the formula may contain different probabilities for all nodes. Obviously, the centrality value is rising for each node with the increasing of the probability of jumping to this node. As the sum of probabilities is equal to 1, the maximal probability for each node is also equal to 1 while other probabilities are equal to 0.



**Fig. 1** Two graphs with proportional weights of edges

In order to assess the influence of node  $i$  on all other nodes, we can calculate personalized PageRank on the graph of directed influences where the probability of a jump to node  $i$  is equal to 1 while other probabilities are equal to 0. The obtained centrality vector can be considered as the influences node  $i$  on other nodes ignoring self-influence.

We note that transition probability in random walks approach is proportional to weights on edges. It means that personalized PageRank does not distinguish between different values of weights. For instance, consider two graphs on Fig. 1, where weights indicate the intensity of a direct influence on node 1 on nodes 2 and 3. According to PageRank, the probability of transition from node 1 to node 2 is equal to  $1/3$  and from node 1 to node 3 is equal to  $2/3$  for both graphs. However, real weights on edges differ significantly.

In order to settle this problem, we change weights in the whole graph at each step of personalized PageRank calculation. Suppose that we estimate the influence of node  $i$  on other nodes. Hence, we apply personalized PageRank with  $E_i = 1$  and  $E_j = 0 \forall j \neq i$ . Additionally, we change weights of in-coming edges on node  $i$  as follows:  $\forall j \in N w_{ji}^{new} = N - 1 - \sum_k w_{jk}$ , where  $N$  is a set of all nodes in a graph. It means that the new weight of an edge between node  $j$  and node  $i$  is calculated as the difference between maximal potential influence (which is equal to  $N - 1$  when node  $j$  influences all other nodes with intensity 1 except self-influence) and real influence of node  $j$ . This approach allows to redistribute probabilities of transition relatively to real weights. Figure 2 illustrates the changes of graphs above if we calculate personalized PageRank for node 1.

After we add new edges or increase weights of existing edges to node  $i$  we return to node  $i$  more frequently while random walks. This approach increases self-influence of node  $i$  (which is later ignored) and redistribute influence on other nodes with respect to intensities of the direct influence. Moreover, returning edges allow to increase the influence of node  $i$  on its neighbors and on nodes from the closest surrounding of node  $i$ .

When we calculate personalized PageRank with described changes in a network for all nodes we obtain  $N$  vectors on length  $N$ , where vector  $i$  corresponds to the influ-

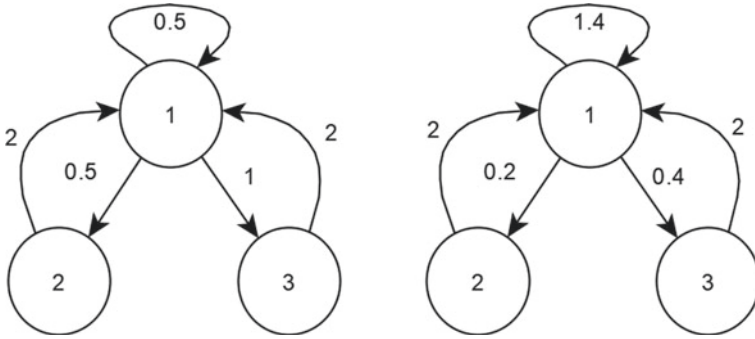


Fig. 2 New graphs of personalized PageRank calculation for node 1

ence of node  $i$  on other  $N$  nodes. These vectors form  $N \times N$  matrix that represents a new graph of indirect influences (ignoring diagonal elements that are self-influences).

The main advantage of this approach is the lower computational complexity compared to LRIC calculation based on simple paths enumeration. Empirical results show a high correlation of LRIC based on personalized PageRank with LRIC based on paths, which means that methods are interchangeable.

After we obtain total node-to-node influence by LRIC on paths or LRIC on personalized PageRank we can aggregate this information into the final influence within the whole network. Aggregation can be done with respect to different attributes as well as to the network structure.

Next, we apply classical centrality measures and proposed methods to export/import networks and compare the results.

### 4 Data Analysis and Results

We use World Integrated Trade Solution (WITS) database for the analysis of trade flows between countries. WITS was jointly developed and compiled by several databases as the UNSD Commodity Trade (UN Comtrade) (trade statistics), the UNCTAD Trade Analysis Information System (TRAINS) (tariffs and non-tariffs measures), the WTO’s Integrated Database (IDB) (Most Favored Nation data), the World Bank, the Center of International Business, Tuck School of Business and Global Preferential Trade Agreements Database (preferential trade agreements) [12].

UN Comtrade database, which provides to WITS detailed trade export and import statistics by commodities, is one of the largest databases of international trade data. The main advantages of using this data for the analysis of trade statistics are that it is free access, it collects data since 1962, covers more than 170 countries and territories, contains various classifications, regularly update and load new information, converts values into one unit (US dollars) using relevant exchange rates [13].

WITS (with UN Comtrade) provides data in various classifications (or, in other words, nomenclatures). There are two main systems that are used in international trade statistics: the Harmonized System (HS) and the Standard International Trade Classification (SITC). Such classifications are used by economics to standardize the values and to make data comparable.

In this research, we use SITC classification. SITC uses 5 aggregation levels, where each product group on each level is coded by some digits: the longer the code is (5 digits max), the more detailed the product division is. All products in SITC revision refer to one 1 digit category (code 0), while 2 digits codes under food category contains 10 subcategories (00–09) [14, 15].

We analyze results both for total food trade and for different product categories from 1996 to 2016. We take the following 10 product commodities: live animals, meat, dairy/eggs, fish, cereals, vegetables/fruits, sugar/honey, coffee/tea/cocoa/spices, feeding stuff for animals and miscellaneous. Firstly, we analyze basic graph characteristics in order to understand a general picture of the structure of networks. We calculated the total number of countries that participate in trading processes, total trade value for all networks, density of graphs and the number of weakly and strongly connected components (wCC and sCC).

We can see positive trends in the number of countries (Fig. 3), in the total amount of trade (Fig. 4) and in the density of trade (Fig. 5) for all product commodities. The number of wCCs is equal to 1 for all graphs, while the number of sCCs varies significantly for considered graphs. However, the giant sCC is unique for all graphs while other sCCs include single nodes.

As we can see, according to density parameter countries mostly trade in vegetables/fruits and coffee/tea categories while the largest volumes of trade (in 1000 USD) are in vegetables/fruits, cereals, meat and fish product categories. After that, we analyze a general picture of trade structure and estimate the most influential countries. Next, we provide results for total trade and 3 product categories (cereals,

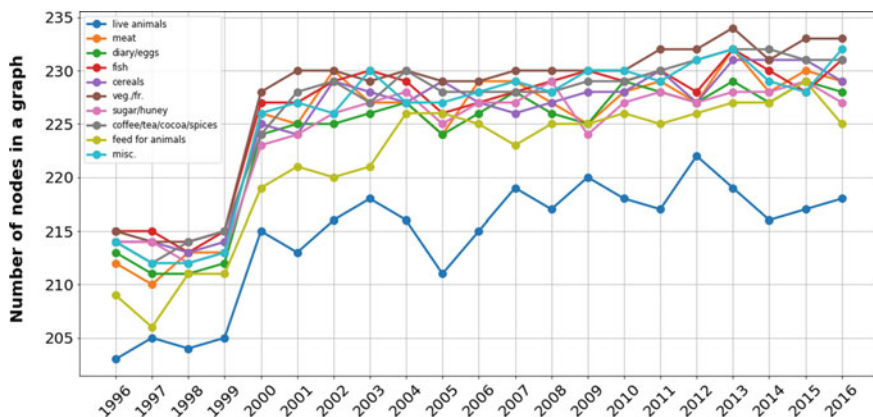


Fig. 3 Dynamics of the number of nodes



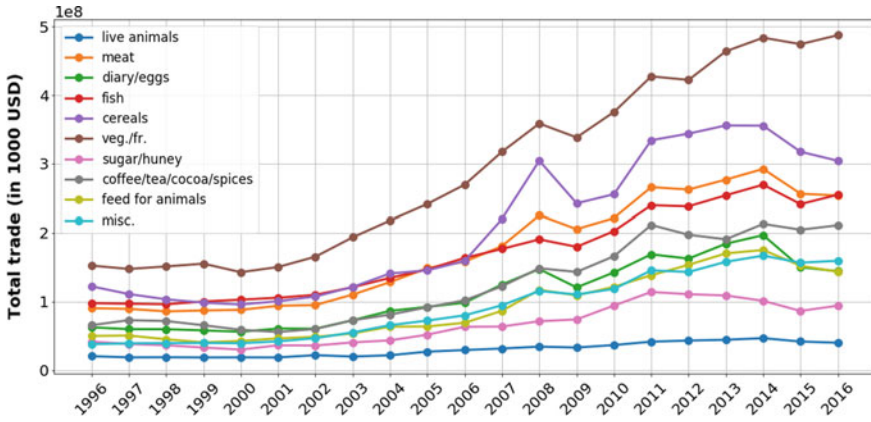


Fig. 4 Total trade dynamics

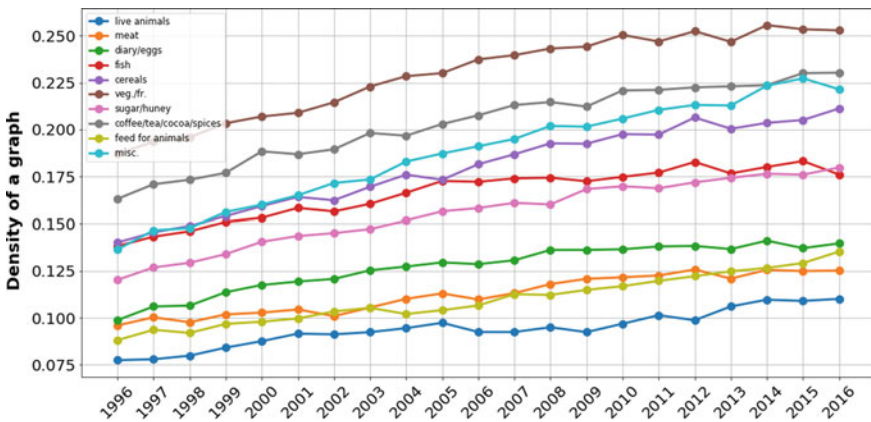


Fig. 5 Density dynamics

vegetables/fruits, and meat) in 2010 but other products and years were analyzed as well.

Table 2 indicates top 10 of the most central elements by different classical centrality measures (weighted/weighted in/out-Degree, eigenvector, PageRank, hubs and authorities).

In-degree centrality indicates the total number of suppliers. Seychelles, United States, France, Canada, Austria, and Hong Kong are major importers in terms of a total of number of trading partners. Out-degree centrality indicated the total number of purchasing countries. United States, Thailand, France, China, Italy, and Brazil have the largest numbers of the exporting partners. Weighted versions of in- and out-degrees indicate total import and total export values of particular products. The United States and Germany are key consumers and suppliers in terms of total trade while each

**Table 2** Classical centralities: top-10 countries by different product categories in 2010

Total trade							
inDeg	w. inDeg	outDeg	w. outDeg	Eigenvec.	PageRank	Hubs	Auth.
SYC	USA	USA	USA	USA	USA	USA	JPN
FRA	DEU	THA	DEU	NLD	BRA	NLD	USA
CAN	JPN	FRA	NLD	DEU	ARG	CAN	DEU
USA	GBR	GBR	BRA	FRA	DEU	CHN	CAN
AUT	FRA	CAN	FRA	CAN	NLD	DEU	GBR
DEU	NLD	BEL	CHN	BEL	CHN	FRA	FRA
GBR	ITA	CHN	ESP	ESP	FRA	MEX	NLD
NLD	RUS	NLD	BEL	MEX	CAN	BRA	ITA
POL	ESP	ITA	CAN	BRA	THA	BEL	MEX
ESP	BEL	MYS	ARG	ITA	AUS	ESP	BEL
Cereals							
inDeg	w. inDeg	outDeg	w. outDeg	Eigenvec.	PageRank	Hubs	Auth.
USA	JPN	USA	USA	CAN	USA	USA	JPN
CAN	USA	CHN	FRA	USA	CAN	CAN	MEX
FRA	DEU	ITA	CAN	FRA	FRA	AUS	KOR
GBR	NLD	THA	DEU	ITA	DEU	FRA	CAN
SYC	EGY	FRA	THA	DEU	ITA	THA	EGY
DEU	MEX	DEU	ARG	MEX	AUS	ARG	NGA
KOR	ITA	BEL	AUS	THA	THA	CHN	USA
ARE	GBR	GBR	ITA	BEL	BEL	ITA	VEN
BEL	FRA	NLD	IND	GBR	ARG	DEU	PHL
CHN	SAU	IND	BEL	NLD	GBR	RUS	CHN
Vegetables/Fruits							
inDeg	w. inDeg	outDeg	w. outDeg	Eigenvec.	PageRank	Hubs	Auth.
SYC	USA	USA	USA	MEX	USA	MEX	USA
FRA	DEU	CHN	CHN	USA	CHN	CAN	DEU
AUT	GBR	THA	ESP	CAN	CHL	CHN	CAN
CAN	FRA	ITA	NLD	CHL	THA	ESP	FRA
DEU	NLD	CAN	MEX	CHN	MEX	CHL	GBR
USA	RUS	NLD	ITA	ESP	CAN	USA	JPN
POL	JPN	FRA	BEL	CRI	ECU	NLD	NLD
GBR	CAN	ZAF	TUR	THA	NLD	CRI	RUS
NLD	BEL	BEL	FRA	NLD	ESP	ITA	BEL
CZE	ITA	DEU	CHL	ECU	ARG	ECU	ITA
Meat							
inDeg	w. inDeg	outDeg	w. outDeg	Eigenvec.	PageRank	Hubs	Auth.
SYC	JPN	USA	BRA	BRA	DEU	USA	JPN
FRA	GBR	BRA	USA	USA	NLD	BRA	MEX

(continued)

**Table 2** (continued)

Total trade							
HKG	DEU	FRA	DEU	DEU	BEL	AUS	RUS
VNM	FRA	NLD	NLD	ARG	FRA	DNK	HKG
CHN	RUS	DEU	AUS	URY	BRA	CAN	CAN
DEU	ITA	AUS	DNK	CAN	DNK	DEU	GBR
USA	NLD	ITA	FRA	AUS	ESP	NLD	NLD
ARE	USA	CHN	CAN	NLD	IRL	CHN	DEU
GBR	HKG	CAN	BEL	DNK	ITA	THA	USA
ESP	MEX	DNK	ESP	NZL	POL	BEL	ITA

product has distinct leaders by total export and import. Eigenvector, PageRank, and Hubs detect the most important exporters taking into account distant connections and Authorities centrality measure elucidates the most important importers considering distant connections.

We also calculate LRIC measures by paths and LRIC by personalized PageRank in order to identify the most influential countries. As to the threshold value, we take  $q\%$  of maximal value between total export and total import. If country  $i$  exports more than imports then no country can influence  $i$  through its import value otherwise country  $i$  can reduce its export and cover the losses. On the other hand, if country  $i$  imports more than exports then the loss of import values cannot be covered by export reduction. Hence, quotas for large exporters are calculated through export values while quotas for large importers are calculated through import values. Using parameter  $q$ , we analyzed different values between 1% and 20% and conclude that results do not distinguish a lot. Table 3 shows Kendall rank correlation of the final influence in total trade by one of the LRIC indices depending on different quotas.

For the next estimations, we choose  $q = 15\%$ . We also compare the results obtained by different centrality measures and LRIC indices with the help of correlation analysis. Table 4 indicates Kendall correlation of ranks calculated by different centrality measures on the network of total trade in 2010.

As we can see from Table 4 different versions of LRIC indices from [10] correlate at a very high level with each other and LRIC by personalized PageRank approach. What is more, for the total trade network LRIC indices correlate with weighted out-degree centrality at level 0.81–0.92. LRIC MaxMin is the least correlated with classical centrality measures and other versions of LRIC indices. We can also notice that authorities’ centrality does not correlate with other measures except weighted in-degree as both these measures detect key importers rather than exporters.

Finally, we provide the results for the networks of total trade, cereals, vegetables/fruits, and meat in 2010 obtained by the proposed long-range interactions measures as shown in Table 5.

We can see that the most influential countries by total trade are different from key exporters by particular product groups. LRIC indices detect Canada, United States,





**Table 5** LRIC centralities: top-10 countries by different product categories in 2010

Total Trade				Cereals			
LRIC MaxProd	LRIC MaxMin	LRIC SumProd	LRIC PPR	LRIC MaxProd	LRIC MaxMin	LRIC SumProd	LRIC PPR
MEX	MEX	BRA	CAN	CAN	CAN	CAN	CAN
CAN	CAN	THA	CHN	USA	USA	USA	USA
NLD	NLD	ARG	USA	FRA	FRA	FRA	FRA
USA	DEU	CAN	MEX	DEU	DEU	DEU	DEU
CHN	CHN	MEX	NLD	AUS	ITA	ITA	AUS
BRA	BEL	NZL	DEU	ITA	HUN	THA	THA
BEL	USA	NLD	BRA	THA	BEL	BEL	ITA
DEU	BRA	BEL	THA	BEL	AUS	NLD	ARG
FRA	FRA	AUS	FRA	NLD	NLD	HUN	BEL
THA	ESP	ESP	ESP	HUN	THA	GBR	NLD
Vegetables/Fruits				Meat			
LRIC MaxProd	LRIC MaxMin	LRIC SumProd	LRIC PPR	LRIC MaxProd	LRIC MaxMin	LRIC SumProd	LRIC PPR
MEX	MEX	CHL	CHN	CAN	CAN	NZL	CAN
CHN	CHN	THA	MEX	BRA	BRA	BRA	AUS
CAN	CAN	MEX	USA	NLD	NLD	NLD	DNK
CHL	CHL	CHN	ESP	AUS	DEU	CAN	USA
USA	THA	CAN	CAN	DEU	AUS	DEU	BRA
ESP	ESP	ESP	CHL	DNK	BEL	DNK	DEU
THA	USA	CRI	NLD	USA	DNK	BEL	NLD
NLD	NLD	NLD	ITA	BEL	NZL	AUS	NZL
CRI	CRI	ARG	THA	NZL	USA	ARG	FRA
BEL	BEL	USA	PHL	FRA	FRA	ESP	BEL

France, Germany, and Italy as the most influential traders of cereals products. Mexico, Chile, China, Thailand, Canada, and United States are key exporters of vegetables and fruits. As to meat products, Canada, New Zealand, Brazil, the Netherlands, and Australia are considered as the most important exporters.

Additionally, if we compare these rankings with food balance sheet indicators we can notice that we obtain new information. Hence, the results that take into account network structure should be also considered and analyzed as they discover new knowledge in trade processes.

## 5 Conclusion

This work tends to analyze the organization of products export/import processes in the world. Most of the well-known indicators of countries representation in terms of food trade are based on pure statistics. In this research, we apply social network analysis in order to join both statistical indicators and bilateral export/import relations between countries. This approach allows to understand how countries influence each other through trading relations between them. We calculated classical centrality measures for the network of total trade and for networks that represent export/import relations with respect to different product categories. The main disadvantage of classical centrality measures is that they do not consider individual attributes of nodes which might be essential. Hence, we also apply measures that consider individual parameters of nodes as well as the possibility of the group influence and long-range connections. Long-range Interaction Centrality (LRIC) indices detect the most influential elements in a network considering all these mentioned factors.

Basically, LRIC indices propose to analyze simple paths between elements in order to estimate the indirect influence of nodes on each other. In this work, we also apply random walks approach based on personalized PageRank calculation with the aim to assess indirect influences. The main advantage of random walks is lower computational complexity compared to simple paths enumeration. Moreover, the results obtained with random walks correlate at the very high level with the results obtained by simple paths enumeration, which means that these methods are replaceable.

We compared the results by both classical centrality measures with different versions of LRIC indices for the networks of total trade, cereals, vegetables/fruits, and meat. The most influential countries distinguish for different product categories. Moreover, LRIC indices detect new hidden leaders compared to classical measures, because a lot of important factors are considered.

**Acknowledgements** The article was prepared within the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) and supported within the framework of a subsidy by the Russian Academic Excellence Project '5-100'. The work related to the analysis of global retail food network (Sect. 4) was prepared within the framework of the Russian Science Foundation under grant No 17-18-01651.

## References

1. Food and Agriculture Organization of the United Nations.: Home <http://www.fao.org/home/en/> (2018). Accessed 29 Oct 2018
2. Who.int.: Home <http://www.who.int/> (2018). Accessed 29 Oct 2018
3. Fao.org.: FAOSTAT <http://www.fao.org/faostat/en/#home> (2018) Accessed 29 Oct 2018
4. Newman, M.E.J.: Networks: An Introduction. Oxford University Press, In Oxford, UK (2010)
5. Bonacich, P., Lloyd, P.: Eigenvector-like measures of centrality for asymmetric relations. Soc. Netw. **23**, 191–201 (2001)

6. Katz, L.: A new status index derived from sociometric analysis. *Psychometrika* **19**, 39–43 (1953)
7. Jackson M.: Representing and measuring networks. In: *Social and Economic Networks*, pp. 20–53 (2008)
8. Shimbel, A.: Structural parameters of communication networks. *Bull. Math. Biol.* **15**, 501–507 (1953)
9. Piraveenan, M., Prokopenko, M., Hossain, L.: Percolation centrality: quantifying graph-theoretic impact of nodes during percolation in networks. *PLOS ONE* **8**(1), e53095. <https://doi.org/10.1371/journal.pone.0053095>
10. Aleskerov, F.T., Meshcheryakova, N.G., Shvydun, S.: Power in network structures. In: Kalyagin, V.A., Nikolaev, A.I., Pardalos, P.M., Prokopyev, O. (eds.) *Models, Algorithms, and Technologies for Network Analysis*. Springer Proceedings in Mathematics & Statistics, vol. 197, pp. 79–85 (2017)
11. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: bringing order to the web. In: Technical report, Stanford Digital Library Technologies Project (1998)
12. Wits.worldbank.org.: WITS - About WITS [https://wits.worldbank.org/about\\_wits.html](https://wits.worldbank.org/about_wits.html) (2018). Accessed 19 Apr 2018
13. Unstats.un.org.: What is UN Comtrade? (Introduction, UN Comtrade, Web browser) <https://unstats.un.org/unsd/tradekb/Knowledgebase/50075/What-is-UN-Comtrade> (2018). Accessed 19 Apr 2018
14. Wits.worldbank.org.: Understanding Nomenclatures [https://wits.worldbank.org/WITS/WITS/WITSHELP/Content/Basics/A5.Product\\_Nomenclatures.htm](https://wits.worldbank.org/WITS/WITS/WITSHELP/Content/Basics/A5.Product_Nomenclatures.htm) (2018). Accessed 19 Apr 2018
15. Standard International Trade Classification Revision 4, Department of Economic and Social Affairs, ST/ESA/STAT/SER.M/34/REV.4



# Facial Clustering in Video Data Using Deep Convolutional Neural Networks



Anastasiia D. Sokolova and Andrey V. Savchenko

**Abstract** This paper presents an automatic system that structures information in video surveillance systems based on the analysis of facial images. We describe the cluster analysis in video data using face detection in each video frame and feature extraction with pretrained deep convolutional neural networks. Different aggregation techniques to combine frame features into a single video descriptor are implemented to organize video data based on clustering techniques. An experimental study with the YouTube Faces dataset that demonstrates the most accurate algorithm matches normalized average frame feature vectors and group them with average linkage agglomerative clustering algorithm.

## 1 Introduction

Nowadays, automatic systems are attracted more attention due to their efficiency in various areas [1]. People need to analyze and group data, make statistics, make a precision based on previous detected information. Some applications, such as Aurora, provide the opportunity to analyze customers, know how long a person look at the certain good and recognize facial attributes (age, gender, race, emotions, etc.) [2]. For the field of public safety data organization is significant in order to save information about each visitor [3]. However, the surveillance systems are characterized by a huge amount of video data because hundred frames are processed in a few seconds [4–6].

Moreover, person recognition and feature extraction from the image became more popular. The system that can detect people and then cluster them helps to find information about a certain person. The only rule of such system is to verify that two video

---

A. D. Sokolova (✉) · A. V. Savchenko  
Laboratory of Algorithms and Technologies for Network Analysis,  
National Research University Higher School of Economics,  
Nizhny Novgorod, Russian Federation  
e-mail: [adsokolova96@mail.ru](mailto:adsokolova96@mail.ru)

A. V. Savchenko  
e-mail: [avsavchenko@hse.ru](mailto:avsavchenko@hse.ru)

© Springer Nature Switzerland AG 2020  
I. Bychkov et al. (eds.), *Network Algorithms, Data Mining,  
and Applications*, Springer Proceedings in Mathematics & Statistics 315,  
[https://doi.org/10.1007/978-3-030-37157-9\\_11](https://doi.org/10.1007/978-3-030-37157-9_11)

tracks contain the images of the same person, it is the task of known face verification methods [6, 7]. Hence, the goal of this paper is to make an analysis of verification algorithms and cluster methods.

The rest of the paper is organized in the following way: in Sect. 2 we formulate the proposed approach and give a mathematical description of used dissimilarity measures between video tracks. In Sect. 3, experimental results for the YouTube Faces (YTF) dataset [8] are presented. In Sect. 4 we give concluding comments and future plans.

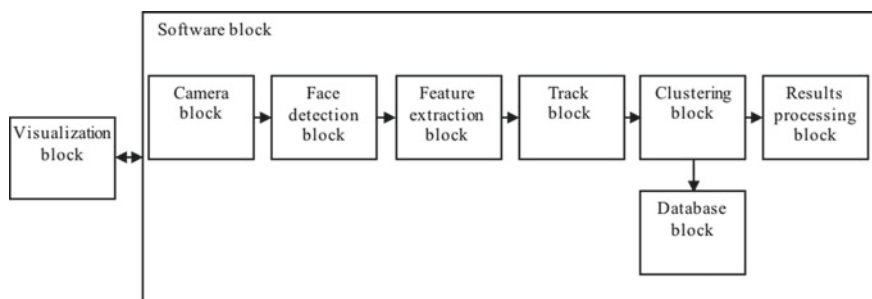
## 2 Video Data Analysis

The task of work is to process video sequences which contain  $T > 1$  frames, divide it in tracks with observations of one person and identify all tracks with the same person. The architecture of the developed system is shown in Fig. 1.

Here, firstly, the faces are detected using one of the pretrained TensorFlow models. This open-source repository contains different preliminary trained models of neural networks and provides an interface for object detection (TensorFlow Object Detection API). The face detection is based on MobileNet SSD (Single-Shot Detector) trained on the WiderFace dataset [9].

The following step is to implement a tracker algorithm that divides the input data into  $M < T$  disjoint tracks  $\{X(m)\}$ ,  $m = 1, 2, \dots, M$ . The track is characterized by the first frame index  $t_1(m)$  and the last frame index  $t_2(m)$  that the duration of track is formulated as follows:  $\Delta t(m) = t_2(m) - t_1(m) + 1$ . Then a clustering algorithm [10–12] is used to group similar tracks. In order to implement the clustering, it is necessary to extract facial features from every frame, aggregate the features of an individual frame to a single descriptor for the whole track and then match these descriptors.

The most popular and efficient way to extract features is to use convolutional neural networks (CNNs) [13, 14]. There are a lot of CNNs trained with external large



**Fig. 1** The architecture of the proposed system

dataset, e.g., Casia WebFaces, MSCeleb-1M of VGGFace datasets. The CNN output for  $t$ th facial image is stored in  $D$ -dimensional feature vector  $\mathbf{x}(t)$ . These features are usually matched with the Euclidean ( $L_2$ ) metric  $\rho(X(m_1), X(m_2))$  [15]. However, when the video sequences are grouped, it is required to compute the distance  $\rho(X(m_1), X(m_2))$  between tracks  $X(m_1)$  and  $X(m_2)$ . In this paper, we examine several different approaches that were presented to compute the distance between tracks [16]:

1. The distance

$$\rho(X(m_1), X(m_2)) = \rho(\mathbf{x}^*(m_1), \mathbf{x}^*(m_2)), \quad (1)$$

between medoids of video tracks

$$\mathbf{x}^*(m_i) = \underset{\mathbf{x}(t), t \in [t_1(m_i), t_2(m_i)]}{\operatorname{argmin}} \sum_{t'=t_1(m_i)}^{t_2(m_i)} \rho(\mathbf{x}(t), \mathbf{x}(t')) \quad (2)$$

2. Average features vectors of each track are matched:

$$\rho(X(m_1), X(m_2)) = \rho(\bar{\mathbf{x}}(m_1), \bar{\mathbf{x}}(m_2)), \bar{\mathbf{x}}(m_i) = \frac{1}{\Delta t(m_i)} \sum_{t'=t_1(m_i)}^{t_2(m_i)} \mathbf{x}(t) \quad (3)$$

3. Comparison of the median features  $\mathbf{x}'(m_i)$  of each track [17]:

$$\rho(X(m_1), X(m_2)) = \rho(\mathbf{x}'(m_1), \mathbf{x}'(m_2)). \quad (4)$$

In order to organize video data several techniques are used:

1. Sequential clustering [18] arbitrarily chooses the first cluster center among all the video tracks and processes the remaining tracks sequentially. The distance between the actual track and its nearest cluster center is computed. If it is not greater than a certain threshold, the track is assigned to its nearest cluster. Otherwise, a new cluster is formed containing one element, namely, an analyzed video track.
2. RankOrder hierarchical clustering [19] is a subtype of hierarchical clustering. It argues that two videos with the same person have many common images (“neighbors”), but the “neighbors” from the videos of different people are very different.
3. DBSCAN [20] is a density-based clustering nonparametric algorithm: given a set of points in some space, it groups together points that are closely packed together (points with many nearby neighbors), marking as emissions points that lie alone in low-density regions (whose nearest neighbors are too far away).
4. Birch [21] is a hierarchical clustering method over particularly large datasets. An advantage of BIRCH is its ability to incrementally and dynamically cluster incoming, multidimensional metric data points in an attempt to produce the best-quality clustering for a given set of resources.

5. MiniBatchKMeans [22] uses mini-batch optimization for k-means clustering, which reduces computation cost by orders of magnitude compared to the classic batch algorithm while yielding significantly better solutions than online stochastic gradient descent.
6. Agglomerative clustering [23] is a hierarchical clustering that uses a bottom-up approach: each observation starts in its own cluster, and clusters are successively merged together. There are different linkage criteria for merge strategy, we will use the contemporary average linkage which is usually characterized by the best quality.

The resulted clusters are displayed to a user (Fig. 1). We prepared a demo application for Android [12], which extracts facial clusters from a gallery of a mobile device. It sequentially processes all photos from the gallery in the background thread. However, the intermediate results are available, so it is not necessary to wait for a long time. The demography category contains the stacked histograms (Fig. 2a) of age and gender of closed persons, which appears in at least 3 photos from the gallery. As usual, tapping the bar will display the list of all photos of a particular subject (Fig. 2b). If there is more than one subject with identical gender and age range, the display form contains a spinner on top, which can be used to choose a particular person by a sequential number (Fig. 2c).

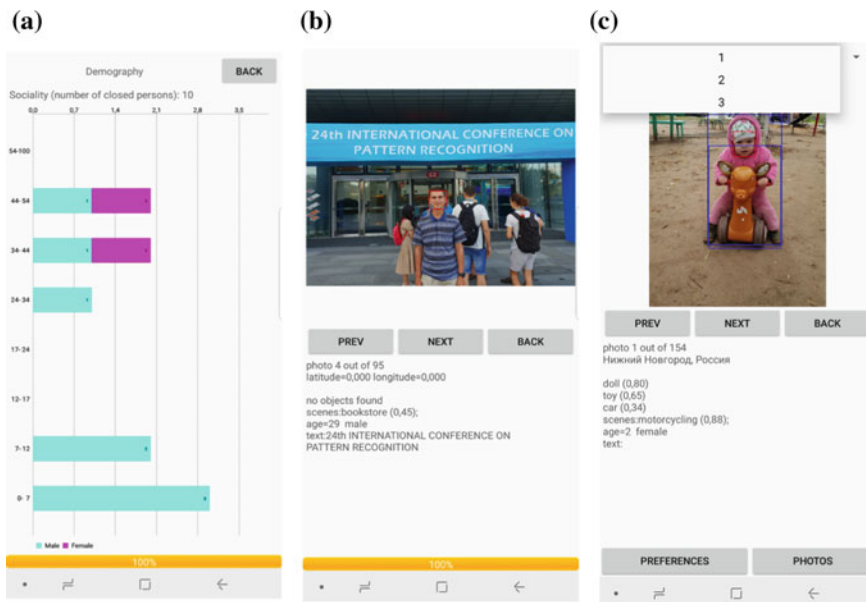
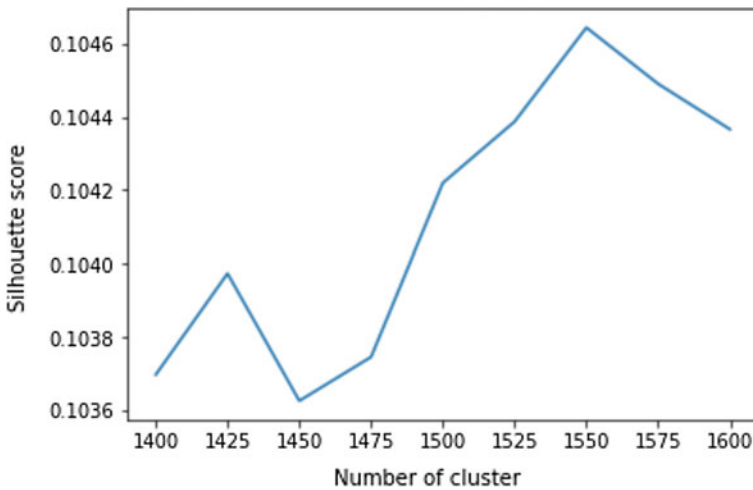


Fig. 2 Facial cluster analysis in the mobile demo

### 3 Experimental Results

The experimental study was conducted with the YTF dataset [8], which contains 3425 videos of 1595 different people. An average of 2.15 videos is available for each subject. The shortest track duration is 48 frames, the longest track contains 6070 frames, and the average length of a video clip is 181.3 frames. We implemented a special software prototype using PyCharm from JetBrains (Python 3.6 language), OpenCV (image and video processing), Caffe (feature extraction) [13], TensorFlow (face detection) and scikit-learn (clustering) libraries. In order to extract features we used three publicly available CNNs suitable for face recognition, namely, the VGGNet [24], Lightened CNN version C (LCNN) [14], VGGFace2 [25]. The VGGNet extracts  $D = 4096$  features vector in the output of “fc7” layer from  $224 \times 224$  RGB images. The Lightened CNN extracts  $D = 256$  features vector (“elt-wise fc2” layer) is computed from  $128 \times 128$  grayscale image. The VGGFace2 is the ResNet50 model, which extracts  $D = 2048$  features vector from “pool5\_7  $\times$  7\_s1”. Their advantages are high velocity of image processing and high accuracy of detection. To make features more robust to observation conditions it is worth to use normalization [26]. In this paper, we use conventional  $L_2$ -normed features.

Various aggregation techniques (1)–(4) and clustering algorithms described in the previous section were studied. The sequential clustering threshold for resulting clusters is set by fixing the FAR (False Acceptance Rate) value to 1%. Some algorithms such as Birch and MiniBatchKMeans require the number of clusters, but for video surveillance systems it is difficult to define a certain number. Therefore, we computed the Silhouette score (Fig. 3) that measures how similar an object is to its own cluster compared to other clusters [27].



**Fig. 3** Silhouette score for agglomerative clustering with average linkage, median features (4) extracted by VGGNet

**Table 1** Results of clustering, Lightened CNN (YTF dataset)

Algorithm	CNN	$D$	$K$	ARI	AMI	$H$	$C$	$V$	FM
Sequential clustering	Medoid (1)	256	2487	0.439	0.449	0.965	0.967	0.966	0.598
		125	2473	0.444	0.454	0.906	0.911	0.909	0.572
	AvePool (2)	256	2492	0.527	0.530	0.972	0.971	0.971	0.663
		140	2467	0.501	0.485	0.964	0.962	0.963	0.592
	Median (3)	256	2503	0.513	0.524	0.968	0.970	0.969	0.647
		135	2478	0.443	0.445	0.966	0.964	0.965	0.593
RankOrder	Medoid (1)	256	2107	0.655	0.677	0.981	0.985	0.983	0.692
		125	2101	0.658	0.673	0.980	0.980	0.980	0.695
	AvePool (2)	256	2105	0.593	0.627	0.972	0.973	0.973	0.674
		140	2101	0.602	0.630	0.981	0.980	0.981	0.679
	Median (3)	256	2049	0.591	0.624	0.968	0.975	0.972	0.670
		135	2051	0.603	0.629	0.980	0.980	0.980	0.681
DBSCAN	Medoid (1)	256	2348	0.548	0.480	0.939	0.986	0.962	0.558
		125	2877	0.368	0.250	0.916	0.999	0.956	0.470
	AvePool (2)	256	2682	0.326	0.321	0.923	0.989	0.955	0.328
		140	2974	0.293	0.194	0.910	0.999	0.952	0.405
	Median (3)	256	2348	0.557	0.496	0.933	0.974	0.967	0.639
		135	2598	0.555	0.407	0.932	0.999	0.964	0.614
Birch	Medoid (1)	256	1400	0.546	0.554	0.952	0.940	0.946	0.547
		125	1400	0.504	0.519	0.950	0.932	0.941	0.508
	AvePool (2)	256	1405	0.440	0.503	0.955	0.922	0.938	0.461
		140	1400	0.381	0.463	0.952	0.912	0.931	0.409
	Median (3)	256	1400	0.608	0.618	0.964	0.946	0.955	0.612
		135	1400	0.558	0.577	0.961	0.938	0.950	0.566
MiniBatchKMeans	Medoid (1)	256	1487	0.029	0.179	0.913	0.792	0.848	0.073
		125	1486	0.060	0.220	0.919	0.826	0.870	0.11
	AvePool (2)	256	1441	0.055	0.263	0.932	0.830	0.878	0.117
		140	1552	0.053	0.278	0.930	0.851	0.889	0.112
	Median (3)	256	1511	0.065	0.284	0.932	0.850	0.889	0.127
		135	1486	0.106	0.329	0.937	0.872	0.903	0.169
Average linkage	Medoid (1)	256	1550	0.585	0.608	0.953	0.952	0.952	0.585
		125	1475	0.562	0.576	0.952	0.945	0.949	0.563
	AvePool (2)	256	1550	0.673	0.695	0.965	0.961	0.963	0.674
		140	1605	0.686	0.713	0.965	0.965	0.965	0.686
	Median (3)	256	1575	0.676	0.702	0.965	0.963	0.964	0.676
		135	1600	0.680	0.708	0.964	0.964	0.964	0.680

Moreover, we associate each input object with a sequence of principal component scores of features extracted by deep neural network [28]. The number of components in each element of this sequence is chosen using 0.9% explained proportion of total variance for the training set.

We have calculated the following metrics of cluster labeling: adjusted rand index (ARI), adjusted mutual information (AMI) score, homogeneity score  $H$ , complete-

**Table 2** Results of clustering, VGGNet (YTF dataset)

Algorithm	CNN	$D$	$K$	ARI	AMI	$H$	$C$	$V$	FM
Sequential clustering	Medoid (1)	4096	2821	0.326	0.325	0.913	0.941	0.927	0.478
		750	2820	0.316	0.409	0.912	0.921	0.917	0.471
	AvePool (2)	4096	2264	0.553	0.560	0.974	0.977	0.976	0.672
		690	2108	0.508	0.492	0.966	0.965	0.966	0.590
	Median (3)	4096	2582	0.462	0.470	0.937	0.948	0.943	0.471
		740	2647	0.465	0.448	0.924	0.935	0.930	0.490
RankOrder	Medoid (1)	4096	2200	0.465	0.525	0.935	0.938	0.937	0.563
		750	2176	0.487	0.530	0.930	0.930	0.930	0.561
	AvePool (2)	4096	2213	0.590	0.604	0.953	0.952	0.953	0.557
		690	2224	0.601	0.605	0.958	0.958	0.958	0.560
	Median (3)	4096	2015	0.542	0.520	0.949	0.935	0.942	0.496
		740	2063	0.544	0.528	0.944	0.935	0.940	0.546
DBSCAN	Medoid (1)	4096	2326	0.113	0.129	0.931	0.914	0.923	0.437
		750	2535	0.115	0.212	0.931	0.920	0.926	0.264
	AvePool (2)	4096	2997	0.235	0.166	0.907	0.995	0.949	0.300
		690	2533	0.109	0.331	0.922	0.962	0.942	0.143
	Median (3)	4096	2583	0.216	0.166	0.899	0.992	0.946	0.205
		740	2611	0.217	0.343	0.910	0.962	0.936	0.196
Birch	Medoid (1)	4096	1400	0.327	0.378	0.935	0.903	0.918	0.339
		750	1400	0.290	0.352	0.933	0.893	0.913	0.310
	AvePool (2)	4096	1400	0.216	0.341	0.938	0.876	0.906	0.261
		690	1535	0.184	0.333	0.934	0.879	0.906	0.231
	Median (3)	4096	1400	0.377	0.436	0.945	0.910	0.927	0.395
		740	1550	0.377	0.450	0.943	0.918	0.931	0.391
MiniBatchKMeans	Medoid (1)	4096	1541	0.035	0.146	0.902	0.793	0.844	0.071
		750	1529	0.039	0.169	0.907	0.812	0.857	0.077
	AvePool (2)	4096	1506	0.055	0.196	0.914	0.814	0.861	0.100
		690	1538	0.111	0.269	0.923	0.865	0.893	0.154
	Median (3)	4096	1477	0.040	0.178	0.912	0.796	0.850	0.840
		740	1532	0.081	0.241	0.920	0.847	0.882	0.129
Average linkage	Medoid (1)	4096	1575	0.386	0.459	0.937	0.929	0.933	0.392
		750	1450	0.439	0.469	0.941	0.928	0.934	0.441
	AvePool (2)	4096	1600	0.486	0.557	0.950	0.942	0.946	0.491
		690	1600	0.564	0.606	0.952	0.952	0.952	0.564
	Median (3)	4096	1575	0.495	0.556	0.950	0.942	0.946	0.499
		740	1600	0.562	0.603	0.951	0.951	0.951	0.562

ness score  $C$ , v-measure score  $V$ , Fowlkes Mallows (FM) score for each algorithm. In addition, we present the resulted number of clusters  $K$  and number of features  $D$ . The results for features from Lightened CNN, VGGFace and VGGFace2 descriptors are presented in Table 1, Table 2 and Table 3, respectively.

Here the most effective results were demonstrated by agglomerative clustering of normalized average features using VGGFace2 descriptor, which has the highest

**Table 3** Results of clustering, VGGFace2 (YTF dataset)

Algorithm	CNN	$D$	$K$	ARI	AMI	$H$	$C$	$V$	FM
Sequential clustering	Medoid (1)	2048	1786	0.492	0.518	0.968	0.971	0.970	0.633
		470	1644	0.495	0.507	0.846	0.923	0.885	0.580
	AvePool (2)	2048	2264	0.553	0.560	0.974	0.977	0.976	0.672
		450	2108	0.508	0.492	0.966	0.965	0.966	0.590
	Median (3)	2048	1849	0.524	0.544	0.962	0.965	0.963	0.679
		480	2245	0.503	0.492	0.962	0.959	0.961	0.661
RankOrder	Medoid (1)	2048	2395	0.428	0.465	0.984	0.985	0.985	0.429
		470	2152	0.453	0.474	0.981	0.982	0.982	0.454
	AvePool (2)	2048	2005	0.594	0.628	0.985	0.882	0.883	0.702
		450	2007	0.593	0.629	0.982	0.877	0.915	0.700
	Median (3)	2048	2001	0.592	0.625	0.972	0.964	0.968	0.681
		480	2015	0.608	0.630	0.982	0.984	0.983	0.675
DBSCAN	Medoid (1)	2048	1978	0.425	0.333	0.945	0.914	0.930	0.291
		470	1854	0.351	0.315	0.932	0.822	0.877	0.305
	AvePool (2)	2048	2267	0.330	0.327	0.935	0.948	0.942	0.333
		450	1649	0.327	0.324	0.949	0.952	0.950	0.329
	Median (3)	2048	2156	0.327	0.327	0.936	0.949	0.943	0.360
		480	1789	0.316	0.329	0.937	0.950	0.944	0.348
Birch	Medoid (1)	2048	1532	0.340	0.407	0.951	0.812	0.882	0.280
		470	1323	0.359	0.406	0.945	0.771	0.858	0.356
	AvePool (2)	2048	1595	0.421	0.453	0.944	0.936	0.940	0.345
		450	1590	0.395	0.412	0.937	0.935	0.936	0.351
	Median (3)	2048	1400	0.433	0.514	0.915	0.927	0.921	0.398
		480	1400	0.386	0.498	0.920	0.934	0.927	0.316
MiniBatchKMeans	Medoid (1)	2048	1416	0.018	0.157	0.937	0.798	0.862	0.161
		470	1502	0.034	0.222	0.93	0.795	0.860	0.144
	AvePool (2)	2048	1407	0.003	0.197	0.932	0.932	0.932	0.118
		450	1518	0.019	0.201	0.948	0.940	0.944	0.128
	Median (3)	2048	1435	0.018	0.145	0.928	0.909	0.918	0.135
		480	1525	0.021	0.159	0.905	0.903	0.904	0.098
Average linkage	Medoid (1)	2048	1425	0.563	0.579	0.947	0.954	0.950	0.563
		470	1525	0.570	0.579	0.943	0.941	0.942	0.571
	AvePool (2)	2048	1595	0.682	0.695	0.958	0.968	0.966	0.701
		450	1575	0.684	0.715	0.961	0.960	0.960	0.694
	Median (3)	2048	1550	0.683	0.690	0.964	0.972	0.969	0.693
		480	1450	0.667	0.673	0.963	0.961	0.962	0.688

score of all clustering metrics. However, DBSCAN and RankOrder algorithms can also be used, their scores are only 2–7% less through all metrics. The Lightened CNN demonstrated better results than the VGGNet. The latter CNN has also much higher runtime and space complexity. Association of each video track with a sequence of principal component scores reduces the features number and computational complexity of tracks matching, but may cause 1–3% lower clustering quality.



## 4 Conclusion

In this paper, we thoroughly examined cluster methods for video tracks described by aggregated features of individual frames extracted using pretrained CNN. An automatic system of grouping video data was built. We focus on efficient computation of dissimilarity of video tracks and comparative analysis of clustering algorithms. An experimental study demonstrated that matching normalized average features and using agglomerative clustering is the most accurate and fast approach.

The main direction for further research is to examine a more sophisticated clustering algorithm in order to reduce the error of clustering data that one cluster contains the images of only one person and there are no different clusters for this person.

**Acknowledgements** The article was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE University) in 2019 (grant No. 19-04-004) and within the framework of the Russian Academic Excellence Project “5–100”.

## References

1. Sokolova, A.D., Kharchevnikova, A.S., Savchenko, A.V.: Organizing multimedia data in video surveillance systems based on face verification with convolutional neural networks. In: Proceedings of the International Conference on Analysis of Images, Social Networks and Texts (AIST), Springer, Cham, pp. 223–230 (2017)
2. Aurora <https://retailnext.net/en/aurora>
3. Zhang, Y.J., Lu, H.B.: A hierarchical organization scheme for video data. *Pattern Recognit.* **35**(11), 2381–2387 (2002)
4. Shan, C.: Face recognition and retrieval in video. In: Video Search and Mining. Springer, Berlin, Heidelberg, pp. 235–260 (2010)
5. Savchenko, A.V.: Search Techniques in Intelligent Classification Systems. Springer International Publishing, Heidelberg (2016)
6. Chen, J.C., Ranjan, R., Kumar, A., Chen, C.H., Patel, V.M., Chellappa, R.: An end-to-end system for unconstrained face verification with deep convolutional neural networks. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 118–126 (2015)
7. Li, H., Hua, G., Shen, X., Lin, Z., Brandt, J.: Eigen-PEP for video face recognition. In: Cremers, D., Reid, I., Saito, H., Yang, M.H. (eds.), Proceedings of the Asian Conference on Computer Vision, Springer, Cham, vol. 9005, pp. 17–33 (2014)
8. Wolf, L., Hassner, T., Maoz, I.: Face recognition in unconstrained videos with matched background similarity. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 529–534 (2011)
9. Yang, S., Luo, P., Loy, C.C., Tang, X.: Wider face: a face detection benchmark. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5525–5533 (2016)
10. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley (2009)
11. Savchenko, A.V.: Clustering and maximum likelihood search for efficient statistical classification with medium-sized databases. *Optim. Lett.* **11**(2), 329–341 (2017)
12. Savchenko, A.V.: Efficient facial representations for age, gender and identity recognition in organizing photo albums using multi-output CNN. arXiv preprint [arXiv:1807.07718](https://arxiv.org/abs/1807.07718) (2018)

13. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 675–678 (2014)
14. Wu, X., He, R., Sun, Z., Tan, T.: A light CNN for deep face representation with noisy labels. *Proc. IEEE Trans. Inf. Forensics Secur.* **13**(11), 2884–2896 (2018)
15. Goodfellow, I., Bengio, Y., Courville, A.: *Deep Learning*. MIT press (2016)
16. Sokolova, A.D., Savchenko, A.V.: Cluster analysis of facial video data in video surveillance systems using deep learning. In: Proceedings of the Computational Aspects and Applications in Large-Scale Networks (NET). Springer Proceedings in Mathematics and Statistics, vol. 247, pp. 113–120 (2018)
17. Rassadin, A., Gruzdev, A., Savchenko, A.V.: Group-level emotion recognition using transfer learning from face identification. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction, pp. 544–548 (2017)
18. Miyamoto, S., Arai, K.: Different sequential clustering algorithms and sequential regression models. In: Proceedings of the IEEE International Conference on Fuzzy Systems, pp. 1107–1112 (2009)
19. Zhu, C., Wen, F., Sun, J.: A rank-order distance based clustering algorithm for face tagging. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 481–488 (2011)
20. Schubert, E., Sander, J., Ester, M., Kriegel, H.P., Xu, X.: DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Trans. Database Syst. (TODS)*, **42**(3), 19 (2017)
21. Zhang, T., Ramakrishnan, R., Livny, M.: BIRCH: an efficient data clustering method for very large databases. *ACM Sigmod Rec. (ACM)* **25–2**, 103–114 (1996)
22. Sculley, D.: Web-scale k-means clustering. In: Proceedings of the 19th International Conference on World Wide Web (ACM), pp. 1177–1178 (2010)
23. Rokach, L., Maimon, O.: Clustering methods. In: *Data Mining and Knowledge Discovery Handbook*. Springer, Boston, MA, pp. 321–352 (2005)
24. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: Proceedings of the British Machine Vision Conference (BMVC), vol. 1, no. 3, pp. 6–17 (2015)
25. Cao, Q., Shen, L., Xie, W., Parkhi, O. M., Zisserman, A.: VGGFace2: A dataset for recognising faces across pose and age. In: Proceedings of the 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG), pp. 67–74 (2018)
26. Savchenko, A.V., Belova, N.S.: Unconstrained face identification using maximum likelihood of distances between deep off-the-shelf features. *Expert Syst. Appl.* **108**, 170–182 (2018)
27. da Silva, S.F., Brandoli, B., Eler, D.M., Neto, J.B., Traina, A.J.: Silhouette-based feature selection for classification of medical images. In: Proceedings of the IEEE 23rd International Symposium on Computer-Based Medical Systems (CBMS), pp. 315–320 (2010)
28. Savchenko, A.V.: Sequential three-way decisions in multi-category image recognition with deep features based on distance factor. *Inf. Sci.* **489**, 18–36 (2019)

# **Network Applications**

# The Existence and Uniqueness Theorem for Initial-Boundary Value Problem of the Same Class of Integro-Differential PDEs



A. I. Egamov

**Abstract** The second initial-boundary value problem for a class of nonlinear PDEs of the second order and an integral operator of a given form is considered. Dependence of its solution with the solution of the standard second linear initial-boundary value problem for a second order hyperbolic equation is shown. The proof of energy inequality presents at the beginning for an auxiliary linear problem and then for a nonlinear problem. The existence and uniqueness theorem of the corresponding initial-boundary value problem is proved with its help. For better understanding of the problems under consideration as particular representatives of the studied class of Integro-differential equations, the examples of Integro-differential equations for various integral operators of this type are given at the final of the article. The author believes that, despite the existing purely theoretical interest, this class of Integro-differential equations will attract great attention for further research. The solution of the nonlinear problem has a curious nontrivial property, and the existence and uniqueness theorem of the original problem will help to justify the application of various applications to her.

**Keywords** The second initial boundary value problem · Integro-differential hyperbolic PDE · Energy inequality · The existence and uniqueness theorem

## 1 Introduction

The integro-differential equations play an increasing role in modern mathematics. The History of their solution methods can be studied for example in [1, 2]. Their solution methods are generally nontrivial and, practically, each case is individual [3, 4]. Recently, the method of separating variables, previously considered to be exclusively applicable to linear equations, has been successfully applied to some

---

A. I. Egamov (✉)

Lobachevsky State University of Nizhny Novgorod, 23 Gagarina Avenue,  
Nizhny Novgorod 603950, Russian Federation  
e-mail: [albert810@yandex.ru](mailto:albert810@yandex.ru)

© Springer Nature Switzerland AG 2020

I. Bychkov et al. (eds.), *Network Algorithms, Data Mining, and Applications*, Springer Proceedings in Mathematics & Statistics 315,  
[https://doi.org/10.1007/978-3-030-37157-9\\_12](https://doi.org/10.1007/978-3-030-37157-9_12)

173

types of nonlinear partial differential equations [5]. However, most often researchers try to find a pattern between a nonlinear problem and the corresponding linear one. The reduction of a nonlinear to a linear problem is considered as a significant step towards its solution, since the solution of a linear problem is usually known and standard [6–8]. This approach is used in this article. A theorem on the connection of a solution for a class of Integro-differential equations with linear hyperbolic PDE is proved. At the end of the article for more understanding examples that discuss particular cases. This paper uses a method similar to that developed by O. A. Kuzenkov, which applies him to initial-boundary value problems for the parabolic and the first order hyperbolic equations with the specific integral operator [9–11]. Earlier there were the attempts to apply a similar method to a second order hyperbolic equation with a special integral operator [12, 13]. In this paper the theorem on the existence of the uniqueness of the original nonlinear problem is proved. Its solution is presented in an explicit form through the solution of a linear problem. At beginning for the auxiliary and then for the original problem the energy inequality is proved in the same way as the method presented in [7]. Its proof is necessary to substantiate the uniqueness of the solution of the original problem for the possibility of correct application to it of various numerical methods and optimization methods. The solution of the nonlinear problem has an interesting property. It satisfies the phase constraint, an analogue of the norm invariant.

## 2 Auxiliary Assertions

It is known [6–8] that for an arbitrary  $T > 0$  in the domain  $\Omega_T = [0; l] \times [0; T]$  there exists a function  $z(x, t)$ , twice continuously differentiable by its variables is the solution of a hyperbolic equation of the 2nd order

$$z''_{tt}(x, t) = a^2 z''_{xx}(x, t) + b(x)z(x, t) + \sigma(t)z'_t(x, t), \quad (1)$$

with second edge

$$z'_x(0, t) = z'_x(l, t) = 0, \quad (2)$$

and initial conditions

$$z(x, 0) = \varphi(x), \quad z'_t(x, 0) = \psi(x), \quad (3)$$

$b(x)$ ,  $\sigma(t)$  are continuous functions, [6] the functions  $\varphi(x)$ ,  $\psi(x)$  are smooth enough times, satisfying the coupling conditions (2). The solution to this problem is unique and can be obtained, for example, by Fourier method.

Denote  $P[w]$  is an integral operator of the form

$$P[w] = \left( \int_0^l F_1(w(x, t), w'_t(x, t), w'_x(x, t)) dx \right)^\alpha, \tag{4}$$

where  $\alpha$  is the positive constant,  $F_1$  is a function twice continuously differentiable with respect to its arguments, such that an integral exists for any suitable twice continuously differentiable function  $w(x, t)$  is the argument of the operator  $P[w]$ . The set of values of the operator  $P[w]$  is twice continuously differentiable function of the variable  $t$ . We introduce the notation  $P_w(t) \equiv P[w(x, t)]$  (if we mean the derivative, we will put a stroke over the operator, for example,  $P'_t[w]$ ).

In the proofs of statements for simplicity and convenience of writing if the argument of the operator  $P[w]$  is a function  $z(x, t)$  is the solution of the problem (1)–(3), then the corresponding function will be written without an index:  $P(t) \equiv P_z(t)$  or just  $P$ . In particular,  $P(0) = P[z(x, 0)] = P[\varphi(x)]$ .

Denote  $R[w]$  is an integral operator of the form

$$R[w] = \int_0^l F_2(w(x, t), w'_t(x, t), w'_x(x, t)) dx, \tag{5}$$

where  $F_2(w(x, t), w'_t(x, t), w'_x(x, t))$  is a continuous function with respect to its arguments, such that the integral exists for any suitable function  $w(x, t)$  is the argument of the operator  $R[w]$ , in addition, there is a continuous derivative with respect to its arguments  $\frac{\partial F_2}{\partial w}(w(x, t), w'_t(x, t), w'_x(x, t))$ . Similar to the above integral operator we introduce the notation  $R_w(t) \equiv R[w]$ . It is easy to see that  $R_w(t)$  is continuous function.

Let the function  $q(t), t \in [0; T]$ , be the solution of the Cauchy problem for the Riccati equation

$$q'_t(t) - \beta(t)q(t) - q^2(t) = R_w(t), q(0) = 0, \tag{6}$$

where  $\beta(t)$  is a continuous function at  $[0; T]$ . The solution (6) exists on some segment  $t \in [0, T_0]$  for any continuous function  $R_w(t)$ . Depending on the function  $R_w(t)$ , the function  $q(t)$  can be [14] as bounded on the interval  $[0; T]$ , and tend to  $\infty$  at  $t \rightarrow T_0 - 0, T_0 \in [0; T]$ . In the latter case, we can take another  $T < T_0$  and thus guarantee the existence of a continuously differentiable (and hence bounded) function  $q(t)$  on the segment  $t \in [0, T]$ .

Denote by  $\gamma_0 = \|q\|_{C[0, T]} = \max_{t \in [0; T]} |q(t)|$  the norm of the continuous function  $q(t)$  in the space of continuous functions on the segment  $[0; T]$ , and

$$\|b\|_2 = \left( \int_0^l b^2(x) dx \right)^{\frac{1}{2}}$$

is norm of function  $b(x)$  in space  $L_2(0, l)$  [7]. Denote

$$\kappa(w, t) = \|w(x, t)\|_2^2 + \|w'_t(x, t)\|_2^2 + \|w'_x(x, t)\|_2^2,$$

and  $T_M = \{t : \max_{t \in [0, T]} \kappa(w, t)\}$ .

We prove an auxiliary statement:

**Lemma 1** *Let the function  $w(x, t)$  be continuous by  $x$  and continuously differentiated by  $t$ ;  $\eta(x, t)$  is a continuous function with respect to its arguments, denote  $\eta_0 = \max_{(x,t) \in \Omega_T} |\eta(x, t)|$ , then*

$$\left| \int_0^T \int_0^l 2\eta(x, t)w(x, t)w'_t(x, t) dx dt \right| \leq \eta_0 T \kappa(w, T_M).$$

**Proof**

$$\begin{aligned} \left| \int_0^T \int_0^l 2\eta(x, t)w(x, t)w'_t(x, t) dx dt \right| &\leq \int_0^T \int_0^l |\eta(x, t)| \cdot |2w(x, t)w'_t(x, t)| dx dt \leq \\ &\leq \eta_0 \int_0^T \int_0^l (w^2(x, t) + w'^2_t(x, t)) dx dt \leq \eta_0 \kappa(w, T_M) \int_0^T 1 dt = \eta_0 T \kappa(w, T_M). \end{aligned}$$

The proof is complete.

### 3 The Main Result

In this paper we consider a class of integro-differential equations of the form

$$y''_t(x, t) = a^2 y''_{xx}(x, t) + b(x)y(x, t) + \beta(t)y'_t(x, t) - R[y]y(x, t), \tag{7}$$

with boundary and initial conditions

$$y'_x(0, t) = y'_x(l, t) = 0, \tag{8}$$

$$y(x, 0) = \varphi(x), y'_t(x, 0) = \psi(x). \tag{9}$$

Let the function  $q(t)$  be the bounded solution of the Cauchy problem (6) for  $w(x, t) \equiv y(x, t)$ , where  $y(x, t)$ —solution of problem (7)–(9).

**Theorem 1** *If on the segment  $[0; T]$  true equality*

$$\sigma(t) = \beta(t) + 2q(t), \tag{10}$$

and the function

$$p(t) = \exp\left(\int_0^t q(\tau) d\tau\right). \tag{11}$$

Then in the domain  $\Omega_T$  there is a unique solution to the problem (6)–(9): the function  $y(x, t)$  and the equality is true

$$y(x, t) = \frac{z(x, t)}{p(t)}, \tag{12}$$

where  $z(x, t)$  is the solution of the problem (1)–(3).

**Proof 1. Existence.** We consider only such  $T$  for which the function  $q(t)$  is bounded by  $t \in [0; T]$ , so due to (11) the function  $p(t)$  on it satisfies the inequality

$$p(t) \neq 0 \tag{13}$$

and  $p(0) = 1$ .

According to the formula (12), the boundary (8) and initial (9) conditions are satisfied:

$$y'_x(0, t) = \frac{\partial}{\partial x} \left( \frac{z(0, t)}{p(t)} \right) = \frac{z'_x(0, t)}{p(t)} = 0,$$

likewise  $y'_x(l, t) = 0$ . Further

$$y(x, 0) = \frac{z(x, 0)}{p(0)} = \varphi(x), \quad y'_t(x, 0) = \frac{z'_t(x, 0)}{p(0)} - \frac{z(x, 0)P'_t(0)}{p^2(0)} = \psi(x).$$

We show that the function  $y(x, t)$ , satisfy equality (12) given the expression (11), is the solution of the equation (7). Differentiate with respect to the equality (12):

$$y'_t(x, t) = \left( \frac{z(x, t)}{p(t)} \right)'_t = \frac{z'_t}{p} - \frac{z}{p} \frac{p'_t}{p}.$$



Doing next, we obtain that

$$\begin{aligned}
 y''_{tt}(x, t) &= \left( \frac{z'_t}{p} - \frac{z}{p} \frac{p'_t}{p} \right)'_t = \frac{z''_{tt}}{p} - 2 \frac{z'_t}{p} \frac{p'_t}{p} - \frac{z}{p} \left( \frac{p''_{tt}}{p} - 2 \frac{p'^2_{tt}}{p^2} \right) = \\
 &= \frac{a^2 z''_{xx} + b(x)z + \sigma(t)z'_t}{p} - 2 \frac{p'_t}{p} \left( \frac{z'_t}{p} - \frac{z}{p} \frac{p'_t}{p} \right) - \frac{z}{p} \frac{p''_{tt}}{p} = \\
 &= \frac{a^2 z''_{xx} + b(x)z}{p} - 2 \frac{p'_t}{p} \left( \frac{z'_t}{p} - \frac{z}{p} \frac{p'_t}{p} \right) + \sigma(t) \left( \frac{z'_t}{p} - \frac{z}{p} \frac{p'_t}{p} \right) - \frac{z}{p} \left( \frac{p''_{tt}}{p} - \sigma(t) \frac{p'_t}{p} \right) = \\
 &= a^2 y''_{xx}(x, t) + b(x)y(x, t) + (\sigma(t) - 2 \frac{p'_t}{p})y'_t(x, t) - \left( \frac{p''_{tt}}{p} - \sigma(t) \frac{p'_t}{p} \right) y(x, t).
 \end{aligned}$$

Hence the need for equality

$$\sigma(t) - 2 \frac{p'_t}{p} = \beta(t).$$

Therefore, given (10),  $q(t)$  is expressed as

$$q(t) = \frac{p'_t(t)}{p(t)}. \quad (14)$$

Note that then  $p_t(0) = q(0) = 0$  and by solving the differential equation (14) with initial condition  $p(0) = 1$ , we obtain the formula (11). From (6), (14) it follows that

$$\begin{aligned}
 R_y(t) &= q'_t - q^2(t) - \beta(t)q(t) = \frac{p''_{tt}}{p} - \frac{p'^2_{tt}}{p^2} - q^2(t) - \beta(t)q(t) = \\
 &\frac{p''_{tt}}{p} - (\beta(t) + 2q(t))q(t) = \frac{p''_{tt}}{p} - \sigma(t)q(t).
 \end{aligned}$$

Therefore, in the implementation of equality

$$R_y(t) = \frac{p''_{tt}}{p} - \sigma(t) \frac{p'_t}{p}, \quad (15)$$

the function  $y(x, t)$  satisfying equality (12) is the solution of the equation (7).

We clarify that the proof implies that if one of the three functions  $p(t)$ ,  $q(t)$  and  $R_y(t)$  is known, then the other two are uniquely determined. Existence is proven.

**2. Energy inequality for a linear problem.** We derive an energy inequality for the problem (1)–(3). Let  $b_0 = \|b\|_{C[0, l]}$ ,  $\sigma_0 = \|\sigma(t)\|_{C[0, T]}$ ,  $\beta_0 = \|\beta(t)\|_{C[0, T]}$ .

In this (second) part of the proof  $w = z(x, t)$ ,

$$\kappa(z, t) = \|z(x, t)\|_2^2 + \|z'_t(x, t)\|_2^2 + \|z'_x(x, t)\|_2^2, \tag{16}$$

$$T_M = \{t : \max_{t \in [0, T]} \kappa(z, t)\}.$$

Acting similarly to [7], multiply both parts (1) by  $2z'_t(x, t)$  and integrate as shown below.

$$\begin{aligned} & \int_0^l \int_0^T (2z'_t(x, t)z''_{tt}(x, t)) dt dx = \\ & = \int_0^l \int_0^T (2z'_t(x, t)a^2z''_{xx}(x, t) + 2b(x)z(x, t)z'_t(x, t) + 2\sigma(t)z_t'^2(x, t)) dt dx. \end{aligned} \tag{17}$$

For sufficiently smooth initial conditions there are mixed continuous  $z''_{tx}(x, t)$  and  $z''_{xt}(x, t)$  and they are equal, therefore, integrating piecemeal taking into account (2), we have

$$\begin{aligned} & \int_0^T \int_0^l 2z'_t(x, t)z''_{xx}(x, t) dt dx = 2 \int_0^T z'_t(x, t)z'_x(x, t) \Big|_0^l dt - \\ & - \int_0^l \int_0^T 2z'_x(x, t)z''_{xt}(x, t) dx dt = - \int_0^l (z_x'^2(x, T) - z_x'^2(x, 0)) dx. \end{aligned}$$

In addition,

$$\int_0^l \int_0^T 2z'_t(x, t)z''_{tt}(x, t) dt dx = \int_0^l z_t'^2(x, T) dx - \int_0^l z_t'^2(x, 0) dx.$$

And, further,

$$\begin{aligned} & \int_0^l \int_0^T 2b(x)z(x, t)z'_t(x, t) dt dx = \int_0^l b(x)(z^2(x, T) - z^2(x, 0)) dx \leq \\ & \leq b_0(\|z(x, 0)\|_2^2 + \|z(x, T)\|_2^2). \end{aligned}$$

Last term:

$$\begin{aligned} \int_0^l \int_0^T 2\sigma(t)z_i'^2(x, t) dt dx &= \int_0^T 2\sigma(t) \int_0^l z_i'^2(x, t) dx dt \leq 2\sigma_0 \int_0^T \kappa(z, t) dt \leq \\ &\leq 2\sigma_0 \kappa(z, T_M) \int_0^T 1 dt \leq 2T\sigma_0 \kappa(z, T_M). \end{aligned}$$

Therefore, by converting (17), we obtain an estimate

$$\begin{aligned} \|z_i'(x, T)\|_2^2 + a^2 \|z_x'(x, T)\|_2^2 &\leq \|z_i'(x, 0)\|_2^2 + a^2 \|z_x'(x, 0)\|_2^2 + b_0 \|z(x, 0)\|_2^2 + \\ &+ b_0 \|z(x, T)\|_2^2 + 2T\sigma_0 \kappa(z, T_M). \end{aligned} \quad (18)$$

Estimate the square of the norm

$$\begin{aligned} \|z(x, T)\|_2^2 &= \int_0^l z^2(x, T) dx = \int_0^l \left( z(x, 0) + \int_0^T z_i'(x, t) dt \right)^2 dx \leq 2 \int_0^l z^2(x, 0) dx + \\ &+ 2 \int_0^l \left( \int_0^T z_i'(x, t) dt \right)^2 dx \leq 2 \|z(x, 0)\|_2^2 + 2 \int_0^T \int_0^l z_i'^2(x, t) dx dt \leq \\ &\leq 2 \|z(x, 0)\|_2^2 + 2 \int_0^T \kappa(z, t) dt \leq 2 \|z(x, 0)\|_2^2 + 2T\kappa(z, T_M), \end{aligned}$$

that is

$$\|z(x, T)\|_2^2 \leq 2 \|z(x, 0)\|_2^2 + 2T\kappa(z, T_M). \quad (19)$$

By adding (18) and (19) multiplied by  $(b_0 + 1)$ , we obtain, taking into account (16),

$$(a^2 - 1) \|z_x'(x, T)\|_2^2 + \kappa(z, T_M) \leq C_0 \kappa(z, 0) + T(2\sigma_0 + 2(b_0 + 1)) \kappa(z, T_M), \quad (20)$$

where  $C_0 = \max\{3b_0 + 2, a^2\}$ . If  $a^2 \geq 1$ , the first term of the left part of the inequality is replaced by zero, so the left part will not increase if  $0 < a^2 < 1$ , then replace the left part of the inequality (20) with a smaller expression:

$$a^2 \kappa(z, T) \leq C_0 \kappa(z, 0) + T(2\sigma_0 + 2(b_0 + 1)) \kappa(z, T_M),$$

and divide both parts by  $a^2$ . Here and below  $C_i, i = \overline{1, 5}$ ,—are constants depending only on  $a^2, T, b_0, \sigma_0$  (instead  $\sigma_0$  we can take  $\gamma_0$  and  $\beta_0$  because the inequality  $\sigma_0 \leq \beta_0 + 2\gamma_0$ , follows from (20)). Let  $T$  satisfies to inequality

$$C_3 = \frac{T}{\min\{1, a^2\}}(4\gamma_0 + 2\beta_0 + 2b_0 + 2) < 1.$$

from the very beginning of the second part of the proof. For  $T = T_M$  the inequality

$$\kappa(z, T_M) \leq C_1\kappa(z, 0) + C_3\kappa(z, T_M)$$

is valid, this means that for  $t \in [0, T]$  the energy inequality is also true

$$\begin{aligned} & \|z(x, t)\|_2^2 + \|z'_t(x, t)\|_2^2 + \|z'_x(x, t)\|_2^2 \leq \\ & \leq C_2(\|z(x, 0)\|_2^2 + \|z'_t(x, 0)\|_2^2 + \|z'_x(x, 0)\|_2^2). \end{aligned} \tag{21}$$

If  $T_M = 0$ , the inequality (21) is automatically satisfied when  $C_2 = 1$ .

3. **Uniqueness.** Let us justify the uniqueness of the solution. Suppose the opposite: there are at least two different solutions to the problem (6)–(9):  $y_1(x, t)$  and  $y_2(x, t)$ .

Without limiting generality, we can assume that they are different already at some interval  $t \in (0, T_A)$ . Assume that this is not the case, let then  $t_0 = \inf(\{t : \exists x \in [0, l], y_1(x, t) \neq y_2(x, t)\})$ . It follows from the continuity of the initial functions that the initial conditions for the problems are similar for  $t_0$ .

$$y_1(x, t_0) = y_2(x, t_0) = \varphi(x), \quad y'_{1t}(x, t_0) = y'_{2t}(x, t_0) = \psi(x), \tag{22}$$

So, it can be assumed that  $t_0 = 0$ ,  $y_1(x, t)$  and  $y_2(x, t)$  are two different solutions to the problem (6)–(9). For each of them there is a corresponding solution to the Cauchy problem (6)  $q_1(t)$  and  $q_2(t)$ . The existence of functions follows from (10)  $\sigma_1(t)$  and  $\sigma_2(t)$ , which define two different solutions  $z_1(x, t)$  and  $z_2(x, t)$ , by  $\sigma(t) = \sigma_1(t)$  and  $\sigma(t) = \sigma_2(t)$  respectively and as shown above, the formula (12):

$$y_i(x, t) = \frac{z_i(x, t)}{p_i(t)}, \quad i = \overline{1; 2}. \tag{23}$$

While we suppose existence of two solutions, let

$$\gamma_0 = \max\{\max_{t \in [0; T]} |q_1(t)|, \max_{t \in [0; T]} |q_2(t)|\}.$$

According to (10), (11), let's evaluate with a reserve the functions  $p_i(t)$ :

$$\exp((-\gamma_0 - 1)T) \leq p_i(t) \leq \exp((\gamma_0 + 1)T), \quad i = \overline{1; 2}. \tag{24}$$

From the formula (23) and inequalities (21), (24) it follows that with  $T$  selected, both the function  $y_1(x, t)$  and the function  $y_2(x, t)$  will be bounded, moreover, the majoring estimates are true:

$$|y_i(x, t)| \leq C_4, \quad |y_{ix}(x, t)| \leq C_5, \quad |y_{it}(x, t)| \leq C_5, \quad i = \overline{1; 2}. \quad (25)$$

Since functions  $F_2(w(x, t), w'_i(x, t), w'_x(x, t))$  and  $\frac{\partial F_2}{\partial w}(w(x, t), w'_i(x, t), w'_x(x, t))$  are continuous in their arguments adjacent to a bounded closed set (parallelepiped) (25), then they reach their extreme values on it, and, moreover, for any solution (23) inequalities are valid

$$|R[y]| = \left| \int_0^l F_2(y(x, t), y'_i(x, t), y'_x(x, t)) dx \right| \leq C_6, \quad \text{for } y = y_i, \quad i = \overline{1; 2}, \quad (26)$$

and

$$\left| \frac{\partial F_2}{\partial y}(y(x, t), y'_i(x, t), y'_x(x, t)) \right| \leq C_7, \quad \text{for } y = y_i, \quad i = \overline{1; 2}. \quad (27)$$

Here and further constants  $C_i, i \geq 6$ , depend only on  $T, l, \gamma_0$  and functions  $b(x), \beta(t), \varphi(x)$  and  $\psi(x), F_2(w(x, t), w'_i(x, t), w'_x(x, t))$ .

Let  $\theta(x, t) = y_2(x, t) - y_1(x, t)$ , then  $\theta(x, t)$  is the solution of the equation

$$\begin{aligned} \theta''_{tt}(x, t) &= a^2 \theta''_{xx}(x, t) + \theta(x) y(x, t) + \beta(t) \theta'_t(x, t) - \\ &\quad - R[y_2] \theta(x, t) - (R[y_2] - R[y_1]) y_1(x, t), \end{aligned} \quad (28)$$

with boundary and initial conditions (see (22))

$$\theta'_x(0, t) = \theta'_x(l, t) = 0, \quad (29)$$

$$\theta(x, 0) = 0, \quad \theta'_t(x, 0) = 0. \quad (30)$$

Next, apply the technique of the 2nd part of the proof. The difference from the 2nd part is the presence of the last two “terms” and the need for their evaluation.

Since the function  $\kappa(\theta, t)$  in this part of the proof depends on  $\theta$ , here  $T_M = \{t : \max_{t \in [0, T]} \kappa(\theta, t)\}$ . From Lemma 1 and (26) it follows that

$$2 \int_0^T \int_0^l R[y_2] \theta(x, t) \theta'_t(x, t) dx dt \leq C_{10} T \kappa(\theta, T_M).$$

Applying Lagrange’s theorem and then Lemma 1 to the next integral, we obtain taking into account (25) and (27)

$$\begin{aligned}
 & 2 \int_0^T \int_0^l (R[y_2] - R[y_1])y_1(x, t)\theta'_t(x, t) dx dt = \\
 & = 2 \int_0^T \int_0^l y_1(x, t) \frac{\partial F_2}{\partial y}(y_{cp})\theta'_t(x, t) \int_0^l \theta(\xi, t) d\xi dx dt \leq C_{11}T \kappa(\theta, T_M).
 \end{aligned}$$

Further, acting completely analogously to part 2, we obtain an energy inequality for the problem (28)–(30): for  $t \in [0, T_1]$ ,

$$\begin{aligned}
 & \|\theta(x, t)\|_2^2 + \|\theta'_t(x, t)\|_2^2 + \|\theta'_x(x, t)\|_2^2 \leq \\
 & \leq C_{12}(\|\theta(x, 0)\|_2^2 + \|\theta'_t(x, 0)\|_2^2 + \|\theta'_x(x, 0)\|_2^2),
 \end{aligned} \tag{31}$$

where  $T_1$  is some positive constant. However, for  $\theta(x, t)$ , the right-hand side (3) is zero. Therefore  $y_1(x, t)$  and  $y_2(x, t)$  on some segment  $[0, T_1]$  coincide. Contradiction. Thus, the solution of the problem (6)–(9) is unique.

### 4 The Properties of the Studied Class of Integro-Differential Equations

**Corollary 1** *Suppose that for some operator  $R[y]$  it was possible to find a function  $F_1(w(x, t), w'_t(x, t), w'_x(x, t))$  such that the operator  $P[z]$  given by the expression (4) satisfies the identity  $P[z] \equiv p(t)$  at  $t \in [0, T]$ , where  $z(x, t)$  is the solution of the problem (1)–(3). The initial functions  $\varphi(x)$  and  $\psi(x)$  are constrained as follows, to fulfill the conditions*

$$P(0) = 1, \quad P'_t(0) = 0. \tag{32}$$

Moreover, for any  $t \in [0, T]$ ,

$$P[z] \neq 0. \tag{33}$$

Then Theorem 1 is valid. for  $P[z] \equiv p(t)$ , and the expression (12) is rewritten as

$$y(x, t) = \frac{z(x, t)}{P[z]}. \tag{34}$$

**Proof** Conditions (32) the equations  $p(0) = 1$  and  $p'_t(0) = 0$  (and  $q(0) = 0$ , see (14)) are required to be executed. The inequality (33) under these assumptions is equivalent to the inequality (13).

*Example 1* Let's the initial conditions be imposed:

$$\int_0^l \varphi(x) dx = 1, \int_0^l \psi(x) dx = 0, \quad (35)$$

and true inequality

$$\int_0^l z(x, t) dx \neq 0, \quad (36)$$

for any  $t \in [0, T]$ , where  $z(x, t)$  is the solution to the problem (1)–(3).

We assume

$$R[y] = \int_0^l b(x)y(x, t) dx, \quad (37)$$

where  $y(x, t)$ —the solution of the problem (6)–(9), (37). The conditions for the corollary of the theorem are fulfilled (see (35), (36)). This solution exists, at least on some segment  $[0, T_0]$ .

From the formulas presented in the first part of the Theorem proof can be found function

$$p(t) \equiv P[z] = \int_0^l z(x, t) dx. \quad (38)$$

We obtain that on the segment  $[0, T]$  there exists a unique solution problems (6)–(9), (37), which is represented (see (34), (38)) as

$$y(x, t) = \frac{z(x, t)}{\int_0^l z(x, t) dx}.$$

At that

$$P[y] = \int_0^l y(x, t) dx = \int_0^l \frac{z(x, t)}{\int_0^l z(x, t) dx} dx = 1.$$

*Example 2* Let the operator  $P[z]$  be given as

$$P[z] = \left( \int_0^l z^2(x, t) dx \right)^{\frac{1}{2}} > 0,$$

where  $z(x, t)$  is the solution of the problem (1)–(3), and  $z(x, t) \neq 0$ . To fulfill the conditions (32), you must require that the initial functions satisfy the equations

$$\int_0^l \varphi^2(x) dx = 1, \int_0^l \varphi(x)\psi(x) dx = 0.$$

The conditions for the corollary of the theorem are fulfilled.

From the formulas presented in the first part of the Theorem proof can be found the operator

$$R[y] = \int_0^l (b(x)y^2(x, t) - y_x^2(x, t) + y_t^2(x, t)) dx. \tag{39}$$

The solution of the original nonlinear problem (6)–(9), (39) is associated with the solution of the linear problem (1)–(3) equality:

$$y(x, t) = \frac{z(x, t)}{\left(\int_0^l z^2(x, t) dx\right)^{\frac{1}{2}}}.$$

It is easy to see that for any  $t \in [0, T]$  there is a phase constraint  $P[y] = 1$ , which is equivalent to

$$\int_0^l y^2(x, t) dx = 1.$$

## References

1. Weinberg, M.M.: “Integro-Differential Equations”, Results of Science. Series of Mathematical Analysis. Probability Theory. Regulation. 1962, pp. 5–37. VINITI (1964)
2. Orlov, S.S.: Generalized Solutions of High-Order Integro-Differential Equations in Banach Spaces, p. 150. IGU, Irkutsk (2014)
3. Samarsky, A.A., Mikhailov, A.P.: Mathematical Modeling: Ideas. Methods Examples, p. 320. FIZMATLIT (2005)
4. Kalinin, A.V., Morozov, S.F.: On a nonlinear boundary problem of the theory of radiation transfer. J. Comput. Math. Math. Phys. **30**(7), 1071–1080 (1990)
5. Zaitsev, V.F., Polyanin, A.D.: Method of Separation of Variables in Mathematical Physics, Educational edn, p. 92. St. Petersburg (2009)
6. Ladyzhenskaya, O.A.: Mixed Problems for the Hyperbolic Equation. State Publishing House of Technical and Tactical Literature, p. 282 (1953)



7. Ladyzhenskaya, O.A.: Boundary value problems of mathematical physics. Science, 408 (1973)
8. Polyanin, A.D.: Handbook of Linear Equations of Mathematical Physics, p. 576. FIZMATLIT (2001)
9. Kuzenkov, O.A.: On the properties of a class of integro-differential equations in Lebesgue space. Nonlinear Dyn. Control. Issue 1. Digest of articles/ed. S.V. Yemelyanov, S.K. Korovin, 347–354 (2001)
10. Kuzenkov, O.A.: On a class of integro-differential equations in Lebesgue space. Differ. Equ. **38**(1), 134–135 (2002)
11. Kuzenkov, O.A.: The Cauchy problem for a class of nonlinear differential equations in a Banach space. Differ. Equ. **40**(1), 24–32 (2004)
12. Kuzenkov, O.A., Egamov, A.I.: “Optimal control for a class of integro-differential equations,” news of the Russian academy of natural sciences. Ser. Math. Model., Mech. Control. **1**(2), 140–145 (1997)
13. Kuzenkov, O.A., Egamov, A.I.: Optimal control for the oscillatory process. UNN Bull. Math. Model. Optim. Control **2**(19), 174–179 (1998)
14. Egorov, A.I.: Riccati Equations, 2nd edn., supplemented, p. 448. SOLON-Press (2017)

# Mapping of Politically Active Groups on Social Networks of Russian Regions (On the Example of Karachay-Cherkessia Republic)



Galina Gradoselskaya, Ilia Karpov and Tamara Shcheglova

**Abstract** The article shows which segments constitute social and political activity in online social networks in the Karachay-Cherkessia Republic (KChR) and the width of their representation. The author's technique allows to collect data on politically active groups of KChR. The segments of the social and political activity of the Republic on the social networks are shown. Eight main clusters of political activity in social networks of KChR were obtained by the author's method of grain clustering. Each cluster was analyzed by social network analysis methods. The most influential persons and social movements are shown, and features of their network activity were investigated.

## 1 Introduction

### 1.1 Description of the Research Problem

The space of social networks has many meaningful dimensions: social, political, economic. Each of them has its own particular representation in the country and regional aspect. This is dictated by the cultural, historical, informational features of the territory. Therefore, the study of social networks in a region, republic, or autonomous region of Russia is of particular research interest. In fact, the study of social networks in regions should begin with a primary mapping—understanding which set of groups represents a particular content area. The problem is that often even regional experts and managers hardly present the general picture of the activity of social networks.

Therefore, the mapping of social networks in the region is of interest from both the methodological (development of the algorithm) and informative points of view.

---

G. Gradoselskaya · I. Karpov · T. Shcheglova (✉)  
National Research University Higher School of Economics, 101000  
Moscow, Russian Federation  
e-mail: [tshcheglova@hse.ru](mailto:tshcheglova@hse.ru)

© Springer Nature Switzerland AG 2020  
I. Bychkov et al. (eds.), *Network Algorithms, Data Mining, and Applications*, Springer Proceedings in Mathematics & Statistics 315,  
[https://doi.org/10.1007/978-3-030-37157-9\\_13](https://doi.org/10.1007/978-3-030-37157-9_13)

It was necessary to understand the segments of social and political activity on social networks in the KChR and the width of its representation. It was also significant to identify key groups and actors on social networks, which should be noted while working in the information space of the Republic.

The study shows that the oppositional activity on social networks is significantly related to the actions of representatives of the Republic's elite. Social and political activity on social networks is diverse, it includes many streams and differs from other national regions of the Russian Federation.

Links to the groups, individuals and media on different social networks were provided by experts. On our part, we used a methodology for mapping social networks at the federal and regional levels of the Russian Federation, tested in several projects (it is reflected in Part 2.2 of this article). The mapping of social networks was carried out using the method of grain clustering.

Data were collected from the online social networks: Facebook, Vkontakte, Instagram, Odnoklassniki, and LiveJournal in April 2017.

## ***1.2 Research Problems of Information Gathering on Social Networks in the KChR***

The main problem was the lack of initial information from the representatives of the Republic about the situation on the social media of the Republic. Therefore, we used an additional method of initial information extension. The feature of this method is the minimum manual information collection. Having no more than 15 objects from all social networks on the start, we received a database with more than 5000 objects by our algorithm of grain clustering. As far as the gathering of initial data was done from five social networks, we had an opportunity to compare the results between them. Different social networks have diverse nature of social and technical character:

- Clusters on KChR groups on Facebook are the most heterogeneous, they characterize the social processes in the information space. They also became the basis for the search for similar resources in other social networks—Vkontakte and Instagram.
- There is a big expansion of personal connections of oppositionists via friendship networks on Facebook. In other networks, personal accounts are not so tightly represented.
- The structure of clusters on social networks Vkontakte and Instagram is much more friable and blurred than on Facebook. The threshold strength of communication between groups is also lower than on Facebook.
- The peculiarity of Facebook is political network engagement. When typing keywords in the Facebook search box, a link is sent to oppositional groups (even not the largest). For example, for the word “KChR” there is a reference to the oppositional group “For the rights of the KChR” with the size of only 298 participants. The word “Karachay...” refers to the oppositional group “Blogosphere of

Karachay-Cherkessia” (size: 5892 participants). The word “Cherke...” refers to the oppositional group “Circassian Renaissance” (size: 6922 participants).

- Facebook provides a consolidation of protest and nationalist activity from all the republics of the Northern Caucasus and the federal opposition. There is an active presence of foreign groups, first of all, Turkish.

In this article, we will consider the results of Facebook groups clustering.

## **2 Literature Review of the Regional Network Researches**

### ***2.1 Network Approaches to Regional Political Mapping on the Internet***

Mapping of political resources on the Internet has a very long tradition, and a wide range of applications to different technological platforms, such as email, chat rooms, blogs, social networks (Facebook, Twitter). With the development of technology, the types of relationships transformed, but the objects of mapping remained unchanged: either the political preferences of individual actors, or the political preferences of collective actors (groups, public, etc.). The goals of mapping have also not changed over time: it was necessary to understand which political attitudes are present in the political segment of the Internet of a particular region (country), and how they are relatively positioned to each other. And this information is not self-valuable—with its help it is necessary to understand the social processes taking place in society, to predict the development of the political situation.

Two main approaches to network mapping in political networks are automatic and manual expert coding of objects.

The most famous paper on automatic coding is an article of Adamic and Glance [1] on the automatic mapping of the American blogosphere in the elections of the 2004 year. These scholars visualized relationships between social and political blogs and have obtained different clusters. With a large amount of work spent on the collection of information, meaningful conclusions are likely to occur. There were discovered two main clusters: one supported the democratic party, the other—the republican party. There was a small number of links and intermediary blogs between them.

An interesting paper in which the author utilizes an expert approach is “Public Discourse in the Russian Blogosphere” by Etling et al. [4]. The mapping of the Runet here is carried out in order to understand the processes of political mobilization in Russia, under the leadership of Berkman Klein Center for Internet & Society. In this case, a lot of work has also been done on collecting empirical information and expert coding of blogs. The picture of political groups is more complex and variegated than in the study of Adamic and Glance. However, when only an expert approach is used, the resulting multidimensional space is difficult to interpret from a structural point

of view. The same principles and methodological techniques were used elsewhere by Kelly and Etling [8].

*It could be argued that the best approach for conducting mapping is a combination of automatic data collection and expert coding for understanding the social mechanisms which lie under the political processes taking place on the Internet.*

The technical issue of collecting information is constantly discussed in this kind of research collecting and visualizing the connections between objects in the virtual space. The work of Lin et al. [11] is devoted to this question.

The question of the collection algorithms is still debated, as in the study of Borgatti and Cross [3].

*An approach that combines search algorithms and clustering methods is the most promising for mapping social networks and is used in our author's algorithm in the study of KChR.*

The issue of measuring communication strength in virtual social networks is also relevant and was studied by Petróczi et al. [12].

In addition to the network approach, the study of political processes in social networks addresses another significant aspect of the study. It should be understood that the processes that group actors in social networks for political preferences do not always occur for natural reasons. Over time, manipulative processes of information dissemination and social projection are becoming increasingly active.

The works of Rusch [14], Atkins and Huang [2] are devoted to the social engineering of Internet fraud. Although they consider not political processes, the manipulation techniques are the same: authority, tradition, attraction, urgency, fear/threat, liking and similarity, reciprocation, social proof, etc. Rusch [14] emphasizes the psychological aspects of the manipulation via Internet fraud and describes the used principles of social psychology. Separately, he describes the principles of “reverse-engineering”, aimed at the philistine psychology and the typical situational reaction.

## **2.2 Regional Policy Network Studies**

In the Russian-language literature, the mapping of the political space in the regions is primarily of practical importance. Sharkov discusses the relationship between network (virtual) and real identities, as well as the transition to the phenomenon of the “mass person”. The theoretical foundations of social design processes in the virtual space are considered. The peculiarity of the perception of political processes is their immersion in the communicative context [15].

The authors of other articles conducted regional studies in many regions of Russia. In the work of Gradoselskaya [5], sociopolitical processes were studied in eight regions of Russia, differing in the level of innovative development: from Yakutia to the Moscow region. In the other article of these authors [6], the methodology is written in detail.

In general, in Russia, the Institute of Sociology of the Russian Academy of Sciences is engaged in detailed research of the regions. The results are shown in a number of publications [10, 13].

A study conducted by Kolosov and Sebentscov [9] describes the geo-political discourse about the North Caucasus in Russia. They distinguish nationalistic, oppositional, Islamist discourses, which partially coincides with the results of our study presented in this article.

The most obvious consequences of social design in the virtual space are manifested in mass public actions, as shown by Kelasiev et al. [7]. They consider the mechanisms of influence on target audiences, their mobilization, as well as the organization of dialogue (or confrontation) of the government and public activists.

### 3 A Brief Description of the Mapping of Social Networks Using Grain Clustering Algorithm

The grain clustering method was proposed by G. Gradoselskaya in 2014 when solving the research problem of structuring politically active groups of Russia. The standard mathematical and linguistic approaches that existed at that time were of no practical use and/or proved to be too costly in terms of time and finances. At the moment an article is being prepared with a detailed description of the methodology.

From that time to date, more than 20 scientific and commercial studies have been conducted on network mapping at different levels: federal, regional, city and district. These were studies of political activity, social, nationalist, criminal, professional networks, etc. In all cases, surprising results were obtained.

The study on the KChR is given as an example in this article, the automatic method of collecting information gave very interestingly, nontrivial structure—the ring, where clusters are grouped according to the principle of “strung beads” (see Fig. 2). This is the fundamental difference between the mapping using the grain clustering method and the research of social networks, where groups were selected manually. In the case of manual collection, usually one cluster is obtained, with a pronounced core, loosely coupled peripherals and no complex structures.

1. Briefly, the grain clustering algorithm can be described as a sequence of several steps:
2. The selection of a small number of grain groups. Experts select a small number (from 5 to 10) of the groups that most closely correspond to the research topic—political, financial, criminal groups, etc. In this case, there were regional groups of Karachay-Cherkessia. As noted above, the number of grain groups was extremely small, since the experts had no idea what was happening in his region.
3. Between grain groups, the density of bonds is recalculated. This is considered a zero-cluster increment cycle.

4. All groups of all users of the grain groups are gathered and ranked by a decrease in the frequency of common users. Thus, the connection between cluster objects (groups) will be the number of joint users between groups.
5. Next, the density change is recalculated when each new group is added to the cluster. As soon as the cluster density begins to decline rapidly, the addition of new groups ceases, and the first cycle of cluster accumulation is considered to be completed.
6. After that, we go back to step 2, and start the next increment cycle. Based on our experience, in order to map most of the groups in the region, there are only 2–3 cycles of increment are needed.

In addition to controlling the density of the cluster, when implementing the cycles of the algorithm, you can control the threshold values of the links, add a vector of correspondence of new groups already included in the cluster in previous cycles.

The scheme of actions of the grain clustering algorithm is shown below in Fig. 1.

After all groups have been selected and visualized and structural clusters have been found, the automatic collection phase is considered complete. The next stage is expert coding, when all collected groups are reviewed by experts and categorized in accordance with the goals and objectives of the study. Further, a comprehensive analysis is carried out and conclusions and recommendations are prepared.

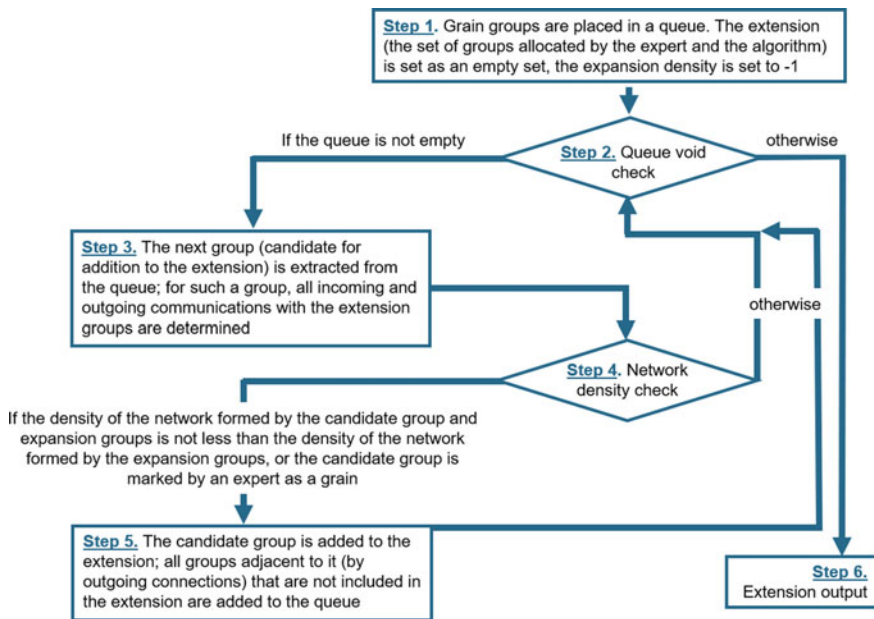


Fig. 1 Grain clustering algorithm is a method with nondecreasing density

## **4 Structuring Clusters of Politically Active Groups of KChR on Facebook**

### ***4.1 General Clustering of Groups that Show Political Activity in the Information Space of the KChR and Neighboring Caucasian Regions***

Hundreds of thousands of users and groups are active in social networks. The method of mapping network activity helps to understand what is happening. On the basis of joint actions of users (participation in communities, adding friends, commenting) links are built between groups in the network. The greater the number of connections, the higher the likelihood that different groups are united into a common network and coordinated from one center. The user who is interested in politics will be included in several groups, and the set of groups for each actor may be different. Combining personal activities, we get a picture of the overall structure of relations between political groups.

To identify the structure, 5–10 large “model” groups are sufficient, and through the connections of participants with other groups, the entire cluster is identified (opposition, pro-government, regional or national-oriented, etc.).

Next, the experts determine the boundaries of the cluster (to reduce the density of bonds); they distinguish specific types of groups (opposition, pro-government, nationalist, interests, advertising, services, etc.), key actors (professionals who create groups and attract new members into them). Then an analysis of the social replenishment mechanisms of clusters and the links between them is carried out.

According to the primary structural mapping on Facebook, eight clusters were identified. General characteristics of clusters are in Table 1.

Figure 2 shows the results of visualizing the links between the 370 Facebook groups. The threshold is 150 mutual participants in groups.

We obtained clearly defined eight clusters, each of which is easily interpreted due to the homogeneous content of their groups. They can be described as follows: Abkhazian, Adyghe, Kabardino-Balkarian, Caucasian (Islamic), Karachay-Cherkess, Federal Opposition, Stavropol, Pan-Turkism in the Caucasus. Pan-Turkism and Caucasian clusters are the largest and influence all others.

The cluster “Pan-Turkism” is based on the activity of Turkish information resources. Partly in Turkish, partly in Russian, partly in the national languages of the Republic. Promote the idea of uniting all the Turks and territorial claims to the Russian territories (Southern Volga, Siberia, the Caucasus). About a third of the groups in this cluster are directed to work in the KChR (it contains the name of the Republic or its constituent peoples in the name).

Lists of publications were obtained for all clusters and their content was analyzed. The texts of the publications of the Pan-Turkism cluster are the most interesting in content, since they represent a number of basic queries in relation to the Russian Federation. But they are voiced through groups dedicated to the KChR.



**Table 1** Main structural clusters in the social network Facebook on KChR

No.	Cluster name	Description of the cluster
1	Abkhazian	A secular cluster, partly in the national language. Devoted to the problems of Abkhazia and life conditions of Abkhazians in other republics of the Caucasus
2	Adyghe	Most of it is nationalistic, half of it is in the national language. It also contains sociopolitical groups and social movements. The main idea is “the unification of divided Nations”
3	Kabardino-Balkar	Mostly nationalistic, partly in the national language. It also contains sociopolitical groups and social movements
4	Caucasian, Islamic	There are groups of Caucasian republics: Chechnya, Ingushetia, Dagestan. Main content—integration on the basis of religion. Some groups are purely Islamic, promoting lifestyle (“Islamic family”, “Islamic values”, “Islamic medicine”, “Oppression of Muslims”, etc.). Partly in national languages
5	Karachay-Cherkesia	The cluster includes sociopolitical groups dedicated to the KChR, media, and groups in support of regional oligarch T. (as well as public movements such as “RCPC”, “Elbrusoid”, etc.). The cluster is associated with both the Caucasus-Islamic cluster, the oppositional cluster, and the Stavropol cluster
6	Federal opposition	Federal groups have met before in mapping oppositional groups at the Federal or Moscow level. Coordinate the activity of the opposition throughout the country (including the KChR)
7	Stavropol	The cluster is dedicated to the neighboring region—Stavropol. Mostly informational
8	Pan-Turkism in the Caucasus	The cluster is based on the activity of Turkish information resources. Partly in Turkish, partly in Russian, partly in the national languages of the Republic. They promote the idea of uniting all the Turks and territorial claims to the Russian territories (Southern Volga region, Siberia, the Caucasus). About a third of the cluster’s groups are aimed to work in the KChR (contains the name of the Republic or its people in the title). A third of groups from Turkey are on the national language of KChR (Cherkess language)

It is possible to allocate the main subjects from the Turkish groups working especially for KChR:

- Territorial claims: almost all territories of modern Russia is considered as gifts of the Tatar khans;
- Linguistic claims: The North Caucasian languages are considered as basic, primogenitors of all modern European and Indian languages;



Fig. 2 Mapping the KChR groups, threshold—150+ mutual members

- Genetic claims: investigating DNA, it has been established (unclear by whom—scientific visibility of these statements is more important) that Caucasians are primogenitors of all Europeans.

In the cluster “Caucasian (Islamic)”, there are groups of Caucasian republics: Chechnya, Ingushetia, Dagestan. According to the content, they are unified by one religion. Some groups are purely Islamic, the propaganda of the way of life (“Islamic family”, “Islamic values”, “Islamic medicine”, “oppression of Muslims”, etc.). Partially in national languages. Quite a lot of opposition groups, sociopolitical movements, pan-Turkic groups, Internet media.

Apart from automatic clustering, expert coding was carried out for all 370 groups, and informative typologies in each cluster were identified.

Thus, in the Abkhazian cluster, a significant part of groups in the national language was identified. In the Adyge cluster, in addition to groups in the national language, a subgroup of resources dedicated to the nationalist movement “Adyge-Khase” was identified. In the “Caucasian, Islamic” cluster, two subgroups dedicated to the Caucasian republics of Chechnya and Dagestan were identified.

In addition to the groups characterizing the content of each cluster, the types of groups in each of the 8 clusters were identified. These are sociopolitical and oppositional groups and sites representing the media.

## 4.2 Cluster of Politically Active Groups of KChR

Figure 3 shows the groups in the cluster directly devoted to the KChR.

Groups dedicated to KChR are present in all clusters. If we combine all these groups in a separate scheme, we can distinguish 4 main types that dominate in the KChR:



**Table 2** Main structural clusters in the social network Facebook on KChR

No.	Group name	Degree centrality	Group type
1	Tarihhibiz	1.000	Sociopolitical
2	Kyarachilila	0.919	Oppositional
3	Elbrusoid	0.915	Sociopolitical
4	Aliy Totorkulov	0.915	Region oligarch's
5	Blogosphere of Karachay-Cherkesia	0.859	Sociopolitical
6	CIRCASSIAN RENAISSANCE	0.641	Sociopolitical
7	Çerkes TV	0.638	Turkish
8	Circassian Information Channel	0.605	Media
9	ÇERKESYA	0.605	Turkish
10	Military exploits of Karachai and Balkar	0.553	Republican
11	Çerkes (Adıge) Tarih Kulübü	0.531	Turkish
12	Çerkes Haber-Paylaşım ve Kamuoyu Oluşturma Grubu	0.483	Turkish
13	GÖLCÜK GÜMÜŞ.ÇERKES TAKILARI	0.455	Turkish
14	Çerkes Milliyetçileri (Circassian Nationalists)	0.451	Turkish
15	ÇERKES (ADIGE) GENETİĞİ	0.380	Turkish
16	Adygea “Adyghe Khase”	0.376	Sociopolitical movement “Adyghe Khase”
17	Çerkesler	0.373	Turkish
18	Helping compatriots from Syria	0.369	Helping compatriots from Syria
19	Unity and Development	0.365	Sociopolitical
20	Circassian Media	0.351	Media
21	Fund of the Adyghe culture “Heritage” (oshad.ru)	0.341	Adyghe
22	Çerkes Haber	0.340	Turkish
23	ÇERKES SOYKIRIMI ve DEMOKRATİK HAKLARI	0.332	Turkish
24	<a href="http://www.natpressru.info/">http://www.natpressru.info/</a>	0.328	Media
25	Kabardino-Balkarian “Adyghe Khase” RSM	0.322	Sociopolitical movement “Adyghe Khase”

It is necessary to emphasize foreign sources of opposition and nationalist content, mainly from Western and Turkish information resources. The quality of content is very professional, expensive, immersed in the realities of the Republic—very complete. They are also used to model future social events (marking special dates, organizing marches, supporting social movements, etc.).

In all networks, there are nationalist, Islamic, opposition groups in Russian, English, Turkish, Arabic, national languages of the Republic. Some of these resources were created specifically to work for the KChR audience (the group names contain either the name of the Republic or the name of the people living on its territory).

If we consider interconnection (compare the representation of accounts and groups between different networks), we can note their different roles in the coordination of communication and information processes in the KChR:

- Facebook: consolidation and coordination of the main opposition and nationalist resources within the Republic, federal and foreign sources. The structure is more pronounced, clear clusters are distinguished, related to their substantive specialization;
- Vkontakte: district groups, groups by interest, youth groups. The structure is more friable, there are few clusters;
- Instagram: a source of video materials, scattered into personal accounts and tightly connected structures of information resources, media;
- Odnoklassniki: groups of household and leisure purposes, have a long history and are supported by large social inertia;
- LiveJournal: points of publication of negative materials against the leadership of the Republic.

From the point of view of the power of influence on the population of the Republic, Facebook and Vkontakte networks are almost equal, although they are aimed at different purposes. The network of Facebook is focused on international opposition and nationalist contacts, through this network, the main content is produced, prepared on international resources (including off-grid, for example, "Cherkessia.net"). Foreign languages of communication are also actively used: English, Turkish, sometimes Arabic. The Vkontakte network concentrates on local resources, both have public and political content, and neutral, in relation to the region and human settlements.

Odnoklassniki represents groups mostly for domestic and leisure purposes, but they also contain political materials placed by special accounts. To reveal such materials on the statistical level is rather difficult, it is necessary to periodically monitor the information resources collected in this study.

Other networks are more entertaining and communicative. The links where are more scattered and quickly break up into personal accounts (as, for example, on Instagram). In LiveJournal, many accounts are either obsolete (inactive since the 2016 year), or they are used as points of throwing dirt on the leadership of the republic (to then disperse information via Facebook, Vkontakte, Twitter) and they can be regarded as bots.

The LiveJournal is not currently an active network, the vast majority of KChR accounts have moved from LiveJournal to other social networks. In fact, Instagram now took over the functions of LiveJournal for posting current observations of the user over himself and the world around him, in addition, he significantly lowers the threshold level, since the placement of materials in the Instagram does not require the preparation of texts, just a smartphone.

The opposition's personal networks are a single and tightly connected, through which information of nationalist, anti-state content is quickly disseminated. Many of them are connected with information technologies, possess methods of dissemination of information and goods on the Internet, which increases their potential danger in information confrontation.

Protest resources in almost all networks occupy a central position and accumulate around themselves resources in basically two types: the media and social movements. The potential impact that these resources can have on the information space of the Republic can be assessed as very significant.

Special attention should be paid to public movements that openly support the opposition: "Elbrusoid", "The Russian Congress of the Peoples of the Caucasus" (RCPC), "Adyghé Khase" (with branches in the Karachay-Cherkessia Republic and neighboring republics), "Tarkhibiz", "Karachay Renaissance", "Unity and Development", etc. Potentially, these movements can be used to conduct social actions and/or protest actions.

There are quite a few groups (relatively small in size) that are associated with the world of cinema, directing, showing films, hiring actors for productions, etc., in the Vkontakte network of the KChR. Conclusions on this phenomenon: First, the cultural elite of the Republic is actively involved in the process of producing nationalist, anti-government content. Secondly, production clips of nationalistic, anti-government content are produced. Thirdly, actors are needed to conduct actions, implement manipulative methods of crowd control, and so on.

In all networks, there is a phenomenon of "information mining"—preparation for the future information campaign of 2017. Bots are created, accounts are pumped through, accounts are activated, which until now have been abandoned.

**Acknowledgements** The study has been funded by the Russian Academic Excellence Project "5-100".

## References

1. Adamic, L.A., Glance, N.: The political blogosphere and the 2004 US election: divided they blog. In: Proceedings of the 3rd International Workshop on Link Discovery. ACM (2005)
2. Atkins, B., Huang, W.: A study of social engineering in online frauds. *Open J. Soc. Sci.* **1**(3), 23–32 (2013)
3. Borgatti, S.P., Cross, R.: A relational view of information seeking and learning in social networks. *Manag. Sci.* **49**(4), 432–445 (2003)
4. Etling, B., Alexanyan, K., Kelly, J., Faris, R., Palfrey, J., Gasser, U.: Public discourse in the Russian blogosphere: mapping RuNet politics and mobilization. Berkman Center Research Publication No. 2010-11, Oct 19 2010
5. Gradoselskaya, G.V.: Analiz sotsial'nykh setey v kontekste vyyavleniya sotsiokul'-turnykh razlichiy vo vnutri regional'nykh sotsial'nykh otnosheniyakh. *Rossiya: re-formirovaniye vlastno-upravlencheskoy vertikalii v kontekste problem sotsio-kul'turnoy modernizatsii regionov.* 2017. Razdel 4. Str. 272–348

6. Gradoselskaya, G.V., Karpov, I.A., Scheglova, T.E.: Information space of social networks as reflection of social installations of the population on the relation to bodies of authority and management. *Rossiya i mir: global'nyye vyzovy i strategii sotsiokul'turnoy dinamiki*, Materialy Mezhdunarodnoy nauchno-prakticheskoy konferentsii (Moskva, 12–13 oktyabrya 2017 g.), 172–179
7. Kelasiev, O.V., Kazakov, S.V., Leies, A.U.: Spetsifika kommunikatsii vlasti i naseleniya v kontekste massovogo publichnogo protesta. *Zhurnal sotsiologii i sotsial'noy antropologii*. Tom IX. № I (34), 103–122 (2006)
8. Kelly, J., Etlings, B.: Mapping Iran's online public: politics and culture in the persian blogosphere. The Berkman Center for Internet & Society at Harvard Law School, Research Publication No. 2008-01, 6 Apr 2008
9. Kolosov, V.A., Sebentsov, A.B.: Severnyy Kavkaz v rossiyskom geopoliticheskom diskurse. *Polis. Politicheskiye issledovaniya* **2**, 146–163 (2014)
10. Kozlov, N.D.: Politicheskiye kul'tury regionov Rossii: uravneniye so mnogimi neizvestnyimi. [http://www.civisbook.ru/files/File/Kozlov\\_2008\\_4.pdf](http://www.civisbook.ru/files/File/Kozlov_2008_4.pdf)
11. Lin, J., Halavais, A., Zhang, B.: The blog network in America: blogs as indicators of relationships among US cities. *Connections* **17**, 15–23 (2007)
12. Petróczy, A., Nepusz, T., Bacsó, F.: Measuring tie-strength in virtual social networks. *Connections* **27**(2), 39–52 (2007)
13. Power and Society in Russian Regions: Vlast' i obshchestvo v regionakh Rossii: praktiki vzaimodeystviya. Moskva Institut sotsiologii RAN 183 (2015)
14. Rusch, J.J.: The social engineering of internet fraud. Paper Presented Internet Society Annual Conference. [http://www.isoc.org/isoc/conferences/inet/99/proceedings/3g/3g\\_2.htm](http://www.isoc.org/isoc/conferences/inet/99/proceedings/3g/3g_2.htm). Accessed 26 Mar 2007
15. Sharkov, F.I.: Visualization of political media space. *Polis. Polit. Stud.* **5**, 97–107 (2016)

# Social Mechanisms of the Subject Area Formation. The Case of “Digital Economy”



Oxana Mikhailova , Galina Gradoselskaya  and Alexander Kharlamov 

**Abstract** The structure of natural language could be considered a semantic network. This implies the allocation of the speech markers, which describe the subject and semantic areas. In this article, a wide range of texts about the digital economy was analyzed, making it possible to show the thematic structure of this subject area. Central and peripheral concepts were identified to characterize theoretical core concepts and related topics clarifying the application of the digital economy. Identification of the thematic areas was performed in two ways—through the construction of a thematic tree (neural network modeling in the Text Analyst) and the analysis of semantic networks. The results, approaches, and methods of this study could be used during the investigation of the other large thematic fields related to new ideological currents, being developed as an element of social design and management.

**Keywords** Digital economy · Social networks · Social network analysis · Text analyst

## 1 Introduction

The digital economy is a term that has become popular in various sectors of the economy, public administration, entrepreneurship, and modern communications. This concept was used in governmental programs. It is actively included in the management tools worldwide. Thus, it is possible to interpret the “digital economy” as one of the new ideologies, which proliferated far beyond the managerial sphere.

A set of concepts accompanying and explaining the key term “digital economy” is published in expert texts, articles, books. Books devoted to (or containing the term)

---

O. Mikhailova (✉) · G. Gradoselskaya · A. Kharlamov  
Higher School of Economics, Moscow, Russian Federation  
e-mail: [oxanamikhailova@gmail.com](mailto:oxanamikhailova@gmail.com)

A. Kharlamov  
Institute of Higher Nervous Activity and Neurophysiology of RAS, Moscow, Russian Federation  
Moscow State Linguistic University, Moscow, Russian Federation

© Springer Nature Switzerland AG 2020  
I. Bychkov et al. (eds.), *Network Algorithms, Data Mining, and Applications*, Springer Proceedings in Mathematics & Statistics 315,  
[https://doi.org/10.1007/978-3-030-37157-9\\_14](https://doi.org/10.1007/978-3-030-37157-9_14)



“digital economy” have been actively produced over the past 20 years (for example, [1–4]). These publications that determine the development of the digital economy as a subject area, became the object of our study. 312 publications were selected expertly (books, articles, monographs).

The base with the names of publications could be found here: <https://drive.google.com/file/d/1wmz9Y37vpo9dWGmZN-xgFd-JMaEymBLT/view?usp=sharing>. We attempted to structure the thematic space, surrounding the concept of the “digital economy”. To do this, key speech markers were positioned in space in such a way, that the semantic structures became visible [5–7].

The structurization of the information field surrounding and absorbing the key concept of the “digital economy” was carried out in two ways. First, the mechanism of neural networks was involved (using the Text Analyst program). We received a structure in the form of a tree. Each of the nodes on the top level of the thematic tree is described by its own group of speech markers. Secondly, we have built a network that directly connects all the speech markers. Such a network is called semantic. The construction and visualization of the network were carried out using Automap and ORA, respectively. The advantage of network modeling is the ability to calculate centrality indices. These indices helped us to test the main hypothesis of our study.

The main hypothesis of the study was that the “digital economy” is an ideological concept (the alternative was that this concept is instrumental). Ideological terms are usually located in the center of the network and have an impact on thematic areas that follow the development of the main concept. Instrumental terms, on the contrary, are located between the center and the periphery (marginal, remote peaks), with the help of which the main concept is realized. In this sense, ideological concepts possess “power” (in terms of Van Dijk T. A.) [8]. In addition to testing of the main hypothesis, we expertly structured the semantic field and compared the grouping with the thematic groups found by the Text Analyst program. Another characteristic that identifies a subject area as ideological is the presence of emotional, socially programming markers.

In this article, we firstly address the approaches of modern conceptual analysis. Then we describe the thematic content of the digital economy field. Next, we characterize the change of epistemes in the field of the digital economy. After that, we present the methodological toolkit and show the results of testing the main hypothesis of the study. The conclusion discusses the findings and outlines the prospects for the use of the developed methodology.

## 2 Literature Review

### 2.1 *Approaches in the Modern Conceptual Analysis*

The structurization of the information field is usually performed in the studies of conceptual spaces. The modern conceptual analysis could be carried out in the framework of historical, scientific, and critical approaches [9, pp. 161–162]. We describe each of the approaches and give examples of studies that were conducted by the representatives of each of them.

Historical approach reveals the temporal context of the concept. It gives a better understanding of the changes in the concept's position in the public sphere throughout the history of the concept. Such studies are not limited to etymology. They explain the causes and contexts of transformations of the meaning of a concept.

For example, Wei and his colleagues analyzed the evolution of concepts, approaching the conceptual field as a dissipative system [10]. The term “dissipative system” was taken by computer scientists from thermodynamics, because it allowed them to explain a wide range of modifications that can occur during the “life” of the conceptual field (emergence of new concepts, disappearance of old, semantic transfer, and semantic diversification).

The next approach is scientific. It differs from the historical due to the interest in the practical improvement of the conceptual apparatus. This means that researchers, which represent this approach tend to more clearly define concepts. Therefore, they aim to find out acceptable contexts for concepts usage, relationships between different concepts, and shades of meanings. Examples of the scientific approach to the analysis of the concept and its environment can be considered as generalizing approaches in technical and human sciences, works devoted to the automatic processing of definitions of terms, the formation of brief contents of articles.

Automatic synthesis of texts is now in demand in medicine. The compilation of brief contents of medical articles can be carried out by clustering weighted networks. In these networks, the nodes are sentences and the links are the degrees of similarity of sentences among themselves [11]. For each cluster, the sentence is selected that is the most similar to the other sentences of the cluster. This sentence receives the status of a generalizing sentence. The result is a summary of the article, equal in length to the number of clusters in the network. Machine processing of definitions of terms is aimed to reflect the network structure of such concepts and connections with the other ones [12]. Studies of the possibility of integration of the new concepts into conceptual systems are being carried out in a similar way [13]. The authors of such studies calculate the similarity of words, on the basis of which the decision is made to include/not include a new concept in the ontology.

The last approach to working with conceptual fields and concepts we want to mention is the critical. Knowledge in it is conceptualized as a source of power. Representatives of the approach study how concepts are formed under the influence of social forces. These researchers seek to shed light on the impartial aspects of the meanings invested in commonly used concepts. Works performed within the

framework of the critical approach are characteristic of postmodern social sciences, the most famous of which is Foucault. Usually, the genealogical approach is used to reconstruct phenomena, but in principle it can be used to study individual concepts [14]. The main disadvantage of the approach compared with the scientific and historical is the weak formalization of the procedure.

Our approach to the analysis of the social mechanisms of the digital economy combines scientific and critical approaches. The scientific approach is implemented through thematic modeling and visualization of semantic networks. Elements of the critical approach could be seen in the expert analysis of the formation of the subject area and verification of the main hypothesis of the study.

## ***2.2 Expert Analysis of the “Digital Economy” Field Formation***

In this study, a desk research of 312 publications on the digital economy serves two functions. Firstly, it describes the empirical object of research. Secondly, it demonstrates the proposed research directions in the study domain. The articles included in our database on the digital economy covers the period 1996–2017. These articles were selected to represent thematic diversity. We did not want to quantitatively represent the research on the digital economy. Therefore, we selected for analysis the most cited articles belonging to various subtopics within the digital economy to reflect the research field of meanings around the concept of “digital economy”. The variety of sources of narratives is also present (our database of articles contains works by authors from 47 countries).

Studies in the digital economy can be considered evolving in some fashion. We understand fashion waves as epistemes (changes in the discursive contexts of using the concept, reflecting the stages of its institutionalization [9, p. 168]). Discursive contexts are thematic configurations that use a concept in a specific time period. We identified the following topics within the framework of which the “digital economy” concept was used: digital marketing, digital price, digital economy modeling, management, mass media, smart cities, labor market, e-commerce, tourism, digital culture, electronic services, electronic strategies, dissemination of the digital economy, e-business, intellectual property rights, e-government, and information technology. Descriptions of some of the listed topics are given below.

The topic “Dissemination of the digital economy” reflects changes in the economies of countries that faced the introduction of digital technologies. For instance, Kostakis, Rus, and Bavenes describe the socio-environmental consequences of the establishment of two competing value models in the digital economy [15]. The authors concluded that both models have a negative impact on the ecological situation. One of the models could lead to the maximization of capitalist profits and capital accumulation. The second model, although it has softer effects on the economy, has serious transaction costs.

The theme “Mass Media” includes narratives, related to the functioning of the media in the digital economy. Salamon describes the formation of a community of freelance journalists, who teamed up to fight for the interests of the feminized class of online journalists. This class was formed as a result of cheaper journalists in the digital economy [16]. The author concludes that the traditional methods of struggle (boycotts, strikes, and legal actions) lose their effectiveness in the modern economy. Thus, digitalization reduces the abilities to fight for the rights of media workers.

The subject of “Intellectual Property Rights” is devoted to various kinds of difficulties which arise from the need to protect copyright in the transition from the analog type of information to digital. On the one hand, the authors recognize cheaper methods of transporting information flows. On the other hand, they talk about the problems of control over unauthorized copying and distribution of information. Violation of ownership of information leads to a decrease in revenue of information content producers [17].

It is difficult to say that the conclusions drawn by researchers in the framework of thematic fields are positive. The authors expect changes, the degree of favorableness of which is controversial for all spheres of society. In addition, the articles do not present a unified approach to the definition of the digital economy.

Here we want to describe the dynamics of epistemic changes that occurred after the appearance of the term “digital economy” in the work of John Tapscott in 1995 [1]. The study of this concept was made first in relation to e-commerce and business processes [18, 19]. These themes remained dominant throughout the development of all areas.

In the 2000s, the subject spectrum of the digital economy was supplemented with cross-country studies and the study of the impact of the digital economy on the ecological situation [20, 21]. Also, issues related to digital inequality, the influence of digitalization on an ordinary person as a consumer and a participant in noneconomic relations, gradually began to rise in quantity.

By 2005, together with the actualization of new topics (tourism, legal support of the digital economy) [22, 23], there was a diversification of the spectrum of research discussing the impact of the digital economy on the social sphere. First of all, due to the articles on the integration of digital technologies in the educational process in schools and higher educational institutions [24].

The trend in analyzing the digital impact on the information society persisted in 2010. In particular, scholars began to write about the impact of the digital economy on medicine and the labor market [25, 26]. In addition, efforts to calculate the macro- and micro-indicators of the success of the introduction of the digital economy and its operation have intensified. That is, the subject of the research interest that has taken over was the indicators of the success of the digital economy, the possibility of accelerating development and the readiness of various societies to digital changes [27, 28].

Toward the middle of the 2010s, the vocabulary and research tools for studying the response of the labor market to the digital economy have significantly expanded [29, 30]. In addition, issues of digital security and penalties for crimes in online space have become significant [31–33]. These issues, along with a new topic on

the specifics of the digital state and the axial issues of e-commerce and e-business, continue to occupy major positions in the research field of the digital economy even now [34].

### ***2.3 Ideological Character of the Term “Digital Economy”***

Consulting agencies have attempted to study the digital economy. However, reports submitted by the largest companies McKinsey, Deloitte, Boston Consulting Group [35–37] have serious drawbacks. The methodology is not transparent. Secondary data and expert estimates were used. The geographical range of countries was limited. The negative effects of the digital economy were missed out. In addition, they do not have a single definition of the digital economy.

The term’s creator (Tapscott) defined the digital economy through twelve characteristics: knowledge, digitization, virtualization, molecularization, integration, disintermediation, convergence, innovation, prosumerism, immediacy, globalization, and the discordance [1]. In the definition proposed by this author, the controversial nature of the concept is already noticeable. For example, along with innovation and integration, the disorder and prosumerism are predicted.

## **3 Data Processing with the Help of Automatic Networks**

### ***3.1 Toolkit for the Text Analysis***

When analyzing the texts in the framework of this work, two classes of tools were used. Automap (<http://www.casos.cs.cmu.edu/projects/automap/>) for text analysis and ORA (<http://www.casos.cs.cmu.edu/projects/ora/>) for visualization when processing the information about the frequency of occurrence and coexistence of words in the text. The program for automatic semantic text analysis Text Analyst (<http://www.analyst.ru/>) on the basis of frequency information reveals the key concepts of the text, calculates their weight characteristics, and forms the semantic network of the text as a semantic portrait of the text, from which it extracts the thematic structure of the text in as a minimal tree subgraph of the mentioned semantic network [38].

Replacing the frequency portrait of the text with its semantic portrait leads to a tremendous change in the analysis results: during the analysis, real hubs (vertices of the semantic network with the greatest connectivity in the text) are revealed in the text, unlike imaginary ones based on the frequency characteristics of the text.

## 3.2 *Conceptual Statistical Analysis of Texts*

Unlike existing approaches to text analysis based on the unigram text model (LSA, pLSA—when the text is considered as a set of unrelated words—“bag of words”), Text Analyst implements an approach based on the n-gram text representation. In this case, the text is considered as a set of n-grams (in the specific case  $n = 10$ ), the statistical characteristics of which (sets of n-grams) are calculated using an iterative procedure based on statistics of the frequency of occurrence and pairwise occurrence of words of the source text. In this case, identification coupled within this text pairs of words allows building a network on which to further implement an iterative procedure for rearranging the weight characteristics of the vertices of the network. This allows identification of the vertices associated with the greatest number of other vertices with the greatest weight. These vertices are the most significant in this text.

### 3.2.1 **Primary Processing**

Preprocessing is not needed in this algorithm. That is why it works well with all European languages (texts in the Czech language are as well analyzed by the English version of Text Analyst—provided that the diacritical marks are removed—as well as the English texts). But primary processing reduces information noise. It splits into: (1) removing non-textual information from the text—tables, figures, numbers, formulas, abbreviations (except for special cases of their usage); (2) text segmentation into words and sentences; (3) removal of stop-words, prepositions, commonly used words (by building appropriate dictionaries—hence the dependence on the language); and, finally, (4) lemmatization. The text was normalized and after that processed statistically: the frequency network of the text was constructed.

### 3.2.2 **The Frequency Network**

At the stage of frequency processing, the text is indexed: the frequency of occurrence of words and word pairs is calculated taking into account the division of the text into sentences. That is, words from adjacent sentences are not combined in pairs. After indexing, the pairs of words are combined into a network using the following algorithm. In each pair of words, there is the first, and there is the second, as they are found one after another in the text. All pairs line up one after another according to the principle “the first word of the next pair is combined with the second word of the previous pair.” It is important to understand that pairs, thus, are arranged in chains, which can be locked to other pairs, as well as branch. The resulting network will be directed: you can move along these chains only in one direction. Breaks of chains, and separately standing couples are possible. Each vertex of such a network is marked by the frequency of occurrence of the corresponding word, and each connection is marked by the frequency of occurrence of the corresponding pair.

It is very important to understand that there may be couples with the same first and different second words. This is the basis for the subsequent re-ranking of the network nodes. That is, such sets of pairs are combined in the so-called asterisks with one common “first” vertex and a set of “second” vertices—semantic features of the “first” vertex.

### 3.2.3 Rearranging

Such a representation of pairs of words naturally inspires an account of the joint occurrence of words in the text: under each word a bunch of first-level stars appear, on which the next level stars hang, and so on to any depth (the process ends naturally when the appropriate stars from the arsenal). It is possible, therefore, for each word of the text to take into account the weight of the stars hanging on it to any depth. At the same time, the new weight of the “first” word is calculated as the sum of the weights of the “second” words hanging on it:

$$w_i(t + 1) = \left( \sum_{i,i \neq j} w_i(t)w_{ij} \right) \sigma(\bar{E}) \quad (1)$$

where  $w_i(0) = p_i$ ,  $w_{ij} = p_{ij}/p_j$  и  $\sigma(\bar{E}) = 1/(1 + e^{-k\bar{E}})$ —a function that normalizes to the average energy of all vertices of the network  $E$ , where  $p_i$  is the frequency of occurrence of the  $i$ th word in the text,  $p_{ij}$  is the frequency of joint occurrence of the  $i$ th and  $j$ th words in text sentences.

For consistency of representation, the new values of the weight characteristics obtained at each iteration of the process are non-linearly normalized in the layer to the function  $\sigma(E)$ . In the program Text Analyst, such a renormalization is carried out 10 times (in 10 layers).

### 3.2.4 A Homogeneous Semantic (Associative) Network

The resulting semantic network is a semantic portrait of the text. In this case, the weight of the bonds remains unchanged. The weight characteristics of the vertices of the network transform. It can be noted that the greater the number of “second” vertices is that the “first” vertex is connected, the more its weight becomes to the detriment of the weight of the other vertices of this layer. The vertices with the highest weight characteristics are the key concepts of the text. They are all other peaks.

Such a network is useful in itself: it represents a means of associative text navigation. We can move from vertex to vertex (in the direction of word connections), moving from one concept to another through their connections in the text. It is possible to find the main peak in this network—the peak with the greatest weight.

### 3.2.5 Thematic Modeling

This main summit is undoubtedly the main theme of this text. But this means that all the vertices of this bunch of stars represent the table of contents of the part of the text associated with this main theme. Removing weak feedbacks and vertices with small weights we get the minimal tree subgraph of the semantic network, which turns out to be a thematic tree. In addition to the main vertex, it has child vertices of a lower level, which are subthemes of the main theme, and they have their own child vertices, which are interpreted as sub-subtopics.

It may happen that in the minimum tree subgraph there will be more than one main topic. Therefore, if you delete the root top—the main topic, the tree falls apart into several subtrees, each of which belongs to its main topic. This happens if the text contains several nonintersecting subject areas simultaneously.

It must be remembered that since the thematic analysis uses weights obtained by rearranging to several levels of network connections ( $n$  is actually equivalent to grammatic—syntagmatic connectedness of words in the text), this identification of the thematic structure is more correct than using only frequency analysis Automap and ORA) when using a unigram text model (bag of words).

## 4 Empirical Results

An empirical database of research comprised key publications on the digital economy. These publications were selected expertly. They cover 15 years of field development (in total 312 articles). Abstracts, annotations, and introductions of that articles were analyzed. These materials show the ideological vector of the development of the digital industry. The following characteristics of the text were expertly marked:

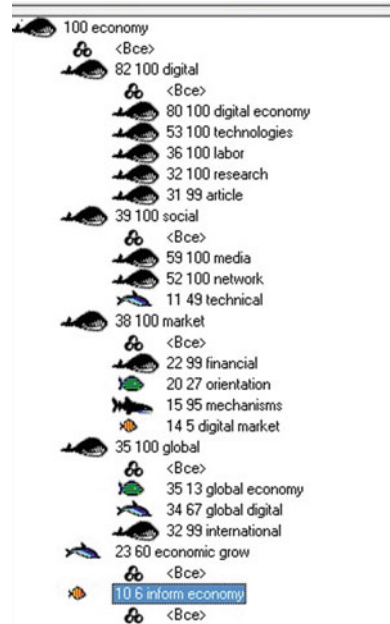
- related concepts (thematic areas) characterizing the text;
- keywords;
- abstract;
- introduction;
- authors;
- organization;
- bibliography;
- references.

All speech markers were subjected to the normalization procedure. Then this array of texts was subjected to two types of processing: first, the thematic ranking of speech markers was carried out using the program Text Analyst; secondly, the semantic network is built using the tools ORA and Automap.

The results of the Text Analyst, in the form of a thematic tree, are shown in Figs. 1 and 2. The higher the position of the speech marker in the tree structure, the higher its value, and the more “daughter” speech markers depend on this node.



**Fig. 1** The top of the thematic tree



In the subject tree, there are 5 main “root” clusters, each of them is characterized by the group of words.

The root word “digital” is described in speech markers and concepts “digital economy”, “technologies”, “labor”, “research”, “article”.

The root word “social” is defined by the speech markers and concepts “media”, “network”, and “technical”.

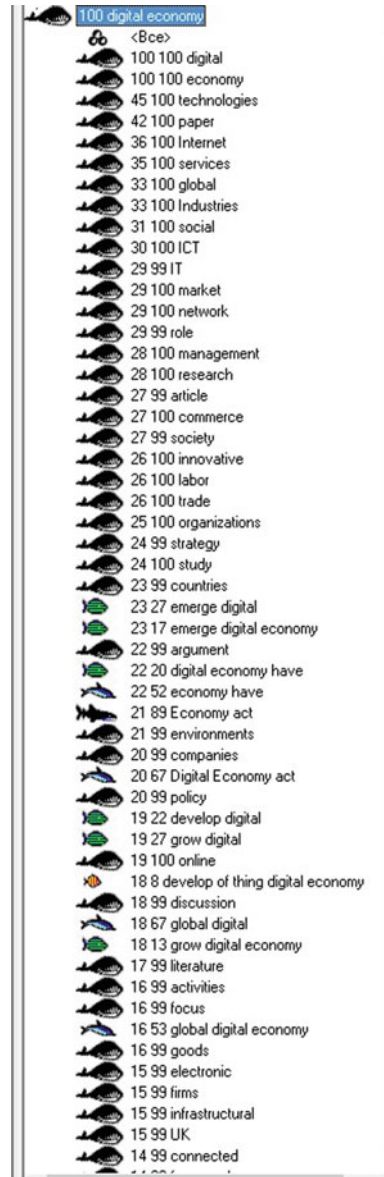
The root word “market” is revealed in speech markers and concepts “financial”, “orientation”, “mechanisms”, and “digital market”.

The root word “global” is expressed in speech markers and concepts “global economy”, “global digital”, and “international”.

According to the results of thematic modeling, there were fewer clusters than those that were selected on the basis of expert analysis. Perhaps the differences in the outcome of machine classification and manual are due to the fact that the Text Analyst classification combines the selected expert groups into larger ones. For example, the “social” thematic group is filled with two expert topics highlighted: “Media” and “Information Technology”. In addition, the words that are used in the titles of the thematic groups are less complex since the algorithm did not divide the words into parts of speech. If we compare the resulting groups with epistemic changes, the clusters themselves reflect a collective picture of the entire subject area, because there are words related to different epistemes.

The procedure for building a semantic network included several stages. Firstly, the deletion of stop-words, numbers, extra spaces, and punctuation in Automap, the network contained 15,059 words. Then the network was further cleaned in ORA: the

**Fig. 2** The nearest neighbors of the concept of “digital economy”



stop-words that remained after automatic processing, non-textual characters, adverbs and verbs were removed manually. Nouns have been reduced to the nominative case. Depending on the frequency of use of a word in the plural or in the singular, the forms of the same words were reduced to the plural or singular, thus the size of the network was 357 concepts. Further, the size of the network was reduced to 248

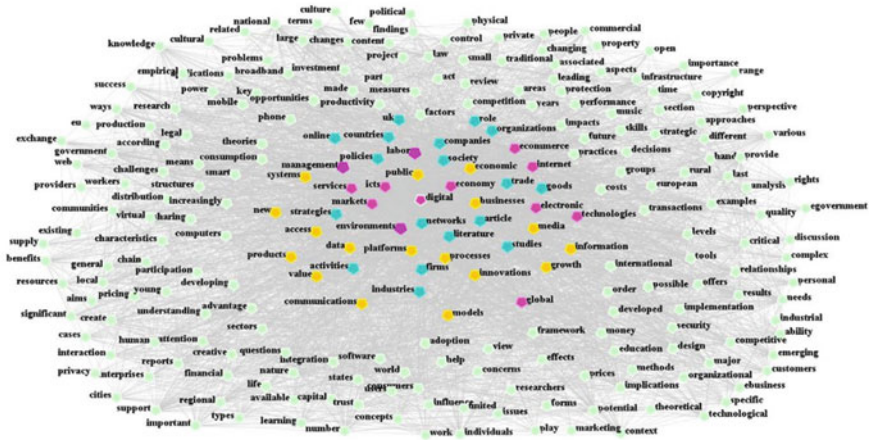


Fig. 3 Semantic network (ORA)

words. The minimum frequency for which the network size was reduced was 68 (with a maximum frequency of 1569 words).

The total network for the totality of all texts, visualized in ORA, is shown below in Fig. 3. The color-coding of speech markers—the vertices of the semantic network shows that the speech marker is in the top of the list—either in the subject tree or in the semantic network by the total degree:

- raspberry markers exist both in the thematic tree and the semantic network;
- yellow—speech markers, which exist only in the semantic network;
- blue—speech markers, which exist both in the thematic tree and in the semantic network, but have extremely low centrality measures.

Index “Total degree” is calculated by the formula:

$$C_{D(n_i)} = d(n_i) = x_{i+} = \sum_j x_{ij} = \sum_j x_{ji} \tag{2}$$

where,  $d(n_i)$  is the total number of links entering and leaving the node [39, p. 178].

In the center there are speech markers “digital”, “economy”, which are surrounded by several thematic segments.

The economic terms connected with the digital economy: business, market, trade, firm, companies, goods, value, etc.

A separate direction represents a new technological and communicative infrastructure: “Internet”, “e-commerce”, “electronic”, “technologies”, “platforms”, “online”, “networks”, “icts”, etc.

The consequence of such technologies is the new rules of information interaction: “media”, “public”, “communications”, etc.

Social terms related to the digital economy: “countries”, “society”, “policies”, “strategies”, etc.

Related to the digital economy management technologies: “management”, “services”, “organizations”, “role”, etc.

A separate direction is connected with the analysis of data: “data”, “models”, “processes”, “studies”, “access”, etc.

There are terms that can be attributed to emotional, ideological projecting: “innovations”, “growth”, “new”, etc.

It is significant that almost all the terms surrounding the concept of the digital economy are primarily associated with changes in the way of life, the economy, and public administration. Both in the thematic tree and in the social network there are practically no speech markers denoting real sectors of the economy. The only term that is tied to the real sector of the economy in the semantic network is “industries”.

After conducting two stages of text array analysis, a comparison of the results was carried out. The table below shows the centrality for all speech markers highlighted on the semantic network, as well as those included in the thematic tree. They are marked by color-coding, depending on the presence in the results of two types of analysis: a thematic tree from Text Analyst and a semantic network. The color-coding of speech markers is the same as on the semantic network.

In our network, some of the features of the digital economy that were originally incorporated into it have emerged. These are concepts of globalization and innovation. Innovations were identified as a significant term in the network, only on the basis of the analysis of the semantic network. Whereas globalization turned out to be a more important term. The concepts associated with the negative aspects of the phenomenon are not reflected. The color-coding of speech markers is the same as on the semantic network. In the Table 1: T.D.—total degree, A.—authority, B—betweenness, B.P.—Bonachich Power, C—closeness, E.V.—eigenvector.

## 5 Conclusion

The methods of grouping speech markers and visualization of their relative position in space are just the starting point for more substantial conclusions. The groupings of speech markers help to determine the subtopics into which the analyzed conceptual field is divided. The mutual arrangement of speech markers shows which concepts subthemes consist of, which are more central—key, and which are auxiliary.

It is especially exciting to look at the corpus of texts representing the theoretical direction in terms of structurally informative visualization. If the central speech markers are not supported by the operational concepts, the theory which is studied has no way out to the practical level.

In the light of the universal digitalization of the economy, the concept of “digital economy” was extremely interesting to consider from the angle of “theoretical-practicality”. An empirical study of the corpus of fundamental theoretical texts on the digital economy revealed its dual nature. Of course, this direction is more theoretical. The place of operational concepts is occupied by concepts marking market relations and consumption. There are practically no concepts representing social

**Table 1** Centrality measures of speech markers on the topic of “digital economy”

Concept	T. D.	A.	B.	B.P.	C.	E. V.
digital	1.000	-	0.029	1119.000	-	1.000
economy	0.699	1.000	0.037	229.000	0.532	0.984
businesses	0.337	0.011	0.023	365.000	0.469	0.020
services	0.239	0.025	0.012	82.000	-	0.032
Internet	0.226	-	0.031	214.000	0.493	0.030
markets	0.223	0.011	0.030	146.000	0.494	0.029
economic	0.195	-	0.020	200.000	0.468	-
technologies	0.192	0.133	0.016	-	0.470	0.136
ICT	0.164	-	0.016	132.000	0.458	-
models	0.147	-	0.009	-	-	-
media	0.141	0.032	0.010	144.000	0.456	0.059
global	0.139	-	0.015	154.000	0.472	0.063
public	0.136	0.006	-	130.000	-	-
processes	0.126	-	-	-	0.461	-
innovations	0.113	-	0.015	85.000	-	-
electronic	0.110	-	-	125.000	-	-
systems	0.110	-	0.012	-	0.454	-
platforms	0.107	0.033	-	-	-	0.036
labor	0.105	0.050	-	81.000	-	0.053
commerce	0.104	-	0.011	73.000	-	-
growth	0.104	-	-	83.000	0.478	0.024
communication	0.104	0.027	-	79.000	-	0.039
access	0.104	-	0.017	90.000	0.466	-
data	0.101	0.018	-	-	-	0.020
environments	0.100	0.024	-	-	-	0.026
management	0.098	-	-	92.000	-	-
value	0.098	-	0.016	80.000	0.459	-
new	0.095	-	-	-	-	-
IT	-	-	-	-	-	-
network	-	0.016	-	-	-	0.017
role	-	-	-	73.000	0.453	-
research	-	-	-	-	-	-
article	-	-	-	-	-	-
society	-	0.028	0.010	-	0.453	0.044
trade	-	-	-	-	-	-
organizations	-	-	0.017	-	0.459	-
strategy	-	0.021	-	-	-	0.025
study	-	-	-	-	-	-
countries	-	-	-	-	-	-
companies	-	-	0.010	-	-	-
policy	-	-	-	-	-	-
literature	-	-	-	-	-	-
activities	-	-	-	-	-	-
goods	-	0.020	-	-	-	0.022
firms	-	-	-	-	-	-
UK	-	-	-	-	-	-
industries	-	-	-	-	-	-

and production trends. The fact that these findings are confirmed in both methods used (thematic modeling using neural networks and structuring with the help of semantic networks) speaks about the immanent-ideological nature of the direction of the digital economy. Actually, this is confirmed at the start, when the author of the concept of “digital economy” Tapscott does not give its clear direct definition. It gives only a number of indirect characteristics, each of which is also difficult to operationalize [1], however, half of them are widely used as self-justified value: “Knowledge”, “Digitization”, “Virtualization”, “Innovation”, “Globalization”. The concepts of “Digitization”, “Innovation”, “Globalization” in the media space have already become independent ideological directions.

The study of linguistic structures, conducted in several ways of structural analysis of the text, revealed the semantic priority of the directions in the thematic field. Sub-themes were highlighted: economic, social, new technological and communication infrastructure, information interaction, management technologies, data analysis. In these areas, approximately the same results were obtained in the construction of a thematic tree (using the Text Analyst program) and in the construction of a semantic network (using the Automap and ORA).

The construction of a social network showed a specific result that highlighted the ideological component that accompanies the term “digital economy”—emotional, socially programming markers. But, since these terms had centralities somewhat smaller than the term “digital economy”, one can accept the null hypothesis that the digital economy is an ideological term rather than an instrumental one.

Thematic modeling showed results differing from expert thematic and epistemic groupings, which may be associated with a small number of articles and a plurality of topics identified by experts. Thematic modeling takes into account the volume of words included in the cluster. Small groups of words that differ from the majority in meaning can be mixed with larger groups for consistency reasons. It is likely that if we select an equal number of articles for each expert group, the results of thematic modeling will more closely correspond to the theoretical field classification.

The results, approaches, and methods of this study can be used when considering other large thematic fields devoted to new ideological trends that are being developed as an element of social design and management.

**Acknowledgements** The study has been funded by the Russian Academic Excellence Project “5-100”.

## References

1. Tapscott, D.: *The Digital Economy: Promise and Peril in the Age of Networked Intelligence*, vol. 1. McGraw-Hill, New York (1996)
2. Samuelson, P.: Intellectual property and the digital economy: why the anti-circumvention regulations need to be revised. *Berkeley Technol. Law J.* 519–566 (1999)
3. Orlikowski, W.J., Iacono, C.S.: The truth is not out there: an enacted view of the ‘digital economy.’ *Underst. Digit. Econ. Data Tools Res.* 352–380 (2000)

4. Nalebuff, B.J., Brandenburger, A.M.: Co-opetition: competitive and cooperative business strategies for the digital economy. *Strategy Leadersh.* **25**(6), 28–33 (1997)
5. Gee, J.P.: *An Introduction To Discourse Analysis: Theory and Method*. Routledge (2004)
6. Kharlamov, A.A., Yermolenko, T.V., Zhonin, A.A.: Modeling of process dynamics by sequence of homogenous semantic networks on the base of text corpus sequence analysis, pp. 300–307. Springer (2014)
7. Hofmann, T.: Probabilistic latent semantic analysis, pp. 289–296. Morgan Kaufmann Publishers Inc. (1999)
8. Van Dijk, T.A.: *Discourse and power*. Macmillan International Higher Education (2008)
9. Berenskoetter, F.: Approaches to concept analysis. *Millenn. J. Int. Stud.* **45**(2), 151–173 (2017)
10. Wei, X., Zeng, D.D., Luo, X.: Concept evolution analysis based on the dissipative structure of concept semantic space. *Future Gener. Comput. Syst.* **81**, 384–394 (2018)
11. Nasr Azadani, M., Ghadiri, N., Davoodijam, E.: Graph-based biomedical text summarization: an itemset mining and sentence clustering approach. *J. Biomed. Inform.* **84**, 42–58 (2018)
12. Wang, Y., et al.: Formal ontology generation by deep machine learning, pp. 6–15. IEEE (2017)
13. Ngom, A.N., et al.: A method to validate the insertion of a new concept in an ontology, pp. 275–281. IEEE (2016)
14. Bogumił, Z.: *Gulag Memories: The Rediscovery and Commemoration of Russia's Repressive Past*, 248 pp. Berghahn Books (2018)
15. Kostakis, V., Roos, A., Bauwens, M.: Towards a political ecology of the digital economy: socio-environmental implications of two competing value models. *Environ. Innov. Soc. Transit.* **18**, 82–100 (2016)
16. Salamon, E.: E-lancer resistance: precarious freelance journalists use digital communications to refuse rights-grabbing contracts. *Digit. J.* **4**(8), 980–1000 (2016)
17. Ciocou, C.N.: Considerations about intellectual property rights, innovation and economic growth in the digital economy. *Econ. Ser. Manag.* **14**(2), 310–323 (2011)
18. Doukidis, G.I.: Introduction to the special issue. In: Doukidis, G.I. (eds.) *Developing the Business Components of the Digital Economy*, vol. 3, pp. 3–6. M E Sharpe Inc. (1999)
19. Poon, S., Swatman, P.: A longitudinal study of expectations in small business internet commerce. *Int. J. Electron. Commer.* **3**(3), 21–33 (1999)
20. Bajaj, K.K.: Asia's leap into e-commerce: analysis of developments in some countries. *Prometheus U. K.* **19**(4), 363–375 (2001)
21. Dunn, S.: Micropower. *J. Corp. Citizsh.* (32), 42 (2000)
22. Sui, D., Rejeski, D.: Environmental impacts of the emerging digital economy: the e-for-environment e-commerce? *Environ. Manage.* **29**(2), 155–163 (2002)
23. Petrovic, O., et al.: Vertrauen in digitale Transaktionen. *Wirtschaftsinformatik.* **45**(1), 53–66 (2003)
24. Vătuui, T., Popeangă, V.: The utilization of information and communication technologies in educational area. *Ann. Univ. Petrosani Econ.* **6**[object Attr], 199–206 (2006)
25. Ainsworth, J.D., Buchan, I.E.: e-labs and work objects: towards digital health economies. In: Mehmood, R., et al. (eds.) *Communications infrastructure. Systems and applications in Europe*, vol. 16, pp. 205–216. Springer Berlin Heidelberg, Berlin, Heidelberg (2009)
26. Grindrod, C.B.E.P.: Mathematical modelling for the digital society. *IMA J. Appl. Math.* **76**(3), 475–492 (2011)
27. Hanson, V.L.: Influencing technology adoption by older adults. *Interact. Comput.* **22**(6), 502–509 (2010)
28. Salvado, J.: Travel experience ecosystem model: building travel agencies' business resilience in Portugal [Electronic resource] <http://hdl.handle.net/10400.8/445> (2011). Accessed 21 Mar 2017
29. Shade, L.R.: 'Give us bread, but give us roses': gender and labour in the digital economy. *Int. J. Media Cult. Polit.* **10**(2), 129–144 (2014)
30. Zhang, L., Fung, A.Y.: Working as playing? Consumer labor, guild and the secondary industry of online gaming in China. *New Media Soc.* **16**(1), 38–54 (2014)

31. Petit, N.: Technology Giants, the Moligopoly Hypothezis and Holistic Competition: A Primer (2016)
32. Ettliger, N.: Open innovation and its discontents. *Geoforum* **80**, 61–71 (2017)
33. Larson, C. (2015). Live publishing: the onstage redeployment of journalistic authority. *Media Cult. Soc.* 0163443714567016 (2015)
34. Al-Khouri, A.M.: Digital identity: transforming GCC economies. *Innovation* **16**(2), 184–194 (2014)
35. Czifrovaya Rossiya. Novaya real 'nost'. McKinsey (2017)
36. Global human capital trends the rise of the social enterprise. Deloitte (2017)
37. «Rossiya 25: ot kadrov k talantam». Boston Consulting Group (2017)
38. Blei, D.M.: Probabilistic topic models. *Commun. ACM* **55**(4), 77–84 (2012)
39. Wasserman, S., Faust, K.: *Social Network Analysis: Methods and Applications*, 852 pp. Cambridge University Press (1994)



# Methodology for Measuring Polarization of Political Discourse: Case of Comparing Oppositional and Patriotic Discourse in Online Social Networks



Tamara Shcheglova, Galina Gradoselskaya and Ilia Karpov

**Abstract** The paper analyzes speech markers and semantic concepts typical for patriotic and oppositional discourse in social networks. About 100 000 posts from Facebook, VKontakte, and LiveJournal were analyzed, and 35 000 most frequent speech markers were processed, of which 1800 markers were selected for analysis. The alternative method to TF-IDF metric for specific text markers identification is proposed. The features of oppositional discourse in comparison with the patriotic discourse were formulated. On the one hand, the analysis of sets of speech markers that characterize political groups allows us to understand social models and attitudes embedded in the discourse and the subsequent behavior of representatives of these groups. On the other hand, it is possible to extend a set of keywords for text search of a certain political orientation, based on the obtained results.

## 1 Introduction

The Internet space and social media have a dual nature. On the one hand, it is as a structural formation, where actors (persons, groups, pages, etc.) are connected by information flows and social ties. On the other hand, information flows from a kind of general discursive space, where speech markers merge into higher level concepts. Speech markers and concepts have a significant semantic load. They have a social function by implementing models of social influence and manipulation like “us-them” model [15], and ideological function by demonstrating values, projecting models of the future, etc.

Therefore, another view on the space of social networks is possible—as a constructed space of meanings, which is a generalized reflection of the discourse of social groups that influences real sociopolitical processes. A similar view on the role of discourse and communicative space is presented in the works of Lotman

---

T. Shcheglova (✉) · G. Gradoselskaya · I. Karpov  
National Research University Higher School of Economics, 101000 Moscow, Russian Federation  
e-mail: [tshcheglova@hse.ru](mailto:tshcheglova@hse.ru)

© Springer Nature Switzerland AG 2020  
I. Bychkov et al. (eds.), *Network Algorithms, Data Mining, and Applications*, Springer Proceedings in Mathematics & Statistics 315,  
[https://doi.org/10.1007/978-3-030-37157-9\\_15](https://doi.org/10.1007/978-3-030-37157-9_15)

[10], Makeeva [11], Tsyvyan [16]. We should also mention the founders of this approach—the classics of the Geneva linguistic school: Saussure [14] and Bally [1].

Our study will show key speech markers and semantic concepts that characterize the discursive space of opposing groups in politics on social networks. One group represents a pro-government position, which in the modern political field is defined as “patriotic”. Another group represents an oppositional position.

It is necessary to emphasize that the terms “oppositionists” and “patriots” in our study are conventionally accepted, by the principle of self-identification by representatives of these groups, and by their labeling of groups of opponents. That is such identification is also a derivative of “collective intelligence” in a discursive Internet space.

The aim of the study is to determine the speech markers and concepts peculiar for the identified political groups—the bearers of certain political attitudes (in our study—“patriotic” and “oppositional”).

The results of the research can have both scientific-methodological and applied significance. To identify speech markers that characterize the discourse of political groups, we developed a special method. The analysis of sets of speech markers that characterize political groups is independent value, it allows us to understand social models and attitudes embedded in the discourse and the subsequent behavior of representatives of these groups. The applied value of the research is that “reverse search technology” is possible—texts search of a certain political orientation or designed social processes (for example, protest actions, strikes, pickets, etc.) according to established specific speech markers.

Solving the problem of comparing the speech behavior of two different political groups required the development of special tools. A familiar linguistic tool for most significant words search showed uninterpreted results on our corpus of texts. Perhaps the peculiarity of the object of our research—two large politically opposed corpora of texts—was not taken into account. Therefore, we had to develop our own method of text polarization and highlight keywords that mark this differentiation—discourse differentiation index.

## **2 Literature Review of Discourse Research and Allocating Key Speech Markers**

### ***2.1 Theoretical Background for the Study of Discourse***

In critical discourse analysis (CDA), discourse is much more than a sequence of linguistic signs and symbols. Even more, discourse is a multidimensional substance, that includes texts itself, discursive practice, and sociocultural practice. It is a text as it is the product of language. It is discursive practice because it is associated with an established type of discourse connected to a particular kind of activity. And it is

sociocultural practice so as it explains the relationship between discursive and social processes [5].

T. van Dijk pays special attention to the functioning of the language in the mass media. Van Dijk examines the impact of sociocultural factors on the mechanism of language use. An important component of the general theory of communicative-linguistic interaction, according to van Dijk, is the cognitive theory of language use, which not only give access to the processes and structures that provide cognitive processing of sentences and statements, but also explains how planning, production, and understanding of speech is happening [2].

Van Dijk adopts the idea of presenting positive-self and negative others by using specific speech markers in discourse. He studies the strategies of foregrounding positive practices of oneself and de-emphasizing any positive aspect of the other [3].

## 2.2 *Studies of Political Language Features*

Politics is a struggle for power in order to achieve certain political, economic and social goals. The analysis of political discourse should treat discourse as an instrument of doing politics. In this context, language plays a significant role since every political action is born, prepared, controlled, influenced, and performed by language [7]. Internet and social media have dramatically changed the study of political communication as researches access massive feeds of data on online social media behavior, networks and language [6].

In one research author investigated ideological structures of polarized discourse coded in the reports of two online news websites: *egyptindependent* and *Ikhwanweb*. The author found out features of the ideologies of polarized discourse and concluded with a discussion on how both websites establish a dichotomy of “we” versus “them” [4].

In the other research, Obama’s political discourse is investigated. The authors oppose liberal discourse to conservative and highlight its main features. Concepts of *freedom* and *justice* constitute liberal discourse in the US. *Freedom* is defined as the social and political rights of individuals that protect them from interference by others in their lives. *Justice* is understood in terms of equal rights and the end of oppression in the social world [7].

## 2.3 *Linguistic Features of Measuring Distinctive Words*

It is common practice in computational linguistics to model documents by the words that have been weighted by their term frequency–inverse document term (TF–IDF). It has been the most commonly adopted document representation method for various text-processing tasks. It provides a weight to each word in a document according to the frequency of its occurrence in text and the rareness of its use in the other

documents on the corpus of texts. TF-IDF metric works on the basis of a bag of words, which involves the assumption that the document is simply a collection of words and a vector can be computed by estimating the relative distance between words [9]. TF-IDF reliably captures what is distinctive about a particular document and it could be interpreted as a feature evaluation technique. According to the logic of this approach, most distinctive words are the ones spoken by one party and not spoken once by the other [13].

The problem with that metric is that it allows us to pick out distinctive but not widely used words. It is also should be noted that the standard linguistic approach ignores nonstandard word use, considering them marginal and erroneous, while the cognitive approach allows interpreting nonstandard uses as specific operations on knowledge. Thus, it becomes possible to detect the hidden intentions of the speaker [8].

### 3 Method for Differentiating Speech Markers

#### 3.1 Data Description

To understand the quantitative trends in the discourse of patriots and oppositionists, the corresponding publications in social networks were investigated. The study was conducted in October–November 2015. At the first stage, 230 patriotic and 240 oppositional resources were expertly selected in three social networks: Facebook, VKontakte, and LiveJournal.

The resources were groups and open pages, from which about 100 000 posts were downloaded (over the previous 6 months). Based on the downloaded frequencies of texts (speech markers) for each political group (patriots and oppositionists) were counted, as well as the total frequency for all groups. 35 000 most frequent speech markers were processed: all indexes, marking the peculiarities of the discourses of patriots and oppositionists were calculated. The final basis for analysis, containing socially significant and differentiating discourses of patriots and oppositionists, was over 1800 words. Traditionally, the metric TF-IDF is used to identify specific text markers or subsamples [12]. However, according to the data obtained in the study, this metric showed not very adequate, rare words (with a frequency of about 5–10 words in the entire array). Perhaps this result would be acceptable for linguists, but for our purposes (further use of markers for targeted material search and classification of texts) such a result will not be relevant. Therefore, to determine specific speech markers we developed an alternative method for differentiating speech markers.

### 3.2 *Characteristics of Frequency Distributions*

In the database uploaded from the primary data of social networks, there is a list of speech markers that are used in posts and comments of patriotic and oppositional resources. Speech markers were counted after the normalization of texts (putting the words in the nominative case of the singular).

For each speech marker, the baseline data was calculated as the initial data, which were subsequently used to calculate differentiation indices:

- total frequency of use (in all texts);
- frequencies for each group of texts (patriotic and oppositional);
- percentage of occurrence of a given speech marker in the entire discourse (relative frequency);
- relative frequencies of use for each group of texts (normalization is carried out by dividing the frequency of the speech marker by the total volume of the discourse in this group of texts).

### 3.3 *Methodological Issues*

After weighing the words with TF-IDF, metric large absolute frequencies and small relative frequencies of speech markers did not allow to objectively compare the prevalence of speech markers in a particular discourse. Therefore, there was a need for developing an indicator (or a system of indicators) that shows the predominance of the speech marker in a particular discourse, as well as a general indicator that allows to identify key speech markers polarizing the discourse of groups of two different political orientations.

Here we propose the system of indicators that measure the difference between word use in two discourses. That system consists of two basic indices and one final index (as a combination of two basic ones): *PO Index*, *OP Index*, and *Total Index*.  $WF_{\text{patriotic}}$  is the number of occurrences of the speech marker in the patriotic discourse and  $WF_{\text{oppositional}}$  is the number of occurrences of the speech marker in the oppositional discourse:

- *Index of prevalence of patriotic discourse over the oppositional (PO Index)* is calculated as the ratio of the relative frequency of the occurrence of the speech marker in the oppositional discourse to the patriotic discourse. This index shows how many times the word prevails in the oppositional discourse in relation to the patriotic:

$$PO\ Index = \frac{WF_{\text{patriotic}}}{WF_{\text{oppositional}}} \quad (1)$$

- *Index of prevalence of oppositional discourse over patriotic (OP Index)* is calculated as the ratio of the relative frequency of the occurrence of the speech marker in the patriotic discourse to the oppositional one. It shows how many times this word prevails in patriotic discourse in relation to the opposition:

$$OP\ Index = \frac{WF_{oppositional}}{WF_{patriotic}} \quad (2)$$

- *Index of differentiation of speech markers between discourses (Total Index)* is calculated as the square root of the difference between the *PO Index* and the *OP Index*. It shows the degree of discrepancy between the usage of the given word in different discourses—patriotic and oppositional. Usually, the size of the index is 1 or more (which corresponds to the predominance of the word in some discourse more than twice):

$$Total\ Index = \sqrt{(PO\ Index - OP\ Index)^2} \quad (3)$$

## 4 Practical Results of the Study

Key speech markers common to all political groups (speech markers ranked in descending order of the total absolute frequency) are shown in Table 1. The importance of key geopolitical concepts for patriots and oppositionists coincides.

In all discourses, there are the country names—in the present and past tense: “Russia”, “RF” (Russian Federation), “country”, “state”, “USSR”. Equally significant are the references to the people of the country: “people”, “population”, “national”. Also, there are references to major international actors: “West”, “European”, “International”.

The name of the Crimea peninsula after the events of 2014 can be designated as situational speech markers. This event became significant in both discourses—oppositional and patriotic.

It is significant that in the second and third places there are speech markers “us” and “them”, which indicates the prevalence of the model of sociopolitical differentiation “us-them” [15] in the discourses of all political groups.

Key speech markers of patriotic discourse (the speech markers are first selected from the most frequent words, and then ranked in descending order of the *PO Index*—the predominance of patriotic discourse over the oppositional discourse) are shown in Table 2.

In the patriotic discourse, the first place takes situational lexicon which is associated with the events at the time of the research (autumn 2015) taking place in

**Table 1** Key speech markers common to all political groups

Speech marker	Total frequency	OP index	PO index	Total index
Russia	478 924	1.025	0.976	0.048
Them	458 490	1.018	0.982	0.036
Us	397 217	1.065	0.939	0.125
Country	234 049	1.148	0.871	0.277
World	130 245	0.877	1.140	0.263
State	92 923	1.163	0.860	0.302
People	85 771	1.228	0.815	0.413
RF (Russian Federation)	84 449	1.142	0.875	0.267
History	69 487	0.923	1.083	0.160
Politics	63 348	1.093	0.915	0.178
Crimea	60 819	0.811	1.233	0.421
West	57 781	1.187	0.843	0.344
Worldwide	57 675	0.933	1.072	0.139
Victory	51 499	0.772	1.295	0.523
Government	51 119	1.191	0.840	0.351
Western	50 814	0.779	1.284	0.504
USSR	49 706	0.804	1.244	0.441
International	49 678	1.076	0.929	0.146
Population	46 188	1.062	0.941	0.121
European	42 081	0.912	1.097	0.185
National	40 162	0.976	1.025	0.049

the southeast of Ukraine: “Donetsk”, “DPR”, “Novorossia”, “Donbass”. Also, ideological opponents of different levels are recalled: “Poroshenko”, “American”. Military terms predominate: “battle”, “combat”, “tank”, “military”, “front”, “defense”, “troops”, “army”, etc.

The features of patriotic discourse can be formulated as follows:

- Discourse is directed to the past: the history of the Soviet Union, its achievements, victories are recalled;
- Discourse is militarized: the names of weapons and military terms prevail;
- In the patriotic discourse situational speech markers devoted to current events in Ukraine and Donetsk.

Key speech markers of oppositional discourse (speech markers are first selected from the most frequent words, and then ranked in descending order of the OP Index (the predominance of oppositional discourse over patriotic) are shown in Table 3.

In addition to the high-frequency words reflected in Table 3, that characterize the oppositional discourse, there are less frequent, but very popular terms (more than 300 in the subsample) that can also be grouped in meaning. A large semantic group of

**Table 2** Key speech markers of patriots

Speech marker	Total frequency	OP index	PO index	Total index
Battle	64 366	0.099	10.117	10.018
Donetsk	37 140	0.119	8.409	8.290
Tank	42 179	0.149	6.691	6.542
Combat	67 959	0.162	6.166	6.003
Poroshenko	42 909	0.169	5.913	5.744
DPR (Donetsk People's Republic)	55 316	0.173	5.771	5.598
Novorossia	41 806	0.189	5.299	5.110
Fire	40 994	0.200	5.010	4.810
Enemy	42 848	0.202	4.952	4.750
Rocket	33 553	0.223	4.481	4.258
Troops	74 335	0.246	4.072	3.826
Army	93 353	0.253	3.951	3.698
Defense	42 233	0.301	3.327	3.027
Hero	38 318	0.308	3.250	2.943
Donbass	72 301	0.309	3.236	2.927
Front	37 562	0.319	3.131	2.812
Military	142 767	0.438	2.283	1.845
Ukraine	295 467	0.482	2.076	1.594
American	79 282	0.498	2.007	1.509
Force	111 356	0.586	1.707	1.121

speech markers that characterize power in Russia (for example, “Kremlin”, “federalism”, “clamp”, “kleptocracy”, etc.) and the head of the country (“VVP” (Vladimir Vladimirovich Putin), “putler”, etc.). Separately the supporters of power are characterized by “Kremlebot”, “troll”, “Edinaya Rossia” etc. And the information space of the country (“zombiebox”, “propaganda”, “pro-Kremlin”, “hurray-patriotism”, etc.). Specific names of state corporations, names of state officials, names of regions of the country that are in the zone of attention of the opposition are also listed. Also, there are specific persons of influence, resources of influence. The directions of the opposition’s actions are listed, and as usual the protest action (“appeal”, “petition”, “action”, “picketing”, “hunger strike”, “rally”, “procession”, “unauthorized”, etc.) and active action (“terror”, “violence”, “revolutionary”, “lustration”, “anarchist”, “ultra-right”, “bolotnyi” (after protest events in May 2012 on the Bolotnaya square), etc.).

The opposition resources are actively discussing the activities of large state corporations. The names of companies are often mentioned: “Rosbank”, “Gazprom-Media”, “Lukoil”, “Rosneft”, “VTB”, “RZD”. The context of the discussion concerns situations that are possible carriers of corrupt practices: government contracts, government procurement.



**Table 3** Key speech markers of oppositionists

Speech marker	Total frequency	OP index	PO index	Total index
Court	34 822	4.120	0.243	3.877
Oil	25 145	3.164	0.316	2.848
Kremlin	21 328	3.094	0.323	2.771
Ruble	47 952	3.040	0.329	2.711
Society	35 268	2.770	0.361	2.409
Price	41 361	2.756	0.363	2.393
Elections	30 797	2.743	0.365	2.378
Network	23 305	2.504	0.399	2.105
Action	25 981	2.346	0.426	1.920
Freedom	28 703	2.330	0.429	1.901
Law	44 680	2.295	0.436	1.859
Civil	31 019	2.283	0.438	1.845
Bank	26 569	2.255	0.444	1.811
Putin	135 541	2.083	0.480	1.603
Company (firm)	40 783	2.082	0.480	1.602
Social	26 581	2.043	0.490	1.553
Crisis	28 091	1.972	0.507	1.466
Sanction	42 667	1.762	0.567	1.195
Power	109 463	1.688	0.592	1.096
Article	33 171	1.634	0.612	1.022

Representatives of oppositional discourse actively link to resources—significant and respected for them sources of information: “SvobodaNews”, “Libernews”, “Opir”, “Rabkor”, “Open Russia”, “Forbes”, “TVrain”, “Grani”, “Snob”, “Slon”, “Novaya Gazeta”, “Echo”, “Obozrevatel”, “Vedomosti”, “Rosbalt”, “Transparency International”, “Inosmi”, “Kommersant”, “Meduza”, “Euronews”, “Interfax”. At the same time, there are no pro-government sources on the list.

Among the representatives of the oppositional discourse, a clear self-identification is built: political prisoner, dissident, dissenter. There are also target groups that are carriers of opposition views: Democrat, political prisoner, dissent, youth, student, intelligence. The social positioning of the opposition is accompanied by emotional speech marks: self-defense, protest, hybrid, anti-Putin. The actions of the authorities in relation to the opposition are characterized in such a way as to justify their own protest actions. They are characterized by the following negative markers: arbitrariness, prohibit, dispersal, discrimination, redistribution, illegal, police, censorship.

The features of oppositional discourse in comparison with the patriotic discourse can be formulated as follows:

- The discourse is more specific than the discourse of patriots—the names of modern politicians, ministries, and state corporations are much more common in use.
- Most modern economic terms predominate—the discourse of oppositionists claims a monopoly of scientific character and objectivity.
- Economic evaluation of the country’s future is depressing: “sanctions”, “oil”, “crisis”, etc.
- Legal terms prevail as a guarantee of the legal basis of political activity. The discourse of the oppositionists represents both legal terminology and prison slang (“pakhan” (crime boss), “zek” (convict), “skhod” (descendant of thieves), etc.).
- The terms that became the ideological norm in the 90s are fully used: “society”, “public”, “civil”, “freedom”, etc.
- Social technologies of manipulation are mentioned: “action”, “picket”, etc.

We can draw a general conclusion that oppositional discourse is the result of careful sociolinguistic reflection and social design. Patriotic discourse is formed spontaneously, there is no ideological basis and an organizational component of work with patriotically minded groups of the population. In general, patriotic discourse loses to the opposition on ideological and methodological grounds.

## 5 Conclusion

According to the results of the study, conclusions can be drawn in two directions: informative and methodological.

### 5.1 Informative Conclusions

Polarization of discourse is observed in the Russian political information space of social networks. There are two political groups opposing each other (the names are given according to their self-determination): “patriots” supporting the actions of the authorities, and “oppositionists” challenging the activities of the authorities. There are practically no overlapping and common topics for discussion between them; they live in an alternative, parallel social reality.

A comparison of these discourses allows us to say that the “patriotic” discourse is extremely poor in comparison with the “opposition” discourse, it is socially led. “Opposition discourse”, on the contrary, is active, it constantly updates the dictionary in accordance with the current sociopolitical situation. A priori advantage to the oppositional discourse is given by the presence of the dominant liberal-democratic ideology, the key concepts in Russian society, that considered as a basic value. Oppositional discourse projects social reality, and not only states events.

An extremely interesting result of the analysis of oppositional discourse is the identification of clear social projection techniques in verbal form. These are the

methods of building up the identification of a political group and its mobilization, the image of the enemy, the moral justification of their own protest actions, etc.

## ***5.2 Methodological Conclusions***

Classical linguistic approaches to the identification of key distinctive words are not always suitable for solving sociological problems. The reason is most likely in different objects of study. For linguistics, the focus of research is on texts, and for sociologists—social processes that are labeled or are accompanied by these texts. Therefore, the breadth of distribution of the identified keywords, their representativeness in a sociological sense is crucial. For sociolinguistic research, the main thing is the understanding of language as a participant in the social process both in a theoretical and applied sense. So, the found keywords will help to identify the manipulative, socially projective actions from the side of different political groups.

## ***5.3 Final Thoughts***

The proposed methodology makes it possible to identify speech markers specific to a particular discourse from an array of widely used words. The method of differentiation of speech markers can be used not only for analysis of oppositional and patriotic discourse but also for any other opposing social groups: for analyzing the discourses of different generations, nationalities, religions, movements, etc.

Identified words that differentiate opposed discourses (discourses of various social and political groups) can be subjected to additional types of statistical analysis and expert coding. It is possible to group differentiating speech markers according to the roles they perform in the overall sociolinguistic projection of the activity of the groups under study. For example, it could be ideological markers that build a group's identity, slogans that motivate proactive social actions, etc.

It is possible to include the proposed method in more complex types of sociolinguistic analysis, identify socio-projecting models hidden in the texts. The identification of such sociolinguistic models and manipulative techniques in radical social movements could help counteract the spread of these movements in society.

**Acknowledgements** The study has been funded by the Russian Academic Excellence Project “5-100”.

## References

1. Bally, S.: *Language and Life*. URSS, Moscow (2003)
2. van Dijk, T.: *Discourse and Communication*. Walter de Gruyter, Berlin (1985)
3. van Dijk, T.: *News as Discourse*. Hillsdale, NJ, Erlbaum (1988)
4. Eissa, M.: Polarized discourse in the news. *Procedia Soc. Behav. Sci.* **134**, 70–91 (2014)
5. Fairclough, N.: *Critical Discourse Analysis. The Critical Study of Language*. Longman, London (1995)
6. Gonawela, J.: Studying political communication on Twitter: the case of small data. *Curr. Opin. Behav. Sci.* **18**, 97–102 (2017)
7. Horvath, J.: Critical analysis of Obama's political discourse. In: *International Conference of Language, Literature and Culture in a Changing Transatlantic* (2009)
8. Issers, O.: *Communicative Strategies and Tactics of the Russian Language*. URSS, Moscow (2008)
9. Kim, D., Seo, D., Cho, S., Kang, P.: Multi-co-training document classification using various document representations: TF-IDF, LDA, and Doc2Vec. *Inf. Sci.* (2018)
10. Lotman, Y.: *The Semiosphere*. Art-SPB, Saint-Petersburg (2010)
11. Makeeva, L.: *Language, Ontology and Realism*. Publishing House of the Higher School of Economics, Moscow (2011)
12. Manning, C., Raghavan, P., Schütze, H.: Scoring, term weighting, and the vector space model. In: *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA (2008)
13. Monroe, B., Colaresi, M., Quinn, K.: Fightin' words: lexical feature selection and evaluation for identifying the content of political conflict. *Polit. Anal.* **16**(4), 372–403 (2008)
14. De Saussure, F.: *Works on Linguistics*. Progress, USSR (1977)
15. Shipilov, A.: Opposition "us–them" in sociocultural development. *Philos. Leg. Thought: Alm.* **5**, 280–304 (2003)
16. Tsivyan, T.: *Model of the World and its Linguistic Basis*. URSS, Moscow (2009)

# Network Analysis Methodology of Policy Actors Identification and Power Evaluation (The Case of the Unified State Exam Introduction in Russia)



Dmitry Zaytsev, Gregory Khvatsky, Nikita Talovsky and Valentina Kuskova

**Abstract** In this paper, we presented a methodology for identifying policy actors for policy fields and evaluate their power. Presented methodology is based on text parsing and mining, and producing networks with analysis of the text processing results. We used the example of the Russian Unified State Exam, as the real case of policy formulation and implementation, to test the proposed methodology. The methodology was shown to have great potential for verifying the theories of policy studies and for a broader application in the areas where analysis of policy actors and their power, influence, and impact is needed.

## 1 Introduction

The field of public policy studies is a relatively newly emerged discipline, established in the 1980s, which traces its history from Lasswell's pathos to established policy science as a domain of knowledge about democratic rational policy-making [1]. The main goal of public policy scholars is to study different public policy fields and subfields in search for regularities and laws in policy formulation and policy implementation processes [2].

During the past few decades, the interest in policy processes generated a number of formal theoretical explanations. Among them are the Advocacy Coalition framework [3], Punctuated equilibrium theory [4], Multiple streams approach [5], Policy styles theories [6], Policy Design and Policy Capacity framework [7],

---

D. Zaytsev (✉) · G. Khvatsky · N. Talovsky · V. Kuskova (✉)  
National Research University Higher School of Economics, Moscow, Russian Federation  
e-mail: [dzaytsev@hse.ru](mailto:dzaytsev@hse.ru)

V. Kuskova  
e-mail: [vkuskova@hse.ru](mailto:vkuskova@hse.ru)

G. Khvatsky  
e-mail: [gkhvatsky@hse.ru](mailto:gkhvatsky@hse.ru)

N. Talovsky  
e-mail: [ntalovsky@hse.ru](mailto:ntalovsky@hse.ru)

© Springer Nature Switzerland AG 2020  
I. Bychkov et al. (eds.), *Network Algorithms, Data Mining, and Applications*, Springer Proceedings in Mathematics & Statistics 315,  
[https://doi.org/10.1007/978-3-030-37157-9\\_16](https://doi.org/10.1007/978-3-030-37157-9_16)

Pragmatic Approach to Public Policy [8], and some others. Each of them makes its own unique contribution, explaining the multidimensional and complex nature of public policy and policy changes. In doing so, they attempt to grasp the multi-actor nature of policy-making. The variety of actors in public policy is reflected in such terms as policy communities, policy coalitions, and, finally, policy networks. The results of policy-making are dependent on the activity and configuration of such policy networks and various conditions, or factors, that are situational, depending on configuration of certain combinations. Social Network Analysis, as a methodological framework, provides the terms, methods, and quantitative statistical techniques that allow us to model the complicated processes for a given policy, and provide an opportunity to and develop theories of policy processes that go above and beyond what other instruments provided. At the present time, however, public policy studies lack empirical research—verification and validation of developed theories.

One of the key research problems in public policy studies as well as applied policy practices is to define and analyze main policy actors in the field. The following step after you define the key policy actors in some policy field is to evaluate their influence or impact on policy changes. Traditional question for political scientists: Who governs—is transformed into the more relevant for public policy studies and decision-making practice: who can influence, let us say, education policy, how they can influence, and what is the (short-) long-term effects from influential policy actors activity in a particular policy field such as education or innovation policies. And the following questions become particularly important both for policy practice and for the public policy theories development and their verification: Why one policy actors succeed in impacting policy-making, and others are not?

To define who the main policy actors in the policy field is a tricky and not obvious question. Usually scholars who are specialized in a particular policy field as they know the subject pretty well, they can provide the list of policy actors who are influential in the field without any difficulties. Doing qualitative research, it is not a problem and is a domain of the scholar. In quantitative research, we know the classical studies of Robert Dahl [9] where he defined the list of power-holders in the city to conduct further research of power distribution at the local level. The method he proposed was based on the survey of the experts (opinion leaders, journalists, and other experts in local politics), who produce the initial list of decision-makers in the city. In both cases, the construction of the list of important policy actors in a particular policy field is based on expert opinions.

The experts' capacity to define policy actors important to the particular policy field can be challenged by several problems. First is how do experts define "importance"? Is it a formal position as a criteria to define important policy actors, his or her reputation, or activity in diverse stages of the policy process? Each criterion defines only one type of policy actors: formal one (based on position), latent (based on reputation), and public (based on activity). Hence, second, the proper mixture of criteria becomes a real challenge for the research design of such an expert survey. Third, when we ask experts, we receive declarations that can be far from the real situation and suffer from the experts' biases. When we ask experts to assess power, influence, or impact of policy actors we faced with the same or even more problems connected

with the high complexity and multidimensionality of the power as social phenomena. Therefore, it is very hard to measure the power of policy actors quantitatively, and both qualitative and quantitative attempts will be hardly free from the subjective opinions of experts which survey are very hard to avoid for studying policy actors' impact in policy-making.

As a result, it will be wonderful to have a more objective, opinions-free method to define policy actors of a particular policy field and have more unbiased technique to measure their power in its various manifestations. In this paper, we present a methodology for policy actors' identification and evaluation of their power (influence and impact). This methodology is based on text-parsing, text-mining, and network analysis. For application, we choose education policy and the case of introduction of the Unified State Exam (USE) in Russia.

The USE is an obligatory nation-wide test for all contemporary school-leavers in Russia. Based on this test, school-leavers are automatically certified as being finished the contemporary school and can be enrolled in the universities. The idea was first introduced in 1990s as a way to get rid of the outdated "soviet" system when pupils were forced to pass school exams to finish it, and enrollment exams to the university, which was a great stress for families, accompanied with bribes and corruption as the exams were mostly oral with results mostly depended on the examiner. In addition, the USE was supposed to become a policy tool to fight inequality in access to higher education of two kind: first, when pupils from reach families have more opportunities to enroll in the best Russian universities then the school-leavers from middle-class and below middle-class families; and, second, when pupils from the province have less opportunities to enroll the leading institutions of higher education than students from Moscow and Saint Petersburg.

By producing the network of policy actors related to the introduction of the USE, where links are the joint appearance of the policy actors in the mass media texts, and analyzing the network characteristics of the network, we are demonstrating a new method for identification and power evaluation of policy actors in a particular policy field. This methodology is universal and can be expanded and used to identify and assess policy actors in a variety of policy fields.

## 2 Methodology

The chosen unit for data collection and analysis were texts of news. To collect them we used well-known Russian agency specialized in media monitoring and analysis "Medialogia". It is a developer of the automatic system for monitoring and analysis of mass media. It covers 47,400 mass media from Russia and foreign countries. The database of Medialogia includes all types of mass media, including TV, radio, newspapers, magazines, websites, etc. We used the Citation Index, developed by Medialogia, which is an indicator of the quality of the media content distribution. It takes into account: (1) the number other media referenced the source article, and the authoritativeness of the publisher of the source; (2) social impact of the mass

**Table 1** Number of parsed articles per year

Periods for analysis	Number of parsed news articles
2001	134
2003	1408
2007	4692
2011	10,000
2016	10,000
Total	26,234

media (amount of likes and shares in social media). The index is calculated using mathematical and linguistic analysis of texts from 47,400 sources and 800 million social media accounts. To avoid self-citation, cross-references from media owned by the same company are not analyzed. The traffic, circulation and audience are not analyzed as well [10]. Medialogia has one of the deepest archives—about 20 years.

As the Unified State Exam was formulated as a policy for the new President of Russia Vladimir Putin in 2000 and appeared in the governmental documents of 2000 and 2001, the year of 2001 was chosen as a starting point for our analysis. We decided to analyze news by years as the formulation and implementation of the USE came through the turbulence of political process which influences on the adoption and further development of the USE very much. The USE became one of the most contentious topics and the struggle is that its future was very much associated with electoral cycles, and its adoption and implementation with the victory of liberal parties and politicians [11]. Therefore, our sample of years relates to the beginning of electoral cycles or elections in the Russian parliament—the State Duma. Consistently these are 2003, 2007, 2011, and 2016. The news were collected for those 5 years. To collect them we used the following keywords or request to the Medialogia system: “edinyi gosudarstvenyi ekzamen”, which means “unified state exam”. The numbers of collected texts of news by years are presented in Table 1. Overall, we have collected 26,234 news articles related to the implementation of the Unified State Exam in Russia.

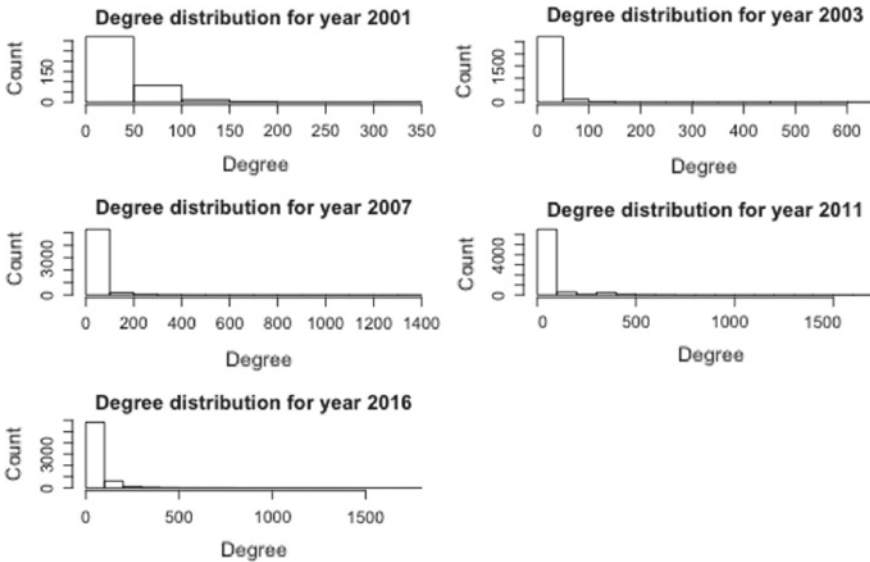
From the data analysis of the collected news articles, we expected to receive the list of policy actors related to the introduction of the USE and measures of their importance. To achieve these goals, we decided to separate the data analysis process into two distinct steps: exploratory and confirmatory. At exploratory part, we planned to obtain the list of policy actors related to the introduction of the USE by machine recognition of the names in all 26,234 news articles separated by years. For each year, we have used a named entity recognition library “Natasha” to identify policy actors mentioned in every article. For the numbers of received names see Table 2.

We constructed graphs based on the mentions of the names. To count these graphs, we used NetworkX library. “NetworkX is a Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks” [12]. It allows to manipulate graphs in particular to create them, add and delete nodes



**Table 2** Numbers of recognized names per year

Period of analysis	Names/Nodes	Edges
2001	417	6557
2003	2869	22,819
2007	5525	80,039
2011	7198	164,848
2016	6680	145,578
Total	22,689	419,841



**Fig. 1** Degree centrality distribution for the networks of policy actors by years

and edges, read and save in files. Also, it allows to count network statistics, but for this we used Pajek as it is faster. In resulting networks, the nodes were individuals mentioned in the news articles. The nodes had a link between them if two individuals were mentioned in one article, its weight being the number of such articles.

Then, we calculated a certain descriptive statistics for the five networks we have obtained such as degree and betweenness centrality. The distribution of degree centrality is shown in Fig. 1. We see that in observed networks there are a lot of less important nodes with a low level of centrality, and few nodes with the relatively high centrality. The power law is executed, the network is not random and has some pattern or structure.

The same conclusions can be made while observing the betweenness centrality distribution by years (Fig. 2). We calculated betweenness centrality because we were more interested in identifying such policy actors as policy advisors who were

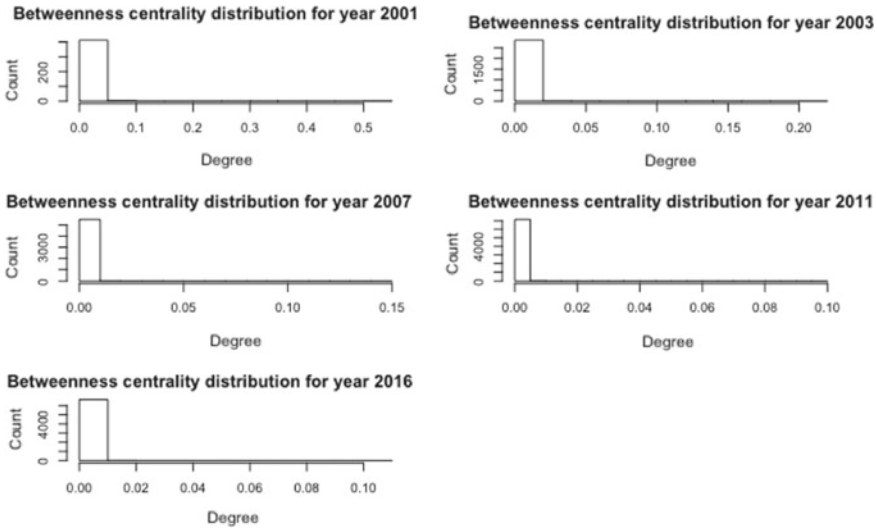
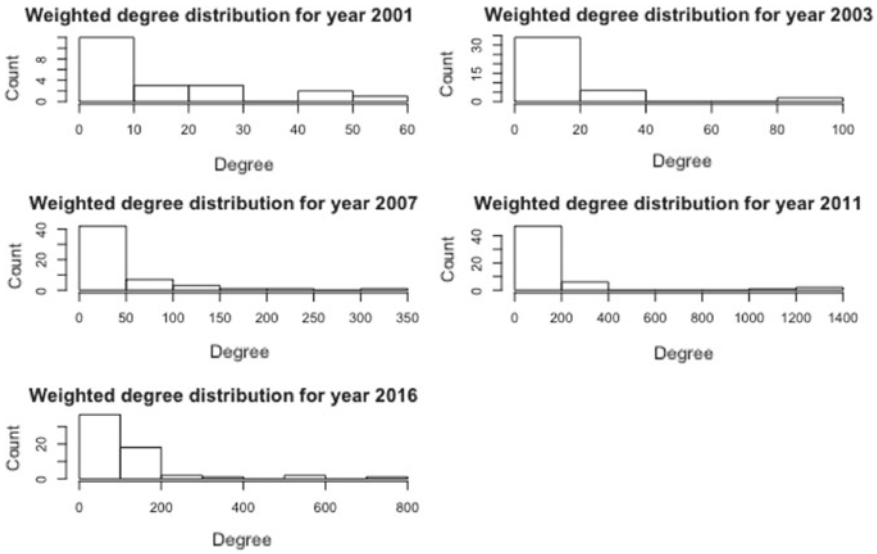


Fig. 2 Betweenness centrality distribution for the networks of policy actors by years

extremely active in the process of the implementation of the USE. We will talk about it in the next part about the methodology application that we know from the theory that policy advisors in the policy process can have a position of brokers. And within the case of introduction of the USE in Russia, our hypothesis is that due to such brokering position specialists in education policy or policy advisors in this field manage to influence and impact a lot into the formulation and implementation of the USE policy. That is why in further analysis, we will mostly use betweenness centrality to analyze the policy impact of policy actors.

However, at the exploratory part we faced several problems. The first major issue was the fact that since the library we have used to recognize named entities in the texts sometimes failed to properly find a nominative case of last names (for example, “Kuzminov” and “Kuzminova”) were often confused by the library. The second issue was that a lot of “noise” was generated in addition to the useful data, because, for example, names of the USE tutors were extracted from the articles as well. Therefore, we have decided to do a second step we called “confirmatory”.

Using results from the exploratory part we created a dictionary of last names of individuals we considered important policy actors. To do this, we sort the initial machine list of policy actors by descending order of betweenness centrality. And we google each name, we have from the highest to the lower betweenness centrality to understand their relation to the Russian education policy and the USE policy-making. Based on this coding we leave presidents, prime ministers, ministers of education, public servants of education policy agencies, rectors, journalists, and other policy actors related to the policy formulation and implementation of the USE. We deleted those names, who were the USE tutors, authors of the textbooks how to prepare for



**Fig. 3** Betweenness centrality distribution for the networks of policy actors by years

the USE, and others that are hardly related to the education policy-making. Through these, we created the dictionaries for each year with policy actors in education policy who had high level of betweenness centrality. Then we combine those dictionaries into the joint, universal one.

After that, we constructed another set of graphs of co-mentions of the terms or names from the dictionary. In order to lemmatize the texts, we have used Mystem which is a freeware morphological analysis application from Yandex. Then, we have computed another set of descriptive statistics for the new graphs with degree and betweenness centrality (Figs. 3 and 4).

There are still more important and less important nodes, but their distribution becomes a little bit flatter. So, at this stage we were able to get rid of “noise”. The network is still not random and has a structure. We can clearly see that there is a center and periphery. Therefore, to analyze such networks have some more sense than those received on the first exploratory part.

The outcomes of the confirmatory part of the implementation of the proposed methodology for the USE case are such networks that already can be interpreted and relevant to test public policy theories. The next part is demonstrating the application of the proposed methodology for policy studies.

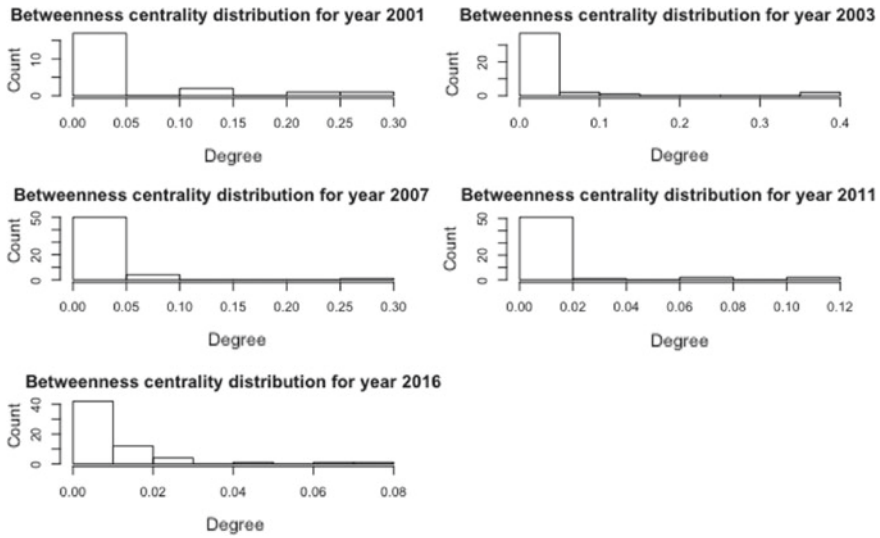


Fig. 4 Betweenness centrality distribution for the networks of policy actors by years

### 3 Application

Education policy is one of the rare examples of the consistently being formulated and implemented public policy in Russia. This happened due to the coherent and proactive position of policy advisors or epistemic communities. They are networks “of professionals with recognized expertise and competence in a particular domain and an authoritative claim to policy-relevant knowledge within that domain or issue-area” with “(1) a shared set of normative and principled beliefs, which provide a value-based rationale for the social action of community members; (2) shared causal beliefs, which are derived from their analysis of practices leading or contributing to a central set of problems in their domain and which then serve as the basis for elucidating the multiple linkages between possible policy actions and desired outcomes; (3) shared notions of validity, that is, intersubjective, internally defined criteria for weighing and validating knowledge in the domain of their expertise; and (4) a common policy enterprise, that is, a set of common practices associated with a set of problems to which their professional competence is directed, presumably out of the conviction that human welfare will be enhanced as a consequence” [13].

In public policy theories, policy advisors have the role of mediators between external to government advice producers or providers as academia does, from the one hand, and within government knowledge consumers such as policy-makers, on the other hand. The role of policy advisors is to span the boundary between producers and consumers, or academia people and policy-makers [14].

But in certain conditions, policy advisors or epistemic community can take the more important role beyond the mediation and transfer and translation of academic

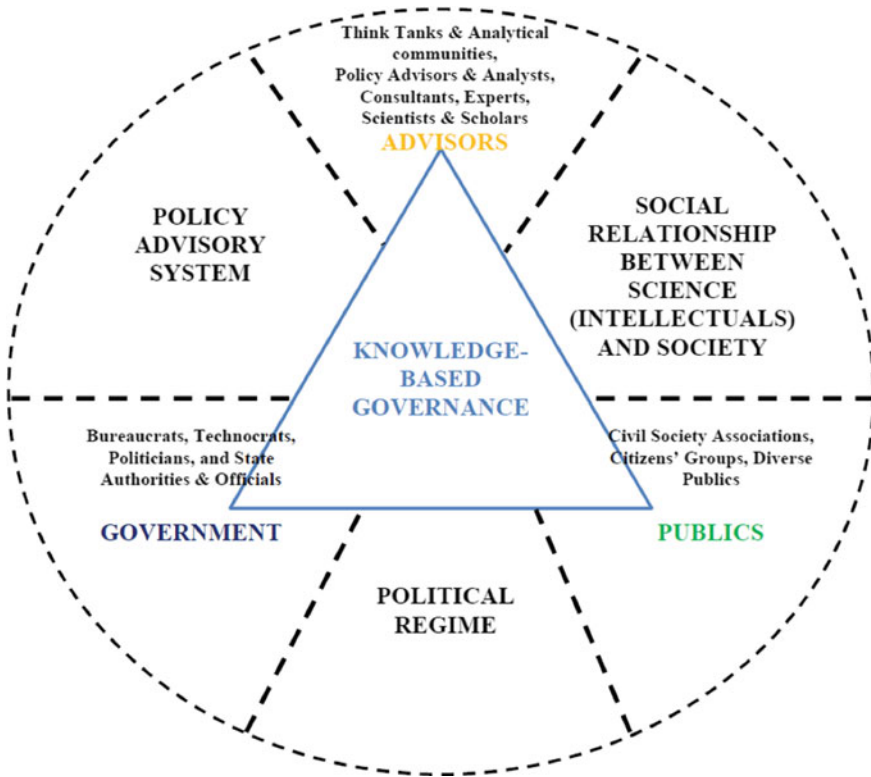


Fig. 5 Knowledge-based governance (ideal model)

information into the practical one. They can become the drivers for policy changes. The increasing influence of policy advisors reflects in many thinks. One of it is the appearance of the concept “knowledge-based governance” [15]. It can be summarized in the Fig. 5.

The ideal model of knowledge-based governance is that the governmental decisions in a policy field are formulated and implemented based on epistemic communities’ advise. So, advisors appeared on-the-top of the triangle, and they provide advice to the government utilized into the respective policy decisions and they provide information to the public explaining the sense of the current and future policies. Such a model presumes developed policy advisory system [16], respectable relations between science and society, and democratic political regime.

However, the ideals are a rare case in the real world (Fig. 6). Mostly, we have to deal with the situations when the policy advisors have roles to legitimize or oppose the already taken decision by the government or the role can limit by technical advice only. In this case, we are dealing with the government-driven model of policy change, when the government appears on the top providing directives to the both policy advisors and public. Another far-from-ideal situation of absence of knowledge-based



**Fig. 6** Knowledge-based governance (ideal model)

governance is when the public become drivers of change. Usually, it is protests against some policy measures that break down the whole policy design that government was trying to establish. Then public appeared on the top and government and advisors are forced to react and satisfy the claims of the protesters. In both the situations, we deal with the failures of the establishment of the knowledge-based governance.

There are plenty of examples of failure on knowledge-based governance approach to the policy design and policy-making in Russia: Monetization of Benefits Reform (was aimed “to replace subsidies on transport, education, health care, and other needs for pensioners, students, invalids, veterans, and other with cash benefits” [16], Restrictions in Housing Policy (Restrictions in Housing Policy—case of “dol’shiki” (“investors in real estate projects who lost their savings and, in most cases, their homes”); case of forced evictions of residents of South Butovo district in Moscow; case of building new high-speed toll road that cut through the Khimki Forest in Moscow region), and Reform of banning all Right-hand-drive Cars in Russia.

Monetization of Benefits Reform and Reform of Banning all Right-hand-drive Cars in Russia are the examples of ill-conceived reforms from the government with abuse of democratic deliberation principles and procedures, reasoning both from public and intellectuals, which led to the mass protests of public and urgent mobilization of financial and PR resources to correct the situation in the regime of “manual government”. Government failed to assess the risks, to build up communication based on common value system with public and analysts, ignored negative perception and disapproval of the proposed reforms both from public and intellectuals [17].

Reform of Banning all Right-hand-drive Cars in Russia lead to the “Russia’s Automotive Rebellion”, developing to social movement not only against banning all right-hand-drive cars, but also “nomenklatura” privilege, like use of “migalki” (vehicle lighting used by state officials). Fighting with “migalki” is an example of “public-driven” attempts to establish knowledge-based governance, as well as case of “dol’shiki”. But this attempt was not successful because they faced restriction or ineffective policy from the government, and lack of support from epistemic communities.

In this row, the introduction of the USE was a rare example of knowledge-based governance. The case of Russian education policy is a combination of favorable and unfavorable condition for the establishment of knowledge-based governance (like it is shown in Fig. 7). But because education policy, especially in the part of the

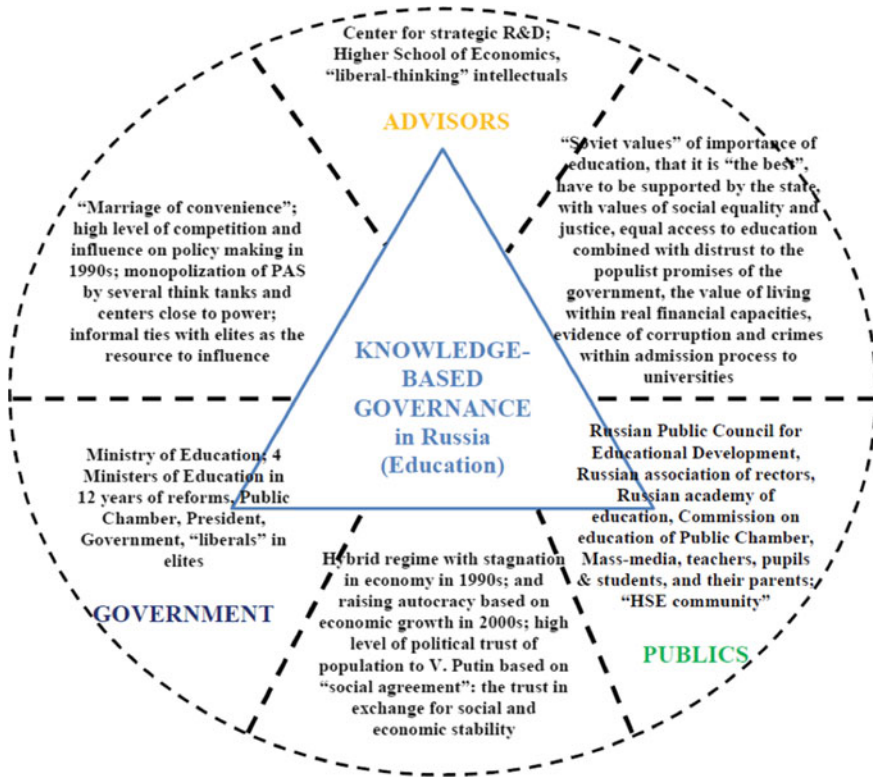


Fig. 7 Knowledge-Based Governance in Russian Education policy

introduction of the USE, remained the domain of policy advisors, they drive policy change in the consistent evidence-based policy.

That is why it is worth to prove that it is happened due to the strong, capable, and proactive position of the policy advisors able to impact policy change. Or to put it in the terms of network analysis, due to the high central position. As an evidence of such central position, let us provide the results of our network analysis which was developed based on the described above methodology (Table 3).

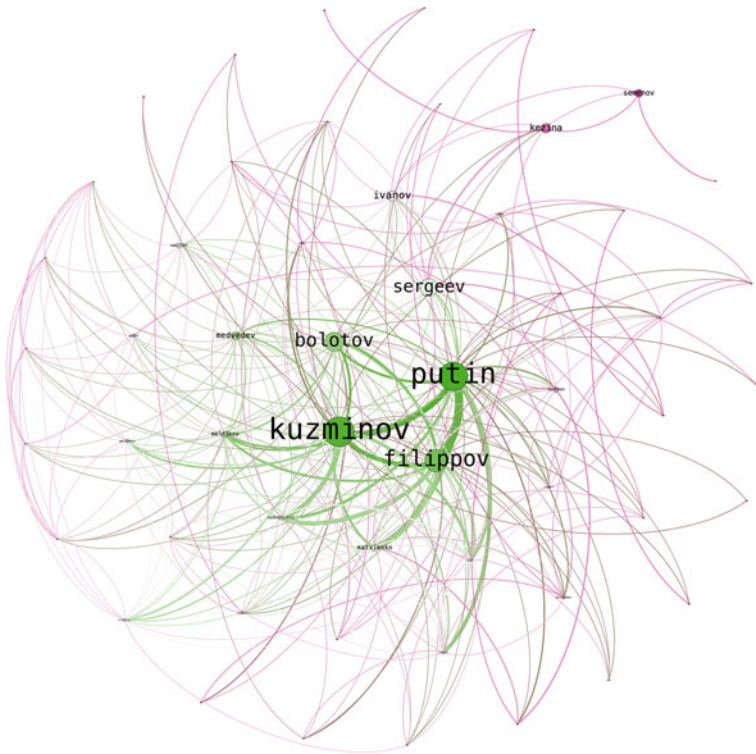
From the data we can see that Yaroslav Kuzminov, rector of the Higher School of Economics with whom the introduction and further development of the USE is closely associated, and the informal leader of epistemic community in the field, take the highest betweenness centrality position with high degree and closeness centralities too. He has the highest betweenness centrality position then Putin, and the second degree and closeness centrality position just after Putin.

In Fig. 8, we also see the network with connections between policy actors. The volume of the circle is proportional to betweenness centrality measure. In the picture, we also can observe that Kuzminov has ties with different types of policy



**Table 3** Centrality measures of top-10 policy actors in the USE policy in 2001 (for each measure the top-10 is in bold)

Label	All Degree of N1 (5)	All closeness centrality in N1 (54)	Betweenness centrality in N1 (54)
Kuzminov	<b>34</b>	<b>0.726</b>	<b>0.165</b>
Putin	<b>35</b>	<b>0.736</b>	<b>0.158</b>
Filippov	<b>31</b>	<b>0.688</b>	<b>0.123</b>
Bolotov	<b>28</b>	<b>0.663</b>	<b>0.103</b>
Sergeev	<b>20</b>	<b>0.616</b>	<b>0.095</b>
Ivanov	16	0.582	<b>0.053</b>
Kezina	7	0.530	<b>0.048</b>
Medvedev	<b>26</b>	<b>0.646</b>	<b>0.040</b>
Semenov	4	0.405	<b>0.038</b>
Matvienko	<b>21</b>	<b>0.609</b>	<b>0.030</b>
Melnikov	<b>25</b>	<b>0.639</b>	0.028
Vasilev	19	<b>0.589</b>	0.021
Sadovnichij	<b>21</b>	<b>0.609</b>	<b>0.016</b>
Golubkov	<b>20</b>	0.582	0.010



**Fig. 8** Network of policy actors in the USE policy in 2001



actors: president (Putin), his administration (Mdevedev), minister of education (Filipov), government and governmental agencies (Ivanov and Bolotov), and even rector community (Sadovnichij).

## 4 Conclusion

In conclusion, let us formulate steps for further development of the proposed methodology to identify policy actors for policy fields and evaluate their power. Further developments of the proposed methodology and its capacity being applicable to test public policy theories are connected with several problems being resolved.

First, is to increase the accuracy of names entity recognition. If with names here the proposed methodology consists of two steps (exploratory and confirmatory) works well, with organizations it is worth to be improved because the name of organization consists of several words that can be hard for the machine to recognize them.

Second, the tie formation is also the question. Here we proposed to form the tie between policy actors if they appeared together in a document. But why not in the sentence, or paragraph? Also, we do not make any difference between the sentiments of appearance: were they appear together within positive or negative contexts. And if evidence to chose new article is more than to chose sentence or paragraph, because the article is an element of public discourse but not random sentences and paragraphs; to distinguish between sentiments seems to be important from the perspectives of policy actors' reputation which is the part of their potential power.

Third, to test the impact of policy advisors on policy changes we have to go beyond the descriptive statistics and build more sophisticated models such as social influence and social selections models, using ERGM and SIENA in longitudinal social networks research design.

Fourth, the impact of policy advisors also has to be tested for its sustainability, how it changes through time and space. Change in time can be traced from the longitudinal models. Change in space can be grasped through the comparative analysis of policy advisory systems across policy fields and countries.

**Acknowledgements** The article was prepared within the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) and supported within the framework of a subsidy by the Russian Academic Excellence Project '5-100.'

## References

1. Irving, J., Lerner, D., Lasswell, H.: The policy sciences: recent developments in scope and method. *Int. J.* (1952). <https://doi.org/10.2307/40197737>
2. Mazmanian, D., Sabatier, P.: *Implementation and Public Policy*. Scott Foresman, Glenview, Illinois (1983)

3. Sabatier, A., Jenkins-Smith, H.: *Policy Change and Learning: An Advocacy Coalition Approach*. Westview Press, Boulder, Colorado (1993)
4. Baumgartner, F., Jones, B.: *Agendas and instability in American politics*. University of Chicago Press, Chicago (1993)
5. Kingdon, J., Thurber, J.: *Agendas, Alternatives, and Public Policies*. Little, Brown, Boston (1984)
6. Richardson, J.: *Policy Styles in Western Europe*. George Allen & Unwin, London (1982)
7. Howlett, M., Mukherjee, I.: *Routledge Handbook of Policy Design*. Routledge, New York (2018)
8. Zittoun, P.: *The Political Process of Policymaking: A Pragmatic Approach to Public Policy*. Palgrave Macmillan, London (2014)
9. Dahl, R.: *Who Governs?: Democracy and Power in an American City*. Yale University Press, New Haven, CT (1961)
10. Medialogia: O kompanii (About the company). <http://www.mlg.ru/about> 30 Nov 2018
11. Startsev, B.: *Khroniki obrazovatelnoi politiki: 1991–2011*. HSE, Moscow (2012)
12. Hagberg, A., Chult, D., Swart, P.: Exploring network structure, dynamics, and function using NetworkX. In: Varoquaux G, Vaught T, and Millman J (eds.) *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pp. 11–15. Pasadena, USA (2008)
13. Haas, P.: Introduction: Epistemic Communities and International Policy Coordination. *International Organization* (1992). <https://doi.org/10.1017/S0020818300001442>
14. Dunn, W.: The two-communities metaphor and models of knowledge use: an exploratory case survey. *Knowledge* **4**(1), 515–536 (1980)
15. Chou, K.: Conflicts of technology policy and governance paradigm in a knowledge-based economy: A case analysis of the construction of the Taiwan biobank. *Issues Stud.* **43**(3), 97–130 (2007)
16. Craft, J., Howlett, M.: Policy formulation, governance shifts and policy influence: location and content in policy advisory systems. *J. Pub. Policy* (2012). <https://doi.org/10.1017/S0143814X12000049>
17. Greene, S.: *Moscow in Movement: Power and Opposition in Putin’s Russia*. Stanford University Press, Stanford (2014)

# Correction to: User Preference Prediction in a Set of Photos Based on Neural Aggregation Network



Kirill V. Demochkin and Andrey V. Savchenko

**Correction to:**  
**Chapter “User Preference Prediction in a Set of Photos Based on Neural Aggregation Network” in:**  
**I. Bychkov et al. (eds.), *Network Algorithms, Data Mining, and Applications*, Springer Proceedings in Mathematics & Statistics 315,**  
[https://doi.org/10.1007/978-3-030-37157-9\\_8](https://doi.org/10.1007/978-3-030-37157-9_8)

In the original version of this chapter, the title of Chapter 8 has been changed from “Visual Product Recommendation Using Neural Aggregation Network and Context Gating” to “User Preference Prediction in a Set of Photos Based on Neural Aggregation Network”. The chapter and book have been updated with the changes.

---

The updated version of this chapter can be found at  
[https://doi.org/10.1007/978-3-030-37157-9\\_8](https://doi.org/10.1007/978-3-030-37157-9_8)

© Springer Nature Switzerland AG 2020  
I. Bychkov et al. (eds.), *Network Algorithms, Data Mining, and Applications*, Springer Proceedings in Mathematics & Statistics 315,  
[https://doi.org/10.1007/978-3-030-37157-9\\_17](https://doi.org/10.1007/978-3-030-37157-9_17)