9th International Conference on Network Analysis NET 2019

Moscow, May 18-19, 2019

Laboratory of Algorithms and Technologies for Network Analysis
National Research University
Higher School of Economics, Nizhny Novgorod
Laboratory of Advanced Combinatorics and Network Applications,
National Research University
Moscow Institute of Physics and Technology
Location: 16, Lva Tolstogo street, Yandex office, Entrance 4, basic floor, conference hall "Moulin Rouge"

NATIONAL RESEARCH
UNIVERSITY

Springer
Proceedings
in Mathematics
& Statistics

Springer
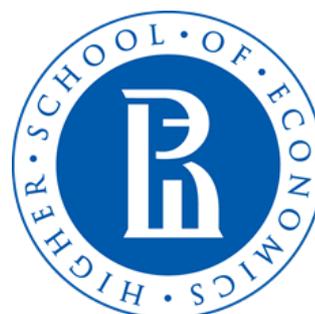
Opening lecture
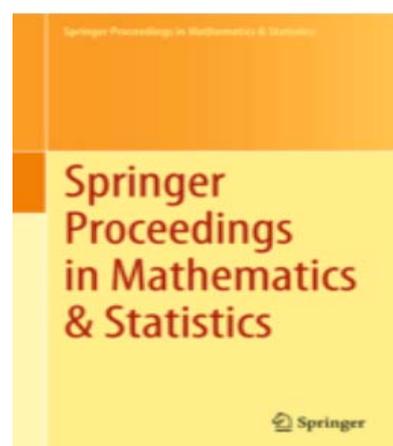
Panos M Pardalos, University of Florida and HSE
https://www.ise.ufl.edu/pardalos/   https://nnov.hse.ru/en/latna/
Networks of Networks

Plenary and invited talks

Vladimir Batagelj, Department of Mathematics, University of Ljubljana,
http://vlado.fmf.uni-lj.si/vlado/vlado.htm
Analysis of large networks

Alexander Dolgui, IMT Atlantique, Nante, France,
https://www.mines-stetienne.fr/~dolgui/
Combinatorial design of manufacturing systems: new approaches and real life applications

Mario Guarracino, ICAR-CNR, Naples, Italy,
http://www.na.icar.cnr.it/~mariog/
Supervised and unsupervised classification of networks

Anuška Ferligoj, Faculty of Social Sciences, University of Ljubljana
https://www.fdv.uni-lj.si/en/news-and-information/contacts/researchers/info/anuska-ferligoj/
The analyses of the WoS data on clustering networks

Andrei Leonidov, Lebedev Physical Institute of the Russian Academy of Sciences,
http://www.lebedev.ru/ru/people/alphabet.html?view=person&id=1766
Noisy discrete choice graphical games

Oleg Prokopyev, University of Pittsburg, USA and HSE NN
https://www.pitt.edu/~droleg/
On Bilevel Extensions of Network Optimization Problems

Maxim Zhukovskii, MIPT and Yandex,
https://mipt.ru/education/chairs/dm/staff/zhukovsky.php
Logical limit laws for attachment model

Location: 16, Lva Tolstogo street, Yandex office, Entrance 4, basic floor, conference hall "Moulin Rouge"

Saturday, May 18

09:30 – 10:00 Registration
10:00 – 10:30 Opening lecture. Panos M. Pardalos
Networks of Networks
10:30 – 11:20 Alexandre Dolgui
Combinatorial design of manufacturing systems: new approaches and real life applications
11:20 – 11:40 Coffee Break
11:40 – 12:30 Maxim Zhukovskii
Logical limit laws for attachment models

12:30 – 13:00 Anuška Ferligoj
The analyses of the WoS data on clustering networks
13:00 – 14:00 Lunch Break
14:00 – 14:50 Vladimir Batagelj
Analysis of large networks
14:50 – 16:20 Session 1
Chernenko Elizaveta *Network determinants of conflict in business*
Gradoselskaya Galina *Information waves on social networks: identification, specification and profiling*
Mikhaylova Oxana *Network approach for the analysis of the satanic milleu in the Russian social network "Vkontakte"*
Organizations
Shcheglova Tamara *Identification of patterns of social behavior in political groups in online social networks*
Žnidaršič Anja *How to estimate missing ties due to actor non-response in networks when blockmodeling structure is analysed?*
16:20 – 16:40 Coffee Break
16:40 – 18:10 Session 2
Karpov Ilia *Generating discussion network of social media using text mining*
Korzhenko Anna *Transformation of Communication Channels: from Pager to Instant Messengers*
Lushnikova Polina *Application of Data Envelopment Analysis and Network Methodology towards International Sanctions Effectiveness Evaluation*
Talovsky Nikita *Classification of the Named Entities: Testing the Advocacy Coalition Theory by Document Analysis (the Case of Russian Civil Society Policy)*
Vladimirova Alina *Missing, Asymmetric, Mixed: Selecting Data for International Trade Network Models*

Location: 16, Lva Tolstogo street, Yandex office, Entrance 4, basic floor, conference hall "Moulin Rouge"

Sunday, May 19

09:30 – 10:00 Registration

10:00 – 10:50 Andrei Leonidov
Noisy discrete choice graphical games

10:50 – 11:20 Mario Guarracino
Supervised and unsupervised classification of networks

11:20 – 11:40 Coffee Break

11:40 – 13:00 Session 3

Khvatsky Grigory *Multilevel Exponential Random Graph Models Application to Civil Participation Studies*

Kiselev Dmitrii *Prediction of new itinerary markets for airlines via attributed network embeddings*

Makarov Ilya *Predicting Collaborations via Weighted Link Prediction based on Co-authorship Network Embedding*

Makrushin Sergey, Blokhin Nikita *Generation of associative series of hashtags by building semantic paths on a co-occurrence network*

Sidorov Sergei, Melnichuk Dmitriy *Analysis of Financial Network Topological Dynamics of the Russian Stock Market from 2012 to 2019*

13:00 – 14:00 Lunch Break

14:00 – 14:50   Oleg Prokopyev
On Bilevel Extensions of Network Optimization Problems

14:50 – 16:20 Session 4

Katargin Nikolai *Solving network tasks in Excel: the traveling salesman problem, the placement of power substations*

Miasnikof Pierre *Graph Clustering via Digital Annealer*

Veremyev Alexander *Dense Subgraphs in Random Graphs*

Rubchinsky Alexander *Dijkstra algorithm with modifies weight functions*

Ketkov Sergey *On the shortest path problem with probability constraints*

16:20 – 16:40 Coffee Break

16:40 – 18:10 Session 5

Gribanov Dmitry *Integer minimization of conic functions*

Koldanov Petr *Testing new property of elliptical model for stock returns distribution*

Mokeev Dmitry *Polynomial-time algorithm of finding the minimum k-path vertex cover and the maximum k-path packing in some graphs*

Nikolaev Kirill *Classification of User Interests by Text Messages with Recurrent Neural Networks*

Sirotkin Dmitry *Local transformations of the graphs for an independent set problem and associated constructive theorem*

Opening Lecture

Networks of Networks

Panos M. Pardalos

University of Florida and HSE

Many complex systems in nature (or man made) are represented not by single networks but by sets of interdependent networks. Such networks of networks (NoN) include the internet, airline alliances, biological networks, and smart city networks. There is no doubt that NoN will be the next frontier in network sciences. In my lecture I will address some recent developments (robustness, diversity) and discuss some challenging problems in NoN.

Plenary and invited talks

Analysis of large networks

Vladimir Batagelj

IMFM Ljubljana, IAM UP Koper, NRU HSE Moscow

A large network is a network with several thousands (millions, ...) of nodes that can be entirely stored in a computer's memory. Large networks are usually sparse (Dunbar's number). To deal with large networks only subquadratic algorithms are acceptable. Some special types of networks are considered: two-mode networks, miltirelational networks, multimodal networks (collections of linked networks), and temporal networks. These types can be combined. We present some general and efficient approaches and algorithms to deal with large networks: node properties or link weights detecting important parts of network, cores and generalized cores, cuts and islands, short cycle connectivity, multiplication of networks, and illustrate them with examples from applications.

For details see:

Batagelj V, Doreian P, Ferligoj A, Kejžar N: Understanding
Large Temporal Networks and Spatial Networks: Exploration,
Pattern Searching, Visualization and Network Evolution.
Wiley, 2014.

De Nooy W, Mrvar A, Batagelj V: Exploratory Social Network
Analysis with Pajek; Revised and Expanded Edition for
Updated Software. Cambridge UP, 2018.

Combinatorial design of manufacturing systems: new approaches and real life applications

Alexander Dolgui

IMT Atlantique, Nante, France

A complex machine or machining line consists of a sequence of work positions through which products move one way in order to be processed. Designing such a production system represents a long-term decision problem involving different crucial decision stages. Combinatorial design is one of them; it mostly deals with assigning the set of indivisible units of work (named tasks or operations) to work positions (or stations). In literature, the most attention was paid for combinatorial design of assembly lines (assembly line balancing problems). In our work, we develop approaches and formulations of combinatorial design for machining lines and complex machines. All types of machining lines are considered: mass production transfer lines, flexible lines based on machining centers and reconfigurable manufacturing systems.

Supervised and unsupervised classification of networks

Mario Guarracino,

ICAR-CNR, Naples, Italy

Networks represent a convenient model for many scientific and technological problems. From power grids to biological processes and functions, from financial networks to chemical compounds, the network representation of data makes it possible to highlight both topological and qualitative characteristics. In this talk, we report recent developments in supervised and unsupervised classification of network data. Given a dataset whose members are networks, we show how to cluster them. In case a class label is available, we show hot to build a mathematical model for their supervised classification. We focus on networks with labeled nodes and weighted undirected edges, defining distances between networks. We provide empirical results on datasets of biological interest.

The analyses of the WoS data on clustering networks

Anuška Ferligoj,

University of Ljubljana, NRU HSE, Moscow

Vladimir Batagelj,

University of Ljubljana, NRU HSE, Moscow

Patrick Doreian,

University of Ljubljana, University of Pittsburgh

There is a large literature on clustering networks. We collected bibliographic data for the network clustering literature including both community detection and blockmodeling works through to February 22, 2017. The primary data source was the Web of Science. From the obtained data we created a citation network among works. In addition, we included data on authors, journals and keywords to generate some two-mode networks featuring works $\times$ authors, works $\times$ journals, and works $\times$ keywords. The boundary problem is discussed as was a treatment ensuring the studied citation network is acyclic. Lists of the most prominent journals where works in the network clustering literature appeared were created. Components of the studied network were identified and examined. The CPM path through the main component was identified. It revealed a clear transition from the social network part of the literature to the community detection part. The key-route paths revealed the same transition but with more works and a more nuanced view of it. Ten link islands, as clusters, were identified. Detailed discussions were provided for four including one with a clear distinction between the community detection and social networks literatures as being connected through a cut.

Noisy discrete choice graphical games

Andrey Leonidov

Lebedev Physical Institute of the Russian Academy of Sciences,

Modern understanding of static and dynamic equilibria in noisy discrete choice games on graphs is discussed. New results on quantal response equilibria in static and dynamic binary and static ternary choice games on graphs are presented.

On Bilevel Extensions of Network Optimization Problems

Oleg Prokopyev

University of Pittsburg, USA and HSE NN

Bilevel optimization models capture decision-making processes that involve two (or more) independent decision-makers, namely, the leader and the follower(s), who act in a hierarchical manner. The involved decision-makers may be collaborative or conflicting. Bilevel optimization approaches has been applied in several important application domains. In general, the idea of extending single-level combinatorial optimization problems into the bilevel framework has attracted considerable research attention in past decades. In this talk we overview some of the work in the area, in particular, focusing on bilevel problems in network settings.

Logical limit laws for attachment models

Maxim Zhukovskii

MIPT and Yandex,

A sequence of random graphs G_n on n labeled vertices obeys first order (FO) zero-one law, if every first order sentence is true on G_n with an asymptotical probability either 0, or 1. The FO convergence law holds, if, for every FO sentence, the probability that it is true on G_n converges. It was independently proven by Glebskii, Kogan, Liogon'kii and Talanov in 1969 and Fagin in 1976 that an undirected random graph having uniform distribution on the set of all graphs on {1,..,n} obeys FO zero-one law. This random graph model has the same distribution as the binomial random graph G(n,p=1/2). It is known that the same result holds for every constant edge probability p. Many other random graph models were considered in the context of logical laws: sparse binomial random graphs, random trees, random regular graphs, and others. We consider the uniform attachment model: at every step, we add a vertex with m edges chosen uniformly at random. An interest in studying attachment models is motivated by results of Bollobas et al. that demonstrate similarities between certain attachment models and real networks. We have proved that, for m=1, the random graph obeys FO zero-one law, while, for m>1, it fails. Nevertheless, for m>1, the FO convergence law holds.

Regular talks

Network determinants of conflict in business organizations

Elizaveta Chernenko,

National Research University - Higher School of Economics

Researchers in the fields of Organization Studies and Management have traditionally looked at such variables of interest as productiveness, employee turnover and team conflict as functions of attributes of employees, management units or organization as a whole. However, in the past 15 years, researchers have acknowledged that many of these outcome variables are largely shaped by the informal structure of an the organization, or, its social network. As coined by Rob Cross, organizational network analysis is an X-ray into the actual processes evolving in the organization. Organizational Network Analysis can hence be an effective tool to address many management tasks in organizations, such as measurement of employees' performance, creating a collaborative environment, identifying organizational roles of employees for targeted interventions in change initiatives; succession planning; increasing team viability through diagnosis and management of conflicts.

This research is devoted to the investigation of network predictors of personal and group-level conflict in an organization. One such important predictor is employee centrality. There is mixed evidence regarding whether high centrality is desirable. For example, high betweenness centrality of a leader in the advice network was shown to be associated with higher conflict and lower team viability. Along with other aspects, my paper addresses such phenomena by an empirical study spanning three countries over four years.

I use longitudinal data on one industrial organization, operating in Belarus, Poland, and Russia. The data were collected by the author via an online questionnaire, filled by the employees in 3 waves (June 2015, December 2017 and December 2018). Network data are based on employees' nominations in response to questions on different working interactions, including reporting relationships, asking for advice, conflicts, and favorable attitudes. A relationship is defined as 'conflict' whenever one employee indicated that he/she "has had to turn to a third party for help or mediation" to resolve the difficulty with another employee. The resulting networks are middle size (825 employees in 2015, 754 in 2017, 883 in 2018) multi-relational directed graphs, weighted by intensity of interactions. The data available also include employees' position, department affiliation, seniority, peer-evaluated competences, social skills as well as demographics such as gender and age.

I use these data to test the set of hypotheses about network factors of organizational conflict on an individual (H1 and H2) and department level (H3 and H4):

H1. Asymmetric perception of organizational role in a subordination network increases chances of conflict between the corresponding employees.

H2a. Organizational structures allowing reporting relationships between one employee to multiple managers are associated with increased organizational conflict.

H2b: The associated organizational conflict is moderated by soft skills and managerial competencies of involved managers.

H3. Higher managerial skills and competencies of leaders are negatively associated with intragroup and intergroup conflict of involved departments.

H4a. Higher indegree centrality of a leader in advice the network is associated with lower levels of the intragroup conflict.

H4b. Higher betweenness centrality of a leader in advice network is associated with higher levels of the intragroup conflict.

The set of Machine Learning techniques is applied to solve the classification problem for the purpose of predicting a conflict link between a pair of nodes. Demographic characteristics, measures of formal organizational position, peer-evaluated competencies and skills were used together with network-based characteristics of the involved employees to predict a conflict in the corresponding pair of nodes.

To test the stated set of hypotheses I used such network predictors as link reciprocity, responsible for the symmetric perception of organizational roles and a set of centralities (indegree, closeness, betweenness and PageRank) and clustering coefficient of nodes on the obtained networks to provide a detailed picture of the network position of the involved nodes. Furthermore, for the purposes of exploratory analysis, I also test the predictive power of graphlet spectra of involved nodes on a subordination network. In other words, I used frequency distributions of a node, "touching" graphlets of size 2 to 5 at different orbits. Previous research on the organizational data has shown an outstanding explanatory power of graphlet distributions to explain the role played by an individual within the organizational network.

The data are currently being processed. As the preliminary the result I can say several network predictors have shown their significance, providing evidence in support of some hypotheses, while other are not supported by the data. The resulting model is currently being calibrated. However, it has already shown to have high predictive power with AUC of above 0.9

The results of the tests and exploratory analysis will be proposed for discussion at the conference.

Information waves on social networks: identification, specification and profiling

Galina Gradoselskaya, NRU - HSE

Information waves have recently become an increasingly sophisticated mechanism for the controlled dissemination of information, combining both technological and social resources. The technical possibilities for the dissemination of information through social networks and the social nature of the dissemination of this information raised the question of the discovery of these waves by researchers. The groundwork on the part of mathematicians (cascades, etc.) brought little benefit in practical terms, precisely because the technical component of the mathematical models missed the social component, as well as the manipulative component of information waves. A content and social characteristic of the authors determine how information is distributed. This, in fact, is taken into account by researchers who study the dissemination of information as the propagation of information waves. Thus, it is possible to state the scientific problem of determining the information wave as an object of study, conducting the operationalization of this concept and developing a methodology for studying it. In connection with the formation of a new information and media paradigm, the need to clarify the conceptual and terminological apparatus arises. Information waves are the ways of propagating manipulative models in the information space. It can be defined in another way: information waves as a controlled way to spread information to target audiences. An important direction of controlled dissemination of information is working with target audiences. But the design of targeted access streams is not just to broad layers of consumers of the information product, but access to a narrowly targeted group, accurate reporting of information designed specifically for this group, and designing a social response to the message received. A wave presupposes a response, two-way communication, and if there is none, there is no wave either. This suggests the social nature of information waves, which distinguishes them from simple information flows.

The main problem in studying information waves is the lack of a systematic approach to analysis due to their diversity: waves can be strong and weak in terms of their impact on the public consciousness, long or short in duration, and also have different shapes and patterns of propagation. Depending on these characteristics, wave profiles can be natural, artificial, mixed or technical. Knowledge of the wave profile can help to correctly predict further information dissemination and decision-making on the management of this propagation. Here we propose a methodological technique for identification and analyzing information waves. The method of identifying information waves consists of two analytical steps: automatic and expert (manual). At the first stage, information waves are extracted from the entire array of publications due to automatic analysis of texts and their clustering by Birch algorithm. And at the second stage, an expert verification of the clusters takes place: texts that fall into one cluster are compared and analyzed. The method of profiling the waves includes the time-frequency distribution analysis, dissemination structure analysis and analysis of user's types, who participate in the dissemination.

We have analyzed nine information waves and compare them. As a result of this study, we answer the question: how does the time-frequency, network and information characteristics affect the information wave profile. By analyzing the time-frequency distribution, it is possible to see outliers, jumps and other anomalous patterns in the distribution, as well as to determine the wave propagation profile. The analysis of the distribution structure that we conducted allowed us to take into account the links not only between different types of resources (groups, public pages, websites and people), but also between different platforms (VKontakte, Facebook, Odnoklassniki, Twitter, YouTube, media sites, etc.). We assumed that the quality of the model is related to the wave profile. If the wave is natural, the curve fitting method works, and the model describes the data well. To simulate artificial waves, it is necessary to resort to other methods of analysis.

Integer minimization of conic functions

Dmitry Gribanov,

National Research University Higher School of Economic,

Lab LATNA

In this talk we consider algorithms for the problem of integral conic function minimization. The class of conic functions is a wide subclass of quasiconvex functions that contains strictly quasiconvex functions, convex functions and quasiconvex polynomials as proper sub-classes. Moreover, it is known that the constraint variant of the minimization problem with conic functions can be polynomially reduced to the unconstrained problem of minimization of only one conic function. Additionally, we will consider the problem of discrete conic function minimization. A discrete conic function can be defined only in the points of integral lattice, so it is impossible to evaluate this function in an arbitrary rational point.

In our talk we assume that the conic functions to be minimized are presented by the comparison oracle and an optimal minimum point is contained in the ball of a radius R. The algorithms that will be presented have oracle complexities of the type $f(n) \log(R)$, where $f(n)$ is a function that depends exponentially from the dimension n. The known lower complexity bounds and efficient algorithms for the case n=2 are also will be considered.

Generating discussion network of social media using text mining

Ilia Karpov,

National Research University Higher School of Economic

This report describes the mehod for generating discussion network of social media using text mining, based on the environmental topics for 2019. For this, the initial list of ecological keywords was used to search for posts in VKontakte network. The resulting posts array was clustered. After the expert selection of clusters corresponding to the subject of environmental problems and gathered information about the most active groups. We built For the language model to observe protests lexic in their posts. After that, we ranked the groups by percentage of protest posts and compared the most and the least involved groups. We also built a network of groups based on the intersection of users of these groups.

Solving network tasks in Excel: the traveling salesman problem, the placement of power substations

Nikolai Katargin,

Financial University under the Government of Russian Federation

Fundamentally new algorithms for solving two problems in Excel are proposed: the choice of the route in the road network (the traveling salesman problem), the placement and connection to consumers of power substations with the provision of minimizing losses in power grids. The author's know how: "short plan" in the problem of choosing a route, which allows to dramatically reduce the number of variables varied by the computer, non-trivial use of the service Solver using the gradient descent method to vary binary variables, as well as the joint variation of real and binary variables in the problem of placement. The problem of the appearance of "islands" in the problem of choosing a route – nodes that are not connected with the main route is solved. The algorithm is based on the prevention of return on the same way, except in special cases – star routes from some points. The algorithms are implemented in MS Excel, for their use does not require programming, but only filling in the tables of the original data and simple steps: copying and summarizing. Route selection and placement of substations are tested on networks with 15 nodes, which is enough for practice. The optimal routes are laid through the real road network, both without returning to the original point and the ring route, as in the classical traveling salesman problem. The problem of placing objects also uses real maps (Yandex). Two options have been worked out- with the limitation of substations in capacity and without limitation. This algorithm may be used to optimize the location of, for example, fuel and commodity supply bases. The results may be used for practical work in transport logistics and placement of new objects, as well as training students in solving economical problems using mathematical modeling and information technology.

On the shortest path problem with probability constraints

Sergey Ketkov,

National Research University Higher School of Economics,

Lab LATNA

We consider a weighed directed graph, where the nominal costs are not known to the decision-maker in prior. Specifically, the initial information consists of the intervals containing the unknown nominal cost and some additional information about the underlying distribution. For example, one may take into account information that extremely high costs are quite rare, or use some confidence bounds for the mean. We minimize the total expected cost with respect to the worst possible probability distribution. The resulting distributionally robust optimization problem is formulated as a mixed-integer programming problem using duality theory. Some structural properties of this problem are discussed and connections with budget constraints are revealed.

Multilevel Exponential Random Graph Models Application to Civil Participation Studies

Grigory Khvatsky,

National Research University Higher School of Economics

Due to the development of the Internet and social networks, civil participation in political processes is taking new forms, inexplicable by in the classical theoretical frameworks. In this study, we present a new methodology for researching these new, unconventional forms of civil participation. We use data collected from a social network site VK.com. Social network analysis methods, such as multilevel exponential random graph models (ERGM) were used to analyze the data. First proposed by Wang in 2013, multilevel ERGM allow us to model tie formation in a wide class of social networks. This method allows us to efficiently model group membership based on node attributes. Since the collected data is fundamentally 2-mode, this model allows us to identify important factors that lead to people joining online protest communities.

Prediction of new itinerary markets for airlines via attributed network embeddings

Dmitrii, Kiselev,

National Research University Higher School of Economics, Moscow, Russia

Airline business is very competitive, so airlines try to act in optimal way to gain profits. There are different ways to do it: marketing, operations or better network planning. Network planning is one of the most important problems that airlines solve every day. It deals with finding optimal schedule and appropriate plane types and choosing proper prices for a specific route. This paper considers how airline could efficiently enter new itinerary markets for airline network using modern approaches in network science. It is important to mention that there are few research works in the chosen field because most data remain confidential. Lack of data also induces us to choose United States market for research. There are several available for us data sources for US aviation market. In terms of network science new itinerary route problem can be stated as a link prediction problem (LPP). Link prediction is a problem of predicting new edges in network or reconstructing missing edges in noisy graph data. Basic approach suggested to find several similarity measures between vertices [Liben-Nowell et al., 2007] (for e.g., common neighbors). If nodes are similar by certain threshold then new edge should exist. This approach can be expanded using topology and social network analysis theory bringing more complex metrics such as generalized Jaccard coefficient, hitting time between nodes or homophily and structural balance [Wang et al., 2015]. From machine learning point of view, link prediction problem can be solved by standard classification approach using standard frameworks over feature space for pairs of nodes and predicting whether the edge between them will appear/exist in the network. The answer on how to construct such features without domain expert and independent of task was found via formulation of graph embedding. These methods allow to find automatically low-dimensional description of nodes and edges [Hongyun Cai, 2018; Peng Cui, 2018; Goyal and Ferrara, 2018; Haochen Chen et al., 2018; Daokun Zhang et al. 2018; Ziwei Zhang et al., 2018]. It helps efficiently solve LPP because it does not require calculation of all the similarities or manually engineering graph-related features. For dynamic networks there exist graph embedding models taking into account temporal component in structural data [Goyal et al., 2018]. Detailed survey of approaches on link prediction problem can be found in

[Wang et al., 2015]. We define Airline network by set of nodes as cities and flights between them as edges. We describe nodes with airport productivity statistics. It contains such data as flight delays, number of passengers. Each edge is described by different combinations of the following options: set of marketing carriers (categorical), mean price (numeric), set of planes (categorical), amenities (vector of Boolean values), set of departure time (periodical numeric) and mean duration (numeric). It is also important to notice that network is dynamic, because of markets opening and closing, price temporal dependence, changing in schedule. Network data is obtained from US Department of Transportation Airline Origin and Destination Survey (DB1B). It is 10% subsample of all airline tickets reported by carriers. Minimal available time delta for DB1B data is quarter. This data contains edge attributes: fares, carriers, number of passengers on market. The second data source is the Marketing Carrier On-Time Performance from US DOT. It contains information about flight schedule, delays, cancellations, duration can be calculated from actual departure and arrival time. City and airport features and statistics are also available from above data sources. In this paper we compare methods for solving link prediction problem based on network structure and based on network node and edge attributes. We analyze only data for the latter year (2018). Overall number of cities connected by graph was 414 and number of edges was 73482. For train period we choose first quarter of 2018. The construction of test sample will be described below. Test positive examples are created as routes (edges) opened in the periods after first quarter of 2018. Test negative examples are generated as edges that connect nodes with path along two incident edges. There are two types of models tested: end-to-end link prediction models and unsupervised embedding learning followed by classifier. First type of models was evaluated in two cases: unsupervised embedding learning on train sample followed by supervised classifier learning on half of test sample and semi-supervised learning on union of train sample and half of test set. For models of second type only first case was evaluated. We compare different models using area under ROC curve and average precision metrics.

Testing new property of elliptical model for stock returns distribution

Petr Koldanov,

NRU Higher School of Economic, lab LATNA

Wide class of elliptically contoured distributions is a popular model of stock returns distribution. However the important question of adequacy of the model is open. There are some results which reject and approve such model. Such results are obtained by testing some properties of elliptical model for each pair of stocks from some markets. New property of equality of Kendall correlation coefficient and probability of sign coincidence for any pair of random variables with elliptically contoured distribution is proved in the paper. Distribution free statistical tests for testing this property for any pair of stocks are constructed. Holm multiple hypotheses testing procedure based on the individual tests is constructed and applied for stock markets data for the concrete year. New procedure of testing the elliptical model for stock returns distribution for all years of observation for some period is proposed. The procedure is applied for the stock markets data of China, USA, Great Britain and Germany for the period from 2003 to 2014. It is shown that for USA, Great Britain and Germany stock markets the hypothesis of elliptical model of stock returns distribution could be accepted but for Chinese stock market is rejected for some cases.

Transformation of Communication Channels: from Pager to Instant Messengers

Anna Korzhenko,

National Research University Higher School of Economics

The article provides an analytical review of the transformation of modern communication channels: from pager to messenger. The systematization of the basic theoretical positions and practical reasons for the migration of users from one communication channel to another has been carried out. The reasons for the popularity of the Telegram messenger and its influence on the information agenda in Russia occupy a special place in the review. A number of popular claims regarding Telegram are considered, some of which can be considered mythological. A special plot is devoted to the processes of blocking Telegram in Russia.

By their functions, many messengers are not inferior to each other. In addition, the concept of modern instant messaging software is very similar to its predecessor – ICQ, which won the attention of users for its simplicity and convenience. All messengers are built on a single interface: conversations with users and group chats are sorted by the time of the last conversation, messages within the chat appear in strict chronology. Nevertheless, the key role in choosing a messenger is not only clear interface, but also a list of convenient and high-quality features. Based on a comparative analysis, the most functional, at the moment, is the Telegram messenger.

Telegram quickly gained popularity among the mass consumer. Was it a well-built marketing plan or P. Durov, the creator of the messenger, launched the application on time?

Application of Data Envelopment Analysis and Network Methodology towards International Sanctions Effectiveness Evaluation

Polina, Lushnikova, HSE

In this paper we propose to use data envelopment analysis (DEA) to evaluate effectiveness of international sanctions. Data envelopment analysis is a performance measurement technique which is used for comparing the performances of similar units of an organization. DEA has wide applications in all industries including hospitals, banks, universities etc. The most important advantage of this technique is that it can handle the multiple input and output variables which are generally not comparable to each other. We chose Threat and Imposition of Sanctions (TIES) Data 4.0 as a base for analysis. This database was chosen because it contains the largest number of cases (1412 from 1945 to 2005), takes into account which countries impose sanctions at the same time, and also estimates the cost of all sanctions both for those who receive and for those who impose them. As an input variable for DEA model we use the impact of institutions on sanctions, type of issue, threat, economic embargo, sanction type. And as the output variable, we chose the outcome of sanctions for sender and recover. We use DEA cross-efficiency outputs to build the network of sanctions episodes depend on the similarity between them. These allow us to cluster sanctions episodes depend on their outcomes and analyze why one group of sanction episodes are more efficient than others.

Predicting Collaborations via Weighted Link Prediction based on Co-authorship Network Embedding

Ilya Makarov,

National Research University Higher School of Economics

In this paper, we study the problem of predicting collaborations in co-authorship network. We formulated our task in terms of link prediction problem on weighted co-authorship network, in which authors play the role of nodes, and weighted edges connecting two authors are formed by storing a number of research papers co-authored by these authors. Our task is then formulated as regression machine learning model based on network features constructed using network embedding and new family of edge embedding operators. We evaluate our model on large AMiner co-authorship network for (un)weighted node2vec network embeddings. We also evaluate our models on the dataset containing temporal information on National Research University Higher School of Economics (HSE). We showed that our model of network edge representation has better performance on both, AMiner and HSE co-authorship networks.

Generation of associative series of hashtags by building semantic paths on a co-occurrence network

Sergey Makrushin and Nikita Blokhin,

Financial University under the Government of the Russian Federation

In the current research we considered the problem of building associative series of hashtags for the Instagram, because hashtags are an important mechanism of navigation in contemporary social media. These series should meet two criteria: they should be relatively short and shouldn't have wide semantic gaps between sequential hashtags. Such series could be used for creating recommendations to increase the quantity of hashtags in messages.

We gathered information about using hashtags in Instagram and built a co-occurrence network with 1.7 million nodes (hashtags) and 63.9 million of weighted links. We have proposed a definition for the optimization problem of a semantic path building on this co-occurrence network, which leads to building associative series of hashtags. For the two-criteria optimization problem we propose a combined optimization function. Parameters of this function allow to balance the criterion of minimizing the widest semantic gap and the criterion of minimizing the length of the path.

To measure semantic similarity between hashtags, we used a metric based on vector embedding of hashtags by word2vec algorithm. Using empirical paths built by different algorithms, we graduated the parameters of the combined optimization function. We built examples of semantic paths using the Dijkstra's algorithm and the fast greedy algorithm and performed subjective judgment about the possibility of utilizing these paths as associative series.

Graph Clustering via Digital Annealer

Pierre, Miasnikof,

University of Toronto and Fujitsu Co-Creation Lab

Graph clustering is an NP-hard problem. In addition, graphs are not in metric space, so distances between vertices are typically unknown. We begin by computing vertex-to-vertex distances and then formulate the k-means problem as a QUBO. We use a Fujitsu Digital Annealer chip to obtain a solution within reasonable time.

Network approach for the analysis of the satanic milleu in the Russian social network "Vkontakte"

Oxana Mikhaylova.

International laboratory for Applied Network Research, Higher School of Economics

In the Russian social networks now there are many communities, which could be characterized as belonging to the satanic milieu. These communities have more than thousands of subscribers. Danish sociologist of religion A. Petersen argued that satanism should be understood as a field which consists of different discourses: rationalistic, esoteric and reactive. Among these three ideal types only reactive satanism represents the explicit intentional sometimes aggressive rebellion against Christianity. In the spring of 2018 the representative of the Russian Security Council argued that Russian Internet is now flooded by satanic destructive cults and that we have to save Russian youth from the influence of the dangerous ideologies. The protection of Russian youth from the satanism presumes the blockage of communities that seem suspicious. In the European history there was satanic moral panic, so that it was supposed that probably not all communities which call themselves satanic are – reactive. To understand which communities exist in the Russian Internet exploratory social network analysis of the groups which constitute the kern of Satanic milieu in "Vkontakte" was conducted. In addition, the semantic networks for the publications in the groups were constructed to illustrate the process of satanism framing. The employed methodological procedure could be used for the analysis of the whole structure of the satanic milieu and the other discursive fields of interest in Russian social networks.

Polynomial-time algorithm of finding the minimum k-path vertex cover and the maximum k-path packing in some graphs

Dmitry Mokeev,

NRU High School of Economics,

Laboratory of Algorithms and Technologies for Networks Analysis (Nizhny Novgorod)

For a graph G and a positive integer k, a subset C of vertices of G is called a k-path vertex cover if C intersects all paths of k vertices in G. For a graph G and a positive integer k, a subset M of pairwise vertex-disjoint paths of k vertices in G is called a k-path packing. We describe some graphs, having the equal values of cardinalities of a maximum k-path packing and a minimum k-path vertex cover (for k > 4) and give a polynomial-time algorithm of finding the minimum k-path vertex cover and the maximum k-path packing in such graphs.

Classification of User Interests by Text Messages with Recurrent Neural Networks

Kirill Nikolaev,

National Research University Higher School of Economics,
Laboratory of Algorithms and Technologies for Networks Analysis (Nizhny Novgorod)

This talk is devoted to the formulation and solution of a new text classification problem: automatic determination of user interests by their text messages. This can be relevant for various recommender systems and targeted advertising, especially on mobile devices and / or when user background is unknown (no information on age, location, marital status. An original dataset is collected: almost 90 thousand text messages from forums on various subjects, each of which is related to one of ten interests: anime, books, food, football, games, movies, music, nature, painting, traveling. Using the obtained data, we performed computational experiments on predicting the category of text using deep learning methods – in particular, recurrent neural networks. We evaluated classification accuracy for different representations of text, algorithms and architectures. Part of our contribution is that we propose a lexis-centered interest modeling technique for text representation that augments the more widely-used doc2vec method. Based on the results of a number of computational experiments, we proposed a complete pipeline for solving the problem of labeling user interests by text messages. It includes a series of preprocessing techniques, followed by a combined approach to feature extraction (doc2vec + keywords and key character trigrams), with feature vector then input into a bidirectional LSTM neural network that predicts the user interest. The classification accuracy achieved using this method was 78.5%. The dataset and code of the proposed solution are publicly available.

Dijkstra algorithm with modified weighted functions

Alexander Rubchinsky, HSE, MISIS

The shortest path problem, consisting in finding paths of minimal summary weight, is one of the most investigated and required problems of discrete optimization. It is due to widespread occurrence of diverse applied statements that are reduced to shortest path problems. In such initial meaningful statements weights of arcs or edges are distance, time, cost, penalty, load (quotient of flow to capacity) – in a word, any minimizing value, which has some additive or other decomposional properties (considered in the work in more detail). Usually the sum of the weights of all the edges (or arcs) of the path P has been considered as a weight $W(P)$ of a path P. Some other dependences $W(P)$, both simple and complicated ones, have been considered in the literature but the review of different specific generalizations and modifications of the shortest path problem is out of scope of the work. Here some general requirements to dependence of the path weight from weights of its arcs are considered. The weight functions, satisfying the suggested conditions, allow elaborating such modifications of the classical Dijkstra algorithm that guarantee finding of shortest (in the above sense) paths.

The known general conditions are presented in several publications. They are formulated as follows.

1. A weight of a path is no less than a weight of the initial segment of the path, consisting of all its edges except the last one.

2. A weight of a path can be presented as a non-decreasing function of two variables: a weight of the initial segment of the path and a weight of the last edge of the path.

Both conditions were essentially used in evident proofs of applicability of classical Dijkstra algorithm for problems with weight functions, satisfying these conditions.

However, even in several relatively simple examples weight functions do not satisfy condition 2, though does satisfy condition 1. Let us consider the following weight function: $W(P)$ is equal to sum of weight in all the edges of the path P, if it

exceeds 4, and it is equal to 4, otherwise.

Assume path P consists of 3 consecutive edges with weights 1, 1, 1; segment P_1 consists of the 2 first edges and segment P_2 coincides with the 3rd edge. By the construction, $W(P_1) = 4$, $W(P_2) = 4$ и $W(P) = 4$. Assume now that weights of the 2 first edges forming segment P_1 are equal to 2 (instead of 1), and the weight of the last edge is equal to 1 as before. By the construction, $W(P_1) = 4$, $W(P_2) = 4$ и $W(P) = 5$. It means that in this simple case the weight of the initial segment and the weight of the last edge remain the same, while the weight of a whole path changes (5 instead of 4), that contradicts to condition 2.

Several other examples of shortest path problems whose weight functions do not satisfy condition 2 and satisfy condition 1 are given in the work. The main theoretical result is formulated as follows.

Statement 1. Assume that weights in all the edges be non-negative and the weight function satisfies condition 1. In this case the classical Dijkstra algorithm is applicable.

It is possible to say that condition 1 is necessary and sufficient for applicability of classical Dijkstra algorithm, because its infringement clear demonstrates that it does not work in this case.

Two practical examples of shortest path problem with modified weight functions are considered in the presented work.

Identification of patterns of social behavior in political groups in online social networks

Tamara Shcheglova, NRU HSE

The information space of social networks, as a special social and communicative environment, is a subject to social engineering. In the Internet space, attempts are regularly made to influence attitudes and social behavior on a large scale through online media or individual groups and individuals in order to have the desired effect on the target population. And online social networks play a key role in this influence, because, firstly, they provide access to specific target audiences, and, secondly, they open up the possibility of using the latest advertising mechanisms and promoting ideas and attitudes. For a long time already communities in social networks (especially in Facebook) work according to a debugged scheme. The structure of roles existing in a group allows one to create and maintain user interest in certain political topics. There are not only administrators, moderators and "all other members" who constitute an online group. First of all, users are divided into active and inactive. The latter can also be of several types: who joined the group by chance, who is in it for a while and forgets to quit or who just quietly watches what is published and discussed in the group. Among active users may be moderators, content creators, information hubs, likers, reposters, commentators, bots, etc. Usually there are opinion leaders in almost every group, which set topics and moderate discussions on them, and often these are not official group administrators and moderators. Identifying the structure of roles and positions, firstly, helps to reveal manipulation schemes that exist in online groups, and secondly, allows political communication members to understand how to work effectively with specific groups. In our research we analyze and compare structures of online political groups by defining the roles and positions of group members with a help of blockmodeling technique. Modeling of block structures allows to select cohesion, center and periphery in the structure and to carry out ranking. On the one hand, blockmodeling is a mathematical method that helps to reduce a large, potentially incoherent network to a smaller comprehensible structure that can be interpreted more readily. On the other hand, it refers to several major theoretical concepts in sociology such as positions, roles and role sets, which are in the focus of our research. The method of blockmodeling is combined with other network methods like multilayer and multimodal analysis due to the fact that in online social networks there are several types of entities that can be measured in different ways.

Analysis of Financial Network Topological Dynamics of the Russian Stock Market from 2012 to 2019

Sergei Sidorov, Dmitriy Melnichuk, Alexey Grigoriev,

Saratov State University

The link to the draft of the talk is https://drive.google.com/file/d/18U0JwZ7rZvL3Pt9z_lmFlY4Zh3WwVf5r/view?usp=sharing

Abstract. During the past ten years, the Russian economy suffered a series of shocks, such as the sharp fall of oil prices in 2014 and the sanctions imposed by governments of European Union countries and the United States. Both shocks brought widespread and long-lasting negative impacts and led to a slowdown in growth of the Russian economy. These dramatic events have been reflected by the Russian stock market behavior. This paper studies the properties of the Russian stock market by employing the data-driven science and network approaches. The theory of complex networks allows us to build and examine topological network structures of the market with the further identification of relationships between stocks and the analysis of hidden information and market dynamics. Our research examines the impact of these shocks on financial network topological dynamics of the Russian stock market in the period from 2012 to 2019. In this paper we will present an analysis of the structures, topological properties and dynamics properties of the Russian stock market using market graph, hierarchical tree, minimum spanning tree, planar maximally filtered graphs, asset graph approaches. Moreover, we compare topological properties of the networks constructed for the US and China stock markets in [1] with the properties of corresponding networks constructed for the Russian stock market using a dataset spanning over eight years.

[1] Tang, Y., Xiong, J.J., Jia, Z.Y., Zhang, Y.C.: Complexities in nancial network topological dynamics: Modeling of emerging and developed stock markets. Complexity 2018, 056110 (2018).

Local transformations of the graphs for an independent set problem and associated constructive theorem

Dmitry Sirotkin,

National Research University Higher School of Economics,

lab LATNA, Nizhny Novgorod

An independent set problem is to find the size of the maximum set of pairwise non-adjacent vertices. Numerous cases of this problem being NP-complete or solvable in polynomial time are known for the different classes of graphs. The local transformations of graphs are widely used for determining which of these two cases holds. We consider some certain class of local replacements of one subgraph with another one, such that these replacements change an independence number by known value. Every such local transformation can be defined by a certain pattern – a set of subsets of certain vertices of replaced subgraph. In our talk we show that for any such pattern, which falls under a certain restriction, there exists a replacing subgraph. This result contributes to the algorithmic graph theory and theory of constructing the graphs with the given properties.

Classification of the Named Entities: Testing the Advocacy Coalition Theory by Document Analysis (the Case of Russian Civil Society Policy)

Nikita, Talovsky,

National Research University Higher School of Economics

We present a methodology for identification and classification of policy actors. We used network analysis, rule-based named entity recognition on a computational cluster for actor identification and Chinese whispers algorithm with pre-specified clusters to identify probable coalitions between the identified actors. We test this methodology on the case of Russian policy towards civil society. The theoretical framework we have chosen is the Advocacy Coalition Framework which is a public policy theory aimed at explaining long-term policy change by understanding how and why people engage in policy-making. One of the key ideas of the theory is that people engage in politics to translate their beliefs into action, and then gather into advocacy coalitions based on the shared "beliefs system". Identification of actors is one of the most fundamental issues in political science, as it is often the first step in the research process. The classification of the actors based on latent characteristics such as shared beliefs is also a problem. Most of the research paper apply qualitative methodology for both of the steps. By applying our methodology, we identify two coalitions — pro-governmental and oppositional. The leading actors in each coalition correspond with quantitate research in the field.

Dense Subgraphs in Random Graphs

Alexander Veremyev, University of Central Florida

The problem of identifying large dense subgraphs received a lot of attention in the recent literature due to its various practical applications. Since the recognition version of this problem is NP-complete, good analytical bounds may play an important role for the design and development of effective heuristic algorithms as well as evaluating their performance. In this presentation, we develop asymptotic analytical bounds on the size of the largest subgraphs with a certain edge density (referred to as quasi-cliques) in binomial random graphs G(n,p). Specifically, we show that the size of the largest quasi-clique with high probability concentrates around two explicitly defined integers. Moreover, we present our recently developed MIP models and conduct computational experiments to investigate the quality of presented bounds which suggest that they are quite accurate even for relatively small values of n. Some interesting observations, relations of the obtained results to other problems and directions for future research will be also discussed.

Missing, Asymmetric, Mixed: Selecting Data for International Trade Network Models

Alina V. Vladimirova,

Institute of Oriental Studies of the Russian Academy of Sciences

A network approach to analysis of international trade became a well-established practice a long time ago; however, it is developing fast until now. Scholars constantly keep introducing new research questions, new models and new data. Taking into account the current replication crisis and data revolution we would like to focus on the last element of this list in our conference paper. While there are brilliant studies published by network scientists on the topic of missing data influence on results, little was said about effects of data quality problems specific for international trade statistics such as asymmetry. To fill this gap and to assist the process of data selection for network models we have conducted a project that consists of a few stages. We have started with a literature review to identify databases used for international trade analysis and then we have assessed their quality. During these stages we mostly were interested in comparing these sources from points of coverage and accuracy, however, as scientists, we have taken not only criteria traditional for data management but also incorporated the FAIR data principles. Our findings can be presented in a few parts:

1) results of API monitoring

2) measurement of asymmetry levels and identification of the most deviant cases

3) accounting for the influence of a particular dataset quality issues on standard centrality indices.

Thus, even though in the end it is still impossible to name one the best source of international trade mirror statistics, our study can help researchers to navigate better between datasets that are made to a greater or lesser extent with missing, asymmetric or mixed data.

How to estimate missing ties due to actor non-response in networks when blockmodeling structure is analysed?

Anja, Žnidaršič, University of Maribor, Faculty of Organizational Sciences

Patrick Doreian, University of Pittsburgh, Department of Sociology and University of Ljubljana, Faculty of Social Sciences;

Anuška Ferligoj, University of Ljubljana, Faculty of Social Sciences.

In social networks, units are most often actors, where the most frequently reported and published relations are: be a friend, be a classmate, a co-worker, providing or receiving help, gossiping, attending an event together, etc. Similarly as in standard social science research, surveys are still the most common instrument to gather network data and therefore, the gathered network data are prone to measurement errors. An error in social network occurs when there is an extra tie or a missing tie according to the true underlying and unobservable structure. This definition is suitable for binary networks and has to be adapted to valued networks where the magnitude of tie values represents their great advantage. In valued networks each changes of a tie value can be considered as measurement error.

Types of errors found in the literature can be classified into three categories: the boundary specification problem, errors caused by questionnaire design, and errors caused by actors. The latest category, errors due to actors, can be further divided into complete actor non-response, non-response on specific ties, and (random) measurement errors in recorded ties. This presentation focuses on actor non-response which occurs when actor or actors refused to participate in the research and provide data on other network members. In matrix representation of the network, each non-respondent leads to a row of missing ties (ideally denoted by NA) meaning that all out-going ties are missing.

The goal of the network analysis is to create from raw relational data a useful description of relationships. One useful and widely used technique for finding structural patterns in networks is (generalized) blockmodeling. The goal of the blockmodeling is to reduce a large incoherent network to a smaller comprehensible and simply interpretable structure. In more detail, the purpose of the blockmodeling procedure is to partition the network actors into clusters (discrete subgroups also called positions), and, at the same time, to partition the set of ties into blocks which are determined by the positions. Actors are

partitioned into clusters based on selected type of equivalence (the most known is the structural equivalence) or simply with selection of allowed block types according to philosophy of generalized equivalence. The concept of generalized blockmodeling was proposed in 2005 by Doreian and coworkers.

Combining both facts, wide usefulness of generalized blockmodeling and existence of non-respondents in social networks, results in the decision to investigate the impact of actor non-response in valued networks to the results of blockmodeling.

Similarly as in ordinary social surveys, also in network data there exist several attempts how to estimate or impute missing data. The most commonly used treatments are: reconstruction, imputations based on modal values of incoming ties, combination of reconstruction and imputations based on modal values for ties between non-respondents, imputations of the total mean, imputations of the mean value of incoming ties, null tie imputations, and median of the k-nearest neighbours based on indegree of actors.

One of the best treatments for missing network data is the k-nearest neighbours approach which searches for k nearest actors according to their incoming ties and then calculates the outgoing ties of each non-respondent as a median of outgoing ties of selected nearest neighbours. In the previous studies, usually, 3-nearest neighbours were used to estimate the missing outgoing ties. This presentation will present the performance of the treatment for the k-nearest neighbours based on incoming ties for various k's (e.g. from 1 up to 10 nearest neighbours). More precisely, it will try to determine what the optimal number of nearest neighbours is, to successfully estimate missing data when revealing network's blockmodel.

To answer the above question, results of the simulations based on several real networks will be presented. The obtained results with the k-nearest neighbours will be compared to reconstruction treatment, where missing outgoing ties of non-respondents are replaced by the available incoming ties to those actors (Stork and Richards, 1992; Huisman, 2009). To emphasize, that ignoring missing data of non-respondents is an extremely bad approach, the results will be compared to the null tie imputation treatment, where instead of missing ties zeros (indicating that no relationship exist between two actors) are imputed. The null tie imputation treatment fails to reveal both micro (position membership) and macro level (block structure) depictions of the network.