Workshop on Clustering and Search techniques for large scale networks

October 23rd – October 25th, 2015

Russian Science Foundation (RSF), Russia

РОССИЙСКИЙ НАУЧНЫЙ ФОНД

Friday, October 23rd

Room 209 HSE, 136 Rodionova Str.

9:30-10:00 Registration of participants

10:00 -10:50 Panos Pardalos

Opening talk: Constrained Subspace Classifier For High Dimensional Datasets

10:50 -11:20 Coffee break

11:10 -12:00 Giuseppe Nicosia

Plenary talk: Pareto Optimization and Data Analysis for Biological Networks

12:10 -13:00 Mario Guarracino

For biological experimental data to regulatory and interaction networks

13:00 -14:30 Lunch

14:30 -15:20 Giuseppe Nicosia

Lecture: Multi-objective Optimization for Large Scale Networks

15:30 -16:20 Alexey Nikolaev

Efficient approach for the maximum clique problem based on machine learning

16:20 -16:40 Coffee break

16:40 -17:30 Valery Kalyagin

Robust identification in large scale network

Saturday, October 24th

Room 209 HSE, 136 Rodionova Str.

10:00-10:50 Katerina Papadaki

Plenary talk: Patrolling Games

10:50 -11:20 Coffee break

11:10-12:00 Nenad Mladenovic

Variable neighborhood programming – a new automatic programming method in artificial intelligence

12:10-13:00 Alexander Ponomarenko

Techniques for overlapping community detection

13:00-14:30 Lunch

14:30-15:20 Katerina Papadaki

Lecture: Patrolling Games for special graphs

15:30-16:20 Ashwin Arulselvan

Lecture: Exact and approximation algorithms for designing optical access networks

16:20 -16:40 Coffee break

16:40-17:00 Karpov Nikolay

Detecting New a Priori Probabilities of Data Using Supervised Learning

17:00-18:00 Student session

Sunday, October 25th

Room 209 HSE, 136 Rodionova Str.

10:00-10:50 Ashwin Arulselvan

Plenary talk: Recent advances in critical element detection in networks

10:50 -11:20 Coffee break

11:10-12:00 Andrey Savchenko

Statistical classification of a sequence of objects based on a fuzzy approach

12:10-13:00 Mikhail Batsyn

Applying bitwise operations for solving combinatorial optimization problems

13:00 – 14:00 Lunch

14:00-15:00 *Open problems session*

15:00 Closing

Constrained Subspace Classifier For High Dimensional Datasets

Panos M. Pardalos

CAO, University of Florida, USA & LATNA, NRU HSE, Russia

pardalos@ise.ufl.edu

In this work, we propose a new binary classification method called constrained subspace classifier (CSC) for high dimensional datasets. CSC improves on an earlier proposed classification method called local subspace classifier (LSC) by accounting for the relative angle between subspaces while approximating the classes with individual subspaces. CSC is formulated as an optimization problem and can be solved by an efficient alternating optimization technique. Classification performance is tested in publicly available datasets. The improvement in classification accuracy over LSC shows the importance of considering the relative angle between the subspaces while approximating the classes. Additionally, CSC appears to be a robust classifier, compared to traditional two-step methods that perform feature selection and classification in two distinct steps.

This is joint work with Petros Xanthopoulos and Orestis Panagopoulos.

Pareto Optimization and Data Analysis for Biological Networks

Giuseppe Nicosia

Dept. of Mathematics and Computer Science

University of Catania, Italy

nicosia@dmi.unict.it

This talk presents a Pareto-oriented approach using ad-hoc data mining for the design and optimization of large-scale biological networks. The goal is to solve the metabolite over-production problem in the *E. coli* bacteria and in other living molecular machines for the industrial production of chemicals, bioenergy and biofuel. The methodology is based on a multi-objective optimization algorithm, which balance biomass formation and metabolite of interest overproduction, finding a number of Pareto-optimal and Pareto-epsilon-optimal strains. Optimization capabilities are strongly enhanced by global Sensitivity Analysis, approximated Pareto-optimal strains and Robustness analysis, furnishing both biological insight to the problem and precious information for industrial purposes. We throughout fully demonstrate the solidity and flexibility of our approach, comparing it with state-of-the-art techniques in the over-production problem of *i)* 1,4-butanediol (+662.7% compared to wild-type), *ii)* myristoyl-CoA (improvement of +21.43% and +5.19% for biomass and myristoyl-CoA compared to other approaches), *iii)* malonyl-CoA, *iv)* acetate and *v)* succinate; considering various environmental conditions, and simulation of the genetic manipulations allowed. Our approach demonstrates to be a solid tool for the analysis and design of large-scale biological networks.

For biological experimental data to regulatory and interaction networks

Mario Guarracino

ICAR-CNR, Italy

mario.guarracino@cnr.it

Although there is not much difference in the genetic material own by a mouse and a human being, these two organisms appear and are quite different. The complexity of an organism is strictly related to the complexity of interactions among molecules and its cells, and the study of such interactions can greatly help to have more insight in the processes and functions governing phenotypic characteristics of an organism. One of the most powerful tools to model such phenomena are networks and related analysis tools. In this lecture we will see how to use publicly available experimental data to build networks and which are the most common techniques to mine them. Finally, we will discuss open research problems.

Multi-objective Optimization for Large Scale Networks

Giuseppe Nicosia

Dept. of Mathematics and Computer Science

University of Catania, Italy

nicosia@dmi.unict.it

Recent advances in complex networks call for robust, flexible and efficient optimization methodologies. We present a class of multi-objective optimization algorithms for the analysis and design of large scale networks. Our method efficiently explores the high dimensional space, finding a number of robust trade-offs, hence furnishing a deeper insight to the addressed problems.

The lecture presents 1) Multi-objective Optimization, 2) Local and Global Sensitivity Analysis, and 3) Local and Global Robustness analysis. More specifically, I will present single- and multi-objective optimization algorithms. I will show that the condition of Pareto optimality can be relaxed (e.g., epsilon-dominance) to include suboptimal points that can be used to boost the algorithm in its convergence process. The Sensitivity Analysis is used to compute an index for each parameter that indicates its influence in the model. The Robustness Analysis (RA), Local, Global and Glocal robustness, proves useful to assess the fragileness and robustness of the Pareto optimal solution (or of a given feasible solution) as a result of a perturbation occurring in the model. Our methodology is suitable for (i) any model consisting of ordinary differential equations, differential algebraic equations and for (ii) any simulator. In the lecture, I will show how these techniques offer avenues to systematically explore, analyse, optimise, design and cross-compare large-scale network.

Efficient approach for the maximum clique problem based on machine learning

Alexey Nikolaev

NRU HSE, Russia

ainikolaev@hse.ru

In this talk a new algorithm is presented that uses machine learning technique for choosing the fastest algorithm from the set of considered maximum clique problem solvers. Computational results show that the proposed algorithm chooses the fastest algorithm from five considered state-of-art algorithms with high accuracy. The average speedup of the proposed algorithm compared to the considered solvers reaches 35%.

Robust identification in large scale network

Valery Kalyagin

NRU HSE, Russia

vkalyagin@hse.ru

A class of distribution free multiple tests is proposed for the identification of network structures in random variables networks. The tests are based on estimations of probability of sign coincidence of pairs of random variables. The quality of the tests is measured by conditional risk with additive loss function. It is proved that in this case multiple tests for threshold graph identification are distribution free in the class of elliptically contoured distributions. It is demonstrated in simulations that this class of tests is distribution free for identification of other important network structures as well. Some applications to market network analysis are discussed.

Patrolling Games

Katerina Papadaki

LSE, UK

k.p.papadaki@lse.ac.uk

This talk describes a class of patrolling games on graphs, motivated by the problem of patrolling a facility (for example in order to defend an art gallery against theft of a painting, or an airport against terrorist attack). The facility can be thought of as a graph Q of interconnected nodes (e.g. rooms, terminals) and the Attacker can choose to attack any node i of Q within a given time T: He requires m consecutive periods there, uninterrupted by the Patroller, to commit his nefarious act (and win). The Patroller can follow any path on the graph. Thus the patrolling game is a win-lose game, where the Value is the probability that the Patroller successfully intercepts an attack, given best play on both sides. We determine analytically optimal (minimax) patrolling strategies for various classes of graphs.

Variable neighborhood programming – a new automatic programming method in artificial intelligence

Nenad Mladenovic

LAMIH, University of Vallenciennes, France &

Mathematical Institute SANU, Belgrade, Serbia

nenad.mladenovic@univ-valenciennes.fr

Automatic programming is an efficient technique that has contributed an important development in the field of artificial intelligence. Genetic programming (GP), inspired by genetic algorithm, is among the few evolutionary algorithms used to evolve population of programs. To the best of our knowledge, there are no many techniques in the literature, after the appearance of GP, whose solutions are represented as programs. In this paper we present a new technique of that type, called Variable neighborhood Programming (VNP) that was inspired by GP and Variable Neighborhood Search (VNS) metaheuristics. VNS is based on systematic change of neighborhood structures within a local search. VNP starts with a single solution presented by a program, and the search for the good quality global solution continues by exploring different neighborhoods. To show the effectiveness of our method, we tested it on benchmark problems drawn from time series prediction and classification areas, and we compared it with the related techniques.

This is joint work with Souhir Elleuch and Bassem Jarboui.

Techniques for overlapping community detection

Alexander Ponomarenko

NRU HSE, Russia

aponomarenko@hse.ru

The talk will be devoted to techniques overlapping community detection. Clustering is a powerful means for analyzing complex networks. In spite of the absence of the rigorous mathematical formulation of the term “cluster”, many techniques were proposed. We will discuss several common approaches, their strong and weak points.

Patrolling Games for special graphs

Katerina Papadaki

LSE, UK

k.p.papadaki@lse.ac.uk

I describe a class of patrolling games on graphs, motivated by the problem of patrolling a facility (for example in order to defend an art gallery against theft of a painting, or an airport against terrorist attack). The facility can be thought of as a graph Q of interconnected nodes (e.g. rooms, terminals) and the Attacker can choose to attack any node i of Q within a given time T: He requires m consecutive periods there, uninterrupted by the Patroller, to commit his nefarious act (and win). The Patroller can follow any path on the graph. Thus the patrolling game is a win-lose game, where the Value is the probability that the Patroller successfully intercepts an attack, given best play on both sides.

In this lecture I will describe proofs of dominance results, strategy reduction techniques and bounds on the value of some graphs. Further, I will demonstrate that some strategies are optimal for graphs like the Hamiltonian graph, the star graph and the line graph.

Exact and approximation algorithms for designing optical access networks

Ashwin Arulselvan

University of Strathclyde, UK

ashwin.arulselvan@strath.ac.uk

In classical facility location problem, we are given a set of customers to be served by a set of potential facilities. It is often perceived as a clustering problem, where in, we seek clusters of customers each with a designated facility that minimizes the overall connectivity and location cost. We will present some recent work in this field with an application to optical access network design. In an optical access network, we have a set of customers with a demands that should be routed from their serving facilities. Optical cables need to be laid along these routes and this gets translated as costs. Due to the time and money involved in building such networks, its often deployed over a period of time and incremental solutions are often sought. The optical cables come with varying capacities and costs obeying economies of scale. We discuss several models and algorithmic solutions whose computational efficiency is demonstrated on real world and benchmark instances.

Detecting New a Priori Probabilities of Data Using Supervised Learning

Nikolay Karpov

NRU HSE, Russia

nkarpov@hse.ru

We are interested in the problem of estimation a priori probabilities (relative frequency or prevalence) of the classes in a dataset of unlabelled items. This is usually calls as a quantification problem and solves with the help of supervised learning techniques. There are many works which applies general-purpose classifiers to solve this problem but till now there is no total evaluation of known algorithms. We implement an approach based on expectation-maximization (EM) algorithm to estimate a priori probability of classes in text quantification task. Obtained results we compare with existing state-of-the-art quantification methods and find that our one has better accuracy than others.

Recent advances in critical element detection in networks

Ashwin Arulselvan

University of Strathclyde, UK

ashwin.arulselvan@strath.ac.uk

There has been an increased interest in identifying critical elements (nodes or edges) of a network. Networks have become an integral part of our modern daily lives starting from the apparent ones like the internet and social networks to the not so conspicuous ones including global supply chain and finance networks. The day-to-day operations of these networks rely on their robustness, which in turn is governed by the inclusion or removal of their critical elements. The definition of robustness and hence the methods used to identify these critical elements are predominantly application driven. For example, in a telecommunication network or a power grid we have to ensure a certain percentage of continued service in the event of any node (or edge) failure, while in a social network we are interested in determining the key members of its largest coalition. Since the inception of graph theory, we have been traditionally studying the critical element detection in various forms including disjoint paths, cliques, dominating sets, Steiner trees, multicuts, etc. We have been modifying these problem definitions in order to adapt ourselves to the evolving modern networks that are by nature massive and not necessarily static. In this talk, we will discuss this trend and shed some light on the obvious necessity for these modifications. We will also discuss some of the progress we have made and the direction we are headed.

Statistical classification of a sequence of objects based on a fuzzy approach

Andrey Savchenko

NRU HSE, Russia

avsavchenko@hse.ru

In this talk we focus on the problem of statistical classification of a sequence of identically distributed objects (e.g., the video-based image recognition, the phoneme recognition). The classification method on the basis of a fuzzy approach is discussed in details. Its preliminary phase includes the association of each reference object with the fuzzy set of classes. At first, each object (e.g., the frame) in a classified sequence is put in correspondence with the fuzzy set, which grades are defined as the posterior probabilities. Next, this fuzzy set is intersected with the fuzzy set, corresponding to the nearest neighbor reference object. The final decision for the whole sequence of objects is the arithmetic mean of these fuzzy intersections. We experimentally demonstrate, that our approach makes it possible to implement a voice control system, which does not need the general acoustic model if the latter does not fit to the user voice due to known variability sources (childhood, voice diseases, non-nativeness, etc.).

Applying bitwise operations for solving combinatorial optimization problems

Mikhail Batsyn

NRU HSE, Russia

mbatsyn@hse.ru

In many combinatorial optimization problems input data are represented by a binary matrix. For example in graph optimization problems a graph is given by its binary adjacency matrix. At the same time modern computer processors can work efficiently with blocks of bits processing 64 or even 128 bits simultaneously by one bitwise operation. This idea can be used to develop efficient heuristic and exact algorithms for combinatorial optimization problems. In this talk we present our heuristic algorithm for the vertex coloring problem (Komosko, Batsyn, Segundo, Pardalos, 2015) and an exact algorithm BBMC for the maximum clique problem (Segundo, Rodriguez-Losada, Jimenez, 2011). Both these algorithms apply bitwise operations for processing the adjacency matrix of the input graph. The computational results show that such an approach can significantly increase the speed of calculations.