Mathematical Data Science

Matematički institut SANU, 22. 6. 2015.

9:50-10:00	**Opening of the workshop**
10:00-10:20	***Predrag Nešković****, Program Officer Mathematical Data Sciences, Office of Naval Research* **Mathematical Data Science program**
10:20-10:40	***Zoran Obradović****, Laura H. Carnell Professor of Data Analytics, Temple University (Computer and Information Sciences Dept., Statistics Dept., Data Analytics and Biomedical Informatics Center)* **Predictive Analytics in Complex Dynamic Networks**
10:40-11:00	***Natasa Pržulj****, Computer Science, Imperial College London* **Predictive Integration of Networked Big Data: From Biology to Economics**
11:00-11:30	**Coffee break**
11:30-11:50	***Milan Vukićević****, Faculty of Organizational Sciences, University of Belgrade* **Predictive, preventive, personalized medicine - data, and knowledge driven approach**
11:50-12:10	***Veljko Milutinović****, School of Electrical Engineering,University of Belgrade* **DataFlow SuperComputing**
12:10-12:30	***Zoran Ognjanović, Miodrag Rašković, Zoran Marković****, Matematički institut SANU* **Probabilistic logic based unceratain reasoning**

12:30-13:00 Coffee break

13:00-13:20 ***Miodrag Mihaljević***, *Matematički institut SANU*

Advanced Encryption Approaches for Machine-to-Machine Communications and Big Data Processing

13:20-13:40 ***Nenad Mladenović***, *Matematički institut SANU*

Clustering Community Networks by Variable Neighborhood Search

13:40-14:00 ***Gordana Pavlović-Lažetić, Jovana Kovacević***, *Faculty of Mathematics, University of Belgrade*

Predictive models based on support vector machines for structured outputs

Speaker: **Pedrag Nešković**, Program Officer, Office of Naval Research, Arlington, USA

Title: Mathematical Data Science program

Abstract: I will provide an overview of research sponsored by ONR’s Mathematical Data Sciences program and specific thrust that the program is addressing. I will also describe the ONR funding guidelines and possible funding opportunities for international researchers.

Title:

Structured Regression in Evolving Health Networks

Speaker: **Zoran Obradovic**, Laura H. Carnell Professor of Data Analytics
Data Analytics and Biomedical Informatics Center, Computer and Information Sciences Department, Statistics Department Temple University, PA, USA

Abstract:

Predictive modeling in health networks is a challenging problem due to partially observed node attributes and links that often evolve over time. Additional challenges involve presence of multiple types of links among nodes that should be considered jointly where various nodes have different temporal dynamics. In this talk we will present an overview of the results of our ongoing big data project aimed to address some of these challenges by developing effective methods for structured regression with propagating uncertainty in evolving networks. The proposed methods will be discussed in context of applications to predicting admission and mortality rate for high impact diseases at a large number of hospitals.

Biography:

Zoran Obradovic is a L.H. Carnell Professor of Data Analytics at Temple University, Professor in the Department of Computer and Information Sciences with a secondary appointment in Statistics, and is the Director of the Center for Data Analytics and Biomedical Informatics. He is the executive editor at the journal on Statistical Analysis and Data Mining, which is the official publication of the American Statistical Association and is an editorial board member at eleven journals. He is the chair at the SIAM Activity Group on Data Mining and Analytics and was co-chair for 2013 and 2014 SIAM International Conference on Data Mining and was the program or track chair at many data mining and biomedical informatics conferences. His work is published in more than 300 articles and is cited more than 15,000 times (H-index 48). For more details see http://www.dabi.temple.edu/~zoran/

Mathematical Institute of the Serbian Academy of Sciences and Arts
Workshop on Mathematical Data Science
Monday, June 22, 2015 at 11:30 - 11:50, room 301f

Predictive, Preventive, and Personalized Medical Platform based on Smart Health Records and Enriched Clinical Knowledge

Speaker: Milan Vukicevic, University of Belgrade, Faculty of Organizational Sciences, Belgrade, Serbia

Abstract: Recent efforts in personalizing health care and increased availability of personal devices (activity trackers, wrist bands etc.) allows closer interaction between doctors and patients through telehealth, home visits systems, etc. This resulted in improved primary and secondary healthcare, as well as an increased generation of healthcare data. Currently developed data analysis platforms are mostly providing support for data collection and monitoring, leaving all the analysis and decision making to the clinicians and healthcare givers. Unfortunately, physicians are often overwhelmed and are not in a position to constantly monitor and process the large volumes of data that each patient generates (it is anticipated that in 2025 there will be a shortage of 90000 doctors in the US alone) and this is often an important reason for wrong or late diagnoses and care. High complexity of healthcare problems (e.g. multi-modality of the data, partially observed data, frequent context changes etc) often leads to unreliable predictive algorithms that cannot be used as a decision support for medical decision making, where potentially wrong decisions have high human and financial costs. This leads to a situation where doctors and patients cannot benefit from data driven models and data driven models do not exploit existing domain knowledge, leaving a **large gap between actual data usage and potential data usage** in healthcare that prevents a paradigm shift from delayed interventional to predictive person-tailored medicine. An important reason for this is that predictive models try to learn such complex systems from scratch based only on data, while disregarding available domain knowledge. We address this problem by fusion of domain knowledge sources (e.g. ontologies or expert systems) with heterogeneous data sources (e.g. Electronic Health Records or data streams) and utilize such enriched data in pre-processing and modeling phases of predictive analytics process. Since, interpretability of developed models is a must for healthcare applications the main focus on applications of interpretable models such as Generalized Regression models and Shaplet based stream analyses.

This is joint research with Data Analytics and Biomedical Informatics Center, Temple University and Center for Business Decision Making, University of Belgrade, Faculty of Organizational Sciences.

Prof. Veljko Milutinović

Life Member of the ACM
Fellow Member of the IEEE
Member of Academia Europaea
Member of the Serbian Academy of Engineering
Member of the Scientific Advisory Board of MindGenomics
Member of the Scientific Advisory Board of MaxelerTechnologies

DataFlow SuperComputing for BigData PetaAnalytics

This presentation analyses the essence of DataFlow SuperComputing, defines its advantages and sheds light on the related programming model. DataFlow computers, compared to ControlFlow computers, offer speedups of 20 to 200 (even 2000 for some applications), power reductions of about 20, and size reductions of also about 20. However, the programming paradigm is different, and has to be mastered. The talk explains the paradigm, using Maxeler as an example, and sheds light on the ongoing research in the field. Examples include DataEngineering, DataMining, FinancialAnalytics, etc. A recent study from Tsinghua University in China reveals that, for Shallow Water Weather Forecast, which is a BigData problem, on the 1U level, the Maxeler DataFlow machine is 14 times faster than the Tianhe machine, which is rated #1 on the Top 500 list (based on Linpack, which is a smalldata benchmark). Given enough time, the talk also gives a tutorial about the programiing in space, which is the programming paradigm used for the Maxeler dataflow machines (established in 2014 by Stanford, Imperial, Tsinghua, and the University of Tokyo). The talk is concluded with selected examples (appgallery.maxeler.com and webIDE.maxeler.com).

Zoran Ognjanović, Miodrag Rašković, Zoran Marković

Mathematical Institute, Serbian Academy of Sciences and Arts, Belgrade

Abstract. Formal reasoning with uncertain knowledge is an ancient problem dating, at least, from Leibnitz. Recently we are experiencing a growing interest in the field motivated by applications to CS and AI. Some of the formalisms for representing, and reasoning with, uncertain knowledge are based on probabilistic logics, that extend the classical logic calculus with probabilistic operators. We give a survey of different probability logics, original mathematical techniques, and the results of the authors including solutions of some problems from literature. We also present formalizations of more general types of probability functions (with values not in the unit interval of reals), possible applications of probability logics to real-world reasoning (default and spatiotemporal reasoning, measuring of inconsistency etc.) and computationally efficient heuristic procedures for testing satisfiability of probability formulas are described.

Zoran Ognjanović is a research professor at the Mathematical Institute of the Serbian Academy of Sciences and Arts. His research interests concern: applications of mathematical logic in computer science, artificial intelligence and uncertain reasoning, automated theorem proving, applications of heuristics to satisfiability problem and digitization of cultural and scientific heritage. He is a recipient of the Serbian Academy of Sciences and Arts Award in the field of mathematics and related sciences for 2013 and the annual award of Serbian Ministry of Science for results in fundamental research in 2004. He has authored or coauthored over 60 technical papers and chapters in monographs.

Miodrag Rašković is a research professor at the Mathematical Institute of the Serbian Academy of Sciences and Arts. He spent several months at the University of Wisconsin, under supervision of J. Keisler. His research interests concern: mathematical logic and uncertain reasoning. He is one of the authors of the following books: Probability quantifiers and operators, Stories about small and big numbers (in Serbian), Non-standard analysis (in Serbian). He is the chairmen of the seminar for Probability logic at the Mathematical Institute.

Zoran Marković is a research professor at the Mathematical Institute of the Serbian Academy of Sciences and Arts. He received the B.A. (in 1971) and M.A. (in 1974) degrees from Faculty of Mathematics of University of Belgrade, and Ph.D. (in 1979) from University of Pennsylvania, Philadelphia. He had visiting positions at University of California, Berkeley, University of California Davis, and University of Amsterdam (Institute for Logic, informatics and linguistics). His research interest in mathematics concerns intuitionistic, probabilistic and non-monotonic logics, and their applications in intelligent reasoning. He is a recipient of the Serbian Academy of Sciences and Arts Award in the field of mathematics and related sciences for 2013.

Miodrag Mihaljević

Mathematical Institute, Serbian Academy of Sciences and Arts, Belgrade

Advanced Encryption Approaches for Machine-to-Machine Communications and Big Data Processing

Abstract:

Current information-communication technologies (ICT) heavily involve machine-to-machine (M2M) communications and big data processing. These two issues open a number of mathematical challenges regarding data privacy and secrecy. A particular challenge is developing of algorithms which support reduction of the overheads implied by employed techniques for data security. Frequent employment of the security mechanisms could result into a heavy cumulative overhead regarding implementation complexity, computational complexity and power consumption. At the same time we face a request for a high and preferably provable security of the employed cryptographic techniques. Accordingly, the advanced cryptographic techniques should at the same time provide high security and as small as possible overheads to the main functionality of a system. This talk considers an encryption approach for fulfilling the claimed goals based on involvement elements of coding theory into traditional compact cryptographic techniques for encryption in order to provide a secure and lightweight processing of data for the privacy and secrecy purposes. We discuss certain gains which can be achieved based on employment results from coding theory regarding channels with synchronization errors and the wire-tap channel coding. It is shown that joint employment of pseudo-randomness generated by compact finite-state machines, randomness and dedicated coding provide a framework for developing of lightweight and provable secure encryption algorithms for data privacy and secrecy.

Mathematical Institute of the Serbian Academy of Sciences and Arts
Workshop on Mathematical Data Science
Monday, June 22, 2015 at 11:30 - 11:50, room 301f

Clustering comunity networks by Variable neighborhood search

Speaker: Nenad Mladenovic, Mathematical Institute, SANU, Belgrade, Serbia. (joint work with S. Caffieri and P. Hansen)

Abstract: Complex systems in a variety of domains are represented by networks. The most prominent examples include social networks, describing individuals and their interactions and relationships, telecommunication networks, transportation networks, biological networks, and many more. A modular structure characterizes many complex systems, which contain subgroups of entities sharing some common properties. A topic of particular interest in the study of complex networks is therefore the identification of modules, also called *clusters* or *communities*. This is very useful to identify some properties of the system described by the studied network starting from its structural features.

Let us consider a graph $G = (V, E)$, with V the set of vertices and E the set of edges, used to represent a network. Several models and clustering criteria have been proposed. One often maximizes or minimizes a criterion function. The most used is *modularity*, based on the idea of comparing the fraction of edges falling within communities to the expected fraction of such edges. Recall that the degree k_i of a vertex i belonging to V is the number of its neighbors (or adjacent vertices). Let $S \subseteq V$ be a subset of vertices. Then the degree k_i can be separated into two components $k_i^{in}(S)$ and $k_i^{out}(S)$, i.e., the number of neighbors of i inside S and the number of neighbors of i outside S. A set of vertices S forms a community in the *strong sense* if and only if every one of its vertices has more neighbors within the community than outside: $k_i^{in}(S) > k_i^{out}(S), \quad \forall i \in S$. A set of vertices S forms a community in the *weak sense* if and only if the sum of all degrees within S is larger than the sum of all degrees joining S to the rest of the network: $\sum k_i^{in}(S) > \sum k_i^{out}(S)$. This is equivalent to the condition that the number of edges within S is at least half the number of edges in the cut of S. These concepts inspired the definition of the *edge-ratio* criterion. More precisely, the definition of community in the weak sense is extended into a criterion for a bipartition to be optimal: one seeks to maximize the minimum for both classes of the bipartition of the ratio of inner edges to cut edges. Specifically, the ratio of the number of edges within a community to the number of cut edges which have one end point only within that community is considered: $r(S) = \sum_{i \in S} k_i^{in}(S) / \sum_{i \in S} k_i^{out}(S)$.

In this paper we propose a heuristic based on (basic) Variable Neighborhood Search to perform the bipartitioning step in a hierarchical divisive algorithm based on the edge-ratio network clustering criterion, which was shown to be an alternative approach to modularity criterion. Neighborhoods to be used in the VNS are defined using the Hamming distance. We develop a local search for the addressed problem and evaluate its complexity. Computational results show that VNS allows us to obtain good quality results significantly reducing the computational time needed to perform bipartitioning steps.

Mathematical Institute of the Serbian Academy of Sciences and Arts
Workshop on Mathematical Data Science
Monday, June 22, 2015 at 13:40 - 14:00, room 301f

Predictive models based on support vector machines for structured outputs

Speaker: Jovana Kovacevic (co-author: Gordana Pavlovic-Lazetic), University of Belgrade, Faculty of Mathematics, Belgrade, Serbia

Abstract: Structured output prediction defines a family of problems where the aim is to teach a function to predict complex objects such as sequences, trees or graphs. More formally, it can be stated as follows: given a set of training examples $S = \{(x_1, y_1),...,(x_n, y_n)\} \in (X \times Y)^n$, we want to learn a prediction function $f(x) = \arg\max_{y \in Y} F(x, y; w)$, where the function F is a scoring function that measures how well a particular output $y \in Y$ matches an input $x \in X$. It is parameterized by the vector w which is learned during the training process. Choosing the algorithm for calculation of $\arg\max$ is especially challenging since the output set Y is often of high cardinality. In this research, the focus is on SSVM algorithm that belongs to the group of discriminative modelling techniques. It represents a generalization of the well-known SVM algorithm on structured outputs. We applied SSVM on two different domains: protein function prediction and hierarchical text categorization.

With large number of genomes being sequenced every year, there is a growing number of newly discovered proteins. Fast and accurate information on protein function is especially important in context of human diseases, since many of them can occur due to functional mutation. One approach to protein function prediction is to find one or more functions that the protein performs in a cell using only its primary sequence as input data. The space of all known protein functions is structured as a directed acyclic graph, known as Gene Ontology (GO), with several thousand nodes, where each node is labelled by one function and each edge encodes the so called “is-a” relationship. Every output y represents the subgraph of GO, consistent in a sense that it contains protein’s functions propagated to the root. We have implemented a SSVM based predictor that determines protein function from histogram of amino acid 4-grams that appear in the protein primary sequence.

Hierarchical text categorization refers to assigning a text document to one or more most suitable categories from a hierarchical category space. Such a process has different useful applications including document organization, text filtering, spam detection, mail routing, news monitoring, automatic document indexing and a hierarchical catalogue of web resources. We have developed a hierarchical and a multiclass classifier, both SSVM based, using byte n-gram based documentation representation techniques, and applied them on several corpora of texts in different languages.

This is joint research between Bioinformatics research group, Department of Computer Science and Informatics, Indiana University and Bioinformatics research group, Faculty of Mathematics, Belgrade University.