SOE 22: Focus Session: Big Data (joint with jDPG)

The availability of large-scale data invades all areas of econophysics, sociodynamics, as well as bioinformatics and poses methodical challenges for data analysis, visualization and modeling. This session provides an overview how methods adapted from statistical physics and network analysis deepen the understanding of the interaction of humans through language and social media, their emergent collective behaviour the assessment of risks and the detection of crises. (Session compiled by Kerstin Kämpf, TU Darmstadt and Jens Christian Claussen, U Lübeck.)

Time: Thursday 9:30–12:30 Location: H37

Topical Talk SOE 22.1 Thu 9:30 H37

Physics and the Information Society: Turning Big Data into Big Insight — •René Pfitzner — ETH Zurich, Chair of Systems Design, Switzerland

As of today 2.3 Billion people are online, spend 1 Billion hours on the web, write 400 Mio. tweets and produce a total of 65 terabytes of Facebook content – every day.

Inspired by these impressive numbers, in this talk I will illustrate two main points about "Big Data" research. First, based on examples I will show how the availability of large amounts of "online" data facilitates research to gain insights into "offline" phenomena like disease spreading. Second, I will show that in complex information systems, composed of interacting social and technical components, non-trivial questions about information propagation, the emergence of memes or measuring the relevance of content occur.

I will point to methodologies that have been developed, and continue to be developed, to cope with these research challenges and opportunities - often inspired by theories well known from the Physics literature.

Invited Talk SOE 22.2 Thu 10:00 H37

Network analysis literacy — •Katharina Anna Zweig — TU Kaiserslautern, Computer Science Department, Graph theory and complex network analysis, Gottlieb-Daimler-Str. 48, 67663 Kaiserslautern, Germany

Big data often comes in a form that relates objects or subjects to each other. Examples for this kind of data describe interactions between proteins or people, plane connections between cities, or references from articles to other articles. Relational data is best analyzed by network analytic measures which have been proven useful in very different disciplines; high hopes have been put in them to finally understand the complex systems surrounding us. While network analysis is often very successful, in this talk I will show that not all relational data should actually be represented as a network and that not all measures are likely to give reasonable results in all contexts. I will discuss the "trilemma of social network analysis" which puts an emphasis on matching the data and its network representation, the method to use, and the question to be answered.

Invited Talk SOE 22.3 Thu 10:30 H37

From Noise to Signal. Stories about big data. — •Sune Lehmann[1], Yong-Yeol Ahn[2], Alan Mislove[3], Jukka-Pekka Onnela[4], and Niels James Rosenquist[5] — [1]Technical University of Denmark, Kgs Lyngby, Denmark — [2]Indiana University, Bloomington Indiana — [3]Northeastern University, Boston, MA, USA — [4]Harvard School of Public Health, Boston, MA, USA — [5]Mass General Hospital, Boston, MA

This talk tells the story of how we used over 300 million tweets (Sep 2006 - Aug 2009) to map the collective mood of the United States. The mood of each tweet was inferred using a simple word-list (ANEW), and the results are represented as density-preserving cartograms. A cartogram is a map in which the mapping variable (in this case, the number of tweets) is substituted for the true land area. Thus, the geometry of the actual map is altered so that the shape of each region is maintained as much as possible, but the area is scaled in order to be proportional to the number of tweets that originate in that region. For the final part of the talk, we will discuss the importance of visualization in analysis of Big Data as well as new developments in the area of Big Data.

SOE 22.4 Thu 11:00 H37

Geopolitical risk-index derived from 60 million news articles predicts war — •Thomas Chadefaux — Chair of sociology, modeling and simulation, ETH Zurich, Clausiusstrasse 50, 8092 Zurich, Switzerland

There have been more than 200 wars since the start of the 20th century, leading to about 35 million battle deaths. However, efforts at forecasting conflicts have so far performed poorly for lack of fine-grained and comprehensive measures of geopolitical tensions. Here, we developed a weekly risk-index by analyzing a comprehensive dataset of historical newspaper articles for 166 countries over the past century, which we then tested on a data of all conflicts within and between countries recorded since 1900. Using only information available at the time, we could predict the onset of a war within the next year with up to 85% confidence; we also forecasted over 70% of large-scale wars, while issuing false alarms in only 16% of observations. Predictions were improved up to one year prior to interstate wars, and six months prior to civil wars, giving policy-makers significant additional warning time.

SOE 22.5 Thu 11:15 H37

Big data; fame and money, box office prediction based on Wikipedia activity data — Márton Mestyán[1], •Taha Yasseri[1,2,3], and János Kertész[1,3,4] — [1]Institute of Physics, Budapest University of Technology and Economics, Budapest, Hungary — [2]Oxford Internet Institute, University of Oxford — [3]Department of Biomedical Engineering and Computational Science, Aalto University, Aalto, Finland — [4]Center for Network Science, Central European University, Budapest, Hungary

Use of socially generated Big Data to predict the collective reaction of individuals in societies to a certain event or product has become of great interest in recent years. In this work [1], we investigate the possibility of making precise predictions for the financial success of movies, by monitoring activity and the traffic on Wikipedia articles on the movies. We consider a sample of 312 movies released in the USA market in 2010, and show that, by using a minimalistic linear regression model, one could easily outperform the existing prediction methods. Our model, free of any content analysis, reaches a coefficient of determination of 0.92, one month prior to the movie release.

[1] Márton Mestyán, Taha Yasseri, and János Kertész, Early Prediction of Movie Box Office Success based on Wikipedia Activity Big Data, preprint available at: arXiv:1211.0970.

Topical Talk SOE 22.6 Thu 11:30 H37

Information Retrieval, Applied Statistics and Mathematics on BigData — •Romeo Kienzler — IBM Innovation Center Zurich, Switzerland

Although the majority of algorithms used for BigData Analytics have been developed decades ago, their application on BigData currently experiences a renaissance. In this talk a selection of algorithms and their application will be discussed in the context of BigData. A set of selected Use Cases in the field of Social Network Analysis, Bioinformatics, Financial Fraud Detection and Information Retrival will be discussed. Besides theoretical viewpoints this talk covers also runtime environments for BigData application and explains concepts of data parallelism, partition skew, aggregated storage to CPU bandwidth and fault tolerance on commodity hardware. Besides the omnipresent MapReduce/Hadoop example, which seems to be the de facto standard, we will also discuss massive parallel data warehousing and stream computing. Finally, a technical outlook tries to separate theory from reality, future from presence and hype from vision.

Invited Talk SOE 22.7 Thu 12:00 H37

Web-Based Cognitive Science: Harnessing the Power of the Internet to Study Human Cognition — •Christopher Y. Olivola — University of Warwick, UK

The Internet provides a unique and powerful tool for the social sciences, allowing researchers to collect data and carry out experiments at scales that were previously unfeasible. So far, web-based social science has mainly focused on aggregate behaviors and large-scale phenomena. In contrast, the enormous potential of the Internet has been much less utilized by behavioral scientists studying individual behaviors and their underlying cognitive processes. In this presentation, I will discuss several examples of how web-based research methods can both aid the study of cognition and directly clarify its contents. In particular, I will highlight 4 distinct ways in which cognitive scientists can utilize the Internet: (1) running web-based experiments with large samples of human participants (at low cost); (2) using online games to collect data from intrinsically motivated participants (for free); (3) studying *naturally* occurring online individual behaviors; (4) measuring the contents of memory and the dynamics of attention over time. I will conclude by discussing how students and researchers with a quantitative background (e.g., physicists) can utilize the web to advance our understanding of human cognition.