TUT 2: Statistical Physics Methods for Data Science in Physics (joint session SOE/DY/TUT)

Big Data is an ubiquitious buzzword, but beyond storing and processing large datasets, challenges in applications often lie in high-dimensional and nontrivial structures (= the content!) within the datasets. Artificial intelligence approaches, paired with statistical physics methods, can provide powerful tools which can go far beyond standard statistical methods. The tutorial gives an overview both on methods from stochastic blockmodeling, network analysis, inference and machine learning and on applications ranging from socioeconomic networks to biomolecular simulations. (Session organized by Jens Christian Claussen and Andreas Fery with the divisions of SOE, DY, BP and CPP.)

Time: Sunday 16:00–18:30 Location: H3

Tutorial TUT 2.1 Sun 16:00 H3

Statistical network inference and community detection — •Tiago Peixoto — University of Bath, UK

Network structures are shaped by evolutionary mechanisms and determine the central aspects of how a system functions. However, differently from systems that are naturally embedded in space, we cannot simply "look" at network in order to extract its most important structural patterns. Instead, we must rely on well-founded algorithmic methods to extract this information from data in an interpretable way. In this tutorial, we review a principled approach to this problem based on the elaboration of probabilistic models of network structure, and their statistical inference from empirical data. We focus in particular on the detection of modules (or "communities") in networks.

We aim to cover the following topics: 1. The stochastic block model (SBM) and its variants (degree correction, overlapping groups, etc.); 2. Bayesian inference and model selection: Distinguishing structure from noise; 3. Generalizing from data: Prediction of missing and spurious links; 4. Model extensions: Layered, dynamic SBMs, and generalized models on continuous latent spaces; 5 Fundamental limits of inference, and the undetectability phase transition; 6. Efficient inference algorithms; 7. Network reconstruction from noisy or indirect data.

Tutorial TUT 2.2 Sun 16:50 H3

Network filtering for big data — •Tiziana Di Matteo — Department of Mathematics - King's College London

In this lecture I will present network-theoretic tools to filter information in large-scale datasets and I will show that these are powerful tools to study complex datasets. In particular I will introduce correlation-based information filtering networks and the planar filtered graphs (PMFG) and I will show that applications to financial data-sets can meaningfully identify industrial activities and structural market changes. It has been shown that by making use of the 3-clique structure of the PMFG a clustering can be extracted allowing dimensionality reduction that keeps both local information and global hierarchy in a deterministic manner without the use of any prior information. To advance the PMFG (currently O(N3)), I will introduce a new algorithm, the TMFG (Triangulated Maximally Filtered Graph), that efficiently extracts a planar subgraph which optimizes an objective function. The method is scalable to very large datasets and it can take advantage of parallel and GPUs computing. The method is adaptable allowing online updating and learning with continuous insertion and deletion of new data as well changes in the strength of the similarity measure. Finally I will also show that filtered graphs are valuable tools for risk management and portfolio optimization too and they allow to construct probabilistic sparse modeling for financial systems that can be used for forecasting, stress testing and risk allocation.

Tutorial TUT 2.3 Sun 17:40 H3

Multiscale simulations of soft matter augmented by data-driven methods — •Tristan Bereau — Max Planck Institute for Polymer Research

Multiscale simulations, all the way from quantum chemistry to continuum mechanics, probe a variety of length and time scales relevant to soft-matter systems. In this tutorial, I will describe different strategies to help improve physics-based simulations with recently-developed data-driven methods and concepts. Applications discussed will include Bayesian inference for molecular kinetics, machine learning to vastly improve force-field transferability, and high-throughput screening to explore chemical compound space.