NetInf'17: SDM Workshop on Inferring Networks From Non-Network Data[*]

Ivan Brugere[†][‡] Rajmonda Caceres[§] Brian Gallagher[¶] Tanya Y. Berger-Wolf[†]

1 Introduction

Network representation learning is an emerging field exclusively focused on developing the understanding and rigor for learning useful network representations from *noisy*, *indirect*, and *diverse* data measurements. These data are prevalent in science and industry in the form of genetic expression, brain imaging, geotagged and geo-social data, and user behavioral/interaction data on the web, among others.

Researchers aim to learn a network representation of these data because a network is a convenient and powerful model for investigating the system or population in question. However, researchers are often faced with having to make arbitrary decisions on how to construct networks from underlying data. Such arbitrary decisions and the ad-hoc representations may lead to have important implications on the performance of subsequent learning tasks and decisions. There is a need for novel, rigorous methods in network construction and validation, and an investigation of the challenges and pitfalls along the network science pipeline, from underlying data to domain investigation.

This workshop will explore recent developments in the area of network representation learning including:

1. Multi-modal and heterogeneous techniques that lead to robust network representations
2. Strategic and adaptive data collection mechanisms for discovering the network
3. Task-oriented network representation learning
4. Validation of network representations in the absence of ground truth

This area is related to several communities working in broader network science, including dynamic networks, parametric network modeling and relational learning and statistical inference *on* networks.

2 Keynotes

2.1 On the Impact of Network Inference on Network Science: Propagation of Uncertainty

Eric Kolaczyk (Boston University)

The inference of network connectivity from low-level measurements on a complex system of interest is both a long-standing problem and a problem of intense current interest. There has been much formal work done in certain areas on certain variants of this problem. But, nevertheless, a substantial amount of current practice rests on approaches to network inference that are, in part or in whole, ad hoc. Whether formal or informal, however, in either case there is generally uncertainty in the networks we infer. And particularly undeveloped to date is an understanding of just how that uncertainty propagates to high-level tasks in network science (e.g., summary statistics, community detection, decision making, etc.). In this talk I take a statistical perspective on these issues, first providing some context and background on formal approaches to network inference, followed in turn by a summary of our recent work on post-inference uncertainty quantification.

2.1.1 Bio

Eric Kolaczyk is Professor of Statistics, and Director of the Program in Statistics, in the Department of Mathematics and Statistics at Boston University, where he also is an affiliated faculty member in the Program in Bioinformatics, the Program in Computational Neuroscience, and the Division of Systems Engineering. Prof. Kolaczyk's main research interests currently revolve around the statistical analysis of network-indexed data, and include both the development of basic methodology and inter-disciplinary work with collaborators in bioinformatics, computer science, geography, neuroscience, and sociology.

Besides various research articles on these topics, he has also authored two books in this area. He has given various short courses on material from his books in recent years, including for the Center for Disease Control (CDC) and the Statistical and Applied Mathematical Sciences Institute (SAMSI) in the US as well as similar venues in Belgium, England, and France. Prior to his working in the area of networks, Prof. Kolaczyk spent a decade working on statistical multi-scale modeling.

[*] compbio.cs.uic.edu/netinf17/

[†] University of Illinois at Chicago

[‡] Correspondence: ibruge2@uic.edu

[§] MIT Lincoln Laboratory

[¶] Lawrence Livermore National Laboratory

Prof. Kolaczyk has served as associate editor on several journals, including currently the Journal of the American Statistical Association and the IEEE Transactions on Network Science & Engineering. He has also served as co-organizer for workshops focused on networks and network data. He is an elected fellow of the American Statistical Association (ASA), an elected senior member of the Institute for Electrical and Electronics Engineers (IEEE), and an elected member of the International Statistical Institute (ISI).

2.2 Collective Graph Identification

Lise Getoor (University of California, Santa Cruz)

Graph data (e.g., communication data, financial transaction networks, data describing biological systems, collaboration networks, the Web, etc.) is ubiquitous. While this observational data is useful, it is usually noisy, often only partially observed, and only hints at the actual underlying social, scientific or technological structures that give rise to the interactions. For example, an email communication network provides useful insight, but is not the same as the "real" social network among individuals. In this talk, I introduce the problem of graph identification, i.e., the discovery of the true graph structure underlying an observed network. This involves inferring the nodes, edges, and node labels of a hidden graph based on evidence provided by the observed graph. I show how this can be cast as a collective probabilistic inference task and describe a scalable approach to solving this problem.

2.2.1 Bio

Lise Getoor is a Professor in the Computer Science Department at the University of California, Santa Cruz. Her research areas include machine learning, data integration and reasoning under uncertainty, with an emphasis on graph and network data. She has over 200 publications and extensive experience with machine learning and probabilistic modeling methods for graph and network data. She is a Fellow of the Association for Artificial Intelligence, an elected board member of the International Machine Learning Society, serves on the DARPA ISAT Study Group (2016-2019) and the board of the Computing Research Association (CRA), and was co-chair for ICML 2011. She is a recipient of an NSF Career Award and eleven best paper and best student paper awards. In 2014, she was recognized by KDD Nuggets as one of the emerging research leaders in data mining and data science based on citation and impact. She received her PhD from Stanford University in 2001, her MS from UC Berkeley, and her BS from UC Santa Barbara, and was a Professor in the Computer Science Department at the University of Maryland, College Park from 2001-2013.

3 Accepted Papers

3.1 A General Framework For Task-Oriented Network Inference

Ivan Brugere (University of Illinois at Chicago)

Chris Kanich (University of Illinois at Chicago)

Tanya Y. Berger-Wolf (University of Illinois at Chicago)

We present a brief introduction to a flexible, general network inference framework which models data as a network space, sampled to optimize network structure to a particular task. We introduce a formal problem statement related to influence maximization in networks, where the network structure is not given as input, but learned jointly with an influence maximization solution.

3.2 Network-Based Structure Analysis and Event Detection in the New York City Taxi Dataset

Jose Cadena (Virginia Tech)

Goran Konjevod (Lawrence Livermore National Laboratory)

Giuliana Pallotta (Lawrence Livermore National Laboratory)

Urban datasets have become available to the public in recent years. Among those, taxi trip records are of particular interest because taxis can act as sensors and provide insights into the economical and social activity in a city. Researchers have already been able to extract from taxi data useful indicators for a variety of tasks, including the inference of pollution emissions and the estimation of the resilience of the traffic system to disruptions (e.g., weather events). However, the underlying structure of mobility interactions has not been studied. We analyze a large dataset of taxi trips in New York City covering a span of 7 years and address the following questions: Can we partition the city into groups of locations with similar mobility patterns? Do these groups persist at different spatio-temporal resolutions? Can we detect interesting events based on changes in the mobility patterns? To address these questions, we present a network-based approach for analyzing taxi trip data in terms of interactions in a suitably constructed graph.

3.3 An Iterative Graph-Theoretic Approach for Filtering Noisy Relationships in Correlation Networks

Kathryn Cooper (University of Nebraska at Omaha)

Sanjukta Bhowmick (University of Nebraska at Omaha)

Hesham Ali (University of Nebraska at Omaha)

Inferring domain-specific signals from large-scale raw data is one of the most challenging problems in many application domains. The analysis and representation of biological data obtained from todays advanced high-throughput experimentation presents a glaring example of this challenge. In the absence of straightforward deterministic approaches to identify simple causative signals in complex biological data, correlation network analysis represents an attractive alternative method for inferring regulatory relationships and extracting meaningful knowledge from the available data. However, there exists a distinct lack of recognized validation in many studies utilizing correlation networks that has limited trust in their impact.

The uncertainty regarding applicability of the correlation network in biological applications exists in whole or in part due to a failure to study important parameters in this process such as sample number, count, or type, and how these parameters affect the final network model. We believe that studying these parameters is critical to developing robust networks that will provide consistently the same results. In particular, recent studies suggest that up to 50% of the relationships modeled by correlation networks created with small samples are highly sensitive to changes in sample counts. Due to the possibility of such wide-ranging differences, it is critical to ensure the robustness of the networks under noise and the quality of the analysis results.

In this paper, we propose a methodology to measure the quality of the results obtained from the analysis of correlation networks. We investigate how individual samples affect the robustness of the network model, and develop a schema to test the impact of individual sample removal as the computational load grows. Results obtained from the proposed approach can be utilized to improve trust level of knowledge extracted from correlation analysis and allow biomedical researchers to take advantage from the trusted information extracted from correlation network models.

3.4 Network-based Anomaly Detection for Insider Trading

Adarsh Kulkarni (George Mason University)
Priya Mani (George Mason University)
Carlotta Domeniconi (George Mason University)

Insider trading is one of the numerous white collar crimes that can contribute to the instability of the economy. Traditionally, the detection of illegal insider trades has been a human-driven process. In this paper, we collect the insider trade filings made available by the US Securities and Exchange Commissions (SEC) through the EDGAR system, with the aim of initiating an automated large-scale and datadriven approach to the problem of identifying illegal insider trading.

The goal of the study is the identification of interesting patterns, which can be indicators of potential anomalies. We use the collected data to construct networks that capture the relationship between trading behaviors of insiders. We explore different ways of building networks from insider trading data, and argue for a need of a structure that is capable of capturing higher order relationships among traders. Our results suggest the discovery of interesting patterns.

3.5 Making #Sense of #Unstructured Text Data

Lin Li (MIT Lincoln Laboratory)
William M. Campbell (MIT Lincoln Laboratory)
Cagri Dagli (MIT Lincoln Laboratory)
Joseph P. Campbell (MIT Lincoln Laboratory)

Many network analysis tasks in social sciences rely on pre-existing data sources that were created with direct relations or interactions between entities in consideration. Examples include email logs, friends and followers network on social media, communication networks, etc. In these data, it is relatively easy to identify who is connected to whom and how they are connected. However, most of the data that we encounter on a daily basis are unstructured free-text data, e.g., social media posts, forums, online marketplaces, etc. It is considerably more difficult to extract network data from unstructured text. In this work, we present an end-to-end system for analyzing unstructured text data and transforming the data into structured graphs that are directly applicable to the downstream application. For text analysis, we focus on unsupervised methods to extract informative content or topic words. Specifically, we look at social media data and attempt to predict hashtags for users posts. The resulting hashtags can be used for downstream processing such as graph construction and analysis. With that goal in mind, we apply our methods to the application of cross-domain entity resolution. The performance of the resulting system using automatic hashtags shows improvement over the system using the original user-annotated hashtags.

3.6 Graph Model Selection via Random Walks

Lin Li (MIT Lincoln Laboratory)
William M. Campbell (MIT Lincoln Laboratory)
Rajmonda S. Caceres (MIT Lincoln Laboratory)

In this paper, we present a novel approach based on the random walk process for finding meaningful representations of a graph model. Our approach leverages the transient behavior of many short random walks with novel initialization mechanisms to generate model discriminative features. These features are able to capture a more comprehensive structural signature of the underlying graph model. The resulting representation is invariant to both node permutation and the size of the graph, allowing direct comparison between large classes of graphs. We test our approach on two challenging model selection problems: the discrimination in the sparse regime of an Erdös-Renyi model from a stochastic block model and the planted clique problem. Our representation approach achieves performance that closely matches known theoretical limits in addition to being computationally simple and scalable to large graphs.

3.7 Estimation and Inference of Network Centrality for Covariance based Network Models

Manjari Narayan (Stanford University)

Network science offers a variety of centrality functions to rank the importance or "centrality" of each node in the network. However, in many biological applications, such as molecular biology or functional neuroimaging, networks are not directly observed. Rather, networks are statistical relationships estimated from other observations such as time-series. Thus, from a statistical viewpoint, classical network centrality defined on a known adjacency matrix is in fact an unknown population parameter. Instead, sample network centrality obtained using network estimates are a function of noisy measurements, and should be treated as random variables with an unknown sampling distribution. Consequently, inferring that one node is more central than another node based on uncertain sample network centrality values can result in flawed scientific inferences. To address this problem, we propose a novel nonparametric stability-ranking procedure to quantify the uncertainty in sample network centrality of individual nodes.

Using a dataset from the Human Connectome (HCP) project and eigenvector centrality as a canonical instance of network centrality, we illustrate the difficulty of distinguishing node ranks when networks are estimated from finite sample sizes. The HCP dataset highlights the large empirical variability in ranking nodes by network centrality in brain networks. Furthermore, the presence of overlapping confidence intervals between pairs of empirical ranks reveals that many nodes are not distinguishable with respect to their rankings. In instances where individual node ranks cannot be distinguished, our procedure alternatively quantifies the stability of node ranks within the top-k set.

3.8 Scalable Inference of Functional Networks from Time Series Data

Tara Safavi (University of Michigan)
Chandra Sripada (University of Michigan)
Danai Koutra (University of Michigan)

Among scientific data, which are being generated at an ever-increasing rate, graph structures occur frequently and naturally. Such graphs are not always directly observed. For instance, in imaging neuroscience, monitored brain activity based on changes in blood flow can be leveraged to model the brains organization as a functional network, wherein time series of brain activity across regions of interest are connected according to measures of similarity or dependency. However, existing methods to this end suffer from a lack of scalability, which is an increasingly pressing concern as technology advances and the volume of medical data increases: in the future it is expected that functional connectomes will have hundreds of thousands or even millions of nodes. Beyond scalability, other challenges of functional network construction in neuroscience include the inherent noisiness of the series data, the often-arbitrary thresholding of edge weights or similarity measures, and the related issue of differing output graph representations based on the methods or parameters chosen in the network construction process.

Using the study of brain network organization as a focal point, we thus propose to explore and compare methods of inferring functional graphs from time series by removing the effect of the series magnitude prior to graph construction, as well as defining distance metrics on the preprocessed series that facilitate fast approximate similarity search while avoiding all pairwise comparisons. This work has the potential to contribute to improved identification of biomarkers for mental disease, as research has shown that human cognitive dysfunction may be indicated and/or explained by anomalous or disturbed functional brain connectivity.

4 Program Committee

Sanjukta Bhowmick (University of Nebraska at Omaha)
David Gleich (Purdue University)
Christine Klymko (Lawrence Livermore National Laboratory)
Benjamin Miller (MIT Lincoln Laboratory)
Evangelos E. Papalexakis (University of California, Riverside)
Dane Taylor (University of North Carolina-Chapel Hill)
Elena Zheleva (AAAS Fellow at National Science Foundation)

5 Acknowledgements

We would like to thank the SDM'17 organizers and SDM'17 Workshop Chairs, our keynote speakers, the Program Committee for providing tremendous feedback to authors, and the authors and participants.