Anima Anandkumar (UC Irvine)

Learning mixed membership community models via spectral methods

Abstract: Learning hidden communities in networks is a problem that has applications in numerous domains such as social networks, national security and computational biology. Mixed membership models relax the assumption that a node belongs to a single community, but are harder to train. Popular approaches such as variational inference or Gibbs sampling are not scalable to large datasets and do not have guarantees of recovering the correct model parameters. In contrast, spectral methods are parallelizable and provide consistency guarantees under mild assumptions. Our recent results show that decomposition of tensors involving counts of subgraphs such as 3-stars can learn mixed membership models under tight conditions. I will also discuss extensions to hypergraph models such as social tagging networks.

Chrisitian Borgs (Microsoft Research, New England)

Applications of L^p sparse graphons to machine learning of large networks

Abstract: When analyzing large networks, statisticians often assume a generative model in which the observed graph is assumed to come from a stochastic block model, i.e., a random graph with inhomogeneous edge probabilities given in terms of a small block matrix. A non-parametric version of these stochastic block models are so-called W-random graphs, given in terms of an integrable, symmetric two variable function W over some feature space equipped with a probability measure encoding a prior over features. In this talk we discuss the question of how to recover a good approximation to W from just a single sample of a W-random graph, and relate it to the theory of convergence of sparse graphs. We also show how to use this theory to obtain node differentially private versions of the non-parametric stochastic block model, enabling one of the strongest results on node differential privacy.

Joint work with Jennifer Chayes, Henry Cohn, Shirshendu Ganguly, and Adam Smith.
Jennifer Chayes (Microsoft Research, New England)

L^p theory of sparse graphons

Abstract: We introduce and develop a theory of limits for sequences of sparse graphs based on L^p graphons, which generalizes both the existing L^∞ theory of dense graph limits and its extension by Bollobas and Riordan to sparse graphs without dense spots. In doing so, we replace the no dense spots hypothesis with weaker assumptions, which allow us to analyze a much larger class of graphs, including those with power law degree distributions. This gives the first broadly applicable limit theory for sparse graphs with unbounded average degrees. We also show the equivalence of many nonlocal notions of sparse convergence. This talk assumes no prior knowledge of graphons.

Joint work with Christian Borgs, Henry Cohn, and Yufei Zhao.

Tina Eliassi-Rad (Rutgers)

Multi-armed Bandits for Enhancing Incomplete Networks

Abstract: Networked representations of physical and social phenomena are often incomplete because the phenomena is partially observed. For example, most of the publicly available data from online social networking services (such as Facebook and Twitter) are collected via apps, whatever the API allows, users who make their accounts public, and/or how much resources the researcher/practitioner has. Such incompleteness can lead to inaccurate findings. Consider, for example, the resultant community structure on a fraction of an online social network that was observed via random edge sampling. Hoping to acquire the complete data is often unrealistic, but we may be able to collect data selectively to enhance the incomplete network. We introduce the Adaptive Edge Probing problem. Suppose that one has observed a networked phenomenon via some form of sampling and one has a budget to enhance the incomplete network by asking for additional information about specific nodes, with the ultimate goal of obtaining the most “valuable” information about the network as a whole. Which nodes should be further explored? We show that multi-armed bandits are a good solution to this problem. Moreover, we present epsilon-WGX, a networked-based multi-armed bandit algorithm for identifying which nodes in the incomplete network should be probed. We compare epsilon-WGX to several competing algorithms on various datasets. Aggregated over multiple datasets and a wide range of probing budgets, we find that epsilon-WGX’s performance regularly outperforms other multi-armed bandit strategies and baseline probing strategies.

Joint work with Sucheta Soundarajan (Syracuse), Brian Gallagher (LLNL), and Ali Pinar (Sandia).
**Lise Getoor (UC Santa Cruz)**

Scalable Collective Reasoning in Graphs

Abstract: One of the challenges in big data analytics lies in being able to reason collectively about extremely large, heterogeneous, incomplete, and noisy interlinked data. We need mathematical frameworks and software tools that can represent and reason effectively with this rich form of multi-relational graph data. I will describe some common inference patterns needed for graph data including: collective classification (predicting node labels), link prediction (predicting potential edges), and entity resolution (determining when two nodes refer to the same underlying entity). I will describe some key capabilities required to solve these problems, and finally I will describe a mathematical framework for describing these problems and a highly scalable open-source probabilistic programming language being developed within my group to solve these challenges.

---

**Sharad Goel (Stanford)**

“Going Viral” and the Structure of Online Diffusion

Abstract: New products, ideas, norms and behaviors are often thought to propagate through a person-to-person diffusion process analogous to the spread of an infectious disease. Until recently, however, it has been prohibitively difficult to directly observe this process, and thus to rigorously quantify or characterize the structure of information cascades. In one of the largest studies to date, we describe the diffusion structure of billions of events across several domains. We find that the vast majority of cascades are small, and are characterized by a handful of simple tree structures that terminate within one degree of an initial adopting "seed." While large cascades are extremely rare, the scale of our data allows us to investigate even the one-in-a-million events. To study these rare, large cascades, we develop a formal measure of what we label "structural virality" that interpolates between two extremes: content that gains its popularity through a single, large broadcast, and that which grows via a multi-generational cascade where any one individual is directly responsible for only a fraction of the total adoption.

We find that online diffusion is characterized by surprising structural diversity, with popular events regularly growing via both broadcast and viral mechanisms, as well as essentially all conceivable combinations of the two. Finally, we attempt to replicate these findings with a model of contagion characterized by a low infection rate spreading on a scale-free network. We find that although several of our empirical findings are consistent with such a model, it fails to replicate the observed diversity of structural virality, thereby suggesting new directions for future modeling efforts.
Alexander Ihler (UC Irvine)

Graph-based reasoning for crowdsourcing

Abstract: Obtaining expert information has always been a major bottleneck for machine learning methods. "Crowdsourcing" techniques such as Amazon's Mechanical Turk have become a popular mechanism to access the power of human intelligence, for example to label large datasets. However, this raises the computational task of properly aggregating the crowdsourced predictions provided by a collection of unreliable and diverse annotators.

We transform crowdsourcing into a standard inference problem on a graphical model, allowing us to apply scalable, graph-based message passing algorithms such as belief propagation (BP). We show that BP-based aggregation can be competitive with state-of-the-art methods, and in fact several existing methods can be viewed as special cases of BP. We also use the graph structure to study the effect of expert information collected either before the crowd ("control items") or afterwards ("clean up"). Our analysis gives some theoretical results as well as practical guidance on how to balance several inherent trade-offs in the use of expert advice in crowdsourcing.

Dmitri Krioukov (Northeastern)

Clustering means geometry in sparse networks

Abstract: Two common features of many large real networks are that they are sparse and have strong clustering, i.e., large number of triangles. In many growing networks, the average degree and clustering are roughly independent of the growing network size. Recently, latent-space random graph models have been used successfully to model these features of real networks, to predict missing and future links in them, and to explain their navigability, with applications to designing optimal routing strategies in large telecommunication networks such as the Internet. Yet it remains unclear to what degree these latent-space models are adequate representations of real networks, as they may have properties that real networks do not have, or vice versa.

We show that maximum-entropy random graphs in which all edges are independent Bernoulli random variables, and the expected numbers of edges and triangles at all vertices are fixed to the same values independent of the graph size, are approximately soft random geometric graphs on the real line. The approximation is exact in the limit of standard random geometric graphs with a sharp connectivity threshold. This result implies that a large number of triangles uniformly distributed across all vertices is not only a necessary but also sufficient condition for the presence of a latent/effective metric space in a large sparse graph.
Aleksander Madry (MIT)

Continuous Optimization: the "Right" Language for Fast Graph Algorithms?

Abstract: Traditionally, we view graphs as purely combinatorial objects and tend to design our graph algorithms to be combinatorial as well. In fact, over the years, "combinatorial" became a synonym of "fast" in the context of algorithms, in general.

Recent work, however, shows that one can obtain faster algorithms for a number of such "traditionally combinatorial" graph problems - most notably, the matching and flow problems - by approaching them with tools and notions borrowed from linear algebra and, more broadly, continuous optimization. This raises an intriguing question: is continuous optimization a more suitable and principled optics for fast graph algorithms than the classic combinatorial view?

In this talk, I will discuss this question as well as the developments that motivated it.

Richard Peng (Georgia Tech)

Algorithm Frameworks Based on Structure Preserving Sampling

Abstract: Sampling, or computing on a subset of the data, is a simple yet effective method for designing efficient algorithms. Routines that produce sparse instances from dense ones are used in many recent developments in graph theoretic and numerical algorithms.

While many data sets are inherently sparse, computation on them often involve dense intermediate objects such as matrix inverses and random walk transition matrices. This talk will focus on the design and application of methods that directly produce sparse approximations of these objects. These routines lead to provably efficient algorithms for matrix roots, sparse Cholesky factorizations, and multilevel/multigrid methods.
Padhraic Smyth (UC Irvine)

Statistical Latent Variable and Event Models for Network Data

Abstract: Social network analysis has a long and successful history in the social sciences, often with a focus on relatively small survey-based data sets. In the past decade, driven by the ease of automatically collecting large-scale network data sets, there has been significant interest in developing new statistical and machine learning techniques for network analysis. In this talk we will focus on two general modeling themes in this context: the use of latent variables for low-dimensional vector-based network representations models and event-based models for temporal network data. We will review the representational capabilities of these models from a generative perspective, discuss some of the challenges of parameter estimation that arise, and emphasize the role of predictive evaluation. The talk will conclude with a brief discussion of future directions in this general area.

Based on joint work with Zach Butler, Chris DuBois, Jimmy Foulds, and Carter Butts

Johan Ugander (Stanford)

Graph Clustering for Network Experiments

Abstract: Causal inference under interference is the challenge of administering field experiments in the presence of strong spillover interactions between treatment units. Recent work has demonstrated the efficacy of randomizing the experimental assignment of treatment units through the use of graph cluster randomization, an approach that employs graph clustering techniques to cluster the graph of spillover interactions and then randomize treatment assignments at the cluster level. This talk will discuss how advanced techniques for graph partitioning can be usefully applied to the randomization of experiments on networks. It will also give an overview of other graph-theoretic and algorithmic challenges that arise in causal inference on networks.