Introduction

Network science is an emerging field at the intersection of computer science, electrical engineering, statistics, and applied mathematics, as well as numerous application domains (e.g., sociology, neuroscience, proteomics, epidemiology, communication networks, etc.), which studies structure, formation, and function of complex networks.

My research primarily focuses on the theoretical foundations of network and data science, always with an eye toward applications: that is, my focus is on the formulation and fundamental limits of learning/statistical inference and data compression/transmission problems involving networks, as well as efficient algorithms to achieve those limits. I am also interested in the modeling and mathematical analysis of complex networks as random graphs, and in the application of information theoretic tools to the above problems. I am, furthermore, presently exploring machine learning applications in which graph features play an important role. Finally, I am interested in collaborating with researchers in other areas to translate my results to concrete applications.

In subsequent sections, I highlight a few representative lines of work, and then discuss future directions. I am (and have in the past been) involved in several other projects of an interdisciplinary nature (all are linked by the fact that they involve probabilistic/statistical/information theoretic modeling and analysis of objects and processes arising in, e.g., computer science [3] and biology [12]). These are documented in my CV.

Network archaeology: Inference problems in dynamic networks

Many networks in the real world are dynamic in nature: nodes and edges are added and removed as time progresses (e.g., as species evolve, new proteins with new interactions may be added to protein interaction networks; as a social network evolves, more people join and connect to new friends, etc.). I am generally interested in formalizing notions of learning the dynamic processes driving the evolution of time-varying graphs. Practical problems in this area present substantial challenges, as, for example, it is frequently difficult to procure many independent sample networks from a given source, and the edges of many networks cannot be known with complete certainty. Thus, novel theoretical approaches are needed to circumvent these challenges and provide rigorous statistical guarantees. Below, I describe some of my work along these lines. In addition to publications, I have contributed significantly to a grant proposal (currently under review) on these and other topics. See below for a section on potential funding sources.

Goodness of fit tests for dynamic network models

Statistical models of complex networks are ubiquitous in both theoretical and applied areas and are used to draw scientific conclusions, detect anomalies in technological networks, etc. Hence, it is natural (and important) to have a rigorous method to assess the extent to which a model fits observed data. To this end, in ongoing work, I have formulated the goodness of fit problem for dynamic network models and have explored efficiently computable, general test statistics which
can provably distinguish between sample network trajectories coming from any pair of sufficiently “well-separated” models coming from a large class.

More precisely, we formulate a discrete-time dynamic network model as a Markov process on graphs. In the simplest case, one is given a sequence of snapshots from a growing network and a candidate network model \(M_0\), and the task is to determine whether the snapshots came from \(M_0\) or from some model that is far from \(M_0\) in an appropriate metric. Note that the snapshots form a single sample trajectory from the Markov process, so that multiple independent, identically distributed samples from the process at a given time point are not available. However, many plausible network models have the important property that the conditional distribution of the graph at any given timestep changes slowly with time. We exploit this feature to devise a general test statistic that works well in practice and has a theoretically tractable analysis.

**Inferring network history from a snapshot, enabling temporal pattern discovery**

Another line of my work considers inferring properties of a network’s history (such as previous states, changes in its rate of growth, etc.) from a single snapshot at a given time. For instance, with coauthors, starting in [13], I have rigorously formulated and studied the feasibility of the statistical problem of labeling of nodes in a growing network with the order in which they arrived.

Given the ubiquity of available dynamic network data, several possible applications of node arrival order inference algorithms suggest themselves: e.g., in biology, the evolutionary history of proteins is of interest. Various methods based on physical considerations and genomic/phylogenetic data have been proposed for estimating ages on an evolutionary time scale (and, hence, approximate temporal ordering) of proteins [4]. Such information is of interest, because, for example, evolutionarily early proteins have been implicated in cancers and other diseases [22]. Leveraging protein interaction network data could provide additional validation, as well as greater specificity, of predictions using these other data sources.

More generally, the ability to infer the order of node arrivals in a network would enable the study questions about the history of that network’s formation: for instance, how stable is the set of most highly connected individuals in a social network over time? Moreover, it enables the detection of temporal patterns in data associated with nodes and edges. This, in turn, enables various statistical analyses, such as prediction of the values of these attributes for nodes and edges subsequently added to the network, changepoint/temporal anomaly detection, etc.

**Formulation and results:** The problem, in its most general form, is insoluble without some assumptions: we assume that the network in question is generated as a labeled graph \(G\) from some probabilistic model over graphs on \(n\) nodes (e.g., Barabási-Albert or duplication-divergence). The label of the \(j\)th vertex to arrive in the network is \(j\). The nodes are then randomly relabeled, producing a graph \(\pi(G)\) – that is, the only information about the arrival order is contained in the structure of \(\pi(G)\). The goal, then, is to recover the original arrival order \(\pi^{-1}\), given only knowledge of the relabeled graph \(\pi(G)\) (see Figure 1).

In [13], we made this formulation precise and gave general inapproximability results in terms of the symmetries of the graph model and of the sample graphs \(G\). We then applied these results (along with results from [14, 11]) to show inapproximability results for a few random graph models of interest.

In subsequent work [19], we formulated a relaxed version of the problem, in which it is only
Figure 1: Schematic of the node order inference problem. \( \mathcal{G}_n \) is a random graph model, \( G \) is a sample from it, \( \pi \) is a uniformly random permutation, and \( \sigma \) is an estimate of \( \pi^{-1} \) from \( \pi(G) \).

In ongoing work, I am pursuing characterizing the complexity of tight approximations for an optimal solution to this problem for the Barabási-Albert model. This has necessitated solutions to challenging combinatorial, probabilistic, and optimization problems (e.g., the counting of linear extensions of a time-directed directed version of a typical sample graph, viewed as a poset).

I have also recently explored the use of spreading process data to determine likely network histories from a snapshot at a particular time [18].

**Application to real networks:** In ongoing work (partly contained in [19]), we are applying our algorithms and analysis to biological networks to attempt to derive new insights. For instance, using fMRI data, neuroscientists have constructed brain region co-excitation networks, in which nodes represent brain regions (e.g., the corpus callosum, etc.), and two nodes are connected whenever their patterns of excitation are significantly correlated. We hypothesize that information about the order in which different brain structures (hence, functions) evolved can be gleaned from the structure of this network. We applied our methods to such a network derived from human brains (see Figure 3), which yielded certain biologically plausible results (though a full ground truth is unavailable): e.g., the corpus callosum and cerebellum are judged to be the oldest, while areas of the cerebral cortex are judged to be the youngest. Further work, with significant involvement of
domain experts, is needed for rigorous scientific validation of the generated ordering.

Information content and compression of network structures

The prevalence of network and tree-structured data in recent years has motivated information theorists and computer scientists to explore problems of efficient description (i.e., compression) of labeled and unlabeled graphs (i.e., graph structures) [24, 5, 1, 2]. This is in contrast to the focus in previous decades on the theory of source coding for sequential data [6] (the major difference being that sequences representing network and tree-structured data have significantly more long-range correlation than is typically captured by models of sequences that are classically considered; practically speaking, ignoring these long-range correlations yields suboptimal compression schemes). This requires novel analysis to determine the fundamental limits of compression of structured data, as well as efficient algorithms to achieve these limits.

Formulation and results: The general, most basic problem of structural compression can be formulated as follows: design an efficiently computable source code $\phi$ which maps graphs to bit strings, such that, for a graph $G$, an arbitrary graph isomorphic to $G$ can be recovered efficiently from $\phi(G)$. Moreover, we require that, given an arbitrarily relabeled sample graph $G$ from a particular random graph model on $n$ vertices, the length of the bit string is as small as possible in expected value (see Figure 4). In general, the minimum achievable expected length is given by the Shannon entropy of the distribution on graph structures, which is often extremely nontrivial to compute.

![Figure 4: Schematic of the structural compression problem. Note that the output graph is only isomorphic to the input.](image)
As a concrete example of my work along these lines, in recent papers with coauthors [11, 10], I have studied the problem of asymptotically optimal (and algorithmically efficient) compression of structures from the Barabási-Albert model. The analysis of our compression algorithm (and, in particular, the proof of its optimality) required new structural results for the model, including a full characterization of the typical size of the automorphism group (see the two papers above, as well as [14]; see also [23] for work on a different model).

Biological models and applications

I have collaborated with biologists and bioinformaticians on several projects involving biological applications, including modeling and analysis of nanopore sequencing technologies [12], the statistics of the protein folding process [15, 21], and biological networks [17]. Here, I explain one application-oriented work along these lines.

Statistical and functional significance of pairwise interactions between dense subgraphs: My paper [17] considers the problem of detecting statistically significant overlaps between pairs of dense subgraphs in a graph. Given that dense subgraphs in biological networks have been shown to frequently be of functional significance, we naturally hypothesized that higher-order clustering (in particular, overlaps between dense subgraphs) carries functional information as well. To investigate this hypothesis, we formulated a test for significance of the overlap between two subgraphs $X$ and $Y$ that takes into account the number of nodes and density of $X$, $Y$, and $X \cap Y$. We applied our method to a protein interaction network with pathways (clusters of proteins involved in some function) annotated and found it to perform well in comparison with ground truth pathway pairs derived from correlated transcriptional data between pairs of pathways. Moreover, upon constructing a pathway interaction graph (where nodes represent pathways, and there is an edge between two pathways precisely when our method determines that their overlap in the input graph is significant) we found that clusters of pathways were of functional significance (see Figure 5).

Fundamental computer science applications

One major theme of my work has been the study of the typical behavior/performance of processes and structures arising in computer science applications. That is, it is often useful to understand the performance of a data structure on inputs generated by reasonable probabilistic models, in addition to characterizing its worst-case performance (which in many cases is either overly pessimistic or is uninformative for the purposes of comparison of different data structures). For brevity, I only mention that I have worked on analyses of serializations of concurrent processes with conflicts [3] and on performance parameters of various digital trees (i.e., string data structures which are ubiquitous in applications, from data compression to computational biology to Internet routing) via complex analytic tools [16, 7, 8, 9, 20].

Potential funding sources

As mentioned in the introduction, I have contributed significantly to a proposal currently under review by the Army Research Office. Additionally, research in my area is relevant to the missions of various federal funding agencies, including the NSF, NSA, NIH, DHS, DARPA, etc., as well as industry sources, such as the Amazon and Google Faculty Research Awards. In more detail, the NSF has several programs in its CCF, DMS, and CNS divisions to which my research on statistical
Figure 5: Pathway interaction graph with clusters annotated with functional significance (from [17]).

Aspects of network science are relevant: e.g., Communications and Information Foundations and Algorithmic Foundations. Statistical problems on networks are of interest to the NSA, DHS, and other federal agencies involved in national security/defense-related missions, since these frequently arise in cybersecurity and social network analysis applications (e.g., network anomaly detection, uncovering hierarchies and key nodes in enemy social networks, etc.). The NIH has posted numerous Funding Opportunity Announcements involving dynamic networks of various sorts arising in the biological sciences (e.g., organelle communication networks, protein interaction networks, chromosomal contact networks, etc.). Posted announcements explicitly mention various tasks of interest, including data compression, control, quantification of information transmission, and inference problems relating directly to biological/medical impact.

Regarding industry faculty research awards, Google and Amazon both mention graph mining and analysis of structured data as focus areas, into which my research on network science may fit. Additionally, my current projects involve incorporating prior structural knowledge and constraints into deep learning models, which is a focus of Amazon’s research award program.

Future directions

Broadly speaking, my intent is to continue to pursue statistical inference/machine learning questions relating to graph analysis with potential for impact in practical scientific and engineering applications, and, furthermore, to work with other researchers to develop those applications. More specifically, in the near terms, I plan to continue to explore three main thrusts:

- **Statistical tools with theoretical guarantees for dynamic networks:** There remain substantial challenges in the analysis of these networks. E.g., the design of flexible models for quantifying the relative roles of different mechanisms driving the evolution of networks, infer-
ence under uncertainty and incomplete history (e.g., taking into account realistic constraints on the sampling processes used in obtaining graph structure data), use of time-varying graph structure to make inferences and predictions on graph-correlated data, prediction of the formation of complex structures and transitions in global graph properties, etc., are all fertile ground for further exploration and will pose challenges in terms of both computational and statistical efficiency (since, frequently, many sample graphs from a given source are expensive and difficult to procure). The emphasis would be on the development of novel algorithmic methodology for these problems, as well as characterization of optimality.

- **Control of processes on networks:** Here, one would like to efficiently control the state of, say, an epidemic process on a network (e.g., to maintain an infection on only a particular subset of nodes) by adding/removing edges (each time incurring some cost). Such problems become especially challenging when the dynamics of the epidemic process must be learned from observed infections.

- **Machine learning applications/connections:** Concretely, working with colleagues at the University of Michigan with expertise in biochemistry, I am in the early stages of a project that aims to automate the discovery of novel compounds for the disruption of bacterial biofilms, using a deep learning approach. This is essentially a problem in training generative models for compounds, which are representable as multigraphs with node labels. Hence, appropriate representation of the relevant features of these graphs is of immediate practical relevance and is likely to inspire theoretical developments, since there are many ad hoc, empirical methods with widely varying performance which may be improved upon with more principled approaches. Improvements in the rate and performance of automated compound design methods are likely to be impactful, as the process of compound discovery for such applications currently takes on the order of several years.

Along other lines, given the success of neural network approaches in other domains, it would be of interest to devise neural architectures for the solution of statistical problems on dynamic networks (e.g., learning generative models of growing networks) and to compare their performance with theoretically principled methods.

More generally, I intend to seek out collaboration on interdisciplinary projects (e.g., related to applications in computer science/engineering, biology, etc.) which could benefit from probabilistic, statistical, information theoretic, and algorithmic tools (see, e.g., [12, 15, 17, 3] for some of my prior work along these lines).

**References**


