Research Statement

Satyaki Sikdar
Department of Computer Science & Engineering
University of Notre Dame
ssikdar@nd.edu

Overview

Research Objectives. My principal research interest is to develop, study, and evaluate fundamentally new techniques for the discovery of interesting structural patterns and their function within complex real-world networks, and more generally in data mining, formal language theory, and network science. Due to the size of Web-scale social and information networks, most of my work falls within *big data* paradigm, which requires an expanded use of new technology. My long-term research goals include: (1) understanding the underlying mechanism of network formation; (2) mining interesting and useful patterns in such networks; (3) predicting the evolution patterns for dynamic interaction networks; and (4) studying real-world applications and helping decision-makers design better complex networks to enhance their usability.

Uncovering the Hidden Building Blocks of Complex Networks. Teasing out interesting relationships buried within volumes of data is one of the most fundamental challenges in data science research. Increasingly, researchers and practitioners are interested in understanding how individual pieces of information are organized and interact in order to discover the fundamental principles that underlie complex physical or social phenomena. Indeed the most pivotal moments in the development of a scientific field are centered on discoveries about the structure of some phenomena. For example, chemists have found that many chemical interactions are the result of the underlying structural properties of interactions between elements. Biologists have agreed that tree structures are useful when organizing the evolutionary history of life, sociologists find that triadic closure underlies community development, and neuroscientists have found *small world* dynamics within neurons in the brain.

Graphs offer a natural formalism to capture this idea, with nodes representing the individual entities and edges describing the relationships in the system. Arguably, the most prescient task in the study of such systems is the identification, extraction, and representation of the small substructures that, in aggregate, describe the underlying phenomenon encoded by the graph. Furthermore, this opens the door for researchers to design tools for tasks like anomaly/fraud detection, and the anonymization of social network data.

My Academic Preparation. In order to achieve the long-term goals of complex network research, contributions from a broad spectrum of disciplines are needed, such as data mining, statistics, machine learning, network science, social science, economics, and so on. In my opinion, the general steps to approach the class of problems related to network science research consists of: (1) identifying the individual entities (nodes) and discovering the connections between these objects (edges); (2) adapting theories from other domains to understand the networks; (3) building rigorous probabilistic or machine learning models to address mining problems; (4) proposing efficient algorithms to solve the proposed models; and (5) using large-scale real-world applications to test the theories, models, and algorithms. I also make a serious effort to study the latest discoveries related to network science in other disciplines, such as physics, medicine, social sciences, and economics. Moreover, I pay attention to the potential real-world applications of social and information networks, especially along with the widespread availability of large scale, information-rich network data, such as bibliographic repositories, social media websites, government data, information sharing websites, web-based crowd-sourcing, and so on.

Summary of My Achievements. I have been systematically studying a critically important but new field of network science. My dissertation research makes significant progress in filling the void in current research to find scalable and interpretable graph modeling methods by leveraging the new-found link between formal language theory, graph theory, and data mining. Furthermore, my research helps bridge the gap between subgraph mining and graph generation to create a new suite of models and tools that can not only create informative models of real-world data, but also generate, extrapolate, and infer new graphs in a precise, principled way. I have proposed new theories, methodologies, and algorithms for solving interesting problems in complex networks, which can be applied to a wide range of real-world applications. Along this line, I have published research papers in major conferences and journals; which have laid out the foundations for future studies in the field of graph mining and generative models.

Recent Research

My major works are summarized below, many of which have been published in major conferences and journals. Some of my papers have been taught as a part of curriculum in university computer science courses.

Community Detection. My first foray into academic research came during the initial boom of online social networks and network science. I was especially interested in uncovering the mesoscale and large-scale structure of complex networks. I devised efficient and novel algorithms which could uncover groups of users who were strongly connected with other users in the group, and sparsely connected with the rest of the network. Identifying such groups in large complex networks is a challenging task and we found ways to leverage findings from theoretical computer science and graph theory to achieve it. We found that not all links in the graph are created equal, heavier connections exist between nodes in a community, and communities are connected via lighter connections. We exploited this finding from different angles, one where we identified the skeletal connections of the network by using some clever data structures [1]. We also found a way to use a mixture of breadth-first and depth-first searches to find identify densely interconnected regions around seed nodes [2].

Chapters on Spectral Graph Clustering and the NetworkX Graph Library. I contributed a chapter on Spectral Clustering to an NSF Compendium Report on Graph Kernels edited by Dr. Peter Kogge [3]. Graph theory and linear algebra are interrelated in a myriad of ways. Spectral clustering takes advantage of some neat linear algebraic properties concerning eigenvectors of matrices to find clusters. In the report, I give a brief introduction to the topic, lay down the foundations, and discuss various spectral clustering algorithms and study their computational powers and numerical stability.

NetworkX is one of the most popular software libraries for graph processing. I was asked by Dr. Kogge to contribute a chapter to an NSF Compendium Report on Graph Processing Paradigms [4]. I wrote a gentle introduction to the library with some toy examples to portray the simple yet powerful API.

Graph Grammars. Graph grammars provide a simple yet interpretable framework via graphical rewriting rules that can match and replace graph fragments similar to how a context-free string grammar rewrites characters in a string. These graph fragments represent a concise description of the network building blocks and the instructions about how the graph is pieced together. A potentially significant benefit from this model stems from its ability to directly encode local substructures and patterns in the grammar rules. Forward applications of graph grammars may allow scientists to identify previously unknown patterns in graph datasets representing critical natural or physical phenomena. My dissertation research is heavily invested in graph grammars.

First, I presented a method to extract synchronous hyperedge replacement grammar rules from a temporal graph that clearly and succinctly represents the graph dynamics found in the graph process [5]. As a result, I expect that graph grammar rules may be used to discover previously unknown graph dynamics from sizeable real-world networks. Then, I proposed a different formalism called vertex replacement grammars, which are more expressive, computationally faster, and topologically more faithful than hyperedge replacement grammars [6]. Finally, I also helped develop a model called BUGGE: the Bottom-Up Graph Grammar Extractor, which extracts grammar rules that represent interpretable substructures from large graph data sets [7]. Using synthetic data sets I explored the expressivity of these grammars and showed that they clearly articulated the specific dynamics that generated the synthetic data.

Stress Testing Graph Models. Graph generative models, like other machine learning models, have implicit and explicit biases built-in, which often impact performance in nontrivial ways. Yet, in many systems, errors encoded in loss functions are subtle and not well understood. With this in mind, in this work, I introduced the Infinity Mirror test for analyzing the robustness of graph models [8]. This straightforward stress test works by repeatedly fitting a model to its outputs. Through an analysis of thousands of experiments on synthetic and real-world graphs, we show that several conventional graph models degenerate in exciting and informative ways. We believe that the observed degenerative patterns are clues to the future development of better graph models.

Link Prediction. In this project, I generalize the idea of triadic closure—that a friend of a friend is likely to be a friend—one of the key underlying principles of network formation and growth, under an intuitive umbrella generalization: the Subgraph-to-Subgraph Transition (SST) [9]. I presented algorithms and code to model graph evolution in terms of collections of these SSTs. The SST framework can then be used to create link prediction models for both static and temporal, directed and undirected graphs which produce highly interpretable results that simultaneously match the state of the art graph neural network performance.

Future Directions

In the future, I will focus on the development of the field of network science in four dimensions: (1) the development of graph models for attribute-rich networks, especially social networks; (2) the study the underlying mechanisms of heterogeneous knowledge networks; (3) a deeper understanding how complex networks grow and evolve, and (4) the adaptation of these technologies to real-world interdisciplinary applications.

Modeling Attributed Networks. In this direction, I will build better, more interpretable, and scalable generative models for heterogeneous networks where nodes and edges hold attributes. Are there common characteristics across heterogeneous networks from different domains? Will some of the well-known characteristics in homogeneous networks still hold in heterogeneous ones in a variant form? To answer these questions, I am actively developing new methods inspired by my research on graph grammars but also incorporating recent discoveries in reinforcement learning.

Understanding the Mechanisms of Heterogeneous Knowledge Networks. The explosion of digital information offers an unprecedented opportunity to study the dynamics that shape human understanding, investigation, and certainty. By applying the ideas and frameworks from graph grammars to information networks, I will be able to better understand how humans create and organize the artifacts of knowledge. I am particularly interested in the recent discoveries concerning the *Science of Science* project. I am hopeful that the building blocks of these networks will provide unique insights on successful research collaborations and university faculty hiring patterns.

Growth and Dynamics of Complex Networks. Most real-world networks are temporal in nature, i.e., the topology of the graph changes over time via node and edge additions and deletions. The idea of production rules from graph grammars provides a natural analog to this growing process. I will therefore develop more precise, scalable, and interpretable models for modeling temporal networks, which are able to predict the future topology of a graph.

Visions on Real-World Interdisciplinary Applications. Network Science is an inherently interdisciplinary field related to data mining, machine learning, statistics, social science, economics, etc., and has potential applications in a large span of different domains. For example, transformations and interactions between chemical molecules have innumerable impacts on industry and human health. The bonds of molecules are themselves a graph, and the transformations between molecules performed naturally and artificially are large and complex graphs. Theoretical features of these graphs will yield new information about the way scientists explore chemical diversity, and the underlying design principles of chemical networks, and may support developments in the discovery of therapeutic molecules.

I am looking forward to collaboration opportunities with researchers from all these disciplines and seeking funding opportunities from multiple funding agencies and industries.

References

[1] Partha Basuchowdhuri, Satyaki Sikdar, Sonu Shreshtha, and Subhashis Majumder. Detecting community structures in social networks by graph sparsification. In *Proceedings of the 3rd IKDD Conference on Data Science, 2016*, pages 1–9, 2016.

[2] Partha Basuchowdhuri, Satyaki Sikdar, Varsha Nagarajan, Khusbu Mishra, Surabhi Gupta, and Subhashis Majumder. Fast detection of community structures using graph traversal in social networks. *Knowledge and Information Systems*, 59(1):1–31, 2019.

[3] Neil Butcher, Trenton Ford, Mark Horeni, Kremer-Herman Nathaniel, Steven Kreig, Brian Page, Tim Shaffer, Satyaki Sikdar, Famim Talukder, and Tong Zhao. Spectral community detection. In Peter M. Kogge, editor, *A Survey of Graph Kernels*, pages 67–75. University of Notre Dame, 2019. URL `https://dx.doi.org/doi:10.7274/r0-e7wb-da60`.

[4] Neil Butcher, Trenton Ford, Mark Horeni, Kremer-Herman Nathaniel, Steven Kreig, Brian Page, Tim Shaffer, Satyaki Sikdar, Famim Talukder, and Tong Zhao. Networkx graph library. In Peter M. Kogge, editor, *A Survey of Graph Processing Paradigms*, pages 67–70. University of Notre Dame, 2019. URL `https://dx.doi.org/doi:10.7274/r0-z6dc-9c71`.

[5] Corey Pennycuff, Satyaki Sikdar, Catalina Vajiac, David Chiang, and Tim Weninger. Synchronous hyperedge replacement graph grammars. In *International Conference on Graph Transformation*, pages 20–36. Springer, 2018.

[6] Satyaki Sikdar, Justus Hibshman, and Tim Weninger. Modeling graphs with vertex replacement grammars. In *ICDM*. IEEE, 2019.

[7] Justus Hibshman, Satyaki Sikdar, and Tim Weninger. Towards interpretable graph modeling with vertex replacement grammars. In *BigData*. IEEE, 2019.

[8] Satyaki Sikdar, Daniel Gonzalez, Trenton Ford, and Tim Weninger. The infinity mirror test for graph models. *arXiv preprint arXiv:2009.08925*, 2020.

[9] Justus Hibshman, Daniel Gonzalez, Satyaki Sikdar, and Tim Weninger. Joint subgraph-to-subgraph transitions–generalizing triadic closure for powerful and interpretable graph modeling. *arXiv preprint arXiv:2009.06770*, 2020.