

RESEARCH statement

Jure LESKOVEC

RESEARCH OBJECTIVES

My principal research interest is in applied machine learning and large-scale data mining, focusing on the analysis and modeling of large real-world networks as the study of phenomena across the social, technological, and natural worlds. Such graphs frame numerous research problems and high-impact applications. For example, social networks on the internet generate revenue of multiple billions of dollars; detection of virus outbreaks can save lives; regulatory gene networks help us understand how our cells work; anomaly detection in large computer-traffic networks is vital for corporate and national security.

My long term research goal is to harness large-scale networks to understand, predict, and ultimately, enhance social and technological systems. I would like to create explanatory and predictive models of actions of large groups of people and societies, and biological and technological systems. Although the actions of a particular individual or component may be too difficult to model, machine learning and statistics can be applied to large groups or ensembles, which can yield effective models with ability to predict the flow of future events. Based on my recent results and research experience, I believe that the study of large networks is the promising approach to developing such understandings, as graphs capture local dependencies, and also reveal large-scale structure and phenomena arising from the multitude of local interactions. Local seemingly “random” behavior can propagate to the macro scale where global regularities and patterns emerge, *e.g.*, power-law degree distributions and small-diameters.

On the way to achieving this long-term goal, my research consists of (1) analyzing theoretical models of network structure and evolution; (2) developing statistical machine learning models and algorithms to efficiently estimate the model parameters from data; (3) working with massive datasets of gigabyte and terabyte scale, as certain behaviors and patterns are observable only when the amount of data is large enough.

CURRENT ACHIEVEMENTS

Through my work, I have addressed a number of important questions regarding the properties and patterns of large evolving networks by revealing how local behavior and structure leads to large scale phenomena and useful applications. What does a “normal” network look like? How will it evolve over time? Is the network or a community “healthy”? How do information and viruses spread over the network? How can we identify and find influential nodes or select nodes to immunize in networks? Answers to such questions are vital to a range of application areas from the identification of illegal money-laundering rings, misconfigured routers on the Internet, viral marketing, and protein-protein interactions to disease outbreak detection.

Results of my doctoral research have been included in the curricula of several graduate classes on network analysis, advanced data mining, internet algorithms, social media and social networks across universities. For example, William Cohen, Kathleen Carley, Stephen Fienberg at CMU, Lada Adamic at University of Michigan, Jon Kleinberg at Cornell University, Nina Mishra at University of Virginia, Jiawei Han at UIUC, Constantine Dovrolis at Georgia Tech and others make our results part of their courses.

In my dissertation research, I focused on static and evolving networks, and the dynamics of processes, like virus propagation, that take place in networks. The table below gives the overall structure of my thesis research with the mapping to the sections of this document.

	Analysis	Models	Algorithms
Static networks	6	2	2 8
Dynamics of network evolution	1	1	7
Dynamics of processes on networks	4	3 8	5

1 Network evolution Our recent work had influence on thinking about fundamental structural properties of networks varying over time. For example, to date, it was commonly believed that the average degree of graphs of natural phenomena remains constant as they grow over time. Moreover, it was also assumed that the distances in networks slowly (logarithmically) increase with the network size. We showed that in fact networks *densify over time* as the number of edges $e(t)$ at time t is increasing as $e(t) = c n(t)^a$ with the number of nodes $n(t)$. The densification exponent a is non-trivial, $a \approx 1.2-1.6$ [15]. Even more surprisingly, the diameter of the network *shrinks* as it grows. These findings are fundamentally different from what was believed and commonly assumed in the past. A natural question to ask then is why do we observe these regularities? What is the connection between densification and shrinking diameters? As the existing intuitions and models do not explain these types of behavior, we developed a “Forest Fire” generative model that creates graphs with these properties [1]. We also showed that densification itself is not enough to observe shrinking diameters. The work received the best research paper award at ACM KDD 2005 [15].

2 Kronecker graphs A different question is how one can generate realistic looking synthetic graphs. This competency is important as we often need good null-models for simulations, what-if scenarios and hypothesis testing. We developed a Kronecker graphs model that is based on the tensor product of graph adjacency matrices. In contrast to previous models, Kronecker graphs capture greatest number of static and dynamic network properties [16], while being mathematically tractable. Moreover, we developed a maximum likelihood approach for parameter estimation of Kronecker graphs [6] Naïve approaches take super-exponential time, while we developed a linear time parameter estimation algorithm. Using approximation and sampling we efficiently search the space of $10^{1,000,000}$ states, and estimate the model parameters for networks with millions of nodes in a matter of hours. Kronecker graphs have been harnessed by the high performance computing community, e.g., by Jeremy Kepner at MIT Lincoln Lab, and David Bader at Georgia Tech.

3 Diffusion and cascading behavior To model the evolution of large networked systems one also needs to understand how influence and information spread and propagate. Developing insights into such propagations are important for selecting targets for advertising and marketing, finding opinion makers with great influence in shaping people's opinions, and to select nodes to monitor to best detect the potential epidemics. In our work on information propagation between blogs [8] and on product recommendation networks [2, 11], we developed macroscopic models of the spread of influence in networks [8, 22], and found common and abnormal network substructures, called *cascades*, that the propagation process creates [8, 13].

4 Human adoption curve To the best of our knowledge, my research was the first to answer a simple question: What is the probability of a person adopting the behavior (e.g., buying a product) as more friends have adopted [11]. Two competing theories are diminishing returns, which assumes that the probability of adoption slowly increases, and a critical threshold hypothesis, which assumes that the probability of purchase suddenly jumps as a particular number of friends acquire the product. The validation of these competing models is only made possible with sufficient data. We observed 16 million product recommendations between 4 million people on half million products from a large online retailer. We found that probability of adoption follows a diminishing returns property, and that the probability of adoption saturates (and sometimes even starts to decrease) after around 20 network neighbors adopt [2]. These findings are important for advertising and viral marketing. Follow up works by Duncan Watts, Jon Kleinberg, Daniel Huttenlocher and others later confirmed the same behavior in a number of other domains, e.g., the probability of joining a community, sending an email, or editing an article on Wikipedia.

5 Cascade detection The diminishing returns property has also led us to efficient and theoretically sound algorithms for network sensor placement [5]. Submodularity is the diminishing returns property that we exploited to develop new tighter bounds for greedy optimization of submodular functions and to devise new efficient optimization algorithms. Our approach [22] ranked first in the “Battle of the Water Sensor Networks” competition where the task was to place sensors in a city water distribution network to effectively detect contaminants spreading over the network [23]. Beyond the task at hand, we showed that the same sensor placement algorithm can be used to decide the best news sites on the internet to read to not miss important information, i.e., to detect “information epidemics” effectively.

We tracked the information propagation on the blogosphere for 1 year, and used our algorithms to find the most informative blogs. Our work received the best student paper award at ACM KDD 2007 [5], and generated a burst of interest on the blogosphere as we experienced a “Slashdot effect” with more than 20,000 visits to our project website <http://www.blogcascades.org>.

6 Communities Researchers in the social sciences and physics have long been excited about the existence of “network communities”, where the intuition is that networks contain sets of nodes that interact more strongly with each other than with the remainder of the network. Recently, we found behaviors that are fundamentally different from intuitions based on small social networks, spatial graphs or hierarchical community structure that has typically been assumed for social and biological networks. Our observation is that, in large networks, tight communities exist only at smaller scales, but as the community grows it vanishes and blends with the rest of the network [19]. Formalization and models of such behavior would have a wide range of implications for researchers in the social sciences who want to discover communities from network data, and also for graph clustering and partitioning research [24].

7 Large-scale data All my research integrates and is part of a large graph mining code base that I developed during the course of my research. My C++ software library is freely available and scales to massive graphs. For example, I worked on the “planetary scale” Microsoft Instant Messenger network, the largest social network analyzed to date [20]. I collected and analyzed 4.5 Terabytes of network data. The MSN network contains 240 million people, with more than 1 billion conversations per day. I used the graph for generating insights about numerous communication phenomena, including the small world hypothesis, homophily, and patterns of intra- and international conversation.

8 Other achievements I have also worked on numerous auxiliary topics related to networks, like graph sampling [12], web search quality prediction [7], sensor placement [22], virus propagation and epidemic thresholds [3, 9], document summarization using graphical document representations [17], text classification [18] and topic intensity tracking [14]. All of these projects tackled large-scale real-world problems. For example, our approach for text classification on graphs won the KDD Cup 2003 [4].

VISION FOR THE FUTURE

The long-term goal of my research is to build and harness models of natural and synthetic systems to make predictions about future events. I believe that it will be feasible in the long-term to predict events and overall dynamics of the behavior of networked systems, such as large groups of people, web, communication, and biological networks. The idea is that actions of an individual are too hard to model but machine learning can be applied to large groups to predict the general flow of future events. I believe the right approach is through networks where local dependencies and behavior propagate to global patterns and trends. The key is to connect local to global, complement the topology information with other types of data, and choose the right scale where micro propagates to macro.

In my thesis research, I made several steps towards this long term goal. We now better understand microscopic and macroscopic network evolution and models that connect the two. Moreover, we can efficiently fit the network models to the data and predict prior and future states of the network. We also have a better understanding of how information and influence propagate over the network, what are the traces of propagation, and how to use these for selecting influential nodes or detect disease outbreaks.

On the road towards the long-term goal my research will focus across three dimensions: structure and dynamics of networks, designing networked systems and influencing their evolution; encompassing richer types of networked data; and scaling up the analyses to internet-scale computing.

Network structure and communities The online world is a rich testbed for my research as web media and social networking sites contain very detailed traces of human social activity, people’s profiles, interests, groups, etc. I want to understand how network topology, user profiles, and past actions determine the future of particular groups or events, and, more abstractly, how links arise and decay. I want to define and explore the notion of the “health” of a social network or a community. In this context, I am collaborating with LinkedIn and Facebook; we plan to perform large-scale data mining and machine learning for social network analysis.

Influencing evolution Beyond simply observing and characterizing a network, we can try to influence the activities and overall evolution of a community. For example, we can explore mechanisms that would help a community to evolve in a healthy manner and continue to grow organically. The key here is to make a step from passive observation to actively trying to steer and influence the development of the network or community. Explorations in this realm could lead to strategies that help networked communities to survive and serve as richer resources for people.

Information propagation Another important aspect is online media and information propagation: How do people (sites) consume and alter information, and how do they influence propagation? For example, information from New York Times probably spreads in a different way than Slashdot posts. Moreover, when a story is posted on Slashdot, it is given a special boost. A certain community starts discussing it, which further diffuses the story. I wish to analyze and build predictive models of such behavior. Natural case studies here are predicting opinion formation and its outcomes, finding trendsetters, predicting election results, the success of a new product, and the evolution of content on Wikipedia.

Richer types of graphs Most algorithms and models today work on simple undirected graphs. I plan to extend generative models and mining algorithms to graphs with multiple edges between pairs of nodes and, to graphs with different types of nodes and weights or attributes on nodes and edges. Incorporating other data modalities like textual information, and historical and communication data, will allow for richer and more accurate models.

Scalability One of my primary research agendas focuses on large scale data and computing architectures for massive data manipulation and analysis. I plan to explore map–reduce type abstractions for large scale computing and extend my software library to distributed “share nothing” architectures. The question here is what kinds of analyses are suitable for such architectures, and how to parallelize data mining and machine learning algorithms to scale to thousands of machines. Here, I am collaborating with the CMU Parallel Data Lab and Yahoo Research, who recently gave us access to a 5000 node Hadoop cluster. This line of research will allow me to perform near real-time analysis of planetary and internet scale data and find patterns that are practically unobservable at smaller scales.

__ These steps capture my sense for the beginning of an evolving research framework that will allow me to tackle these challenging problems in a unique way. My research on networks is theoretically grounded and spans several areas of computer science as diverse as machine learning, theory and systems. Implications of my research have direct applications well beyond computer science – to social sciences, physics, economics and marketing. I am excited about the influence that my graduate research has already had within industry and academia, and look forward to continuing to make strides on both theoretical foundations and real-world applications.