My main research goal is to understand how social networks form and evolve over time. Network (a.k.a. relational, graph) data have become ubiquitous and accessible, in domains such as online social networks, citation networks, and political campaign contributions. Understanding how these networks form and evolve is a critical data mining problem, with applications in fields such as sociology, marketing, security, and human-computer interaction. My interests focus on three interrelated topics: the global topology of networks, diffusion within a network, and applications of network data.

The first goal is to observe patterns in the topological structure of these graphs. How do new nodes and links form in a network? Are these patterns common to all networks, or only those in certain domains? These observations provide intuition about the mechanisms driving network evolution, allow us to forecast future behavior, and help us spot anomalies.

In addition to knowing how networks form, it is important to understand the dynamics of diffusion inside a network. How do rumors or viruses spread through a network of people or computers? Do certain structural conditions allow for different patterns of propagation? One way to address these questions is to study subgraphs formed by citations, or cascades, and the typical patterns they display.

Finally, I plan to use knowledge of networks and their behavior to build useful applications. For instance, can we use ideas about propagation to spot certain types of communities (blog communities, auction fraudsters, etc.) or forecast product adoption? Can we detect fraud in a cellular phone network by identifying deviations from typical call patterns? In online social networks, do certain structural patterns allow for a sub-group to thrive or cause it to dissolve? If we know what typical diffusion patterns are, how can we optimize marketing plans to target certain users and take advantage of network effects? These are only a few of the many areas that may be aided by a better understanding of networks.

Completed work

Topic 1: Topology

In published work (KDD08 and ICDM08), my collaborators and I have examined many different types of large (millions of nodes) weighted graphs, such as social networks of blogs, political campaign contributions, and patent citations. We make several discoveries across these networks, three of which we briefly cover here: a) the superlinear behavior of edge...
Figure 1: **Superlinearity**: There is a power-law relationship between the number of incoming edges a node has and the total incoming edge weight. In particular, in political campaign donation data, if a candidate has $d$ donors, the candidate will receive superlinearly more money—specifically, $\propto d^{1.17}$ dollars in donations. Each point is a single federal candidate (for example, the outlier noted is John Kerry’s presidential campaign). The inset shows the power law exponent over time.

weights, b) the constant/oscillating behavior of components, and c) the design of a generative model.

**Edge weights**: One surprising finding is the superlinear relationship between the number of unique edges in a graph and the total edge weight (such as packet sizes in network transfers or dollar amount in campaign donations). We observe this superlinear relationship on a local scale as well as on a graph at large—for example, candidates with more donors receive superlinearly more money in total. This finding is illustrated in Figure 1.

**Component sizes**: Our second topological observation is that once the giant connected component forms (as is known to occur in social networks), the sizes of the next largest connected components (NLCC’s—the second- and third-largest components) stabilize. Surprisingly, there appears to be a certain threshold that a component will reach before joining to the largest component—while this threshold varies between networks it appears to remain near-constant over time within each network.

**Generative models**: Being able to model these behaviors is important for understanding the mechanisms causing them and allows us to make predictions. As our third topological finding, we have developed two complementary models that match several patterns observed in real evolving networks. The first is an agent-based model which we call the Butterfly generator, the first model to reproduce the NLCC property in addition to other known patterns—most agent-based models will produce a single component. The second model is the Recursive Tensor Model, which uses tensor multiplication and self-similarity to create a weighted network in time that provably follows the observed power-laws as well as observed bursty behavior. Each of these two models is useful—the agent-based Butterfly model allows us to understand the local mechanisms forming networks and forecast “what if” scenarios, while the tensor-based RTM can be simulated in parallel and is useful for theoretical analysis.
**Topic 2: Diffusion**

In addition to understanding global properties, we have discovered mechanisms that lead to local diffusion. To this end, we have performed a large case study in blog citations. We completed extensive analysis of a set of 2 million blog posts and identified common patterns of information propagation by analyzing cascades, or conversation trees. We were the first to examine the shapes of conversations in blogs, and we have three major discoveries in diffusion that we present: a) the power-law decay of in-links, b) power laws in cascade sizes, and c) that different communities have different cascade patterns. Details of these findings, as well as accompanying generative models, are discussed in SDM07, ICWSM07, and ICWSM09.

*Post popularity decay:* First, we found that in-links to a particular post over time drop off with a power law of exponent $-1.5$. While one would certainly expect a blog post’s timeliness to be important, this finding is surprising as one would expect popularity to decay linearly or exponentially.

*Cascade sizes:* Our second finding is on cascade sizes. We found that sizes of cascades overall follow a Zipf distribution (power law with exponent -2)– most conversations consisting of a single post with a few being larger. Additionally, power laws apply to the sizes of particular cascade shapes. Sizes of “stars”, or conversations involving many links to a single post, have a power-law dropoff exponent of -3.1, and sizes of “chains” (a series of posts, each linking to the previous) an exponent of -8. Again, timeliness is likely a key factor in forming these types of cascades.

*Cascades and communities:* We examined these cascade shapes more closely and, using dimensionality reduction techniques, we showed that some types of blogs (such as “politically conservative” blogs) tend to form longer, more drawn-out conversations, while other types (such as “humorous” blogs) have shallower star-like bursts of activity, with one post being the center of each cascade. Thus, we find that analyzing the structure of blog networks can aid in community detection.

**Topic 3: Applications**

Applications for these findings abound, and will be a key direction of future research. Already I have applied concepts in link analysis to the domain of accounting. In work published at KDD09 with collaborators in industry, I used link analysis for detecting misstatements (either intentional fraud or unintentional errors) in general ledger data, a problem that causes billions of dollars in losses annually. Our goal was to assess risk of an account based on the series of transactions made, and how closely an entity is associated with other “risky” entities. The scalable method we developed is able to identify misstated accounts automatically, allowing auditors to complete their work more quickly and efficiently. We showed this produced a lift of up to 6.5 over random, and a vast improvement over using auditing flags without considering network effects, with ROC curves shown in Figure 2. It is particularly effective for low false positive rates: for the same false positive rate of 5%, our method achieves a 70% true positive rate while the baseline achieves less than 30%.
Future Work and Vision

For the next steps in my research I will focus on applications predicting behavior of networks, which will use knowledge of network formation and diffusion that we acquired in previous work. In the long-term I also plan to design intervention techniques for network behavior, based on these predictions.

Next Steps: The first problem I will address is that of making predictions in communities. The goal is to determine whether an online group will succeed or dissolve. This will lend insight into the design of social network software and how it meets certain users’ needs. To accomplish this task, we will compare network properties of different online groups (such as Usenet, Yahoo! Groups, etc.). Are there certain network characteristics or activity patterns that enable groups to succeed? How early can we distinguish features of thriving or dissolving groups? Are there specific precursors to group dissolution? How does individual user engagement play a role in the overall group vitality?

Next, I will expand the study of predicting behaviors in other networks, both on local and global scales. On a local scale, I plan to perform link prediction—whether two nodes will continue to interact, and with what frequency. It is also of interest to predict when a user will join or leave a given sub-community based on network effects. Such predictions are useful in personalization—such as knowing whether a given user will click on an ad, adopt a product, or continue use of a service.

I also plan to predict network behaviors at large. Models such as the Butterfly model provide a mechanism for understanding what is likely to happen; however, it is important to detect anomalous behavior early, and to predict significant changes in the network. Given a record of phone calls or health care transactions, can we detect fraud early? Can we use our understanding of computer network data to control virus propagation?

Vision: In the longer term, I plan to use these predictions to design intervention measures, to potentially change outcomes. For instance, once we understand what graph properties and user behaviors are precursors to a community dissolving, are there ways to keep users...
engaged, so that they will continue to use a given social network software? Can we incentivize certain links to make communication more efficient, so that users can be best informed? Can we design software that will report early stages of anomalous behavior in a phone call network, to prevent losses from fraud schemes?

Understanding the formation and evolution of networks can have a large impact in a wide range of areas, both within computer science and across disciplines. For instance, detecting anomalies is useful in network security for preventing the spread of viruses; forecasting network growth can aid in operations research problems such as resource allocation; and predicting information diffusion and other network effects can improve marketing efforts such as user targeting and brand loyalty. I believe that these are only a few of the opportunities that a better understanding of network data will offer.