Research Statement

Bo Jiang
bjiang@cs.umass.edu
http://people.cs.umass.edu/~bjiang/

OVERVIEW

My research interests are in modeling and analysis of computer and social systems at large, with an emphasis on computer and social networks. Most complex systems, be they technological or social, are naturally modeled as networks, or, from a mathematical point of view, graphs. In addition to traditional computer networks, prominent examples include online social networks such as Facebook and Google+, which make visible social ties between individuals. Another example is furnished by the online encyclopedia, Wikipedia, which is a substantiation of the structure of human knowledge.

My research goal is to provide in-depth understanding of structures and behaviors of large scale complex systems and then to inform or improve system design based on newly gained insights. My approach combines theoretical modeling and analysis and data analytics, with an emphasis on theoretical analysis. On the other hand, the availability of large scale datasets has made data mining and machine learning techniques valuable for network study, and we often face the hard problem of properly interpreting the results obtained from data. One aspect of my research is to build our understanding of empirical statistics on a firm theoretic basis. My work on reciprocity of social networks with capacity constraints \([3]\) is an example along this line. A brief description of my previous research achievements is presented below, followed by some directions for future research.

PREVIOUS RESEARCH ACHIEVEMENTS

My work covers a broad range of topics as illustrated by the following examples.

- **Dynamics of Cumulative Advantage Competitions** (JSTAT \([1]\), SIGMETRICS Best Paper Award \([2]\)). Cumulative advantage (CA) refers to the rich-get-richer phenomenon that has been proposed as a mechanism to help explain the ubiquity of power laws observed in empirical data such as the power-law degree distribution for many real networks. The online social networking boom greatly facilitates viral marketing that tries to exploit this CA effect. We studied in \([1, 2]\) how CA affects competitions for resources using the generalized Pólya urn model that incorporates both intrinsic fitness of competitors and nonlinear CA feedback. We focused on competition duration and intensity, the former being the time until a winner emerges and the latter the number of leadership changes throughout the course. In particular, we characterized the tail distributions of duration and intensity in different regimes of fitness and CA feedback strength. We found that competition duration always exhibits a power-law distribution when CA feedback is strong, or more precisely, (super)linear, irrespective of the relative fitness of competitors. When CA feedback is weak, or sublinear, however, competition duration may or may not have a power-law distribution depending on whether competitors are equally fit. Perhaps counterintuitively, we found that the CA effect may result in extremely long competitions when one competitor is fitter than the other.
In particular, although a competitor with slightly higher fitness eventually wins when CA feedback is linear, it may have to remain at a disadvantageous status for an extremely long time unless it starts out leading by a sizable gap. When CA feedback is superlinear, the fitter competitors may lose and may again struggle for a long time when it does end up winning. These are in sharp contrast to competitions with no or weak CA effect, where a competitor with higher fitness quickly wins. On the other hand, competitions are rarely intense in terms of changes in leadership unless the competitors are equally fit and start out with comparable initial resources.

- **Reciprocity in Social Networks with Capacity Constraints** (SIGKDD [3]). Reciprocity is a key structural property for directed networks, prominent examples of which include the World Wide Web, Wikipedia, and online social networks such as Twitter and Google+. Previous work has argued about the important role played by reciprocity in various technological, informational and social networks, and reported the empirical values of reciprocity for these networks. The traditional way to interpret these values is to compare them with the expected reciprocity in some random models and classify networks as reciprocal or anti-reciprocal. Most real world networks turn out to be reciprocal. However, we demonstrated in [3] that this alone does not suffice for the purpose of evaluating the tendency for reciprocation. To this end, it is important and informative to know the maximum reciprocity achievable by any network satisfying the same set of constraints as does the network under investigation. In particular, we studied the problem of maximizing reciprocity for networks with prescribed in- and out-degree sequences, which may be intuitively interpreted as capacity constraints, such as fame, popularity and budget of attention. We identified structural motifs largely responsible for making reciprocity in real networks non-maximal. We also provided a simple upper bound on the maximum achievable reciprocity, along with necessary conditions and sufficient conditions for achieving the bound. Our empirical study revealed that social networks exhibit reciprocities surprisingly close to their upper bounds. For example, Google+ has an empirical reciprocity of 34%, compared with an upper bound of 47%. This suggests that, modulo capacity constraints, the tendency for social reciprocation is much stronger than indicated by the empirical reciprocity alone. In contrast, biological networks and some engineered networks exhibit different behavior, with reciprocities much lower than the upper bounds.

- **Cache Networks** (INFOCOM [4], ToN [5], ICN [6]). Caches play an important role in both the current internet and future content-centric networking architectures. An efficient caching policy is critical for achieving good system-level performance and user experience. We investigated in [4] the problem of developing optimal joint routing and caching policies in content distribution networks, where requests can be routed either directly to a backend server, or firstly to in-network caches and then to the server upon a cache miss. We considered two variants of the problem where the delay for the path to the backend server may or may not depend on the server load. When it depends the load, we modeled the delay by that of an M/M/1 queue in [4] and generalized it in [5] to any convex delay rate functions, which include those of G/G/1 queues. We showed that the problem is NP-hard and provided efficient approximation algorithms for its solution. In a slightly different thread, we analyzed in [6] the performance of caches with Pending Interest Tables, a key component in the Named Data Networking architecture for future internet. We considered renewal arrival processes and took into account the time to download contents upon cache misses. We generalized to this scenario and demonstrated the effectiveness of the characteristic time approximation technique, which uses TTL caches to approximate caches that use more traditional policies such as LRU.

- **Information Dissemination on Social Networks** (INFOCOM [7]). Online social networks have become major media for information dissemination. Considerable work focuses on answering the important question of how information propagates on social networks. In [7], we studied how
network structure and user behaviors affect the efficiency of information dissemination as measured by propagation delay. We considered a pull model where individuals actively seek new information but have only limited budgets for attention. For various network topologies, we quantified the efficiencies when individuals act cooperatively and when they act selfishly. We found that if users selfishly optimize their own delay, the resulting Nash equilibrium may lie far from the social optimum in some network topologies. Thus an incentive mechanism is needed to coax users into cooperation. We developed a simple mechanism that incentivizes users into mimicking a stochastic gradient descent algorithm, thus bringing the overall efficiency close to optimality.

In addition to the work described above, I have also been involved in a variety of other work. I led efforts to quantify the efficiency of parallelism in systems that are prone to failures and exhibit power-law processing delays [8] and determine how different multipath congestion controllers affect performance metrics such as link utilization [9]. I also collaborated on the study of percolation in stochastic multilayer networks [10], the investigation of how in-network data aggregation impacts target tracking quality in multi-hop wireless sensor networks under network delays [11], the design of a pacing algorithm that effectively reduces network traffic burstiness [12] and the design of an efficient flooding algorithm for low-duty-cycle wireless sensor networks [13].

**Future Research**

My future work will continue towards my goal of providing in-depth understanding of large scale complex systems and improving system performance. I will generally follow my approach of combining theoretical modeling and data analytics. Below I describe two promising research directions.

- *Dependency between Network Characteristics.* My work in [3] studying reciprocity in networks with capacity constraints is an example of the more general problem of studying dependencies between different network characteristics. In [3], the two characteristics are reciprocity and degree sequence, but many other network characteristics are of interest, such as clustering coefficient, diameter and assortativity. We demonstrated in [3] that specifying the degree sequence imposes a strong limitation on the range of reciprocities that can be found in any network conforming to the given degree sequence. A natural extension is to ask the following more general question. How does the specification of some characteristics restrict the feasible ranges of other characteristics? Most previous empirical studies of real networks generally report different characteristics as independent numbers that are interpreted individually. However, a principled study of their dependency is required for a deeper understanding of the empirical observations. For instance, Twitter and Google+ have been found to have comparable diameters, while their average degrees differ by almost a factor of two. How should we interpret the difference? What might be the cause? Studying the dependency between diameter and degree sequence may provide answers to these questions, which may potentially reshape our overall view of these networks under study as we saw in [3]. This study will combine graph theoretical analysis, algorithmic design and data analysis.

- *When Social Networks Meet Technological Networks.* One important research direction is to study how social networks affect technological networks. Take a cache network as an example. In [4], we studied the joint caching and routing problem with the goal of minimizing average content access delay. There we assumed that requests for different contents are generated at constant rates over time. It is possible for this assumption to be violated, especially considering the word-of-mouth effect exacerbated by flourishing online social network services. Some activities on online social networks may well cause a surge in requests for certain contents, e.g. a YouTube video recommended in a Facebook post or in a Twitter tweet. The cascading effect in online social networks has been well
studied. In my future work, I would like to study how information cascading in social network might affect user demand for videos and how we should change content distribution networks accordingly. In particular, I would like to address the following questions. Does information cascading cause surges in video demands? If yes, can the surges be predicted? How persistent are they? How can we take advantage of the predictions to design caching and routing algorithms for enhancing user experience? This study will necessarily require a combination of data analysis, machine learning, stochastic modeling, optimization and algorithm design.

I am also open to and will enjoy inter-disciplinary collaborations as well as collaborations within information science and technology.

References