Research Statement

Bruno Ribeiro
ribeiro@cs.cmu.edu
http://www.cs.cmu.edu/~ribeiro/
Tel: 617-945-6927  (cell)
412-256-4935 (office)

The overarching theme throughout my work is the principled measurement, analysis, and mining of large-scale, complex social and communication networks. Networks are the natural framework of many of today’s highest impact computing applications: online social networking, Web search, product recommendations, mobile ad-hoc networking, and online dating are just a few examples. My research pioneers the modeling of the long-term dynamics of interactions between user attention and user activity on online social networks. My work was also the first to introduce principled experimental design techniques to the network measurement community: I also provided the first formal characterization of random walks on general dynamic networks and proposed novel sampling techniques for fast and accurate characterization of large networks, overcoming key practical hurdles, while solving long-standing theoretical problems. These and other results are the products of my research philosophy of seeking a deep understanding of problems, often finding solutions that build upon solid theoretical foundations of pioneers such as H. A. Simon, R. A. Fisher, and A. Kolmogorov. My past and current research efforts coalesce into my long-term goal of developing measurement tools and explanatory and predictive models that can measure, explain, forecast, and enhance networked systems where entities interact and evolve in complex ways.

BACKGROUND
The ease with which billions of Internet users unrestrictedly network to simultaneously both create and consume content is transforming economies and societies. Online information systems interact with Internet users to create and sustain the estimated multi-trillion dollar wealth of Internet-based companies and the reach and activity of online social movements. Yet, most of our analysis tools are designed for static network representations, while real networks are large, dynamic, and interacting in non-trivial ways. Our current lack of understanding of the dynamics of these interactions has dire consequences. Take the example of MySpace’s demise – within just a few years of NewsCorp’s $580 million cash acquisition of MySpace, MySpace’s market value all but evaporated. Fortunately, the World Wide Web offers tremendous concrete opportunity to better understand and forecast the complex dynamics of billions of users if we can record their interactions and activities – through principled and scalable methods. With this data, we can build models and metrics to unleash actionable knowledge and innovative algorithms to guide the design and forecast the success of new and enduring networked systems.

CURRENT RESEARCH
I strongly believe that theory and practice go hand-in-hand and that new applications drive new theoretical results. For instance, even my most theoretical results in the study of random walks on dynamic networks are driven by a strong sense of practical purpose. Random walks on static networks find a host of applications in data mining (e.g. ranking and collaborative filtering), machine learning (e.g. spectral learning methods and Markov chain Monte Carlo methods), and computer networking (e.g. ad-hoc routing and distributed consensus). Similarly, I believe that random walks on dynamic networks will play a key role in the development of new data mining tools specifically tailored for dynamic and interacting networks. My work covers a broad range of topics and the following select examples illustrate both the impact of my work and the depth of my research:

Forecasting Internet User Activity in a World with Limited Attention. A 1969 lecture of Herbert A. Simon on the “fallacy of believing that computer systems would necessarily reduce user information overload” got me thinking about the mechanics that drive online social networks to overload their users with information. From this serendipitous thought I created the first model that predicts the success and failure
of online social network startups [WWW’14, WSDM’15]. The key novelty of my model is its use of interacting epidemics to describe user attention, user activity, and user acquisition. I believe that all recent efforts in understanding user influence and acquisition that consider only static network structures are bound to have only limited success. To explain my vantage point, consider that the MySpace network currently has millions of inactive users. A structure-only analysis of MySpace is bound to find many “influencers”. Clearly, however, very little influencing can be done in a network where users pay almost no attention. I believe that limited user attention is paradoxically a widely overlooked and key component in the study of networks. My work on this topic seeks to bridge this gap, designing network tools and methods that take into account the complex way in which user attention interacts with user activity.

IMPACT: This work has received considerable coverage in the public sphere in the form of articles in the specialized and lay media including NSF’s Facebook page (the project is partially funded by an NSF grant in which I am a Co-PI). More details in the Carnegie Mellon University Press Release “In Age of Information Overload, Ability to Sustain Attention Determines Success”.


Developing New Data Mining Tools with Experimental Design Theory. Inspired by R. A. Fisher’s 1935 book on the design of experiments, I introduced the use of the Fisher information metric to the network measurement community in the way Fisher originally intended it to be used: to aid the development of accurate network sampling tools both in frequentist and Bayesian settings [JSAC’13, SIMPLEX’13, IMC’06]. Principled network sampling is a fundamental problem in data mining as networks today have millions or even billions of nodes, making it expensive to run knowledge extraction algorithms on them. Also, sampling using naïve methods can lead to biases and erroneous conclusions.

IMPACT: My experimental design technique has been used by other research groups, with some notable examples including Georgia Tech’s Jim Xu’s (SIGMETRICS’12) and the University of Melbourne’s Darryl Veitch’s, whose works were presented in the Transactions on Information Theory in 2013, the Internet Measurement Conference in 2008 and 2014, and a Ph.D. thesis out of Veitch’s group.

Network Characterization via Transient-free Random Walk Sampling. As noted, random walks find a host of applications in static networks. As online social and technological networks have been growing at a tremendous pace, random walk-based techniques that depend on stationary probability measurements can suffer tremendously long transients, thus making their measurements inaccurate and wasting precious resources when computational resources are constrained. With the goal of reducing random walk transients for network measurements, my work introduces a method that projects a random walk into a high dimensional space to nearly eliminate its undesired transient [IMC’10], opening a new way to design faster and more accurate network characterization tools using random walk sampling.

IMPACT: This original work inspired subsequent works on scalable algorithms to estimate various network metrics such as subgraph concentrations [TKDD’14] along with other network characteristics [INFOCOM’12, CDC’12, WAW’10, IMC’10].

Analyzing Random Walks on Dynamic Networks. A. Kolmogorov pioneered the application of real analysis to the study of stochastic processes. Using Kolmogorov’s framework and results, I provided the first steady-state analysis of random walks on the general class of stationary and ergodic dynamic networks [SIGMETRICS’11, SIGMETRICS’12]. These works came to fill a void in the literature that dates back to Pearson’s introduction of the random walk problem in 1905, where much is known about random walks on static graphs but much less is known about random walks in random environments (e.g. a dynamic network). My subsequent work [Scientific Reports’13] analytically quantifies the biases that time resolution coarsening (a widely used technique to reduce data storage costs of dynamic networks) has on simulated diffusions, paving the way to understand the requirements of new trace compression techniques.

FUNDING
During the last three years, my research efforts have been continuously funded by grants for which I am a Co-PI, totaling over $1.7 million dollars.
3. (Co-PI and Technical Lead) 2-year (2012-2014) Army Research Lab Cooperative Agreement W911NF-09-2-005. Awarded: over $600,000.

**FUTURE RESEARCH**

In a world with limited attention and biased viewpoints, how should fundraiser drives, viral marketing, ranking, and recommendation systems operate? Beyond the widely adopted paradigm of networks and network processes that are fully observable, static, passive, and mostly immutable, there is an exciting new science of networks of partially observable interacting agents, especially when these agents are users with finite attention budgets for interactions. To illustrate the potential game-changing impact of this new paradigm, in what follows I present some of the rewards and challenges associated with my immediate and long-term research goals.

*Forecasting and Influencing the Dynamics of User Attention and Activity.* Imagine having the ability to design social media applications using models that relate user attention (the heart and soul of Web-based businesses) with the way information is routed over the network. For example, in one of my most recent works [WSDM’15], I propose a model that accurately predicted the synchronous decline of user activity of MySpace, Multiply, Hi5, and Friendster after Facebook’s transition from *pull to push* friend activity updates. I showed how Facebook’s *push updates* likely overloaded users, drawing their attention away from Facebook’s competitors, critically upsetting their previously stable balance of user attention and activity.

In my future work, I will incorporate models of attention and activity into a host of data mining tools (e.g. ranking, anomaly detection, and recommendation systems). To illustrate the potential of this research, consider that the amount of attention users give to online social networks depends on the quality and quantity of the content posted by their friends. Facebook, for instance, depends on posts to external webpages to keep users interested in the Facebook webpage; however, these external links also detract user attention from Facebook. Given this symbiotic attention-activity relationship between Facebook and external webpages, how should we incorporate the net attention that a webpage brings to the social network into Facebook’s news recommendation and ranking systems?

*Adapting Random Walk Tools to Dynamic Networks for Unbiased Network Analysis.* Random walks on static networks provide some of the best tools to characterize large networks (e.g. edge correlations [IMC’10] and subgraph concentrations [TKDD’14]), to find network partitions (e.g. spectral partitioning), to uncover communities (e.g. Rosvall & Bergstrom’s Infomaps, 2008), to rank nodes (e.g. PageRank), to perform collaborative filtering (e.g. random walks with restarts), and to recommend new friends to online social network users (e.g. Backstrom & Leskovec, 2011). However, extending random walk-based methods to dynamic and interacting networks is a challenging problem. My prior work shows that random walks on dynamic networks, unlike their static network counterparts, behave in a non-trivial manner [SIGMETRICS’12] and diffusion processes on dynamic networks are sensitive to discretization [Scientific Reports’13]. My work on surveying these obstacles is creating a road map for future breakthroughs. For instance, recently I have been working on principled website crawlers to characterize fast-changing network processes (e.g. Twitter cascades) on large slow-changing networks (e.g. the Twitter follower network), extending my earlier work on estimating immutable network characteristics [IMC’10].

Another noteworthy ongoing work considers the problem of forecasting the evolution of an epidemic process over a large network. I am interested in forecasting, for instance, the evolution of the distribution of infected nodes per infection seed over time. Using random marked point processes and axioms rather than models, Hoang, Singh, and I have developed the first statistical method that can cope with biased
observations and forecast the evolving statistical properties of a real-world epidemic process weeks in advance. Our approach uses Kolmogorov’s axiomatic probability framework to extract information from historical epidemic data [Preprint]. I firmly believe my continued study of random walks and epidemic processes will bring about new data mining tools for dynamic and interacting networks in the near future.

**Performing Targeted Network Dissemination.** The Internet is transforming our society through its ability to quickly mobilize hundreds of thousands of people for a cause or a campaign. However, despite the various existing tools for social mobilization, most of the mobilization mechanics are left to chance. But odds can be adaptively boosted. In a recent work I show that in power law networks it is possible to use a greedy algorithm to maximize the expected number of nodes covered by recruited users, even without prior access to topological information (a myopic variant of the NP-hard connected graph cover problem) [NetSciCom’14]. My other recent studies aim to perform incentivized, targeted, and adaptive step-by-step recruitment over large (unknown) online social networks. The resulting method maximizes the successful recruitment rate, subject to an evolving understanding of the network as the campaign grows. The long-term goal of this research includes building fast model proposal, training, selection, and evaluation procedures and uncovering user latent labor market incentive preferences (based on my previous study on labor market incentives in online social network recruitment [COSN’14]). I also plan to design user features that encode complex blends of network topology and node traits. Currently, Murai, Towsley, Gile, and I have made great progress in developing methods to realize my vision [Preprint]. We are now in the process of implementing these methods into a Facebook App backend that performs scalable and fault-tolerant target disseminations for fundraising campaigns using the Facebook network.

**REFERENCES**