1 Overview

My research interests are in the general areas of networking and distributed systems. Nowadays the Internet is ubiquitous and there are a number of killer applications that run on it. Some examples include Peer-to-Peer systems that enable video streaming (Skype) and file sharing (BitTorrent); Online Social Networks (Facebook, Twitter) and e-mail (Gmail, Hotmail) that facilitate information sharing between users. Many users spend the majority of their time in these systems and generate a large amount of the traffic on the Internet. My research is focused on such internet systems and can be split into the following three subareas.

- measurement techniques and analysis of large-scale data produced by internet systems.
- network modeling and simulation of the processes that take place within internet systems
- design and implementation of networked systems, addressing challenges such as scalability or incentive mechanisms.

Below I provide a brief background and describe my research contributions in each subarea.

2 Research contributions

Network Measurement

• Sampling and Analysis of Massive Online graphs

Graphs are a common way to represent internet systems and the data produced by them. Some examples include social, web, email, and peer-to-peer graphs. In many cases, these graphs are difficult to study because of their massive size or access/collection limitations. Therefore, network-based sampling is required to estimate the properties of such graphs as a step towards understanding them.

In my thesis [7], I addressed questions on how to collect (sampling methodologies) and use (estimation and analysis) a representative sample of a massive graph using network-based sampling.

- In [10, 11], I advocated the use of statistically principled methods, whose bias can be corrected, in the measurement of social graphs. My work demonstrated in a real social graph (Facebook) that common sampling methods that were previously used by the network measurement community, such as Breadth-First-Search, are significantly biased and skew the estimation of local and global graph properties despite a seemingly large sample size. For unbiased estimation, I proposed two random walk-based sampling methods and provided trade-offs and suggestions for their use. Additionally, in the context of graph sampling I used formal convergence diagnostics to assess sample quality during the data collection process.

- In [8], I proposed a novel random walk-based sampling method, Multigraph Sampling, which exploits the existence of multiple relation graphs defined on the same set of nodes in the graph. For example, in social graphs, in addition to the friendship relation, one can consider the group membership relation. Multigraph Sampling obtains a representative sample with faster convergence, even when the individual graphs fail, i.e. are disconnected or highly clustered. Additionally, I designed a computationally efficient algorithm that practically implements Multigraph Sampling as a union of individual graphs.

- In [15], we proposed another novel random walk-based sampling method, Stratified Weighted Random Walk; which takes into account the node properties that are relevant to the measurement objective when setting edge weights during the network sampling. An experimental evaluation in the Facebook social graph shows that a Stratified Weighted Random Walk needs up to 15 times less samples than a simple Random Walk to achieve the same error for the estimation of the college sizes.

During my postdoc, my research focused on the unbiased estimation of graph properties that reveal information about the network structure using a sample of nodes, under a variety of probabilistic sampling techniques (independence sampling and random walks) and designs (induced sampling and star sampling). In [16], we derived estimators that summarize the social graph of an online social network as a category graph. For example, we can use the friendship of users in Facebook that live in different countries to summarize the strength of social relationships between any two countries as a country-to-country category graph. In [12], I derived novel estimators of the degree-dependent clustering coefficient and joint degree distribution. The estimation of such graph properties can be used as an input to network models that generate synthetic graphs resembling the real graph.

The analysis of massive graphs based on crawled samples is another contribution of my work. In [8, 10, 11] the first representative samples from large online social networks, namely Facebook and Last.fm, were collected.
using statistically principled methods and made publicly available to researchers. The datasets have been requested more than 2,500 times by researchers and students in various universities and research institutes.

- I used the Facebook samples to characterize several key topological properties of Facebook (node degree distribution, the degree-dependent clustering coefficient, and assortativity) and user privacy awareness conditioned on geography, degree and neighbor’s privacy awareness.

- I used the Last.fm samples to estimate properties of the relation graphs that comprised the multigraph (friendship, group and event membership) and general properties of the website such as the percentage of paid subscribers and the weekly music charts.

- In [16], we designed a web-based visualization tool, GeoSocialMap (www.geosocialmap.com), that aims to visually represent category graphs embedded in geography and is well suited for data exploration.

- In [13], I studied for the first time the popularity and user reach of Facebook applications. Based on user application installation statistics obtained by crawling Facebook, I proposed a simple and intuitive model to simulate the process with which users install Facebook applications. Using this model one can determine the user coverage from the popularity of applications, without detailed knowledge of how applications are distributed among users. Such prediction can be useful to parties interested in reaching users via applications.

Some of my papers were studied at advanced graduate classes at UMass, University of Minnesota, and HKUST.

• Network security

Denial-of-service attacks pose a real threat in the operation of computer networks. They typically exhaust resources and disrupt the normal operation of a target network host or subnet. In this area, I identified new attacks that can leverage the traffic produced by peer-to-peer systems and improved the traceback of attack sources.

- In [6], I identified vulnerabilities in the design of BitTorrent that can be exploited to use the system as a platform for launching denial-of-service attacks to hosts outside the peer-to-peer swarm. I measured and characterized the attack traffic from live Internet experiments that demonstrate the feasibility of such attacks. As a result, our insights received publicity from the New Scientist [2] and Slashdot [4].

- In [19], we designed a practical traceback scheme that combines probabilistic packet marking with network coding. The purpose of a traceback scheme is to identify the source and node traversal path of the traffic. Inspired by the coupon collector’s problem we mark packets with random linear combinations of router IDs, instead of individual IDs, so as to reduce the number of packets required to reconstruct the attack path and the total traceback time.

• Network tomography

In [9, 20], we developed new techniques to infer loss rates of individual network links by sending and collecting probe packets at the edge of a network that supports network coding, such as wireless networks, peer-to-peer networks and overlay networks. This is the first work in network tomography that leverages network coding capabilities, i.e. allows intermediate nodes to process and combine, in addition to just forward, packets. Our techniques improve estimation accuracy, bandwidth efficiency and link identifiability.

Networked Systems

During my PhD, I designed several novel networked systems, each with specific demands and challenges. I typically build and deploy prototypes of such systems on the Internet. That allows me to evaluate their performance and design tradeoffs in a real environment.

Kangaroo [21, 22] is a scalable peer-to-peer Video-on-Demand streaming system that supports forward and backward seeking that I built in collaboration with Telefonica Research. Using key mechanisms, algorithms, and data structures at the peers (neighborhood manager, scheduler) and tracker (scalable smart tracker), Kangaroo supports seeking operations without the need for over-provisioned peers or server. We evaluated its performance in a wide area network during the 2008 Olympics and accomplished technology transfer to Telefonica. Kangaroo won the New Internet Application (NAI) 2007 award in Spain [3] and received publicity from Gigaom [5].

Adball [1] is a company that I co-founded and directed as the CTO. It provides an online collaboration and social networking platform that crowd-sources the concept, content generation, and dissemination of advertising
campaigns. It brings order to the crowd-sourced creative process, by (i) organizing the collaboration among its users, (ii) enabling a multi-staged collaborative process, (iii) ranking the produced content, and (iv) assigning reputations to users based on their interactions within the platform. Adball won two entrepreneurship awards and served ~20K users during its five-month online presence.

In [11], I built a Social Graph Crawler to tackle technical challenges faced during data collection in large-scale Online Social Networks. The crawler (i) operates in rich content websites, (ii) implements several principled sampling methods and (iii) performs distributed data fetching primarily to reduce the effect of user growth in the collected sample and secondarily to bypass data access limitations. More importantly, the crawler allowed me to perform network measurements on real large-scale graphs such as Facebook, and Last.fm (see section 2).

3 Current and Future Research

In the future I will continue to work in the subareas of network measurement, network modeling and system design. My main area of work will be the development of measurement and data analysis techniques, algorithms, and models in large networks as a means towards understanding and describing their behavior. My goal is to establish connections with other disciplines, in addition to computer science, such as biology, sociology, and economics, which also represent data using graphs. Here are some specific projects that I am currently undertaking in each subarea.

Network Measurement. Recent work has shown that counting the number of local subgraphs (or motifs) within graphs can provide important insights and help us better understand them. We are developing unbiased techniques that can efficiently estimate the frequency of any subgraph through graph sampling [14]. The techniques are flexible to support both directed and undirected graphs, and annotated node attributes. They operate under the presence or absence of node IDs during sampling.

We are examining the similarities and differences between a mobile call graph and a user interaction social graph in regard to the local interactions between users in the temporal and spatial dimensions. Our analysis is based on a private dataset provided by a telecom operator and a dataset collected by crawling Twitter.

Modeling. In [12], I developed a model that builds on the dK-series framework [17] and can be used to generate synthetic graphs with key properties resembling those of a real yet not fully known graph. The model exactly specifies degree correlations and clustering and is shown to be a sweet spot between accurate representation of the original graph and practical constraints. I showed how to estimate the model parameters by sampling the graph and presented new algorithms that efficiently generate synthetic graphs with the specified model parameters. We are now developing efficient generation algorithms of dK-series models [17] beyond 2K, which specifies the joint degree distribution, e.g., 2.5K and 3K models. Previous work typically uses Markov chain Monte Carlo methods to traverse the space of graphs with specified target properties beyond 2K, which can be fairly slow in large graphs.

Networked Systems. We are building a prototype Android app of QuestCrowd [18], a location-based system that uses crowd-sourcing to answer simple day-to-day questions. It supports realtime questions along with questions that can be answered using factual information by people not present at the location. Key design mechanisms that set it apart from similar location-based systems include (i) a reputation system that motivates quality contributors by rewarding high-quality answers, and (ii) a forwarding mechanism that leverages existing social graph relations to increase participation.

4 Summary and Collaborations

In my future research, I will continue to undertake practical real-world problems in the area of networked systems. My approach is to take a crack at a specific problem and identify the challenges involved. Then, I examine whether it is possible to apply existing theory/algorithms in this new context or whether I need to produce new theory/algorithms. Finally, I make the best effort to release together with my publications accompanying software or datasets so as to allow the results of my research to be reproduced and make a contribution to the community.

I look forward to establishing collaborations with researchers from different disciplines, other institutions and the industry. I find that interdisciplinary collaborations can help one get the best perspective to look at a problem and can lead to a more creative and simple solution. Additionally, industry collaborations can become a source of real-world problems, with potential for large impact, that are waiting for a solution. At UC Irvine, I have enjoyed working with people from various departments including Engineering, Computer Science, and Sociology. I also had productive collaborations with other universities (EPFL, UC San Diego, and IT University of Copenhagen); research labs (Telefonica Research, and AT&T Research) and startups (Adball).
References


