Research Statement

Adriana Iamnitchi

My research is rooted in distributed systems, with emphasis on characterizing cyber-social systems and designing, implementing and experimenting with algorithms, services and applications for large-scale networked-systems. In a typical project cycle I quantitatively characterize socio-technical phenomena at scale, model them, apply new understandings to the design of distributed systems, and experimentally measure the outcomes. In the process I often rely on, and contribute to, research from other fields. Recently I have used research from sociology, psychology and political science to build better understandings of quantitative observations or to inform my design and experiments. Joint publications [1, 2, 14, 23, 24] and federal funding with Dr. Skvoretz from the Sociology Department of University of South Florida best capture the results of my interdisciplinary research.

While my recent work is related mainly to online social interactions and big data processing, the same research practice (of quantitatively evaluating socio-technical environments and then applying observations to the design of distributed systems or services) defines my previous work in scientific grids [6, 7] and peer-to-peer systems [8, 9, 3]. In the following I limit to the presentation of my most recent research results, followed by a description of ongoing research.

1 Recent Work

With the plethora of digital records of our social interactions with other Internet-connected users, I became interested in two questions: *What can we learn about us as a society given the records of our many actions?* And *how can these insights influence the design of distributed services for better scalability and performance?* I present below how my recent work is addressing these two questions.

1.1 Measuring Socio-Technical Phenomena At Scale

In studying online social networks I am interested in discovering patterns in user behavior that either unveil important lessons about the offline world or can help solve problems in online environments.

1.1.1 Cheating in Online Games as a Social Contagion

Understanding and quantifying the factors that lead to cheating in society is problematic, due to people's inherent desire to hide socially unacceptable actions. While significant progress has been made in understanding unethical behavior via in-lab experiments, little has been measured at scale, in the wild. Our work [2] was the first to look at cheaters in online gaming from the perspective of social network analysis: we wanted to understand if cheaters[1] are pariahs in their online social network or if they are in positions of power. It turned out that neither was the case: the cheaters were members of the community with no different social network characteristics (number of social contacts, centrality, etc.) than fair players. The only notable difference was that cheaters tended to have more cheater friends than fair players did. This observation led to the discovery that cheating in online gaming spreads as a social contagion [1], a phenomenon that had been observed in offline settings, such as education environments.

The mechanisms of this contagion phenomenon eluded us for long, due to unavailable data. Finally, with the release on the web of timing information on when the cheating label was applied, we built an accurate timing of influence and investigated factors that favor or limit contagion.

[1] Online game players on the Steam gaming platform who are caught cheating have their accounts marked with a cheating label that is permanent, publicly visible, and prevents them from playing the game they cheated in on secure game servers. Cheaters remain members of Steam and may play other games unrestrictedly.

Specifically, we were inspired by psychology studies on cheating influence done by Gino, Ayal and Ariely [4] and processed our dataset to mimic their in-lab experiments. Our empirical analysis [25] confirmed two in-lab results: when cheating takes place and is not visibly punished (yet), there is an evident increase in cheating behavior from the players who may be observing cheating in action. When cheating is punished (by the publicly visible application of the cheating label), fewer players end up cheating. We verified again on different datasets from the same source that cheating behavior is contagious.

1.1.2 Content Abusers in Yahoo Answers

Inspired by our work on cheating in online social games, we became interested in understanding unethical behavior in other online social communities. Thanks to a research grant from Yahoo!, we were able to study a community question-answering platform, Yahoo Answers, with a different type of unethical behavior: violating community rules that forbid spamming, ranting, self-promotion, homophobic comments, etc.

Community question-answering platforms are crowd-sourced services for sharing user expertise on various topics, from mechanical repairs to parenting. They rely on their users for monitoring and flagging inappropriate content. Common wisdom tells that the more flags a user receives, the more toxic the user is for the community and perhaps s/he should be suspended. However, our analysis [13] shows that the number of flags does not tell the full story. On one hand, our results show that users with many flags contribute positively to the community by posting well-appreciated answers and increasing the participation per discussion thread. On the other hand, we observed that users who never get flagged are found (by human moderators or by automatic detection) to violate community rules seriously enough as to get their accounts suspended.

However, our analysis also suggested a way to identify the bad users not solely on the number of flags received. We observed that users whose accounts are suspended have particular social network properties: we found strong evidence of homophilous behavior and used this finding to detect abusive users who go under the community radar and never get flagged. Drawing on this research, we were able to build a classifier that can detect abusive users with high accuracy, which can reduce the human moderator costs significantly.

1.1.3 Privacy Preferences And Cultural Heritage Dictate Our Online Behavior

Our efforts of understanding the deviant users in Yahoo Answers led us to study how privacy concerns relate to online activity and how the cultures we are part of shape our patterns of online contributions. We learned that privacy preference is correlated with engagement, retention, accomplishments and deviance from the norm [11]. We found that privacy-concerned users have higher qualitative and quantitative contributions, show higher retention, report more abuses, have higher perception on answer quality and have larger social circles. However, at the same time, these users also exhibit more deviant behavior than the users with public profiles.

We also confirmed that behavioral cultural differences exist in community question answering platforms [12]. We found that national cultures differ in Yahoo Answers along a number of dimensions such as temporal predictability of activities, contribution related behavioral patterns, privacy concerns, and power inequality.

1.2 Applying Social Knowledge to Distributed Systems Design

The question that continues to fascinate me (and was the core question for my NSF Career Award) is how to exploit social knowledge in the design of distributed systems.

1.2.1 Quantifying and Exploiting the Strength of Indirect Social Relationships

While direct social ties have been intensely studied in the context of computer-mediated social networks, indirect social ties (e.g., friends of friends) have seen little attention. Yet in real life, we often rely on friends of our friends for recommendations (of good doctors, good schools, or good babysitters), for introduction to new job opportunities, and for many other occasional needs. In this interdisciplinary work we 1) quantified the strength of indirect social ties, 2) validated the quantification, and 3) empirically demonstrated its usefulness for two example applications [24]. We quantified the social strength of indirect ties using a measure of the strength of the direct social tie that connects two people and intuition provided by the sociology literature. We evaluated the proposed metric by framing it as a link prediction problem and experimentally demonstrated that our metric accurately predicts link formation. We showed via data-driven experiments that the proposed metric for social strength can be used successfully for, e.g., predicting information diffusion paths [23] and alleviating known problems in socially-incentivized friend-to-friend systems [22].

1.2.2 Socially-Aware Peer-to-Peer Topology Design

Social applications mine user social graphs for many objectives, best known being personalized marketing and personalized search. When such applications are implemented on a fully decentralized, peer-to-peer (P2P) architecture, the social graph is distributed on the nodes of the P2P system. The traversal of the social graph translates into a socially-informed routing in the peer-to-peer layer. We proposed in [18] the model of a projection graph that is the result of decentralizing a social graph onto a peer-to-peer network. We focused on three social centrality metrics: degree, node betweenness and edge betweenness centrality and analytically formulated the relation between metrics in the social graph and in the projection graph. We demonstrated experimentally the usability of the projection graph properties in designing social search applications and unstructured P2P overlays that exhibit improved performance and reduced overhead [19].

1.3 Distributed Processing Solutions in Support of Social Computing

Answers to the two questions that I mainly pursue in my current research need sometimes support to become complete solutions. Two such examples are described below. The first provides the architecture and services to support, for example, the calculation of the strength of indirect social ties in a decentralized way, to avoid concentration of information typical of a centralized system. The second is an algorithmic solution to the computational challenges of calculating node centrality in very large social graphs.

1.3.1 Services and Architecture in Support of Social Applications

An unprecedented information wealth produced by online social networks, further augmented by location/collocation data, is currently fragmented across different proprietary services. Combined, it can accurately represent the social world and enable novel socially-aware applications. We designed and implemented Prometheus [17, 16], a socially-aware peer-to-peer service that collects social information from multiple sources into a multigraph managed in a decentralized fashion on user-contributed nodes, and exposes it through an interface implementing non-trivial social inferences while complying with user-defined access policies. Simulations and experiments on a global distributed testbed (PlanetLab) with emulated application workloads show the system exhibits good end-to-end response time, low communication overhead, and resilience to malicious attacks.

In support of this service, we proposed an architecture [5] that includes social sensors that capture and interpret social signals based on the interaction between two users, a personal aggregator of social information, the Prometheus data management service that builds and maintains the augmented social graph, and a set of social inference functions as this service's API for social

applications. We showed that including social knowledge in the topology overlay is beneficial both for end-to-end performance and for increasing resilience to malicious attacks.

1.3.2 Computing Centrality Metrics in Large Networks

Aware of the computational challenges posed by very large social graphs, we proposed an alternative way to identify nodes with high betweenness centrality in very large networks. In this work [15] we introduced a new metric, k-path centrality, and a randomized algorithm for estimating it, and showed empirically that nodes with high k-path centrality have high node betweenness centrality. The randomized algorithm runs much faster for prohibitively large networks at the cost of limited loss of accuracy. Experimental evaluations on real and synthetic social networks show improved accuracy in detecting high betweenness centrality nodes and significantly reduced execution time when compared with existing randomized algorithms.

2 Ongoing Work

2.1 Anonymizing Large, Dynamic Social Graphs

Recently funded by NSF (under award number IIS 1546453), this project is continuing the interdisciplinary collaboration with the mathematical sociologist Dr. Skvoretz from USF. The work is motivated by the tension between the need for real social graphs datasets, fundamental to understanding a variety of phenomena, such as epidemics, adoption of behavior, crowd management and political uprisings, and the serious privacy risks associated with releasing real datasets: even when humans identities are removed, studies have proven repeatedly that de-anonymization is doable with high success rate. Such de-anonymization techniques reconstruct user identities using third-party public data and the graph structure of the naively anonymized social network: specifically, the information about one's social ties, even without the particularities of the individual nodes, is sufficient to re-identify individuals. The risks are so serious that companies hesitate to share datasets on their customers even with their own in-house research groups, in an attempt to contain user privacy breaches.

The accepted approach now is thus to anonymize social graphs by modifying the graph structure enough as to decouple the particular node identity from its social ties, yet preserving the graph characteristics in aggregate. Various solutions have been proposed, some based on rewiring the original graph structure, others based on clustering, and others based on generating graphs from a graph signature. For all structural graph anonymization techniques, however, the challenge is the tension between providing privacy in the altered graph structure and preserving the accuracy of the structural characteristics of the original graph in the altered graph, which is what matters for their utility for research.

An added challenge, though, is the need for anonymizing longitudinal social networks. If progress in analyzing processes that involve large social networks is to continue, then access to such longitudinal datasets is necessary. Moreover, as we encountered in our previous research [2], social networks often exhibit dual dynamic nature: not only the topology of the graph changes (by node and edge insertion and removal), but also node attributes may change, often informed by the local graph topology.

How to anonymize evolving graphs so as to guarantee privacy while preserving the graph's structural properties for utility is still a challenging problem. Adding labels to nodes increases the privacy challenge. On top of that, adding a dynamic process that changes node labels over time in a natural process makes the problem even more challenging. Addressing these challenges is the objective of our work.

2.2 Privacy as Contextual Integrity

Combining and incorporating rich semantics of user social data, which is currently fragmented and managed by proprietary applications, has the potential to more accurately represent a user's social ecosystems. However, social ecosystems raise even more serious privacy concerns than today's social networks.

The privacy challenge is fundamentally due to the lack of a universal framework that establishes what is right and wrong [21]. Nissenbaum proposed such a framework in her formulation of privacy as contextual integrity [20]. Our preliminary work [10] proposed to model privacy as contextual integrity by using semantic web tools and focused on defining default privacy policies, as they have the highest impact. The model implements the basic concepts of Nissenbaums privacy framework: social contexts, norms of appropriateness, and norms of distribution. Through a real implementation and performance evaluation we showed that such a framework is practical. What is missing is the core of applying Nissenbaums framework to the cyber-social ecosystem: an evaluation of the effectiveness and feasibility of such a solution in a realistic environment.

References

[1] J. Blackburn, N. Kourtellis, J. Skvoretz, M. Ripeanu, and A. Iamnitchi. Cheating in online games: A social network perspective. *ACM Trans. Internet Technol.*, 13(3):9:1–9:25, May 2014.

[2] J. Blackburn, R. Simha, N. Kourtellis, X. Zuo, M. Ripeanu, J. Skvoretz, and A. Iamnitchi. Branded with a scarlet "c": cheaters in a gaming social network. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 81–90, New York, NY, USA, 2012. ACM.

[3] C. Borcea and A. Iamnitchi. P2P systems meet mobile computing: A community-oriented software infrastructure for mobile social applications. In *Self-Adaptive and Self-Organizing Systems Workshops*, pages 242–247. IEEE Computer Society, 2008.

[4] F. Gino, S. Ayal, and D. Ariely. Contagion and differentiation in unethical behavior the effect of one bad apple on the barrel. *Psychological science*, 20(3):393–398, 2009.

[5] A. Iamnitchi, J. Blackburn, and N. Kourtellis. The social hourglass: An infrastructure for socially aware applications and services. *IEEE Internet Computing*, 16:13–23, 2012.

[6] A. Iamnitchi, S. Doraimani, and G. Garzoglio. Filecules in high-energy physics: Characteristics and impact on resource management. In *15th IEEE International Symposium on High Performance Distributed Computing (HPDC)*, pages 69–79, June 2006.

[7] A. Iamnitchi, S. Doraimani, and G. Garzoglio. Workload characterization in a high-energy data grid and impact on resource management. *Journal of Cluster Computing*, 12:153–173, June 2009.

[8] A. Iamnitchi and I. Foster. Interest-aware information dissemination in small-world communities. In *14th IEEE International Symposium on High Performance Distributed Computing (HPDC)*, July 2005.

[9] A. Iamnitchi, M. Ripeanu, E. Santos-Neto, and I. T. Foster. The small world of file sharing. *IEEE Trans. Parallel Distrib. Syst.*, 22(7):1120–1134, 2011.

[10] I. Kayes and A. Iamnitchi. Aegis: A semantic implementation of privacy as contextual integrity in social ecosystems. In *11th International Conference on Privacy, Security and Trust (PST)*, July 2013.

[11] I. Kayes, N. Kourtellis, F. Bonchi, and A. Iamnitchi. Privacy concerns vs. user behavior in community question answering. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2015, Paris, France, August 25 - 28, 2015*, pages 681–688, 2015.

[12] I. Kayes, N. Kourtellis, D. Quercia, A. Iamnitchi, and F. Bonchi. Cultures in community question answering. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media, HT 2015, Guzelyurt, TRNC, Cyprus, September 1-4, 2015*, pages 175–184, 2015.

[13] I. Kayes, N. Kourtellis, D. Quercia, A. Iamnitchi, and F. Bonchi. The social world of content abusers in community question answering. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 570–580, 2015.

[14] I. Kayes, X. Qian, J. Skvoretz, and A. Iamnitchi. How influential are you: Detecting influential bloggers in a blogging community. In *Proceedings of the 4th international conference on Social Informatics*, pages 29–42. Springer Berlin Heidelberg, 2012.

[15] N. Kourtellis, T. Alahakoon, R. Simha, A. Iamnitchi, and R. Tripathi. Identifying high betweenness centrality nodes in large social networks. *Social Network Analysis and Mining*, pages 1–16, 2013. 10.1007/s13278-012-0076-6.

[16] N. Kourtellis, J. Blackburn, C. Borcea, and A. Iamnitchi. Special issue on foundations of social computing: Enabling social applications via decentralized social data management. *ACM Trans. Internet Technol.*, 15(1):1:1–1:26, Mar. 2015.

[17] N. Kourtellis, J. Finnis, P. Anderson, J. Blackburn, C. Borcea, and A. Iamnitchi. Prometheus: User-controlled p2p social data management for socially-aware applications. In *11th International Middleware Conference*, November 2010.

[18] N. Kourtellis and A. Iamnitchi. Inferring peer centrality in socially-informed peer-to-peer systems. In *11th IEEE International Conference on Peer-to-Peer Computing*, September 2011.

[19] N. Kourtellis and A. Iamnitchi. Leveraging peer centrality in the designof socially-informed peer-to-peer systems. *Parallel and Distributed Systems, IEEE Transactions on*, 25(9):2364–2374, Sept 2014.

[20] H. Nissenbaum. Privacy as contextual integrity. *Washington Law Review*, 79(1):119–158, 2004.

[21] H. Nissenbaum. A contextual approach to privacy online. *Daedalus*, 140(4):32–48, 2011.

[22] X. Zuo, J. Blackburn, N. Kourtellis, and A. Iamnitchi. The power of indirect ties in friend-to-friend storage systems. In *Proceedings of the IEEE International conference on Peer-to-Peer Computing*. IEEE, 2014.

[23] X. Zuo, J. Blackburn, N. Kourtellis, J. Skvoretz, and A. Iamnitchi. The influence of indirect ties on social network dynamics. In *Proceedings of 6th International Conference on Social Informatics*. Springer, 2014.

[24] X. Zuo, J. Blackburn, N. Kourtellis, J. Skvoretz, and A. Iamnitchi. The power of indirect ties. *Computer Communications*, 73, Part B:188 – 199, 2016. Online Social Networks.

[25] X. Zuo, C. Gandy, J. Skvoretz, and A. Iamnitchi. Bad apples spoil the fun: Quantifying cheating in online gaming. In *International AAAI Conference on Web and Social Media*, 2016.