Sampling from Social Activity Networks to Understand User Behavior and Systems

Jennifer Neville
Departments of Computer Science and Statistics
Purdue University
West Lafayette, IN 47907
neville@cs.purdue.edu

Online social activity and interaction is becoming embedded into the fabric of our society. From electronic communication (e.g., email, IMS) to social media (e.g., blogs, wikis) to online content sharing (e.g., facebook, flicker, youtube)—we are currently undergoing an explosive growth in the manner and frequency in which people interact online, both with each other and with content.

The ability to collect and analyze large-scale, complex datasets has recently transformed the fields of computational biology and physics—and the explosive growth of social interactions online offer the potential for social science to undergo a similar transformation. Since the traces of electronic activity, including the production and consumption of content, provides a wealth of data that is much more extensive (and more cost effective to collect/observe) than what has previously been studied in social science domains, it could be used to vastly improve our understanding of interpersonal behavior, social processes, and decision making. At the same time, from a systems perspective, the use patterns in social systems (i.e., structure of content and traffic) may be quite different from other networks/traffic on the Internet. The online activity data can thus be used to understand social behavior as it relates to both Internet traffic and load on the infrastructure of online social networks, in order to drive the development of appropriate models, tools and systems to manage and maintain online content and interactions.

However, the size and scope of online interaction data make it impractical to collect and study *complete* datasets. In 2009, for example, Facebook reported that the number of chat messages had exceeded one billion per day. Thus, network *sampling* methods are critical to selecting a subset of the data for study. Much of the previous research on social network sampling has focused on algorithm development, with the aim of accurately and efficiently select nodes/edges/subgraphs from a single large graph (Leskovec & Faloutsos 2006; Hubler *et al.* 2008; Ribeiro & Towsley 2010; Maiya & Berger-Wolf 2011; Ahmed *et al.* 2011).

For example, consider an input graph $G = (V, E)$ of size $n = |V|$. Then the goal is for the sampling algorithm to select a subgraph $G_s = (V_s, E_s)$ with a subset of the nodes ($V_s \subset V$) and/or edges ($E_s \subset E$), such that $|V_s| = \phi n$, where $\phi < 1$ is the sampling fraction. In some cases, the aim is to use G_s to estimate parameters of the full graph (e.g., degree distribution). In other cases, the aim is to have G_s be a *representative* subgraph from the the full graph. Since a complete graph is rarely available for evaluation, the proposed sampling methods are typically assessed by measuring the similarity between characteristics of selected sample and those of the input graph, which is inevitably a sample itself. Key technical challenges that have been investigated include:

- How to sample when the data are heterogeneous and interdependent (e.g., networks are sparse, but heavy-tailed with clustering) (Leskovec & Faloutsos 2006; Hubler *et al.* 2008; Maiya & Berger-Wolf 2011).
- How to sample without knowledge of the full graph (e.g., users are only visible through queries) (Ribeiro & Towsley 2010).
- How to sample in a dynamic environment when there are not enough resources to store the full graph (e.g., in graph *streams*) (Ahmed *et al.* 2011).

However, these efforts have generally not considered the larger issue of how sampling impacts the analysis and understanding of social processes and performance of social systems (e.g., the performance of a new routing protocol for an OSN system, or the accuracy of a viral marketing model). In particular, they have focused more on preserving properties of the network structure, rather than on providing accurate assessment of the properties of processes *overlaid* on the network structure. Although in some cases, preserving aspects of the network topology in a sample may be *sufficient* to accurately estimate the characteristics of processes overlaid on the network, it may not be *necessary*, nor may it be the only manner in which we can accurately estimate performance.

Moreover, there has been relatively little attention paid to developing the theoretical foundation for sampling from *network processes* that would drive the investigation of these types of questions. For example, if the aim is use G_s to evaluate performance of a process $f(.)$ on a larger graph G (where $n >> m$), then the algorithm evaluation should assess sample "representativeness" by estimating an empirical distribution for the process overlaid on the generated samples $\hat{P}(f(G_m))$ and compare it to process in the original graph $P(f(G))$. For example, if $f(.)$ is a diffusion process that models the spread of information in a social network, then we would like our evaluation of f in G_s to accu-

rately reflect the diffusion properties of f that would be observed in the full graph G. However, to begin to formulate and assess sampling algorithms in this manner, we need a more precise description, and better understanding, of graph *populations*—both with respect to the distribution of possible worlds and their dynamics/evolution over time.

A *statistical population* is typically defined as the set of all items that one wishes to study. When the object of study is an entire network, the population should be defined as a set of networks of a specified size (e.g., n), or the set of networks that could be generated by the same underlying process that created the input network G. In practice, we can rarely observe multiple networks from the same social network domain. There is only one Facebook friendship graph, one Flicker graph—although we can down-sample many smaller networks from these large networks, we cannot measure a second, independent instance. Instead, these networks correspond to *complex systems* evolving over time. Therefore, it is more reasonable to define the population through the process that underlies the formation of the networks. Although it is still an open question as to how to model the *generative* processes of network structure probabilistically, this will be critical to the investigation of sampling methods and their impact on subsequent network analysis.

For example, since many graph characteristics are *not independent* of graph size, it is not clear what structure in the smaller subgraphs will give an accurate estimate of the performance in larger graphs (as they evolve over time). Consider the case where the original network consists of n nodes and we construct a 10% sample (i.e., $|V_s| = 0.1n$). Let $|E_o|$ and $|E_s|$ be the number of edges in the original and the sampled network respectively, and let the density in the original network be $\frac{|E_o|}{n(n-1)}$. Then if we match the density in the sampled graph: $\frac{|E_s|}{0.1n(0.1n-1)} = \frac{|E_o|}{n(n-1)}$, the number of sample edges will be: $|E_s| \simeq 0.1^2|E_o| < 0.1|E_o|$. This shows the dependency of graph metrics on graph size—the number of nodes grows linearly, but the number of possible edges grows quadratically (in n). In G, the average degree is $\bar{d}_o = \frac{|E_o|}{n}$. On the sample subgraph G_s, if the density is equal to that of the original graph, then the average degree will be underestimated $\bar{d}_s = \frac{|E_s|}{0.1n} \simeq \frac{0.1|E_o|}{n}$. Similarly, if we aim to capture the original average degree in the sample, then the density will be overestimated. It is not clear which metric to optimize to select "better" sample graphs for evaluation of performance.

Moreover, since none of the recent work on graph sampling includes an explicit definition of the population of interest or a description of the set of events under consideration, this has led to *subjective* evaluations of algorithm performance where the similarity of the sample to the original graph is used as an indirect proxy for representativeness. *Representativeness* of a sample subgraph G_s should be measured through the likelihood of G_s given the underlying process that generated G. The primary assumption with the current proxy evaluation is that when the sampled network exhibits graph metrics similar to the original input network, then the sample is "close" to the mode of the distribution. However, since the underlying distribution is not formally defined, it is not clear whether this assumption holds in practice. Moreover, when the statistics do not match exactly (as is most often the case), a secondary assumption is that "closer" implies "more" representative. Again, it is not clear whether this holds for real world network processes (e.g., we do not know how much variance is expected in real-world network systems).

To develop a better understanding of how sample structure affects the analysis of behavior and performance in larger networks, more attention needs to be paid to the exploration of correlations among graphs properties, how they evolve given a specific generative mechanism, and how to model probability distributions over graph processes. We note that current statistical models of graphs focus on modeling graphs of a specific size (i.e., the number of nodes is fixed). Size independent graph models exist (e.g., Lovász's graphon), but the current state of the art does not address sparse graphs, nor are there any formal notions of the statistical properties of graph *processes* (e.g., stationarity).

Network sampling is critical for analyzing online social interaction data, both for system development and for investigation and refinement of social theories. We note that since almost *every* network dataset is a sample of network data, the sampling method can impact the accuracy of analysis even when researchers have not explicitly considered *how* to sample. The key aspect of the data—the relationships among users, content, and applications—is also the characteristic that makes it difficult to guarantee unbiased "representative" samples, since local dependencies combine in complex ways to produce global structure. Thus, in order to drive both the advancement of computational social science and the development of robust and reliable social computing systems, more research needs to focus on developing:

- A formal framework for sampling from heterogeneous, partially-observed, interdependent data.
- An understanding of various network characteristics and their dependencies.
- Probabilistic models of dynamic graph processes, that can model network structure as it evolves over time.
- An analysis of the impact of sample *representativeness* on the investigation of social processes and/or system protocols overlaid on the networks.

References

Ahmed, N.; ; Neville, J.; and Kompella, R. 2011. Network sampling via edge-based node selection with graph induction. Technical Report 11-016, CS Dept, Purdue University.

Hubler, C.; Kriegel, H.-P.; Borgwardt, K. M.; and Ghahramani, Z. 2008. Metropolis algorithms for representative subgraph sampling. In *ICDM*.

Leskovec, J., and Faloutsos, C. 2006. Sampling from large graphs. In *SIGKDD*, 631–636.

Maiya, A. S., and Berger-Wolf, T. Y. 2011. Benefits of bias: Towards better characterization of network sampling. In *SIGKDD*.

Ribeiro, B., and Towsley, D. 2010. Estimating and sampling graphs with multidimensional random walks. In *ACM SIGCOMM Internet Measurement Conference*.