Impact of Sampling Design in Estimation of Graph Characteristics

Emrah Cem
Department of Computer Science
The University of Texas at Dallas
Richardson, TX 75080
Email: Emrah.Cem@utdallas.edu

Mehmet Engin Tozal
School of Computing and Informatics
The University of Louisiana at Lafayette
Lafayette, LA 70504
Email: metozal@louisiana.edu

Kamil Sarac
Department of Computer Science
The University of Texas at Dallas
Richardson, TX 75080
Email: ksarac@utdallas.edu

***Abstract*—Studying structural and functional characteristics of large scale graphs (or networks) has been a challenging task due to the related computational overhead. Hence, most studies consult to sampling to gather necessary information to estimate various features of these big networks. On the other hand, using a best effort approach to graph sampling within the constraints of an application domain may not always produce accurate estimates. In fact, the mismatch between the characteristics of interest and the utilized network sampling methodology may result in incorrect inferences about the studied characteristics of the underlying system. In this study we empirically investigate the sources of information loss in a sampling process; identify the fundamental factors that need to be carefully considered in a sampling design; and use several synthetic and real world graphs to elaborately demonstrate the mismatch between the sampling design and graph characteristics of interest.**

I. Introduction

Recently, there has been a significant interest in modeling the real world complex systems as large scale graphs. The resulting graphs, in general, lack a simple structure and consist of numerous interconnected or interacting entities. These graphs are then studied to understand the structural and functional characteristics of the underlying real world systems as well as the interactions among the entities appearing in these systems.

World Wide Web graph, Internet topology maps at various levels, protein interaction networks, online social networks, and actor collaboration networks are among the examples to the complex systems that are represented as large graphs. Note that in complex systems context the terms *graph* and *network* are used interchangeably to refer to a graph data structure and we follow the same convention in this paper.

Studying these complex systems might allow us to model the evolution of the Internet; predict the damage of an epidemic; reveal various characteristics of online social cliques; or identify the hub sites in World Wide Web. However, representing these systems as graphs on computers and analyzing their characteristics is usually uneconomical, impractical, or impossible due to the size and/or inherent limitations of these systems. To illustrate, capturing and studying the entire World Wide Web as a graph is extremely challenging if not beyond the capabilities of current computers [4]; discovering the complete router level topology of an Internet Service Provider (ISP) network is rather difficult unless it is provided by the ISP which is typically not the case due to security and privacy concerns [26], [27]; and mapping all interactions in protein interaction networks is experimentally difficult and time consuming [12].

Given that many of these systems are difficult to capture and/or impractical to analyze in their entirety, most studies consult to sampling to gather necessary information for the analysis. In general, the objective of network sampling can be divided into two: (1) designing sampling schemes that are compliant with the process that generates the population graph, i.e., the sampling design should produce sample graphs from the probability distribution over graphs given by the underlying graph generation process and (2) designing sampling schemes that will help us study the characteristics of a given population graph rather than the characteristics of the process generating that graph. The second objective above can be further divided into two: (2a) estimating characteristics of population graphs such as degree distribution, average clustering coefficient, or variation in path length, and (2b) obtaining a subgraph that preserves *important* topological characteristics of the population graph, i.e., *representative subgraph sampling*. Note that the objective (1) above is different from the objective (2b). Note also that it is difficult to come up with a sampling scheme that would satisfy both objectives (2a) and (2b) simultaneously [2].

Another practical issue that may impact network sampling in the context of the second objective above is the available querying mechanisms. In a given application domain, we may have two types of querying mechanisms available as *backend* querying and *frontend* querying. We say that *backend* querying is available if one can obtain the value of an entity characteristic from the *population* graph, e.g., degree of a sampled node refers to its degree in the *population* graph. We say that *frontend* querying is available if one can obtain the value of an entity characteristic from the *sample* graph, e.g., degree of a sampled node refers to its degree in the *sample* graph. Note that in the context of *representative subgraph sampling*, we are mainly using *frontend* querying as the goal in this case is to build a representative subgraph.

Given that graph sampling may have various different types of objectives as mentioned above, it is important to know how to carefully design a sampling method. We believe that the effectiveness of a given graph sampling design depends on the following design considerations: (a) the characteristic of interest under study, (b) topology of the population graph (e.g., whether it follows scale free, small world, random, or semi-hierarchical graph models), and (c) available querying type (*frontend* vs *backend*). Note that design considerations

that we address is comprehensive, but may not be exhaustive. As an example, a given graph sampling method may perform good for one characteristic of interest but may not perform well for some other characteristic under *frontend* querying [2]. Similarly, a given graph sampling method may preserve a given characteristic (e.g., type of degree distribution) of a random population graph in the sample graph but may not preserve the same characteristic of a scale-free population graph in the sample graph [24]. Finally, as we demonstrate in our evaluations, the available querying type may significantly impact the performance of a given sampling design in studying various characteristics of the population graph.

Based on our prior work in Internet topology mapping and our familiarity with studies in online social network (OSN) analysis, we realize that the above mentioned design considerations are not necessarily utilized in many studies that involve sampling from a population graph. In other words, in many studies in Internet topology measurement and OSN analysis (and possibly many others), there is a tendency to use readily available sampling schemes to quickly put together a sample subgraph and use it for analysis. The potential problems with such an approach include (1) use of a sampling method that is not inline with proper design considerations for the analysis study at hand and (2) misuse of the data obtained during the sampling process. An example for the first problem would be the direct use of metropolized random walk with frontend querying, i.e, without carefully designing an estimator, to study shortest path length distribution of the population graph. An example for the second problem is to use a sampling design that is effective in producing samples for studying a particular characteristic (e.g., path length distribution) but use the resulting samples for studying some other characteristic (e.g., degree distribution) without ensuring the suitability of the sampling method for studying those additional characteristics.

In this paper, we conduct a large scale experimental study to empirically demonstrate the relation between various sampling designs and the accuracy of resulting samples in estimating various characteristics of population graphs. We specifically employ the common practice of utilizing a readily available sampling design to generate a sample graph and use that subgraph with available querying to study various characteristics of the population graph. We then use a divergence metric to compare the similarity between the characteristic of interest obtained by the sampling process and the corresponding population characteristic. For the experiments, we use several types of synthetic graphs and several real world graphs as population graphs. Therefore, the knowledge of the population graph enables us compute the divergence scores as a metric to measure the accuracy of using various sampling methods in studying various graph characteristics. Note that if this were possible in practice, obviously there would not be a need for sampling at all.

Based on the conducted experiments, we observed that the performance of a sampling method significantly depends on the above *design considerations*. Therefore, while making inferences about a population graph using a sample graph within the application constraints, these design considerations should not be ignored. This study calls for further research in understanding the theory behind the ability to make inferences about population graphs using sample graphs obtained by a given sampling design.

The rest of the paper is organized as follows. Section II presents the related work. Section III discusses our sampling framework. Section IV demonstrates the experimental results motivating this study. Finally, Section V concludes the paper.

II. Related Work

Early work in network sampling proposed statistical estimation techniques to accurately capture some fundamental characteristics of the underlying networks such as population size [8]. More recently, network sampling is utilized in several popular networking application domains such as Internet topology measurements (ITM), online social networks (OSN), and peer-to-peer networks (P2P) domains.

In ITM domain, sampling is used to estimate the size of the Internet [28]. Various studies also looked at the statistical properties of the sample networks obtained by traceroute-based path samples [1], [6]. In OSN domain, sampling is used in obtaining nodal or topological characteristics of the underlying social network. Most of the studies have focused on obtaining an unbiased sample or removing sampling bias in the collected sample network [16]–[19], [22]. Other work demonstrated that certain biases can be exploited to increase the inclusion probability of desired properties during sampling [20]. In P2P domain, sampling has been studied with the goal of eliminating sampling bias due to skewed node degree distribution in overlay networks [11], [21], [25].

Another goal in network sampling has been to collect *representative* subnetworks from a population graph [10], [13], [15]. In this context, a subnetwork that matches many popular topological characteristics of the population graph is considered representative. Some of the most popular characteristics considered include degree distribution, clustering coefficient and betweenness characteristics. On the other hand, there has been concerns that building a subnetwork that matches an arbitrary set of popular topological characteristics of the population network may not always be feasible [2], [20].

In our work, we experimentally show that the utilized sampling strategy has an important effect on the representativeness of the resulting subnetwork even when we consider a single topological characteristic to match. We consider both backend and frontend querying and utilize a small but diverse set of sampling schemes including (1) simple node and link sampling, (2) walk based sampling, and (3) path based sampling to demonstrate our points.

III. Sampling in Complex Systems

In the context of complex networks, sampling is typically used to collect a subgraph from the underlying system so as to estimate topological and functional properties of the population graph. As we argued in the introduction section, the accuracy of the obtained results in such a sampling scheme depends on various elements that we refer as *design considerations* that contribute to the sampling process. In this section, we present several examples for each design consideration as well as the sampling methods that were considered in this paper. The examples presented below are commonly used in various sampling studies in sampling of complex systems.

A. Topological Characteristics of Graphs

A graph $G(V, E)$ is an ordered pair of a vertex set $V = \{v_1, v_2, \ldots\}$ and an edge set $E = \{e_1, e_2, \ldots\}$ such that $e_k = (v_i, v_j)$ and $v_i, v_j \in V$. In reality V is the set of objects that are meaningful in a particular domain and E is the set representing interactions or relations among these objects. Topological characteristics of graphs are those features that are related to the structural layout of the graph, e.g., degree distribution, clustering coefficient, betweenness, path length distribution, and diameter. In this study we focus on the topological characteristics of simple undirected graphs where $e_k = (v_i, v_j) \Rightarrow e_k = (v_j, v_i)$ without self edges. Below, we formally define the graph characteristics that we used in our experiments namely degree, clustering coefficient, and path length distributions.

Degree of a vertex v_i, also called the number of neighbors, denotes the number of edges incident to v_i and it is represented by d_i. Degree distribution of a graph $G(V, E)$, is the probability distribution of degrees of vertices in V. In many networks, degree of a vertex shows the popularity or importance of that vertex. For example, high degree vertices in OSN are those people having many friends, in AS level Internet topology networks they, most of the time, correspond to Internet Exchange Points (IXPs).

Clustering coefficient is a measure of closed connections appearing among vertex groups in a graph. Clustering coefficient c_i of a vertex v_i is defined as $\frac{2|R_i|}{d_i(d_i-1)}$, where R_i is the number of connected pairs between all neighbors of v_i. Clustering coefficient distribution of a graph $G(V, E)$ is the probability distribution of clustering coefficients of vertices in V. High clustering coefficient indicates existence of tightly associated cliques in OSN and alternative bypassing paths around a particular router in router level Internet topology graphs. In our experiments, we assumed that vertices with degree ≤ 1, have clustering coefficient of 0.

A path between v_i and v_j is a sequence of consecutive edges starting from v_i and ending at v_j. Given that each edge in a graph is assigned a weight, a shortest path between v_i and v_j is one of the paths having the minimum total weight. In the most common case edges in a simple undirected graph are equally weighted which we also adopt in this paper. Note that, there might be more than one shortest path between any two vertices. Path length , $l_{i,j}$, between v_i and v_j is the number of edges appearing on a shortest path between v_i and v_j. Path length distribution of a graph $G(V, E)$ is the probability distribution of path lengths of vertex pairs in V. Path length between any two vertices in OSN is a measure indicating how easy to reach from one person to another via the intermediate people. Maximum path length in the router level Internet topology maps is called the diameter of the Internet and used as an indication of the maximum number of times that a packet can be enqueued/dequeued between two hosts.

Note that degree of a vertex involves 1-hop information around the vertex in the graph. Clustering coefficient of a vertex involves 2-hop information. Path length distribution is a characteristic that depends on the global structure of the network. As a result, we believe that the selected characteristics correspond to a small but diverse set for analysis purposes.

B. Graph Models

Many types of graphs with different distinguishing characteristics have appeared in the literature. Some graph types, however, occur so frequently in natural and man-made systems that they have been studied and used widely. In this part we introduce three popular graph types which frequently appear in the literarure namely random, scale free, small world networks.

Random graphs are type of graphs that are obtained through a random process. Gilbert/Erdos-Renyi [7] (ER) is a widely used model for generating random graphs. ER model, $G(|V|, p)$, builds a random graph $G(V, E)$ by first creating $|V|$ isolated vertices and then, independently connecting each pair of vertices with probability p. All graphs having $|V|$ vertices and $|E|$ edges have the same probability of being generated, $p^{|E|}(1-p)^{\binom{|V|}{2}-|E|}$, in ER model. However, as parameter $p \in (0-1)$ increases dense graphs are more likely to be generated over sparse graphs. At $p = 0.5$ the model generates any possible graph of $|V|$ vertices with the same probability, $(0.5)^{\binom{|V|}{2}}$.

A graph, $G(V, E)$, having a power law degree distribution is called a scale free graph. Power law degree indicates the existence of a few vertices with very high degrees along with many vertices with less degrees. Degree distribution is said to conform power law if it is in the form of $P\{d_i = d\} \propto d^{-\alpha}$ where α is the scaling parameter [5]. Barabasi-Albert [3] (BA) model is a widely used method for generating scale free networks. BA model is based on two principles (i) incremental growth and (ii) preferential attachment. Incremental growth assumes that the final graph, $G(V, E)$, is generated by introducing vertices one by one. Preferential attachment implies that the more connected a vertex is, the more likely it is to grab new links to be added to the network. In BA model, each new vertex is connected to m existing vertices with a probability proportional to the number of edges that the existing vertices already have.

A graph, $G(V, E)$, having a large average clustering coefficient and a small average shortest path is called a small-world network. Large clustering coefficient implies the existence of highly clustered cliques in the graph. Small shortest paths refers to the existence of hub vertices or hub cliques within the graph. Small shortest paths denote that the average shortest path grows with respect to the logarithm of the number of vertices, i.e., $\mu_P \propto log(|V|)$ where μ_P is the average path length and $|V|$ is the vertex set size. Watts-Strogatz [30] (WS) is the most widely used model for generating small-world networks. Given the number of vertices $|V|$ and the mean vertex degree μ_D, the model creates a circular lattice where each vertex is linked to $\mu_D/2$ closest neighbor vertices on clockwise and counter clockwise. Then, with probability β each edge of each vertex v_i is rewired to a randomly selected vertex v_k such that $i \neq k$. WS model guarantees high clustering coefficients by the initial circular lattice and decreases the shortest path length by randomly rewiring edges. The parameter β controls the properties of the final graph. As $\beta = 0$ the graph would remain as a lattice with high clustering coefficient and high average shortest path. As $\beta \to 1$ the final graph resembles to ER graphs with $p = \mu_D/(|V| - 1)$.

Each of these graphs are used as underlying population graph and we use different sampling techniques to collect

network samples from each of these graphs.

C. Sampling Methods

A sampling method refers to a systematic approach that is used to observe a set of sampling units from the population graph, where sampling unit refers to the entities in the population graph that is selected at each step during a sampling process. Typically, vertices and edges are used as observation units but other structural units such as end-to-end paths, star structures, triangle structures, etc., can also be considered as sampling units. In this paper, we use vertices, edges, and end-to-end shortest paths as observation units.

In practice, the available observation units may be decided by the practical and operational limitations in the underlying application domain. As an example, in OSN analysis domain, implementing a random vertex sampling may be extremely costly if the underlying OSN system has a sparsely populated identifier space. Most OSN applications, on the other hand, allow a random walk starting from a chosen vertex in the system allowing us to sample vertices and edges. On the other hand, in ITM domain directly observing a router or a subnet from a remote location is prohibited due to security and privacy concerns.

In network sampling applications, sampling methods are mainly determined based on the data collection mechanisms supported by the underlying application domain. It is often the case that a supported sampling method may not be an ideal one and may often result in sampling bias or high variance for the estimation process of the graph characteristic under study. In this paper, we consider several sampling methods as discussed below.

Random vertex (RV) sampling selects a number of vertices from a population graph with equal probability. Although the technique is simple and effective in estimating the direct properties of vertices, e.g., degree distribution, with backend querying, it may not be readily supported by some complex systems such as the Internet or may be extremely costly to implement in some complex systems such as an OSN. We also use a version of RV sampling, called Induced RV (IRV) sampling which includes the existing edges from the population graph among the selected vertices into the sample graph. IRV sampling is also known as induced subgraph sampling. Random edge (RE) sampling selects a number of edges from the population graph with equal probability and also selects the incident vertices. This sampling method is also known as incident subgraph sampling. Induced random edge (IRE) sampling first performs RE sampling, then includes the existing edges from the population graph among the selected vertices if not selected during RE sampling. In general, RV, IRV, RE, and IRE sampling schemes may produce disconnected sample graphs.

Random walk (RW) sampling [19] initially designates a random vertex as the *current vertex*. At each step, it selects a new vertex among the immediate neighbors of the current vertex and designates it as the new current vertex. Random walk sampling process is governed by the stochastic process $\{X_n, n \geq 0\}$ with transition probability

$$P_{RW}\{X_{k+1} = v_j | X_k = v_i\} = \begin{cases} 1/d_i & \textit{if } (v_i, v_j) \in E \\ 0 & \textit{otherwise} \end{cases} \quad (1)$$

where X_k is the random variable denoting the vertex selected at step k. For connected graphs with enough many steps the process converges to the stationary distribution $\pi(v_i) = d_i/2|E|$. That is, the process introduces bias by selecting larger degree vertices with higher probabilities compared to the smaller degree ones.

Metropolized random walk (MRW) removes the bias introduced by naive RW by rearranging the one step transition probability matrix of the RW stochastic process. Specifically, the transition matrix for MRW would be

$$P_{MRW}\{X_{k+1} = v_j | X_k = v_i\} = \begin{cases} P_{RW}\{X_{k+1} = v_j | X_k = v_i\} min\{\frac{d_i}{d_j}, 1\} & \textit{if } v_i \neq v_j \\ 1 - \sum_{v_i \neq v_j} P_{MRW}\{X_{k+1} = v_j | X_k = v_i\} & \textit{otherwise} \end{cases}$$

Thereby, MRW sampling reduces the probability of transitioning to high degree vertices to remove the bias introduced by the RW sampling.

Random shortest path (RSP) sampling selects paths among all possible shortest paths with equal probability. The total number of shortest paths on a simple graph $G(V, E)$ is at least $|V|(|V|-1)$. Usually shortest path algorithms picks one path over the others in case there are multiple shortest paths between two vertices. Hence, the number of shortest paths is exactly $|V|(|V|-1)$ in practice. Since enumerating all shortest paths in a graph is costly, RSP can be approximated by using shortest paths of randomly selected (source,destination) pairs uniformly without replacement. The sample graph in this scheme is constructed by carefully merging the shortest paths among the pairs of vertices. Variations of this approach in the ITM domain include sampling among K vantage points, i.e., (K,K) sampling, or sampling between K sources and M destinations, i.e., (K,M) sampling where (K$\ll$M).

D. Query Types

Query type is another important aspect of network sampling. One can analyze the topological properties of the sampled components in two ways: *backend querying* and *frontend querying*. In backend querying, a topological characteristic of a sampled unit is queried (or extracted) from the underlying population graph though, the entity appears in the sample graph.

In frontend querying, a topological characteristic of an entity is queried (or extracted) directly from the sample graph.

Since sample graphs do not preserve all information related to the population graph, frontend querying introduces another source of bias as opposed to backend querying. Therefore, the bias quantification and elimination should be handled differently in frontend and backend querying cases.

IV. Experimental Results

In this section, we conduct experiments to observe the impact of the issues listed in the previous section on network sampling. Our goal is to present instances or evidence that

sampling bias may emerge from various directions including the mismatch between sampling method and the characteristic under study; mismatch between the sampling method and the graph model; mismatch between the characteristic of interest and the utilized sampling unit, or frontend vs backend querying capabilities in sampling. The recent sampling literature includes studies that develop estimators for removing sampling bias for certain types of sampling methods and for certain types of topological characteristics [9], [16], [22]. Given that not all sampling schemes have proper estimators defined, in this work, we do not consider the use of such estimators.

We use the graph models listed in Section III-B as population graphs and take samples from them using the sampling methods described in Section III-C. Then, using each query type described in Section III-D, we compute the distribution of characteristic (among the ones listed in Section III-A) under study for each sample. Finally, to measure the difference between the sample and the population distributions, we use Jensen-Shannon (JS) divergence, a modification of Kullback-Leibler (KL) divergence, that is symmetric and robust with respect to noise and the size of histogram bins [23]. JS divergence is defined as

$$d(H||K) = \frac{1}{2}\sum_i \left(h_i log_2(\frac{h_i}{m_i}) + k_i log_2(\frac{k_i}{m_i}) \right) \quad (2)$$

where $m_i = (h_i + k_i)/2$.

In Equation 2, H and K are population and sample distributions defined over the same sample space. h_i and k_i are the relative outcome frequencies for discrete distributions and relative histogram bin frequencies for continuous distributions belonging to H and K. Note that the Jensen-Shannon divergence takes values between 0 and 1 since we use the base 2 logarithm. In the experiments, we take samples of size 10% from each population graph by using sampling methods described in Section III. Each divergence score presented in this study is the average of 30 experiments. In general as the sample size increases the estimation gets closer and closer to its true population values and 10% is a commonly used sample size in the literature.

A. Results on Synthetic Networks

In this section, we generate random, scale-free, and small-world population graphs by using Erdos-Renyi (ER), Barabasi-Albert (BA), and Watts-Strogatz (WS) models, respectively. These models generate graphs with different topological characteristics which enable us to observe the effects of the graph models on the results. Unless otherwise stated, for each model, we generate synthetic graphs of size 50,000 with average vertex degree 20 to compare the graph models properly. To get the desired average degree, we set the model parameters as BA($m = 10$), ER($p = 0.0004$), WS($\mu_D = 20, \beta = 0.5$). We set the number of vantage points to 3% of the population size in K-K sampling method. In K-M sampling, we set the number of source and destination vertices to 3% and 7% of the population size, respectively. Since K-K and K-M sampling methods are mostly utilized in Internet measurement studies and the number of available 'vantage points' is often very limited, we selected parameters so as to reflect the real-world usage better as well as to sample 10% of the population graph.

Figure 1 presents the Jensen-Shannon divergence values of our experiment results for the three characteristics. We omit random vertex (RV) sampling results in the figure because it is not applicable for the frontend querying (FE) since no edges are sampled and it gives same results as IRV sampling for the backend querying (BE). Note that random walk based sampling methods require the information on the neighboring vertices be available to the sampling process in order to sample the next step vertex. Hence, one might question the use of FE querying for such cases. However, companies in various domains use a random walk based method to sample their proprietary networks and publicize the resulting sample graph for analysis purposes without providing any further access to their population networks.

Based on the results in Figure 1, we can conclude that for any sampling method, BE querying outperforms FE querying when it is applicable to the synthetic graph models. Although this is an expected result, the difference can be attributed to the fact that the vertex based sampling methods tend to under-sample the edges of a population graph. On the other hand, computation of many graph characteristics including degree, clustering coefficient, and path length involves edge information as well. The edge related information of the population graphs is available to BE querying while FE querying can only rely on the edge information provided by the sample graph. Because RV, IRV, and MRW sampling methods select vertices uniformly at random, they generally perform well when used with BE querying. On the other hand, RE, IRE, and RW sampling with BE querying does not produce similar results because they tend to over sample high degree vertices compared to low degree ones. RSP and K-M methods perform better than the K-K method though, RSP is slightly better than K-M for most of the cases. Path based sampling bias mostly occurs around the source and destination vertices. That is, the probability of a vertex being sampled is higher for those vertices which are located closer to the source and destination vertex sets in terms of hop count. In our experiments both the source and destination sets are constructed uniformly at random. On the other hand, K-M and RSP possess more diversity in terms of the destination set compared to the K-K method. The bias due to the diversity around the destination sets is reflected in Figure 1.

Note that quite often real world systems do not support (or provide service for) RV and IRV sampling, e.g., WWW and OSN. Hence, Figure 1 suggests that MRW sampling with BE querying is a practical and well-performing method for clustering coefficient, degree, and path length distribution estimation.

1) Degree Distribution: Figures 1-b and 1-e show the degree distribution divergence scores of different sampling methods on our graph models for BE and FE querying, respectively. Figure 1-b shows that RV and MRW with BE querying outperform all other methods. This observation is not surprising because RV with BE querying is always the best sampling method to estimate any local vertex characteristic. Unfortunately, RV sampling with BE querying is not available in many domains including OSN, ITM, and WWW. MRW with BE querying also performs well because it could be

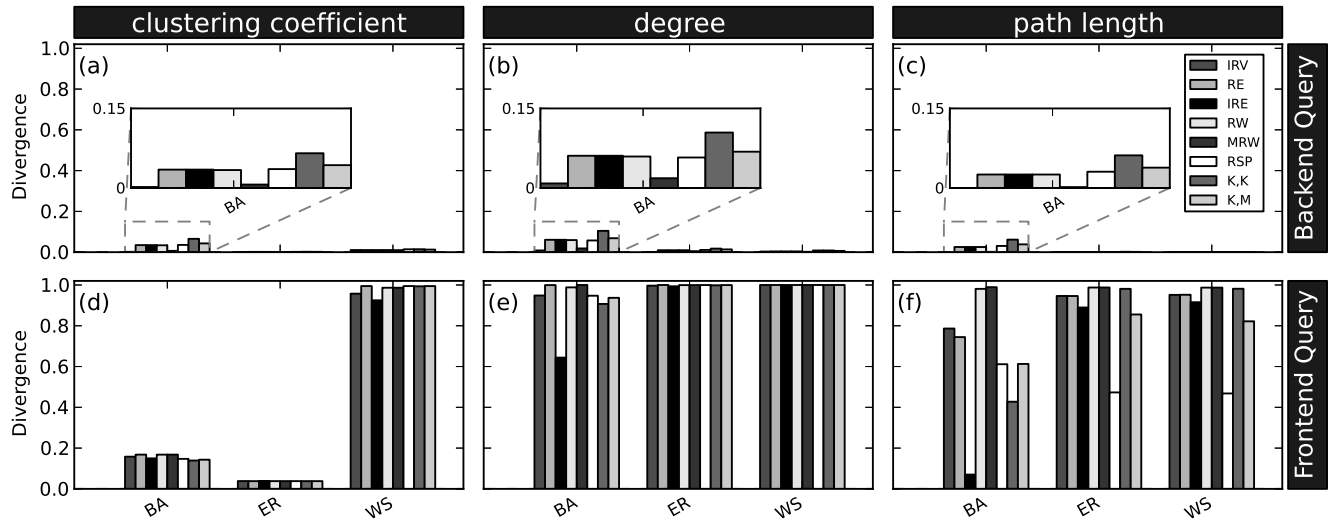


Fig. 1. Jensen-Shannon divergence scores for synthetic graphs. Backend querying results (top) outperform Frontend querying results (bottom).

considered as an approximation of RV with BE querying for degree distribution. MRW sampling with BE querying is supported in WWW domain, but it is not directly supported in OSN and ITM domains due to privacy concerns and technical limitations, respectively.

Interestingly, all sampling methods with BE querying performed worse on BA population graphs compared to ER and WS models. Analyzing the probability mass functions (Figure 2) of those population graphs shows that ER and WS models produce almost symmetric degree distributions while BA model naturally produces a power law graph. That is, BA graphs have a small number of high degree vertices along with high number of small degree vertices. RE and IRE tend to sample the high degree vertices with higher probability. Similarly, naive RW tend to visit high degree vertices with high probability. As shown in Figure 3, in BA graphs, those high degree vertices also have high betweenness centrality values. That is, the number of shortest paths passing over those high degree vertices is significantly larger than the paths passing through the vertices having small degree. Hence, path based sampling methods (RSP, K-K, and K-M) sample high degree nodes with higher probabilities as in the other sampling methods. This analysis clearly shows that the underlying graph models have significant impacts on sampling where all other parameters are held fixed.

Figure 1-e demonstrates that all sampling methods almost equally perform bad with frontend querying on all graph models for degree distribution. The divergence can be attributed to the fact that not all edges of a vertex could be sampled and included into the sample graph. As a result, in all cases degrees of vertices are under-sampled and this artifact is reflected in the divergence results. On the other hand, the sampling methods perform marginally better on BA graphs. The reason is simply BA graphs have many small degree vertices and their edges are sampled better compared to those having large degrees. As a special case in Figure 1-e IRE sampling performs better on BA graphs. In addition to the existence of many small degree vertices in BA graphs, IRE adds additional edges into the

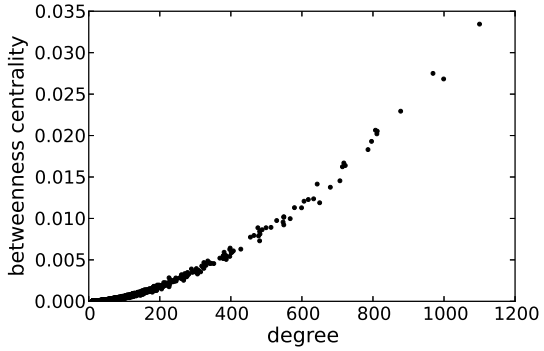
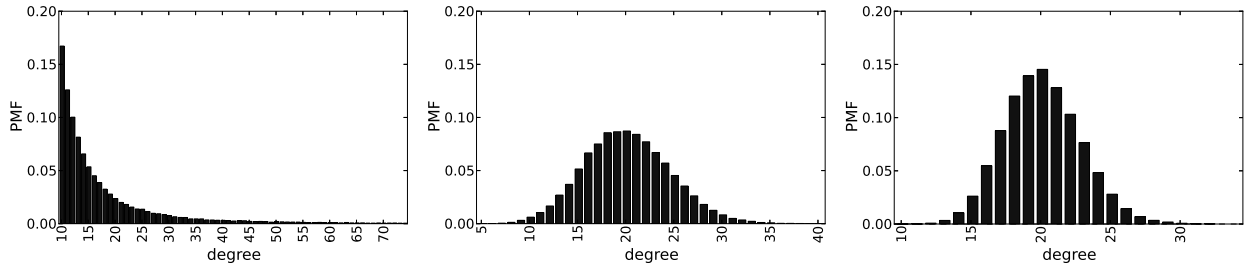


Fig. 3. Barabasi-Albert population graph. Degree vs. betweenness centrality.

sample graph if any two vertices are already sampled. Hence, the degree of a vertex is closer to the population value with IRE sampling.

2) Clustering Coefficient Distribution: Figures 1-a and 1-d show the clustering coefficient distribution divergence scores of different sampling methods on our graph models for backend and frontend querying, respectively. Both figures strongly articulate the underlying graph models' impact on performance of the utilized sampling methods.

Figure 1-a demonstrates that all sampling methods perform acceptably well with BE querying on ER and WS graphs. On the other hand, all methods but IRV and MRW relatively perform worse on BA graphs. Our further analysis shows that the difference between BA graph model and ER and WS graph models can be attributed to the relation between the degree and clustering coefficient distributions of these models. Specifically, BA graphs consist of small number of of high degree vertices and high number of small degree vertices. Furthermore, as shown in Figure 4, low degree vertices in BA graphs have higher clustering coefficient values. In fact, this outcome is quite natural because of the preferential attachment



(a) BA (b) ER (c) WS

Fig. 2. Degree probability mass functions for the population graphs.

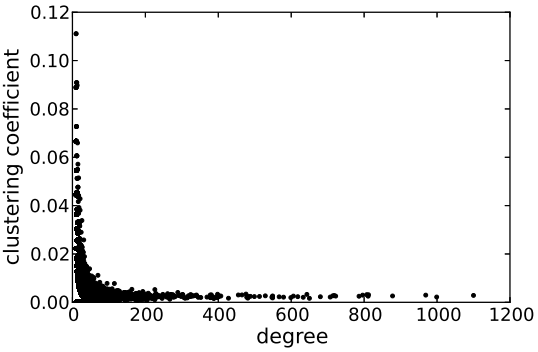


Fig. 4. Barabasi-Albert population graph. Degree vs. clustering coefficient.

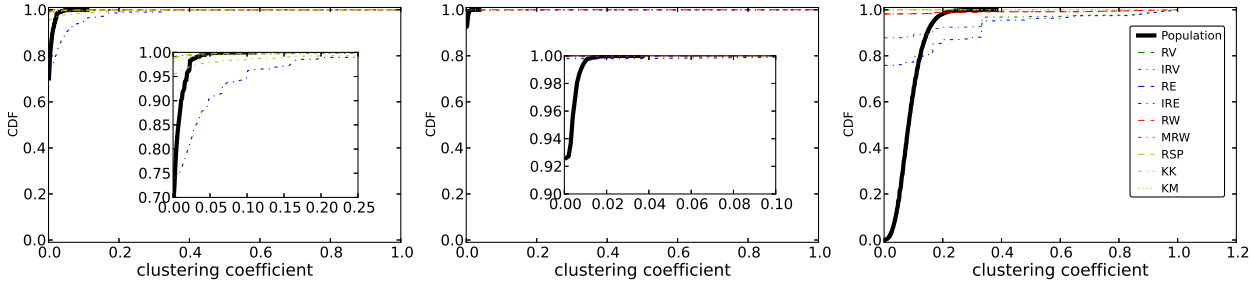
concept incorporated into the BA graph generation algorithm. That is, a newly introduced vertex tends to get connected to a higher degree vertex rather than its relatively lower degree first hop neighbors. RE and IRE methods tend to under-sample lower degree vertices because they have less edges. RSP, K-K, and K-M sampling methods also tend to under-sample low degree vertices because those vertices have lower betweenness centrality (see Figure 3). As a result, the divergence score increases for these methods because they cannot fairly sample vertices having high clustering coefficient. Note that the arguments do not apply to IRV and MRW on BA graphs because IRV samples vertices uniformly at random and MRW removes the degree bias which indirectly reduces the clustering coefficient bias.

Figure 1-d shows that FE querying performs worse than BE querying for clustering coefficient as well. This behavior may again be attributed to the vertex based sampling methods under-sampling the edges hence underestimating the clustering coefficient values in many FE querying cases as shown in Figure 5. Further analysis shows that we can explain the differences of the divergence scores among different population graph models by the existence of many vertices having clustering coefficient zero. The vertices with degree one have zero clustering coefficient by definition. Hence, if a vertex with degree one is sampled and added into the sample graph, it is sampled with all of its edges (one) and its true clustering coefficient value (zero). Moreover, since all sampling methods under-sample the edges in the population graphs, the number of vertices having zero clustering coefficient values increases in the sample graph. As a result, sampling schemes introduce vertices having artificial zero clustering coefficient. As shown in Figure 5, around 70% of the vertices in the population BA graph and 93% of the vertices in the population ER graph have zero clustering coefficient values. Because of edge under-sampling, those vertices with zero clustering coefficient values in the population graph are compensated in the sample graph hence resulting in less divergence. On the other hand, the same argument is not applicable to WS model because the clustering coefficient distribution has a larger range and variance.

3) Path Length Distribution: Figures 1-c and 1-f show the path length divergence scores of different sampling methods on our graph models for BE and FE querying, respectively.

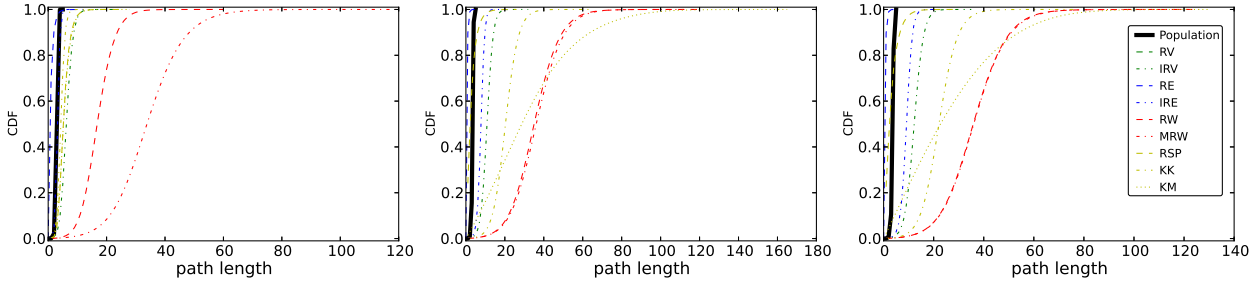
Figure 1-c shows that all sampling methods perform well with BE querying on ER and WS graphs. On the other hand, all methods but IRV and MRW relatively perform worse on BA graphs. Further analysis shows that the difference between BA graph model and ER and WS graph models can be attributed to the vertex degrees. Since BA graphs consist of small number of of high degree vertices and high number of small degree vertices; RE, IRE, and RW sampling methods tend to under-sample low degree vertices occurring at the periphery of the graph. A similar argument is applicable to RSP, K,K, and K,M sampling methods because these methods favor vertices having high betweenness centrality which in turn have high degrees. As a result, the sampled vertices are the ones mostly appearing in the core of the graph rather than periphery and they have comparably shorter path lengths in the population graph.

The divergence scores conveyed in Figure 1-e are especially surprising. All methods except RSP with frontend querying perform bad on ER and WS graphs. Figure 6 shows the cumulative densities of the population graphs as well as the sample graphs obtained by all sampling methods. One counter intuitive observation is that the lengths of observed sample paths as well as the diameter of the sample graph are much higher than the population graph except for RE sampling. The artifact simply occurs due to vertex based sampling methods under-sampling edges in the population graphs. Many of the shortest paths appearing in a population graph are not sampled because of one or more unsampled edges. As a result, path lengths between many sampled vertices in sample graphs get increased. On the other hand, further analysis shows that the reason that RE sampling behaves differently is that this sampling scheme generates highly disconnected graphs having smaller shortest paths. Finally, RSP performs better than K-K and K-M schemes because it does not have the bias due



(a) BA (b) ER (c) WS

Fig. 5. Clustering coefficient cumulative density functions for the population graphs.



(a) BA (b) ER (c) WS

Fig. 6. Path length cumulative density functions for the population graphs.

to sampling vertices closer to the preselected source and destination vertices.

BA graphs have core vertices with high degrees and high betweenness centrality. These vertices are, in general, sampled into the sample graph as well. Specific to IRE sampling method, all edges induced to the neighboring sample vertices of the core vertices are also sampled. Since these vertices carry most of the shortest paths between other vertices, the path length distribution is preserved in the sample graph. As a result, IRE with FE querying performs very well on BA graphs (see Figure 1-f).

B. Results on Real-World Networks

In this section, we present our experimental results on real world graphs shown in Table I. CAIDA, COAUTHOR, ENRON, and EPINIONS graphs can be obtained at *http://snap.stanford.edu*. The overall results are presented in Figure 8 where the rows correspond to BE and FE query types and columns correspond to degree distribution, clustering coefficient distribution, and path length distribution characteristics, respectively. The top row is for BE querying and the bottom row is for FE querying cases. Similar to the synthetic topology based experiments, our results for path length distribution with BE querying assumes that selected nodes know their path length distribution in the topology graph. In practice, this information is rather difficult to collect and therefore the results here show the best case scenario.

The bar charts in Figure 8 present the summary results for the experiments. We use several cases from Figure 8 to illustrate the effects of *design considerations* as well as some interesting observations.

1) Effect of population graph: The variations of the divergence scores across different application domains (see Figure 8-d) show that the underlying population graph has an impact on the results when the same sampling method and the same query type are utilized. This is also visible in other charts in the same figure.

Figure 8-a, 8-b, and 8-c show that, for each network, relative performances of sampling methods for all three characteristics are similar with BE querying. For instance, in EPINIONS network, for all three characteristics, IRV (and hence RV) sampling method performs the best; RE, IRE, and RW sampling methods perform the worst. This observation is also apparent for other networks. With BE querying, the performance of the sampling method depends on which vertices it selects. As long as there is a correlation between different characteristics, we expect to observe similar results. As depicted in Figure 7, as the vertex degree increases, clustering coefficient of the vertex decreases, which is very natural when the definition of clustering coefficient is considered. As the vertex degree increases, to keep the clustering coefficient the same, the number of edges between the neighbors should increase quadratically.However, this is not satisfied, especially for high degree vertices, in any real-networks that we considered. The similar performance results in Figures 8-a and 8-b can be explained in a similar way. However, this result does not imply that if a sampling design performs well for one characteristic, it will also perform well for any other characteristic [2]. Note that this trend is not observed in FE querying. In FE querying, performance of the sampling method depends on the topology of the sampled graph since no information can be extracted from the population. Therefore, observing similar results for different characteristics is not expected especially if the nature

Network Name	Appl. domain	# of nodes	# of edges	degree (min:mode:max)	average clustering coefficient	diameter
CAIDA	Internet topology	26,475	106,762	1:2:2628	0.010	17
CO-AUTHOR	Collaboration (COND-MAT)	21,363	182,628	1:5:281	0.633	15
ENRON	Email exchanges	33,696	361,622	1:3:1383	0.497	12
EPINIONS	Social(Trust)	75,877	508,836	1:2:3044	0.228	13
FACEBOOK [29]	Social(Friendship)	63,392	816,886	1:10:1098	0.279	15
GNUTELLA	P2P	65,561	147,878	1:2:95	0.010	11
YEAST-PPI [14]	Biological	1,458	1,993	1:2:56	0.070	19

TABLE I. DETAILS OF REAL-WORLD NETWORKS USED IN EXPERIMENTS

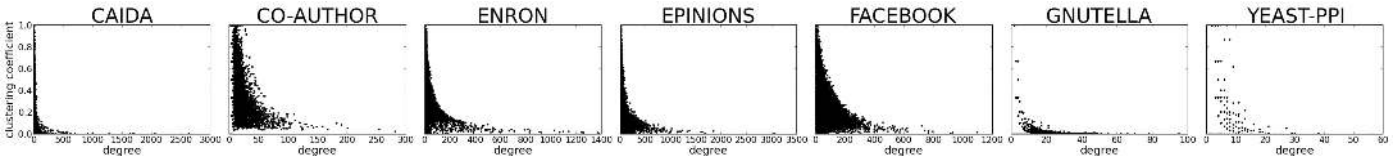


Fig. 7. Scatter plots of degree vs clustering coefficient.

of the characteristics are different.

2) Effect of characteristic: The mismatch between the characteristic of interest and the sampling design will affect the results of a sampling study. For instance, walk based sampling methods perform poorly for the estimation of path length distribution with frontend querying (see Figure 8-f). This is because they mostly create sparse subgraphs by under-sampling edges causing overestimation of shortest path lengths. However, in Figure 8-f we observe a different behavior for YEAST-PPI network. This difference can be attributed to the sparsity of the population graph. Moreover, by further analysis, we noticed that the topology of the YEAST-PPI network has lots of chain structure, which fits to the nature of the walk-based designs since the walk-based designs select only one edge of a vertex at each visit and create sparse sample graphs.

Divergence scores of MRW sampling for GNUTELLA network in Figures 8-d, 8-e, and 8-f are 0.031, 0.236, and 0.950 for clustering coefficient, degree and path length distributions, respectively. This result shows that the characteristic of interest is important for the performance of the utilized sampling design. By analyzing the clustering coefficient distribution of GNUTELLA network, which is not shown here due to space limitations, we observed that more than 90% of vertices have zero clustering coefficient similar to the ER graph. Since the MRW sampling method under-samples the edges, it introduces vertices having artificial zero clustering coefficient. The vertices with zero clustering coefficient values in the population graph are compensated in the sample graph by those vertices hence resulting in less divergence. High divergence score for the path length distribution can be attributed to the mismatch between the sampling design and the characteristic of interest.

3) Effect of query type: Figure 8 depicts that the BE querying (top row) results are significantly different than FE querying results (bottom row) in most cases. For instance, in Figures 8-c and 8-f, for YEAST-PPI network, even though IRV and RE sampling methods perform better than path-based sampling methods under BE querying, the situation is the reverse for FE querying. This result shows that query type also has an impact on the performance of the sampling design.

In addition to above results, we also observe that in some cases FE querying performs better than BE querying, which we did not observe in synthetic graph analysis. For instance in Figures 8-b and 8-e, for CAIDA and YEAST-PPI networks, degree distribution results with FE querying outperform the degree distribution results with BE querying in walk-based sampling methods. We believe that this is a quite interesting result. Existence of such cases may motivate researchers to do further analysis to understand the interaction between the sampling design, the characteristic of interest, and the topology of the network. Understanding this interaction may give clues about the feasibility of making inferences about a given population characteristic using the sample graph.

V. FUTURE WORK & CONCLUSIONS

In this study we demonstrate that the design of a sampling scheme plays an important role in estimating various network characteristics. Our empirical evaluations clearly shows that network characteristic estimation problem depends on many factors and should be handled individually per characteristic and per sampling scheme supported by the underlying system. Even though we consider important factors of sampling biases in this paper, they may not be exhaustive. We plan to conduct a theoretical study supporting our findings in this empirical study. Furthermore, we plan to develop working instances of our proposed sampling framework and device new statistical estimators for different features of real world complex systems including OSN, ITM, P2P, and WWW.

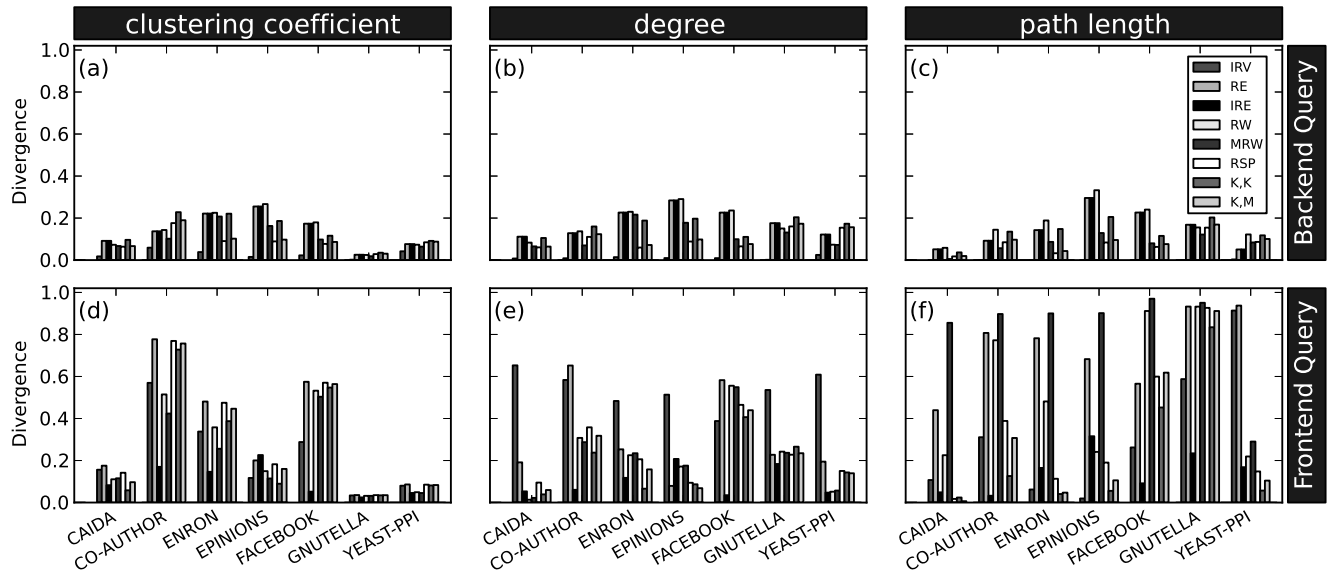


Fig. 8. Jensen-Shannon divergence scores for real-world graphs. The variability in the performance of sampling designs depicts the effects of following *design considerations* : (a) the characteristic of interest under study, (b) topology of the population graph, and (c) available querying type.

REFERENCES

[1] D. Achlioptas, A. Clauset, D. Kempe, and C. Moore. On the bias of traceroute sampling: Or, power-law degree distributions in regular graphs. *J. ACM*, 56(4):21:1–21:28, July 2009.

[2] N. Ahmed, J. Neville, and R. Kompella. Reconsidering the foundations of network sampling. In *Proc. of the 2nd Workshop on Information in Networks*, New York, September 2010".

[3] A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, October 1999.

[4] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Computer Networks*, 33(16):309 – 320, 2000.

[5] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law Distributions in Empirical Data. *SIAM Review*, 51(4):661–703, 2009.

[6] L. Dall'Asta, J. I. Alvarez-Hamelin, A. Barrat, A. Vázquez, and A. Vespignani. Exploring networks with traceroute-like probes: Theory and simulations. *Theor. Comput. Sci.*, 355(1):6–24, 2006.

[7] P. Erdős and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci*, 5:17–61, 1960.

[8] O. Frank. Estimation of graph totals. *Scandinavian Journal of Statistics*, 4(2):81–89, 1977.

[9] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou. Walking in facebook: A case study of unbiased sampling of osns. In *INFOCOM, 2010 Proc. IEEE*, pages 1–9. IEEE, 2010.

[10] M. Gjoka, M. Kurant, and A. Markopoulou. 2.5k-graphs: from sampling to generation. In *Proc. of INFOCOM*, Turin, Italy, April 2013.

[11] C. Gkantsidis, M. Mihail, and A. Saberi. Random walks in peer-to-peer networks. In *Proc. of IEEE INFOCOM*, Hong Kong, China, 2004.

[12] J.-D. Han, D. Dupuy, M. Bertin, N. ahd Cusick, and M. Vidal. Effect of sampling in topology predictions of pritein-protein interaction networks. *Nature Biotechnology*, 23(7):839–844, July 2005.

[13] C. Hubler, H.-P. Kriegel, K. Borgwardt, and Z. Ghahramani. Metropolis algorithms for representative subgraph sampling. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 283–292. IEEE, 2008.

[14] H. Jeong, S. P. Mason, A.-L. Barabasi, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, May 2001.

[15] V. Krishnamurthy, M. Faloutsos, M. Chrobak, J.-H. Cui, L. Lao, and A. G. Percus. Sampling large internet topologies for simulation purposes. *Computer Networks*, 51(15):4284–4302, 2007.

[16] M. Kurant, M. Gjoka, C. Butts, and A. Markopoulou. Walking on a graph with a magnifying glass: Startified sampling via weighted random walks. In *Proc.s of SIGMETRICS*, San Jose, CA, USA, June 2011.

[17] M. Kurant, M. Gjoka, Y. Wang, Z. Almquist, C. Butts, and A. Markopoulou. Coarse-grained topology estimation via graph sampling. In *Workshop on OSNs*, Helsinki, Norway, August 2012.

[18] S. H. Lee, P.-J. Kim, and H. Jeong. Statistical properties of sampled networks. *Physical Review E*, 73:016102/1–016102/7, November 2006.

[19] L. Lovasz. Random walks on graphs: A survey. *Combinatorics (Paul Erdos is Eighty)*, 2(1):1–46, 1993.

[20] A. S. Maiya and T. Y. Berger-Wolf. Benefits of bias: Towards better characterization of network sampling. In *Proc. of KDD*, San Diego, CA, USA, August 2011.

[21] A. Rasti, M. Torkjazi, R. Rejaie, N. Duffield, W. Willinger, and D. Stutzbach. Respondent-driven sampling for characterizing unstructured overlays. In *Proc. of IEEE INFOCOM Mini-conference*, Rio de Janeiro, Brazil, April 2009.

[22] B. Riberio and D. Towsley. Estimating and sampling graphs with multidimensional random walks. In *ACM IMC*, Melbourne, Australia, November 2010.

[23] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vision*, 40(2):99–121, Nov. 2000.

[24] M. P. H. Stumpf and C. Wiuf. Sampling properties of random graphs: The degree distribution. *Phys. Rev. E*, 72:036118, Sep 2005.

[25] D. Stutzbach, R. Rejaie, N. Duffield, S. Sen, and W. Willinger. On unbiased sampling for unstructured peer-to-peer networks. *IEEE/ACM Transactions on Networking*, 16(6):377–390, April 2008.

[26] M. E. Tozal and K. Sarac. TraceNET: An Internet Topology Data Collector. In *Proceedings of ACM Internet Measurement Conference*, Melbourne, Australia, Nov 2010.

[27] M. E. Tozal and K. Sarac. Subnet level network topology mapping. In *IEEE IPCCC International Performance Computing and Communications Conference*, Orlando, FL, USA, November 2011.

[28] F. Viger, A. Barrat, L. Dall'Asta, C.-H. Zhang, and E. D. Kolaczyk. What is the real size of a sampled network? the case of the internet. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 75(5):056111, 2007.

[29] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi. On the evolution of user interaction in facebook. In *Proc. of the 2nd ACM SIGCOMM Workshop on Social Networks (WOSN'09)*, August 2009.

[30] D. J. Watts and S. H. Strogatz. Collective dynamics of small-worldnetworks. *nature*, 393(6684):440–442, 1998.