Examining Shortest-Path Distances to Samples in Graphs

Alan Zhu,[1] **Jiaqi Ma,** [2] **Qiaozhu Mei** [3]
[1] Carnegie Mellon University
[2] University of Illinois Urbana-Champaign
[3] University of Michigan

Abstract

With very large graph data becoming commonplace across many fields, sampling is often needed to reduce the graphs into practical sizes. This procedure raises critical questions about representativeness as a sample cannot capture the properties of the original graph perfectly, and different parts of the graph are not equally affected by the loss. Recent work has shown that the distance between other nodes to the sampled nodes can be a quantitative indicator of bias and fairness issues in graph machine learning contexts. In this paper, we present a systematic analysis of popular node sampling methods on different types of synthetic graphs and real-world graphs from the perspective of shortest-path distances to the sample. Additionally, we propose a theoretical framework for estimating the distribution of shortest-path distances to the sample. We examine how the choice of sampling method and graph configuration affects the distribution of the shortest-path distances and demonstrate the accuracy of our theoretical framework. This study makes an early step towards understanding the behavior of graph sampling methods.

Introduction

Graph sampling is commonly used in a variety of fields that are concerned with large scale graph data (e.g., the Web graph or a real-world social network). In many cases, working with such a full graph or even knowing the full graph is infeasible, and one has to effectively reduce the size of the graph (often through sampling the nodes) so that sophisticated data science methods can be practically deployed. For example, Web crawlers travel the graph of Web pages to model the Internet [3], epidemiologists sample a population to model the spread of diseases [6], and biologists sample cell interactions to understand behavior of cellular networks [1]. Sampling can also be used to simplify complex graph structures that would otherwise be difficult to work with [16]. In particular, when dealing with large social networks, sampling reduces the size and structural complexity of the graph and allows for easier analysis [31].

However, graph sampling is not as simple as randomly selecting nodes (though this can be a valid approach). An ideal sample would be tailored to the research question at hand, and it should avoid any unwanted bias. For example, sampling hard-to-reach communities poses unique challenges in that those who are successfully sampled may not be representative of the larger population [25]. In addition, samples would ideally retain relevant properties of the original graph, which is not always straightforward to generate [28]. Careless sampling is also prone to unwanted bias [21], especially in more complex graphs where the importance of nodes is not necessarily equal [29]. Carefully choosing a sampling method can help address these representation concerns.

We aim to better understand the representativeness of graph sampling methods from the perspective of the distribution of shortest-path distances (DSPD) from nodes to a sample. The DSPD provides valuable insights into how well the sample covers the graph, a measure of representativeness known as network reach [21]. Additionally, in graph neural networks [27], distance from sample has been shown to be an indicator of fairness [20], as the machine learning models perform better on nodes closer to the labeled nodes than on those further away. When dealing with hard-to-reach communities, models perform worse on those nodes further away from the sample [11]. However, determining the DSPD is difficult: the naive method of performing sampling then empirically calculating the shortest-path distances can be computationally expensive.

This paper provides, to our knowledge, the first systematic study of the impact of sampling methods on the DSPD as well as a framework for theoretically estimating the DSPD to a sample. We demonstrate the effectiveness of our method through comparison against empirically calculated distributions on synthetic graphs. In particular, we show that our method is able to accurately make predictions about which of two sampling methods would yield a shortest-path distance distribution that more heavily favors smaller distances. Additionally, we verify the differences we find between sampling methods on real-world graphs.

Related Work

Graph Sampling

Graph sampling is a common technique used in cases where working with a full graph is impractical or impossible. These cases can arise when collecting data from the full network or working with the full network is computationally infeasible,

or in social science contexts when working with a hard-to-reach communities, to name a few. Maiya and Berger-Wolf [21] discusses how well a sample covers a graph through *network reach*, measuring it through *discovery quotient*, which is the proportion of nodes that are in the sample or is a neighbor of a sample. However, we examine a more general idea via the full distribution of distances.

Hu and Lau [13] conduct a survey of graph sampling techniques and provide ad hoc analysis of the impacts of the sampling techniques. Costenbader and Valente [7] look at the stability of centrality metrics under sampling, including the centrality metrics we use for centrality-based sampling (degree and betweenness). Stumpf, Wiuf, and May [28] conclude that the scale-free property of a graph would not be preserved under random sampling. These studies all look at whether certain properties of the original graph would be preserved by the sampled graph. This paper instead examines the representativeness of nodes in a sample using the DSPD of other nodes to that sample. This property is not intrinsic to the original graph, but depends on both the graph and the sample, and therefore is not something to be preserved "before and after" sampling.

Shortest Path Distance

Classical network science literature measures the average of shortest path distances in a graph and uses it to categorize networks [32]. A few existing studies look at the DSPD beyond the mean, including Katzav, Biham, and Hartmann [14], Ventrella, Piro, and Grieco [30], and Katzav et al. [15]. Katzav, Biham, and Hartmann [14] obtain an analytical expression for the DSPD between nodes in subcritical Erdös-Rényi graphs, while Katzav et al. [15] do so for Erdös-Rényi graphs in general. Ventrella, Piro, and Grieco [30] propose models that can be used to find the distribution for scale-free networks. These studies do not consider network sampling and only investigate the properties of the full graph. As such, the distributions investigated are sample independent, which is not what we track.

Ma, Deng, and Mei [20] find that the distance of an unlabeled node to the subgroup of labeled nodes in a graph is a good indicator of the performance of a graph neural network on that node, which motivates our investigation of the DSPD. However, their work does not conduct analysis of the effect of sampling methods.

This paper measures the DSPD *to a given sample of nodes* such that we are only interested in the distances of nodes outside a selected sample to any node within that sample.

Methodology

We conduct a systematic investigation of the DSPD from nodes to samples by various sampling techniques in various types of graphs. We use common graph generating techniques to obtain synthetic graphs, and examine some real world graphs. For each graph, we apply common sampling techniques, examine the resulting distribution, and compare against the distribution predicted by our theoretical framework.

Overview of Sampling Methods

Random Sampling Random sampling is a straightforward way to sample from a graph. To generate a sample of size n, we select n nodes from the graph uniformly at random, without replacement.

Snowball Sampling Snowball sampling [8] is a form of traversal-based sampling commonly used in sociology when sampling on hard-to-reach communities [13], where random sampling is not feasible. In snowball sampling, we begin with a seed sample, then produce a frontier of all nodes adjacent to the seed. At each step, a node in the frontier is removed from the frontier and added to the sample with a certain probability. If the node is added, it is removed from the frontier and the frontier is expanded with that node's neighbors. This simulates the process of using referrals to find new subjects, which may be necessary if the population being sampled is stigmatized and standard surveys are ineffective.

Centrality-Based Sampling Methods We also investigate centrality-based sampling methods. These methods select high-centrality nodes into the sample, which are considered nodes with high importance/influence [26].

For example, degree centrality sampling takes the n nodes with largest degree as the sample. Degree sampling is often used as a metric for importance in a network, so selecting high degree nodes for a sample is a potential way to select the most important/influential nodes for predictions. [5] found that in social networks, a small sample of high degree users disseminating true information could effectively counteract false information.

Betweenness centrality [4] sampling takes the n nodes with largest betweenness metric. Betweenness of a node v is measured as the sum of the fraction of all shortest paths between s and t that pass through v, for all s and t. A larger value again indicates a more central node. Similar to other centrality measures, this can be used as a metric of importance, as [19] did to examine interdisciplinarity of scientific journals.

Overview of Graphs

We applied the aforementioned sampling methods on both synthetic and real world graphs for our study. Synthetic graphs can be obtained more easily than real world graphs, and the regularity of their degree distributions makes it simpler to apply our theoretical framework. However, they may not always be able to capture all the intricacies of real world networks. To better compare between synthetic and real-world graphs, we examine synthetic graphs with and without community structures. All synthetic graphs were of size $n = 2000$. We examined three graph generation techniques: binomial graphs, power-law graphs [2], and Stochastic Block Model (SBM) graphs [12].

Binomial Graphs Binomial graphs are graphs with n nodes, where every edge has a chance of being in the graph, independent of every other edge. For a given probability p, the expected number of edges is then $\binom{n}{2}p$. We generated

binomial graphs with $n = 2000$ nodes and $p = 0.005$ probability of an edge. These graphs are simple and intuitive to analyze compared to other types of graphs.

Power-Law Graphs The power-law model is a way to generate scale free graphs, where the degree distribution of nodes follows a power law distribution. Denoting the probability of a node having degree k as $p(k)$, $p(k) = ak^{\gamma}$ where γ is a parameter chosen and a is a normalizing constant. Minimums and maximums can be enforced on the support of p. We generated two types of power-law graphs. Both were of size $n = 2000$, with $\gamma = -2.5$; one had a support of $3 \leq k \leq 29$ and the other had a support of $13 \leq k \leq 37$. We will refer to the first configuration as Power Law A and the second configuration as Power Law B. These configurations were chosen as they lead to different relative results between sampling methods. Many social networks are scale free, making power-law graphs a potentially good way to replicate real world graphs.

Stochastic Block Model The Stochastic Block Model (SBM) is another way to generate graphs resembling real world social networks. The graph is split into blocks with each edge within a block appearing with probability p_1 and each edge across two blocks appearing with probability p_2. We generated SBM graphs with 4 blocks of 500 nodes. The inside-block edge probability was 0.02 while the outside-block edge probability was 0.002. Unlike the power-law model, these graphs aim to replicate the community structure, where there are multiple communities that are highly connected within themselves, with sparse connections between communities.

We are able to carefully control the properties of synthetic graphs, allowing us to more easily make conclusions; finding enough real-world graphs that satisfy the properties we need to be able to make strong conclusions would be prohibitively difficult. We take advantage of this fact when choosing the two configurations of the power law graphs. Additionally, the diversity of synthetic graphs, mimicking different types of real world graphs, allows us to generalize results seen in synthetic graphs.

Real World Graphs We examined four real world graphs from the Stanford Network Analysis Project (SNAP) [17]. These were the Facebook Large Page-Page Network dataset [24], and the California, Texas, and Pennsylvania Road Networks [18].

The Facebook graph contains 22,470 nodes representing official pages on Facebook, with 171,002 undirected edges representing mutual likes between pages.

In the road network graphs, nodes represent intersections between and endpoints of roads, with undirected edges representing the road segments connecting these intersections or endpoints. The networks were very large, with California's graph having nearly 2 million nodes and nearly 3 million edges, Texas's graph having around 1.3 million nodes and nearly 2 million edges, and Pennsylvania's graph having just over 1 million edges and around 1.5 million edges. To reduce the size of the graphs for easier analysis, we took a connected subgraph of 50,000 nodes, generated by randomly picking a root point, and adding recursive adding all neighbors of nodes already selected until we reached 50,000 nodes. This process was repeated 10 times, each time starting with a different random node.

Theoretical Framework

Next, we present a framework to theoretically examine the DSPD resulting from different sampling methods on different graphs.

Preliminaries

As preliminaries, we begin with a simpler question: the DSPD to a *single random node* in a graph, the single node DSPD problem. In this setting, this is equivalent to determining the DSPD in the entire graph. Nitzan et al. [23] derive such a distribution analytically for configuration model graphs [22], where the distribution of the degrees of nodes is specified. For example, a binomial graph with n nodes and probability p may be viewed as a configuration model graph where the distribution of degrees follows a binomial distribution $B(n-1, p)$.

Intuitively, the distribution is obtained by looking at "shells" around a single node recursively. The probability that the shortest path distance between two random points is greater than ℓ can be calculated recursively: for the shortest-path distance between nodes i and j to be greater than ℓ, the shortest-path distance between any neighbor of i and j must be greater than $\ell - 1$. If the probability that the distance between any two nodes is greater than $\ell - 1$ is known, then by accounting for the distribution of the number of neighbors of i, the probability that the distance between any i and j is greater than ℓ can be determined. An additional complication is that the degree distribution of a randomly selected node is different from the degree distribution of a node known to have a neighbor, and this is resolved by maintaining two recursive equations.

Specifically, for graphs with N nodes,

$$P(d > \ell) = P(d > 0) \prod_{\ell'=1}^{\ell} m_{N,\ell'},$$

where $P(d > \ell)$ is the probability of a pair of nodes i, j having shortest-path distance greater than ℓ, and $m_{N,\ell}$ is defined as, in a graph with N nodes, $P(d > \ell | d > \ell - 1)$.

First,

$$m_{N,\ell} = \sum_{k=1}^{N-2} p(k)(\tilde{m}_{N-1,\ell-1})^k,$$

where $p(k)$ is the degree distribution and $\tilde{m}_{N-1,\ell-1}$ is the probability that for a pair of nodes i and j, node r a neighbor of i, in a graph of size $N-1$ (i.e., excluding i), the shortest-path distance from r to j is greater than ℓ given that the distance is greater than $\ell - 1$.

Next,

$$\tilde{m}_{N,\ell} = \sum_{k=1}^{N-2} \frac{k}{c} p(k)(\tilde{m}_{N-1,\ell-1})^{k-1},$$

where $c = \sum_{k=1}^{\infty} kp(k)$ is a normalizing constant. Here the distribution $\frac{k}{c}p(k)$ is used as the degree distribution of i's neighbor r is not drawn from $p(k)$: the probability of a node being a neighbor of i is proportional to its degree. The exponent is $k-1$ as one of r's neighbors is i.

The base cases are

$$m_{N,1} = \sum_{k=1}^{N-1} p(k) \left(1 - \frac{1}{N-1}\right)^k,$$

and

$$\tilde{m}_{N,1} = \sum_{k=1}^{N-1} \frac{k}{c} p(k) \left(1 - \frac{1}{N-1}\right)^{k-1}.$$

Multi-Node Sample Framework

This analysis of the DSPD to a collection of sampled nodes is considerably more challenging than the single node DSPD problem presented above. Simply sampling multiple distances from the single node DSPD does not suffice as the distances are not independent. Additionally, our framework must account for sample nodes not selected randomly, while the single node DSPD framework requires the source node to be selected randomly.

We now describe how we overcame these challenges. We retain the idea of looking at shells centered around a single node, except the center of the shell is a *supernode* consisting of all the nodes in the sample contracted into one. More formally, let an (undirected) graph be $G = (V, E)$, where V is the set of nodes and $E \subseteq V \times V$ is the set of edges. Denote a set of sample nodes as $S \subseteq V$, all the edges of G involving a node in S as $E_S \subseteq E$. Let $N := \{v \in V \setminus S \mid (u, v) \in E\}$, which is the set of neighbors of nodes in S that are outside S. Then the graph (which we call a *contracted graph*) after contracting the sample S into a supernode u_S is $G' = (V', E')$, where

$$V' := V \setminus S \cup \{u_S\}, E' := E \setminus E_S \cup \{(u_S, v) \mid v \in N\}).$$

We then need to adjust the formula for $m_{N,\ell}$. In particular, instead of the degree distribution $p(k)$ in the original graph G, we draw from the degree distribution of the supernode u_S in the contracted graph G', denoted as $p_S(k)$. Thus,

$$P(d > \ell) = P(d > 0) \prod_{\ell'=1}^{\ell} m_{N,\ell'},$$

as before. However, the recursions are now

$$m_{N,\ell} = \sum_{k=1}^{N-2} p_S(k)(\tilde{m}_{N-1,\ell-1})^k,$$

and

$$\tilde{m}_{N,\ell} = \sum_{k=1}^{N-2} \frac{k}{c} p(k)(\tilde{m}_{N-1,\ell-1})^{k-1}.$$

The base cases are

$$m_{N,1} = \sum_{k=1}^{N-1} p_S(k) \left(1 - \frac{1}{N-1}\right)^k,$$

and

$$\tilde{m}_{N,1} = \sum_{k=1}^{N-1} \frac{k}{c} p(k) \left(1 - \frac{1}{N-1}\right)^{k-1}.$$

Note that p_S is affected by both the structure of the original graph G and the sampling method used to draw S. In the following sections we discuss how the sampling methods and graph structures we use affect this distribution.

Sampling Methods

Random Sampling To determine the degree distribution of the sample supernode for a random sample of size n, we need to determine the distribution of a sum of n independent random variables drawn from the single node degree distribution. Assuming $p(k)$ is the probability of a single node having degree k in the original graph G, then the degree distribution of the supernode u_S in the contracted graph G' is

$$p_S(k) = \sum_{\substack{\sum_{i=0}^{N-1} n_i = n \\ \sum_{i=0}^{N-1} i n_i = k}} \binom{n}{n_1, n_2, \cdots, n_{N-1}} \prod_{i=0}^{N-1} p(i)^{n_i}.$$

This estimation is accurate under the assumption that there are no edges between nodes in the sample, i.e. that all degrees counted lead to a node not in the sample. In cases where the sample size is small compared to the overall size of the graph, this assumption generally holds. Additionally, we may zero-out highly unlikely single-node degrees (i.e., round $p(i)$ to 0) to reduce computation expense. Not only does this avoid calculating terms that will not influence the final distribution much, this also allows us to calculate p_S using polynomial multiplication: representing p as a polynomial $q(x)$ where the coefficient of x^i is $p(i)$, $q(x)^n$ is a polynomial where the coefficient of x^i is $p_S(i)$.

Snowball Sampling To estimate the degree distribution of the sample supernode, p_S, under snowball sampling, similar to the method for random sampling, we still determine the distribution of a sum of random variables. However, these random variables are drawn from the single node degree distribution, p, weighted by the degree. This represents the fact that the probability of a node being selected via the snowball process is proportional to the degree of that node. However, we also apply a correction of $-2n$ for a sample of size n, representing the fact that a sample of size n will have about n edges between nodes in the sample.

Recall that $p(k)$ is the probability of a single node having degree k in the original graph G. Assuming $c = \sum_{i=0}^{N-1} ip(i)$ and a sample S has size n, then the degree distribution of the supernode u_S in the contracted graph G' is

$$p_S(k) = \sum_{\substack{\sum_{i=0}^{N-1} n_i = n \\ \sum_{i=0}^{N-1} i n_i = k+2n}} \binom{n}{n_1, n_2, \cdots, n_{N-1}} \prod_{i=0}^{N-1} \left(\frac{ip(i)}{c}\right)^{n_i}.$$

As with random sampling, this estimation is accurate under the assumption that there are no edges between nodes in the sample, other than the edges traversed during the snowball process. Similarly, we may zero-out highly unlikely single-node degrees and use polynomial multiplication to reduce computation cost. Representing the normalized weighted p as a polynomial $q(x)$ where the coefficient of x^i is $\frac{ip(i)}{c}$, $q(x)^n$ is a polynomial where the coefficient of x^i is $p_S(i-2n)$.

An additional assumption this estimation makes is that the number of seed nodes is small: seed nodes would have their degree distribution come from $p(k)$, not $\frac{ip(i)}{c}$. In general, this assumption holds: many more nodes are added via the snowball process in comparison to seed nodes.

Centrality-Based Sampling Methods Determining the degree distribution of nodes sampled with centrality based methods appears challenging due in large part to the difficulty in estimating the degree distribution of the nodes selected. In this paper, we only examine these methods empirically and leave theoretical analysis to future work.

Graphs

Here we discuss the degree distribution $p(k)$ for the synthetic graphs described earlier. For binomial graphs, the degree distribution of a single node simply follows a binomial distribution: for a graph with n nodes and probability of an edge p, $p(k) = \binom{n}{k}p^k(1-p)^{n-k}$. Power-law graphs are generated according to $p(k) = ak^{-\gamma}$, with the normalizing constant a described earlier.

However, applying our framework to SBM graphs is more difficult than binomial or power-law graphs as our original framework is built off of the configuration model, which treats all edges and nodes equally: when moving from one shell to the next, every edge leads to a new node with identical degree distribution. This is no longer true for SBM graphs due to the block structure. We discuss how we overcome this challenge below.

Stochastic Block Model Graphs Recall that a critical observation for our framework is that the degree distribution of a node which is a neighbor of another node is not the same as the degree distribution of a random node; neighbors are more likely to have large degree. However, for SBM graphs, there are two degree distributions we must consider: the inside-block degree distribution and the outside-block degree distribution. We thus must consider whether a neighbor is within the same block or across blocks. For example, if a node is known to be a neighbor via a inside-block edge, then we would expect its inside-block degree distribution to be different from that of a randomly selected node, but its outside-block degree distribution would be the same. Similarly, if a node is known to be a neighbor via an outside-block edge, then we would expect its outside-block degree distribution to be different from that of a randomly selected node, but its inside-block degree distribution would be the same.

To account for this, we must modify the recursive formulas. For an SBM graph G, denote the p_i as the inside-edge degree distribution and p_o as the outside-edge degree distribution.

As before,

$$P(d > \ell) = P(d > 0) \prod_{\ell'=1}^{\ell} m_{N,\ell'}.$$

However, now

$$m_{N,\ell} = \sum_{k_i+k_o<N-1} p_{si}(k_i)p_{so}(k_o)\cdot \left(\tilde{m}^i_{N-1,\ell-1}\right)^{k_i}\left(\tilde{m}^o_{N-1,\ell-1}\right)^{k_o},$$

where p_{si} is the sample supernode degree distribution for inside-block edges in G', p_{so} is the sample supernode degree distribution for outside-block edges in G', $\tilde{m}^i_{N-1,\ell-1}$ is the probability that a node in the next shell reached via an inside-block edge is greater than ℓ away from the node given that it is greater than $\ell-1$ away, and $\tilde{m}^o_{N-1,\ell-1}$ is the probability that a node in the next shell reached via an outside-block edge is greater than ℓ away from the node given that it is greater than $\ell-1$ away. The recursions for $\tilde{m}^i$ and $\tilde{m}^o$ are

$$\tilde{m}^i_{N,\ell} = \sum_{k_i+k_o<N-1} \frac{k_i p_i(k_i)}{c_i} p_o(k_o)\cdot \left(\tilde{m}^i_{N-1,\ell-1}\right)^{k_i-1}\left(\tilde{m}^o_{N-1,\ell-1}\right)^{k_o},$$

$$\tilde{m}^o_{N,\ell} = \sum_{k_i+k_o<N-1} p_i(k_i)\frac{k_o p_o(k_o)}{c_o}\cdot \left(\tilde{m}^i_{N-1,\ell-1}\right)^{k_i}\left(\tilde{m}^o_{N-1,\ell-1}\right)^{k_o-1},$$

with c_i and c_o being $\sum_{0<k<N-1} kp_i(k)$ and $\sum_{0<k<N-1} kp_i o(k)$, the normalizing constants after weighting p_i and p_o by the degree. Depending on whether the node in the next shell was arrived at via an inside-block edge or outside-block edge, the degree distribution for the matching class of edge is drawn from the weighted degree distribution rather than the unweighted degree distribution. The base cases are then

$$m_{N,1} = \sum_{k_i+k_o<N} p_{si}(k_i)p_{so}(k_o)\left(1-\frac{1}{N-1}\right)^{k_i+k_o},$$

$$\tilde{m}^i_{N,1} = \sum_{k_i+k_o<N} \frac{k_i p_i(k_i)}{c_i} p_o(k_o)\left(1-\frac{1}{N-1}\right)^{k_i+k_o-1},$$

$$\tilde{m}^o_{N,1} = \sum_{k_i+k_o<N} p_i(k_i)\frac{k_o p_o(k_o)}{c_o}\left(1-\frac{1}{N-1}\right)^{k_i+k_o-1}.$$

Determining p_{si} and p_{so} under random sampling is straightforward: we need only apply the random sampling process for other graph types to p_i and p_o separately. It is more difficult to determine the distributions for snowball sampling as we must consider the distribution of the type of edges the snowball process takes. The first step is to determine the probability that a randomly selected node reached

during the snowball process was arrived at via an inside-block edge. Denote this probability p. We may solve for p:

$$\begin{aligned}
&P(\text{node via inside-edge}) \\
&= P(\text{node via inside-edge } | \text{parent node via inside-edge}) \\
&\quad P(\text{parent node via inside-edge}) \\
&+ P(\text{node via inside-edge } | \text{parent node via outside-edge}) \\
&\quad P(\text{parent node via outside-edge}),
\end{aligned}$$

with the observation that $P(\text{parent node via inside-edge}) = p$ and $P(\text{parent node via outside-edge}) = 1 - p$. To determine the conditional probabilities, we have

$$P(\text{node via inside-edge} \mid \text{parent node via inside-edge}) = \sum_{0<k_i<N-1} \sum_{0<k_o<N-1} \frac{k_i p_i(k_i)}{c_i} p_o(k_o) \frac{k_i - 1}{k_i + k_o - 1},$$

$$P(\text{node via inside-edge} \mid \text{parent node via outside-edge}) = \sum_{0<k_i<N-1} \sum_{0<k_o<N-1} p_i(k_i) \frac{k_o p_o(k_o)}{c_o} \frac{k_i}{k_i + k_o - 1},$$

from which we may solve for p.

Next, to determine the distribution of the number of inside/outside-block neighbors of the supernode, we need to account for those inside/outside-block neighbors that are part of the snowball process. For example, for a sample size of s of which s_i are reached via inside-block edges, the distribution of the number of effective inside-block neighbors is the distribution of a sum of values drawn from $\frac{k_i p_i}{c_i}$ s_i times (as s_i of the nodes were reached via an inside-block edge) and p_i $s - s_i$ times (as the remaining nodes were reached via an outside-block edge), minus $2i$ (to remove those edges which are inside the sample as part of the snowball process). Efficiently calculate this using the same polynomial multiplication idea as before: let the distribution p_i be represented by $q(x)$ and the normalized weighted p_i be represented by $r(x)$. Then for each s_i we calculate $u_{s_i}(x) = q(x)^{s_i} r(x)^{s-s_i} / x^{2s_i}$.

Finally, over all possible values of s_i, we take a sum of the resulting distributions, weighted by $\binom{s}{s_i} p_i^s (1-p)^{s-s_i}$. The resulting distribution approximates the distribution for the number of inside-block neighbors of the sample. Continuing the polynomial multiplication method from the previous paragraph, we obtain the final polynomial of $\sum_{0<s_i<s} \binom{s}{s_i} p_i^s (1-p)^{s-s_i} u_{s_i}(x)$ A similar process is done for outside-block neighbors.

One important assumption is that the degree distribution of a node in the sample is not seriously affected by conditioning on the number of inside-block edges in the sample. This assumption is what allows us to first fix the number of inside-block edges then determine the distribution of the number of inside-block neighbors through a simple sum of existing distributions.

This general idea for snowball sampling in SBM graphs is similar to snowball sampling in binomial and power law graphs, in that we first sum up the degrees over all nodes, then apply a correction for snowball process edges. The difference here is that the correction is not the same for all cases but rather depends on the snowball process itself, which necessitates the second assumption above.

The Mean Effective Sample Node Degree

By applying our theoretical framework to generate estimated DSPDs we can predict differences in the DSPD from different sampling methods. Additionally, a heuristic which can be used is the *mean effective sample node degree*. This metric is defined as the mean degree contribution towards the degree of the sample supernode in G' by a single node in the sample. For random sampling, this is simply the mean degree of a node. For snowball sampling, this is the mean degree of the weighted degree distribution minus 2. A larger mean effective sample node degree indicates that the DSPD will have a lower mean distance.

We note that the process of calculating the heuristic is slightly more involved for snowball sampling on SBM graphs. We take the average of the effective degree of a node reached via an inside-block edge and of a node reached via an out-side block edge, weighing by the probability of reaching via an inside/outside-block edge.

Computational Complexity

Computing our theoretical framework is significantly faster than empirically calculating the DSPD. The complexity of empirically calculating the DSPD involves graph search algorithms on the full graph. Our framework's complexity is only dependent on the size of the support of p, the sample size s, and the maximum distance to recurse to. Additionally, our recursions can be implemented using vectorized operations for further efficiency.

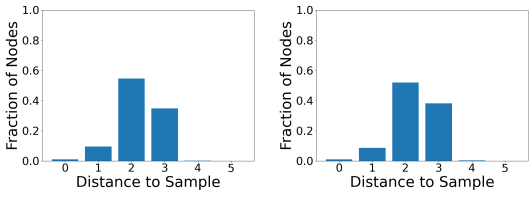
Experimental Study

Study Design

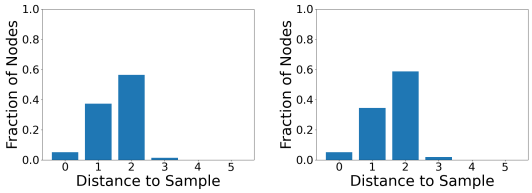
As we mentioned earlier in the Introduction section, the shortest-path distances from nodes to the sample have been shown as a good quantitative indicator of the representativeness of the sample. In this study, we examined both the *shape* of the shortest-path-distance distribution as well as the *mean shortest-path distance* when comparing sampling methods. In terms of the distribution shape, a more heavily right-skewed distribution would indicate that there are fewer nodes that are far from the sample, indicating that the sample is "fairer". In terms of the mean shortest-path distance, a smaller mean distance indicates that the sample does a good job of "covering" all of the nodes.

We conducted experiments on the synthetic graphs presented in the previous section using random sampling and snowball sampling, with the aim of comparing the sampling methods as well as verifying the accuracy of our theoretical framework.

In obtaining empirical distributions, as the sampling methods and graph generation process includes randomness, we had multiple instances of graph/sample combinations as we would generate multiple graphs and/or sample multiple times. For each type of synthetic graph, we generated 10 graphs; for each sampling method, we repeat the sampling 10 times on each generated graph. Thus we would have 100

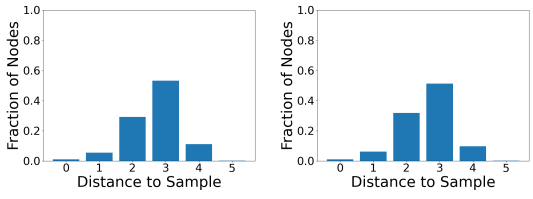


(a) Rand. $s = 20$ (b) SB. $s = 20$

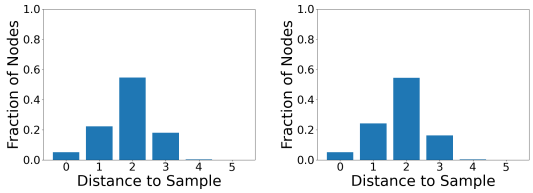


(c) Rand. $s = 100$ (d) SB $s = 100$

Figure 1: Comparison of sampling methods on binomial graphs, sample size s. Rand. = random sampling, SB = snowball sampling.



(a) Rand. $s = 20$ (b) SB. $s = 20$



(c) Rand. $s = 100$ (d) SB $s = 100$

Figure 2: Comparison of sampling methods on Power Law A graphs, sample size s. Rand. = random sampling, SB = snowball sampling.

trials of the experiments (sampling 10 times on each of the 10 graphs). For each trial, we could obtain a histogram distribution of the shortest-path distances, and we took the average on the histogram distributions of all the trials as the final experiment results for the graph/sample combination.

Applying the theoretical framework, meanwhile, was a one step process: with knowledge of the degree distribution of a node in the graph and the sampling method, one pass through the recursive formulas was sufficient.

Experiments on Synthetic Graphs

We conduct experiments on synthetic graphs to examine the effect of sampling methods, as well as verify the accuracy of our theoretical framework.

Comparing Between Sampling Methods Depending on the configuration of the graph, random sampling or snowball sampling could perform better. From Figure 1, in binomial graphs, the DSPDs for random sampling and snowball sampling are similar in shape. However, random sampling gives a lower mean distance than snowball sampling: at $s = 20$, random sampling has a mean distance of 2.237 with snowball sampling at 2.280, at $s = 100$ random sampling has a mean distance of 1.541 with snowball sampling at 1.576.

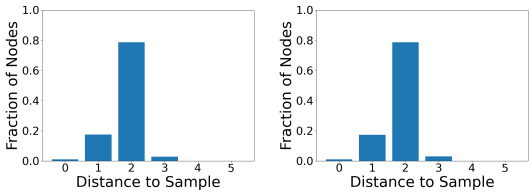
Similarly, from Figure 2 and Figure 3, we see that both Power Law A and Power Law B graphs have very similar distributions for random sampling and snowball sampling. However, for Power Law A graphs snowball sampling gives consistently lower mean distances: at $s = 20$, random sampling has a mean distance of 2.684 while snowball sampling has a mean distance of 2.627, at $s = 100$ random sampling has a mean distance of 1.863 while snowball sampling has a mean distance of 1.824; while for Power Law B graphs random sampling gives consistently lower mean distances: at $s = 20$, random sampling has a mean distance of 1.833 while snowball sampling has a mean distance of 1.836, at $s = 100$ random sampling has a mean distance of 1.315 while snowball sampling has a mean distance of 1.320.

From Figure 4, we see that for SBM graphs, the shape differences between random sampling and snowball sampling are more noticeable. Here random sampling consistently has both a noticeably lighter right tail and a smaller mean distance. For completeness, we report the exact mean distances: at $s = 20$, random sampling has a mean distance of 2.049 while snowball sampling has a mean distance of 2.145, at $s = 100$ random sampling has a mean distance of 1.507 while snowball sampling has a mean distance of 1.448.

In summary, on graphs without community structure the DSPDs for random sampling and snowball sampling were very similar, while on graphs with community structure there was a more noticeable difference. One sampling method is also not consistently better than the other; rather the comparison depends on the graph configuration, and sampling size does not affect this comparison.

Accuracy of Theoretical Framework Figure 5 and Figure 6 demonstrate that there is good agreement between the distributions predicted by our theoretical framework and the empirically observed distributions for all types of synthetic graphs. For all graphs, the predicted distributions match well with the empirical distributions in terms of both shape and center. Examining the exact densities, the predicted densities are more accurate for shorter distances, with discrepancies becoming more apparent at larger distances. These discrepancies could be due to the approximations used by applying our framework. Slight errors introduced by our assumptions can accumulate through the recursive calculations.

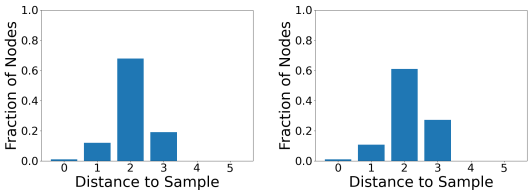
Our mean effective sample node degree heuristic is accurate as well. For example, for Power Law A graphs, the effective mean degree of random sampling is calculated to be 5.55, while the effective mean degree of snowball sampling is calculated to be 6.64, and we do see snowball sampling have a lower mean distance. For Power Law B graphs the effective mean degree of random sampling is calculated to be



(a) Rand. $s = 20$ (b) SB. $s = 20$

(c) Rand. $s = 100$ (d) SB $s = 100$

Figure 3: Comparison of sampling methods on Power Law B graphs, sample size s. Rand. = random sampling, SB = snowball sampling.



(a) Rand. $s = 20$ (b) SB. $s = 20$

(c) Rand. $s = 100$ (d) SB $s = 100$

Figure 4: Comparison of sampling methods on SBM graphs, sample size s. Rand. = random sampling, SB = snowball sampling.

19.47, while the effective mean degree of snowball sampling is calculated to be 19.36, and we do see snowball sampling have a lower mean distance.

For SBM graphs the heuristic is less effective due to the community structures involved, but does work in the example we examined: the effective mean degree of random sampling is calculated to be 12.78, while the effective mean degree of snowball sampling is calculated to be 11.60, and we do see random sampling give a lower mean distance.

Experiments on Real-world Graphs

Though synthetic graphs are commonly used in the network science literature to understand the behavior of algorithms and natural graphs, it is still necessary to verify that they serve as good proxies for real world graphs with respect to DSPDs. We therefore run similar experiments on real world graphs and compare the results with synthetic graphs. The experiment results on the Facebook graph with varying sample sizes are visualized in Figure 7, and the results on three different road networks are visualized in Figure 8.

Comparing Sampling Methods For all the real world graphs, we saw clear differences between the DSPDs for all sampling methods. Between random sampling and snowball sampling, random sampling produced distributions that had both lower mean distances and a lighter right tail. This difference was not observed in the synthetic binomial and power-law graphs, where both the theoretical distributions and the empirical distributions exhibited little difference between random sampling and snowball sampling. There were small differences that could be predicted by our theoretical framework, but not to the scale observed in these real world graphs.

On the other hand, such differences were noticeable in SBM graphs. We are not able to obtain an exact comparison with the real world graphs as the communities in the real world graphs are not of equal size, but SBM graphs do exhibit the trend of random sampling producing distributions with lower mean distances and lighter right tails. This indicates that the community structure is important to DSPDs, as binomial and power-law graphs lack the community structure of SBM and real world graphs. Additionally, the patterns predicted by our theoretical framework and verified by experimentation on synthetic graphs are preserved when examining similar real world graphs.

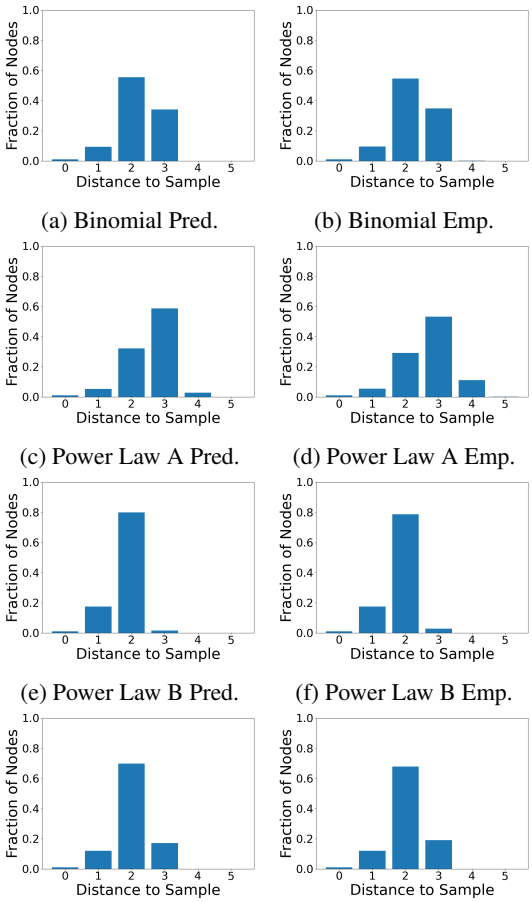
Examination of the centrality based sampling methods indicates that there is much nuance to explore. In the case of the Facebook graph, we see that the centrality based sampling methods all have a lower mean distance than random sampling. However, the opposite is true for the road network graphs. This points to the added complexity of centrality based sampling methods being worthwhile in some, but not all, circumstances. Here, centrality based sampling methods likely perform better in the Facebook graph due to the higher degrees of nodes compared to the road networks, where there are physical limitations on the number of roads that can join at an intersection. Moreover, there are differences in the shapes of the DSPDs between betweenness based sampling and degree based sampling.

In summary, we find that the trends observed in the real world graphs are most consistent with the trends observed in the synthetic SBM graphs, and that graph type affects the relative effectiveness of centrality based sampling methods.

Discussion and Conclusion

Finally, we conclude this paper with further discussions of our theoretical framework and implications of our comparison between synthetic and real world graphs.

Theoretical framework. Our theoretical framework has been shown to be very accurate on graphs without community structure, e.g. binomial graphs and power-law graphs. On graphs with a community structure, it is necessary to



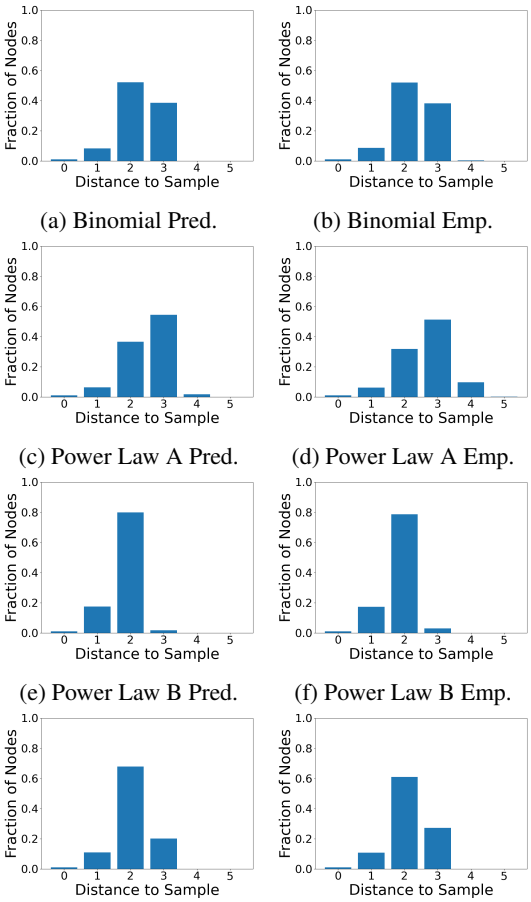
(a) Binomial Pred. (b) Binomial Emp.

(c) Power Law A Pred. (d) Power Law A Emp.

(e) Power Law B Pred. (f) Power Law B Emp.

(g) SBM Pred. (h) SBM Emp.

Figure 5: Comparison of predicted (Pred.) and actual (Emp.) distributions. Distributions from random sampling size $s = 20$ on respective graphs.



(a) Binomial Pred. (b) Binomial Emp.

(c) Power Law A Pred. (d) Power Law A Emp.

(e) Power Law B Pred. (f) Power Law B Emp.

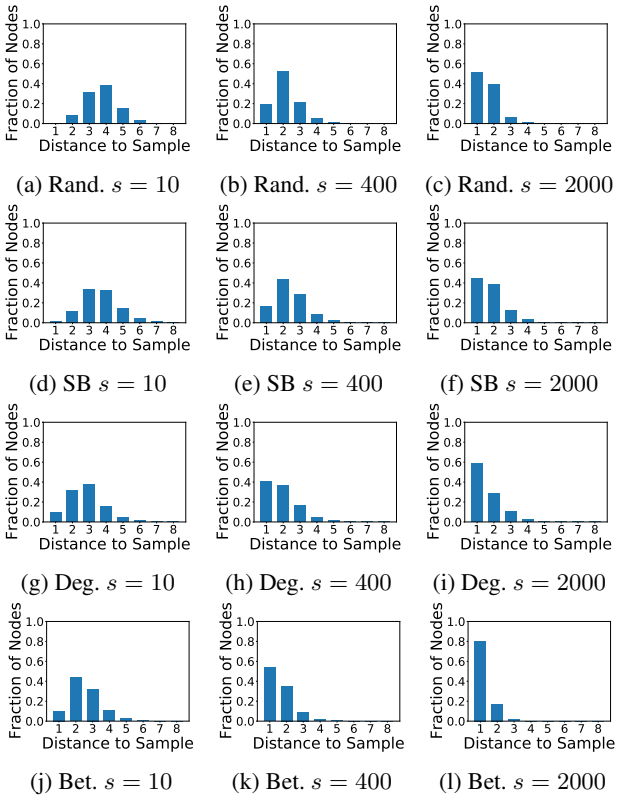
(g) SBM Pred. (h) SBM Emp.

Figure 6: Comparison of predicted (Pred.) and actual (Emp.) distributions. Distributions from snowball sampling size $s = 20$ on respective graphs.

make the recursive formulas more fine-grained to achieve the same level of accuracy. Additionally, the heuristic we propose is effective at predicting the comparison between sampling methods given a model configuration. However, for both the framework and the heuristic, adjustments were needed to account for the community structure in SBM graphs with identical blocks. Further research is needed to adapt our framework to graphs with more complex community structures.

Future work deriving the degree distribution of samples generated via other sampling methods would expand the applicability of our framework. We present methods for random sampling and snowball sampling, but other sampling methods like centrality based sampling are also commonly used. For these methods, it difficult to determine the degree distribution of a single node in the sample and to estimate the correction needed due to edges within the sample. There would be much value in investigating these sampling methods, as in some cases they outperform random sampling and snowball sampling.

Influence of graph properties. Through the experiments on synthetic graphs, we have shown that the types of graphs have a major influence on the DSPD to the sample. This suggests that we need to pay attention to the choice of graph generation models when investigating the representativeness of graph sampling methods using synthetic graphs. Comparing the results on synthetic graphs and those on real world graphs, SBM graphs are closer to those of real world graphs than binomial or power law graphs. In the case of binomial graphs, we believe this is because the synthetic graphs are too dense ($O(n^2)$ edges) compared to real-world graphs, resulting in very short shortest-distances. In the case of power law graphs, we believe this is because the generation method results in fewer high-centrality nodes than the real-world graphs, and that the high-centrality nodes are more closely connected than in real-world cases. This again results in shorter shortest-path distances than those in real world graphs. The community structure in social networks or the planar property of road networks are important factors to be considered.



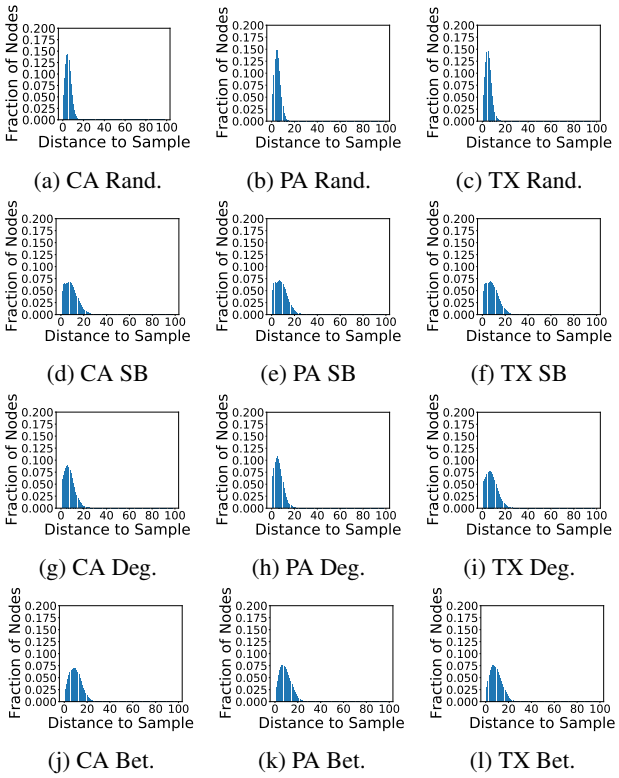
(a) Rand. $s = 10$ (b) Rand. $s = 400$ (c) Rand. $s = 2000$

(d) SB $s = 10$ (e) SB $s = 400$ (f) SB $s = 2000$

(g) Deg. $s = 10$ (h) Deg. $s = 400$ (i) Deg. $s = 2000$

(j) Bet. $s = 10$ (k) Bet. $s = 400$ (l) Bet. $s = 2000$

Figure 7: Comparison of different sampling sizes and sampling methods on the Facebook graph. Rand. = random sampling, SB = snowball sampling, Deg. = degree based sampling, Bet. = betweenness based sampling.



(a) CA Rand. (b) PA Rand. (c) TX Rand.

(d) CA SB (e) PA SB (f) TX SB

(g) CA Deg. (h) PA Deg. (i) TX Deg.

(j) CA Bet. (k) PA Bet. (l) TX Bet.

Figure 8: Comparison of sampling methods on Road Network graphs. Sample size = 1000. CA = California, PA = Pennsylvania, TX = Texas. Rand. = random sampling, SB = snowball sampling, Deg. = degree based sampling, Bet. = betweenness based sampling.

Ethical considerations. This paper investigates the representativeness of graph sampling methods through the lens of shortest-path distances. We believe this is an important problem with considerable ethical impacts on many social science studies. However, we note that the desired representative properties is very application dependent. Our study of the shortest-path distances is motivated by results on bias and fairness in a graph machine learning context. Those using our work should verify that shortest-path distances are similarly important in their context. Nevertheless, we hope that this study can help attract more attention to the problem of the representativeness of graph sampling methods in general.

References

[1] Aittokallio, T.; and Schwikowski, B. 2006. Graph-based methods for analysing networks in cell biology. *Briefings in bioinformatics* 7(3): 243–255.

[2] Barabási, A.-L.; and Albert, R. 1999. Emergence of scaling in random networks. *science* 286(5439): 509–512.

[3] Boldi, P.; Santini, M.; and Vigna, S. 2004. Do your worst to make the best: Paradoxical effects in pagerank incremental computations. In *International Workshop on Algorithms and Models for the Web-Graph*, 168–180. Springer.

[4] Brandes, U. 2008. On variants of shortest-path betweenness centrality and their generic computation. *Social networks* 30(2): 136–145.

[5] Budak, C.; Agrawal, D.; and El Abbadi, A. 2011. Limiting the spread of misinformation in social networks. In *Proceedings of the 20th international conference on World wide web*, 665–674.

[6] Cohen, R.; Havlin, S.; and Ben-Avraham, D. 2003. Efficient immunization strategies for computer networks and populations. *Physical review letters* 91(24): 247901.

[7] Costenbader, E.; and Valente, T. W. 2003. The stability of centrality measures when networks are sampled. *Social networks* 25(4): 283–307.

[8] Goodman, L. A. 1961. Snowball sampling. *The annals of mathematical statistics* 148–170.

[9] Hagberg, A. A.; Schult, D. A.; and Swart, P. J. 2008. Exploring Network Structure, Dynamics, and Function using NetworkX. In Varoquaux, G.; Vaught, T.; and Millman, J., eds., *Proceedings of the 7th Python in Science Conference*, 11 – 15. Pasadena, CA USA.

[10] Harris, C. R.; Millman, K. J.; van der Walt, S. J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N. J.; Kern, R.; Picus, M.; Hoyer, S.; van Kerkwijk, M. H.; Brett, M.; Haldane, A.; del Río, J. F.; Wiebe, M.; Peterson, P.; Gérard-Marchant, P.; Sheppard, K.; Reddy, T.; Weckesser, W.; Abbasi, H.; Gohlke, C.; and Oliphant, T. E. 2020. Array programming with NumPy. *Nature* 585(7825): 357–362. doi:10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

[11] Heckathorn, D. D.; and Cameron, C. J. 2017. Network sampling: From snowball and multiplicity to respondent-driven sampling. *Annual review of sociology* 43: 101–119.

[12] Holland, P. W.; Laskey, K. B.; and Leinhardt, S. 1983. Stochastic blockmodels: First steps. *Social networks* 5(2): 109–137.

[13] Hu, P.; and Lau, W. C. 2013. A survey and taxonomy of graph sampling. *arXiv preprint arXiv:1308.5865* .

[14] Katzav, E.; Biham, O.; and Hartmann, A. K. 2018. Distribution of shortest path lengths in subcritical Erdős-Rényi networks. *Physical Review E* 98(1): 012301.

[15] Katzav, E.; Nitzan, M.; Ben-Avraham, D.; Krapivsky, P.; Kühn, R.; Ross, N.; and Biham, O. 2015. Analytical results for the distribution of shortest path lengths in random networks. *EPL (Europhysics Letters)* 111(2): 26006.

[16] Kurant, M.; Gjoka, M.; Wang, Y.; Almquist, Z. W.; Butts, C. T.; and Markopoulou, A. 2012. Coarse-grained topology estimation via graph sampling. In *Proceedings of the 2012 ACM workshop on Workshop on online social networks*, 25–30.

[17] Leskovec, J.; and Krevl, A. 2014. SNAP Datasets: Stanford Large Network Dataset Collection. http://snap.stanford.edu/data.

[18] Leskovec, J.; Lang, K. J.; Dasgupta, A.; and Mahoney, M. W. 2008. Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters. *CoRR* abs/0810.1355. URL http://arxiv.org/abs/0810.1355.

[19] Leydesdorff, L. 2007. Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. *Journal of the American Society for Information Science and Technology* 58(9): 1303–1319.

[20] Ma, J.; Deng, J.; and Mei, Q. 2021. Subgroup generalization and fairness of graph neural networks. *Advances in Neural Information Processing Systems* 34.

[21] Maiya, A. S.; and Berger-Wolf, T. Y. 2011. Benefits of bias: Towards better characterization of network sampling. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 105–113.

[22] Newman, M. E. 2003. The structure and function of complex networks. *SIAM review* 45(2): 167–256.

[23] Nitzan, M.; Katzav, E.; Kühn, R.; and Biham, O. 2016. Distance distribution in configuration-model networks. *Physical Review E* 93(6): 062309.

[24] Rozemberczki, B.; Allen, C.; and Sarkar, R. 2019. Multi-scale Attributed Node Embedding.

[25] Salganik, M. J.; and Heckathorn, D. D. 2004. Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological methodology* 34(1): 193–240.

[26] Saxena, A.; and Iyengar, S. 2020. Centrality measures in complex networks: A survey. *arXiv preprint arXiv:2011.07190* .

[27] Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; and Monfardini, G. 2008. The graph neural network model. *IEEE transactions on neural networks* 20(1): 61–80.

[28] Stumpf, M. P.; Wiuf, C.; and May, R. M. 2005. Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proceedings of the National Academy of Sciences* 102(12): 4221–4224.

[29] Stutzbach, D.; Rejaie, R.; Duffield, N.; Sen, S.; and Willinger, W. 2006. Sampling techniques for large, dynamic graphs. In *Proceedings IEEE INFOCOM 2006. 25TH IEEE International Conference on Computer Communications*, 1–6. IEEE.

[30] Ventrella, A. V.; Piro, G.; and Grieco, L. A. 2018. On modeling shortest path length distribution in scale-free network topologies. *IEEE Systems Journal* 12(4): 3869–3872.

[31] Wang, T.; Chen, Y.; Zhang, Z.; Xu, T.; Jin, L.; Hui, P.; Deng, B.; and Li, X. 2011. Understanding graph sampling algorithms for social network analysis. In *2011 31st international conference on distributed computing systems workshops*, 123–128. IEEE.

[32] Watts, D. J.; and Strogatz, S. H. 1998. Collective dynamics of 'small-world'networks. *nature* 393(6684): 440–442.

Libraries Used

This study used the following libraries/datasets:

1. NetworkX [9], distributed with the 3-clause BSD license.
2. Numpy [10], distributed with the 3-clause BSD license.
3. SNAP [17], distributed with the 3-clause BSD license.