A Measurement Framework for Directed Networks

Mostafa Salehi, *Student Member, IEEE,* and Hamid R. Rabiee, *Senior Member, IEEE*

Abstract—Partially-observed network data collected by link-tracing based sampling methods is often being studied to obtain the characteristics of a large complex network. However, little attention has been paid to sampling from directed networks such as WWW and Peer-to-Peer networks. In this paper, we propose a novel two-step (sampling/estimation) framework to measure nodal characteristics which can be defined by an average target function in an arbitrary directed network. To this end, we propose a personalized PageRank-based algorithm to visit and sample nodes. This algorithm only uses already visited nodes as local information without any prior knowledge about the latent structure of the network. Moreover, we introduce a new estimator based on the approximate importance sampling to estimate average target functions. The proposed estimator utilizes calculated PageRank value of each sampled node as an approximation for the exact visiting probability. To the best of our knowledge, this is the first study on correcting the bias of a sampling method by re-weighting of measured values that considers the effect of approximation of visiting probabilities. Comprehensive theoretical and empirical analysis of the estimator demonstrate that it is asymptotically unbiased even in situations where stationary distribution of PageRank is poorly approximated.

Index Terms—Directed Networks, Link-tracing Sampling, Estimation, PageRank.

I. INTRODUCTION

MANY real-world communication systems such as the Internet, World Wide Web (WWW), Peer-to-Peer systems, and Online Social Networks (OSNs) can be modeled as a complex network of interacting dynamical nodes. In the last decade, a considerable amount of research has been done on measuring the characteristics of complex networks in various domains [1]. To extract useful knowledge from a network, one should study the collected network data. However, the tremendous growth of the Internet and its applications in recent years has resulted in creation of large-scale complex networks involving tens or hundreds of millions of nodes and links. Thus, it may be impossible or costly to obtain a complete picture of these large networks, and partially-observed data is often being studied for network characterization. The network resulting from such measurements may be thought of as a sample from a larger underlying network. However, sampling in a network context introduces various potential complications such as the bias of selecting the nodes in the procedure of sampling [2].

Many network characterization metrics we are aware of can be expressed as averages of some target functions. For example, the density of infected computers with a virus in the Internet, the ratio of sport websites in the WWW networks, and the average number of copies of a file in a Peer-to-Peer

The authors are with Department of Computer Engineering, Sharif University of Technology, Tehran, Iran. e-mail: mostafa_salehi@ce.sharif.edu, rabiee@sharif.edu

network, can be computed by the average of the target function f: $f(x) = 1$, if node x has desired characteristic (respectively, is infected with that virus, is a sport website, has a copy of that file), and $f(x) = 0$, otherwise.

A measurement framework for computing the average of such functions can be achieved in two steps; (1) Sampling, and (2) Estimation. In many cases (e.g. the Internet), the global structure of a network is initially unknown and there is no sampling frame (i.e., a list of all nodes within a network, from which nodes can be directly sampled). Therefore, using classical sampling methods such as random sampling are either impossible or impractical, and link-tracing based sampling methods are the only feasible solutions to collect data from such networks. In these methods, one can see the neighbors of already sampled nodes and make a decision on which nodes to visit next. Snowball (which is similar to breadth-first search in computer science) [2] and Random walk based methods [3] are among the most popular link-tracing based methods.

In the Estimation step, an estimator is used to approximate the network characteristics. An estimator is a function that accepts a summary of the sampled data as input and outputs an estimate of an unknown network characteristic. Specifically, the selection bias of a sampling method can be corrected by using the general idea of Hansen-Hurwitz estimator [4], i.e. the measured value is weighted inversely proportional to the visiting probability of the corresponding node [5]. Therefore, we need to compute (or approximate) the visiting probabilities of each sampled node. In particular, in undirected networks, when data from snowballs with more than one wave are available, it might be comparatively easy to compute the visiting probabilities of each sampled node based on the first few waves, but it is more complicated to find the required visiting probabilities for snowballs of many waves [2], [6]. For random walk based sampling methods, in a sufficiently long ergodic walk on an undirected network, we would expect that the probability that a node is visited is proportional to its degree [7].

Many of the networks around us such as WWW, Twitter, and Peer-to-Peer networks contain directed links, or links which do not have the same strength in each direction. However, despite recent efforts on characterizing undirected networks based on the sampled data [8], [9], [10], [11], little attention has been made on the statistical properties of sampled directed networks [12], [13]. The directed links pose a considerable challenge to the process of network measurement (i.e., unknown bias in the sampled nodes). In summary, computing (or approximating) the visiting probabilities of sampled nodes for directed networks still remains as an open problem. For example, the stationary distribution for random walks on a directed network is no longer determined by the degree sequences. We can only show that a random walk process on

a strongly connected directed network has a unique stationary distribution (but, there is no analytical solution for it yet). However, most real-world directed networks are not strongly connected. Without this property, the limiting distribution of a random walk is sensitive to the starting distribution, and its analysis can be very complicated. A modification of random walk which overcomes this problem allows for a jump with small probability to a node at random from the whole network. That is called PageRank in the Internet search literature [14]. It's clear that according to our assumptions about the absence of sampling frame, jumping to a random node from the whole network may not be possible. Therefore, PageRank can not be applied to our problem in its global definition.

In this paper, we propose a complete framework to measure nodal characteristics which can be defined by average target functions in an arbitrary directed network. In the sampling step, we introduced a novel link-tracing algorithm by utilizing the general idea of PageRank. We employ a modified version of PageRank, called personalized PageRank [15], using a specific set of nodes as a seed vector. Jumping is limited to this seed vector. Thus, personalized PageRank determines the importance of every node to the seed by using local information. In particular, the proposed method samples the underlying network by moving from a node to one of its neighbors through an outgoing link, based on the approximated probability of personalized PageRank.

Since these probabilities represent an approximation of the exact visiting probabilities, we utilize the idea of approximate ratio importance sampling to propose a new estimator. The previous well-known work [5], [16] on the estimators for correcting the bias, does not consider the effect of approximation of visiting probabilities which we address in this paper. Importance sampling is one of the methods of Monte Carlo simulation [17] in which instead of generating a sample from the target distribution (e.g., uniform distribution in our case), the estimator generates a sample from a different trial distribution (e.g., the distribution generated by the proposed sampling algorithm). We evaluate our framework over large synthetic and real datasets in terms of estimated average of a target function (here, we focus on the infection ratio and outdegree distribution) and its bias. Our results show that the proposed method performs well even in low sampling rates.

II. Related Work

Link-tracing based sampling methods in the context of networks are somewhat distinct from classical sampling methods [2]. Random walk (RW) [7] is one of the most important and widely used sampling methods in different kind of network contexts such as uniformly sampling Web pages from the Internet [18], content density in Peer-to-Peer networks [19], degree distributions of the Facebook social graph [10], and collecting data from information diffusion networks [20], [21]. A classic RW, samples a graph by moving from a node, x, to a neighboring node, y, through an outgoing link, chosen uniformly at random from the neighbors of x. By this process links and nodes are sampled. In any given connected and non-bipartite undirected graph, the classic RW is biased towards node with higher degree.

In general, the probability of selecting the next node determines the probability that nodes are visited in sampling procedure. Metropolis-Hastings technique [22] can be used to modify the probabilities of the node selection in random walk in order to have the uniform stationary distribution for visiting each node. This technique is a general Markov Chain Monte Carlo (MCMC) method [23] for sampling from probability distributions based on constructing a Markov Chain that has the desired distribution as its stationary distribution. This approach, known as Metropolis-Hastings Random Walk (MHRW), has been applied to Peer-to-Peer networks [19], Facebook [10], and Twitter [24].

Alternatively, one can use the unmodified classic RW method to sample from a graph and correct the degree bias by re-weighting the sampled values. The Respondent-driven sampling (RDS) [3], [16] is an example of this approach. Actually, this is a framework in the field of social sciences to sample and infer in hard-to-reach populations such as injection drug users. In these populations, a sampling frame for the target population is not available. Sampling from OSNs or WWW network graph is analogous to the sampling of hidden population in the social sciences. In the context of graph sampling, the RDS method has been used for Facebook [10], Twitter [24], and Peer-to-Peer networks [19].

One of the main assumptions of the above mentioned studies is that the underlying network is undirected. Since most real-world networks (such as WWW, Twitter, and Peer-to-Peer networks) are directed, this assumption is not easily met in real life. However, little attention has been paid to directed networks by considering the effect of directed links. The authors in [25] consider directed links in the Peer-to-Peer network while maintaining desirable statistical properties (i.e., uniformity) for the sampling procedure. The main idea of this RW based method is to avoid the calculation of each node's visiting probability, and instead to adjust the transition probabilities iteratively, converging to a state in which the sum of transition probabilities into each node equals 1. The resulting transition matrix is said to be doubly stochastic, and induces uniform stationary distribution.

In [12], it is shown that when the sample size is relatively small, the RDS method may generate relatively large biases and errors if the studied networks are directed, indicating that estimates from previous RDS studies should be interpreted and generalized with caution. In [13], a random walk sampling algorithm with jumps, called Directed Unbiased Random Walk (DURW), is proposed that achieves asymptotically unbiased estimates of the outdegree distribution of a directed graph. The authors construct an undirected graph using the nodes that are selected by the random walker on the directed graph. The undirected graph is built to allow the walker to traverse known outgoing links backwards, and guarantees that the probability of sampling a node can be approximated, even though incoming links are not observed. However, this method is based on the assumption that nodes can be sampled uniformly at random from the original graph (i.e., there is some means of obtaining a list of nodes in the network) which is not always feasible. For example, while it is feasible for Wikipedia and Twitter, it is not feasible for the WWW graph.

In conclusion, while the effect of directed links have been studied, there is no efficient measurement framework for directed networks in the literature. Moreover, no analytical solution for the visiting probabilities of the nodes is available for a general directed network.

III. PRELIMINARIES

A. Basic Notations and Definitions

Let $G = (V, E)$ with $n = |V|$ and $m = |E|$ be the graph representing a complex network, where V is the set of nodes, and $E = \{(i, j) : i, j \in V, A_{ij} = 1\}$ is the set of unweighted directed links between pairs of nodes. A_{ij} indicates the presence ($= 1$) or absence ($= 0$) of relation of interest (such as a hyper-link in WWW, friendship in a social network, and file transferring in a Peer-to-Peer network) from i to j. The above $n \times n$ matrix A is called an adjacency matrix. The indegree of a node x, $d_{in}(x)$, is the number of distinct ingoing links $(y_1, x), ..., (y_k, x)$ into x, and its outdegree, $d_{out}(x)$, is the number of distinct outgoing links $(x, y_1), ..., (x, y_k)$ out of x.

A link is called reciprocal if a connected pair of nodes have both an ingoing and an outgoing link between each other. The proportion of reciprocal links in a network called reciprocity, r (i.e. $r = 1$ when the network is undirected, and $r = 0$ when the network is directed in a way such that there are no reciprocal links).

Let S be a sample of nodes where $S \subset V$, and $G(S)$ is the induced subgraph of G based on the sample S. Let $N(S)$ denote the neighborhood of S;

$$N(S) = \{w \in V - S : \exists v \in S \ s.t. \ (v, w) \in E\} \quad (1)$$

Therefore, $N(S)$ is the set of nodes that we know exist due to other outgoing links but have not yet visited.

Let L be a set of real-valued labels. For instance, a label can be the degree of a node. We assume that a label $l_x \in L$ is assigned to each node $x \in V$ by a target function $f : V \to L$, i.e. $f \subseteq \{(x, l_x) : x \in V, l_x \in L\}$. We can express many network characterization metrics as averages of some target functions over V. Given a target function f, the average of f is: $avg(f) = \sum_{x \in V} f(x)/|V|$.

Our main goal in this paper is to propose a measurement framework to compute $avg(f)$ (based on the sampled nodes S) for a general target function f (that defines a nodal characteristic) in an arbitrary directed network G. Our only assumption on G is that it is weakly connected, i.e. every node can be reached from every other node by traversing the links without considering the direction of the links.

B. Personalized PageRank Vectors

PageRank was first introduced by Brin and Page [14] for search algorithms in the WWW network. However, it can be defined for any graph as a stationary distribution of a certain random walk over that graph. At each step, with probability $(1-\alpha)$, the random walk follows a randomly outgoing link of a node, and with probability α the random walk makes a jump to a new node chosen uniformly among all nodes in the network. The jumping constant α ($0 < \alpha \leq 1$), controls the diffusion rate. With smaller α, the random walks spread further from the initial nodes before performing a random jump. Traditionally, the value of α is chosen to be 0.15.

Let A be the adjacency matrix of a graph G. The PageRank vector $\mathbf{p}(\mathbf{s})$ of this graph is a visiting probability distribution on the nodes of G that can be defined as the solution of the following equation:

$$\mathbf{p}(\mathbf{s}) = \alpha \mathbf{s} + (1 - \alpha)\mathbf{p}D_A^{-1}A \quad (2)$$

where $\mathbf{s}$ is a seed vector that includes the initial distribution for starting nodes. We use the notation D_A^{-1} to indicate the diagonal matrix with the inverse outdegree for each node. Globally, uniform $\mathbf{s} = \mathbf{1}/n$ is considered for computing PageRank. However, an arbitrary distribution $\mathbf{s}$ can lead to creation of personalized PageRank [26]. Specifically, the random walk in the personalized PageRank only jump to a few nodes of personal interest. It can be shown that for any $\mathbf{s}$ and α, there is an unique vector $\mathbf{p}(\mathbf{s})$ satisfying Equ.(2) [15].

There are various algorithms for computing global PageRank and personalized PageRank [15], [27], [28]. As a standard PageRank algorithm, we can efficiently compute the solution of Equ.(2) by applying the power method [29]. Given some initial distribution $\mathbf{s}$, the power method is defined as following iteration:

$$\mathbf{p}^{(k+1)} = (1 - \alpha)\mathbf{p}^{(k)}D_A^{-1}A$$
$$\mathbf{p}^{(k+1)} = \mathbf{p}^{(k)} + (1 - \|\mathbf{p}^{(k+1)}\|_1)\mathbf{s}$$

where $\|.\|_1$ denotes the 1-norm. As an algorithm, the power method continues this iteration until $\|\mathbf{p}^{(k+1)} - \mathbf{p}^{(k+1)}\|_1 < \delta$ for a user-provided stopping tolerance δ. Based on the idea of the power method, Gleich and Polito [28] present an efficient algorithm, called boundary-restricted personalized PageRank algorithm (BRPPR), to compute an approximation of the personalized PageRank vector $\mathbf{p}(\mathbf{s})$ by examining only a small fraction of the input web graph near the starting vector $\mathbf{s}$. Specifically, this algorithm iteratively divide the web pages into an active and inactive set. At each iteration, the set of active web pages is expanded to include more web pages that are likely to have a high personalized PageRank value. Specifically, it expands pages until the total rank on the frontier set (the set of pages that we know exist due to other outgoing links but have not yet visited) of pages is less than an expansion tolerance. Therefore, only the set of active pages and their outgoing link information are actually involved in the computation. Moreover, a stopping tolerance threshold is considered in order to determine the termination of the approximation algorithm. The analysis of BRPPR leads to the following theorem (theorem 3.3 and remark 3.4 of [28]):

Theorem 1: The BRPPR algorithm with expansion tolerance κ and stopping tolerance δ yields an approximate PageRank vector $\hat{\mathbf{p}}$, where $\|\mathbf{p} - \hat{\mathbf{p}}\|_1 \leq \frac{2(1-\alpha)}{\alpha}.\kappa + \frac{2-\alpha}{\alpha^2}.\delta$

where $\|\mathbf{p} - \hat{\mathbf{p}}\|_1$ gives the Manhattan distance between two distribution vectors of $\mathbf{p}$ and $\hat{\mathbf{p}}$. Although, the authors in [28] address the problem of calculating the PageRank score of a Internet webpage in the field of search engines, we utilize their general idea here in the context of network sampling.

C. Importance Sampling

As mentioned, we are interested in measurement of quantities that can be written as averages of some target functions f over a finite set V (i.e., the set of all nodes). Let $\pi : V \to [0, \infty)$ be the target distribution on a set V. In our problem, the desired π is the uniform distribution on V, then all nodes x are equally likely to be selected, i.e. $\pi(x) = 1/|V|$. Let $\hat{\pi} : V \to [0, \infty)$ be a target measure that is unnormalized form of π, i.e. $\hat{\pi}(x) = \pi(x).Z_{\hat{\pi}}$ where $Z_{\hat{\pi}} > 0$ is the normalization constant of $\hat{\pi}$. We say that measure $\hat{\pi}$ induces a corresponding probability distribution on V. Therefore, $\hat{\pi}$ is a relative weight, which represents the probability of x to be chosen in the distribution π. In our problem, we have $\hat{\pi} = 1$ (for all $x \in V$) and $Z_{\hat{\pi}} = |V|$.

The average of a target function relative to $\hat{\pi}$ is essentially a sum where the target measure is the probability distribution π. Hence, we can write

$$avg(f) = avg_{\hat{\pi}}(f) = sum_{\pi}(f) = \sum_{x \in V} f(x).\pi(x) \quad (3)$$

For example, in our previous example of infection in a population, the infection ratio is an average of the target function relative to an uniform target measure. Alternatively, it is a sum of the same function relative to the uniform distribution on V (i.e., $\pi(x) = 1/|V|$ for all x).

The naive estimator for $avg_{\hat{\pi}}(f)$ works as follows [17]: (1) Generate a random sample X from the distribution π induced by $\hat{\pi}$, (2) Compute the function of f for sampled X (i.e., $f(X)$). It is easy to check that this estimator is unbiased. However, since we assume that there is no sampling frame in our problem, uniform sampling from the distribution π may be hard or costly. Consequently, this simple estimator is inefficient. To solve this problem, we can use *Importance Sampling (IS)*, as one of the methods of monte carlo simulation [17]. In the following, we state some theorems about the general idea of importance sampling [30]. Although, the authors in [30] address the problem of externally measuring aggregate functions over documents indexed by search engines, we use them here in the context of network sampling.

1) Importance Sampling (IS): The basic idea of importance sampling is: instead of generating a sample Y from the target distribution π, the estimator generates a sample X from a different trial distribution p on V to directly estimate statistical sums relative to the target measure $\hat{\pi}$. Specifically, given a target function f, importance sampling estimator estimates $sum_{\hat{\pi}}(f) = \sum_{x \in V} f(x).\hat{\pi}(x)$. The trial distribution p can be any distribution, as long as $supp(\pi) \subseteq supp(p)$ (here, $supp(p) = \{x \in V | p(x) > 0\}$ and $supp(\pi)$ is defined similarly). In particular, we can choose p to be a distribution that is easy to sample from. By considering $\hat{p}$ as unnormalized form of p, the importance sampling estimator is then defined as follows:

$$E_{IS}(X) = f(X).\frac{\hat{\pi}(X)}{\hat{p}(X)} = f(X).w(X), \quad X \in V \quad (4)$$

where $w(X)$ is called the importance weight. Function f is computed by samples of distribution p. It can be shown that $E_{IS}(X)$ is an unbiased estimator for $sum_{\hat{\pi}}(f)$. Implementation of an importance sampling estimator requires: (1) Ability to sample efficiently from the trial distribution p; and (2) Ability to compute the importance weight $w(x)$ (or its estimator) and the function value $f(x)$, for any given node $x \in V$. There is no need to know the normalization constant $Z_{\hat{\pi}}$ or to be able to sample from π. This basic importance sampling estimator is only suitable for estimating sums. We next extend it for estimating averages.

Importance Sampling for average: Recall that the estimator for averages estimates $avg_{\hat{\pi}}(f) = sum_{\pi}(f)$. If we are able to compute $\pi(x)$ for each $x \in V$, we can use the importance sampling estimator described above to estimate average of a function relative to $\hat{\pi}$. However, in many cases we can only compute $\hat{\pi}(x)$ and not $\pi(x)$. For example, in our case, $\pi(x) = 1/|V|$, where $|V|$ is the number of nodes, that is typically unknown. Ratio importance sampling is a technique for estimating averages relative to any target measure $\hat{\pi}$, even when $\pi(x)$ is unknown.

The ratio importance sampling estimator computes two sum estimators: M_1 of $sum_{\hat{\pi}}(f)$, and M_2 of $sum_{\hat{\pi}}(w)$, and then outputs M_1/M_2. However, there is one problem: the expectation of a ratio is not the ratio of the expectations, i.e., $\mathbb{E}(M_1/M_2) \neq \mathbb{E}(M_1)/\mathbb{E}(M_2)$. To solve this problem, ratio importance sampling utilize the following idea: if we replace $\mathbb{E}(M_1)$ and $\mathbb{E}(M_2)$ by averages of multiple independent instances of the estimator of $\mathbb{E}(M_1)$ and $\mathbb{E}(M_2)$, the difference between $\mathbb{E}(M_1/M_2)$ and $\mathbb{E}(M_1)/\mathbb{E}(M_2)$ diminishes to 0. Therefore, the ratio importance sampling estimator for $avg_{\hat{\pi}}(f)$ is then defined as follows:

$$E_{RIS}(X_1, X_2, ..., X_n) = \frac{\frac{1}{n}\sum_{i=1}^{n} E_{IS}(X_i)}{\frac{1}{n}\sum_{i=1}^{n} w(X_i)} \quad (5)$$

where $X_1, X_2, ..., X_n$ are n independent samples from the trial distribution p.

2) Approximate Importance Sampling (AIS): Unfortunately, in many cases we are unable to accurately compute the importance weight function w. This problem arises, e.g. in our case, due to complexity of analysing a RW based sampling method on a directed network. In the following, we address the effect of the approximate importance weights on importance sampling. Approximate importance sampling method employ an "approximate importance weight function" $u(x)$ rather than the exact one $w(x)$. This estimator can be defined as:

$$E_{AIS}(X) = f(X).u(X) \quad (6)$$

where X is distributed according to the trial distribution $\hat{p}$. The estimation generated by approximate importance sampling is close to the true value as long as the importance weight function $w(x)$ and the approximate importance weight function $u(x)$ are similar. To analyse the bias of this estimator, we consider the following theorem.

Theorem 2: ([30]) Suppose $supp(u) \subseteq supp(w)$
$$\mathbb{E}(E_{AIS}(X)) = sum_{\hat{\pi}}(f(Y)).\mathbb{E}(\frac{u(Y)}{w(Y)}) + Z_{\hat{\pi}}.cov(f(Y), \frac{u(Y)}{w(Y)})$$

where $X \sim \hat{p}$, $Y \sim \pi$, and $Z_{\hat{\pi}}$ is the normalization constant. Therefore, there are two types of bias in this estimator: (1) multiplicative bias, depending on the expectation of u/w relative to π. It may have a significant effect on the estimator's bias, and thus must be removed. (2) additive bias, depending on the correlation between f and u/w and on the normalization constant $Z_{\hat{\pi}}$. It is typically less significant, as in many practical situations f and u/w are uncorrelated (e.g., when f is a constant function, as is the case with infection ratio example).

To remove the multiplicative bias, we assume that it is possible to estimate the multiplicative bias, and divide $E_{AIS}(X)$ by this estimate. For an instance $x \in V$, the ratio $u(x)/w(x)$ is called the weight skew at x. The multiplicative bias factor is the expected weight skew relative to the target distribution π. To remove this bias, we need to estimate the expected weight skew. Let $E_{WSE}(X)$ be an unbiased weight skew estimator for $\mathbb{E}(\frac{u(Y)}{w(Y)})$. It follows from Theorem 2 that:

$$\frac{\mathbb{E}(E_{AIS}(X))}{\mathbb{E}(E_{WSE}(X))} = sum_{\hat{\pi}}(f(Y)) + \frac{Z_{\hat{\pi}}.cov(f(Y), \frac{u(Y)}{w(Y)})}{\mathbb{E}(\frac{u(Y)}{w(Y)})} \quad (7)$$

Thus, the ratio of the expectations of the two estimators, $EST_{AIS}(X)$ and $EST_{WSE}(X)$, gives us the desired result $sum_{\hat{\pi}}(f)$, modulo an additive bias factor that depends on the correlation between f and u/w. The additive bias can be considered 0 since f is often a constant function. Ignoring this additive bias, it would seem that the ratio $EST_{AIS}(X)/EST_{WSE}(X)$ is a good estimator for $sum_{\hat{\pi}}(f)$. Although the expectation of a ratio is not the ratio of the expectations, one applies the ratio estimation technique similarly to what we did in the ratio importance sampling. Therefore, the approximate ratio importance sampling estimator for $sum_{\hat{\pi}}(f)$ can be defined as follows:

$$E_{ARIS}(X_1, X_2, ..., X_n) = \frac{\frac{1}{n}\sum_{i=1}^{n} E_{AIS}(X_i)}{\frac{1}{n}\sum_{i=1}^{n} E_{WSE}(X_i)} \quad (8)$$

where $X_1, ..., X_n$ are n independent samples from the trial distribution $\hat{p}$.

Approximate importance sampling for averages: Here, we employ E_{ARIS} to design an importance sampling estimator for averages. Recall that the estimator for averages estimates $avg_{\hat{\pi}}(f) = sum_{\pi}(f)$. Thus, we need to compute the importance weight function $w_{avg}(x) = \pi(x)/\hat{p}(x)$. However, $\pi(x)$ cannot be calculated exactly. For solving this problem, we use an approximate weight function $u(x) \approx w_{avg}(x)$ instead of $w_{avg}(x)$, and then fixing the bias using approximate ratio importance sampling estimator (i.e., E_{ARIS}) described above. To this end, we need to come up with the weight skew estimator whose expectation equals $\mathbb{E}(u(Y)/w_{avg}(Y))$, where Y is distributed according to π. It can be shown that with a constant function f, the approximate weight $u(X)$ itself is an unbiased estimator for $\mathbb{E}(u(Y)/w_{avg(Y)})$ [30] . We thus set $E_{WSE}(X) = u(X)$. Therefore, approximate importance

sampling for averages, E_{AVG}, is then defined as follows:

$$E_{AVG} = \frac{\frac{1}{n}\sum_{i=1}^{n} f(X_i).u(X_i)}{\frac{1}{n}\sum_{i=1}^{n} u(X_i)} \quad (9)$$

where $X_1, ..., X_n$ are n independent samples from the trial distribution $\hat{p}$. The following theorem shows the bias and the variance of this estimator.

Theorem 3: ([30]) The bias and variance of E_{AVG} (equ. 9) are defined as
$Bias(E_{AVG}) = E_{AVG}(X_1, X_2, ..., X_n) - avg_{\hat{\pi}}(f) = \frac{Z_{\hat{\pi}}.cov(f(Y), \frac{u(Y)}{w_{avg}(Y)})}{\mathbb{E}(\frac{u(Y)}{w_{avg}(Y)})} + O(\frac{1}{n})$
$var(E_{AVG}(X_1, X_2, ..., X_n)) = O(\frac{1}{n})$

IV. PROPOSED MEASUREMENT FRAMEWORK

The pseudo code of proposed framework, called Directed-Network-Measurement (DNM), is shown in Algorithm IV.1. Our main goal is to approximate the probability of each visited node in sampling procedure, and correct the bias of sampling by using the idea of approximate importance sampling. The proposed framework is designed through two steps; sampling and estimation. The sampling step initially starts with an empty sample set S, and adds a random node y to it as a seed. Subsequent nodes of the sample are chosen based on the following procedure. In each iteration, the set of neighborhood of already sampled nodes, i.e. $N(S)$, and its corresponding active link matrix, L, is generated. The matrix L is the adjacency matrix for all outgoing links from nodes in S, but may reference nodes that are not in S. We then compute one PageRank iteration on L from a vector $\hat{p}$ to a vector $\tilde{p}$. Next, we add nodes (that are likely to have a high personalized PageRank value) to S until the total probability on the remainder of the neighborhood is less than an expansion ratio of κ.

As we mentioned earlier, the essential property of a sampling method that makes it appropriate for network inference is that its visiting probabilities are known or estimable for the sampled nodes. In the estimation step, we use computed PageRank value of each sampled node (i.e. $\hat{p}$), as an approximation of the exact visiting probability. Then, we utilize the idea of approximate importance sampling to design a new estimator and study its performance for a general target function. Specifically, we consider the obtained $\hat{p}(x)$ as an approximate weight function $u(x)$.

According to the general ratio importance sampling estimator (refer to Equ.9), we propose the following estimator for computing the average of a general target function f in a directed network.

$$E_{DIR} = \frac{\frac{1}{n}\sum_{i=1}^{n} f(X_i).\hat{p}(X_i)}{\frac{1}{n}\sum_{i=1}^{n} \hat{p}(X_i)} \quad (10)$$

where $X_1, X_2, ..., X_n$ are n samples extracted by sampling procedure.

Algorithm IV.1: $\text{DNM}(y, \alpha, \kappa, \delta)$

$S = \{y\}$; L = matrix of y to its outgoing links
SUM = 0
WSE = 0
$\hat{\mathbf{p}} = 0; \hat{\mathbf{p}}(y) = 1$
% Sampling step
repeat

$$
\textbf{do}
\begin{cases}
\tilde{\mathbf{p}} = (1 - \alpha)\hat{\mathbf{p}} D_L^{-1} L \\
\tilde{\mathbf{p}} = \tilde{\mathbf{p}} + (1 - \|\tilde{\mathbf{p}}\|_1) \\
\textbf{while } \tilde{\mathbf{p}}(N(S)) > \kappa \\
\quad \textbf{do}
\begin{cases}
\text{Find } x \in N(S) \text{ with max value in } \tilde{\mathbf{p}} \\
\text{Add } x \text{ to } S \\
\text{Remove } x \text{ from } N(S) \\
\text{Update } \tilde{\mathbf{p}}(N(S))
\end{cases} \\
\text{Update } D_L \text{ and } L \\
\omega = \|\tilde{\mathbf{p}} - \hat{\mathbf{p}}\|_1 \\
\hat{\mathbf{p}} = \tilde{\mathbf{p}}
\end{cases}
$$

until $\omega < \delta$
% Estimation step
for $x \in S$

$$
\textbf{do}
\begin{cases}
\text{SUM = SUM} + f(x).\hat{\mathbf{p}}(x) \\
\text{WSE = WSE} + \hat{\mathbf{p}}(x)
\end{cases}
$$

return (SUM/WSE)

The following theorem shows the bias and variance of the proposed estimator.

Theorem 4: The bias and variance of E_{DIR} (equ. 10) are defined as

$Bias(E_{DIR}) = E_{DIR}(X_1, X_2, ..., X_n) - avg_{\hat{\pi}}(f) = \frac{|V| \cdot cov(f(Y), \hat{\mathbf{p}}/\mathbf{p})}{\mathbb{E}(\hat{\mathbf{p}}/\mathbf{p})} + O(\frac{1}{n})$

$var(E_{DIR}(X_1, X_2, ..., X_n)) = O(\frac{1}{n})$

where $Y \sim \pi = 1/|V|$, and $Z_{\hat{\pi}}$ is the normalization constant.

proof: By substituting $u = 1/|V|$, $w_{avg} = \mathbf{p}$, and $Z_{\hat{\pi}} = |V|$ into Theorem 3, the theorem is proved.

We conclude from this theorem that if we use sufficiently large number of samples, then we are likely to obtain an estimate of $avg_{\hat{\pi}}(f)$, which has only additive bias that depends on the correlation between f and $\hat{\mathbf{p}}/\mathbf{p}$.

V. EXPERIMENTAL EVALUATION

A. Datasets

Synthetic Networks: We utilize three kinds of random models for generating directed networks:

Directed Erdos-Renyi (DER) networks [31]: The model starts with $|V|$ nodes, which link to each other with probability $p_0 \in (0, 1)$. The reciprocity of these network equals to 1. In order to decrease the reciprocity, reciprocal links in the base network are randomly chosen and converted to irreciprocal links. We generate networks by using this model with $|V| = 2000$, $p_0 = 0.1$, and $r = 0.6$.

Directed Watts-Strogatz (DWS) networks [32]: The model starts from a completely regular network with identical degree (k-nearest neighbors' connection) and clockwise links. Each link will be rewired with two randomly selected nodes with probability $p_0 \in (0, 1)$. In simulation, we generate networks with $|V| = 2000$ nodes by setting $k = 20$ and $p_0 = 0.1$.

Directed Scale Free (DSF) [33]: Network starts with m_0 nodes, which link to each other with probability p_0 (as in an Erdos-Renyi random graph). At each time step, one node and m links are added to the network. The endpoints of the links are randomly selected among all nodes according to the following probability: $p_1(x) = \beta_1 \frac{d_{in}(x)}{|E|} + \beta_2 \frac{d_{out}(x)}{|E|} + \beta_3 \frac{1}{|V|}$, where $\beta_1 + \beta_2 + \beta_3 = 1$. We generate networks with $|V| = 1000$ and average degree of nodes = 50. We leave the number of links, $|E|$, unconstrained. For attaching probability, we set $\beta_1 = 0.7, \beta_2 = 0.2, \beta_3 = 0.1$. Moreover, we set $m_0 = 2$ for generating initial Erdos-Renyi graph and consider $p_0 = 1$ to have them fully connected.

Real-world Networks: Moreover, we considered some real-world directed network:

An Online Social Network (OSN): This is a Facebook-like Social Network [34] that originate from an online community for students at University of California, Irvine. The dataset includes the 1,899 users that sent or received at least one message. There are 20,296 directed links among these users.

A Peer-to-Peer Network (P2P): We use a snapshot of the Gnutella Peer-to-Peer file sharing network collected in August 2002 [35]. Nodes and links represent hosts and connections between them in this network, respectively. The dataset includes the 10,876 hosts and 39,994 directed links.

The Internet Autonomous System (AS): The network of routers comprising the Internet can be organized into sub-networks called Autonomous Systems (AS). Each AS exchanges traffic flows with some neighbors. We use a CAIDA AS graph collected in February 2004 [36]. The dataset contains 16,493 nodes and 66,744 links.

B. Target Function

The fast growth of Internet and other communication networks makes them a suitable target for malicious activities. An infection spreads through the links of such networks constructed by computers and their communication channels. The process by which malicious objects such as worms, trojan horses, and computer viruses travel through computer networks is analogous to the process of spreading epidemics through a population. In this paper, we study the infection ratio in spreading of a disease in a population as a network characteristic. It is an example of family of binary properties of a network where each node is tagged by label 0 or 1. In particular, infection ratio can be considered as the average of a target function f: $f(x) = 1$, if individual x is infected, and $f(x) = 0$, otherwise.

To spread disease on the test datasets, we use the susceptible-infectious-recovery (SIR) epidemic spreading model [37]. This is an epidemiological model widely used to simulate the spreading of epidemics, i.e. number of people infected with a contagious disease, in a population as a function of time. At every time step, the model assumes a transition rate θ_1 for a susceptible person to become infected, if an infected neighbor is present, and a rate θ_2 for an infected person to become recovered or die. The recovered person will
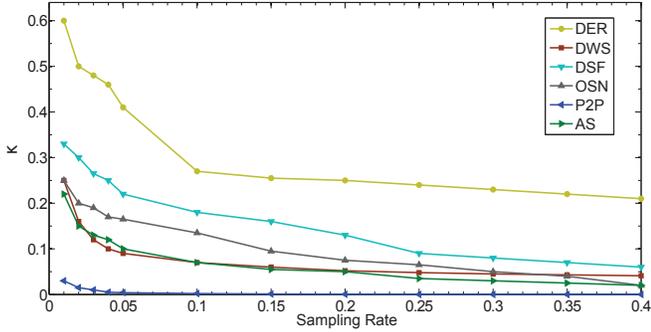
Fig. 1. The relation between expansion tolerance (κ) of the proposed sampling method and sampling rate for all test networks.

never be infected again. In the simulations, we use $\theta_1 = 0.2$ and $\theta_2 = 0.05$. We run the simulation until desired ratio of infected nodes obtained.

C. Experimental Setups

We evaluate the performance of the DNM framework in various situations and in terms of different aspects. We consider two metrics; 1) estimated infection ratio, and 2) bias (which is defined by the absolute difference between estimated and true infection ratio). For each sampling rates in the range of 0.01% to 40%, we repeated the simulations 100 times by selecting random nodes as seed, and finally, averaged the obtained values. We set jumping factor $\alpha = 0.15$ and stopping tolerance $\delta = 10^{-7}$ as the inputs of the proposed framework.

Moreover, we set expansion tolerance κ based on the desired sampling rates. Since, the parameter κ controls the accuracy of the approximation of PageRank values (i.e., the visiting probabilities of the sampled nodes), different values of κ provides different sampling rates. Specifically, the less κ leads to more samples, and consequently the more accuracy is obtained. We study the obtained κ per various sampling rates for all test networks. Our result is demonstrated in Figure 1. As we can see, the desired sampling rate depends on the structure of underlying network. According to Theorem 1, the maximum error of the approximating visiting probabilities is in the range of $(0, 6.8)$. This means that the proposed sampling method computes the visiting probabilities of the sampled nodes with reasonable accuracy even in low and medium sampling rates.

D. Evaluation Results

1) Main Results: The results of our study on the performance of proposed framework is shown in Figure 2. As we can see, in general, the error in terms of estimated infection ratio and bias decreases significantly by increasing the sampled nodes. In particular, we reach desirable result in medium sample rates (estimated infection ratio and bias converges to 20% and 0, respectively). However, we found different behaviour in DWS networks; the estimated infection ratio is close to true value in all sampling rates (i.e. it's independent of number of samples). Moreover, the bias decreases with increasing the sampling rate.
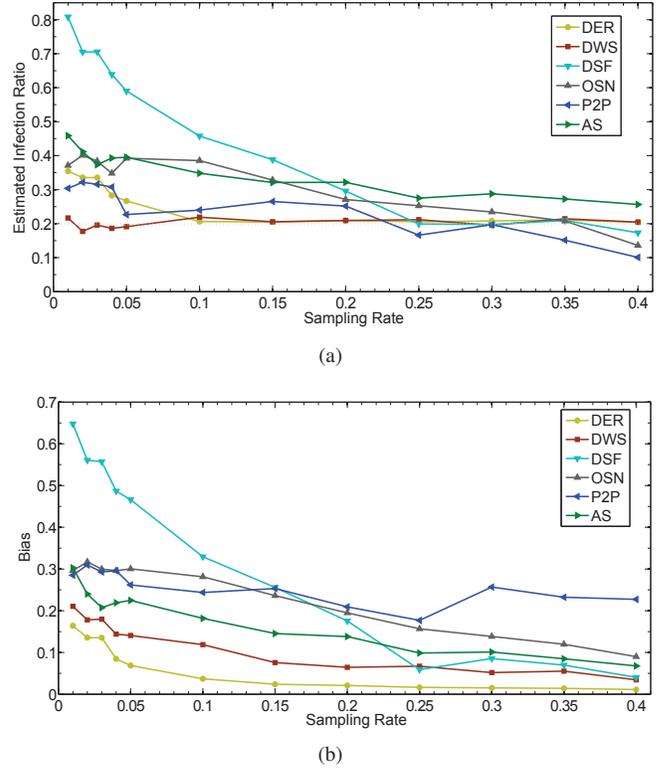

(a)

(b)

Fig. 2. The performance of the proposed framework in terms of (a) Estimated Infection Ratio (the true value is 20%) and (b) Bias.

2) Convergence analysis: Deriving valid population estimates from sampled nodes is based on the assumption that the samples are derived from the equilibrium distribution, which is true asymptotically. In this section, we study the convergence of the DNM estimates to equilibrium distribution during the data collection step.

In general, the starting nodes (the seeds) are not selected from a sampling frame, but instead are ad-hoc samples. One way to reduce the dependence of final estimates to seed nodes is to use a burn-in period by discarding large numbers of initial sampled nodes before analyzing the collected data. Given the high cost in terms of time and effort of collecting data, this may not be a desirable approach. Moreover, the authors in [38] demonstrate that in a without-replacement sampling setting, this approach can even introduce more bias. In fact, the only real way to apply a burn-in to ensure accuracy would be to repeat the burn-in after every sampled node.

An alternative approach would be to estimate the relative visiting probabilities of all sampled nodes, conditioned on the composition of the seeds, and compute the estimates based on those probabilities. We follow this approach in the proposed framework. In particular, since jumping in personalized PageRank is limited to the seed vector, we approximate the importance of every node to the seed, i.e. the visiting probability, which is used in the Estimation step to correct the bias to the seeds.

To monitor the convergence of the proposed framework, we use a standard diagnostic test developed within the MCMC literature, namely Geweke [39]. This test was applied for

the first time in the context of network sampling in [10]. The Geweke diagnostic detects the convergence of a single Markov chain by comparing the location of the sampled parameter on two different time intervals of the chain. The test is a standard Z-score with the standard errors adjusted for autocorrelation. We used this diagnostic in different runs of the proposed framework (generated by selecting some random nodes as seed) for the OSN dataset, and compares the difference between the first 10% and the last 50% of the samples. Figure 3 presents the results of the Geweke diagnostic for the infection ratio as a network characteristic. We observe that after sampling approximately 200 nodes, the Z-scores are strictly between $[-1, 1]$. This indicates that the proposed framework has achieved a good mixing with our initial selection of the random seeds.
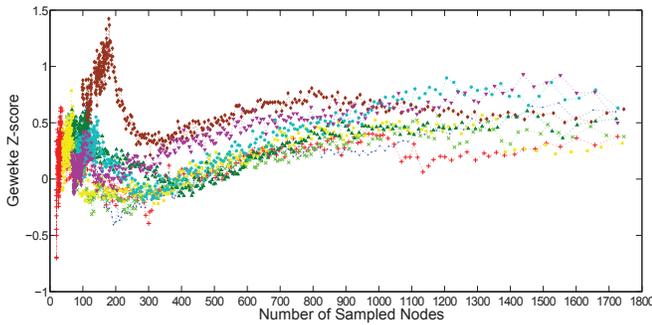

Fig. 3. The Geweke convergence diagnostic of DNM for infection ratio in the OSN dataset. The lines show the Geweke Z-score for some runs with different random nodes as seed.

3) The effect of network reciprocity: Here, we study the performance of the proposed framework in networks with different levels of reciprocity r which is an important parameter in characterization of directed networks. To this end, we generate some underlying directed networks with $r \in \{0.6, 0.7, 0.8, 0.9\}$ by DER model. As we can see the results in Figure 4, in lower sampling rates, the amount of error in estimation of infection ratio decreases with decreasing reciprocity (we observe the same pattern in terms of bias). As a general conclusion, if we have to sample a few nodes from a network, lower proportion of reciprocal links in that network leads to more accurate outputs.

4) Sensitivity to true infection ratio: Figure 5 shows the sensitivity of DNM to infection ratio in DER, DWS, DSF, and OSN datasets (due to page limitation, we did not demonstrate the results for P2P and AS networks). We run the simulation of disease spreading until the true ratio of infected nodes reaches to desired values (specifically, 20%, 30%, 40%, and 50%). These ratios are used as the ground truth to performance evaluation.

According to the results, one can categorize the networks into three groups. The first group includes the networks generated by the DER model. In this group, the estimated infection ratio is more farther from the true value by increasing the true infection ratio. This error converge to zero for higher sampling rate in various infection ratios (In terms of bias, we observe the same pattern; higher infection ratio leads to higher
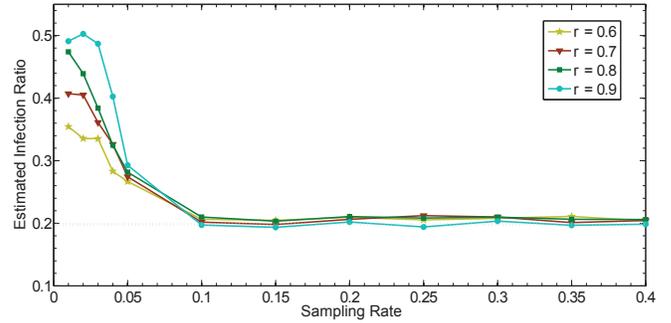

Fig. 4. The effect of reciprocity (r) on the performance of DNM in underlying networks generated by DER model. True value of infection ratio $(= 0.2)$ shown in the grey dotted line.

bias in lower sampling rates). In second group, that includes the DWS networks, the infection ratio dose not have major impact on the estimated values and their biases.

We found an inverse pattern in third group which is composed of the DSF and OSN (we observe the same results in the P2P and AS networks). In particular, lower true infection ratio leads to higher error in estimation of infection ratio (similarly, for obtained bias) for lower sampling rates. However, this error reduces by increasing the number of samples. The deduction rate is higher for the populations with lower infection ratio. As a general conclusion, in a scale free network with higher infection ratio, lower sampling rates are sufficient and result in a reasonable bias. More specifically, increasing the sampling rate does not have significant impact on the bias.

5) Estimating Outdegree Distributions: In what follows we compare the performance of DNM against the DURW sampling algorithm proposed by Ribeiro et al. [13] (presented in Section II). This is the most closely related method to ours that achieves asymptotically unbiased estimates of the outdegree distribution of a directed graph. However, there is a fundamental difference between the proposed framework and DURW; DNM only uses already visited nodes as local information without any prior knowledge about the latent structure of the network. In addition, the DURW algorithm is based on the assumption that nodes can be sampled uniformly at random from the original graph, which is not always feasible.

There are two controlling parameters in DURW. The first one is denoted by c, which is the cost of a random jump (i.e. the average number of sampling steps required to perform the jump). The second one, w, is the random jump weight that controls the probability of performing a random jump. Figure 6 shows the comparison between estimates of the bias (in log scale) of DNM and DURW for all outdegrees in the OSN dataset (we observed the same results for other test networks). The bias was obtained over 100 runs. In all simulations we sampled 10% of the network. The DURW random jump weight and cost was $w = 10$ and $c \in \{1, 10, 50\}$, respectively.

We found that both methods obtain accurate estimates. However, the fraction of nodes with large outdegrees can be estimated more accurately than the ones with small outdegrees. This is because both of them tend to sample nodes with larger
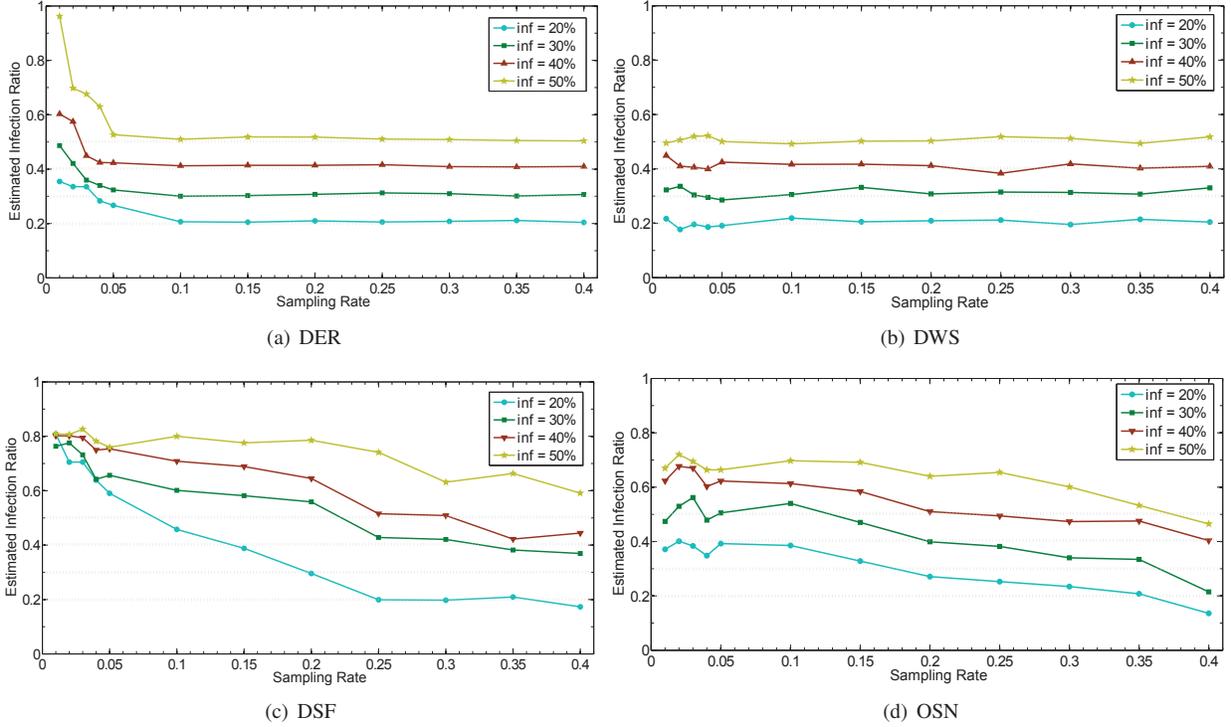
Fig. 5. The performance of the DNM framework in the networks with different infection ratios; (a) DER, (b) DWS, (c) DSF, and (d) OSN.

outgoing links more frequently; sometimes causing lower estimation errors for the large outdegree nodes. Moreover, as demonstrated in Figure 6, DURW outperforms DNM in terms of bias when the cost of a random jump is not significant. The cost of a jump effectively reduces the number of total nodes that can be sampled, which increases the bias. Therefore, we observe that the bias of the outdegree distribution estimates in DURW increases with c, and the DNM method is more accurate than DURW for larger costs.
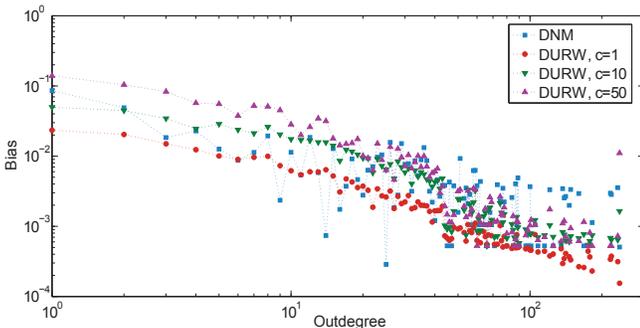

Fig. 6. Bias of DNM compared with DURW ($w = 10$ and $c \in \{1, 10, 50\}$) for all outdegrees in OSN dataset for sampling rate= 10%.

VI. CONCLUSION

In this paper, we introduced a novel two-step framework to measure nodal characteristics of large scale directed networks which can be defined by average target functions. We proposed a novel link-tracing network sampling algorithm by utilizing

the idea of personalized PageRank. In particular, this method samples the underlying network by moving from a node to one of its neighbors through an outgoing link based on the approximated probability of Personalized PageRank. Since these probabilities can be considered as an approximation of the exact visiting probability, we proposed a new estimator based on the idea of approximate importance sampling. Our estimator is able to overcome the effect of approximate stationary distribution of random walk-based sampling methods on the accuracy of the estimation. We showed both analytically and empirically that the proposed framework is asymptotically unbiased.

REFERENCES

[1] M. Newman, A. Barabasi, and D. J. Watts, "The structure and dynamics of networks," *IEEE J. Sel. Areas Commun. on Measurement of Internet Topologies*, vol. 107, no. 24, p. 624, 2006.
[2] O. Frank, "Survey sampling in networks," in *Handbook of Social Network Analysis*, J.Scott and P.Carrington, Eds., 2011.
[3] D. D. Heckathorn, "Respondent-driven sampling: a new approach to the study of hidden populations," *Social problems*, vol. 44, pp. 19–24, 1997.
[4] M. Hansen and W. Hurwitz, "On the Theory of Sampling from Finite Populations," *Annals of Mathematical Statistics*, vol. 14, no. 3, 1943.
[5] E. Volz and D. Heckathorn, "Probability based estimation theory for respondent-driven sampling," *Official Statistics*, vol. 24, p. 79, 2008.
[6] M. Kurant, A. Markopoulou, and P. Thiran, "Towards unbiased bfs sampling," *JSAC special issue on Measurement of Internet Topologies*, vol. 29, no. 9, 2011.

[7] L. Lovasz, "Random Walks on Graphs: A Survey," *Combinatorics,*, vol. 99, pp. 1–46, 1993.

[8] M. Salehi, H. R. Rabiee, and A. Rajabi, "Sampling from complex networks with high community structures," vol. 22, no. 023126, 2012.

[9] M. Gjoka, C. T. Butts, M. Kurant, and A. Markopoulou, "Multigraph sampling of online social networks," *IEEE J. Sel. Areas Commun. on Measurement of Internet Topologies*, vol. 389, no. 23, 2011.

[10] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, "Practical Recommendations on Crawling Online Social Networks," *IEEE J. Sel. Areas Commun. on Measurement of Internet Topologies*, vol. 29, no. 9, pp. 1872–1892, 2011.

[11] B. Ribeiro and D. Towsley, "Estimating and sampling graphs with multidimensional random walks," in *Proceedings of IMC*, 2010, pp. 390–403.

[12] X. Lu, L. Bengtsson, T. Britton, M. Camitz, B. J. Kim, A. Thorson, and F. Liljeros, "The Sensitivity of Respondent-driven Sampling Method," *Journal of the Royal Statistical Society Series A Statistics in Society*, vol. 175, pp. 191–216, 2011.

[13] B. Ribeiro, P. Wang, F. Murai, and D. Towsley, "Sampling Directed Graphs with Random Walks," in *Proceedings of IEEE INFOCOM*, 2012, pp. 1692–1700.

[14] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Computer Networks and ISDN Systems*, vol. 30, no. 1-7, pp. 107–117, Apr. 1998.

[15] G. Jeh and J. Widom, "Scaling personalized web search," in *Proceedings of the conference on World Wide Web*, 2003, pp. 271–279.

[16] M. J. Salganik and D. D. Heckathorn, "Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling," *Sociological Methodology*, vol. 34, no. 1, pp. 193–240, 2004.

[17] J. S. Liu, *Monte Carlo Strategies in Scientific Computing*. Springer, Jan. 2001.

[18] M. R. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork, "On near-uniform URL sampling," in *Proceedings of the World Wide Web conference on Computer networks*, 2000, pp. 295–308.

[19] D. Stutzbach, R. Rejaie, N. Duffield, S. Sen, and W. Willinger, "On Unbiased Sampling for Unstructured Peer-to-Peer Networks," in *Proceedings of IMC*, 2008, pp. 27–40.

[20] M. Eslami, H. R. Rabiee, and M. Salehi, "Diffusion-Aware Sampling and Estimation in Information Diffusion Networks," in *Proceedings of ASE/IEEE Social Computing*, 2012, pp. 176–183.

[21] M. Eslami, H. R. Rabiee, and M. Salehi, "Sampling from Diffusion Networks," in *Proceedings of ASE Social Informatics*, 2012.

[22] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculation by fast computing machines," *Chemical Physics*, vol. 21, no. 6, pp. 1087–1092, 1953.

[23] S. Chib and E. Greenberg, "Understanding the Metropolis-Hastings Algorithm," *The American Statistician*, vol. 49, no. 4, pp. 327–335, 1995.

[24] M. Salehi, H. R. Rabiee, N. Nabavi, and S. Pooya, "Characterizing Twitter with Respondent-Driven Sampling," in *International Workshop on Cloud and Social Networking (CSN2011) in conjunction with SCA2011*, vol. 29, no. 9, 2011, pp. 5521–5529.

[25] C. Hall and A. Carzaniga, "Doubly stochastic converge: Uniform sampling for directed P2P networks," Faculty of Informatics, University of Lugano, Technical Report 2009/02, Jan.

[26] T. Haveliwala, "Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 4, pp. 784–796, Jul. 2003.

[27] P. Berkhin, "Bookmark-Coloring Algorithm for Personalized PageRank Computing," *Internet Mathematics*, vol. 3, pp. 41–62, 2006.

[28] D. Gleich and M. Polito, "Approximating Personalized PageRank with Minimal Use of Web Graph Data," *Internet Mathematics*, vol. 3, no. 3, pp. 257–294, Jan. 2006.

[29] G. H. Golub and C. F. Van Loan, *Matrix computations (3rd ed.)*. Baltimore, MD, USA: Johns Hopkins University Press, 1996.

[30] Z. Bar-Yossef and M. Gurevich, "Efficient Search Engine Measurements," *ACM Transactions on the Web*, vol. 1, no. 1, p. 401, 2011.

[31] P. Erdös and A. Rényi, "On random graphs, I," *Publicationes Mathematicae (Debrecen)*, vol. 6, pp. 290–297, 1959.

[32] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, Jun. 1998.

[33] D. M. Pennock, G. W. Flake, S. Lawrence, E. J. Glover, and C. L. Giles, "Winners don't take all: Characterizing the competition for links on the web," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 8, pp. 5207–5211, Apr. 2002.

[34] T. Opsahl and P. Panzarasa, "Clustering in weighted networks," *Social Networks*, vol. 31, no. 2, pp. 155–163, 2009.

[35] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graph evolution: Densification and shrinking diameters," *ACM Trans. Knowl. Discov. Data*, vol. 1, no. 1, pp. 2+, Mar. 2007.

[36] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations," in *In ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2005, pp. 177–187.

[37] Y. Moreno, R. Pastor-Satorras, and A. Vespignani, "Epidemic outbreaks in complex heterogeneous networks," *The European Physical Journal B*, vol. 26, no. 4, pp. 521–529, Apr. 2002.

[38] K. Gile and M. S. Handcock, "Respondent-Driven Sampling: An Assessment of Current Methodology," *Sociological Methodology*, vol. 40, pp. 285–327, 2010.

[39] J. Geweke, "Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments," in *In Bayesian Statistics*. University Press, 1992, pp. 169–193.

Mostafa Salehi received his B.Sc. (2005) degree in Computer Engineering from Iran University of Science and Technology, and his M.Sc. (2008) and Ph.D. (2012) degrees in Computer Architecture from Amirkabir University of Technology and Sharif University of Technology, respectively. He was also involved in a number of large scale network design and consulting projects in the telecom industry. His research interests include graph sampling in online social networks, study of partially-observed complex network structures, and multimedia networks.

Hamid R. Rabiee (SM'07) received his BS (1987) and MS (1989) degrees in Electrical Engineering from CSULB, his EEE degree in Electrical and Computer Engineering from USC, and his PhD (1996) in Electrical and Computer Engineering from Purdue University, West Lafayette, USA. From 1993 to 1996 he was a Member of Technical Staff at AT&T Bell Laboratories. From 1996 to 1999 he worked as a Senior Software Engineer at Intel Corporation. He was also with PSU, OGI and OSU universities as an adjunct professor of Electrical and Computer Engineering from 1996-2000. Since September 2000, he has joined Sharif University of Technology, Tehran, Iran. He is the founder of Sharif University Advanced Information and Communication Technology Research Center (AICT), Advanced Technologies Incubator (SATI), Digital Media Laboratory (DML), and Mobile Value Added Services Laboratory (VASL). He is currently a Professor of Computer Engineering at Sharif University of Technology, and the Director of AICT, DML, and VASL. He has been the initiator and director of international and national level projects in the context of UNDP International Open Source Network (IOSN), and Iran National ICT Development Plan. He has received numerous awards and honors for his industrial, scientific and academic contributions. He has acted as chairman in a number of national and international conferences, and holds three patents.