Sampling from Complex Networks with high Community Structures

Mostafa Salehi,[a] Hamid R. Rabiee,[b] and Arezo Rajabi[c]

Digital Media Lab, AICTC Research Center, Department of Computer Engineering, Sharif University of Technology, Tehran, Iran

(Dated: 10 March 2012)

In this paper, we propose a novel link-tracing sampling algorithm, based on the concepts from PageRank vectors, to sample from networks with high community structures. Our method has two phases; (1) Sampling the closest nodes to the initial nodes by approximating personalized PageRank vectors, and (2) Jumping to a new community by using PageRank vectors and unknown neighbors. Empirical studies on several synthetic and real-world networks show that the proposed method improves the performance of network sampling compared to the popular link-based sampling methods in terms of accuracy and visited communities.

Keywords: Complex Network, Social Network, Link-tracing, Sampling, Community, PageRank

To characterize large-scale complex network such as online social interactions, where the global structure is initially unknown, one should study the collected network data by link-tracing sampling methods[1]. In these methods, one can see the neighbors of already sampled nodes and make a decision on which nodes to visit next. However, recent research[2,3] indicates that community structures[4] in a network, which correspond to densely connected groups of nodes with only sparser connections between groups, lead to bias in the network analysis. To the best of our knowledge, no attempt has yet been made to propose a sampling method to solve this problem. In this paper, we propose a novel link-tracing sampling method that takes into account the community structures with limited information about the network. The main idea is based on the expansion of a given community (which is a node initially), to include a set of nodes with the highest proximity to the initial node by using the PageRank (a classic random walk on the graph with jumping). Simulations on artificial and real-world networks are carried out to evaluate the performance of the proposed method.

I. INTRODUCTION

Many real-world systems can be modeled as complex networks of interacting dynamical nodes. Internet, World Wide Web, social interactions and biological systems are only a few examples of such networks. Indeed, we live in a world of networks. A network is often represented as a graph with a set of nodes and links. In a social network, for instance, the nodes represent individuals and links may represent associations such as social interactions.

In recent years, a considerable amount of research has been done on characteristics of complex networks in various domains. These studies range from the calculation of simple measurements summarizing structural properties such as degree distribution[5] to the extraction of functional properties such as cooperativity[6]. Previous research has revealed that one of the common properties in many real-world complex networks is the community structure[7]. A network is said to have community structure if there exists densely connected groups of nodes, with only sparser connections between groups[4]. Communities are significant structures in the networks, as they correspond to real social groups, similarity, or a common function.

To extract useful knowledge from a network, one should study the collected network data. However, with the advancement of technology, today we are confronted with very large scale networks. For example, Twitter as the most popular micro-blogging Online Social Network (OSN), has attracted more than 200 million users worldwide. Moreover, for many networks, their global structure is initially unknown, i.e. there is no sampling frame (a list of all elements we can sample from). The aforementioned factors may prevent access to the entire data in large networks.

To solve this problem, one can use network sampling to infer the characteristics of the original network using small number of samples (subsets of nodes and links) from the network. Therefore, the data that most previous studies use to study the real-world networks were collected by a sampling method[8]. Since the gathered data is often incomplete, studying the accuracy of sampling methods is necessary to perform accurate network analysis. In the lack of a sampling frame, link-tracing sampling methods are the only feasible solutions. In these methods, one can observe the neighbors of already sampled nodes and make a decision on which nodes to visit next. Snowball[1], Classic Random Walk[9], MHRW (Metropolis-Hastings Random Walk)[10] and RDS (Respondent Driven Sampling)[11] are examples of link-tracing methods.

The essential property of a sampling method that makes it appropriate for network inference is that its vis-

[a] Electronic mail: mostafa_salehi@ce.sharif.edu
[b] Electronic mail: rabiee@sharif.edu
[c] Electronic mail: rajabi@ce.sharif.edu

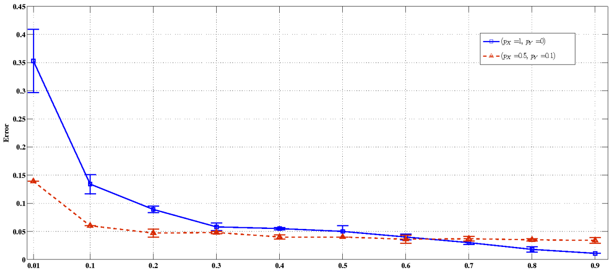


FIG. 1: The effect of community structure on classic random walk; Number of nodes = 200, Sampling rate = 0.4, and Number of runs = 50.

iting probabilities should be known or predictable for all nodes. This allows sample data to be weighted so that they accurately represent the network data. A probabilistic sampling method not only makes inference possible; it also makes it possible to specify how inference uncertainty can be quantified. To this end, an estimator is used to obtain the network characteristics. An estimator is a function that takes a summary of the sampled data as input and provides an estimate of unknown parameters (refer to Section III D).

Recent studies[2,3] indicate that the accuracy of using a link-tracing method depends on the underlying complex structure of the network. In particular, they show that community structures lead to bias in network analysis if the initial nodes (seeds) are not selected randomly among all nodes and the number of visited nodes is not enough to provide uniform random samples. However, according to our assumptions about the scale of the network and absence of sampling frame, these conditions may not be satisfied. The effect of community structures on classic random walk as a well-known link-tracing method is illustrated by the following example (refer to Figure 1). Consider a set of nodes V consisting of two communities, X and Y, of equal size. Let links exist between every pair of nodes, however within-community links have weight $(1 - w)$ while between-community links have weight w where $0<w<1/2$. It's clear that w control the community structure; as w increases, the tendency for within-community links decreases. Let variable x is a node characteristic that is 0 or 1 and $x(v)$ is its value of node v. Let p_X and p_Y denote the proportion of nodes v within the two communities with $x(v) = 1$. Since $|X| = |Y|$, the proportion of nodes with $x(v) = 1$ in the entire network is $p = (p_X + p_Y)/2$. We measure the error by averaging the $|x - p|$, where x is the fraction of sampled nodes with value 1. As we can see in Figure 1, higher community structures lead to higher error.

To the best of our knowledge, no attempt has yet been made to propose a method for sampling the networks with considering community structures. In this paper, we present a sampling algorithm based on the PageRank[12,13], called PageRank-Sampling (PRS), to solve this problem. In a network, links between nodes can be interpreted as the node's importance, and PageRank can capture the relative importance of each node within the network by taking a sequence of random walk steps starting from the initial node(s). Here, we employ a modified version of PageRank, called personalized PageRank[14], using a specified set of nodes as a seed vector. Personalized PageRank determines the importance of every node to the seed by using local information and is used to solve the problem of local community detection[15,16]. Recently, fine-grained relationships between nodes that is necessary for applications such as ranking, was modeled by using Personalized PageRank[17].

The proposed method has two phases. Phase (1): Sampling the closest nodes to the seed node by approximating personalized PageRank vectors. In this phase, the idea is based on expansion of a community, which is a node initially, to include a set of nodes with the highest proximity to the seed nodes. Starting with an initial node, this iterative process is repeated until the specified number of nodes is collected from this community. Phase (2): Jumping to a new community by using PageRank vectors and unknown neighbors. The idea of this phase is to choose a node with lowest PageRank value and highest number of unknown neighbors, as a candid node that guides the walker to a new community. We empirically demonstrate our sampling method improves the performance of the network sampling, compared to typical link-based sampling methods in terms of accuracy and visited communities. Moreover, we study various aspects of the proposed method such as overhead and time complexity.

The rest of the paper is organized as follows. Section II presents related work in network sampling and local community detection. In this section, we summarize most popular link-tracing sampling techniques and various community detection methods with local information of graph. The preliminaries are given in Section III. Section IV presents the proposed algorithm. Section V provides the performance evaluation, and the concluding remarks are presented in Section VI.

II. RELATED WORK

Since in this paper we focus on solving the problem of network sampling by using local community detection approach, the related work in two areas should be considered: (1) Network Sampling, and (2) Local Community Detection.

A. Network Sampling

Network sampling methods are somewhat distinct from classical sampling methods. In general, there are two approaches for constructing and analyzing samples of networks in different fields; the graph sampling and the model-based sampling.

Graph sampling, can be categorized into two subgroups according to the amount of available information about the network graph[18]; the coarse graining methods and graph exploration methods. In the coarse graining methods[19], initially the graph of network can be observed (i.e., the sampling frame is available), and its focus is on development of a smaller sub-graph of a network such that the subgraph keeps at least some of the most relevant properties of the original network. In order to achieve this goal, most existing techniques are based on the idea of either grouping nodes together or removing some nodes. These techniques are very common in the statistical physics literature. On the other hand, reducing the network complexity by graph sampling may cause some information about the initial network to be lost. In the second sub-group, graph exploration techniques, the graph of network is unknown initially, except for some limited number of nodes (i.e., no sampling frame is available). The focus of these techniques is on deriving sample statistics for characteristics measured at each node of the sub-graph sample (e.g., node degree). Graph sampling techniques in this sub-group are based on the link-tracing methods. For unbiased estimation of the characteristics of an unknown network, traditional statistical inference has to be modified by assigning the appropriate weights to the values observed at each node in the sample. These techniques are very common in sociology and computer science.

In the model-based sampling[20], we initially consider a probability model with unknown parameters for the target network. The focus of the techniques in this approach is to use the observed samples from the network to estimate the model parameters, and to extract the network characteristics based on this model. The main challenge in this approach is to specify probability models that adequately incorporate structures of the target network. Model-based approaches are very common in mathematics and statistics.

In this paper, we assume that the graph of network is unknown initially and our goal is to extract sample statistics for measured characteristics. Therefore, our focus in this paper is on the graph exploration methods. Random Walk (RW)[9] is one of the most important and widely used exploration sampling methods in different kind of network contexts such as uniformly sampling Web pages from the Internet[21], content density in peer-to-peer networks[22,23], degree distributions of the Facebook social graph[24,25] and in general large graphs[18]. A classic RW samples a graph by moving from a node u, to a neighboring node v, through an outgoing link (u, v), chosen uniformly at random from the neighbors of node u. By this process links and nodes are sampled. The probability of selecting the next node determines the probability that nodes are sampled. In any given connected and non-bipartite graph G, the probability of being at a node u converges at equilibrium to the stationary distribution $\pi(u) = deg(u)/2|E|$, where $deg(u)$ and E are the degree of node u and are the set of links of the network graph. Thus, the classic RW is biased towards node with higher degree.

Metropolis Hasting technique[10] can be used to modify the probabilities of the node selection in RW in order to have the uniform stationary distribution for visiting each node, i.e. $\pi(u) = 1/|V|$ where V is the set of nodes. This technique is a general Markov Chain Monte Carlo method[26] for sampling from probability distributions based on constructing a Markov chain that has the desired distribution as its stationary distribution. This approach, known as MHRW (Metropolis-Hastings Random Walk)[10], has been applied to peer-to-peer networks[23], Facebook[24,25], and Twitter[27].

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. RDS (Respondent Driven Sampling)[11] method presents an approach to correct the bias[11]. In each iteration of the RDS method, as a Random Walk based sampling method, the walk selects randomly n neighbours of the current node v as the next hop node. When $n = 1$, RDS is equivalent to the classic random walk. In RDS, nodes with more neighbours in the graph of network are more likely to be selected[28]. Estimating the probability of visiting each node in RDS is based on a Markov Chain representation of the sampling process[2]. RDS method is an approach in the field of social sciences to sample and inference in hard-to-reach populations such as injection drug users[29]. In these populations, a sampling frame for the target population is not available. Sampling from complex network with no sampling frame is analogous to the sampling of hidden population in the social sciences. In the context of graph sampling, RDS method has been used for Facebook[24], a variety of synthetically generated graphs[30], and Peer-to-Peer networks[23].

Maiya et al.[31] consider sampling subgraphs to optimize the preservation of community structures based on concepts from expander graphs. They show that subgraph samples produced by the proposed methods are more representative of community structure in the original network. However, the selection procedure of next node in this work is not based on a mathematical framework and one can not compute the probability of visiting sampled nodes. Therefore, there is no explanation on how to correct the sampling bias.

Recently, the effect of community structure on the accuracy of sampling by using a link-tracing method has been studied[2,3]. However, to the best of our knowledge, no attempt has yet been made to propose a method for sampling the networks with high community structures. In this paper, we focus on this issue based on a concrete framework by considering the idea of PageRank in local community detection field. In the proposed method, we correct the sampling bias by using the approximated visiting probability of each sampled node.

B. Local Community Detection

Detecting communities in networks is very important to understand the structure, function and evolution in various areas (such as World Wide Web, social and biological networks). There are two different definitions for a community within a complex network. One definition is that a community is a group of nodes more strongly connected than would occur randomly[5,32]. In the second definition, it's a group of nodes in which there are more links between nodes within the group than those outside of it[33]. In this paper, we consider the latter definition.

Although the community detection in a network is a well-studied problem, it's still considered as an open problem. To solve this problem, many methods have been done in recent years that are based on different ideas and with various capabilities. One can refer to references[4,34] for a detailed review of the large body of work in this area. In general, community detection methods can be divided in two categories; Global and Local approaches. The methods in the former approach[35–37] require the knowledge of the entire structure of the network, which is not feasible for some large-scale and dynamic networks (such as current online social networks and Peer-to-Peer networks).

On the other hand, the main goal of the methods in the local approach is identifying community structures by focusing on a portion of the network under study. Therefore, local community detection methods[38–42] provide a means to relieve scalability challenges and lack of global information about the network. In practice, such methods start a network exploration process from a single (or a small number of) seed node(s) and add an appropriate node from the neighborhood nodes to the community by considering some quality measures. This process continues until it is not possible to identify any other node that should be added to the community. Most of the above methods only detect non-overlapping communities, i.e., a node can only belong to one community. Recently, Lancichinetti et al.[43] proposed a local community detection method, called the Order Statistics Local Optimization Method (OSLOM), that automatically detects overlapping communities and hierarchies of communities. This method optimizes a local fitness function that measures the statistical significance of the detected communities compared, with respect to a global randomized null model.

Andersen et al.[15] study the problem of expanding an initial set into a community. To this end, they compute the probability distribution of random walk steps starting from the seed nodes. The idea of this paper is to sweep across the sorted nodes by a degree-weighted probability and select the nodes of least conductance as the community. In another paper[16], the authors propose a local spectral partitioning algorithm, called PageRank-Nibble, which computes a personalized PageRank vector and chooses the least conductive set as the community involving initial set. A personalized PageRank vector computes the stationary distribution of random walks starting from the input seeds. In this paper, we use the idea of PageRank to perform the local community detection in the first phase of the proposed sampling method and iteratively expand an initial node to a community.

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 is the set of unweighted undirected links between pairs of nodes. Let A denote the adjacency matrix of G, where $A(u, v) = 1$ if and only if there is an link between u and v, otherwise $A(u, v) = 0$. For a node $v \in V$, let $deg(v)$ denote the degree of v. let $D = diag(deg(1), ..., deg(n))$ be the diagonal degree matrix.

Let S be a sample of nodes where $S \subset V$ and $G(S)$ is the induced subgraph of G based on the sample set S. That is, $G(S) = (S, E_S)$ where the link set $E_S = (S \times S) \cap E$. Let $N(S)$ denote the neighborhood of S;

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

Let $UN(v)$ denote the set of new neighbors contributed by v (the unknown neighbors of v);

$$UN(v) = N(v) - S \tag{2}$$

Let the volume of the subset S be $vol(S) = \sum_{v \in S} deg(v)$. It's clear that $vol(V) = 2m$. We write $support(p)$ to denote the set of nodes on the vector p which it is nonzero, i.e., $support(p) = \{v|p(v) \neq 0\}$. Here, the edge border is denoted by $\partial(S) = \{\{u, v\} \in E|u \in S, v \notin S\}$, and the conductance of a set of nodes is given by;

$$\phi(S) = \frac{|\partial(S)|}{min(vol(S), 2m - vol(S))} \tag{3}$$

B. Problem Formulation

Our primary goal in this paper is to consider community structures in developing an accurate network sampling method. We represent community structures in a network G as a group of disjoint subsets whose union is the set V. Specifically, suppose V is partitioned into a set $C = \{C_1, C_2, \dots C_r\}$ of r non-overlapping communities, with union $\bigcup_{C_i \in C} C_i = V$. We select uniformly at random n_i nodes from community C_i, as long as $n_s = \sum_i n_i$, where n_s is the total number of sampled nodes.

Although there are many possible allocations n_i, we are interested in the optimal allocation that minimizes the measurement error. Kurant et al.[44] show that if one is interested in comparing the network properties such as average node degree, in a sampled network rather than the entire network, an equal number of samples from every community is needed (i.e. $n_i = \frac{n_s}{r}$), in order to obtain the optimal allocation. In this paper, we impose the same requirement on the sampled subgraph S, by selecting equal number of nodes from communities in the original network.

C. Approximate Personalized PageRank Vectors

PageRank was first introduced by Brin and Page[12,13] for search algorithms in the graph of WWW network. However, it can be defined for any graph as a stationary distribution of a random walk on 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. Traditionally, the value of α is chosen to be 0.15[45]. With smaller α, the random walks spread further from the initial nodes before performing a random jump.

The PageRank of G is defined as the vector $\mathbf{p}_\alpha(\mathbf{s})$ that is the solution of the following equation:

$$\mathbf{p}_\alpha(\mathbf{s}) = \alpha\mathbf{s} + (1-\alpha)\mathbf{p}_\alpha D^{-1}A \tag{4}$$

Where $\mathbf{s}$ is a seed vector that includes initial distribution for starting nodes. 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[46]. It can be shown that for any starting vector $\mathbf{s}$, and any constant α, there is an unique vector $\mathbf{p}_\alpha(\mathbf{s})$ satisfying Equation (4)[14]. Moreover, we write $\mathbf{p}_\alpha(\mathbf{s}, u)$ If only uth element of $\mathbf{s}$ is 1, and 0 otherwise[46,47].

It is infeasible to calculate exact PageRank vectors for large-scale networks by solving the PageRank Equation (4). The algorithm $ApproximatePR(\mathbf{s}, \epsilon)$ introduced by Andersen et al.[48] can be used to compute ϵ-approximate PageRank vectors with a starting vector $\mathbf{s}$. An ϵ-approximate PageRank vectors for $\mathbf{p}_\alpha(\mathbf{s})$ is a PageRank vector $\mathbf{p}_\alpha(\mathbf{s} - \mathbf{r})$ where the vector $\mathbf{r}$ is non-negative and satisfies $r(u) \leq \epsilon.deg(u)$ for every node u in the network. The approximation error of an ϵ-approximate PageRank vector on any set S is bounded by $\epsilon.vol(S)$[48].

The analysis of $ApproximatePR(\mathbf{s}, \epsilon)$ leads to the following conclusion[48]:

Theorem 1 *for any seed vector* $\mathbf{s}$ *with* $\|\mathbf{s}\|_1 \leq 1$*, and any constant* $0 < \epsilon \leq 1$*, the algorithm computes an* ϵ*-approximate PageRank vector* $\mathbf{p}$ *for* $\mathbf{p}_\alpha(\mathbf{s})$*. The support of* $\mathbf{p}$ *satisfies* $vol(support(\mathbf{p})) \leq \frac{2}{(1-\alpha).\epsilon}$*, and the running time of the algorithm is* $O(\frac{1}{\epsilon.\alpha})$*.*

The proposed method uses $ApproximatePR$ algorithm as a subroutine.

D. Correcting the bias

The selection bias of a sampling method can be corrected by re-weighting of the measured values. This can be done using the Hansen-Hurwitz estimator[49], i.e. nodes are weighted inversely proportional to their visiting probability.

Almost all network characterization metrics we are aware of can be expressed as sums or averages of some functions. For example, the network size, $|V|$, is the sum of the constant 1 function ($f(u) = 1$ for all u). The number of copies of a file in a peer-to-peer network is the sum of the f ($f(u) = 1$, if peer u has a copy of that file, and $f(u) = 0$, otherwise). The density of nodes with in-degree $j \in L$ is the average of the function $f_j(u)$: $avg(f_j) = sum(f_j)/|V| = \sum_{u \in V} f_j(u)/|V|$, where $f_j(u) = 1$, if in-degree of node u is j, and $f_j(u) = 0$, otherwise.

For any function $f : V \to \mathbb{R}$ that defines a nodal characteristic, the estimator of Equation 5 provides an asymptotic estimate of the population mean μ of f under some assumptions[50].

$$\hat{\mu} = \frac{\sum_{i=0}^{n-1} \frac{f(X_i)}{\pi(X_i)}}{\sum_{i=0}^{n-1} \frac{1}{\pi(X_i)}} \tag{5}$$

Where X_i is visited node on i^{th} draw. All the results in this paper are presented after this re-weighting step, whenever necessary.

IV. THE PROPOSED METHOD

The pseudo code of proposed method, called PageRank-Sampling (PRS), is shown in Algorithm 1. This method includes two phases. In the first phase (inner while loop), desired number of nodes, NC, are sampled from the current community, C_i. In the second phase (outer while loop), a new community is selected.

Initially we randomly select a node v and add it to the sample set S. Subsequent nodes of the sample are chosen based on the following procedure. First, $N(S)$, the neighbor set of S, is created. Then, by using $ApproximatePR(\mathbf{s}, \epsilon)$ algorithm[48], approximate personalized PageRank vector for this set is computed with initiate vector $\mathbf{s} = [s_1, s_2, s_3, \cdots, s_n]$ such that $s_i = \frac{1}{|S|}$ if $s_i \in S$, otherwise 0. Then, a node v with maximum PageRank value which is not selected yet, is chosen. According to the concept of PageRank, this node has the highest probability among all non-visited nodes to be in the same community of initial node. This procedure is repeated until the desired number of samples, NC, are collected from each community.

Then, in the second phase, PRS jumps to another community. To this end, PRS computes PageRank values

Algorithm 1 $S = PageRank - Sampling(G, k, t)$

```
Input :
$G = (V, E)$, Neighborhood graph of the network
$k$, the sampling size
$t$, the maximum number of communities in $G$
Output :
$S$, sampled subgraph

$C_i = \varnothing, i = [1 \cdots t]$
$i = 1$
$NC = \lfloor k/t \rfloor$
$v = random(V)$
$S = v$
while $i < t$ do
    while $|C_i| < NC$ do
        $\mathbf{s} = [s_1, s_2, s_3, \cdots, s_n] \quad s.t. \quad s_i = \frac{1}{|S|} \quad if \quad s_i \in S$
        $P = ApproximatePR(\mathbf{s}, \epsilon)$
        Select new node $u \in N(S)$ with maximum value in $P$
        $C_i = C_i \cup u$
        $S = S \cup u$
    end while
    $i = i + 1$
    Select new node $u \in N(S)$ based on minimization of $\frac{P(u)}{UN(u)}$
end while
```

for all gathered samples and their neighbors. Then, the node that minimize $\frac{P(u)}{UN(u)}$ is selected. This node is an initiator node for a new community. Note that a node v with lower PageRank value and higher number of unknown neighbors, i.e. $arg\max_{v \in N(v)} UN(v)$, may lead to a new community[31].

We use the Protien network (refer to SectionV A for details) shown in Fig. 2 as a running example to illustrate the effectiveness of the proposed method. The OSLOM algorithm[43] is employed to extract the communities of the original network. The sampling set with parameters $k = 18$ and $t = 6$ is given by

$S = \{\{21, 16, 20\}, \{74, 73, 84\}, \{12, 11, 1\}, \{52, 57, 59\}, \{60, 67, 63\}, \{66, 89, 90\}\}$

As we mentioned, 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. This allows sample data to be weighted such that it accurately represent the network data (refer to Section III D). We use the approximation of visiting probabilities computed by $ApproximatePR(\mathbf{s}, \epsilon)$ to correct the bias of PRS method.

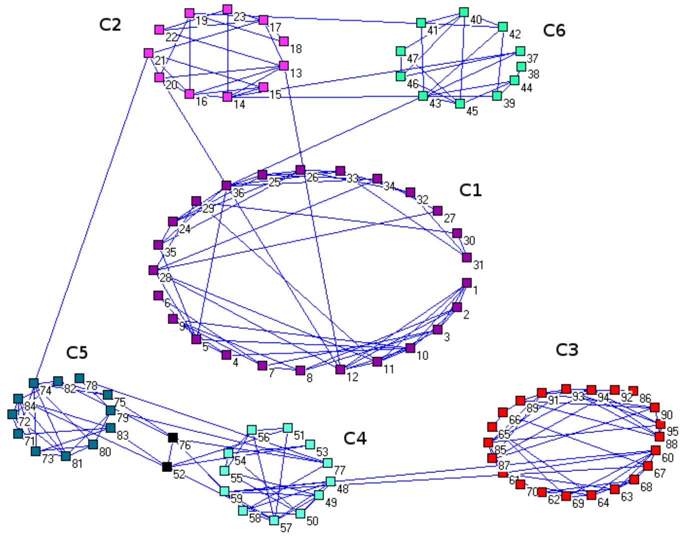


FIG. 2: The extracted communities from the Protien network by using the OSLOM algorithm. Black points represent overlapping nodes. Specifically, Nodes 52 and 76 are overlapping nodes that shared between the communities 4 and 5.

V. EXPERIMENTAL EVALUATION

A. Datasets

To study the performance of proposed method, we consider both synthetic and real-world networks.

Synthetic Networks: We use a well-known benchmark, called LFR[51,52], for generating networks with various levels of community structure. This benchmark is a special case of the planted l-partition model with heterogeneous distributions of node degree and community size. The node degrees are distributed according to a power law with exponent τ_1; the community sizes also obey a power law distribution, with exponent τ_2. To control the level of community structures in the LFR benchmark, we use the mixing parameter μ, which expresses the ratio between the external degree of a node with respect to its community and the total degree of the node. A smaller μ results in networks with higher level of community structures. We assume, for simplicity, that μ is the same for all nodes.

Real-world Networks: Moreover, we consider some well-known real-world undirected and unweighted biological, social, and technological networks as test data; (1) Protein structure[53] and C.elegans Metabolic Network[54] as examples of biological networks (2) Jazz musicians network[55], Zachary's karate club[56], American College Football[57], and Dolphin Social Network[58] as examples of social networks (3) US Air Transportation Network (1997)[59] as an example of technological networks. Table I summarizes the details of these networks. All of these networks are downloadable from the Internet at the web sites provided by the authors of the original works.

In general, higher average conductances (over the whole network) shows lower level of community in the network. Lower conductance of a community rather than the entire network intuitively implies a bottleneck between the community and the rest of the network, and random walks of PageRank tend to stay within the community[60]. According to Table I, Protein and Dolphin have stronger levels of community (Their average conductance of communities equals to 0.15). Moreover,

TABLE I: The details of real-world test data. From left to right, the name of the network, the number of nodes (N) and edges (E), the average degree $< k >$, the number of communities (N_c), the average community size (S_c), and the average conductance of communities (ϕ_c). The communities are extracted by OSLOM algorithm. The corresponding standard deviations are presented in the parentheses.

Networks	N	E	$< k >$	N_c	S_c	ϕ
Protein	95	426	4.48 (1.45)	6	16.2(5.9)	0.15(0.08)
C.elegans	453	2025	8.94(16.76)	25	27.8(31.1)	0.56(0.16)
Jazz	198	2742	27.69(17.45)	10	22.6(9.6)	0.69(0.12)
Karate	34	78	4.59(3.88)	2	17.5(2.1)	0.22(0.02)
Football	115	613	10.66(0.89)	11	10.4(2.5)	0.48(0.08)
Dolphin	62	159	5.13(2.96)	2	36.5(7.8)	0.15(0.13)
Airport	332	2126	12.81(20.13)	8	42.0(58.4)	0.51(0.22)

the community sizes of Airport and C.elegans are widely different (The corresponding standard deviations are 58.4 and 31.1 repectively).

B. Experimental Setups

For each dataset described in the previous subsection, by considering the sampling rate, we select the nodes in the network using PRS and RDS sampling methods. Since, the proposed method is the only sampling algorithm that explicitly considers the community structure, we have compared our performance against RDS as a popular link-tracing based sampling method. Each method is repeated 100 times on each dataset. At each step, RDS method selects one random neighbor of the current nodes to be sampled. For the $ApproximatePR(\mathbf{s}, \epsilon)$ algorithm, we set the accuracy factor $\epsilon = 2^{24}$. Moreover, the results for PRS and RDS are presented after the reweighting step (refer to Section III D). In addition, we utilize the OSLOM algorithm to extract the communities of the original network. These communities are used as ground truth to evaluate the performance of the sampling methods.

To generate the test networks in LFR benchmark, we set the parameters of the graphs as follow: average degree $< k >= 20$, maximum degree $k_{max} = 50$, the mixing parameter $\mu = 0.1$, exponents of the power law distributions are $\tau_1 = 2$ for degree and $\tau_2 = 1$ for community size. Community sizes are in the range $[10, 50]$. We have averaged the value over 100 realizations for each value of the mixing parameter. Moreover, the values of the network size and mixing parameters are initialized according to the scenario of test.

TABLE II: Kolmogorov Smirnov D-Statistic for degree distribution in the PRS and RDS sampling methods compared to degree distribution of original network (Sampling rate=0.15%, $\alpha = 0.05$). Each entry in the table is obtained by averaging the D-statistic over 100 runs per dataset. The corresponding P-values are presented in the parentheses.

Network	PRS	RDS
Protein	0.36(0.05)	0.28(0.08)
C.elegans	0.06(0.01)	0.06(0.01)
Jazz	0.38(0.01)	0.35(0.02)
Karate	0.32(0.01)	0.35(0.04)
Football	0.18(0.05)	0.19(0.04)
Dolphin	0.44(0.09)	0.66(0.07)
Airport	0.15(0.01)	0.20(0.05)

C. Evaluation Results

1. Structural Properties

Structural properties have important effects on the behavior and dynamical properties of networks[5,6], and unless we know something about the structural properties of these networks, we cannot hope to adequately understand how the corresponding system works. The degree distribution, $P(k)$, is one of the most important structural properties that provides a natural summary of the connectivity in the graph of the network. More specifically, $P(k)$ is the fraction of nodes in a network with degree k.

We use the Kolmogorov-Smirnov D-statistic, D, to measure the agreement between the degree distribution in the original network and the sampled network. The D-statistic is a relative measure to compare the distribution of the values in the two datasets[61]. It is defined as $D = max|F'(x) - F(x)|$, where x is over the range of the interested characteristics; F and F' are the two empirical Cumulative Distribution Functions (CDFs) of the data. A value of $D \leqslant \beta$ corresponds to no more than β percentage point difference between CDFs. It is clear that smaller values of D indicate better fits of the sampling distribution to the distribution of original network.

As our results in Table II show, the proposed method nearly provides similar estimation of the degree distributions compared to RDS method. Since the degree distribution is independent of the community concept, it is not strange that we could n't find significant improvment by using the PRS. Notice that degree distribution is best matched in the C.elegans network. Moreover, since both of the sampling methods are based on the random walk (that is biased towards node with higher degree), the sampling distribution can be fitted better to the distribution of the networks with higher average of degree (such as Airport, and Football).

2. Accuracy and Expansion

In the following, we describe the evaluation criteria that is employed to study the performance of the PRS method. The accuracy measure is defined as:

$$Accuracy = 1 - \frac{1}{2p} \sum_{i=1 \; to \; p} |(S_i - S_c)/S_c| \qquad (6)$$

Where p is the number of visited communities. $|S_i|$ and $|S_c|$ are the number of sampled nodes from community i and the number of nodes which must be sampled from each community (i.e. $NC = \lfloor k/t \rfloor$), respectively. Since in computing accuracy, the error resulted from sampling incorrect node is considered twice (for its actual community and the incorrect chosen community), we consider the factor 2 in the denominator. Moreover, we assume that each overlapping node contribute the same amount to its communities. Accuracy measure ranges from 0 to 1 with higher values being better. For example, applying the proposed method on Protein network (refer to Table IV),the extracted node from each communities is: $C_1 = \{1, 11, 12\}$, $C_2 = \{16, 20, 21\}$, $C_3 = \{63, 60, 67, 90, 89, 66\}$, $C_4 = \{52, 57, 59\}$, $C_5 = \{52, 73, 74, 84\}$, and $C_6 = \{\}$. Moreover, node 52 is an overlapping node that belongs to communities C_4 and C_5. Then, the accuracy in this example can be computed as follows

Accuracy $= 1 - \frac{1}{2\times 6}(|\frac{3-3}{3}| + |\frac{3-3}{3}| + |\frac{6-3}{3}| + |\frac{2.5-3}{3}| + |\frac{3.5-3}{3}| + |\frac{0-3}{3}|) = 0.81$

We must not only consider accuracy when evaluating our sampling method, but also consider the number of communities visited in the sample (called *expansion* measure). By considering the both measures, we have a better analysis on the performance of sampling method. Specifically, we compute the expansion of sampling method by calculating the fraction of total communities in the original network represented in each sample, i.e. $\frac{p}{q}$ where p and q is the number of visited communities and the number of communities in original network, respectively. The range of expansion is from 0 to 1 with higher values being better. In the example of applying the proposed method to Protein network (refer to Table IV), the expansion measure is equal to $\frac{5}{6} = 0.83$.

The accuracy and expansion of the PRS and RDS methods in LFR benchmark is shown in Table III. As shown, the proposed method outperforms the RDS method in all the test networks. It improves the accuracy of sampling by 2% to 25% and the expansion by 3% to 49%. The results show that our method is more efficient than RDS method for sampling from the networks with higher level of community structure, i.e. lower μ.

The accuracy and expansion of the PRS and RDS methods in the real-world networks is shown in Table IV. The most improvement in accuracy and expansion is obtained for the Football network that equals to 43% and 38% respectively. Moreover, the Karate has the maximum values of accuracy and expansion (94%, 100%), and the minimum accures in the Protein network (42%, 43%). Notice that the sampling rate has significant role in the analysis the results. Therefor, in next section we address this issue.

TABLE III: Accuracy and Expansion of the sampling methods in the LFR benchmark. The parameters are: network size $N = 1000$, mixing parameter $\mu = 0.1$, sampling rate= 15%, $\alpha = 0.05$. The corresponding standard deviations are presented in the parentheses.

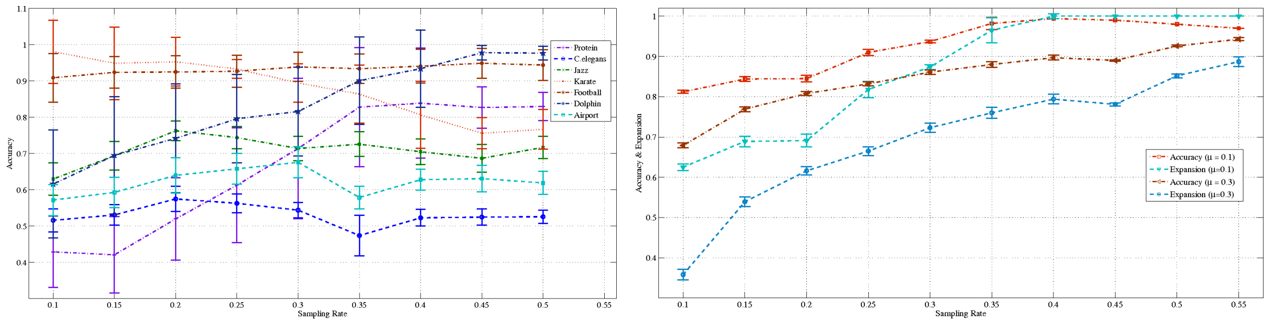
μ	Accuracy		Expansion	
	PRS	RDS	PRS	RDS
0.1	0.83 (0.02)	0.58 (0.02)	0.66 (0.05)	0.17 (0.04)
0.2	0.78 (0.02)	0.61 (0.03)	0.56 (0.04)	0.23 (0.06)
0.3	0.78 (0.01)	0.66 (0.03)	0.55 (0.03)	0.31 (0.06)
0.4	0.79 (0.01)	0.67 (0.03)	0.58 (0.03)	0.35 (0.07)
0.5	0.83 (0.03)	0.71(0.03)	0.66 (0.06)	0.42 (0.07)
0.6	0.86 (0.03)	0.74 (0.04)	0.72 (0.06)	0.48 (0.08)
0.7	0.83 (0.03)	0.76 (0.04)	0.67 (0.06)	0.52 (0.07)
0.8	0.84 (0.03)	0.78 (0.04)	0.68 (0.05)	0.57 (0.07)
0.9	0.81 (0.03)	0.79 (0.04)	0.62 (0.07)	0.59 (0.07)

TABLE IV: Accuracy and Expansion of the sampling methods in the real-world networks. The parameters are: sampling rate= 15%, $\alpha = 0.05$. The corresponding standard deviations are presented in the parentheses.

Network	Accuracy		Expansion	
	PRS	RDS	PRS	RDS
Protein	0.42 (0.09)	0.28 (0.09)	0.43 (0.10)	0.30 (0.11)
C. elegans	0.53 (0.03)	0.45 (0.04)	0.58 (0.04)	0.53 (0.07)
Jazz	0.70 (0.04)	0.56 (0.08)	0.87 (0.06)	0.74 (0.13)
Karate	0.94 (0.08)	0.63 (0.16)	1.00 (0.00)	0.75 (0.25)
Football	0.92 (0.05)	0.49 (0.09)	0.93 (0.04)	0.55 (0.12)
Dolphin	0.72 (0.17)	0.58 (0.10)	0.92 (0.18)	0.80 (0.25)
Airport	0.59 (0.04)	0.40 (0.08)	0.85 (0.06)	0.67 (0.14)

3. The effect of sampling rate

We study the effect of various sampling rates (10% to 55%) on the performance of the proposed method by considering the LFR benchmark and real-world networks. The results is shown in Fig. **??**. In general, since the concentration of our method is on sampling from core of a network, we have to collect a sufficient number of nodes to reach a new community. By focusing on the result of real-word networks, one can categorize the networks into three group; (1)Declining trend (2)Fixed trend, (3)Rising trend: such as Karate. The first group include the networks with overlapping nodes that have considerably higher degree, as compared with the average degree in the network. We called these nodes effective overlapping area. For example, the Karate network is in this group (refer to Fig.3a). It has an overlapping node 3 with degree 10 (the average degree is 4.59). The accu-



(a) Real-World Networks

(b) LFR Benchmark

FIG. 3: The effect of sampling rate on the performance of the PRS method. (a) Real-world networks (α=0.05). (b) LFR benchmark (network size $N = 1000$, mixing parameter $\mu = 0.1$, α=0.05).

racy decreases in this group by increasing the sampling rate.

The second group is composed of the networks with high heterogeneity in the sizes of their communities. The Airport and C.elegans can be considered in this group (refer to Fig.3a). The standard deviations of their community sizes are 58.4 and 31.1 repectively. For the networks in this group, increasing the sampling rate does not have considerable impact on the accuracy of the PRS method. This is because of that on the one hand higher sampling rate leads to sampling more nodes from larger communities. This makes decreasing the accuracy. On the other hand,

The networks in third group, rising trend, not only does not have effective overlapping area, but also their community sizes are not widely different. The Protein, Jazz, Football, and Dolphin as examples of this group (refer to Fig.3a). In these networks, the accuracy increases by increasing the sampling rate. This improvement is more considerable in the networks in which the average conductance is lower (refer to Table I for details of the network conductances). Lower conductances (that shows higher level of community in the network) leads to more accurate probability distribution of PageRank between members of each community. This makes more precise selection of nodes in a community, then takes a more accurate decision for jumping to next community performs more accurate. All of this causes the improvement in the accuracy of PRS method. As we can see in Fig. 3a, the Protein and Dolphin (their average conductance of communities equals to 0.15) has higher improvement. Moreover, our result in LFR benchmark (refer to Fig. 3b) confirm the correctness of this claim. For the lower value of mixing parameters μ, i.e. higher community level, the slope of rising trend of accuracy is greater.

4. The effect of α

At each step of PageRank, the random walk makes a jump to a random node with probability α. This jumping constant controls the rate of diffusion. When α becomes smaller, the random walks spread further from the initial nodes before returning via a random jump. The results of our study about the effect of α on the PRS method in the LFR benchmark are shown in Figure 4. Similar results are obtained for real-world networks.

In general, the values of α (specifically more than 0.15) has minor impact on the accuracy and expansion of the proposed method, because in the first phase of PRS method, only direct neighbors of initial set (one BFS level) are considered. Lower value of α leads to more accuracy and expansion. However, as we will show in Section V C 7 (refer to Equation 7), the time complexity of running the proposed method has a negative relation to the value of α (i.e., time complexity $\propto 1/\alpha$). Therefore, choosing a suitable value can be considered as a tradeoff between time complexity and desired performance. Moreover, by increasing α, the probabilities of the nodes in the PageRank vectors are close to each other and it is difficult to separate the nodes inside and outside of the community. Thus, sampling process is performed by only considering the unknown neighbors and the performance of PRS decreases in terms of accuracy and expansion.

5. The effect of t

In the Section III B, we show that sampling an equal number of nodes from each community leads to obtain the minimum error for characterizing the network properties in a sampled network rather than the entire network. Therefore, the proposed algorithm takes the actual number of communities (t) as input parameter to compute the equal number of nodes that should be sampled from each community. In the situation that the value of

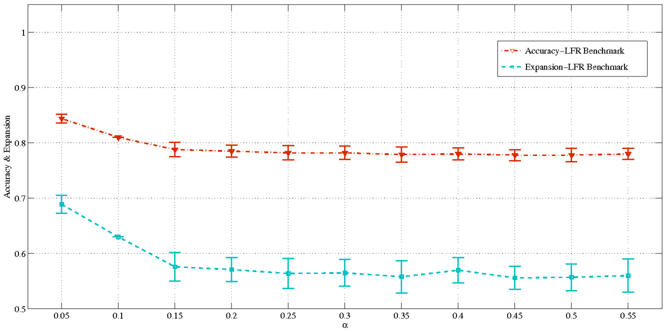


FIG. 4: The effect of α on the accuracy and expansion of the proposed method in the LFR benchmark. The parameters are: network size $N = 1000$, mixing parameter $\mu = 0.1$, sample rate $= 15\%$.

t is unkown, we have to approximate it. Therefore, it can be valuable to determine the sensitivity of the algorithm to error arising from approximation.

In this section, we study the impact of error in t on the performance of the PRS method. To this end, a range of values from $t' = t/2$ to $t' = 2t$ are considered as the approximate number of communities in the networks, where t is the actual number of communities. The results of our study in four LFR benchmark networks with different community level are shown in Figure 5. It can be seen that the maximum accuracy is obtained for $t' = t$, i.e. an equal number of samples from every community leads to optimal allocation. Similar results are obtained for real-world networks (refer to Figure 6 for the Football network).

6. Overhead

Visiting each new node in the network sampling has an associated overhead on resources such as time and bandwidth, and it can be too costly in large-scale networks. To study this issue, we define the overhead of a sampling method as the number of visited nodes in the network. The overhead of the RDS method is equal to the number of sampled nodes, i.e. its order is $O(|S|)$. However, PRS method is an algorithm of order $O(|N(S)|)$ where $N(S)$ denote the neighborhood of S. Moreover, the number of sampled nodes in PRS method has an inverse relation with the number of communities in the original network. In fact, if the network has less number of communities and sampling rate is high, the value of $|N(S)|$ is much lower.

7. Time Complexity

One of the other evaluation factors for a sampling method is the time complexity. The time complexity order of RDS method, as a modified version of classic random walk, is equal to $O(|S|)$. To compute the time complexity of the proposed method, we consider both phases. According to Theorem 1, the running time of computing the PageRank in the first phase is $O(\frac{|S|}{\epsilon.\alpha})$.

In the second phase, we need calculating the number of unknown neighbors of sampled nodes and computing their PageRank values. This repeats t times where t is the number of communities. Let d be the average degree of a node in the network. The order of finding unknown neighbors of a node and computing PageRank in the second phase is $O(t.d.|N(S)|)$ and $O(\frac{t}{\epsilon.\alpha})$ respectively.

According to $\frac{t}{\epsilon.\alpha} \in O(\frac{|S|}{\epsilon.\alpha})$, the time complexity of PRS method is:

$$O(t.d.|N(S)| + \frac{|S|}{\epsilon.\alpha}) \qquad (7)$$

VI. CONCLUSION

In this paper, we introduced PageRank-Sampling (PSR) as an efficient link-based approach to sample large-scale, static, and undirected complex networks with high community structures. The main idea of PSR is based on the expansion of a given community, which is a node initially, to include a set of nodes with the highest proximity to the initial node, by approximating the personalized PageRank. We empirically evaluated the proposed method against the popular Respondent Driven Sampling (RDS), on several synthetic and real-world networks. We showed that the proposed sampling method outperforms the popular sampling methods such as RDS in terms of accuracy and expansion.

VII. ACKNOWLEDGMENT

This research has been partially supported by ITRC (Iran Telecommunication Research Center) under grant number 6479/500 (90/4/22). The authors would also like to thank Mr. Hadi Asheri of DML for his help in preparing the manuscript.

[1] O. Frank, "Survey sampling in networks," in *Handbook of Social Network Analysis*, edited by J.Scott and P.Carrington (2011).

[2] Sh. Goel and M. J. Salganik, "Respondent-driven sampling as Markov chain Monte Carlo," Statistics in Medicine **28**, 2202–2229 (2008).

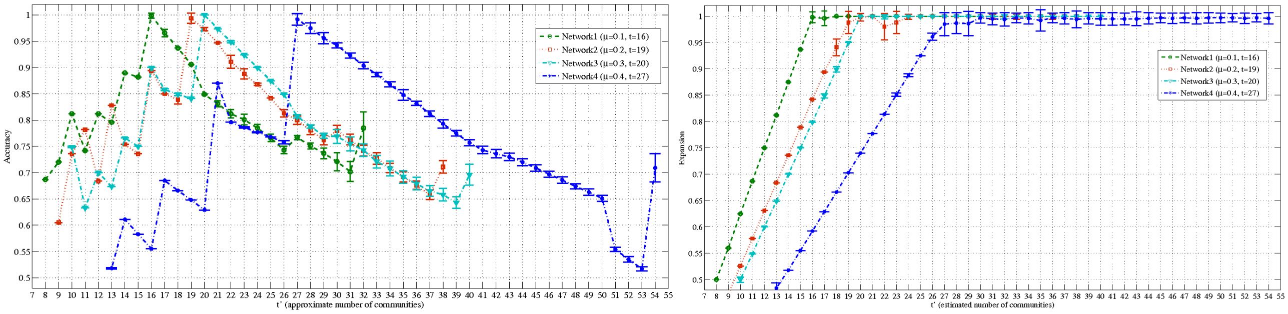
[3] K. Gile and M. S. Handcock, "Respondent-Driven Sampling: An Assessment of Current Methodology," Sociological Methodology **40**, 285–327 (2010).

[4] S. Fortunato, "Community detection in graphs," Physics Reports **486**, 75–174 (Jan. 2010).

[5] M. E. J. Newman, A. L. Barabasi, and D. J. Watts, *The Structure and Dynamics of Networks* (Princeton University Press, 2006).

[6] M. Salehi, H. R. Rabiee, and M. Jalili, "Motif structure and cooperation in real-world complex networks," PhysicaA **389**.

[7] M. E. Newman, "Scientific collaboration networks. I. Network construction and fundamental results," Phys Rev E Stat Nonlin Soft Matter Phys **64** (Jul. 2001).



(a) Accuracy

(b) Expansion

FIG. 5: The effect of approximating the number of communities (t') on the performance of the PRS method in four LFR benchmark networks with different community level. The parameters are: network size $N = 500$, sample rate $= 15\%$, $\alpha = 0.01$, the mixing parametrs of Network1, Network2, Network3, and Network4 is $\mu = 0.1$, $\mu = 0.2$, $\mu = 0.3$, and $\mu = 0.4$ respectively. The actual number of communities of Network1, Network2, Network3, and Network4 is 16, 19, 20, and 27 respectively.

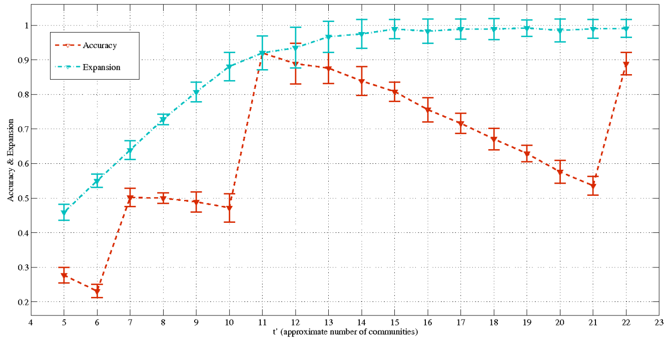


FIG. 6: The effect of approximating the number of communities (t') on the accuracy of the PRS method in Football network. The parameters are: sample rate $=$ 15%, $\alpha = 0.01$.

[8] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proceedings of the ACM SIGCOMM conference on Internet measurement*, IMC '07 (2007) pp. 29–42.

[9] L. Lovasz, "Random Walks on Graphs: A Survey," Combinatorics, **99**, 1–46 (1993).

[10] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculation by fast computing machines," Chemical Physics **21**, 1087–1092 (1953).

[11] D. D. Heckathorn, "Respondent-driven sampling: a new approach to the study of hidden populations," Social problems **44**, 19–24 (1997).

[12] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," Computer Networks and ISDN Systems **30**, 107–117 (Apr. 1998).

[13] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," Technical Report 1999-66 (Stanford InfoLab, 1999).

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

[15] R. Andersen and K. J. Lang, "Communities from seed sets," in *Proceedings of the conference on World Wide Web* (2006) pp. 223–232.

[16] R. Andersen, F. Chung, and K. J. Lang, "Local Graph Partitioning using PageRank Vectors," in *Proceedings of the IEEE Symposium on Foundations of Computer Science* (2006) pp. 475–486.

[17] A. Vattani, D. Chakrabarti, and M. Gurevich, "Preserving Personalized Pagerank in Subgraphs," in *Proceedings of ICML* (2011).

[18] J. Leskovec and Ch. Faloutsos, "Sampling from large graphs," in *Proceedings of the ACM SIGKDD conference on Knowledge discovery and data mining* (2006) pp. 631–636.

[19] D. Gfeller and P. Rios, "Spectral Coarse Graining of Complex Networks," Physical Review Letters **99** (2007).

[20] M.S. Handcock and K. Gile, "Modeling Social Networks with Sampled Data," Annals of Applied Statistics(2007).

[21] 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.

[22] Ch. Gkantsidis, M. Mihail, and A. Saberi, "Random walks in peer-to-peer networks: algorithms and evaluation," *P2P Computing Systems*, Performance Evaluation **63**, 241–263 (Mar. 2006).

[23] 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.

[24] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, "Walking in Facebook: A Case Study of Unbiased Sampling of OSNs," in *Proceedings of IEEE INFOCOM* (2010).

[25] 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(2011).

[26] S. Chib and E. Greenberg, "Understanding the Metropolis-Hastings Algorithm," The American Statistician **49**, 327–335 (1995).

[27] B. Krishnamurthy, Ph. Gill, and M. Arlitt, "A few chirps about twitter," in *Proceedings of the first workshop on Online social networks* (2008) pp. 19–24.

[28] M. J. Salganik and D. D. Heckathorn, "Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling," Sociological Methodology **34**, 193–240.

[29] A. Abdulquader, D. Heckathorn, C. Mcknight, H. Bramson, Ch. Nemeth, K. Sabin, and K. Gallagher, "Effectiveness of Respondent-Driven Sampling for Recruiting Drug Users in New York City: Findings from a Pilot Study," Urban Health **83**, 459–476 (2006).

[30] A. Rasti, R. Rejaie, N. Duffield, D. Stutzbach, and W. Willinger, "Evaluating sampling techniques for large dynamic graphs," Technical Report CIS-TR-08-01 (2008).

[31] A.S. Maiya and T.Y. Berger-Wolf, "Sampling Community Structure," in *Proceedings of Conference on the World Wide Web* (Apr 2010).

[32] D. Bader and J. McCloskey, "Modularity and Graph Algorithms," in *SIAM AN10 Minisymposium on Analyzing Massive Real-World Graphs* (2009).

[33] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, "Defining and identifying communities in networks," Proceedings of the National Academy of Sciences of the United States of America **101**, 2658–2663 (Mar. 2004).

[34] S. Papadopoulos, Y. Kompatsiaris, A. Vakali, and P. Spyridonos, "Community Detection in Social Media," Data Mining and Knowledge Discovery(2011).

[35] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," National Academy of Sciences **105**, 1118–1123 (Jan. 2008).

[36] N. Du, B. Wu, X. Pei, B. Wang, and L. Xu, "Community detection in large-scale social networks," in *Proceedings of the WebKDD* (2007) pp. 16–25.

[37] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. S. Tomkins, "The Web as a Graph: Measurements, Models, and Methods," (1999).

[38] A. Clauset, "Finding local community structure in networks," Physical Review E **72**, 026132 (Aug. 2005).

[39] F. Luo, J. Z. Wang, and E. Promislow, "Exploring local community structures in large networks," Web Intelli. and Agent Sys. **6**, 387–400 (December 2008).

[40] J. P. Bagrow, "Evaluating Local Community Methods in Networks," Statistical Mechanics: Theory and Experiment(2008).

[41] S. Papadopoulos, A. Skusa, A. Vakali, Y. Kompatsiaris, and N. Wagner, "Bridge bounding: A local approach for efficient community discovery in complex networks," (2009).

[42] J. Chen, O. S. Zaiane, and R. Goebel, "Local community identification in social networks," in *International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (2009).

[43] A. Lancichinetti, F. Radicchi, J. J. Ramasco, and S. Fortunato, "Finding Statistically Significant Communities in Networks," PLoS ONE **6**, e18961+ (Apr. 2011).

[44] M. Kurant, M. Gjoka, C. T. Butts, and A. Markopoulou, "Walking on a Graph with a Magnifying Glass: Stratified Sampling via Weighted Random Walks," in *SIGMETRICS* (ACM, 2011).

[45] P. Boldi, M. Santini, and S. Vigna, "PageRank as a function of the damping factor," in *Proceedings of the conference on World Wide Web* (ACM Press, 2005) pp. 557–566.

[46] T. Haveliwala, "Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search," IEEE Transactions on Knowledge and Data Engineering **15**, 784–796 (Jul. 2003).

[47] P. Berkhin, "Bookmark-Coloring Algorithm for Personalized PageRank Computing," Internet Mathematics **3**, 41–62 (2006).

[48] R. Andersen, F. Chung, and K. Lang, "Using pagerank to locally partition a graph," Internet Mathematics **4**, 35–64 (2007).

[49] M. Hansen and W. Hurwitz, "On the Theory of Sampling from Finite Populations," Annals of Mathematical Statistics **14** (1943).

[50] E. Volz and D. Heckathorn, "Probability based estimation theory for respondent-driven sampling," Official Statistics **24**, 79 (2008).

[51] A. Lancichinetti, S. Fortunato, and F. Radicchi, "Benchmark graphs for testing community detection algorithms," Physical Review E **78** (2008).

[52] A. Lancichinetti and S. Fortunato, "Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities," Physical Review E **80**, 016118+ (2009).

[53] R. Milo, Sh. Itzkovitz, N. Kashtan, R. Levitt, Sh. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon, "Superfamilies of Evolved and Designed Networks," Science **303**, 1538–1542 (Mar. 2004).

[54] J. Duch and A. Arenas, "Community detection in complex networks using extremal optimization," Physical Review E **72**, 027104 (Aug. 2005).

[55] P.Gleiser and L. Danon, "Community structure in jazz," Adv. Complex Syst. **6** (2003).

[56] W. W. Zachary, "An information flow model for conflict and fission in small groups," Anthropological Research **33**, 452–473 (1977).

[57] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," Proceedings of the National Academy of Sciences **99**, 7821–7826 (Jun. 2002).

[58] D. Lusseau, "The emergent properties of a dolphin social network," Proceedings of the Royal Society of London. Series B: Biological Sciences **270**, S186–S188 (Nov. 2003).

[59] V. Batagelj and Mrvar A., "Pajek analysis and visualization of large networks," *Proceedings of IEEE INFOCOM '10*, Junger M, Mutzel P, eds. Graph drawing software, 77–103(2003), `http://pajek.imfm.si/doku.php?id=data:index`.

[60] B. Bollobas, *Modern Graph Theory* (Springer, 1998).

[61] M. L. Goldstein, S. A. Morris, and G. G. Yen, "Problems with Fitting to the Power-Law Distribution," The European Physical Journal B **41**, 255–258 (Aug. 2004).