Sampling social networks using shortest paths

Alireza Rezvanian, Mohammad Reza Meybodi

Soft computing laboratory, Computer Engineering and Information Technology Department
Amirkabir University of Technology, Tehran, Iran

Abstract

In recent years, online social networks (OSN) have emerged as a platform of sharing variety of information about people, and their interests, activities, events and news from real worlds. Due to the large scale and access limitations (e.g., privacy policies) of online social network services such as *Facebook* and *Twitter*, it is difficult to access the whole public network in a limited amount of time. For this reason researchers try to study and characterize OSN by taking appropriate and reliable samples from the network. In this paper, we propose to use the concept of shortest path for sampling social networks. The proposed sampling method first finds the shortest paths between several pairs of nodes selected according to some criteria. Then the edges in these shortest paths are ranked according to the number of times that each edge has appeared in the set of found shortest paths. The sampled network is then computed as a subgraph of the social network which contains a percentage of highly ranked edges. In order to investigate the performance of the proposed sampling method, we provide a number of experiments on synthetic and real networks. Experimental results show that the proposed sampling method outperforms the existing method such as random edge sampling, random node sampling, random walk sampling and Metropolis-Hastings random walk sampling in terms of relative error (RE), normalized root mean square error (NMSE), and Kolmogorov-Smirnov (KS) test.

Keywords: Online social networks, Social network analysis, Network sampling, Shortest path.

1. Introduction

In recent years, online social networks (OSN) have emerged as a variety of information about users, their activities, events and news in their real worlds. OSN similar to many real world networks modeled and represented as a graph with a set of nodes (e.g., users of OSN) and edges (a certain type of relationship between users of OSN). It is shown that for different OSN there are common fascinating properties such as small-world and scale-free properties [1], [2]. The nature of OSN is large, dynamic, complex, and also some part of these networks is not fully accessible due to computational or privacy settings and thus most of the time the direct access on these networks is not feasible [3]. Therefore, these networks [4]–[6] are studied and characterized via metrics (i.e., centrality measures) or techniques (i.e., sampling methods) instead of access to the whole network

[7]–[10]. In practice, researchers via sampling from networks can process information in a reasonable amount of time and lower computational effort. From the sampled network which contains only a portion of the whole data, one can reveal some hidden important properties of networks such as user age distribution, user activities, user connectivity, and many more [11], [12].

Let $G= \langle V, E\rangle$ be the input graph where V is the set of nodes with size $n=|V|$ and E is the set of edges. Sampling is a function $f{:}G\rightarrow G_s$ from graph G to sampled graph $G_s=\langle V_s, E_s\rangle$ such that $V_s\subset V$, $E_s\subset E$, and $|V_s|=\phi n$, where $0<\phi<1$ denotes the sampling rate. Sampling methods play an important role in preprocessing, characterizing, and studying real networks [7], [9], [13]. Sampling can be used to study a small part of networks while preserving main features of the original network. In this paper, a sampling method for sampling social networks will be presented using the concept of shortest path. In the proposed sampling method, we first find the shortest paths between several pairs of nodes which are selected according to some criteria. Then the edges in these shortest paths are ranked according to the number of times that each edge has appeared in the set of found shortest paths. The sampled network is then created as a subgraph of the social network which contains only a percentage of highly ranked edges/nodes. The proposed algorithm could be simply implemented using repeated shortest path algorithm. In order to study the performance of the proposed sampling method, a number of experiments on synthetic and real networks are provided. Experimental results show that the proposed sampling method outperforms the existing method such as random edge sampling, random node sampling, random walk sampling, and Metropolis-Hastings random walk sampling in terms of relative error (RE), normalized root mean square error (NMSE), and Kolmogorov-Smirnov (KS) test [14]. The rest of this paper is organized as follows. Section 2 introduces sampling methods in brief. In section 3, the proposed sampling method and some of its improvements are described. The performance of the proposed sampling method is studied through simulations whose results are reported in section 4. Section 5 concludes the paper.

2. Related Works

There are a limited number of recent researches on studying, characterizing and estimating the properties of online social networks via sampling. Several sampling methods have been proposed for sampling networks that can be categorized into three main strategies in terms of collecting samples: random sampling, crawling based sampling, and coarse graining based sampling.

- **Random sampling:** In random sampling, random selection is done based on either nodes or edges without considering its topological structure [15]. Random sampling including two main simple techniques: random edge sampling and random nodes sampling. In random edge sampling (RES), an edge is selected at random and two nodes incident to the edge collected for sampled network. In random node sampling (RNS), each node is selected uniform randomly to form sampled network. RES and RNS due to simplicity are applicable for theoretical investigation. Also, due to simplicity of these methods, some improved methods based on RES and RNS are developed for sampling networks in recent years by researchers [16]–[18].

- **Crawling based sampling:** Crawling based sampling (also called topology based sampling or traversal based sampling) such as breadth-first-search (BFS) [19], depth-first-search (DFS) [20], forest fire sampling (FFS) [21], snowball sampling (SBS) [22], and random walk sampling (RWS) [23] have been used for collecting samples from networks. RWS iteratively selects the next node uniformly at random among all adjacent of that node. In RWS, a node with more edges will have higher probability of being sampled. Therefore, the sample mean tends to overestimate the average degree. It is noted that due to the simplicity and efficiently of RWS a variety of random walk such as Metropolis Hastings Random Walk (MHRW) [24], Weighted Random Walk (WRW) [25], Stratified Weighted Random Walk (SWRW) [25], Respondent Driven Sampling (RDS) [26], and Distributed Learning Automata based Sampling method (DLAS) [27] are improved and widely used in literature.

- **Coarse graining based sampling**: In coarse graining based sampling, the goal is to reduce the scale of networks, which mainly used for network visualization applications to display in a limited screen [28]. In this methods, several approaches presented such as clustering methods [29], k-core methods [30], and fractal based methods [31].

There are some special complexities and challenges in sampling from real networks which can be still discussed as new research fields [15]. A good study for sampling from complex networks presented by *Leskovec et al.* [15]. They proposed sampling method from large graph by introducing several basic methods with two goals: back in time and down scale. The study has shown that RNS and RES do not provide appropriate results. In [19] *Lee et al.* presented that RNS performs better than RES with respect to estimating the clustering coefficient of networks. *Lu et al.* studied sampling on Twitter data and this research reveals that results of RWS is much better than results of RNS or RES methods [32]. An analytical comparison between RWS and BFS sampling has been presented by *Kurant et al.* [8] to sampling from network. Their study indicates that the degree of graph is overestimated by the BFS, while it is underestimated by the RW sampling. Therefore, they suggested analytical solutions to correct the biasness of estimation. A practical framework for uniform sampling from users Facebook has been developed based on crawling in [33]. In this study, the advantage of unbiased estimation of MHRW and Re-WRW (RWRW) over random sampling and BFS has been addressed with comparing various approaches. *Rejaie et al.* tried to estimate the number of users for MySpace and Twitter by generating sequential user id [34], but this technique is failed for those online social networks where the user id is randomly generated. RDS was analyzed in [35] to reduce the biases associated with chain referral sampling of hidden populations. and then later, sampling from Twitter using RDS has been reported to characterize it [36]. They have shown through experimentations on Twitter, RDS has lower error in comparison with MHRW. Analysis of RWS method has been presented by *cooper et al.* [37], where the authors have tried to sample from the high degree vertices and similar graphs regarding the power law distribution. Cumulative distribution of degrees is estimated via sampling based on trace routing and some methods were studied for eliminating bias of the high degrees [38]. *Ribeiro et al.* proposed a sampling method, called Frontier sampling. It is developed from basic random walk which used several dependent random walks. The frontier sampling outperforms conventional random walk and generates small errors in sparse graphs. Based on the different types of relationship between users in OSN, multi-graph has been introduced using random walk [9] and the results of its simulation indicate improvement of the proposed method by *Gjoka.* Random jump in MHRW with prevention of being trapped in local structures has been proposed by *Jin et al* for unbiased estimation [39]. In [7], By sampling they avoid visiting all nodes in the vicinity of a user and thus attain improved performance. The advantage of their algorithm was demonstrated only by experimentations. Sampling for modular structure of networks has been studied by *Maiya et al.* via identifying the communities [40]. It is also structure of a bipartite graph has been considered by *Wang et al.* [41] for some social networks, they utilized MHRW for sampling from bipartite graph. Sampling from directed graphs has been proposed in [42] using random walk. Directed heterogeneous graph [43] is suggested by *yang et al.* in order to semantically sampling. They demonstrated that their sampling technique preserve relational profile property. An study on directed networks for sampling presented in [44] by introducing two-step framework to measure nodal characteristics. Another research on directed networks reported by *Son et al.* [45]. They focused on comparisons between variants of breadth-first search (BFS) sampling on directed networks. Sampling by crawling the edges has been discussed in [46] by the idea of page rank which provides more significant results in comparison with RDS. A sampling method which uses distributed learning automata is reported by *Rezvanian et al.* [27]. This method which is similar to RWS uses a set of learning automata cooperating with each other in order to guide the paths of visiting nodes by updating their action probabilities. The results of their method performed much better than the results of RDS and RWS in terms of RE and KS on well-known real networks. Based on basic snowball sampling a random multiple snowball with cohen process sampling is developed by *Gao et al.* [47]. Their simulations on computer generated networks indicated that this method is able to preserve local and global structure of network.

3. Proposed sampling method

In this section, we describe an algorithm which uses the concept of shortest path for sampling from social network graph. In the proposed sampling method, it is assumed that the graph of given online social networks is accessible by a unique ID of each user, usually generated randomly by social network management services, as nodes of input graph and also their allowed connected users. Figure 1 gives the pseudo code of the proposed sampling method. Let $G\langle V, E\rangle$ be the input graph, where $V=\{v_1, v_2, \ldots, v_n\}$ is the set of nodes, and $E=\{e_1, e_2, \ldots, e_m\}$ is the set of edges. The algorithm iteratively computes shortest paths between l pairs of nodes. At iteration i, it chooses randomly two non-adjacent nodes v_s as source node and v_d *as* destination node and then computes the shortest path π_i between these two nodes using a shortest path algorithm such as *Dijkstra's* algorithm [48]. Then using the computed shortest paths, a rank is assigned to each edge appearing in the computed shortest paths based on the number of shortest paths along which that edge has been appeared. Then the sampled network is constructed by choosing a given percent of the highly ranked vertices.

Experiments conducted by *Newman* and *Milgram* may be evidences of why using the concept of shortest path for sampling social networks is a promising approach. *Newman* [54], through his experiment emphasis the importance of shortest path on the scientific collaboration networks. He showed that on average 64% of one's shortest paths to other scientists pass through one's top-ranked collaborator and 17% pass through the second-ranked one. From the famous experiment of small-world phenomena by *Milgram* [50], [51], one may conclude that a participant does not know the shortest path using which to send a message to the target person, but by delivering the message to the person directly connected to a friend who is supposed to be closer to the target person estimates which of his acquaintances would lead to the shortest path through the full social network. Another important aspect of shortest path is related to selecting the most influential nodes in social networks [52]. In general, in online social networks information propagate along the shortest paths of users as direct and simple way to communicate between each user. For example, smart advertisers try to present their products with minimum cost of propagation via maximum influence path [53]. *Kimura et. al* [54] propose a shortest-path-based influence cascade model and provide efficient algorithms for finding the most influential nodes under these models.

According to the above discussion it seems that by considering the shortest paths as the building blocks of a sampled network we can have a sample which includes the central nodes (such as nodes with high degree or high betweenness) and at the same time preserves the networks functionalities.

The proposed sampling method using the concept of shortest path which we call it SSP is given in Figure 1.

```
Algorithm 1: the proposed sampling method (SSP)
Input: Graph $G=\langle V, E\rangle$
        $l$:  Number of computed shortest paths
        $m$: Number of nodes in the sampled network // $m\leftarrow\phi n$
Output: Sampled graph $G_s=\langle V_s, E_s\rangle$
Assumption
    Let $L$ denotes the maximum number of iteration
Begin
    Let $\pi_t$ denotes the shortest path between $v_s$ and $v_d$ at iteration $t$
    $t \leftarrow 1$
  While ($t < L$)
    Select two non-adjacent $v_s$ and $v_d$ randomly as source and destination nodes
    Find the shortest path $\pi_t$ between $v_s$ and $v_d$
    $t=t+1$
  End While
    Assign a rank to every edge that has been appeared in the shortest paths; The rank of an each is
      computed according to the number of time that the edge has appeared in the computed shortest paths.
    Generate sampled network $G_s$ by considering a subgraph of the input graph which contains a given
      percent of the highly ranked vertices ($m$ number of nodes).
End Algorithm
```

Figure 1. The pseudo code of the proposed sampling method

The time required by the algorithm can be divided into two parts: the time required computing the shortest paths and the time required to sort the edges appearing in the computed shortest paths. Since the number of shortest paths computed is a percentage of the number of nodes in the graph and the computation of a shortest path in a graph with n nodes is proportional to n^2 then the time required by the first part is O(n^3). Sorting the edges appearing in the computed shortest paths in the worst takes O(n^2log$n$) and hence O(n^3) time for the algorithm.

The proposed sampling method can be improved in several ways such as using better heuristics for selecting the source and destination nodes v_s and v_d or better heuristics for ranking the edges of the computed shortest paths. Examples of such improvements will be discussed later in experiment III.

4. Simulation Results

In this section, performance of the proposed sampling method is investigated on several well-known real networks: *Zachary's Karate Club* (Karate) [57], *American College Football* (Football) [58], *E-mail network* (Email) [59], *Jazz musicians network* (Jazz) [60] as reference datasets and *High-energy physics theory citation network* [61], *Social circles* (Facebook) [62], and *Wikipedia vote network* [63] as large networks [64] and some synthetic networks: ER-10000, WS-10000, BA-10000. All of these networks are commonly-used benchmarks for simulation. Table 1 describes the real and synthetic networks are used for the experimentations and their characteristics. For synthetic networks, we use well-known random network of *Erdös-Rényi* model (ER model) [65] which is utilized widely in literature. Other synthetic network is small world networks of *Watss-Strogatz* model (WS model) [1] which reflect common properties of many real networks such as short average path length. Further, the other well-known synthetic networks is *Barabási-Albert* model (BA model) as scale free networks with heavy-tailed degree distribution [2]. We set network parameters as N=10000, P=0.2 for ER model; N=10000, k=4, P=0.2 for WS model; and N=10000, m_0=m=5 for BA model.

Table 1. Description of test networks

Network	Node	Edge	Description
Karate	34	78	Zachary karate club network
Football	115	613	Network of american college football teams
Email	1133	5451	E-mail network of the Univeristy Rovira i Virgili
Jazz	198	2742	Network of jazz musicians
Cit-HepTh	27770	352807	ArXiv High Energy Physics Theory paper citation network
Ego-Facebook	4039	88234	Social circles from Facebook
Wiki-Vote	7115	103689	Wikipedia who-votes-on-whom network
ER-10000	10000	19990294	Synthetic random network
WS-10000	10000	123942	Synthetic small world network
BA-10000	10000	99945	Synthetic scale free network

4.1. Evaluation metrics

In this paper, we use *Kolmogrov-Smirnov Distance* (KS), *Relative Error* (RE) and *Normalized Root Mean Square Error* (NRMSE) for performance studies. The evaluation metrics are described in the rest of this subsection.

4.1.1. Kolmogrov-Smirnov Test (KS)

Kolmogrov-Smirnov D-Statistic is one of the statistical test methods used for calculating the distance between two cumulative distribution functions (CDF). *KS distance* is a commonly used measure for acceptability between distribution of original network and distribution of sampled network. It is calculated as the maximum absolute distance between the two distributions. The result of this test is a value between 0 and 1. As closer as it is to zero, both distributions will have a greater similarity; and as closer as it is to unit, the two distributions will show a greater discrepancy [14]. This metric has been calculated as the following equation

$$KS\,(F,F_s) = \max\left|F\left(x\right) - F_s\left(x\right)\right| \tag{1}$$

where F and F_s are CDF of a given distribution of original network and a given distribution of sampled network, respectively, and x is the range of the random variable. So it is computed as the maximum vertical distance between the two distributions. In this paper, we apply KS test on the degree distribution or shortest path length of original network and sampled network.

4.1.2. Relative Error (RE)

Relative Error (RE) can be applied to assess the accuracy of the results, which is defined by the following equation:

$$\mathrm{RE} = \frac{|\theta - \theta_{\mathrm{s}}|}{\theta} \tag{2}$$

where, θ and θ_s denote the values of a network parameter (i.e., clustering coefficient) in the sampled network and original network respectively [7]. In this paper, we calculated RE for clustering coefficient in the original network and sampled network. The clustering coefficient C_i of node v_i is calculated as follows

$$C_i = \frac{2T_i}{d_i(d_i - 1)} \tag{3}$$

where T_i is the number of edges in its adjacent node of v_i and d_i is the degree of node v_i.

4.1.3. Normalized Root Mean Square Error (NMSE)

Another metric used in this regard is Normalized Root Mean Square Error (NMSE) for values of real and estimated parameters (i.e., clustering coefficient) which is given by the following equation [10], [66]:

$$\mathrm{NMSE} = \frac{\sqrt{\left(\mathrm{E}\,|\,\theta - \theta_s\,|\right)^2}}{\theta} \tag{4}$$

where, θ and θ_s are the values of network parameter of original and sampled network.

4.2. Experimental Results

To show the performance of the proposed sampling method, several experiments are conducted on well-known synthetic and real networks described in Table 1. All experiments for the proposed sampling via shortest path (SSP) are evaluated in terms of RE, NMSE, KS test to assessment of goodness-of-fit and then we are compared it with Random Node Sampling (RNS), Random Edge Sampling (RES), Random Walk Sampling (RWS), Metropolis-Hastings Random Walk (MHRW), and distributed learning automata based sampling (DLAS) [27] for different sampling rates in the experiment II. The results reported in this paper are averages taken over 50 runs.

4.2.1. Experiment I

This experiment is conducted to study the minimum number of shortest paths required to meet a given sampling rate. For this purpose, we run the proposed sampling method on different test networks to calculate number of shortest paths required to meet a specific sampling rate. Results of this experiment are demonstrated in Figure 2 for Ego-Facebook, Wiki-Vote, Jazz, Email, Karate, ER-10000, WS-10000 and BA-10000. From the results, one can conclude that for sampling rates 10%, 15%, 20%, 25% and 30%, depending on the type of the network at least 1% to 4%, 1.5% to 6%, 3% to 9%, 4% to 12% and 7% to 16% of shortest paths, respectively, required to meet a specific sampling rate. Besides, from the results of this experiment, we may also say that for synthetic networks, the required number of shortest paths for synthetic small world networks is less than the required number of shortest paths for synthetic scale-free and synthetic random networks. Moreover, the number of required shortest paths needed for synthetic random networks is higher than the number of required shortest paths needed for real or synthetic networks. The result of this experiment will be used in other experiments in order to choose the required number of shortest paths to meet a given sampling rate.

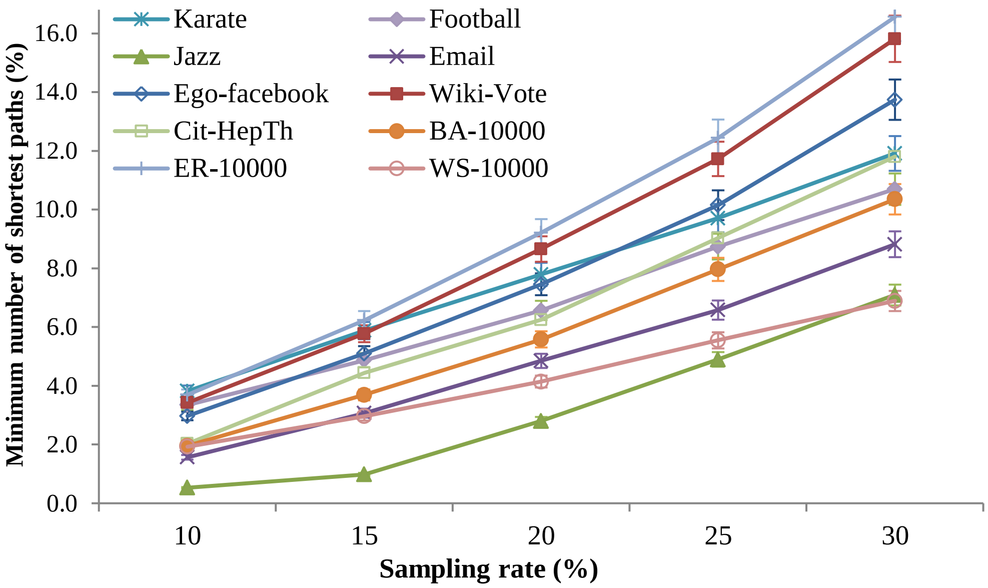


Figure 2. Minimum number of shortest paths required to meet the specific sampling rate

4.2.2. Experiment II

This experiment is carried out to compare the performance of the proposed sampling via shortest path with selecting both source and destination nodes at random as (SSP) with Random Node Sampling (RNS), Random Edge Sampling (RES), Random Walk Sampling (RWS), Metropolis-Hastings Random Walk (MHRW), and distributed learning automata based Sampling (DLAS) [27] for the sampling rate of 10% and 20%. The comparison is made in terms of KS test for degree distribution and shortest path length distribution. The results of this experiment are summarized in Table 2 to Table 5.

Table 2. KS test for degree distribution for sampling rate= 10%

Methods Networks	RES	RNS	RWS	MHRW	DLAS	SSP
Karate	0.45	0.67	0.47	0.43	0.41	**0.39**
Jazz	0.86	0.85	0.67	0.47	0.46	**0.44**
Email	0.57	0.59	0.52	0.36	0.35	**0.23**
Football	0.53	0.61	0.47	0.21	0.22	**0.20**
Wiki-Vote	0.40	0.39	0.37	0.36	0.34	**0.28**
Ego-Facebook	0.47	0.48	0.45	0.41	0.40	**0.35**
Cit-HepTh	0.57	0.58	0.51	0.46	0.45	**0.42**
ER-10000	0.59	0.58	0.60	0.59	0.58	**0.49**
WS-10000	0.58	0.59	0.57	0.55	0.56	**0.48**
BA-10000	0.52	0.53	0.49	0.48	0.49	**0.44**

Table 3. KS test for degree distribution for sampling rate=20%

Methods Networks	RES	RNS	RWS	MHRW	DLAS	SSP
Karate	0.47	0.46	0.51	0.44	0.38	**0.35**
Jazz	0.78	0.76	0.68	0.67	0.35	**0.31**
Email	0.42	0.43	0.41	0.39	0.31	**0.18**
Football	0.46	0.54	0.36	0.19	0.18	**0.17**
Wiki-Vote	0.31	0.32	0.32	0.29	0.30	**0.26**
Ego-facebook	0.39	0.40	0.43	0.36	0.37	**0.29**
Cit-HepTh	0.50	0.51	0.47	0.43	0.42	**0.41**
ER-10000	0.57	0.56	0.57	0.57	0.56	**0.46**
WS-10000	0.55	0.56	0.54	0.52	0.52	**0.43**
BA-10000	0.48	0.49	0.46	0.45	0.46	**0.37**

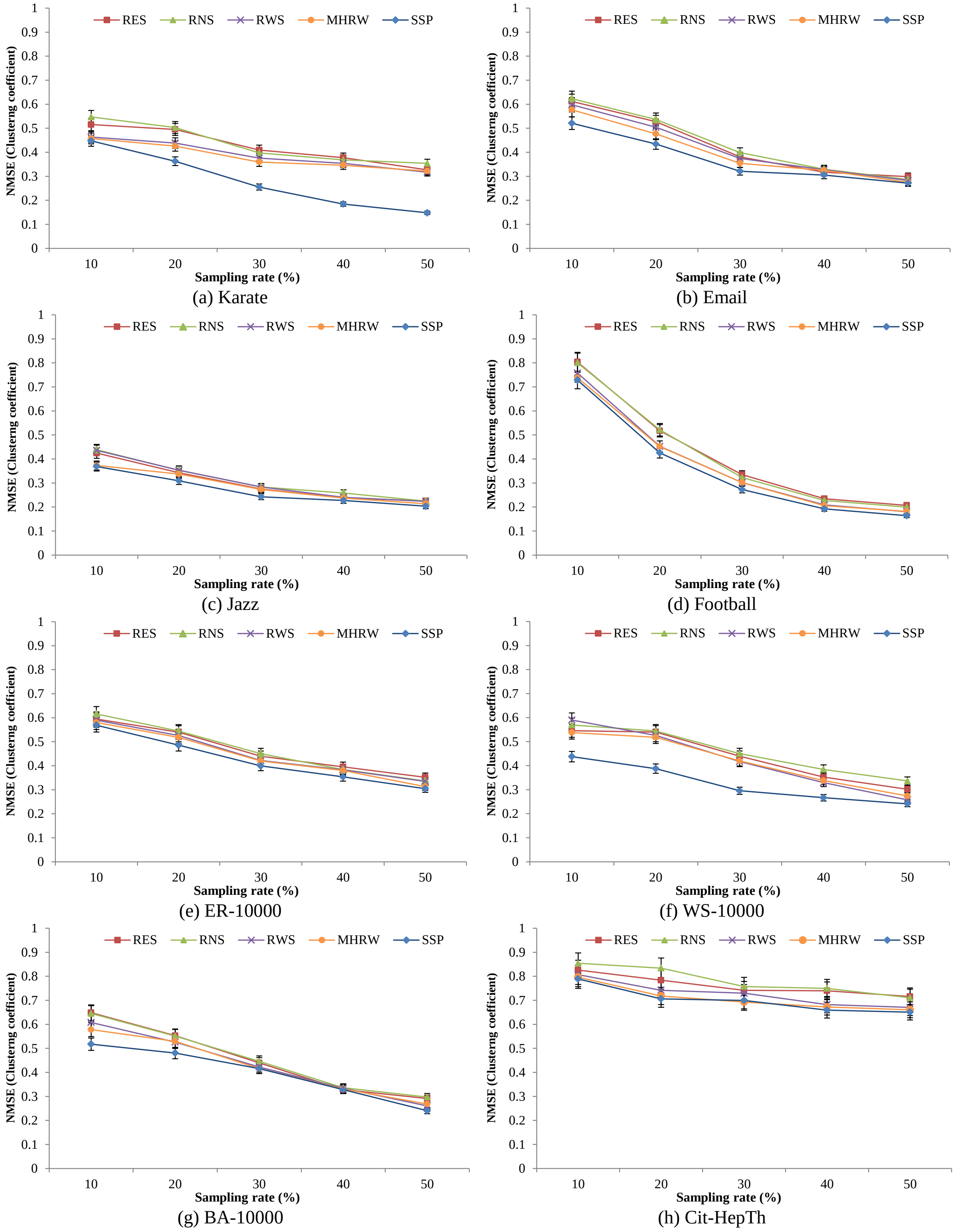
Table 4. KS test for shortest path length distribution for sampling rate=10%

Methods Networks	RES	RNS	RWS	MHRW	DLAS	SSP
Karate	0.33	0.33	0.33	0.33	0.33	0.33
Jazz	0.26	0.26	0.26	0.26	0.26	0.26
Email	0.46	0.46	0.46	0.40	0.40	0.40
Football	0.26	0.26	0.26	0.26	0.26	0.26
Wiki-Vote	0.33	0.33	0.33	0.26	0.26	0.26
Ego-Facebook	0.46	0.53	0.53	0.46	0.46	0.46
Cit-HepTh	0.86	0.93	0.86	0.73	0.73	0.73
ER-10000	0.26	0.26	0.26	0.26	0.26	**0.20**
WS-10000	0.53	0.60	0.53	0.46	0.53	**0.40**
BA-10000	0.73	0.66	0.66	0.63	0.66	**0.61**

Table 5. KS test for shortest path length distribution for sampling rate =0.20%

Methods Networks	RES	RNS	RWS	MHRW	DLAS	SSP
Karate	0.33	0.33	0.33	0.33	0.33	**0.26**
Jazz	0.33	0.33	0.33	0.26	0.33	**0.26**
Email	0.40	0.40	0.40	0.40	0.40	**0.26**
Football	0.26	0.26	0.26	0.26	0.26	0.26
Wiki-Vote	0.20	0.26	0.26	0.20	0.26	**0.20**
Ego-facebook	0.46	0.46	0.46	0.46	0.46	**0.40**
Cit-HepTh	0.73	0.86	0.73	0.66	0.66	**0.60**
ER-10000	0.13	0.13	0.13	0.13	0.13	0.13
WS-10000	0.53	0.53	0.36	0.36	0.36	0.36
BA-10000	0.66	0.66	0.60	0.60	0.60	0.60

Comparisons in terms of NMSE among these sampling methods are also done and the result is reported in Figure 3. According to the results of Figure 3, we may conclude that for all the test networks, higher (lower) value for sampling rate results in smaller (higher) value of NMSE. From Table 2 to Table 5, we also observe that the proposed sampling method outperforms other sampling methods in all test networks. Besides, From Figure 3 and Table 6, 7 and 8, we can say that in terms of NMSE, the proposed method, MHRW, RWS, RES and RNS, has rank 1, 2, 3, 4 and 5, respectively for real and synthetic networks. According to the obtained results, the proposed algorithm preserves the topological properties of the input graph such as clustering coefficient and degree distribution.



(a) Karate

(b) Email

(c) Jazz

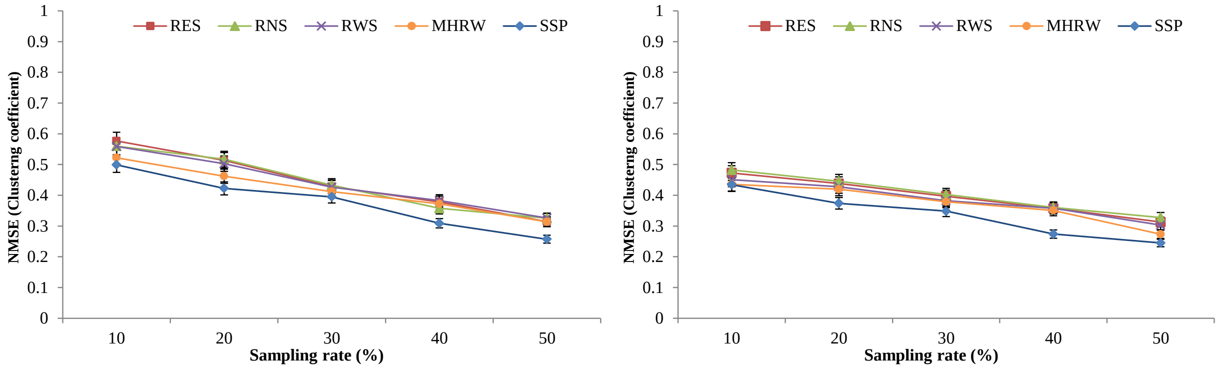
(d) Football

(e) ER-10000

(f) WS-10000

(g) BA-10000

(h) Cit-HepTh



(i) Wiki-Vote

(j) Ego-facebook

Figure 3. Comparison NMSE of clustering coefficient of sampling methods for varying sampling rate

Table 6. Average ranking for different mechanisms with respect to RE (clustering coefficient) for different sampling rates

Sampling rate	10%					20%					30%					40%				
Mechanism Instances	RES	RNS	RWS	MHRW	SSP	RES	RNS	RWS	MHRW	SSP	RES	RNS	RWS	MHRW	SSP	RES	RNS	RWS	MHRW	SSP
Average ranking	4.1	4.8	2.8	2.3	**1**	4.2	4.6	2.9	2.5	**1**	3.9	4.7	2.8	2.5	**1**	4.1	4.5	2.9	2.5	**1**
Ranking	4	5	3	2	**1**	4	5	3	2	**1**	4	5	3	2	**1**	4	5	3	2	**1**

Table 7. Average ranking for different mechanisms with respect to KS (degree distribution) for different sampling rates

Sampling rate	10%					20%					30%					40%				
Mechanism Instances	RES	RNS	RWS	MHRW	SSP	RES	RNS	RWS	MHRW	SSP	RES	RNS	RWS	MHRW	SSP	RES	RNS	RWS	MHRW	SSP
Average ranking	4.1	4.9	2.6	2.4	**1**	4.2	4.5	2.8	2.5	**1**	4.3	4.7	2.7	2.3	**1**	4.1	4.8	2.7	2.4	**1**
Ranking	4	5	3	2	**1**	4	5	3	2	**1**	4	5	3	2	**1**	4	5	3	2	**1**

Table 8. Average ranking for different mechanisms with respect to KS (shortest path length distribution) for different sampling rates

Sampling rate	10%					20%					30%					40%				
Mechanism Instances	RES	RNS	RWS	MHRW	SSP	RES	RNS	RWS	MHRW	SSP	RES	RNS	RWS	MHRW	SSP	RES	RNS	RWS	MHRW	SSP
Average ranking	1.9	2.1	1.9	1.3	**1**	1.8	2	1.7	1.4	**1**	1.9	2	1.8	1.5	**1**	1.8	1.9	1.7	1.3	**1**
Ranking	3	4	3	2	**1**	4	5	3	2	**1**	4	5	3	2	**1**	4	5	3	2	**1**

4.2.3. Experiment III

This experiment is conducted to study the impact of the mechanism used for selecting source and destination nodes on the performance of the proposed sampling method. For this purpose, we study the impact of five different mechanisms **I**, **II**, **III**, **IV** and **V** as described below for selecting source and destination nodes on the performance of the proposed method with respect to RE and KS distance.

I. Selecting both source and destination nodes at random. This mechanism was described in the previous section.

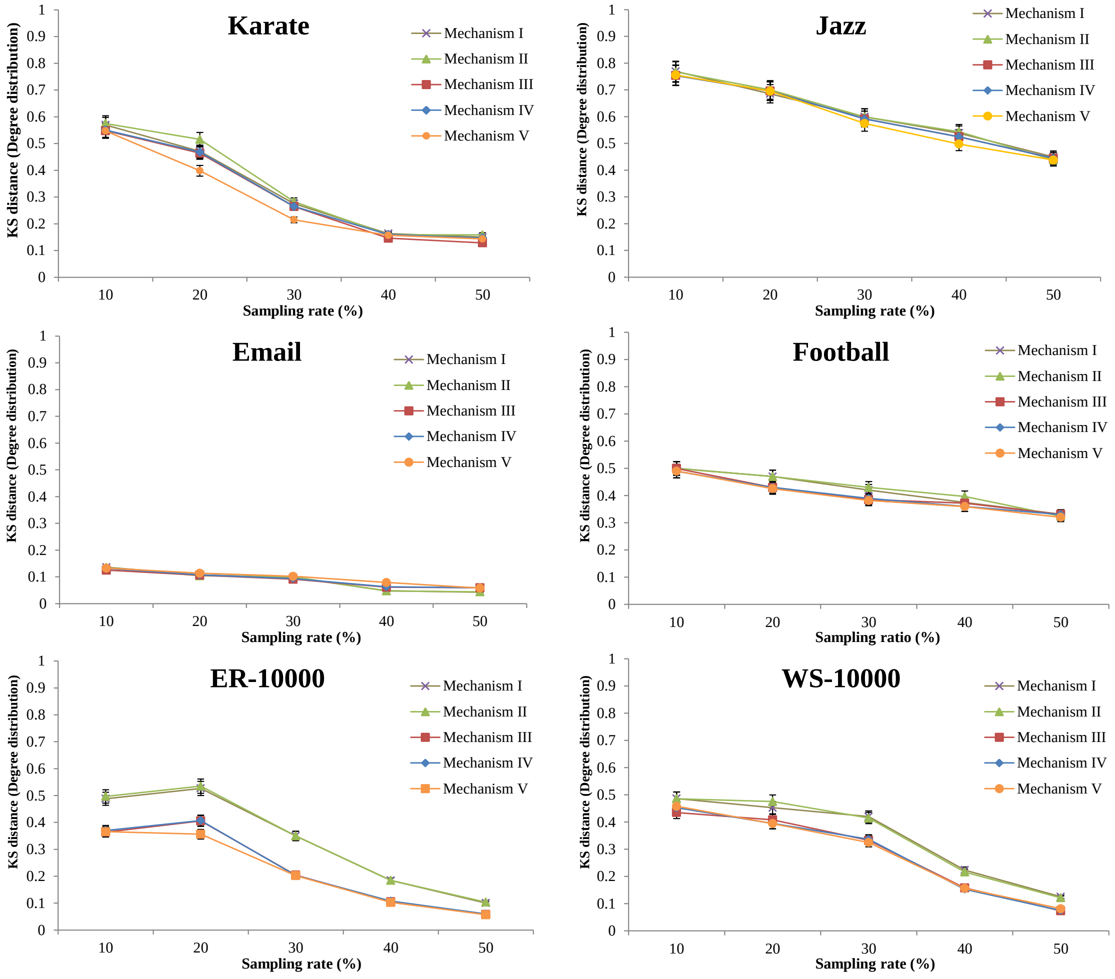
II. Selecting both source and destination nodes at random provided that they are not adjacent. This mechanism was used in the previous experiment.

III. Selecting a central node as a source node and selecting destination node at random. The measure of centrality can be chosen to be any measure of centrality such as betweenness or degree. For this experiment we consider the measure of centrality to be degree centrality.

VI. Selecting a central node as a source node and selecting a non-adjacent destination node at random. For this experiment we consider the measure of centrality to be degree centrality.

V. Considering only the shortest paths with length greater than *k* for assigning rank to edges. The shortest path between two randomly selected nodes is found. If the length of this path is less than *k* then it will be ignored and not used for assigning rank to edges. For this experiment, we set *k* = 2.

The results of this experiment for different test networks are given in Figure 4 and Figure 5. Further, we calculated the average ranking of each mechanism with respect to RE and KS for different sampling rates as given in Table 9, 10 and 11 by inspecting the results given in Figure 4 and Figure 5. Based on Table 9, 10, and 11 we can rank mechanisms I, II, III, IV and V, in 5th, 4th, 3rd, 2nd, and 1st, respectively. From the results, we may conclude that mechanism V outperforms the other four mechanisms in terms of both RE and KS distances.



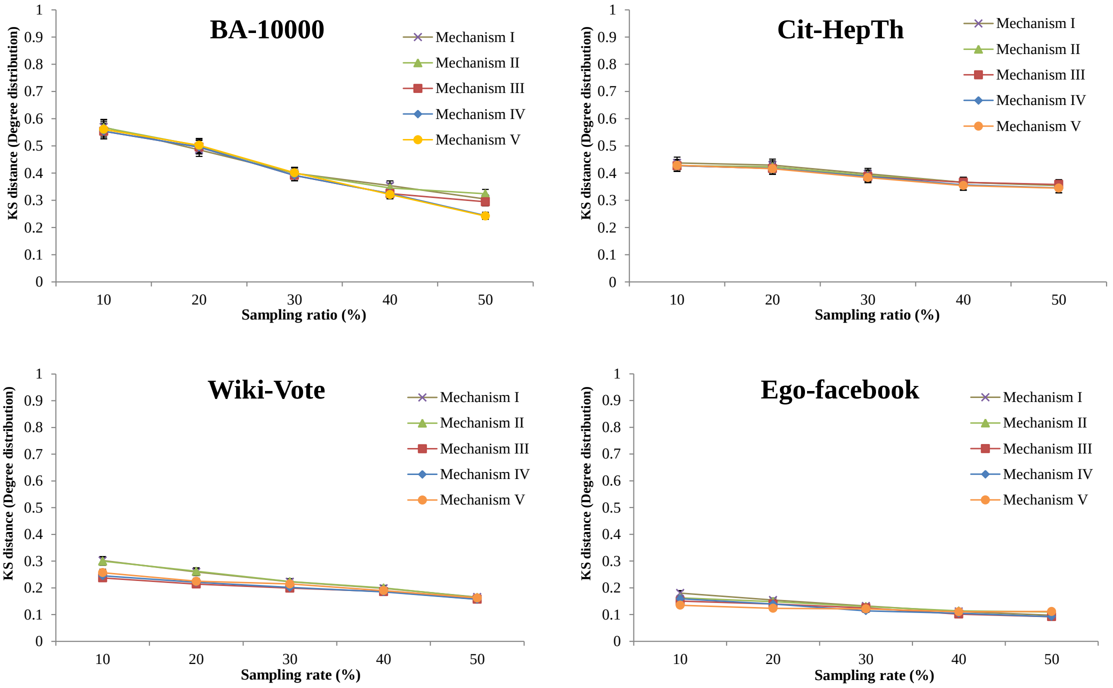
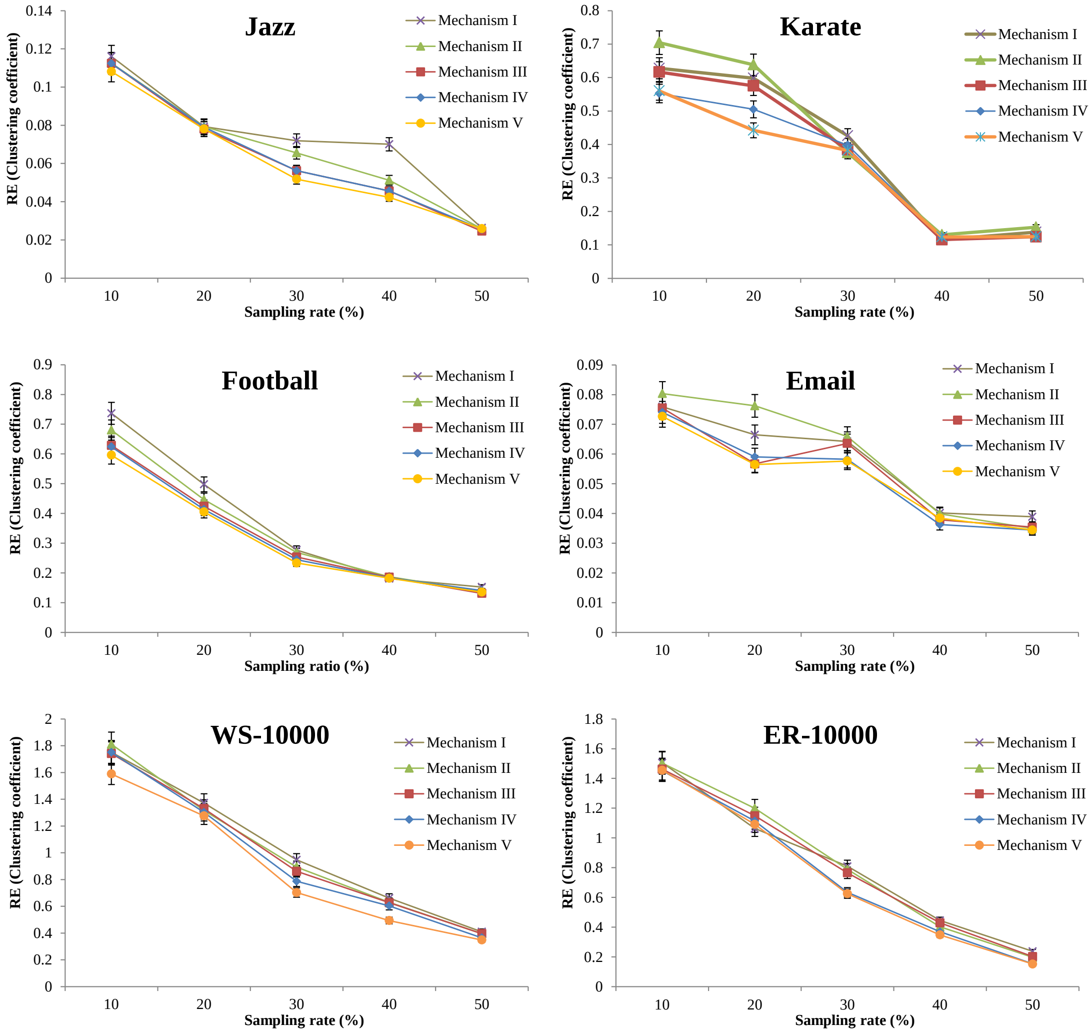


Figure 4. Average KS distance for test networks



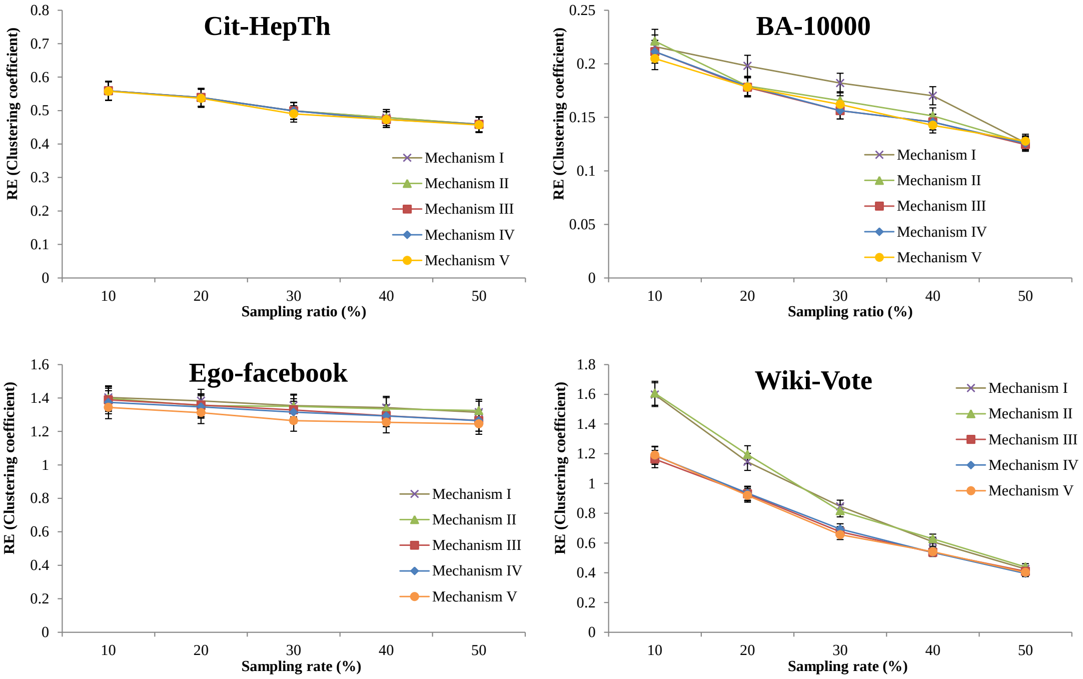


Figure 5. Average RE for test networks

Table 9. Average ranking for different mechanisms with respect to RE (clustering coefficient) for different sampling rates

Sampling rate	10%					20%					30%					40%				
Mechanism Instances	I	II	III	IV	V	I	II	III	IV	V	I	II	III	IV	V	I	II	III	IV	V
Average ranking	4.2	4.3	2.7	2.6	**1.2**	4.6	4.4	2.5	2.3	**1.2**	4.6	3.8	3	2.4	**1.2**	4.4	4.3	2.3	2.1	**1.9**
Ranking	4	5	3	2	**1**	5	4	3	2	**1**	5	4	3	2	**1**	5	4	3	2	**1**

Table 10. Average ranking for different mechanisms with respect to KS (degree distribution) for different sampling rates

Sampling rate	10%					20%					30%					40%				
Mechanism Instances	I	II	III	IV	V	I	II	III	IV	V	I	II	III	IV	V	I	II	III	IV	V
Average ranking	4.2	4.3	2.6	2.5	**1.4**	4.6	4.4	2.4	2.3	**1.3**	4.6	3.8	2.9	2.4	**1.3**	4.4	4.3	2.5	2.2	**1.5**
Ranking	4	5	3	2	**1**	5	4	3	2	**1**	5	4	3	2	**1**	5	4	3	2	**1**

Table 11. Average ranking for different mechanisms with respect to KS (shortest path length distribution) for different sampling rates

Sampling rate	10%					20%					30%					40%				
Mechanism Instances	I	II	III	IV	V	I	II	III	IV	V	I	II	III	IV	V	I	II	III	IV	V
Average ranking	1.7	1.7	1.7	1.6	**1.3**	2.4	2	1.2	1.3	**1.1**	2.3	2.2	2.1	1.8	**1.2**	1.4	1.1	1.1	1.1	1.1
Ranking	3	3	3	2	**1**	5	4	2	3	**1**	5	4	3	2	**1**	2	1	1	1	1

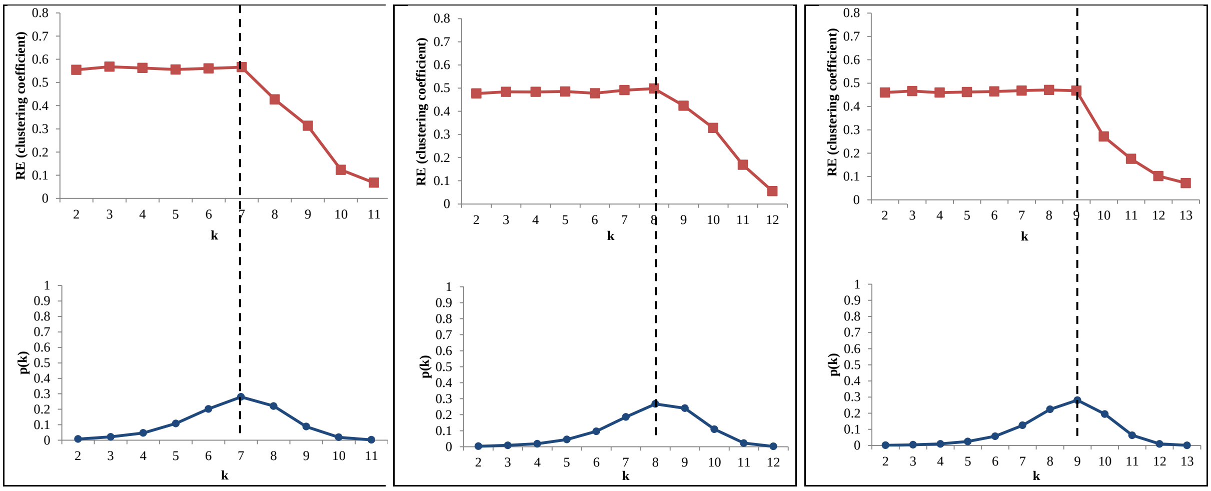
4.2.4. Experiment IV

This experiment is carried out to study the impact of parameter *k* used in mechanism V on RE and KS metrics. To do this, we plot shortest path length distribution and RE versus parameter *k* and then compare these two plots for several synthetic small-world networks and synthetic scale-free networks with different sizes. According to the results reported in Figure 6 and 7, we may conclude the followings:

- For synthetic small world networks increasing parameter *k* causes RE for clustering coefficient remains unchanged up to a point and then starts decreasing. RE starts to decrease when the value of parameter *k* becomes equal to the length of shortest paths occurring the at most in the networks. For example, for WS-2000 RE starts to decrease when parameter *k* is equal to 7. From figures 6a, 6b and 6c we can say that 7 is the length of shortest paths occurring at most in the networks. For WS-5000, RE starts to decrease when the value of parameter *k* becomes equal to the 8.

- For synthetic scale free networks increasing *k* causes RE for clustering coefficient increases up to a point and then starts decreasing. The point at which RE starts to decrease appears when the value of parameter *k* becomes equal to the length of shortest paths occurring the most in the network (figures 7a, 7b and 7c). It seems that in these types of networks longer shortest paths contains more pervasive nodes.

Experimentations with others networks may produce different results.

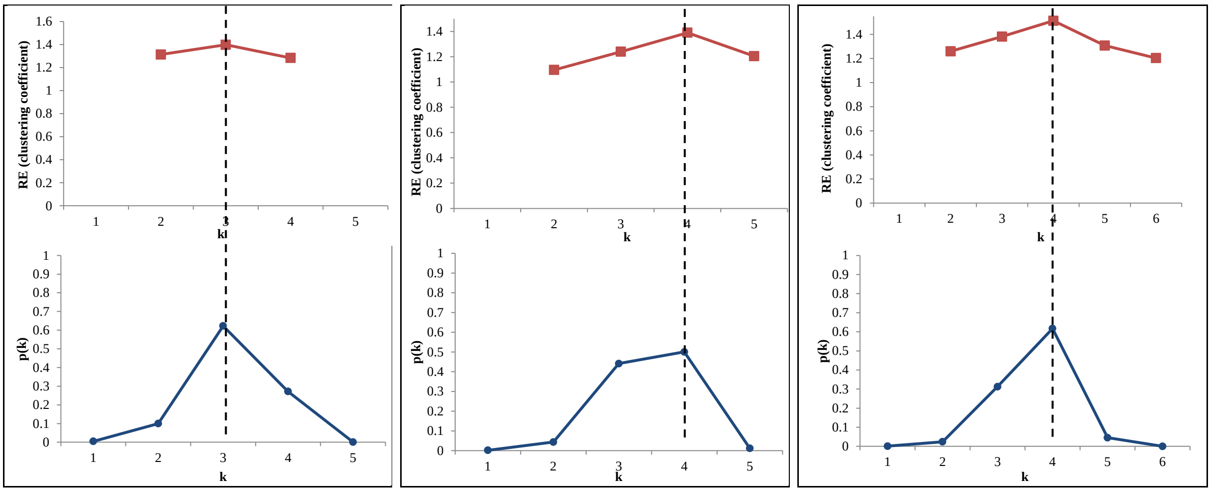


(a) WS-2000

(b) WS-5000

(c) WS-10000

Figure 6. Impact of parameter k on the performance of the proposed method in terms of RE (clustering coefficient) for synthetic small world networks



(a) BA-2000 (b) BA-5000 (c) BA-10000

Figure 7. Impact of parameter k on the performance of the proposed method in terms of RE (clustering coefficient) for synthetic scale free networks

4.2.5. Experiment V

This experiment is conducted to study the impact of sampling rate on the performance of the proposed method in terms of NMSE. For this purpose, for each test network we plot NMSE versus different sampling rates as given in Figure 8. From this figure, we can see that NMSE decreases as the sampling rate increases and also NMSE is higher for small networks. Among synthetic networks, NMSE of random network is higher than NMSE of small world and scale free networks. Also, NMSE of scale free networks is lower than NMSE of other synthetic networks.

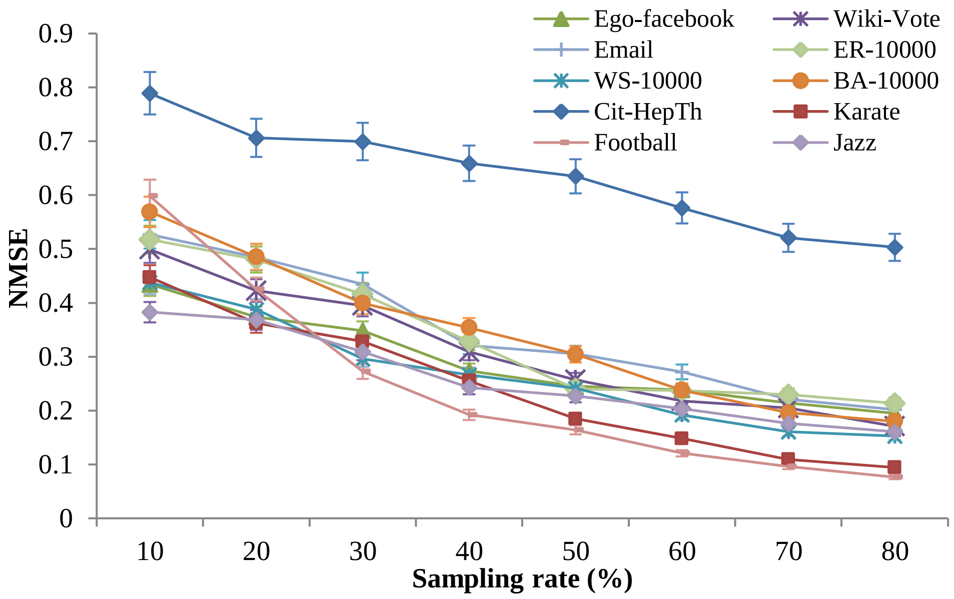


Figure 8. NMSE versus sampling rate for different networks

4.2.6. Experiment VI

This experiment is conducted to study the run time for the proposed sampling method (SSP) when both source and destination nodes are selected at random, Random Node Sampling (RNS) [15], Random Edge Sampling (RES) [15], Random Walk Sampling (RWS) [23], Metropolis-Hastings Random Walk (MHRW) [24], and distributed learning automata based Sampling (DLAS) This experiment is conducted for synthetic random graph and synthetic small world network with 50 to 100 nodes and different densities varying from 0.2 to 0.9 and sampling rate of 10% which is widely used. This experiment was launched on a system with hardware configuration of Intel® Core i5 U520 1.07 GHz processor and 4 GB Ram. The results of this experiment are given in Table 12 for synthetic random graph and Table 13 for synthetic scale free network. It should be noted that unlike most sampling algorithm the proposed algorithm traverses only a portion of the graph rather than the whole graph. Experimentations have shown that the proposed algorithm performs very fast when the input graph is a synthetic random graph with low density. For synthetic random graph with low densities (0.2 to 0.5) the run-time of the proposed algorithm is lowest comparing to other sampling methods. But, for synthetic random graph with high densities (0.6 to 0.9), the proposed algorithm performs only better than DLAS. Experimentations also show that for synthetic small world network the run time for RNS is lowest comparing to other sampling methods. Besides, for synthetic small world network the run time of the proposed sampling method is always lower than the run time of DLAS. Even though the proposed algorithm requires higher running time comparing to other methods but instead it produces better results.

Table 12. **Run-time of different sampling methods for synthetic random graphs with varying densities**

Methods Densities	RES	RNS	RWS	MHRW	DLAS	SSP
0.2	0.4782	0.3502	0.4031	0.8329	1.1772	**0.2785**
0.3	0.5065	0.4386	0.4315	0.9513	1.3045	**0.3579**
0.4	0.5999	0.4772	0.5175	1.0175	1.4452	**0.4461**
0.5	0.7022	0.5405	0.6436	1.1229	1.5497	**0.5204**
0.6	0.7475	**0.6096**	0.7171	1.1838	1.6454	0.6708
0.7	0.7852	**0.6872**	0.7592	1.2752	1.7021	0.8655
0.8	0.8269	**0.7483**	0.7793	1.4588	1.7571	1.2216
0.9	0.9364	**0.8075**	0.8486	1.6779	1.8661	1.9028

Table 13. Run-time of different sampling methods for synthetic small world graphs with varying densities

Methods Densities	RES	RNS	RWS	MHRW	DLAS	SSP
0.2	0.1271	**0.0747**	0.1269	0.2566	2.8525	0.1234
0.3	0.1316	**0.0765**	0.1256	0.3572	3.6182	0.1474
0.4	0.1355	**0.0776**	0.1318	0.4688	4.4147	0.1753
0.5	0.1387	**0.0799**	0.1371	0.4807	4.7167	0.1883
0.6	0.1481	**0.0813**	0.1435	0.5301	5.0714	0.1964
0.7	0.1595	**0.0836**	0.1554	0.5457	5.1746	0.2237
0.8	0.1647	**0.0862**	0.1618	0.5912	5.2214	0.2361
0.9	0.1753	**0.0914**	0.1747	0.6356	5.2528	0.2714

5. Conclusion

In this paper, based on the concept of shortest path we proposed a method for sampling social networks. The proposed algorithm by finding the shortest paths between several pairs of nodes tries to collect highly promising set of nodes and edges for constructing a sampled network. The performance of the proposed sampling method was investigated by conducting several experiments on well-known real and synthetic networks. The experimental results showed that the proposed sampling method outperforms other sampling methods such as RES, RNS, RWS, MHRW, and DLAS in terms of RE, NMSE and KS-test.

Experimental results show that the proposed sampling method outperforms the existing method such as random edge sampling, random node sampling, random walk sampling and Metropolis-Hastings random walk sampling in terms of relative error (RE), normalized root mean square error (NMSE), and Kolmogorov-Smirnov (KS) test

6. References

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

[2] A. L. Barabási and R. Albert, "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509–512, 1999.

[3] M. Huisman, "Imputation of missing network data: Some simple procedures," *Journal of Social Structure*, vol. 10, no. 1, pp. 1–29, 2009.

[4] M. Soleimani-pouri, A. Rezvanian, and M. R. Meybodi, "An Ant Based Particle Swarm Optimization Algorithm for Maximum Clique Problem in Social Networks," in *State of the Art Applications of Social Network Analysis*, F. Can, T. Özyer, and F. Polat, Eds. Springer International Publishing, 2014, pp. 295–304.

[5] M. Soleimani-Pouri, A. Rezvanian, and M. R. Meybodi, "Finding a Maximum Clique Using Ant Colony Optimization and Particle Swarm Optimization in Social Networks," in *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, 2012, pp. 58–61.

[6] F. Amiri, N. Yazdani, H. Faili, and A. Rezvanian, "A Novel Community Detection Algorithm for Privacy Preservation in Social Networks," in *Intelligent Informatics*, vol. 18, A. Abraham, Ed. 2013, pp. 443–450.

[7] M. Papagelis, G. Das, and N. Koudas, "Sampling Online Social Networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 3, pp. 662–676, 2013.

[8] M. Kurant, A. Markopoulou, and P. Thiran, "Towards Unbiased BFS Sampling," *IEEE Journal on Selected Areas in Communications,*, vol. 29, no. 9, pp. 1799–1809, 2011.

[9] M. Gjoka, C. T. Butts, M. Kurant, and A. Markopoulou, "Multigraph sampling of online social networks," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 9, pp. 1893–1905, 2011.

[10] F. Murai, B. Ribeiro, D. Towsley, and P. Wang, "On Set Size Distribution Estimation and the Characterization of Large Networks via Sampling," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 6, pp. 1017–1025, 2013.

[11] U. Pfeil, R. Arjan, and P. Zaphiris, "Age differences in online social networking–A study of user profiles and the social capital divide among teenagers and older users in MySpace," *Computers in Human Behavior*, vol. 25, no. 3, pp. 643–654, 2009.

[12] C. Wilson, B. Boe, A. Sala, K. P. N. Puttaswamy, and B. Y. Zhao, "User interactions in social networks and their implications," in *Proceedings of the 4th ACM European conference on Computer systems*, 2009, pp. 205–218.

[13] E. Volz and D. D. Heckathorn, "Probability based estimation theory for respondent driven sampling," *Journal of Official Statistics-Stockholm*, vol. 24, no. 1, p. 79, 2008.

[14] M. L. Goldstein, S. A. Morris, and G. G. Yen, "Problems with fitting to the power-law distribution," *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 41, no. 2, pp. 255–258, 2004.

[15] J. Leskovec and C. Faloutsos, "Sampling from large graphs," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 631–636.

[16] N. K. Ahmed, F. Berchmans, J. Neville, and R. Kompella, “Time-based sampling of social network activity graphs,” in *Proceedings of the Eighth Workshop on Mining and Learning with Graphs*, 2010, pp. 1–9.

[17] N. K. Ahmed, J. Neville, and R. Kompella, “Space-efficient sampling from social activity streams,” in *Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications*, 2012, pp. 53–60.

[18] 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 *Proceedings of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*, 2011, pp. 281–292.

[19] S. H. Lee, P. J. Kim, and H. Jeong, “Statistical properties of sampled networks,” *Physical Review E*, vol. 73, no. 1, p. 016102, 2006.

[20] S. Even, *Graph Algorithms*, 2nd ed. Cambridge University Press, 2011.

[21] M. Kurant, A. Markopoulou, and P. Thiran, “On the bias of BFS (Breadth First Search),” in *2010 22nd International Teletraffic Congress (ITC)*, 2010, pp. 1–8.

[22] O. Frank, “Survey sampling in networks,” in *The SAGE Handbook of Social Network Analysis*, SAGE publications, 2011, pp. 370–388.

[23] S. Yoon, S. Lee, S. H. Yook, and Y. Kim, “Statistical properties of sampled networks by random walks,” *Physical Review E*, vol. 75, no. 4, p. 046114, 2007.

[24] C. H. Lee, X. Xu, and D. Y. Eun, “Beyond random walk and Metropolis-Hastings samplers: Why you should not backtrack for unbiased graph sampling,” in *Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE joint international conference on Measurement and Modeling of Computer Systems*, 2012, pp. 319–330.

[25] M. Kurant, M. Gjoka, C. T. Butts, and A. Markopoulou, “Walking on a Graph with a Magnifying Glass,” in *Proceedings of ACM SIGMETRICS*, 2011, pp. 1–12.

[26] S. Goel and M. J. Salganik, “Assessing respondent-driven sampling,” *Proceedings of the National Academy of Sciences*, vol. 107, no. 15, pp. 6743–6747, 2010.

[27] A. Rezvanian, M. Rahmati, and M. R. Meybodi, “Sampling from complex networks using distributed learning automata,” *Physica A: Statistical Mechanics and its Applications*, vol. 396, pp. 224–234, 2014.

[28] Y. Jia, J. Hoberock, M. Garland, and J. Hart, “On the visualization of social and other scale-free networks,” *IEEE Transactions on Visualization and Computer Graphics,*, vol. 14, no. 6, pp. 1285–1292, 2008.

[29] S. Lafon and A. B. Lee, “Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,*, vol. 28, no. 9, pp. 1393–1403, 2006.

[30] S. N. Dorogovtsev, A. Goltsev, and J. F. F. Mendes, “K-core organization of complex networks,” *Physical review letters*, vol. 96, no. 4, p. 40601, 2006.

[31] J. S. Kim, K. I. Goh, B. Kahng, and D. Kim, “Fractality and self-similarity in scale-free networks,” *New Journal of Physics*, vol. 9, p. 177, 2007.

[32] J. Lu and D. Li, “Sampling online social networks by random walk,” in *Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research*, 2012, pp. 33–40.

[33] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, “Walking in Facebook: A case study of unbiased sampling of OSNs,” in *2010 Proceedings IEEE INFOCOM*, 2010, pp. 1–9.

[34] R. Rejaie, M. Torkjazi, M. Valafar, and W. Willinger, “Sizing up online social networks,” *IEEE Network,*, vol. 24, no. 5, pp. 32–37, 2010.

[35] K. J. Gile and M. S. Handcock, “Respondent-driven sampling: an assessment of current methodology,” *Sociological Methodology*, vol. 40, no. 1, pp. 285–327, 2010.

[36] M. Salehi, H. R. Rabiee, N. Nabavi, and S. Pooya, “Characterizing Twitter with Respondent-Driven Sampling,” in *2011 IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing (DASC)*, 2011, pp. 1211–1217.

[37] C. Cooper, T. Radzik, and Y. Siantos, “A fast algorithm to find all high degree vertices in power law graphs,” in *Proceedings of the 21st international conference companion on World Wide Web*, 2012, pp. 1007–1016.

[38] B. Ribeiro and D. Towsley, “Estimating and sampling graphs with multidimensional random walks,” in *Proceedings of the 10th annual conference on Internet measurement*, 2010, pp. 390–403.

[39] L. Jin, Y. Chen, P. Hui, C. Ding, T. Wang, A. V. Vasilakos, B. Deng, and X. Li, “Albatross sampling: robust and effective hybrid vertex sampling for social graphs,” in *Proceedings of the 3rd ACM international workshop on MobiArch*, 2011, pp. 11–16.

[40] A. S. Maiya and T. Y. Berger-Wolf, “Sampling community structure,” in *Proceedings of the 19th international conference on World wide web*, 2010, pp. 701–710.

[41] J. Wang and Y. Guo, "Unbiased sampling of bipartite graph," in *2011 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC),* 2011, pp. 357–360.

[42] B. Ribeiro, P. Wang, F. Murai, and D. Towsley, "Sampling directed graphs with random walks," in *Proceedings IEEE INFOCOM,* 2012, pp. 1692–1700.

[43] C.-L. Yang, P.-H. Kung, C.-A. Chen, and S.-D. Lin, "Semantically sampling in heterogeneous social networks," in *Proceedings of the 22nd international conference on World Wide Web companion*, 2013, pp. 181–182.

[44] M. Salehi and H. R. Rabiee, "A Measurement Framework for Directed Networks," *Selected Areas in Communications, IEEE Journal on*, vol. 31, no. 6, pp. 1007–1016, 2013.

[45] S.-W. Son, C. Christensen, G. Bizhani, D. V. Foster, P. Grassberger, and M. Paczuski, "Sampling properties of directed networks," *Physical Review E,* vol. 86, no. 4, p. 046104, 2012.

[46] M. Salehi, H. R. Rabiee, and A. Rajabi, "Sampling from complex networks with high community structures," *Chaos: An Interdisciplinary Journal of Nonlinear Science,* vol. 22, no. 2, pp. 023126–023126, 2012.

[47] Q. GAO, X. DING, F. PAN, and W. LI, "An improved sampling method of complex network," *International Journal of Modern Physics C,* 2013.

[48] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische mathematik,* vol. 1, no. 1, pp. 269–271, 1959.

[49] M. E. Newman, "Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality," *Physical review E,* vol. 64, no. 1, p. 016132, 2001.

[50] S. Milgram, "The small world problem," *Psychology today,* vol. 2, no. 1, pp. 60–67, 1967.

[51] J. Travers and S. Milgram, "An Experimental Study of the Small World Problem," *Sociometry,* vol. 32, no. 4, pp. 425–443, 1969.

[52] P. A. Estevez, P. Vera, and K. Saito, "Selecting the most influential nodes in social networks," in *Neural Networks, 2007. IJCNN 2007. International Joint Conference on,* 2007, pp. 2397–2402.

[53] W. Chen, C. Wang, and Y. Wang, "Scalable influence maximization for prevalent viral marketing in large-scale social networks," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining,* 2010, pp. 1029–1038.

[54] M. Kimura and K. Saito, "Tractable models for information diffusion in social networks," in *Knowledge Discovery in Databases: PKDD 2006,* Springer, 2006, pp. 259–271.

[55] J. Kleinberg, "Small-world phenomena and the dynamics of information," *Advances in neural information processing systems,* vol. 1, pp. 431–438, 2002.

[56] W. Hsu and A. Helmy, "On nodal encounter patterns in wireless LAN traces," *Mobile Computing, IEEE Transactions on,* vol. 9, no. 11, pp. 1563–1577, 2010.

[57] W. W. Zachary, "An information flow model for conflict and fission in small groups," *Journal of anthropological research,* pp. 452–473, 1977.

[58] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences,* vol. 99, no. 12, pp. 7821–7826, 2002.

[59] R. Guimera, L. Danon, A. Diaz-Guilera, F. Giralt, and A. Arenas, "Self-similar community structure in a network of human interactions," *Physical review E,* vol. 68, no. 6, p. 065103, 2003.

[60] P. M. Gleiser and L. Danon, "Community structure in jazz," *Advances in complex systems,* vol. 6, no. 04, pp. 565–573, 2003.

[61] J. Gehrke, P. Ginsparg, and J. Kleinberg, "Overview of the 2003 KDD Cup," *ACM SIGKDD Explorations Newsletter,* vol. 5, no. 2, pp. 149–151, 2003.

[62] J. Leskovec and J. J. Mcauley, "Learning to discover social circles in ego networks," in *Advances in neural information processing systems,* 2012, pp. 539–547.

[63] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Predicting positive and negative links in online social networks," in *Proceedings of the 19th international conference on World wide web,* 2010, pp. 641–650.

[64] J. Leskovec, "Stanford large network dataset collection," *URL http://snap. stanford. edu/data/index.html,* 2014.

[65] P. Erdos and A. Rényi, "On the evolution of random graphs," *Publications of the Mathematical Institute of the Hungarian Academy of Sciences,* vol. 5, pp. 17–61, 1960.

[66] Z. Bar-Yossef and M. Gurevich, "Random sampling from a search engine's index," *Journal of the ACM,* vol. 55, no. 5, pp. 24:1–24:74, 2008.