

Diffusion-Aware Sampling and Estimation in Information Diffusion Networks

Motahareh EslamiMehdiabadi
Sharif University of Technology
Email: eslami@ce.sharif.edu

Hamid R.Rabiee
Sharif University of Technology
Email: rabiee@sharif.edu

Mostafa Salehi
Sharif University of Technology
Email:mostafa_salehi@ce.sharif.edu

Abstract—Partially-observed data collected by sampling methods is often being studied to obtain the characteristics of information diffusion networks. However, these methods are usually done without considering the diffusion process behaviour. In this paper, we propose a novel two-step (sampling/estimation) measurement framework by utilizing diffusion process characteristics. To this end, we propose a link-tracing based sampling design which uses the infection times as local information without any knowledge about the latent structure of diffusion network. To correct the bias of sampled data, we introduce three estimators for different categories of characteristics; links-based, node-based, and cascade-based. To the best of our knowledge, this is the first study to introduce a complete framework measurements for diffusion networks. Our comprehensive empirical analysis over large synthetic and real datasets demonstrates that the proposed framework outperforms common sampling methods (BFS and RW) in terms of link-based characteristics by about 37% and 35% in average, respectively. We also show that an estimator has important role in correcting the bias of sampling from diffusion networks.

I. INTRODUCTION

Information diffusion is one of the noticeable topics that has been considered in large On-line Social Networks (OSN) such as Facebook, Twitter, and YouTube. The networks with pieces of information in different formats such as posts, tweets, and videos are called “information diffusion networks”. The growth in the size of these networks in recent years results in large information networks. For example, in March 2011, Twitter users were sending 50 million tweets per day [2]. Moreover, the latent structure of diffusion networks makes their analysis difficult. Although we usually figure out the time of getting some information by people, we can not find the source of information easily. Furthermore, in epidemic diseases, the infection shows itself when somebody becomes infected without determining who infected whom [1]. Therefore, it may be impossible or costly to obtain a complete picture of a large and latent diffusion network.

Partially-observed network data is often being studied to obtain the characteristics of these networks. The network resulting from such measurements may be thought of as a sample from a larger underlying network. As a result, the accuracy of the studies on diffusion network analysis depends on the estimation of the characteristics based on sampled network data.

One can divide the measurement framework of the network characteristics into two steps: 1) Sampling and 2) Estimation.

In the first step, data is collected from the network using a sampling method. The essential property of a sampling method that makes it appropriate for network inference is that its visiting probabilities should be known for all network elements. This allows sampled data to be weighted so that they accurately represent network data. In the estimation step, 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 outputs an estimate of an unknown population parameter. However, sampling and estimation in a network context introduce various potential complications.

In recent years, a considerable amount of research have been done on analysing the topological characteristics of large OSNs based on the sampled data of different networks such as Facebook [4], [5], Twitter [6], YouTube [7], and other large networks [8], [9]. However, considering the sampling approaches to study diffusion behaviours of social networks, apart from their topologies, is a remarkable issue that should be addressed. The previous work on diffusion data collection [12], [3], [13], [20] have used some well-known sampling methods such as Breadth-First Search (BFS) and Random Walk (RW) without considering the diffusion process behaviour. This leads to gathering redundant data and losing a part of diffusion data, that consequently decrease the performance of these sampling methods (refer to Figure 1). On the other hand, working with diffusion network directly is not applicable in many large real systems in the result of its undiscovered structure. Moreover, the previous studies assert that the characteristics of a sampled diffusion network are indicative of the same characteristics for the whole network. However, in fact, the obtained characteristics represent the sampled graph, not the original graph. Such problems can be compensated for in many cases by using appropriate estimators.

In this paper, we propose a novel two-step (sampling/estimation) framework, called “DNS”, to measure the characteristics of diffusion networks. To this end, we propose a link-tracing based sampling method that utilize diffusion process properties to find the way of traversing the network more accurate. Specifically, this method samples underlying network by moving from a node to one of its neighbours through an outgoing link with the probability of spreading infection over it. We calculate this infection probability by considering the cascades behaviour in diffusion networks. It

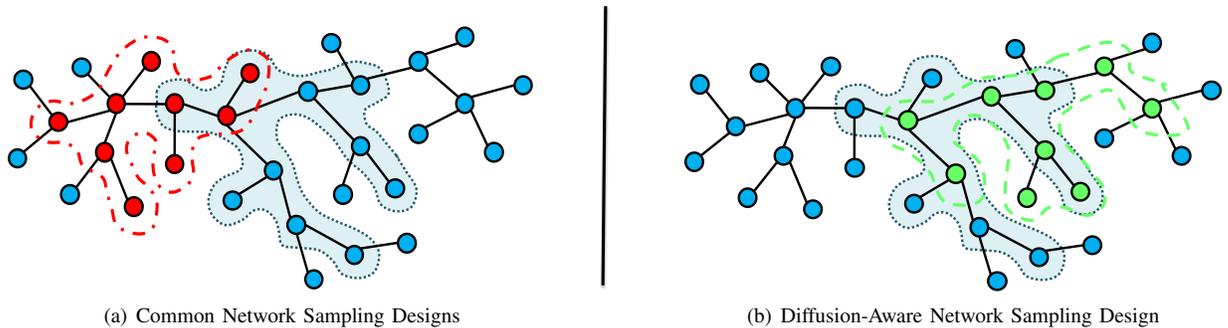


Fig. 1. An example of different sampling designs in diffusion networks. The region specified by dotted lines shows the diffusion network. The red and the green area demonstrates the sampled network obtained by a common and diffusion-aware networks sampling design, respectively. As it is shown, diffusion-aware network sampling design can cover the diffusion network more accurate.

is noteworthy that the algorithm only uses the infection times as local information without any knowledge about the latent structure of diffusion network. Moreover, we extend the well-known Hansen-Hurwitz estimator [38] to correct the bias of sampled data. We propose three efficient estimators related to different categories of network characteristics; links-based, node-based, and cascade-based. To the best of our knowledge, this is the first study to introduce a complete framework measurements for diffusion networks.

We evaluate the proposed framework over large synthetic and real datasets by comparing it with some sampling methods such as BFS and RW. The experimental results demonstrate that DNS outperforms the common sampling methods in terms of link-based characteristics by about 36% in average. We figure out that DNS has decreased the bias of the sampled data by 30% compared to the sampling design without estimation approach. Therefore, finding an appropriate estimator has an important role in correcting the bias of sampling method. Furthermore, the results show that the proposed framework can act very well even in low sampling rates. we also analyse the diffusion rate effect over the performance of the DNS. The analysis shows the independence of our proposed framework to the diffusion process behaviour. Hence, we can use DNS in various diffusion network with different diffusion patterns without any loss in its performance.

In summary, our main contributions can be summarized in the following:

- Proposing a novel sampling design for gathering data from a diffusion network by utilizing the properties of diffusion process.
- Proposing three estimators for correcting the bias of sampled data by computing the visiting probabilities of different types of diffusion characteristics (links-based, node-based, and cascade-based).
- Decreasing the bias of measuring link-based characteristics compared to other common sampling methods

The rest of the paper is organized as follows. Section II presents a classification of data collection approaches in the field of information diffusion networks. The problem formulation is proposed in Section III. The proposed measurement

framework is presented in Section IV. Section V elaborates the experimental evaluation, and the concluding remarks are provided in Section VI.

II. RELATED WORK

Diffusion process as a fundamental phenomenon over OSNs has attracted great attentions in recent years [1], [32], [14], [15], [11], [16], [36], [13], [18]. These attentions necessitate the study of diffusion data collection strategies. Here, we provide a comprehensive survey over the approaches used for collecting diffusion process data.

Complete Data: The first approach that comes to mind is collecting the complete diffusion data. Many diffusion process studies generate some diffusion paths and use them for analysis. Following the Iraq war petitions in the format of e-mail [11], [19], studying communication events between faculty and staff of a university by e-mails [15], and tracking the flow of information by extracting short textual phrases [16] are some examples of researches which were done by this approach. However, gathering diffusion data in many areas has some problems such as missing data, privacy policies, and impossibility of tracing all paths of diffusion. Moreover, large scale of diffusion networks is one of the most important obstacles of gathering diffusion data completely. Existence of these problems results in using sampling methods to obtain partial diffusion data.

Partial Data: Sampling methods can be considered an efficient way to tackle the problem of large-scale diffusion data. Using these methods to collect diffusion data have been studied in some recent work [12], [3], [13], [20]. The majority of these works have utilized one of most common sampling methods; Breadth-First Search (BFS). BFS is a basic graph-based sampling method that has been used extensively for sampling the networks in various domains [7], [4], [21], [6]. At each iteration of BFS, the earliest explored node is selected next. This method discovers all nodes within some distance from the starting node. Inferring diffusion topics from the DBLP database [20] and sampling the Twitter network to study on the resulting diffusion network [12], [13] are some examples which use BFS to collect diffusion data. However, BFS

leads to a bias towards high degree nodes [35] which this bias has not been analyzed so far for arbitrary graphs [22]. Despite the popularity of BFS, the problem of computing the visiting probabilities of network elements (such as nodes and links) in BFS sampling design is still unsolved. This is because of that sampling without replacement in BFS introduces complex dependencies between the sampled elements. To the best of our knowledge, there is no estimator to correct the sampling bias of BFS in a arbitrary network.

In spite of little attention on gathering partial data in the literature of diffusion behaviours, a considerable amount of research have been done on analysing the topological characteristics of the networks in various areas [4], [5], [6], [7], [8], [9]. Random Walk (RW) [23] is also one of the most important and widely used link-tracing sampling methods in different kind of network contexts such as uniformly sampling Web pages from the Internet [24], content density in peer-to-peer networks [33], [34], degree distributions of the Facebook social graph [4], [5] and in general large graphs [8]. 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 $\pi(u) = \text{deg}(u)/2|E|$, where $\text{deg}(u)$ and E are the degree of node u and are the set of links of the network graph. Moreover, the probability that a link is visited is $1/|E|$ (i.e., links are visited uniformly at random) [23].

Using these sampling methods without any attention to diffusion paths will result in some redundant data which are not related to diffusion process. Removing these unnecessary data decreases the efficiency of these sampling methods [3]. To the best of our knowledge, no work has been done before that considers diffusion process characteristics in sampling strategy. In this paper, we propose a diffusion-aware sampling and estimation methods which uses only local information of the underlying network. To the best of our knowledge, this is the first study to introduce a complete framework measurements for diffusion networks. Moreover, we use BFS and RW as baseline for comparison.

III. PROBLEM FORMULATION

A. Basic Notations and Definitions

Let network $G = (V, E)$ be the underlying network where V is the set of nodes, and E is the set of links which $n = |V|$ and $m = |E|$. In diffusion process, some diffusible chunks such as information and epidemic diseases propagate over G . These diffusible chunks are called “infection” when each path of infection will build a “cascade” [1], [18]. When the cascades spread over the underlying network, the diffusion network G^* will be formed.

We define $G_s = (V_s, E_s)$ as the induced sub-graph of G by sampling rate of μ when $V_s \subset V$ and $E_s \subset E$. In order

to analyse diffusion process, we should measure the diffusion characterization metrics from the sampled diffusion data. Since diffusion phenomenon covers many elements of the network (such as nodes, links, and cascades), we determine an “element set”, T , as a set of diffusion network elements [3]. Let L be a finite set of element labels. A label can be, for instance, the degree of a node, the weight of a link, and the length of a cascade. A label l_e is assigned to each element $e \in T$ by a target function $f : T \rightarrow L$, i.e. $f = \{(e, l_e) | e \in T, l_e \in L\}$. For example, infection is a label for each node that shows whether this node is infected during the diffusion process or not. The target function f for this label will match nodes $u \in V$ to the set $L = \{0, 1\}$ ($f(u) = 0$, if node u is not infected and $f(u) = 1$, otherwise).

Almost all network characterization metrics we are aware of can be expressed as some aggregative function. In this paper, we focus on the measurement of diffusion network characteristics. To this end, we consider the average function (η) over diffusion elements as:

$$\eta(f(G)) = \frac{\sum_{e \in T} f(e)}{|T|} \quad (1)$$

In the above infection example, this average shows the percentage of infected nodes by diffusion process to all the nodes of the underlying network.

B. Problem Definition

Our goal is to propose a diffusion-aware measurement framework to collect diffusion data in an efficient way. The diffusion process measurement procedure comprises two steps: (1) samples from the underlying network and computes the desired target function f on the sampled elements, (2) outputs an estimate of $Avg(f)$ by finding an appropriate estimator M . To evaluate the measurement framework, we define the bias metric as the following equation.

$$\rho = \frac{|\eta(f(G^*)) - \eta(f(M))|}{\eta(f(G^*))} \quad (2)$$

Now, our problem becomes equal to find a measurement framework which minimizes the bias, i.e. ρ .

IV. PROPOSED FRAMEWORK

In this section, we propose a diffusion-aware probabilistic measurement framework, called “DNS” (Diffusion Network Sampling), for the first time. This framework contains two parts of sampling and estimation which are explained in the following.

A. Sampling Design

In existing sampling methods such as BFS and RW, we begin at a starting node, and recursively visit (one or some) its neighbours as next nodes without considering the diffusion paths. Here, we try to utilize diffusion process properties to find the way of traversing the network more accurate. By computing the probability of spreading infections over the links of underlying network, we can direct the sampling design toward diffusion paths without any knowledge about diffusion

network structures. Therefore, it can cover a greater part of unknown diffusion network and decrease the redundant data such as some nodes and links which do not attend diffusion process.

To calculate the probability of spreading infection over a link, we focus on the cascades behaviour in diffusion networks. Each cascade c can be assigned to a time vector $t_c = \{t_1, t_2, \dots, t_n\}$ which shows the infection times of nodes by c . If cascade c does not infect a node, this node infection time will be considered ∞ [18]. The cascades with the same structure which propagates over the underlying network is shown by set C with N_c members. We define $C_T = \{t_1, t_2, \dots, t_{N_c}\}$ as the set of cascades' time vectors. We assume the transmission model of cascades follows independent cascade model [37]. In this model, a node gets the chance of transmitting information to its neighbours independently at each time episode.

When a node decides to infect one of its neighbours, it will do the transmission with a waiting time model that shows how far it will take that a node infects a chosen neighbour. In the proposed sampling method, we use exponential model [1] as waiting time model. By defining $\Delta = t_v - t_u$, the infection transmission probability over link $e(u, v)$ at cascade c in this model can be computed as follows.

$$P_c(e) = e^{-\frac{\Delta}{\alpha}} \quad \text{Exponential Model} \quad (3)$$

Where α is an adjustment parameter which determines how fast a cascade spreads. As it can be seen, the probability of spreading an infection have inverse relation to Δ . It is the symptom of a simple fact; when you receive an interesting E-mail, passing time will decrease the probability of forwarding it to your friends. Since diffusion network contains many cascades, each link $e(u, v)$ can attend more than one cascade. Therefore, C_e is defined as the set of cascades which pass over link e . Now, we define for each link e the probability of P_e , named infection probability, by getting the average probability of it over the cascades which attends them:

$$P_e = \frac{\sum_{c \in C} P_c(e)}{|C_e|} \quad (4)$$

The pseudo code of the proposed sampling design is shown in Algorithm IV.1. This method samples the underlying network by moving from a node, u , to a neighbouring node, v , through an outgoing link with the infection probability P_e . It is noteworthy that the algorithm only uses the infection times (i.e. C_T) as local information without any knowledge about the latent structure of diffusion network.

Algorithm IV.1: THE SAMPLING DESIGN(Seed, C_T, k, α)

```

v := Seed      %v is the current node
while (|E_s| < k)  %k is the sampling size
do {
  for each u ∈ Neighbours (v)
  {
    e := (v, u)
    V_s ← V_s ∪ u
    E_s ← E_s ∪ e
    for each c ∈ C_e
    {
      do {
        Δ = t_u - t_v
        P_c(e) = e-Δ/α
        P_e = P_e + P_c(e)
      }
      P_e =  $\frac{P_e}{|C_e|}$ 
      v ← u with probability of P_e
    }
  }
}
G_s := (V_s, E_s)
return (G_s)

```

B. Estimation Approach

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 [38], i.e. elements are weighted inversely proportional to their visiting probability. For any target function $f : T \rightarrow L$ that defines a characteristic (refer to Section III-A), the estimator of Equation 5 provides an asymptotic estimate of the population mean μ of f [39]:

$$\hat{\eta} = \frac{\sum_{i=0}^{k-1} \frac{f(X_i)}{\pi(X_i)}}{\sum_{i=0}^{k-1} \frac{1}{\pi(X_i)}} \quad (5)$$

Where X_i and $\pi(X_i)$ are the visited element (that could be nodes, links or cascades) and its visiting probability on i^{th} draw of sampling method, respectively. Therefore, to use this estimator, we should compute the probability of visiting each element in the proposed sampling procedure. In the following, we address this issue and extend above estimator for three different categories of elements; links-based, node-based, and cascade-based.

1) *Link-based Characteristics*: The links have a great role in spreading infection over the networks. Gaining some information without having any connection to others for propagation will be not valuable in a network. Therefore, link-based characteristics are the most important ones in diffusion process. "Link Attendance", as an example of link-based characteristics, shows the amount of presence in diffusion process for a link. The links with high attendance are significant in some applications such as finding potential paths of infection propagation in the epidemic spreading [3].

Since in the proposed sampling method we move over links with the probability of P_e (Equation 4), the visiting probability of link e will be equal to this probability; i.e. $\pi(e) = P_e$. We can use these visiting probabilities in Equation 5 to estimate the real value of link-based characteristics. As

mentioned before, we only use the local knowledge to compute the visiting probabilities of the links.

2) *Node-based Characteristics*: The number of ‘‘Seeds’’ (the beginners of an infection) [3] and ‘‘participation’’ (the fraction of users involved in the information diffusion) [12] are some examples of node-based characteristics. Diffusion process can be modelled as a Markov random walk over the underlying network G (the details can be found in our previous paper [18]). Therefore, the visiting probability of node u in the proposed sampling method can be defined as:

$$\pi(u) = \sum_{v \in N(u)} \pi(v)\pi(e_{vu}) \quad (6)$$

Where $N(u)$ is the set of node u ’ neighbours. The infection of node u at time t_u depends on the infection of its neighbours at the t_v where $t_v < t_u$. If we define $\pi = \{\pi(0), \pi(1), \dots, \pi(n-1)\}$, calculating π needs global knowledge of a network as it is the stationary distribution of the mentioned Markov chain. Since we do not have global view of the network in sampling procedure, finding the exact value of π is not applicable in real systems. Accordingly, finding an approximation of π can be considered as different research direction for future.

3) *Cascade-based Characteristics*: The cascades as the building blocks of a diffusion network can determine many characteristics of a diffusion process. For instance, the depth of a spreading phenomenon can be determined by the length of its cascades [3], [12]. Owing to the fact that each cascade c has a series of links which it spreads over them, its visiting probability depends on visiting all of its links [1]. Therefore, we can define $\pi(c)$ as the following equation.

$$\pi(c) = \prod_{e \in c} P_c(e) \quad (7)$$

This formula can be calculated by having all the links of a cascade. Since the probability of visiting a cascade needs global knowledge of a network, it should be approximated by using local information.

V. EXPERIMENTAL EVALUATION

A. Setup

As it is discussed in Section IV-B, computing the visiting probability of network elements should be done by only local information. Since calculating nodes and cascades visiting probability need global knowledge about underlying network structure, we evaluate Link-Attendance as a link-based characteristic. We use BFS and RW as baseline methods to compare with DNS.

To build diffusion network, many homogeneous cascades are generated with the same structure over the underlying network. The speed of cascades’ transmission is determined by α . To control the distance that a cascade can propagates, we use β parameter [1]. The fraction of the underlying network G which is covered by the diffusion network G^* is defined as diffusion rate δ .

B. Dataset

We utilize seven synthetic and real networks with different structures which their properties and the cascade generation settings are provided at Table I.

1) *Synthetic Dataset*: We use following models to generate synthetic data:

- Forest Fire model[26] is generated by the parameter matrix [5; 0.12; 0.1; 1; 0] which entries illustrate number of starting nodes, forward burning probability, backward burning probability, decay probability and probability of orphan nodes, respectively.
- The Kronecker graph[25] with three different Kronecker parameter matrix is generated: the Random graph [27] (by Kronecker parameter matrix of [0.9, 0.1; 0.1, 0.9]), the hierarchical network [28] ([0.5, 0.5; 0.5, 0.5]), and the Core-Periphery network [29] by ([0.9, 0.5; 0.5, 0.3]).

TABLE I
THE NETWORK AND CASCADE GENERATION PARAMETERS.

Network	n	m	α	β	δ
Forest Fire	10000	14305	0.7	0.5	0.5
Core-Periphery	8192	15000	0.7	0.1	0.5
Hierarchical	8192	11707	0.4	0.5	0.5
Random(ER)	8192	15000	0.4	0.4	0.5
PolBlog	1490	19090	1.3	0.5	0.5
Football	115	615	0.6	0.5	0.5
NetScience	1589	2742	0.6	0.5	0.5

2) *Real Dataset*: We used three real-world networks to evaluate our experimental results. The first network is based on links and posts of blogs in the political blogosphere around the time of the 2004 presidential election in US [30]. The other network is a network of American football games between Division IA colleges during regular season fall 2000 [40]. The last is co-authorship network of network theory scientists [31] which we named it NetScience.

C. Speed of Cascade

As mentioned, the speed of cascade propagation over the underlying network is controlled by α . In DNS framework, we use this parameter in calculating P_e to determine the direction of sampling design and correct the bias too. As the diffusion network structure is unknown in the most large real systems, the speed of cascades is not available to use in sampling and estimation approach. Therefore, we evaluate the DNS performance in measuring link-attendance characteristic by examining different values of α ($0.1 < \alpha < 3$) over synthetic and real networks in a fixed sampling rate ($\mu = 0.5$). As Figure 2 illustrates, all the networks have similar behaviour respect to α . It demonstrates that most networks achieve the minimum bias (below 10%) in measuring link-attendance characteristics when $0.4 \leq \alpha \leq 0.7$.

Nonetheless, the behaviour of political blog network is different in comparison with other networks to some extent. Analysing this network structure, it shows that the different behaviour is the result of the network density [26] difference.

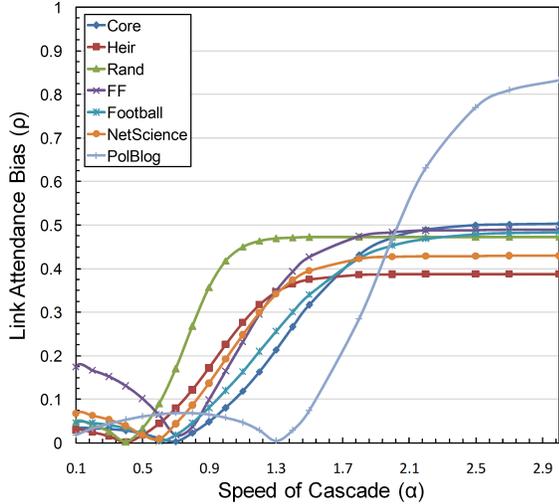


Fig. 2. Speed of Cascade

Comparing the density of political blog network with the other network shows that larger density needs larger α in sampling and estimation procedure. In fact, the higher density necessitates more speed in cascade transmission to visit the elements of the network in a time episode. Therefore, political blog network gets the least bias in $\alpha = 1.3$. The best appropriate α for each network is provided in Table I. We use these values as an input parameter of DNS in our experimental evaluation.

D. Performance Evaluation

In this section, we evaluate the performance of DNS framework in three aspects. First, we compare the bias of DNS and the baseline methods (BFS and RW) in measuring link-based characteristics. Second, we study the importance of estimation approach in the proposed framework. Finally, we analyse the behaviour of these methods in different sampling rates.

Figure 3 shows the results of measuring the link-attendance bias over the sampling rate. As it is observed, the proposed framework can measure this characteristic with very low bias (9% in average). We summarize the average performance difference of DNS with BFS and RW in all networks in Table II. It can be seen that DNS outperforms BFS and RW in terms of link-attendance by about 37% and 35% in average, respectively.

Interestingly, we can see that our proposed framework has decreased the bias by 30% compared to the sampling design of DNS without applying proposed estimation approach. These results confirm that the obtained characteristics from a sampled data represent the sampled graph properties, not the original graph. Therefore, an estimator has an important role in correcting the bias of sampling frameworks. However, this issue has not been considered in previous work on gathering diffusion data. Supporting this idea, we measure a node-based (Seed) and cascade-based (Depth) characteristic by the

sampling design of DNS without estimation. The results show that the proposed sampling design alone, can not perform as good as DNS. Specifically, it can only improve the bias about 12% and 9% in average compared to BFS and RW, respectively.

TABLE II
THE AVERAGE PERFORMANCE DIFFERENCE OF DNS WITH BFS, RW AND DNS WITHOUT ESTIMATION (DNS-WoE).

Network	BFS	RW	DNS-WoE
Forest Fire	49%	21%	14%
Core-Periphery	22%	24%	20%
Hierarchical	43%	45%	37%
Random(ER)	44%	45%	39%
PolBlog	34%	31%	31%
Football	35%	46%	37%
NetScience	30%	31%	22%
Average	37%	35%	30%

Moreover, Figure 3 demonstrates that the proposed method can act very well even in low sampling rates. DNS decreases the bias of measuring diffusion characteristics to 3% when $\mu < 0.3$. This promising result provides an appropriate sampling and estimation framework for the large real networks which only low sampling rates are applicable.

E. Diffusion Behaviour Analysis

The diffusion rate (δ) of infection over the underlying network has a significant role in gathering diffusion data. As this rate decreases, the less parts of the underlying network will be covered by the infection. Therefore, collecting the diffusion data gets more difficult. Here, we analyse the diffusion rate effect over the performance of the proposed method. Figure 4 illustrates that DNS leads to low bias even in low diffusion rates. Additionally, these results demonstrate the independence of our proposed framework to the diffusion process behaviour. Hence, we can use DNS in various diffusion networks with different diffusion patterns without any loss in its performance.

VI. CONCLUSIONS

In this paper, we introduced a novel two-step framework, DNS, to measure the characteristics of large scale and latent diffusion networks. We proposed a sampling algorithm that samples the underlying network by moving from a node to one of its neighbouring nodes through an outgoing link by considering the infection probability. Moreover, we proposed three estimators for correcting the bias of sampled data by extending the well-known Hansen-Hurwitz estimator. To this end, we computed the visiting probabilities of three types of diffusion characteristics; links-based, node-based, and cascade-based.

Our experiments showed that the proposed method outperforms BFS and RW in terms of link-attendance by about 37% and 35% in average, respectively. Moreover, we found out that the proposed estimator can improve the performance of the sampling design about 30%. Therefore, finding an appropriate

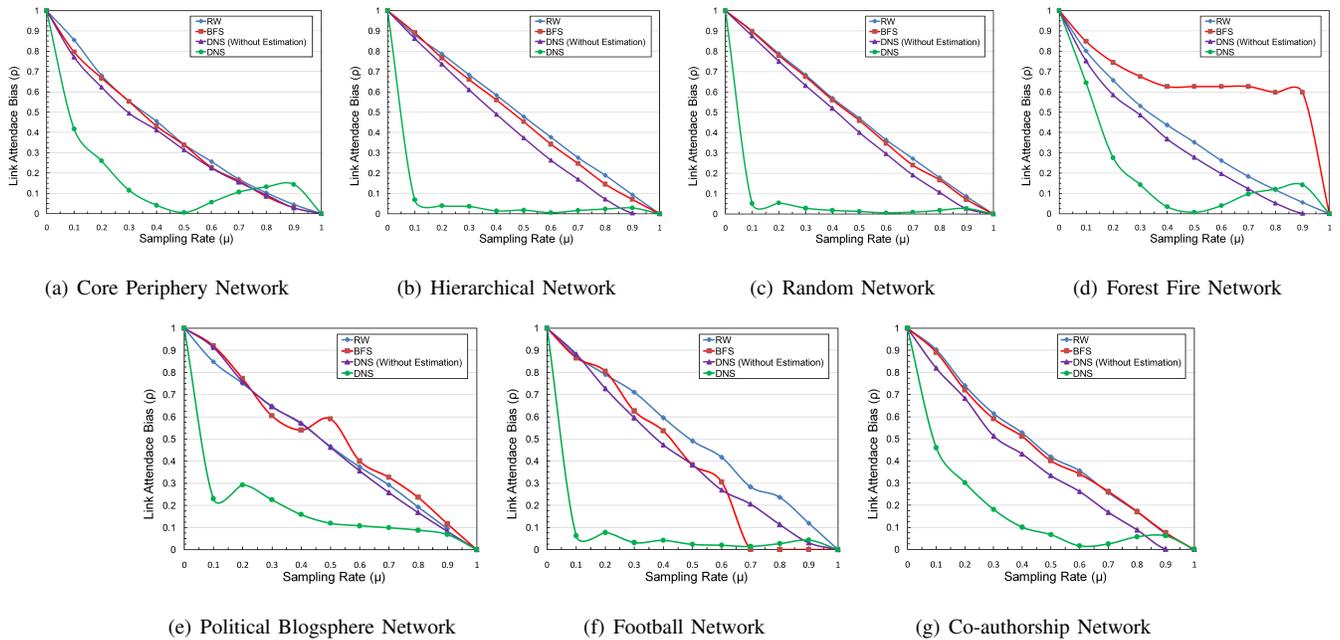


Fig. 3. Link Attendance characteristic evaluation in different sampling rates

estimation has an important role in correcting the bias. Furthermore, the results demonstrated that the proposed method can act very well even in low sampling rates. Additionally, our studies on the diffusion process behaviour shows that DNS leads to low bias even in low diffusion rates.

we believe that our results provide a promising step towards understanding the sampling approaches in analysis and evaluation of diffusion processes. There are several interesting directions for future work. Approximating the visiting probabilities of node-based and cascade-based characteristics is one of our main future goals.

VII. ACKNOWLEDGMENTS

This research has been partially supported by ITRC (Iran Telecommunication Research Center) under grant number 6479/500 (90/4/22).

REFERENCES

- [1] M. Gomez-Rodriguez, J. Leskovec and A. Krause, *Inferring networks of diffusion and influence*, In proc. of KDD '10, pages 1019-1028, 2010.
- [2] Twitter Blog: \neq numbers. [Blog.twitter.com.](http://blog.twitter.com/2011/03/numbers.html), Retrieved 2012-01-20, <http://blog.twitter.com/2011/03/numbers.html>.
- [3] M. Eslami, H.R. Rabiee and M.Salehi, *Sampling from Information Diffusion Networks*, 2012.
- [4] M. Gjoka, M. Kurant, C. T. Butts and A. Markopoulou, *Walking in Facebook: A Case Study of Unbiased Sampling of OSNs*, Proceedings of IEEE INFOCOM, 2010.
- [5] M. Gjoka, M. Kurant, C. T. Butts and A. Markopoulou, *Practical Recommendations on Crawling Online Social Networks*, IEEE J. Sel. Areas Commun, 2011.
- [6] M. Salehi, H. R. Rabiee, N. Nabavi and Sh. Pooya, *Characterizing Twitter with Respondent-Driven Sampling*, International Workshop on Cloud and Social Networking (CSN2011) in conjunction with SCA2011, No. 9, Vol. 29, pages 5521-5529, 2011.

- [7] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel and B. Bhattacharjee, *Measurement and analysis of online social networks*, Proceedings of the ACM SIGCOMM conference on Internet measurement, pages 29-42, 2007.
- [8] J. Leskovec and Ch. Faloutsos, *Sampling from large graphs*, Proceedings of the ACM SIGKDD conference on Knowledge discovery and data mining, pages 631-636, 2006.
- [9] M. Salehi, H. R. Rabiee, and A. Rajabi, *Sampling from Complex Networks with high Community Structures*, Chaos: An Interdisciplinary Journal of Nonlinear Science , 2012.
- [10] J. Leskovec, M. McGlohon, C. Faloutsos, N. S. Glance and M. Hurst, *Patterns of Cascading Behavior in Large Blog Graphs*, In proc. of SDM'07, 2007.
- [11] D. Liben-Nowell and J. Kleinberg, *Tracing information flow on a global scale using Internet chain-letter data*, Proc. of the National Academy of Sciences, 105(12):4633-4638, 25 Mar. 2008.
- [12] M. D. Choudhury, Y. Lin, H. Sundaram, K. S. Candan, L. Xie and A. Kelliher, *How Does the Data Sampling Strategy Impact the Discovery of Information Diffusion in Social Media?*, Proc. of ICWSM , 2010.
- [13] E. Sadikov, M. Medina, J. Leskovec and H. Garcia-Molina, *Correcting for missing data in information cascades*, WSDM, pages 55-64, 2011.
- [14] D. Gruhl, R. Guha, D. Liben-Nowell and A. Tomkins, *Information diffusion through blogspace*, In proc. of the 13th international conference on World Wide Web, pages 491-501, 2004.
- [15] G. Kossinets, J. M. Kleinberg and D.J. Watts, *The structure of information pathways in a social communication network*, KDD '08, pages 435-443. 2008.
- [16] J. Leskovec, L. Backstrom and J. Kleinberg, *Meme-tracking and the dynamics of the news cycle*, KDD '09: Proc. of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 497-506, 2009.
- [17] M.Gomez-Rodriguez, D. Balduzzi and B.Scholkopf, *Uncovering the Temporal Dynamics of Diffusion Networks*, Proc. of the 28 th International Conference on Machine Learning, Bellevue,WA, USA, 2011.
- [18] M. Eslami, H. R. Rabiee and M. Salehi, *DNE: A Method for Extracting Cascaded Diffusion Networks from Social Networks*, IEEE Social Computing Proceedings, 2011.
- [19] F. Chierichetti, J. Kleinberg and D. Liben-Nowell, *Reconstructing Patterns of Information Diffusion from Incomplete Observations*, NIPS 2011.
- [20] C.X. Lin, Q. Mei, Y.Jiang, J. Han and S. Qi, *Inferring the Diffusion and Evolution of Topics in Social Communities*, SNA KDD, 2011.
- [21] Ch. Wilson, B. Boe, A. Sala, K. P. N Puttaswamy and B. Y. Zhao,

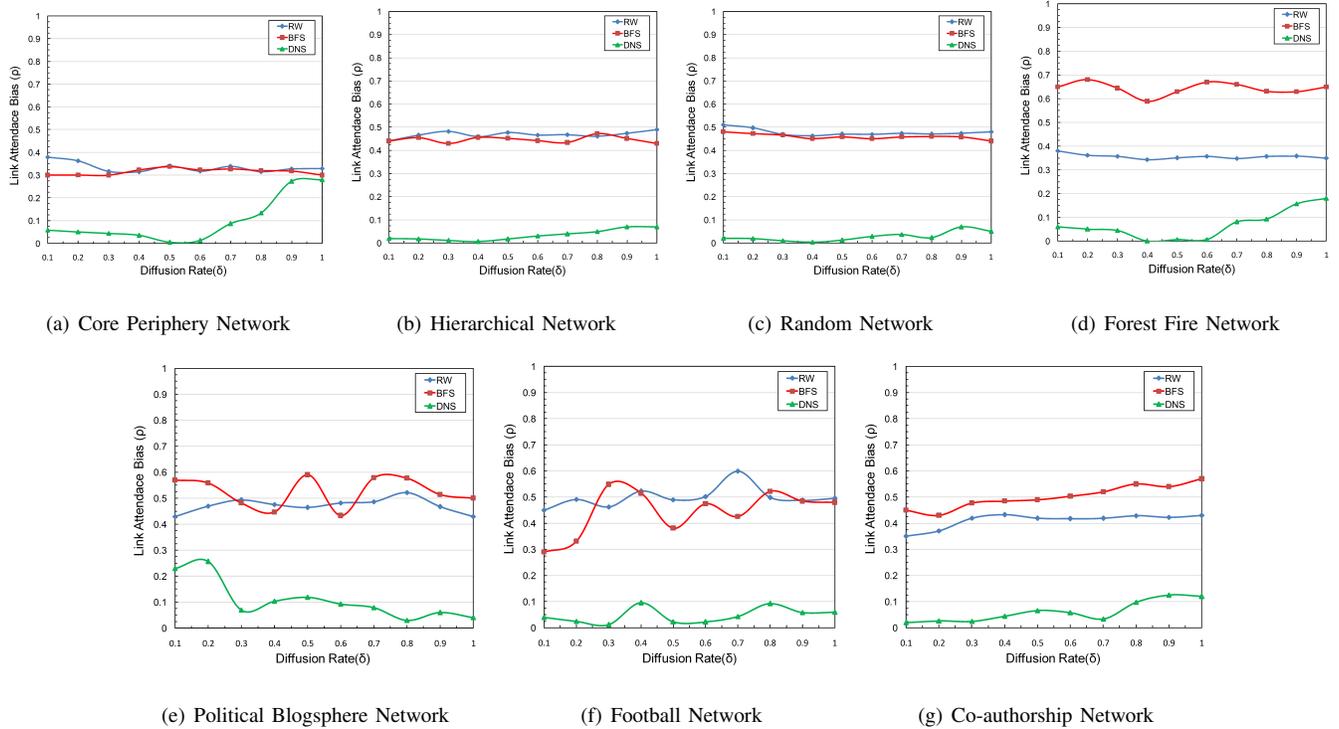


Fig. 4. Analysis of diffusion rate over sampling frameworks

- User interactions in social networks and their implications*, EuroSys '09: Proceedings of the 4th ACM European conference on Computer systems, pages 205–218, 2009.
- [22] M. Kurant, A. Markopoulou and P. Thiran, *On the bias of BFS (Breadth First Search)*, 22nd IEEE International Teletraffic Congress (ITC), pages 1–8, 2010.
- [23] L. Lovas, *Random walks on graphs: a survey*, Combinatorics, 1993.
- [24] M.R Henzinger, A. Heydon, M. Mitzenmacher and M. Najork, *On near-uniform URL sampling*, Proceedings of the World Wide Web conference on Computer networks, pages 295–308, 2000.
- [25] J. Leskovec and C. Faloutsos, *Scalable modeling of real graphs using Kronecker multiplication*, Proc. of ICML, pages 497–504, 2007.
- [26] J. Leskovec, J. Kleinberg and C. Faloutsos, *Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations*, Proc. of KDD, 2005.
- [27] P. Erdős and A. Rnyi, *On the evolution of random graphs*, Publ. Math. Inst. Hung. Acad. Sci., 5: page 17, 1960.
- [28] A. Clauset, C. Moore and M. E. J. Newman, *Hierarchical structure and the prediction of missing links in networks*, Nature, 453: pages 98–101, 2008.
- [29] J. Leskovec, K.J. Lang, A. Dasgupta and M.W. Mahoney, *Statistical properties of community structure in large social and information networks*, WWW, pages 695–704, 2008.
- [30] L.A. Adamic and N. Glance, *The political blogosphere and the 2004 US Election*, Proc. of the WWW-2005 Workshop on the Weblogging Ecosystem, 2005.
- [31] M. E. J. Newman, *Finding community structure in networks using the eigenvectors of matrices*, Preprint physics/0605087, 2006.
- [32] S.A. Myers and J. Leskovec, *On the Convexity of Latent Social Network Inference*, Advances in Neural Information Processing Systems, 2010.
- [33] Ch. Gkantsidis, M. Mihail and A. Saberi, *Random walks in peer-to-peer networks: algorithms and evaluation*, Elsevier Science Publishers B. V., Performance Evaluation, P2P Computing Systems, Vol 63, pages 241–263, 2006.
- [34] D. Stutzbach, R. Rejaie, N. Duffield, S. Sen and W. Willinger, *On Unbiased Sampling for Unstructured Peer-to-Peer Networks*, Proceedings of IMC, pages 27–40, 2008.
- [35] L. Becchetti, C. Castillo, D. Donato and A. Fazzone, *On the bias of BFS (Breadth First Search)*, LinkKDD, pages 1–8, 2006.
- [36] J. Yang and J. Leskovec, *Modeling Information Diffusion in Implicit Networks*, ICDM, IEEE Computer Society, pages 599–608, 2010.
- [37] D. Kempe, J. Kleinberg and E. Tardos, *Maximizing the spread of influence through a social network*, KDD '03: Proc. of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM Press, pages 137–146, 2003.
- [38] M. Hansen and W. Hurwitz, *On the Theory of Sampling from Finite Populations*, Annals of Mathematical Statistics, No. 3, Vol 14, 1943.
- [39] E. Volz and D. Heckathorn, *Probability based estimation theory for respondent-driven sampling*, Official Statistics, pages 79, Vol 24, 2008.
- [40] M. Girvan and M. E. J. Newman, *Community structure in social and biological networks*, Proc. Natl. Acad. Sci. USA 99, pages 7821–7826, 2002.