IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received March 23, 2020, accepted April 17, 2020, date of publication April 23, 2020, date of current version May 7, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2989504

Fast Representative Sampling in Large-Scale Online Social Networks

GUANGREN CAI, GANG LU, JUNXIA GUO, CHENG LING, AND RUIQI LI

UrbanNet Laboratory, College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China

Corresponding authors: Ruiqi Li (lir@mail.buct.edu.cn) and Gang Lu (lugang@mail.buct.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61903020, Grant 61702029, and Grant 61602026, and in part by the BUCT Talent Start-up Fund under Grant BUCTRC201825.

ABSTRACT Online social networks (OSNs) have become important platforms for efficiently connecting people and promoting information dissemination, which is of great importance to our social life and society. However, due to privacy concerns, and access limitations, it is difficult to obtain the whole network of OSNs for analysis, so it is critical to have a representative subgraph. Yet due to the same reasons, we are in lack of the original network as the ground truth which poses great challenges on evaluating sampling methods on the performance on unbiasedness, let alone representativeness. Thus uniform sampling (UNI) [Gjoka *et al.* 2010] was proposed to obtain an unbiased nodal property distribution as of the original network to evaluate the degree of bias of other methods. Yet UNI sampling suffers from its low efficiency, and the representativeness and connectivity of the obtained subgraph, which is formed by the sampled nodes and connections between them, are rarely studied. We propose an adaptive UNI sampling (adpUNI) method to overcome previously mentioned disadvantages of UNI by dividing the userID space into several intervals, whose sampling probability adaptively changes based on its target rate. By adding its neighbors of the targeted node into the sample set (adpUNI+N), we can further improve the performance on sampling efficiency and obtain a more connective and representative subgraph. When applied to Sina Weibo and Twitter, our methods over-perform other classical methods on sampling efficiency, and always have a better performance on connectivity and representativeness than UNI sampling. And we also find that an unbiased sample doesn't guarantee a more representative subgraph.

INDEX TERMS Complex networks, representative sampling, large-scale online social networks, UNI, adaptive method.

I. INTRODUCTION

Nowadays, complex networks are ubiquitous in our world, among which, online social networks (OSNs) play a crucial role in society by efficiently connecting people and promoting social interactions. The spreading of information, ideas, innovations and even behaviors all strongly rely on it [1]–[6]. Social network is the backbone of our society and of great importance to our social life and urban economy [7]–[9]. With the development of information and network technology, OSNs have grown rapidly over the past few decades and already have millions or even billions of users. For example, by the end of 2018, as one of the most popular social micro-blog platforms in the world, Twitter has 321 million monthly active users [10]; for Facebook, this figure is 2.32 billion [11]; Sina Weibo – the most popular micro-blog platform in China — has a size of 462 million monthly active users at the same time, and the total number of all the users is about 1.092 billion worldwide [12]. OSNs are typical instances of complex networks, and after the emergence of Web 2.0, OSNs have become a free-of-cost and efficient mass medium where users can present themselves to and interact with a wider public [7] which goes beyond a simple communicating channel. It attracted great attention from users, researchers and policy makers.

In order to better depict and understand the spreading and interacting dynamics on OSNs, collecting the data of its topology is the crucial first step, yet OSNs operators rarely provide complete data sets to researchers due to user privacy protection and business security; In addition, the giant size of OSNs also pose great technical challenges when handling such big data. Therefore, for many applications or researches

The associate editor coordinating the review of this manuscript and approving it for publication was Zubair Fadlullah.

on massive OSNs, obtaining a relatively small but representative sample network (we also refer to it as subgraph thereafter) is of great value.

There have been a variety of sampling algorithms on complex networks, which can be classified into three categories: graph traversal sampling, random walk sampling, and random selection sampling. Yet one critical issue for all sampling methods is the evaluation of the unbiasedness of the sampled nodes, let alone representativeness of the obtained subgraph (we will clarify these two concepts very shortly), which requires the original networks as the ground truth. However, this is usually not feasible for large scale OSNs due to privacy concerns, data protection, and access limitations, thus uniform sampling (UNI) [13] has been proposed as a solution to obtain a ''ground truth'' of the distribution of the original network to evaluate the bias of other methods. UNI sampling is proved to be an equal probability (i.e., uniform) sampling, which means that for each node, the sampling probability is the same regardless of the topology of the original network. Such uniformity can certainly guarantee the unbiasedness of sampled nodes on any concerned nodal property (such as degree, clustering coefficient, etc.). The *unbiasedness* is measured by the Kolmogorov-Smirnov (KS) distance (see Methods) between the nodal property distribution of the original network and those sampled nodes, whose property are the same as in the original graph (not the value calculated from the subgraph formed by the connections between them) [13]. For example, if there were 1000 nodes with degree 6 in the original network, then UNI sampling with a 10% eventual desired sample size will averagely give you 100 nodes with degree 6. Therefore the nodal property distribution of the sampled nodes obtained by UNI was used as a ground truth to evaluate the unbiasedness of other sampling methods [13], [14].

However, UNI sampling suffers from its low efficiency, when user IDs of the OSN are sparsely allocated in the userID space, which refers to the whole range of possible user IDs (it's usually either $[0, 2^{32}-1]$ or $[0, 2^{64}-1]$ depending on the design of the system). And the *representativeness*, which is measured by the KS distance between the distribution of the subgraph and original network on concerned nodal properties, of UNI method is rarely studied due to the lack of the original network for large scale OSNs. In this work, we find that the obtained subgraph by UNI, as well as MHRW (Metropolis-Hasting Random Walk), is not as representative as ours, though they are unbiased sampling methods [13], [15]. While, for researches and many other practical applications, we need to give a more representative sample network to end users.

In this paper, we propose some fast adaptive methods to overcome the above-mentioned defects of UNI sampling and to obtain a more representative subgraph, since eventually what the end users needed is a representative handy subgraph. We firstly analyze the userID space of Sina Weibo as of 2014, which contains about 470 million valid user IDs. We find that the distribution of valid user IDs of Sina Weibo is quite heterogeneous across its 64-bit userID space, some intervals are very sparse and some are quite dense. Since the sampling efficiency of UNI is mainly affected by the target rate (the total number of targeted nodes divided by the total sampling attempts, when a sampled ID is valid in the OSN system, we will add this node to the sample set, and refer it as a ''targeted'' node), we come up the idea of dividing the userID space into I equal-length intervals, and adaptively changing the sampling probability of each interval according to its current target rate, so that the algorithm samples more frequently in denser intervals (i.e., with more valid IDs in the interval) and less in sparser intervals. We call it adaptive UNI (adpUNI) sampling. In order to further improve the sampling efficiency, we further propose an adpUNI+N method which adds all its neighbors of the targeted node to the sample set, and all other settings are the same with adpUNI. We apply our methods and other classical sampling methods, including UNI, RN (Random Node sampling), MHRW, BFS (Breadth-First Search), and RW (Random Walk), to Sina Weibo and Twitter, and find that the sampling efficiency of our methods are much higher than other methods, and adpUNI+N has the best performance among all methods on sampling efficiency, and always have a much better representativeness and connectivity over UNI sampling method. By testing our methods and other classical ones on the uniformity, unbiasedness and representativeness, we also find that perfect uniformity can ensure an unbiased sampling of nodes, but not a more representative sampled subgraph. For example, our methods are not that uniform, yet the sampled nodes can still be relatively unbiased, and the obtained subgraph is more representative than others.

II. RELATED WORKS

Network sampling algorithms can be classified into three categories: graph traversal sampling, random walks sampling, and random selection sampling [13], [16].

Graph traversal sampling methods attempt to obtain the topology of the original network by traversing it, which mainly include Breadth-First Search (BFS) [17], Depth-First Search (DFS) [18], Forest Fire (FF) [19] and Snow-Ball Sampling (SBS) [20]. These methods all start from a randomly selected initial seed node and vary in the visiting order when sampling the nodes. BFS and DFS are the most basic network sampling algorithms [21]. When performing BFS, if the neighbors of the current node are visited based on a probability p, it becomes FF; BFS can be considered as an extreme case of FF when $p = 1$. Similarly, if only exact n neighbors are chosen randomly when performing BFS, it is called n-name SBS. According to a classic definition by Goodman [22], an n-name SBS is similar to BFS, for every sampled node, not all its neighbors but exactly n neighbors are chosen randomly, and these n neighbors are scheduled to be visited, but only if they have not been visited before. BFS has been widely applied in sampling OSNs [23]–[27], due to the belief that a full view of a particular region in the graph can be representative of the entire network [27].

However, many works have proved that BFS leads to a biased sampling towards high degree nodes [13], [28]–[31]. In [19], Kurant *et al.* quantified the degree bias of BFS sampling and showed that all commonly used graph traversal techniques (i.e., BFS, DFS, FF, SBS) lead to the same bias for random graphs. Here the bias was measured by the Kolmogorov-Smirnov (KS) distance between the nodal property (such as degree) distribution of the original network and sampled nodes, whose property are the same as in the original network (not the value calculated from the subgraph formed by the sampled nodes and connections between them). They also showed that when the original graph is random, it is possible to precisely correct this bias. Additionally, the bias can be reasonably corrected well even in more realistic graphs, but for a very small sample size. Nevertheless, the bias of BFS will be relatively small or even can be ignored without any correction when the sampled graph covers a very large part of the original one. However, it is difficult to interpret the results in all other cases [19].

In contrast to graph traversal sampling methods which can traverse the whole network, random walk sampling methods usually traverse a part of the original graph. Random walk sampling also start from a randomly selected initial seed node, then instead of traversing all or most of its neighbors, the algorithm chooses the next sampled node from its neighbors according to a certain criterion, and during the sampling, nodes can be revisited for several times. Random walk sampling methods have been employed for sampling the WWW [32], peer-to-peer networks [15], [33], [34], some large scale OSNs [16] including Twitter [35], Friendster [36], and Facebook [13]. The most basic random walk sampling (RW) randomly chooses one neighbor of the targeted node at each time step. The bias of RW can be analyzed by Markov Chains and corrected by re-weighting the estimator, and such method is called Re-Weighted Random Walk (RWRW), which has been demonstrated in the context of sampling peer-to-peer networks [33]. Alternatively, the RW can be modified using the Metropolis filter to achieve any desired distribution [37], [38], and if the desired distribution is uniform, the algorithm is called Metropolis-Hasting Random Walk (MHRW). It has been applied by Stutzbach *et al.* to the peer-to-peer network to obtain a nearly uniform representative sample [15]. In [13], the authors prove that RWRW and MHRW both can achieve unbiased sampling, and RWRW is faster than MHRW but also more complicated in the sampling process. While RWRW and MHRW ensure unbiased graph sampling, they suffer from their slow diffusion over the space which can in turn lead to poor estimation accuracy. In particular, their fully random nature in selecting the next node, when making a transition, often cause them to go back to the previous node from where they just come. This produces many duplicate samples for a short to moderate time span, thereby reducing estimation accuracy [39].

The third type is random selection sampling which can be further divided into sampling by random node (RN) selection and by random edge (RE) selection. RN and RE sampling methods differ in the way how the nodes and edges are selected, respectively. The most basic RN sampling selects a set of nodes uniformly at random from the original graph. Yet, there's evidence showing that RN cannot retain the power-law degree distribution of original networks [43]. There are also some improved methods based on basic RN sampling. Thus probability of selecting a node can be proportional to its degree (called Random Degree Node, RDN) or its Page-Rank value (Random Page-Rank Node, RPN) [44]. The idea behind RDN and RPN is to increase the selection probability of important nodes in the graph, either evaluated by degree or Page-Rank centrality [45]. Similar to RN sampling, RE sampling selects a set of edges uniformly at random from the whole edge set of the original graph, and both nodes connected to the selected edge will also be added to the sample set. The sample graphs obtained by RE tend to be biased since the high-degree nodes have more edges and thus have a higher probability to get chosen. Random Node Edge (RNE) sampling solves this problem by selecting a node at random and then selects some edges uniformly among the edges connected to the selected node [16]. Leskovec and Faloutsos [16] *et al.* applied a variety of methods (including RN, RPN, RDN, RE, RNE, RW, FF, etc.) to several large scale networks and showed that in terms of the degree distribution, RN and RPN are closer to the original network, RDN is biased towards nodes with high degree, and RE is farther away than RN and RPN from the original network. Lee *et al.* proved that RN sampling performs better than RE sampling regarding the clustering coefficient of the network [17]. However, all these random selection sampling methods require global information of the network, which hardly can be the case for large scale OSNs.

Thus UNI sampling has been proposed as a solution to obtain a "ground truth" of the distribution of nodal properties of large scale OSNs, when the complete topology is not accessible. UNI sampling has three steps: firstly, set the sampling range as $[0, userID_{max}]$ according to the coding rules of sampling objects (the $userID_{max}$ is usually either $2^{32} - 1$ or $2^{64} - 1$ depending on the design of the system); Secondly, an intended sample ID is randomly selected from the userID space with equal probability each time; Thirdly, make query in the OSN system, and if the ID exists in the actual network, it will be added to the sample set; otherwise, it will be discarded. This method is a textbook technique known as rejection sampling [47], which in general allows to sample from any distribution of interest. In particular, Gjoka *et al.* showed that it guarantees to select uniformly random user IDs from the OSN regardless of the actual distribution in the userID space (i.e., the sampling probability of any node will be the same whatever form of the distribution is) [13]. Therefore, the sampled nodes obtained by UNI will certainly be unbiased on any nodal properties against the original network, and thus the obtained distribution on nodal properties of sampled nodes by UNI was used as a "ground truth" for evaluating the unbiasedness of other sampling methods [13], [14]. In addition, the representativeness of

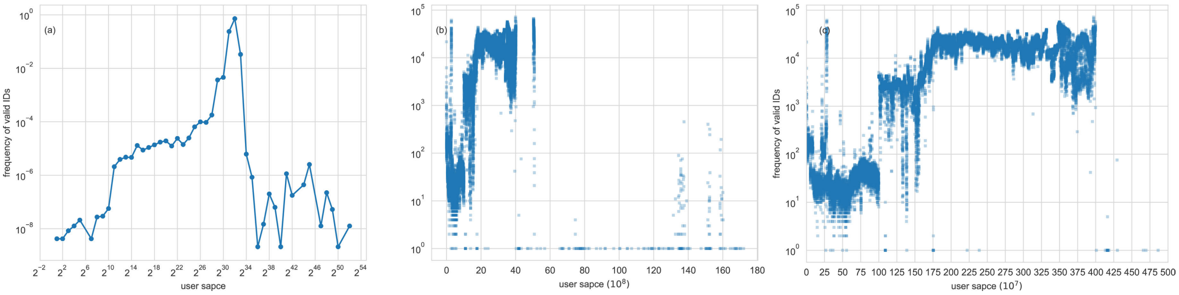


FIGURE 1. **The distribution of valid user IDs of Sina Weibo in (a) the whole userID space with logarithmic bins, (b) the range of [0, 2^{34}] and (c) [0, 5 × 10^{9}] with linear bins on the x-axis. The results indicate that the distribution of valid user IDs is quite heterogeneous. We only plot the intervals with non-zero values in (b) and (c); each dot represents the frequency of valid user ID within the interval.**

the sampled subgraph was rarely studied, and in this paper, we find that its subgraph is not as representative as ours. Besides, UNI sampling suffers from its low efficiency when the userID space is sparse. Now some OSNs, like Facebook and Sina Weibo, have upgraded their userID space from 32-bit to 64-bit, which is quite sparse on average and will lead to the inefficiency of UNI sampling. And as the topological properties of the original network was not considered, UNI sampling can't guarantee the connectivity of the sampled network. These all affect its applicability in practice.

III. METHODS

A. ANALYSIS OF THE userID SPACE OF SINA WEIBO

Sina Weibo was first launched by Sina Corporation in 2009 and filed an IPO as a separate entity in 2014, and now Sina Weibo is one of the biggest media platforms in China. We crawled the userID space dataset of Sina Weibo network as of 2014, which has about 473 million valid user IDs. Its userID space is 64-bit (they made the upgrade of the user system from 32-bit to 64-bit around 2011), so all user IDs are integers in the range of [0, $2^{64} - 1$]. We then analyze the distribution of valid user IDs in the userID space to have a comprehensive understanding (see Fig. 1(a) for the entire distribution in log scale). We find that most of them are in the interval of [2^{29}, 2^{34}] which accounts for around 99% of all valid user IDs; We then further divide [0, 2^{34}] into isometric intervals with a length of 100,000 and show the number of valid user ID in each intervals in Fig. 1(b). We can see that it's very dense in [0, 5×10^{9}], while quite sparse and even empty in some other intervals. We further divide [0, 5×10^{9}] evenly and show the distribution of valid IDs in Fig. 1(c), in which the distribution is still quite uneven.

B. adpUNI AND adpUNI+N SAMPLING METHODS

Based on the heterogeneity of Sina Weibo user ID distribution, we first propose an adaptive UNI (adpUNI) sampling method to improve the efficiency of UNI sampling. First of all, we divide the whole userID space into I equal-length intervals, among which valid IDs may be sparsely or densely allocated. The interval length $l = L/I$, where L is the length of entire user ID space (for Sina Weibo, $L = 2^{64}$). The whole sampling process is without replacement – at each time step t, a new random ID will be generated according to certain rules (see **Algorithm 1**), and if it falls in the interval i, it is called being sampled in the interval i, then the sampling time of the interval $S_i(t)$ will increase by 1. Here we can define the sampling rate of the interval $SR_i(t)$ at time t as

$$SR_i(t) = \frac{S_i(t)}{l}. \tag{1}$$

If this randomly generated ID is valid in the userID space, it is called being targeted, and this node will be added to the sample set, and the interval target time $T_i(t)$, which tells the number of sampled valid user IDs in the interval i, will also be added by 1. The interval target rate $TR_i(t)$ at time t will be

$$TR_i(t) = \frac{T_i(t)}{S_i(t)}, \tag{2}$$

which can tell us at time t, the fraction of targeted nodes among all the attempts in a certain interval i. Here we can see that in general, a higher density of valid user IDs in the interval will correspond to a higher target rate. This is why we need to divide the entire user ID space into I intervals. By dividing intervals, there would be intervals with dense or sparse valid IDs. And then in the sampling process, we design to make the algorithm multisample in the dense intervals, that is, the interval with high target rate, and vice versa.

Yet during the sampling process, there are two problems need to be tackled:

(1) Local optima traps. Intuitively, the higher the interval target rate is, the higher the probability of sampling in the corresponding interval will be assigned. However, if the sampling probability of the interval blindly sticks to its target rate, the algorithm will inevitably fall into some local optima traps – an interval with a high target rate will be sampled more frequently, even if the number of remaining valid user IDs in that interval has become less and less, which may affect the sampling efficiency.

(2) Cold start problem. In contrast to the local optima problem, if the interval sampling probability simply self-adapted according to the interval target rate, the sampling probability of a "cold" interval with a target rate of 0 would also be 0, which means that such an interval wouldn't be able to get "started", causing the valid IDs in that interval to be ignored.

In order to make the sampling probability of each interval change adaptivly with the target rate and overcome the local optima problem, we define the interval sampling probability $P_i(t)$ as

$$P_i(t) = TR_i(t-1) \times [1 - SR_i(t-1)], \tag{3}$$

which is a trade-off between target rate and the fraction of remaining sample-able range of the interval i, and thus can be adaptively changed along the sampling process. On one hand, $P_i(t)$ is proportional to the interval target rate $TR_i(t-1)$, the higher the $TR_i(t-1)$, the greater the $P_i(t)$ will be; on the other hand, the interval sampling rate $1 - SR_i(t-1)$ is an adjustment term which will gradually decrease. Thus the algorithm can avoid continuously sampling in intervals with a higher target rate.

As for the cold start problem, we set an initial minimum sampling probability $\alpha = 1/I$ to all intervals. During the sampling process, if the sampling probability of a certain interval $P_i(t)$ calculated according to Eq. (3) is too small (or even 0), it will be set as α, thus any cold interval can still get a sampling opportunity. And after each iteration, we will update the sampling probability of the targeted interval.

The detailed implementation steps of the adpUNI sampling algorithm are described in **Algorithm 1**.

In the sampling process of adpUNI, according to Eq. (3), the sampling probability of each interval is proportional to the corresponding target rate; therefore, the target rate in intervals with more valid user IDs is higher as well as the corresponding sampling probability, making the number of sampling attempts in dense intervals increase, however, decrease in sparse intervals, and finally the goal of improving the overall sampling efficiency is achieved.

In order to further improve the sampling efficiency and connectivity of the sampled network, when a valid node is targeted, we also add its neighbors into the sample set, and other settings are all the same. We call it adpUNI+N sampling. We performed some analysis on the adding neighbor process for both UNI and our adpUNI. In Fig. 2, we plot the absolute difference between the ID of the targeted node and the IDs of its neighbors of Sina weibo. We can see that the ID difference is not that much (most of them are within the range of 5×10^8 for both of UNI+N and adpUNI+N) and their peeks are on the left, which illustrate that most IDs and their neighbors are within the same or adjacent intervals. Since the dense intervals are usually continuous(see Fig. 7(b)), the sampling probability of corresponding intervals will increase by adding the neighbors of the target node, which can help the algorithm find dense intervals quickly and further improve the sampling efficiency. The code of our algorithms is available at https://github.com/UrbanNet-Lab/FRsamplingOSNs.

IV. EVALUATION METRIC

A. NETWORK STATISTICS

We use three network topology measures to investigate the representativeness of sampled networks.

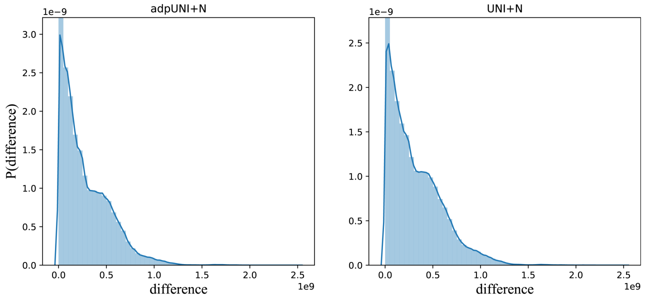


FIGURE 2. The distribution of absolute ID difference of targeted node and its neighbors of adpUNI+N and UNI+N in network of Sina weibo. Most of them are within the range of 5×10^8 for both of UNI+N and adpUNI+N, which is not that much, indicating that there is some correlation between the IDs of users with connections in the OSN.

1) DEGREE DISTRIBUTION

The degree of a node i in a network is the number of connections or edges the node has to other nodes, which can be defined as $k_i = \sum_j A_{ij}$ where A is the adjacency matrix. The degree distribution of a network can be formulated as

$$p(k) = \frac{|v_i \in V|k_i = k|}{n},$$

where v_i denotes the node i, V is the set of all nodes in the network, $|V| = n$ (n is the total number of nodes in the network).

2) CLUSTERING COEFFICIENT DISTRIBUTION

The clustering coefficient of node i is a measure of the extent to which nodes in a graph tend to cluster together, which is the ratio between the number of edges among its neighbors and the total number of all possible edges between them.

$$CC_i = \frac{\#existing\ links\ between\ i's\ neighbors}{\binom{k_i}{2}}.$$

The clustering coefficient distribution of a network G can be formulated as

$$p(c) = \frac{|v_i \in V|k_i > 1, CC_i = c|}{|v_i \in V|k_i > 1|},$$

where $|v_i \in V|k_i > 1|$ is the number of nodes with degree greater than 1.

3) CORENESS DISTRIBUTION

k-core (also called k-shell) decomposition starts by iteratively removing all nodes with minimum degree k_{min} until there is no node left with $k \leq k_{min}$ in the network. The removed nodes are assigned with kc = 1 and are considered in the first layer/shell. In a similar way, nodes with current minimum degree k_{min_2} are iteratively removed and assigned with $kc = 2$. This decomposition process continues removing higher shells until ending up with a complete graph with same degree (i.e., the central core, which will be removed at last) [48]. k-core method decomposes the network into ordered shells from the core to peripheries. Researchers found that core nodes of the network are more influential than the

Algorithm 1: adpUNI(L, I, DSS)

```
Input: $L$ (the range of the userID space), $I$ (the number of intervals), $DSS$ (desired sample size)
Output: A sample network (the set of targeted nodes and edges)
$\alpha = 1/I$; $l = L/I$; $intervalList = []$
Uniformly divide the entire user ID space $L$ into $I$ intervals with equal length $l$
for $i$ in range(I) do
    interval = Interval() #create an instance of Interval class
    $interval.S = 0$ #sampling time of the interval
    $interval.T = 0$ #target time
    $interval.P = \alpha$ #sampling probability
    $interval.lowerLimit, interval.upperLimit = i * L/I, (i + 1) * L/I$
    $intervalList$.append($interval$)
end
$sampledID$ = set() #save all tested IDs to ensure it's a sampling without replacement
$sampledNodes$ = set(); $sampledEdges$ = set()
while len(sampledNodes)<=DSS do
    sort all Interval instances in $intervalList$ according to its sampling probability $P_i$
    $TI$ = None #Targeted Interval
    for $interval$ in $intervalList$ do
        $RP$ = random.random()
        if $interval.P > RP$ then
            $TI = interval$
            break
        end
    end
    if $TI$ then
        $id$ = random.randint($TI.lowerLimit$, $TI.upperLimit$)
        while $id$ in $sampledID$ do
            $id$ = random.randint($TI.lowerLimit$, $TI.upperLimit$)
        end
    else
        $id$ = random.randint(0, $L$)
        while $id$ in $sampledID$ do
            $id$ = random.randint(0, $L$)
        end
        $TI = intervalList$.index(floor($id/l$)) #based on the index of the interval to get the targeted interval
    end
    $sampledID$.add($id$)
    $TI.S += 1$
    if $id$ exists in the system then
        $TI.T += 1$
        $sampledNodes$.update($id$)
    end
    $TI.P = TI.T/TI.S * (1 - TI.S/l)$ if $TI.T/TI.S * (1 - TI.S/l) > alpha$
end
for node $i$ in $sampledNodes$ do
    for node $j$ in $sampledNodes$ do
        make query to see if there's a connection
        if $i! = j$ and $E_{ij}$ then
            $sampledEdges$.update($E_{ij}$)
        end
    end
end
```

periphery ones [48]. The k-core is a metric of the connectivity and community structure of a graph [49], [50]. The core sizes also demonstrate the localized density of subgraphs in the graph [51]. The coreness distribution of the network G can be formulated as

$$p(kc) = (|v_i \in V|Core(v_i) = kc|)/n,$$

where Core(v_i) is the coreness of node v_i.

B. REPRESENTATIVENESS MEASUREMENT

As the goal of our methods is to sample a representative subgraph G_s from the original graph G such that the distance between the properties of G and G_s is small enough, so in this article, we employ Kolmogorov-Smirnov (KS) statistic for the distance measurement, which is widely used as a measure of the agreement between two distribution [52]. The KS statistic is calculated as the maximum vertical distance between two distributions,

$$KS = max_x|F(x) - F'(x)|,$$

where F and F' are two cumulative distribution functions (CDFs), and $0 \leq KS \leq 1$. In this paper, we utilize the KS statistic for computing the distance between the distribution of the original graph and the approximation distribution obtained from the sampled subgraph for the degree, clustering coefficient, and k-core distributions. The value computed by KS D-statistics is between 0 and 1. A smaller KS value indicates that the property of the sample graph is more similar to that of the original graph.

V. RESULTS

In order to verify the effectiveness of our methods, we employ two real-world datasets in this paper: a crawled sample network of Sina Weibo as of 2014, which includes around 3.58 million users and 14.84 million links; the same sample network of Twitter as the one used in [53], which includes 41.6 million users and 1.2 billion links. For both cases, the range of userID space is [0, 2^{32}-1]. And we regard them as the original network for future evaluation. And these two examples are also typical: one is quite sparse and heterogeneous over the whole userID space (Sina Weibo), the other is relatively denser and more uniform (Twitter) (see Fig. 7). Although both of them are directed networks, [54], [55] have shown that we can treat them as an undirected graphs when sampling, i.e., once there is a connection with any direction between two nodes, then it will be regarded as a bidirectional edge, which is the case when we deal with both Sina Weibo and Twitter.

In our experiment, we stored the data of nodes and edges of the original network in the Mysql database in advance. So UNI and our methods access the database to validate the existence of a random ID. In practice, we can directly call API interface provided by most social network platform, such as Sina Weibo and Twitter, to validate if a certain user ID is valid in the system or not [13].

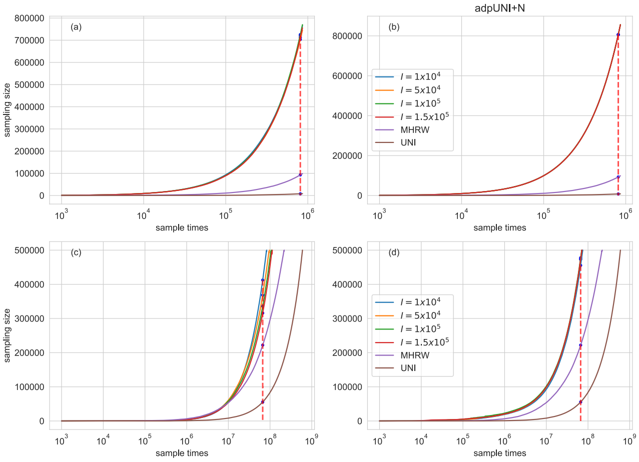


FIGURE 3. The number of targeted nodes as a function of logical sampling times of UNI, MHRW and adpUNI, adpUNI+N sampling with a different number of intervals for Twitter (a, b) and Sina Weibo (c, d). Each solid line is an average of 3 realizations (For MHRW, three crawlers start from three randomly selected nodes, its sampling size is the average of the three crawlers). We can see that, as the red dotted vertical lines indicate, adpUNI and adpUNI+N have higher sampling size under the same logical sampling time steps than UNI and MHRW on both Twitter and Sina Weibo.

A. ROBUSTNESS TESTS

Since the number of intervals I is a parameter of our methods, in this section, we study the impacts of I on sampling efficiency and performance on representativeness of sampled network for adpUNI and adpUNI+N.

1) SAMPLING EFFICIENCY

In Fig. 3, as the sample times increase, we show the change of sampling size (i.e., the number of targeted nodes) of our methods (adpUNI, adpUNI+N) and other rejection sampling methods (UNI, MHRW) for Sina Weibo (a, b) and Twitter (c, d). We set four different interval granularity ($I = 10^4$, 5×10^4, 10^5, 1.5×10^5, respectively). We can see that on both Twitter and Weibo, as indicated by the red dotted vertical lines, adpUNI and adpUNI+N have a much larger sampling size than that of MHRW and UNI with the same logical sampling times, showing that our methods are the most efficient ones. We also report the sampling times needed for each methods to reach a desired sample size (see Table 1), and we can see that adpUNI+N is the fastest one.

From Fig. 3(a), the sampling size of adpUNI slightly decreases with the increase of I (the number of intervals), and when the userID space is divided into 10,000 intervals, its sampling size is the highest for a fixed sampling time steps (see Fig. 3(a)), yet the differences between them on Twitter are almost negligible (see Fig. 3(c)). The slight difference between adpUNI on Weibo (see Fig. 3(a)) and Twitter (see Fig. 3(c)) can be explained by the user ID distributions of the original networks (see Fig. 8). As Twitter is more uniform, so the density heterogeneity of different intervals is not that strong, and thus it's less sensitive to the number of intervals I. And adpUNI+N is more robust to I on sampling efficiency.

TABLE 1. Total sampling times needed (in a unit of 10^8) for adpUNI, adpUNI+N, UNI and MHRW with a sample size of 5×10^5 for Sina Weibo and 10^5 for Twitter. The digits below adpUNI and adpUNI+N are the number of intervals *I*. The impacts of *I* on the logical sampling time of adpUNI+N is almost negligible. adpUNI+N is the fastest one among all these methods.

	adpUNI				adpUNI+N				UNI	MHRW
	10^4	5×10^4	10^5	1.5×10^5	10^4	5×10^4	10^5	1.5×10^5		
Sina weibo	0.83	0.97	1.07	1.13	0.74	0.71	**0.70**	**0.70**	5.99	2.19
twitter	0.108	0.113	0.111	0.109	**0.101**	**0.101**	**0.101**	**0.101**	10.30	0.873

It is worth noting that in Fig. 3 and Table 1, the "sample time" refers to the logical sampling time step, not the real time consumed by implementing each algorithm. In this paper, we perform all our experiments on a DELL server with 48 CPU cores with 2.1GHz and memory size of 256GB. In practical applications, the most time-consuming operation in each logical sampling time step is to access the database to obtain neighbor nodes. MHRW needs to access the database during each sampling step, while UNI, adpUNI, and adpUNI+N do such operation only when they target a node, so MHRW takes much longer time per logical sampling steps. Although MHRW has higher sampling size than UNI with the same sampling logical time steps in Fig. 3, it is actually the slowest one in practice. Even for just sampling 0.1 million nodes on Twitter, it took us more than two weeks.

We didn't make comparisons with other classical sampling methods (such as BFS, RN and RW) on samping efficiency is due to the fact that these methods are not rejection sampling, and thus there is no comparability between them. Overall, adpUNI and adpUNI+N have a much higher sampling efficiency than UNI and MHRW, thus when the desired sampling size is larger, our methods have more comparative advantage.

2) PERFORMANCE ANALYSIS

We then study the impacts of interval number I on the sampling performance on representativeness in terms of KS distance (see Methods) on degree, clustering coefficient and k-core distributions between the original network and sampled networks.

For different values of I, we still report the average situation over 3 realizations with varying sampling sizes ranging from 10^5 to 5×10^5 for Sina Weibo (see Fig. 4), and 10^6 to 5×10^6 for Twitter (see Fig. 5). We find that, generally speaking, the impacts of I are much less significant than sample size. adpUNI has the smallest KS distance when $I = 1.5 \times 10^5$ and the largest KS distance when $I = 10^4$ on almost all three topology measures for Sina Weibo (see Fig. 4(a-c)), while it is almost the opposite for Twitter (see Fig. 5(a-c)). The reason behind it requires further investigations in the future. Yet for most cases, such differences are not that significant. As for adpUNI+N, the impact of interval numbers I is negligible (see Fig. 4(d-f) and Fig. 5(d-f)). And adpUNI+N has a much smaller KS distance on all three measures than adpUNI on both Sina Weibo (see Fig. 4) and Twitter (see Fig. 5). And we can clearly see that with the increase of sampling size, the performance of both methods are all becoming better.

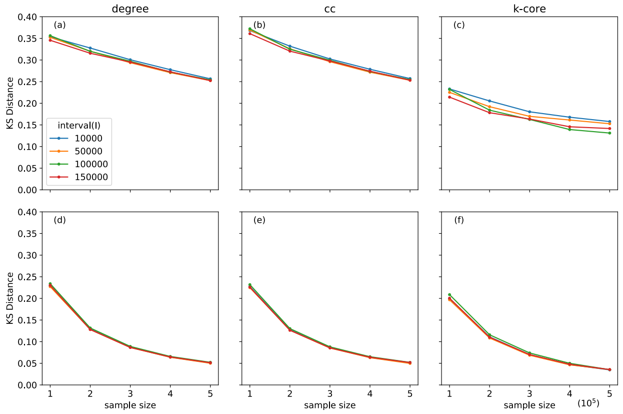


FIGURE 4. The impact of interval numbers *I* on sampling performance on representativeness in terms of KS distance of (a-c) adpUNI and (d-f) adpUNI+N on degree, clustering coefficient, and coreness distributions, respectively, for Sina Weibo. By comparing the results of adpUNI and adpUNI+N, we can see that adpUNI+N is quite robust to the free parameter *I* (i.e., the number of intervals), and it has a much smaller KS distance compared to adpUNI on all three indicators. When sampling size increases, the performance of both methods become better, and adpUNI+N has much greater improvement than adpUNI, whose KS distance decrease about twice as of adpUNI on Weibo.

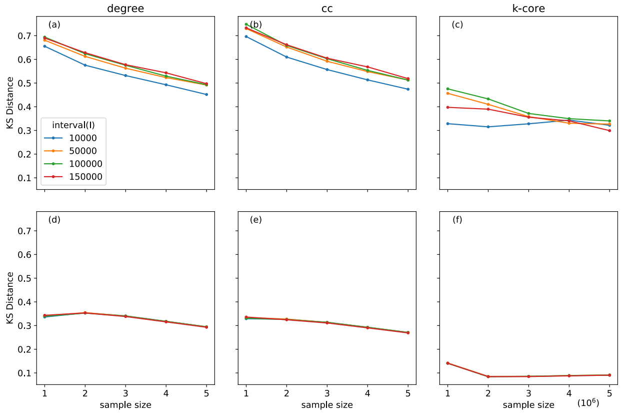


FIGURE 5. The impact of interval numbers *I* on sampling performance on representativeness in terms of KS distance of (a-c) adpUNI and (d-f) adpUNI+N on degree, clustering coefficient, and coreness distributions, respectively, for Twitter.

B. THE RELATIONSHIP OF UNIFORMITY, UNBIASEDNESS AND REPRESENTATIVENESS

UNI sampling has been theoretically proven to be an uniform sampling (i.e., the sampling probability for every node is the same), and thus the distribution of sampled nodes on any nodal properties (such as degree, clustering coefficient, etc.) will be unbiased against the original network. The obtained distribution on nodal properties of sampled nodes by UNI was used as a ground truth for evaluating the unbiasedness of other sampling methods for large scale OSNs. And many other algorithms have been proved to be unbiased as well, such as MHRW and RWRW [15]. Yet the question worth studying is that whether such uniformity or unbiasedness can guarantee higher representativeness or not. Therefore in this section, we will explore the relationship of uniformity, unbiasedness and representativeness for sampling methods. Here we explain the meaning of the three conceptee in the previous sentence again. The "uniformity" means the sampling probability for every node is the same; The "unbiasedness" is the KS distance between the set of sampled nodes (the properties of nodes are directly obtained from the original network) against the original network; and the "representativeness" is regarding the KS distance between the sampled subgraph (the properties of nodes are calculated from the subgraph) against the original network. Next, we first compare the uniformity and unbiasedness of our methods with other classical sampling algorithms.

In Fig. 7, we plot the ratios of the number of targeted IDs to the number of valid user IDs in each interval in the original network of Twitter and Weibo. If the result is a perfect horizontal line, then it means that the uniformity is achieved by the method. From Fig. 7, we can clearly observe that UNI method has almost the same sampling ratio in each interval on both Twitter and Weibo, so it is indeed an uniform sampling; While adpUNI has higher sampling ratios in denser intervals, and lower ratios in sparser ones, which, on the one hand, validates our adaptive design idea, on the other hand, shows adpUNI is not an uniform sampling. And uniformity is a stricter requirement than unbiasedness, as MHRW is proved to be unbiased [15], it still can be not that uniform (see Fig. 7(b)).

The unbiasedness is measured by the KS distance between the distribution of the original network and the sampled nodes (i.e., we directly use the nodal properties of sampled nodes as in the original network, not the properties calculated from the subgraph, which is consisted of the sampled nodes and connections between them). We test our methods on unbiasedness against other classical algorithms (see Fig. 6 and Table 2), and we can see that UNI is of the lowest bias as predicted, but our methods also have a quite good performance, and adpUNI is always more unbiased than MHRW on both Twitter and Weibo. And this further indicate that although perfect uniformity can guarantee unbiasedness, the degree of uniformity is not directly related to the unbiasedness. Even a method is not that uniform, it still can be unbiased (see the comparison between adpUNI and MHRW on Weibo, though MHRW is closer to uniformity than adpUNI, yet it's more biased). And from the Twitter case, when the sample size is 10^6, most methods have similar uniformity (see Fig. 7), yet their unbiasedness varies much larger (see Table 2).

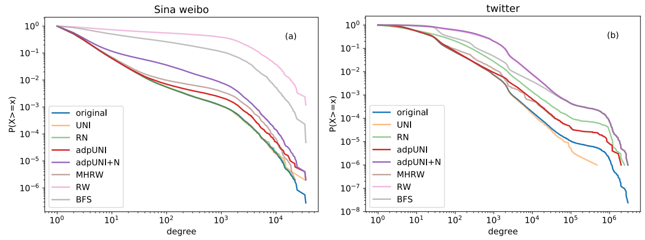


FIGURE 6. The comparisons between the original network and sampled nodes on cumulative distribution of degree with a sample size of 5 × 10^4 for Sina Weibo and 10^6 for Twitter. We can see that UNI is always the closest one to the original network, and adpUNI is also quite comparable. And it's worth noting that some methods are not that stable on unbiasedness on different networks. Note that $P(X >= x)$ in (a, b) refers to Complementary Cumulative Distribution Function (CCDF).

Then we come to the more important question: Can uniformity or unbiasedness guarantee higher representativeness? The representativeness is measured by the KS distance between the distribution of the original network and the sampled network (i.e., for sampled nodes, we calculate their nodal properties from the obtained subgraph). We still compare our methods with other classical sampling methods, including UNI, MHRW, BFS, RW, RN. In Fig. 8, we plot the cumulative distributions of degree, clustering coefficient and k-core of the original network and sampled networks, and report their KS distances in Table 3 with a sample size of 5 × 10^5, which is around 15% of the original Weibo network. We can see that adpUNI+N are among the closest ones to the original network on all three measures. The results on Twitter (see Fig. 9 and Table 4) also show that our methods over-perform UNI, though our methods cannot keep uniformity and have relatively larger bias compared with UNI. This indicate that representativeness is not monotonously related to uniformity or unbiasedness.

C. REPRESENTATIVENESS ANALYSIS

In this section, we compare and analyze the representativeness of UNI, adpUNI, adpUNI+N and other classical network sampling methods, including RN, BFS, RW and MHRW. We plot the cumulative distributions of degree, clustering coefficient and k-core of the original network and sampled networks of Sina Weibo in Fig. 8 and Twitter in Fig. 9. And also we calsulate the corresponding KS distance of Weibo in Table 3 and Twitter in Table 4.

For degree distribution, though the sample network obtained by adpUNI+N has smaller fraction of nodes with degree range from roughly 10 to 100 (corresponds to a slower decrease in CCDF, see Fig. 8(a)), and relatively larger fraction of high degree ones (i.e., a steeper decrease), and it is still much closer to the original network than other classical methods measured by KS distance (see Table 3). RN, UNI,

TABLE 2. Unbiasedness analysis of different methods. In this table, we report the KS distance between the degree distribution of the sampled nodes and the original network with a sample size of 5×10^4 for Sina Weibo and 10^6 for Twitter. We also tested our methods on Twitter with a sample size of 5×10^6, and find the KS distance of adpUNI and adpUNI+N drops to 0.0062 and 0.3851, respectively, which shows that with an increasing sample size, our methods also become unbiased. It's worth noting that some methods are not that stable on unbiasedness on different networks. Top three methods with comparable performance are marked in bold.

	UNI	RN	adpUNI	adpUNI+N	MHRW	BFS	RW
Sina weibo	**0.0008**	**0.0021**	**0.0056**	0.0525	0.0432	0.4829	0.6696
Twitter	**0.0009**	0.1758	**0.0269**	0.5837	**0.0785**	0.5225	0.5595

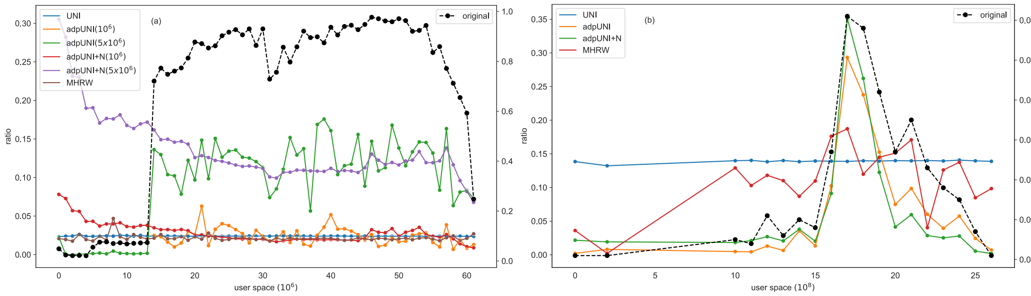


FIGURE 7. The ratio of the number of targeted IDs to the number of valid user IDs in each interval with linear bins for (a) Twitter with a sample size of 10^6 (around 2.5% of the original network) for UNI and MHRW due their low sampling efficiency, and 10^6 (orange line, 2.5%), 5×10^6 (green line, 12.5%) for adpUNI, respectively; (b) Sina Weibo with a sample size of 5×10^5 (around 15%) for both methods. The results indicate that the sampling ratios of UNI are quite uniform over all intervals regardless of original user ID density, while adpUNI has a higher sampling ratio in denser intervals and vice versa. The left y-axis corresponds to the sample ratio of different methods, and the right y-axis depicts the ratio of the number of valid user ID in a certain interval to the length of an interval I in the original network. We can see that valid user IDs of Twitter is relatively dense in most intervals. In this figure, we only plot the intervals with non-zero values, and lines are just a guidance to eyes. For adpUNI and adpUNI+N, with each setting of interval number $I = 10^4$, 5×10^4, 10^5, 1.5×10^5, we perform three independent experiments, and get the average of all realizations as shown in the figure.

and adpUNI they have quite similar performance on both degree and k-core distributions. As we know that clustering coefficient distribution captures a local structure, while coreness distribution captures a more global structure. We can see that the maximum coreness of the sampled networks by UNI, adpUNI, and RN are much smaller than the original one, which indicate that they fail on preserving such global structure. The sampled networks by BFS, RW, and MHRW clearly have smaller fraction of low degree or even leaf nodes and connections between different shells which leads to larger deviations than adpUNI+N on coreness distribution and clustering coefficients distribution at the smaller value region. Among all methods, adpUNI+N is the closest one to the original network. In Fig. 9, we also plot the cumulative distributions of degree, clustering coefficient and k-core of the original network and sampled networks with a sample size of 10^6 for Twitter (around 2.5% of the original network). Due to the fact that MHRW takes too long to even just sample one million nodes on Twitter, MHRW is not compared here. We can see that adpUNI+N is the closest one on clustering coefficient distribution (see Fig. 9(c) and Table 4), and always significantly over-perform UNI sampling (usually the sample network obtained by adpUNI+N is at least twice closer to the original network, see Table 4). By comparing the results on Weibo and Twitter, we can see that when the size of original network is larger and denser in user ID space(the case of Twitter), RN have a better performance on degree and k-core, yet due to its small sample size (2.5%) on Twitter, the local feature such as clustering Coefficients distribution is not preserved that well. And it's worth noting that RN need the global information of the network which is usually infeasible in large scale OSNs, while our methods don't require such global information but only the range of the userID space. Even on Twitter, our methods still have similar performance with the best one, whose KS distance is quite comparable.

D. CONNECTIVITY ANALYSIS

adpUNI + N is proposed on the one hand to further improve the sampling efficiency and on the other hand to improve the connectivity of the sampled network. Thus in this section, we aim to compare the connectivity of sampled network of UNI, adpUNI and adpUNI+N.

The selection process of adpUNI+N didn't explicitly consider the connectivity, but the process of adpUNI+N is in line with explosive percolation [56], [57], which has a critical phase transition point(ie.percolation threshold), and when the fraction of added nodes or links is beyond the critical value, there will be a giant component emerging. It has been proved that on networks having power law degree distributions (scale-free networks) with exponent λ smaller than 3,

TABLE 3. KS distance for all methods on degree, clustering coefficient (cc) and coreness distribution with the same sampling size 5 × 10^5 for Sina Weibo.

	degree	rank	cc	rank	k-core	rank
UNI	0.2381	3	0.1147	3	0.2390	3
adpUNI	0.2524	4	0.1417	4	0.2532	4
adpUNI+N	**0.0501**	**1**	**0.0355**	**1**	**0.0497**	**1**
RN	0.2335	2	0.1022	2	0.2340	2
BFS	0.3502	5	0.2690	7	0.3502	5
RW	0.3599	6	0.2159	6	0.3598	6
MHRW	0.3791	7	0.1822	5	0.3791	7

TABLE 4. KS distance for all methods on degree, clustering coefficient (cc) and coreness distribution with the same sampling size 1 × 10^6 for Twitter.

	degree	rank	cc	rank	k-core	rank
UNI	0.6700	6	0.4312	4	0.7358	6
adpUNI	0.6554	5	0.3285	3	0.6963	5
adpUNI+N	0.3361	4	**0.1412**	**1**	0.3289	3
RN	**0.2392**	**1**	0.4867	5	**0.2548**	**1**
BFS	0.3046	2	0.7304	6	0.3070	2
RW	0.3301	3	0.2148	2	0.3349	4

TABLE 5. Comparision of the fraction of edges(P) in the sampled network with the percolation threshold(*Pc*) for UNI, adpUNI and adpUNI+N. λ is the exponent of the degree distribution of the original network for Sina weibo and Twitter. S is the fraction of nodes in the giant commponent of the sampled network of each sampling methods for Weibo and Twitter. It can be seen the edge fraction of adpUNI+N is quite greater than the corresponding percolation threshold for both Weibo and Twitter, while UNI and adpUNI are the opposite. And accorrding to the fraction of nodes(S) in the giant component of the sampled network of each method, it is obvious that the connectivity of adpUNI+N is the best.

	Weibo($\lambda = 2.18, Pc = 0.03$)			Twitter($\lambda = 2.33, Pc = 0.05$)		
	UNI	adpUNI	adpUNI+N	UNI	adpUNI	adpUNI+N
P	0.019	0.012	**0.257**	0.445×10^{-3}	0.674×10^{-3}	**0.127**
S	0.29	0.31	**0.99**	0.25	0.32	**0.99**

the fraction of nodes or links to be removed from the graph for it to have no giant component tends to 1 in the limit of infinite network size. From the perspective of percolation and focusing on links, this also means: a scale-free network with λ<3 is kept connected by a vanishing fraction of randomly chosen links; i.e., the percolation threshold is zero. For λ>3, instead, a finite threshold appears [57]. We show the fitting of the complementary cumulative distribution(ccdf) of the original network for Sina weibo and Twitter in Fig. 10. In the Table 5, we report the exponent λ(λ) of the degree distribution of the original network for Sina weibo and Twitter and the corresponding percolation threshold(*Pc*) according to the result of [57]. It can be seen that the exponent of them are all smaller than 3 and the relative percolation thresholds are also very small. Then we calculate the ratio of edges in the sampled network of each sampling method to

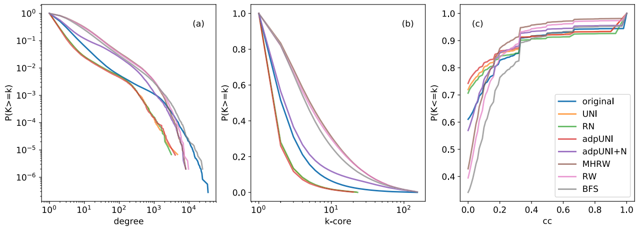


FIGURE 8. (a) Degree, (b) k-core and (c) clustering coefficient (cc) distribution of the original network, and sampled networks obtained by our methods (adpUNI, adpUNI+N) and other classical methods (UNI, RN, RW, MHRW, BFS) with the same sample size 5 × 10^5 for Sina Weibo. AdpUNI+N is the closest ones to the original network on all three topology measures. Note that $P(K >= k)$ in (a, b) refers to Complementary Cumulative Distribution Function (CCDF), and $P(K <= k)$ in (c) refers to CDF.

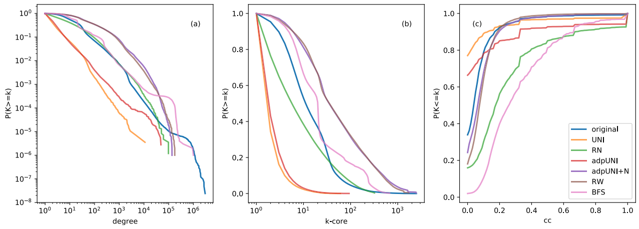


FIGURE 9. (a) Degree, (b) k-core and (c) clustering coefficient (cc) distribution of the original network, and sampled networks obtained by our methods (adpUNI, adpUNI+N) and other classical methods (UNI, RN, RW, BFS) with the same sample size 1 × 10^6 for Twitter. adpUNI+N is still quite comparable to the closest ones to the original network on all three topology measures. Note that $P(K >= k)$ in (a, b) refers to Complementary Cumulative Distribution Function (CCDF), and $P(K <= k)$ in (c) refers to CDF.

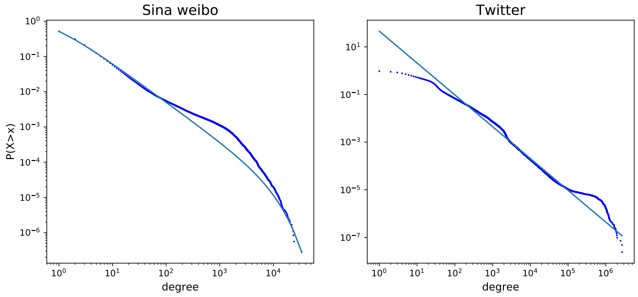


FIGURE 10. Fitting of the complementary cumulative distribution(ccdf) of the original network for Sina weibo and Twitter. The exponents λ are 2.18 and 2.33 for Sina weibo and Twitter.

the total number of edges in the original network of Sina weibo and Twitter and report them(P) in the Table 5. For Sina weibo and Twitter, the edge fraction of adpUNI+N is quite greater than the percolation thresholds, while UNI and adpUNI are the opposite. Thus we can assum that the connectivity of adpUNI+N is better than UNI and adpUNI. Next, we calculate and show the fraction of the nodes in the giant component of the sampled network by UNI, adpUNI and adpUNI+N for Sina weibo and Twitter in Table 5, which shows adpUNI+N is almost completely connected for both of Sina weibo and Twitter, while UNI and adpUNI are far behind.

VI. CONCLUSION AND DISCUSSION

In this article, we proposed some fast representative sampling methods (adpUNI and adpUNI+N) when dealing with large-scale OSNs, which have significant improvement on sampling efficiency and performance based on the observation of heterogeneous userID space. The key idea of our methods are dividing the entire userID space into several equal-length intervals, whose sampling probability adaptively adjust with its real time target rate.The contributions of this paper are as follows:

1) We propose and verify two fast adaptive methods (adpUNI and adpUNI+N) to overcome the defects of UNI which is of low efficiency, this is important for sampling large scale OSNs.

2) The subgraph obtained by our methods is of much higher representativeness and connectivity than UNI, which is more important for practical applications, since eventually we need to give a representative subgraph to end-users.

3) In the paper, we also reveal the relationship of three key concepts involved in all sampling methods: perfect uniformity can ensure an unbiased sampling of nodes, but not necessarily a more representative sampled subgraph.

Though there's a free parameter I (the number of intervals) in our methods, we find that our methods are quite robust with respect to different settings. However, there are also some requirements for being able to implement our methods. First, the range of userIDs assigned to users of OSN, that is MAXuserID, must be known or estimated in advance, so that a new ID can be randomly generated; Second, the OSN need to be allowed to query to return data for the selected userID, if it is valid, or return an error message if it does not exist.

The future research work mainly has the following aspects: First, how to develop more effective adaptive changing rule is an interesting topic. Secondly, we also want to study how to apply our methods to OSNs whose user IDs are pure letters or mix of numbers and letters. Moreover, we only apply our methods to Sina weibo and Twitter, which are regarded as undirected and unweighted static graphs in the paper. And in the future, we aim to study the sampling effect of our methods on other types of networks, such as weighted networks, dynamic networks, multiplex networks, multi-layer networks, etc.

ACKNOWLEDGMENT

Guangren Cai acknowledges technical support from Ankang Luo at UrbanNet Lab.

REFERENCES

[1] D. Centola, "The spread of behavior in an online social network experiment," *Science*, vol. 329, no. 5996, pp. 1194–1197, Sep. 2010.

[2] D. Centola and M. Macy, "Complex contagions and the weakness of long ties," *Amer. J. Sociology*, vol. 113, no. 3, pp. 702–734, Nov. 2007.

[3] R. Li, M. Tang, and P.-M. Hui, "Epidemic spreading on multi-relational networks," *Acta Phys. Sinica*, vol. 62, no. 16, p. 168903, 2013.

[4] R. Li, W. Wang, and Z. Di, "Effects of human dynamics on epidemic spreading in Côte d'Ivoire," *Phys. A, Stat. Mech. Appl.*, vol. 467, pp. 30–40, Feb. 2017.

[5] H. Liao, M.-K. Liu, M. S. Mariani, M. Zhou, and X. Wu, "Temporal similarity metrics for latent network reconstruction: The role of time-lag decay," *Inf. Sci.*, vol. 489, pp. 182–192, Jul. 2019.

[6] S. Zhang, M. Medo, L. Lü, and M. S. Mariani, "The long-term impact of ranking algorithms in growing networks," *Inf. Sci.*, vol. 488, pp. 257–271, Jul. 2019.

[7] J. Heidemann, M. Klier, and F. Probst, "Online social networks: A survey of a global phenomenon," *Comput. Netw.*, vol. 56, no. 18, pp. 3866–3878, Dec. 2012.

[8] R. Li, L. Dong, J. Zhang, X. Wang, W.-X. Wang, Z. Di, and H. E. Stanley, "Simple spatial scaling rules behind complex cities," *Nature Commun.*, vol. 8, no. 1, p. 1841, Dec. 2017.

[9] L. Dong, R. Li, J. Zhang, and Z. Di, "Population-weighted efficiency in transportation networks," *Sci. Rep.*, vol. 6, no. 1, Sep. 2016, Art. no. 26377.

[10] *Twitter: Fiscal Year 2018 Annual Report*. Accessed: Jun. 6, 2019. [Online]. Available: https://s22.q4cdn.com/826641620/files/doc_financials/ar/2018/AnnualReport2018.pdf

[11] *Facebook Q4 2018 Results*. Accessed: Jun. 6, 2019. [Online]. Available: https://s21.q4cdn.com/399680738/files/doc_financials/2018/Q4/Q4-2018-Earnings-Presentation.pdf

[12] *Sina Weibo: Q4 and Annual Financial Report*. Accessed: Jun. 6, 2019. [Online]. Available: https://tech.sina.com.cn/i/2019-03-05/doc-ihrfqzkc1446626.shtml

[13] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, "Walking in facebook: A case study of unbiased sampling of OSNs," in *Proc. IEEE INFOCOM*, Mar. 2010, pp. 1–9.

[14] M. Salehi, H. R. Rabiee, N. Nabavi, and S. Pooya, "Characterizing Twitter with respondent-driven sampling," in *Proc. IEEE 9th Int. Conf. Dependable, Autonomic Secure Comput.*, Dec. 2011, pp. 1211–1217.

[15] D. Stutzbach, R. Rejaie, N. Duffield, S. Sen, and W. Willinger, "On unbiased sampling for unstructured Peer-to-Peer networks," *IEEE/ACM Trans. Netw.*, vol. 17, no. 2, pp. 377–390, Apr. 2009.

[16] J. Leskovec and C. Faloutsos, "Sampling from large graphs," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining - KDD*, 2006, pp. 631–636.

[17] S. Yoon, S. Lee, S.-H. Yook, and Y. Kim, "Statistical properties of sampled networks by random walks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 75, no. 4, Apr. 2007, Art. no. 046114.

[18] S. Even, *Graph Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[19] M. Kurant, A. Markopoulou, and P. Thiran, "On the bias of BFS (Breadth first Search)," in *Proc. 22nd Int. Teletraffic Congr. (LTC)*, Sep. 2010, pp. 1–8.

[20] O. Frank, "Survey sampling in networks," in *The SAGE Handbook of Social Network Analysis*. London, U.K.: SAGE, 2011, pp. 389–403.

[21] A. Rezvanian and M. R. Meybodi, "A new learning automata-based sampling algorithm for social networks," *Int. J. Commun. Syst.*, vol. 30, no. 5, Mar. 2017, Art. no. e3091.

[22] L. A. Goodman, "Snowball sampling," *Ann. Math. Statist.*, vol. 32, no. 1, pp. 148–170, Mar. 1961.

[23] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proc. 7th ACM SIGCOMM Conf. Internet Meas. IMC*, 2007, pp. 29–42.

[24] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong, "Analysis of topological characteristics of huge online social networking services," in *Proc. 16th Int. Conf. World Wide Web WWW*, 2007, pp. 835–844.

[25] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi, "On the evolution of user interaction in facebook," in *Proc. 2nd ACM workshop Online social Netw. - WOSN*, 2009, pp. 37–42.

[26] A. Mislove, H. S. Koppula, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Growth of the flickr social network," in *Proc. 1st Workshop Online Social Netw. WOSP*, 2008, pp. 25–30.

[27] C. Wilson, B. Boe, A. Sala, K. P. N. Puttaswamy, and B. Y. Zhao, "User interactions in social networks and their implications," in *Proc. 4th ACM Eur. Conf. Comput. Syst. EuroSys*, 2009, pp. 205–218.

[28] S. H. Lee, P.-J. Kim, and H. Jeong, "Statistical properties of sampled networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 73, no. 1, Jan. 2006, Art. no. 016102.

[29] L. Becchetti, C. Castillo, D. Donato, A. Fazzone, and I. Rome, "A comparison of sampling techniques for Web graph characterization," in *Proc. Workshop Link Anal. (LinkKDD)*, Philadelphia, PA, USA, Aug. 2006, pp. 2–9.

[30] S. Ye, J. Lang, and F. Wu, "Crawling online social graphs," in *Proc. 12th Int. Asia–Pacific Web Conf.*, Apr. 2010, pp. 236–242.

[31] D. Wang, Z. Li, G. Xie, M.-A. Kaafar, and K. Salamatian, "Adwords management for third-parties in SEM: An optimisation model and the potential of Twitter," in *Proc. IEEE INFOCOM 35th Annu. IEEE Int. Conf. Comput. Commun.*, Apr. 2016, pp. 1–9.

[32] M. R. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork, "On near-uniform URL sampling," *Comput. Netw.*, vol. 33, nos. 1–6, pp. 295–308, Jun. 2000.

[33] A. H. Rasti, M. Torkjazi, R. Rejaie, N. Duffield, W. Willinger, and D. Stutzbach, "Respondent-driven sampling for characterizing unstructured overlays," in *Proc. IEEE INFOCOM 28th Conf. Comput. Commun.*, Apr. 2009, pp. 2701–2705.

[34] C. Gkantsidis, M. Mihail, and A. Saberi, "Random walks in peer-to-peer networks," in *Proc. IEEE INFOCOM*, Mar. 2004, p. 130.

[35] B. Krishnamurthy, P. Gill, and M. Arlitt, "A few chirps about Twitter," in *Proc. 1st workshop Online social Netw. WOSP*, 2008, pp. 19–24.

[36] A. H. Rasti, M. Torkjazi, R. Rejaie, D. Stutzbach, N. Duffield, and W. Willinger, "Evaluating sampling techniques for large dynamic graphs," Univ. Oregon, Eugene, Oregon, Tech. Rep. CIS-TR-08, 2008, vol. 1.

[37] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *J. Chem. Phys.*, vol. 21, no. 6, pp. 1087–1092, Jun. 1953.

[38] W. R. Gilks, S. Richardson, and D. Spiegelhalter, *Markov Chain Monte Carlo in Practice*. London, U.K.: Chapman & Hall, 1995.

[39] 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," *SIGMETRICS Perform. Eval. Rev.*, vol. 40, no. 1, p. 319–330, 2012.

[40] B. Ribeiro and D. Towsley, "Estimating and sampling graphs with multidimensional random walks," in *Proc. 10th Annu. Conf. Internet Meas. IMC*, 2010, pp. 390–403.

[41] K. Avrachenkov, B. Ribeiro, and D. Towsley, "Improving random walk estimation accuracy with uniform restarts," in *Proc. Int. Workshop Algorithms Models for Web-Graph*. Berlin, Germany: Springer, 2010, pp. 98–109.

[42] J. Zhao, P. Wang, J. C. S. Lui, D. Towsley, and X. Guan, "Sampling online social networks by random walk with indirect jumps," *Data Mining Knowl. Discovery*, vol. 33, no. 1, pp. 24–57, Jan. 2019.

[43] M. P. H. Stumpf, C. Wiuf, and R. M. May, "Subnets of scale-free networks are not scale-free: Sampling properties of networks," *Proc. Nat. Acad. Sci. USA*, vol. 102, no. 12, pp. 4221–4224, Mar. 2005.

[44] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," Stanford InfoLab, Stanford, CA, USA, Tech. Rep. 1999-66, Nov. 1999. [Online]. Available: http://ilpubs.stanford.edu:8090/422/

[45] A.-L. Barabasi, *Linked: How Everything is Connected to Everything Else What it Means*. New York, NY, USA: Plume, 2003.

[46] M. Ghavipour and M. R. Meybodi, "Irregular cellular learning automata-based algorithm for sampling social networks," *Eng. Appl. Artif. Intell.*, vol. 59, pp. 244–259, Mar. 2017.

[47] A. Leon-Garcia, *Probability, Statistics, and Random Processes for Electrical Engineering*. Hoboken, NJ, USA: Pearson Education, 2008.

[48] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, "Identification of influential spreaders in complex networks," *Nature Phys.*, vol. 6, no. 11, pp. 888–893, Nov. 2010.

[49] J. I. Alvarez-Hamelin, L. Dall'Asta, A. Barrat, and A. Vespignani, "K-core decomposition of Internet graphs: Hierarchies, self-similarity and measurement biases," 2005, *arXiv:cs/0511007*. [Online]. Available: https://arxiv.org/abs/cs/0511007

[50] S. Carmi, S. Havlin, S. Kirkpatrick, Y. Shavitt, and E. Shir, "A model of Internet topology using k-shell decomposition," *Proc. Nat. Acad. Sci. USA*, vol. 104, no. 27, pp. 11150–11154, Jul. 2007.

[51] C. Seshadhri, A. Pinar, and T. G. Kolda, "An in-depth analysis of stochastic kronecker graphs," *J. ACM*, vol. 60, no. 2, pp. 1–32, Apr. 2013.

[52] M. L. Goldstein, S. A. Morris, and G. G. Yen, "Problems with fitting to the power-law distribution," *Eur. Phys. J. B-Condens. Matter Complex Syst.*, vol. 41, no. 2, pp. 255–258, Sep. 2004.

[53] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in *Proc. 19th Int. Conf. World Wide Web WWW*, New York, NY, USA: ACM, 2010, pp. 591–600.

[54] B. Ribeiro and D. Towsley, "On the estimation accuracy of degree distributions from graph sampling," in *Proc. IEEE 51st IEEE Conf. Decis. Control (CDC)*, Dec. 2012, pp. 5240–5247.

[55] T. Wang, Y. Chen, Z. Zhang, P. Sun, B. Deng, and X. Li, "Unbiased sampling in directed social graph," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 40, no. 4, p. 401, Aug. 2010.

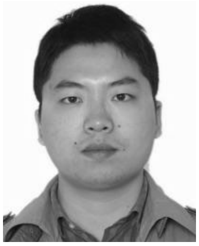
[56] D. Achlioptas, R. M. D'Souza, and J. Spencer, "Explosive percolation in random networks," *Science*, vol. 323, no. 5920, pp. 1453–1455, Mar. 2009.

[57] F. Radicchi and S. Fortunato, "Explosive percolation in scale-free networks," *Phys. Rev. Lett.*, vol. 103, no. 16, Oct. 2009, Art. no. 168701.

GUANGREN CAI received the B.S. degree in computer science from the Beijing University of Chemical Technology, Beijing, China. She is currently pursuing the bachelor's degree with the Department of Computer Science, College of Information Science and Technology, Beijing University of Chemical Technology. She is also a Research Assistant with the UrbanNet Laboratory, led by Dr. R. Li.

GANG LU received the B.S. degree in computer science and the Ph.D. degree in control theory and control engineering from the Beijing University of Chemical Technology, Beijing, China, in 2003 and 2008, respectively. He was a Visiting Scholar with MIT, in 2013, for six months. He is currently a Lecturer with the Department of Computer Science, College of Information Science and Technology, Beijing University of Chemical Technology. He is also a Research Scientist with the UrbanNet Laboratory, which focuses on gaining better understanding on urban systems with network science, big data, and advanced technologies. His main research interests include complex networks for social computing, social network analysis, and computing technology for large-scale graph computing.

JUNXIA GUO received the B.S. degree in computer science from the Taiyuan University of Technology, Taiyuan, China, in 1999, the M.S. degree in computer science from the China University of Mining and Technology Beijing, Beijing, China, in 2002, and the Ph.D. degree in computer science from the Tokyo Institute of Technology, Tokyo, Japan, in 2013. She is currently an Associate Professor with the College of Information Science and Technology, Beijing University of Chemical Technology, China. Her primary research interests include software testing and big data analysis and modeling.



CHENG LING received the B.Eng. degree in electronic engineering from the joint training program held by Shenzhen University and the University of Central Lancashire, in 2008, and the Ph.D. degree in electronic engineering from Edinburgh University, in 2012. From 2013 to 2015, he was a Postdoctoral Researcher with the State Key Laboratory of Intelligent Technology and Systems, Tsinghua University. He has been an Associate Professor with the Beijing University of Chemical Technology, since 2018. His research interests include computational biology and bioinformatics.

RUIQI LI received the B.S. degree in computer science from University of Electronic Science and Technology of China, in 2013, and the Ph.D. degree in systems science from Beijing Normal University, in 2018. He was a Visiting Ph.D. Student with MIT, in 2016, and a Visiting Scholar with Boston University, in 2017. He is currently an Associate Professor with the Department of Computer Science, College of Information Science, Beijing University of Chemical Technology, Beijing, China. He is also the Director of the UrbanNet Laboratory, which focuses on gaining better understanding on urban systems with network science, big data and advanced technologies. His main research interests include urban modeling and computation, social network analysis, and epidemic spreading dynamics.