IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 17, 2019, accepted September 12, 2019, date of publication September 18, 2019, date of current version October 7, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2942216

2-Hopper: Accurately Estimate Individual and Social Attributes of Social Networks With Fewer Repeats via Random Walk

LINGLING ZHANG[1,2], (Student Member, IEEE), HONG JIANG[3], (Fellow, IEEE), FANG WANG[1,2], (Member, IEEE), AND DAN FENG[1,2], (Member, IEEE)

[1]Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan 430074, China
[2]Key Laboratory of Information Storage System, School of Computer Science and Technology, Ministry of Education of China, Huazhong University of Science and Technology, Wuhan 430074, China
[3]Department of Computer Science and Engineering, The University of Texas at Arlington, Arlington, TX 76019, USA

Corresponding author: Fang Wang (wangfang@hust.edu.cn)

This work was supported in part by the NSFC under Grant 61832020 and Grant 61772216, in part by the National Key Research and Development Program of China under Grant 2018YFB10033005, in part by the National Defense Preliminary Research Project under Grant 31511010202, in part by the Hubei Province Technical Innovation Special Project under Grant 2017AAA129, in part by the Wuhan Application Basic Research Project under Grant 2017010201010103, and in part by the Fundamental Research Funds for the Central Universities.

ABSTRACT Random-walk based sampling is widely used to characterize large graphs by producing samples in the form of nodes. However, existing random-walk based sampling methods only focus on the estimation accuracy of structural properties but suffer from repetitive samples which have adverse effects on obtaining accurate information about the structures over social networks represented by large graphs. Furthermore, these existing methods mainly characterize individual attributes while ignoring the social attributes of the nodes. In this paper, a new random-walk based method, called 2-hop neighbors based random walk or 2-Hopper, is proposed to obtain accurate estimations of both basic and social attributes with fewer repetitive samples. Specifically, 2-Hopper is able to greatly reduce redundant paths among nodes during the sampling process and thus produces few repeats. Based on 2-Hopper's sampling process, a re-weighted estimator is proposed to accurately obtain both the individual and social properties while the latter is obtained by a newly proposed algorithm. Experimental results driven by real-world datasets show that on average 2-Hopper can reduce 4.5 times repetitive samples of the state-of-the-art random-walk based methods and obtain more accurate information about the individual and social attributes while 2-Hopper is able to estimate the structural properties of these attributes accurately over large graphs.

INDEX TERMS Random-walk based sampling, few repeats, accurate estimations, basic and social attributes of social networks.

I. INTRODUCTION

Due to increasingly large volumes of data in online social networks (OSNs) represented by ever larger graphs, it is necessary to use sampling methods to estimate the structural properties of OSNs efficiently [1]–[5]. Existing sampling methods designed to characterize large social networks can be divided into three categories, namely, random sampling on nodes or edges, traversal-based sampling and random-walk based sampling. Although random sampling can estimate the properties of OSNs accurately, their reliance on user IDs or pairs of user IDs [6] makes them ineffective because it is almost impossible to successfully infer the true user IDs or pairs of user IDs. On the other hand, traversal-based sampling methods produce biased samples and the biases cannot be remedied by any estimator [4]. In contrast, random-walk based sampling [6]–[8] is highly efficient into producing samples and then estimating the structural properties accurately by employing unbiased estimators.

Existing random-walk based sampling methods characterize the structural properties of large social networks from two angles, namely, the individual attributes and the social attributes. The former refers to the personal traits of users (e.g., the age, gender, number of friends, etc.) while the

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

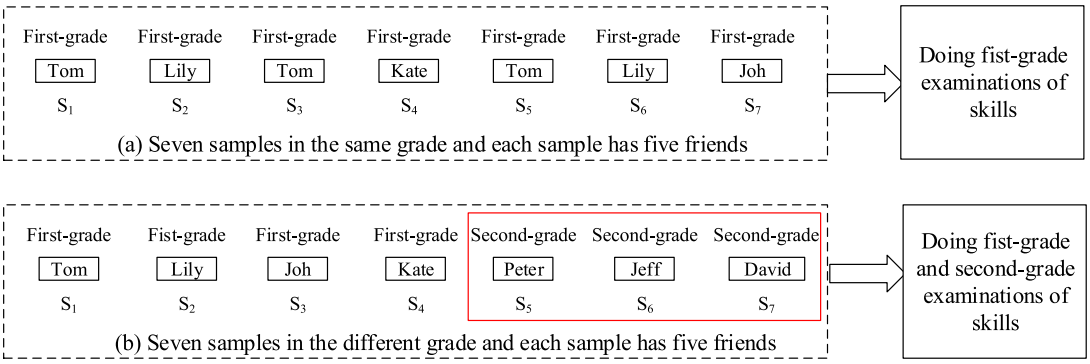


FIGURE 1. Illustrating two different groups of samples, with (Fig.1(a)) and without (Fig.1(b)) repeats, when using the samples to do experiments.

latter characterizes the social activities or common interests between users and their respective friends. Thus, the social attributes of users lie in the connectivity among their neighbors [9], [10]. Both the individual and social attributes are very important in applications that differentiate the users in OSNs for the purpose of data mining [11] and visualizing large graphs [12]–[14].

However, the existing random-walk based sampling methods characterize the social networks in terms of either the individual attributes or the social attributes but not both. They characterize the individual attributes by producing samples in the form of the nodes without further analyzing the connectivity among the users. For the social properties they use patterns of locally connected subgraphs (also called motifs)to characterize the connectivity among a given number n of nodes. Since the value of n is actually evaluated to be less than six by the state-of-the-art sampling techniques, there may not be common social attributes among the loosely connected subgraphs formed by a small number of nodes. As a result, social attributes in the form of these loosely connected subgraphs cannot be uncovered comprehensively and accurately. Worse still, different values of n necessitate different sampling processes, resulting in huge sampling costs. Nevertheless, the motifs in the form of the completely connected subgraphs, also called cliques, can be used to mine the social attributes of users. Therefore, in this paper, we leverage the structures of cliques which can be formed by any number of nodes to obtain the social attributes of users. In other words, with a single sampling process, we focus on characterizing the individual attributes by analyzing the samples themselves while obtaining the social attributes by analyzing the cliques that the samples have participated in.

Furthermore, while mainly focusing on obtaining the structural properties of the individual or social attributes, such as the distribution of a given property, the existing random-walk based sampling methods largely ignore the fact that the samples they produce typically have many repeats, which has a seriously adverse effect on the useful information extractable from these samples to further analyze the formations of the structural properties. Taking Figure 1 for example, although the seven samples have the same five friends for estimating the distributions of the number of friends, the samples in Figure 1(a) which have multiple repeats (3 of Tom, 2 of Lily), can only be used to do the first-order examinations while those in Figure 1(b), which have no repeats, can be used to do the fist-order and the second-order examinations when these samples are used to assess the quality of different examinations or do precise product promotions for different orders.

Therefore, even if the existing random-walk based sampling methods can obtain the individual and social attributes by producing node samples, the following two problems make them inefficient and ineffective.

- The key step of the existing methods is to select the next sample randomly from the neighbors of the currently sampled node, which can lead to many repetitive samples. These repetitive samples prevent more useful information from being extracted from these samples given the same prescribed total number of samples, i.e., sampling budget.
- The social attributes in the form of cliques are obtained by analyzing the connectivity of the neighboring nodes of the currently sampled node. However, these neighbors are also collectively considered as the sampling space for the next sample, meaning that the consecutive samples may have common neighbors and share the same social attributes. Thus, with a limited sampling budget, such a strategy for sampling the next sample tends to severely underestimate the diversity of the social attributes.

In this paper, we observe that the root cause of the repetitive samples is the redundant paths from one node to another node while the main culprit for the inaccurate estimation of the social attributes is the limited sampling space for the next sample. Therefore, to address these two problems, we propose a new random-walk based method by designing a strategy of employing the neighbors of the neighbors, i.e., two-hop neighbors, of users to sample the social networks. This new sampling method, called 2-Hopper, can efficiently enlarge the sampling space while simultaneously

reducing the redundant paths between two nodes. Furthermore, to estimate the structural properties of both the individual and social attributes accurately, we design a re-weighted estimator for the samples produced by 2-Hopper. With the design and prototype implementation of 2-Hopper, we make following contributions.

1) We uncover the root cause for repetitive samples produced by the existing random-walk based sampling methods, namely, the redundant paths from one node to another during the sample selection process. (Section II)
2) 2-Hopper is proposed to estimate the individual and social attributes efficiently by considering two-hop neighbors of the currently sampled node and eliminating the redundant paths from one node to another during the process of preparing the sample selection spaces. A re-weight estimator is proposed to accurately estimate large graphs. (Section III and Section IV)
3) To uncover the social attributes of the users in social networks in the form of cliques, we propose a recursive strategy to find all the cliques corresponding to the node. This is in contrast to the existing algorithms that mainly focus on finding the maximum cliques of the graph and are not adequate for describing the social attributes of the specific users. (Section IV.)
4) Extensive experiments conducted on real-world datasets show that 2-Hopper produces samples with much fewer repetitive samples than existing state-of-the-art sampling methods while estimating both the structural properties of the individual and social attributes accurately. (Section V)

The reminder of this paper is organized as follows. Section II describes the necessary background, which motivates our 2-Hopper study. Section III introduces the design of 2-Hopper in detail and a proper estimator to re-weight the samples produced by 2-Hopper for accurate estimations. Section V presents the evaluations driven by real-world datasets while Section VI concludes our work.

II. BACKGROUND AND MOTIVATION

In this section, we first study the existing random-walk based sampling methods from the angle of sampling paths and then analyze the root causes of the repetitive samples they generate. The insight from the analysis and the need to accurately obtain both the individual and social attributes motivate us to propose 2-Hopper, a new random-walk based sampling method.

A. EXISTING RANDOM-WALK BASED SAMPLING METHODS MOST RELEVANT TO 2-HOPPER

SIMPLE RANDOM WALK (SRW)

SRW's sampling paths are formed by a set of node pairs, each of which is an edge of a large graph [4]. If a node has a large number of neighbors, there must be a large number of paths converging on the node, meaning that this node has a very high probability of being sampled. Conversely, if a node has a small number of neighbors, there is a small probability of this node being sampled. Therefore, during the process of SRW, there is a very likely bias in that nodes with higher degrees tend to be more repeatedly sampled than those with lower degrees, resulting in both over-sampled nodes (i.e., of higher degrees) and under-sampled nodes (i.e., of lower degrees), leading to a severe lack of diversity among the samples. **Non-backtracking random walk (NBRW)**, proposed in [3] and **Circulated Neighbors random walk (CNRW)**, proposed in [15] are based on the idea of non-backtracking to a very small fraction of the sampled paths. In this context, a sampling path refers to an edge through which the random walker goes from the current sample to the next. In NCRW, the currently sampled path of the random walk is eliminated from its candidate paths to obtain the next sample. Whereas, in CNRW, any two consecutive sampling paths which have been visited by the random walk, are blocked. For example, suppose that CNRW has walked along the sampling path of $\mu \rightarrow \nu \rightarrow \omega$, which is now blocked. If it walks from $\mu \rightarrow \nu$ again, the neighbors of ν except for the node ω will be sampled randomly. However, the kind of backtrack-path blocking only prevents a very small fraction of sampled paths being repeated, failing to effectively and fully address the problem of high ratio of repetitive samples.

Skipping random walk (SkipRW), proposed in [16], skips some candidate samples to reduce repetitive samples while following the SRW process. For example, when SkipRW walks through the path (μ, ν), ν is selected as a sample with a pre-defined probability (e.g., 0.5.). Although SkipRW changes the strategy of producing the samples, it does not change the sampling process of SRW at all, failing to reduce the repetitive samples effectively.

Although the existing random-walk based sampling methods (i.e., NBRW and CNRW) try to change the paths of the random walk of SRW, they do not address the root cause of the redundant-path problem. Furthermore, in these methods, the consecutively sampled nodes may be neighbors with each other, which may have the same social attributes and are not conducive to mining important information for characterizing the social attributes. Therefore, they cannot estimate the individual and social attributes accurately.

B. SAMPLING PATHS

In general, the repetitive samples are produced by the existing random-walk based sampling methods because the random walks can backtrack to the already-sampled nodes by walking along the redundant paths among nodes. There are two types of such paths, namely, direct paths and indirect paths, as described below.

A path between two nodes is a *direct path* if there is an edge between the two nodes. Backtracking to the sampled nodes through direct paths is referred to as ***direct backtracking***. Most of the existing random-walk based methods are variations of simple random walk (SRW) in that the next sample is selected from the neighbors of the currently sampled node.

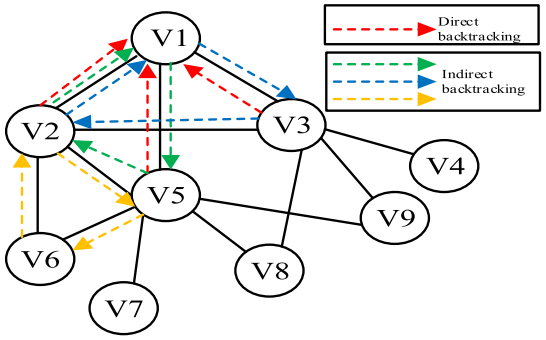


FIGURE 2. Direct backtracking and indirect backtracking to the node $V1$ where the former is referred as to the sampling process backtracking to the sampled node via one of its neighbors (i.e., $V1 \rightarrow V2 \rightarrow V1$) while the latter is referred as to the process backtracking to the sampled node through more than one bridge nodes (i.e., $V1 \rightarrow V5 \rightarrow V2 \rightarrow V1$).

Therefore, there are direct paths from the currently sampled node to the candidate nodes to obtain the next sample. As shown in Figure 2, no matter which of node $V1's$ neighbors is to be selected as the next sample from the candidate samples $V2$, $V3$,and $V5$, there is a chance for SRW's process to backtrack to node $V1$ again as there are three direct paths, labeled as, $\{V2 \rightarrow V1, V3 \rightarrow V1, V5 \rightarrow V1\}$. A path between two nodes is an *indirect path* if they are connected by two of more edges between *'bridge' nodes*. Backtracking to the sampled nodes through indirect paths is referred to as ***indirect backtracking***.Even if a random-walk based sampling method (e.g., NBRW) avoids backtracking to the already sampled nodes through direct paths, there are many indirect paths for the method to backtrack to the sampled nodes again. As shown in Figure 2, SRW backtracks to the sampled node $V1$ through an indirect path, $V1 \rightarrow V5 \rightarrow V2 \rightarrow V1$. Furthermore, as shown in Figure 2, the number of indirect paths is more than that of direct paths from a node to $V1$.

Furthermore, the experiments driven by real datasets are conducted to learn the respective ratios of the repetitive samples due to the two types of backtracking. In this paper, the ratio of repetitive samples (RRS) is defined as $RRS = \frac{B-U}{B}$, where B is the total number of samples and U is the number of unique samples among B. Experimental results based on the sampling process of SRW on the datasets of com-DBLP and amazon0601, described in Section V show that the repetitive samples caused by indirect backtracking are notably more than those caused by direct backtracking. Therefore, it is necessary to avoid indirect backtracking to reduce the repeats significantly.

C. MOTIVATION

From the above analysis, we observe that there are a large number of redundant paths through which a random-walk process can traverse during the sampling processes of the existing random-walk based sampling methods as shown in Table 1, resulting in many repetitive samples. On the other hand, these repetitive samples, once obtained, cannot be removed arbitrarily because they are necessary for obtaining

TABLE 1. Summary of the existing sampling methods from the perspective of redundant paths.

Methods	Redundant paths
SRW [4], [6]	Lots of directly and indirectly redundant sampling paths
NBRW [3]	Avoiding directly redundant sampling paths
CNRW [15]	Avoiding a little bit of indirectly redundant sampling paths
SkipRW [16]	Skipping a bit of repetitive sampling nodes obtained by directly redundant sampling paths

estimations on the structural properties of a large graph as shown in [3], [4], [6], [17]. Therefore, it is necessary to design a new sampling strategy to produce samples with fewer repeats.

As described above, the indirect backtracking is the main root cause of the repetitive samples. Thus, when designing a new random-walk based sampling method, it is necessary to cut down the redundant paths traversed through a large graph by the random walker. To obtain the social attributes it is necessary to reach the neighbors of neighbors, or two-hop neighbors, of the sampled node, to learn the connectivity of the neighbors of the nodes in the form of cliques [18], [19]. Therefore, the cost of obtaining the social attributes of the sampled node includes that of obtaining the two-hop neighbors. Since the probability of a sampled node being re-sampled through indirect backtracking decreases with the increase in the length of indirect paths traversed by the random walker, sampled nodes are more likely to be indirectly backtracked via 2-hop indirect paths than 3-or-more-hop indirect paths, although the latter are far more costly to maintain and keep track of. Therefore, in this paper, to balance the costs of estimating both individual and social attributes, we employ the two-hop neighbors to design a new random-walk based method described in the next section, called 2-Hopper that is able to produces samples with much fewer repeats than the state-of-the-art methods.

III. 2-HOP NEIGHBORS BASED RANDOM WALK

In this section, we first introduce the definitions of relevant terms and notations to facilitate the description of 2-Hopper that follows. Then, we analyze 2-Hopper to validate the high quality of the samples it produces.

A. DEFINITIONS

We refer to an undirected and acyclic graph as $G = (V, E)$, where V denotes the set of nodes and E denotes the set of edges between nodes. The set of (one-hop) neighbors of a node μ is defined as $nei(\mu)$. The 2-hop neighbors of node μ can be described in terms of edge-based neighbors or node-based neighbors.

Edge-based 2-hop neighbors of a node μ, labeled $edge2Nei(\mu)$, are defined as follows: if there exist $(\nu, \omega) \in E$ and $\nu \in nei(\mu)$, $\omega \in edge2Nei(\mu)$. The number of nodes in $edge2Nei(\mu)$ is the sum of the neighbors of the nodes in the set $nei(\mu)$ and the edge-based neighbors do not exclude

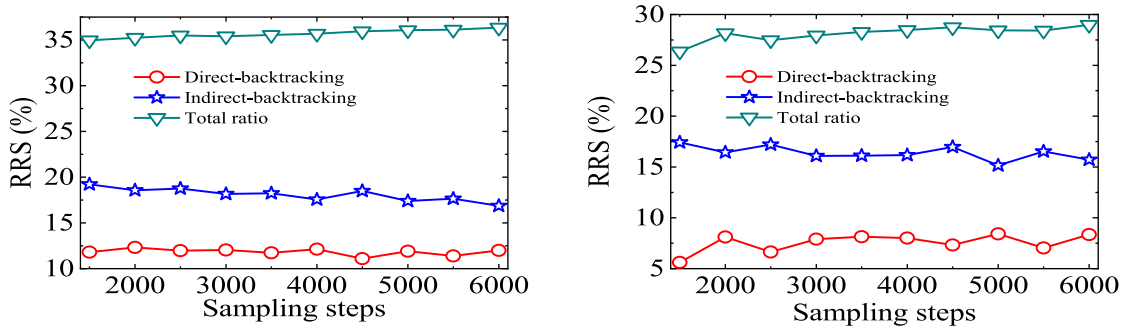


FIGURE 3. The respective ratios of repetitive samples produced by SRW's direct backtracking and indirect backtracking over the DBLP and amazon0601 as a function of the number of sampling step. (a) com-dblp. (b) amazon0601.

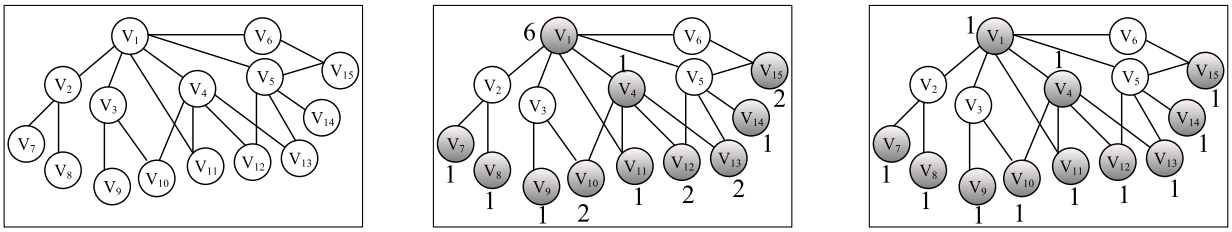


FIGURE 4. (a) is an original graph while (b) and (c) are the 2-hop neighbors (labeled gray) of the node V_1 in two forms, edge-based and node-based respectively. The numbers on the nodes in (b) and (c) denotes the appearance times in the 2-hop neighbor sets respectively and then the candidate sampling paths in (b) and (c) are formed by the paths from V_1 to its 2-hop neighbors where the respective appearance times denote the number of its corresponding sampling path. (a) An original graph. (b) Edge-based neighbors. (c) Node-based neighbors.

repeats. Thus, the size of $edge2Nei(\mu)$ is the sum of the node degrees (nd) of nodes in $nei(\mu)$. For example, Figure 4(b) shows that V_1 has 6 neighbors $\{V_2, V_3, V_{11}, V_4, V_5, V_6\}$ and 20 edge-based 2-hop neighbors (candidate sampling paths) (i.e., $nd(V_2)+nd(V_3)+nd(V_{11})+nd(V_4)+nd(V_5)+nd(V_6) = 3+3+2+5+5+2 = 20$), $edge2Nei(V_1) = \{(V_1 : 6), (V_4 : 1), (V_7 : 1), (V_8 : 1), (V_9 : 1), (V_{10} : 2), (V_{11} : 1), (V_{12} : 2), (V_{13} : 2), (V_{14} : 1), (V_{15} : 2)\}$, where $(V_a : b)$ denotes that node V_a appears b times in $V_1's$ edge-based 2-hop neighbors.

Node-based 2-hop neighbors of a node μ, labeled $node2Nei(\mu)$, are defined as follows: if there exist $\omega \in nei(\nu)$ and $\nu \in nei(\mu)$, $\omega \in node2Nei(\mu)$. However, significantly different from the edge-based 2-hop neighbors, node-based 2-hop neighbors do not allow any repeats in $node2Nei(\mu)$ and thus $node2Nei(\mu)$ contains only unique nodes that are the neighbors of the neighbors of μ. Specifically, if there exist $\omega \in nei(\nu_1)$, $\omega \in nei(\nu_2)$, $\nu_1 \in nei(\mu)$, $\nu_2 \in nei(\mu)$ and $\nu_1 \neq \nu_2$, then $\omega \in node2Nei(\mu)$ and ω appears only once in $node2Nei(\mu)$. In contrast to the edge-based 2-hop neighbors with 20 candidate sampling paths, Figure 4(c) shows that 11 candidate sampling paths are $\{V_1 \rightarrow V_1, V_1 \rightarrow V_4, V_1 \rightarrow V_7, V_1 \rightarrow V_8, V_1 \rightarrow V_9, V_1 \rightarrow V_{10}, V_1 \rightarrow V_{11}, V_1 \rightarrow V_{12}, V_1 \rightarrow V_{13}, V_1 \rightarrow V_{14}, V_1 \rightarrow V_{15}\}$. Whereas, in contrast to 1-hop neighbors based sampling methods with at most 6 candidate sampling nodes (i.e., $\{V_2, V_3, V_4, V_5, V_6, V_{11}\}$), 2-Hopper sampling method has 11 candidate sampling nodes that are referred to as $node2Nei(V_1) = \{V_1, V_4, V_6, V_7, V_8, V_9, V_{10}, V_{11}, V_{12}, V_{13}, V_{14}, V_{15}\}$.

From the descriptions of edge-based and node-based 2-hop neighbors of a node μ, the set of node-based 2-hop neighbors can be used to cut down the redundant indirect paths between the sampled node and the candidate node. For example, Figure 4(b) implies that there are six paths from V_1 to $V1$: $\{V_1, V_2, V_1\}, \{V_1, V_3, V_1\}, \{V_1, V_{11}, V_1\}, \{V_1, V_4, V_1\}, \{V_1, V_5, V_1\}, \{V_1, V_6, V_1\}$, which are revealed by the edge-based 2-hop neighbors. Whereas, the node-based 2-hop neighbors of Figure 4(c) shows exactly one path from V_1 to V_1.

B. SAMPLING METHOD

As described in Section I, the 2-hop neighbors can help better extract information of the social attributes of the social networks. Since the edge-based 2-hop neighbors of a node cannot change the paths among the nodes fundamentally, existing sampling methods (i.e., SRW, NBRW, CNRW and SkipRW) can be easily extended to the techniques by employing the the definition of edge-based 2-hop neighbors. Therefore, from the perspective of sampling paths, the existing methods can be seen as edge-based sampling methods. In other words, edge-based 2-hop neighbors contain many repeats of nodes and redundant paths that do not help address the indirect-backtracking problem at all, and thus in this paper we

TABLE 2. The probabilities of the nodes sampled by the edge-based strategy and the node-based strategy.

Node	V_1	V_4	V_7	V_8	V_9	V_{10}	V_{11}	V_{12}	V_{13}	V_{14}	V_{15}
Edge-based	$\frac{6}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{2}{20}$	$\frac{1}{20}$	$\frac{2}{20}$
Node-base (2-Hopper)	$\frac{1}{11}$	$\frac{1}{11}$	$\frac{1}{11}$	$\frac{1}{11}$	$\frac{1}{11}$	$\frac{1}{11}$	$\frac{1}{11}$	$\frac{1}{11}$	$\frac{1}{11}$	$\frac{1}{11}$	$\frac{1}{11}$

propose a strategy, called 2-Hopper, leveraging node-based, rather than edge-based, 2-hop neighbors to simultaneously extend the sampling space for each step and cut down the indirect paths from one node to another. The two key steps of 2-Hopper, (1) generation of sampling space $node2Nei(\mu)$ for the next sample when the random walk is residing on node μ, and (2) selection of the next sample from $node2Nei(\mu)$, are described in detail via an example as follows.

Take Figure 4 for example, if the current sample is V_1, then the next sample is selected randomly from the node-based 2-hop neighbors, i.e., $\{V_1, V_4, V_7, V_8, V_9, V_{10}, V_{11}, V_{12}, V_{13}, V_{14}, V_{15}\}$, where each node has an equal probability of $\frac{1}{11}$ of being selected. On the other hand, if an edge-based strategy is used, the sampling space will be the edge-based 2-hop neighbors ($edge2Nei(V_1)$) of V_1 with $V1$ appearing 6 times, while V_{10}, V_{12}, V_{13} and V_{15} each appearing twice. Thus, these nodes will have 6 or 2 times higher probability of being selected than the rest of nodes, as indicated in Table 2, making the sampling process highly biased toward V_1. From the sampling process, the necessary cost for 2-Hopper is to obtain the two-hop neighbors of the current sample. However, such cost can be overlapped by the costs of characterizing the social attributes which necessitate analyze the connectivity of the neighbor of the currently sampled node.

A FORMAL DESCRIPTION OF 2-HOPPER

From the sampling process of 2-Hopper, the transition probability from the $(i-1)^{th}$ $(1 \le i \le n)$ state to the i^{th} state is just relevant to the $(i-1)^{th}$ state but irrelevant to the front $i-2$ states, where n is the sampling budget (i.e., total number of samples). Therefore, the sampling process of 2-Hopper can be described as an irreducible and time-reversible Markov chain. $\mu's$ transition probability of 2-Hopper from the node μ to the node ν, labeled as $p(\mu)$, is described as below.

$$p(\mu) = \begin{cases} \dfrac{1}{|node2Nei(\mu)|} & \text{if } \nu \in node2Nei(\mu), \\ 0 & \text{if } \nu \notin node2Nei(\mu). \end{cases} \quad (1)$$

Stationary distribution means that the probability of selecting the node μ converges to a fixed value when sufficient sampling steps have been taken. It is used to explain that 2-Hopper can be used to produce samples. According to the knowledge of the Markov chain based graph sampling [20], the sampling process of 2-Hopper converges to $\pi_\mu = \frac{|node2Nei(nei(\mu))|}{\sum_{v \in V} |node2Nei(v)|}$.

IV. ESTIMATIONS BASED ON 2-HOPPER

In this section, we first propose a re-weighted estimator to estimate structural properties of a large graph by using the sampled nodes produced by 2-Hopper. Furthermore, to estimate both the basic and social attributes of the large graph, we propose a recursive algorithm to obtain the social attributes of the sampled nodes during a sampling method as the basic attributes of the nodes can be obtained directly by analyzing these nodes directly.

A. ESTIMATOR

From the above description, the probability of a node in 2-Hopper is related to the number of the nodes in its node-based neighbors. In other words, the samples produced by 2-Hopper are not obtained with the same chances. Thus, when these samples are used to estimate the characteristics of a large graph with a small estimation error, it is necessary to use an unbiased estimator to remedy the deviation of the samples. Let the value of a property, labeled $pro(\mu)$, ranges among $\{\alpha_1, \ldots, \alpha_k\}$ where k is the number of different values of the the property in a large graph. If the value of $\mu's$ property is equal to α_j $(1 \le j \le k)$, $F(pro(\mu) = \alpha_j) = 1$. Otherwise, $F(pro(\mu) = \alpha_j) = 0$. In this paper, we propose a re-weight estimator described as follows to estimate the structural properties of attributes of a large graph accurately.

$$\tilde{\omega}_j = \frac{1}{W} \sum_{i=1}^{|B|} F(pro(\mu_i) = \alpha_j) \cdot p(\mu_i), \quad (2)$$

where $W = \sum_{i=1}^{|B|} p(\mu_i)$, $\mu_i \in G$, B denotes the number of the total samples and $p(\mu_i)$ is the transition probability of μ_i in 2-Hopper sampling process which is described in Equation 1.

Theorem 1: If the graph G is non-bipartite and connected, then $\tilde{\omega}_j$ is an asymptotically unbiased estimator of ω_j. $\tilde{\omega}_j$ is called *the re-weighted estimator* for ω_j based on the samples produced by 2-Hopper.

Before proving the theorem, the related proposition is given below.

Theorem 2: The sampling probability of the item μ based on the Markov chain is as $\pi(\mu)$, then for any function, $\sum f(\mu) \cdot \pi(\mu) < \infty$, we have

$$\lim_{n \to \infty} \sum_{i=1}^{i=n} f(\mu_i) = \sum_{i=1}^{|V|} f(\mu_i) \cdot \pi(\mu_i).$$

The proof of of this proposition is similar to that presented in [21, Proposition 17.3.4] when $f(\mu_i)$ is set $F(pro(\mu_i) = \alpha_j)$. The proof of Theorem 1 is presented as follows. Based on Proposition 2, the proof of Theorem 1 given below is based

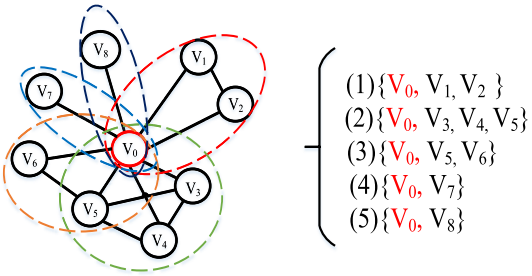


FIGURE 5. Five cliques related to the node V_0 and the maximum clique of V0 is formed by V_0, V_3, V_4,and V_5. $V_0's$ degree, NumClique and MaxClique are 8, 5, and 4 respectively.

on the sampling process of 2-Hopper.

$$\lim_{B\to\infty}\frac{1}{B}\sum_{i=1}^{|B|}\frac{F(pro(\mu_i)=\alpha_j)}{\pi(\mu_i)}$$
$$=\lim_{B\to\infty}\frac{1}{B}\sum_{i=1}^{|B|}F(pro(\mu_i)=\alpha_j)\cdot p(\mu_i)$$
$$\times(\sum_{i=1}^{|B|}|node2Nei(\mu_i)|)$$
$$=\lim_{B\to\infty}\frac{\sum_{i=1}^{|B|}|node2Nei(\mu_i)|}{|B|}\times\sum_{i=1}^{|B|}F(pro(\mu)=\alpha_j)\cdot p(\mu_i)$$

where $p(\mu_i)$ is μ_i's transition probability of 2-Hopper. Then, there is an equation as follows based on the Proposition 2.

$$\lim_{B\to\infty}\frac{\sum_{i=1}^{|B|}|node2Nei(\mu_i)|}{|B|}\xrightarrow{a.s.}\frac{1}{\sum_{i=1}^{|B|}p(\mu_i)}$$

So we have $E[\tilde{\omega}_j]\xrightarrow{a.s.}\tilde{\omega}_j$.

B. INDIVIDUAL AND SOCIAL ATTRIBUTES

In this paper, when a sample is produced by a random-walk based sampling method, it is analyzed to obtain its individual and social attributes represented by the following two aspects.

Degree, reflected by the number of the neighbors of a node, is used to represent the individual attribute of a user in social network. For example, Figure 5 shows that the degree of node V_0 is 8.

Cliques that a node participates in are used to reflect the social attributes of a user in social networks. We consider the minimum clique is made up of two nodes and exclude the isolated node because it is valueless to reflect the social attributes of users. Furthermore, we reflect the social attributes of the users from two aspects: the number of cliques and the size of the maximum clique that a node participates in, referred as NumClique and MaxClique respectively. For example, Figure 5 shows there are five cliques related to the node V_0, representing $V_0's$ five different social attributes, while the size of the maximum clique (formed by V_0, V_3, V_4,and V_5) that node V_0 participates in is four.

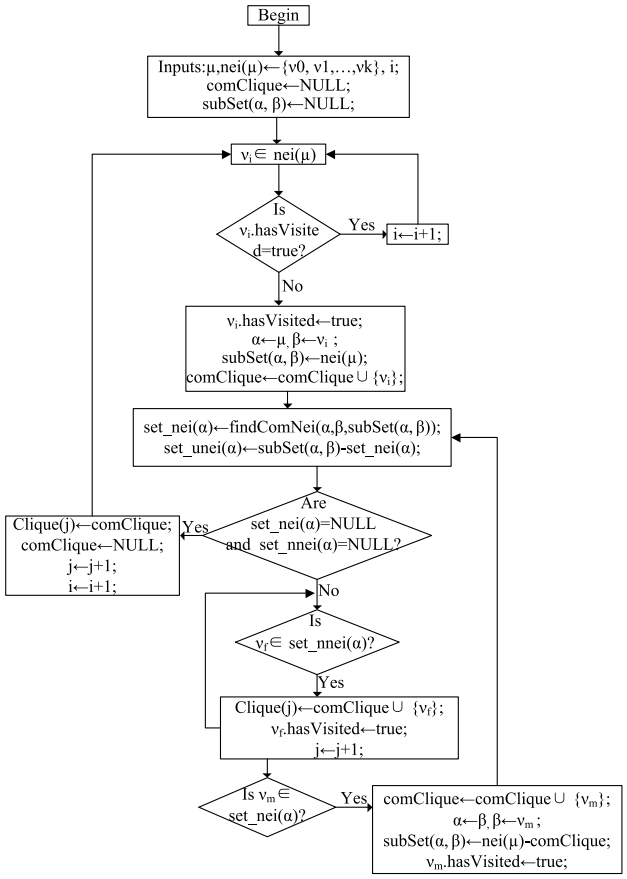


FIGURE 6. The flow chart of the process of finding the cliques related to node μ.

Since existing methods [18], [19] mainly focus on finding the maximum clique by branch-and-bound coloring algorithms without finding all the cliques, they can not be used to obtain the social attributes in the form of cliques directly. Therefore, in this paper, we propose a recursive strategy to identify the cliques one node participate in. The recursive strategy works like that: Let node $\mu's$ neighbors are described as the set $nei(\mu)=\{\mu_0,\ldots,\mu_k\}$, where $k+1$ is the number of $\mu's$ neighbors. The task of obtaining the cliques related to a node can be done by the recursive strategy described as follows. If the cliques related to μ_1 defined as $Clique(\mu_1)$ in the node set $\{\mu_2,\ldots,\mu_k\}$ is known, the cliques contain $\{\mu,\mu_1\}$ are identified. Similarly, before finding $Clique(\mu_i)$, all of the cliques formed by the nodes in the node set $\{\mu_{(i+1)},\ldots,\mu_k\}$ should be found. To avoid repetitive cliques, if a node in the neighboring set has been found to participate cliques, it is labeled with *hasVisited*. Thus, the recursive algorithm just backtracks to the nodes without labels of *hasVisited* to find the cliques. Figure 6 shows the specific process of finding the clique of node μ. Furthermore, the recursive strategy of finding all the cliques that a node participates in is depicted in Algorithm 1 and Algorithm 2, where the function of 'Find-NeiClique' is to find the cliques of $\mu's$ neighboring nodes from $subGraph(\mu)$, which is composed of its neighbors and the edges among the neighbors.

Algorithm 1 Algorithm of FindClique

Input: $\mu, nei(\mu) = \{v_0, v_1, v_2, \ldots, v_k\}$ and i where i is i^{th} neighbors of μ, $i = 0$ for the fist time of running FindClique and $k+1$ is the number of μ's neighbors; $ComClique \leftarrow NULL$ that is used to find the common items among cliques and $subSet(\alpha, \beta) \leftarrow NULL$ that is used to find the common neighbors of α and β;

Output: the set of cliques labeled as $Clique$;

if $v_i.hasVisited == false$ **then**
  $ComClique \leftarrow \emptyset$;
  $FindNeiClique(\alpha, v, nei(\mu), ComClique)$;
  $v_i.hasVisited \leftarrow true$;
  $\alpha \leftarrow \mu, \beta \leftarrow v_i$;
  $subSet(\alpha, \beta) \leftarrow nei(\alpha)$;
  $comClique \leftarrow comClique \cup \{v_i\}$;
  $FindNeiClique(\alpha, \beta, subSet(\alpha, \beta), comClique)$;
else i<k
  $i \leftarrow i + 1$;
  $FindClique(\mu, nei(\mu), i)$;
end if

Algorithm 2 Algorithm of FindNeiClique

$set\_nei(\alpha) \leftarrow findComNei(\alpha, \beta, subSet(\alpha, \beta))$ /* $set\_nei(\alpha)$ is the set of the common neighbors of α and β in $subSet(\alpha, \beta)$ using the Function $findComNei$ */;
$set\_unei(\alpha) \leftarrow subSet(\alpha, \beta) - set\_nei(\alpha)$;
$j \leftarrow$ number of set $Clique$;
if $set\_nei(\alpha) == NULL \&\& set\_unei(\alpha) == NULL$ **then**
  $Clique(j) \leftarrow comClique$;
  $comClique \leftarrow NULL$;
  $i \leftarrow i + 1$;
  $j \leftarrow j + 1$;
  $FindClique(\mu, nei(\mu), i)$;
else
  for each item (i.e., v_f) in $set\_unei(\alpha)$ **do**
    $v_f.hasVisited \leftarrow true$;
    $Clique(j) \leftarrow comClique \cup v_f$;
    $j \leftarrow j + 1$;
  end for
  for each item (i.e., v_m) in $set\_nei(\alpha)$ **do**
    $comClique \leftarrow comClique \cup \{v_m\}$;
    $\alpha \leftarrow \beta, \alpha \leftarrow v_m$;
    $subSet(\alpha, \beta) \leftarrow nei(\mu) - comClique$;
    $v_m.hasVisited \leftarrow true$;
    $FindNeiClique(\alpha, \beta, subSet(\alpha, \beta), comClique)$;
  end for
end if

TIME COMPLEXITY

From the above description, the time complexity for obtaining the social attributes of a node (μ) in the form of cliques is $O(\sum_{i=1}^{i=B} deg(\mu_i) \times deg(\mu_i))$, where $deg(\mu_i)$ denotes the $\mu_i's$ degree. Therefore, it is highly cost to analyze the whole datasets to obtain the social attributes and the limited number of sampled nodes obtained by a typical sampling method can reduce the cost for obtaining the social attributes greatly.

TABLE 3. Summary of graph datasets.

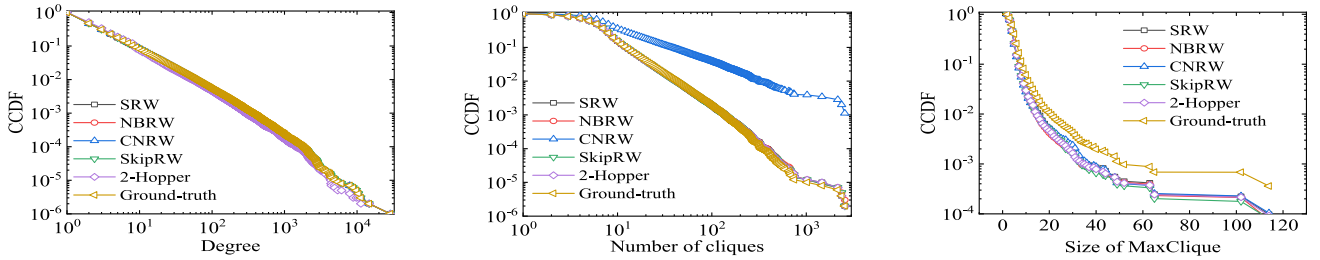
Graph	Slashdot	com-DBLP	amazon0601	com-Youtube
$\|V\|$	77,360	317,080	403,394	1,134,890
$\|E\|$	905,468	1,049,866	2,443,408	2,987,624
Average Degree	23.4	6.6	12.1	5.3

V. EVALUATION

This section describes simulation experiments conducted on a computer with Intel Xeon E5620 processors and 64-bit Ubuntu Linux OS over four real graph datasets, which are downloaded from [22] and and depicted in Table 3. Four state-of-the-art random-walk based sampling methods, namely, SRW, NBRW, CNRW and SkipRW which are described in Section II, are selected as the baselines of 2-Hopper. Furthermore, the probability (P_{SkipRW}) for selecting a node as a sample is $P_{SkipRW} = 0.5$ when SkipRW is residing on the node. All the simulations are executed more than 100 times to ensure the soundness of the experimental results. These random-walk based sampling methods are evaluated from three aspects: the accuracy of estimations on the individual and social properties, the percentage of unique samples and the costs when estimating these properties.

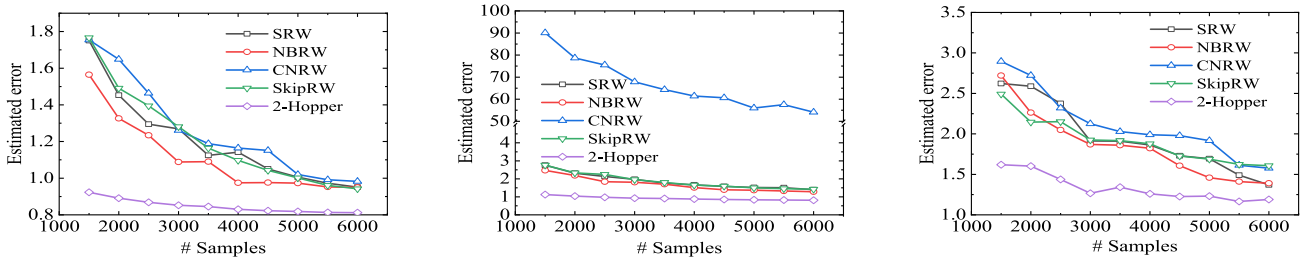
A. ACCURACY

In this paper, the accuracy of a typical random-walk based sampling method is evaluated from two aspects: the usefulness and the estimated error. The usefulness is evaluated by the comparison between the estimated values and the ground-truth values. Besides, the complementary cumulative distribution function (CCDF) is used to describe the specific estimations of a given property, such as the degree of a node which is referred as the number of the neighbors of the node and is employed as the representative of the basic attribute of a large graph. Furthermore, the distributions of NumClique and MaxClique, which are referred as the number of the cliques and the size of the maximum clique respectively that one node participates in, are used to describe the social attributes of the nodes in the large graph. Figure 7 shows that the distributions of the three structures over three respective graphs estimated by 2-Hopper are very close to the ground-truth values. Furthermore, except for CNRW that there is a large gap between its estimation on NumClique distribution and the true value, Figure 7 shows that the other existing random-walk based sampling methods can estimate the three structural properties relatively accurately. Accurate estimations on the structural properties are the premise for analyzing the formations of the structures of these samples. The reason for CNRW's inaccurate estimation lies in that CNRW has blocked the visited sampling paths arbitrarily and then misses some important samples to reflect the structural properties in the form of NumClique. The closeness between



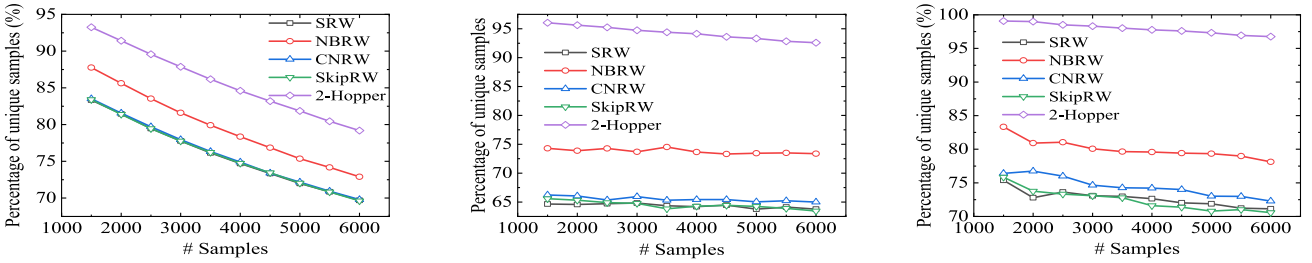
(a) Degree distributions over com-Youtube (b) NumClique distributions over amazon0601 (c) MaxClique distributions over com-DBLP

FIGURE 7. The distributions of the basic attributes represented by degrees and the social attributes represented by the number of cliques and the maximum clique that nodes participate in. Note that each data point (x, y) in the figures indicates that $|V| \times y$ nodes are of degree in (a), NumClique in (b) and MaxClique in (c) equal or smaller than x where $|V|$ is the total number of nodes among the respective graph. 2-Hopper can estimate the structural properties accurately that is the premise for analyzing the formations of these structures.



(a) com-Youtube (b) amazon0601 (c) com-DBLP

FIGURE 8. The mean estimated errors when the five random-walk based sampling methods are used to estimate the distributions of degree, NumClique and MaxClique over com-Youtube, amazon0601 and com-DBLP respectively.



(a) Slashdot (b) com-DBLP (c) com-Youtube

FIGURE 9. The ratio of unique samples obtained by the five sampling methods as a function of sampling budget.

the estimated values and the ground-truth values illustrate the usefulness of 2-Hopper as well as other random-walk based sampling methods.

However, besides the usefulness, when to further evaluate to what extend one sampling method can be used to estimate a large graph, the quantitative estimated errors among different sampling methods are necessary. **Normalized mean square error (NMSE)** defined below is used to evaluate the estimation error [6].

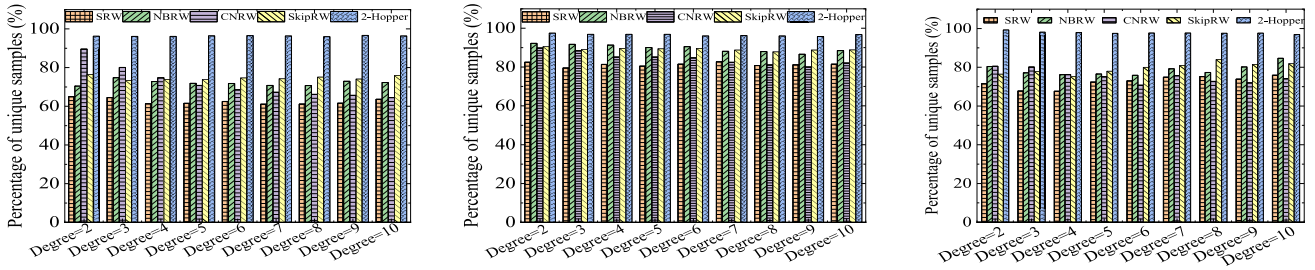
$$NMSE(\tilde{\omega_k}) = \frac{\sqrt{E[(\tilde{\omega_k} - \omega_k)^2]}}{\omega_k},$$

where ω_k and $\tilde{\omega_k}$ are the respectively true and estimated values about the graph characteristic labeled as k. Figure 8 shows that 2-Hopper exhibits the smallest mean estimated errors of the three distributions over the three respective graphs.

B. UNIQUENESS

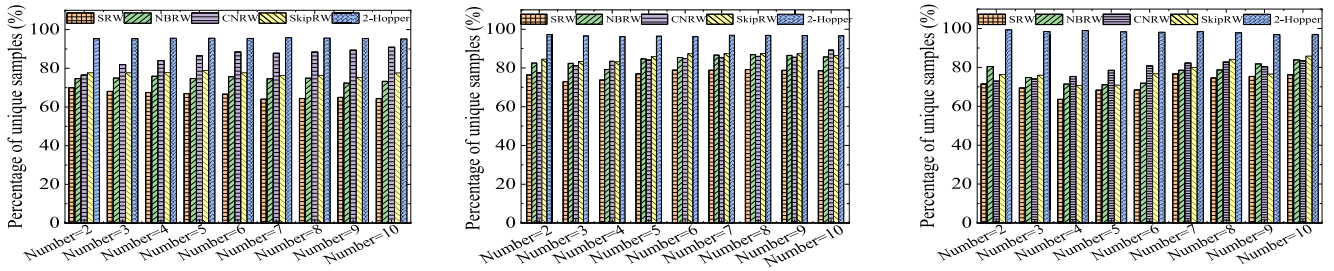
Figure 9 shows the ratio of unique samples produced by the four baseline methods ranges from 65% to 87% over the three datasets which is much lower than that of 2-Hopper, between 81% and 97%. Although 2-Hopper does not use the strategy of non-backtracking to the previously sampled nodes or the two consecutively sampled paths, which is employed by NBRW and CNRW respectively, it is able to significantly reduce the number of repetitive samples across the two datasets as a function of the number of samples, by a factor ranging from 1.8x to 8.6x, with an average of 4.5x, in contrast to the four baseline methods.

As described in Section I, even if one sampling method can obtain the accurate estimations on the structural properties, it is will be more effective when it can obtain more unique



(a) com-DBLP (b) amzon0601 (c) com-Youtube

FIGURE 10. The unique samples used to estimate the ***degree structure*** over com-DBLP, amazon0601 and com-Youtube with 3000, 4000 and 6000 samples respectively. The data point (x, y) in the figures means that there are $|B| \times y$ unique nodes whose degree are equal to x. The experimental results on the three datasets show that 2-Hopper can obtain more information when to analyze the formation of the degree property.



(a) com-DBLP (b) amazon0601 (c) com-Youtube

FIGURE 11. The unique samples used to estimate the ***NumClique structure*** over com-DBLP, amazon0601 and com-Youtube with 3000, 4000 and 6000 samples respectively where the NumClique is referred to as the number of cliques that a node participates in. The data point (x, y) in the figures means that there are $|B| \times y$ unique nodes whose NumClique are equal to x. The experimental results on the three datasets show that 2-Hopper can obtain more information (i.e., the career factor in social networks affecting the clique formations) when to analyze the formation of the NumClique property.

samples with a limited sampling budget. To further uncover the more unique samples obtained by 2-Hopper, we evaluate the percentage of the unique samples which have the same structural property. Such percentage is obtained by the number of the unique samples with a given structure divided by the total number of the samples with the same structure. For example, m samples with all their respective degrees equal to 5 among n samples, then the distribution of the property of the samples with degree $= 5$ is $\frac{m}{n}$. To uncover to what extend the unique samples obtained by different sampling methods can be used to obtain the information of the different structural properties in a large graph, we describe the percentage of the samples related to the three structural properties respectively.

1) DEGREE

When these samples are used to infer the information (i.e., formations) of the property with a certain value (i.e., degree $= 5$), Figure 10 shows that more unique samples produced by 2-Hopper than that by the other four baseline methods, mean more information can be obtained to further analyze the formation of the graph structure in the form of degrees (also called degree structure).

2) NumClique

As described in Section IV, NumClique which refers to the number of the cliques that one node participates in, just reflect the structural properties of the node social attributes quantitatively. For example, the two nodes have the same values of NumClique while the the motifs of these cliques are different. Figure 11 shows that 2-Hopper can produce at least 20% more unique samples to obtain information about the given structures over the three different graphs.

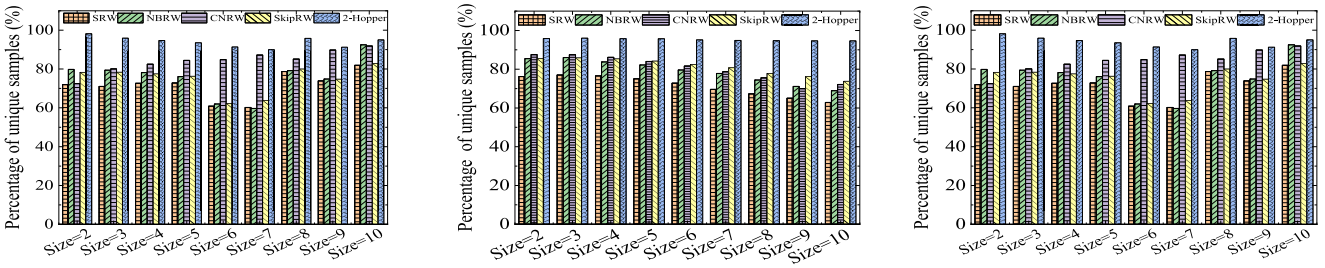
3) MaxClique

Similar to NumClique, Maxclique which refers to the number (size) of the maximum clique one node participates in, is a quantitative description about the node social attribute. The maximum clique that one node (user) of a social network participates in may reflect the major or the interest that the user has. Such information is important to provide hierarchical information about the social network. Therefore, more unique samples means that more information can be obtained from the perspective of the MaxClique structure. Figure 12 shows that 2-Hopper can produce at least 20% more unique samples to obtain information about the given structures.

C. SAMPLING COSTS

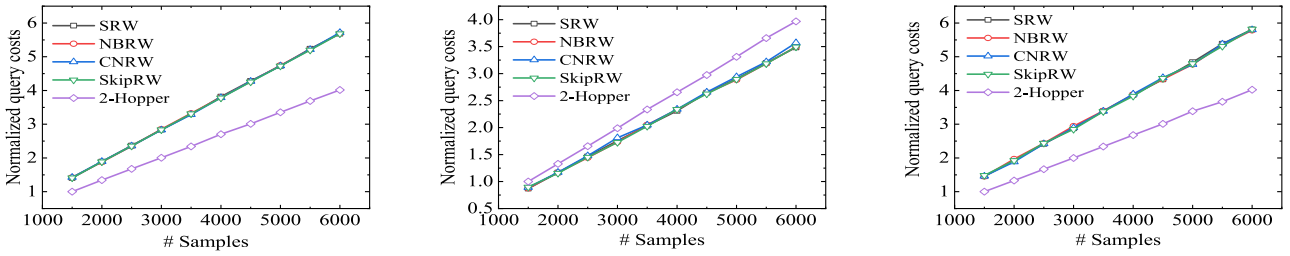
1) QUERY COST

Since the query cost is a key factor when to estimate the properties of social networks by employing a random-walk based sampling method, we use the query costs to evaluate the costs of obtaining the individual and social attributes of social networks during the process of sampling. As described in [15], obtaining a node along with its neighbor nodes can



(a) com-DBLP (b) amazon0601 (c) com-Youtube

FIGURE 12. The unique samples used to estimate the ***MaxClique structure*** over com-DBLP, amazon0601 and com-Youtube with 3000, 4000 and 6000 samples respectively. The data point (x, y) in the figures means that there are $|B| \times y$ unique nodes whose MaxClique are equal to x. The experimental results on the three datasets show that 2-Hopper can obtain more information when to analyze the formation of the MaxClique property.



(a) Slashdot (b) com-DBLP (c) amazon0601

FIGURE 13. The normalized query costs when using different sampling methods to obtain the individual and social attributes over Slashdot, com-DBLP and amazon-0601.

be consider one query from social networks. We simulate the query costs over com-DBLP and amazon-0601 as a function of the number of samples when using the five sampling methods to obtain the individual and social attributes. Figure 13(b) shows that 2-Hopper consumes slightly more query costs than the four methods while Figure 13(a) and Figure 13(c) shows that 2-Hopper consumes fewer queries than the four baseline methods. Because the four baseline methods are more biased to the nodes with higher degrees than 2-Hopper as described in Section II, these methods need lots of queries to estimate the social attributes which necessitate to acquire the neighbors of the neighbors of the sampled node. Thus, when the four baseline methods are used to estimate the dense graph reflected by a larger average degree of the nodes in Slashdot and amazon0601, they consume more query costs. Whereas, when these methods are used to estimate the sparse graph of com-DBLP, 2-Hopper consumes slightly more costs than the other four methods which can be compensated by much more unique samples.

VI. CONCLUSION

In this paper, we propose a new random-walk based sampling method, named 2-Hopper, to uncover the individual and social properties of social networks. 2-Hopper can efficiently cut down the redundant paths from one node to another to reduce the chances of the sampling process backtracking to the already sampled nodes. Consequently, 2-Hopper produces samples of social networks with fewer repeats. The experimental results driven by real-world datasets show that 2-Hopper estimates the structural properties of both the individual and social attributes accurately, if not more accurately than existing state-of-the-art sampling methods while it can obtain more information to analyze the formations of these structures.

REFERENCES

[1] M. Gjoka, C. T. Butts, M. Kurant, and A. Markopoulou, "Multigraph sampling of online social networks," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 9, pp. 1893–1905, Oct. 2011.

[2] M. Kurant, M. Gjoka, C. T. Butts, and A. Markopoulou, "Walking on a graph with a magnifying glass: Stratified sampling via weighted random walks," in *Proc. ACM SIGMETRICS*, 2011, pp. 281–292.

[3] C.-H. Lee, X. Xu, and D. Y. Eun, "Beyond random walk and metropolis-hastings samplers: Why you should not backtrack for unbiased graph sampling," in *Proc. ACM SIGMETRICS*, 2012, pp. 319–330.

[4] R.-H. Li, J. X. Yu, X. Huang, and H. Cheng, "Random-walk domination in large graphs," in *Proc. IEEE ICDE*, Mar./Apr. 2014, pp. 736–747.

[5] 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, 2019.

[6] B. Ribeiro and D. Towsley, "Estimating and sampling graphs with multidimensional random walks," in *Proc. ACM SIGCOMM*, 2010, pp. 390–403.

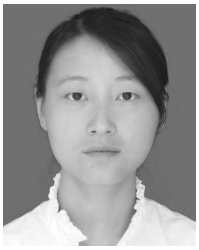
[7] B. Ribeiro, P. Wang, F. Murai, and D. Towsley, "Sampling directed graphs with random walks," in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 1692–1700.

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

[9] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proc. ACM KDD*, 2016, pp. 855–864.

[10] L. F. R. Ribeiro, P. H. P. Saverese, and D. R. Figueiredo, "*struc2vec*: Learning node representations from structural identity," in *Proc. ACM KDD*, 2017, pp. 385–394.

[11] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel, "You are who you know: Inferring user profiles in online social networks," in *Proc. ACM WSDM*, 2010, pp. 251–260.
[12] N. M. M. K. Chowdhury and R. Boutaba, "A survey of network virtualization," *Comput. Netw.*, vol. 54, no. 5, pp. 862–876, Apr. 2010.
[13] I. Herman, G. Melançon, and M. S. Marshall, "Graph visualization and navigation in information visualization: A survey," *IEEE Trans. Vis. Comput. Graphics*, vol. 6, no. 1, pp. 24–43, Jan. 2000.
[14] W. Khaouid, M. Barsky, V. Srinivasan, and A. Thomo, "K-core decomposition of large networks on a single PC," *Proc. VLDB Endowment*, vol. 9, no. 1, pp. 13–23, 2015.
[15] Z. Zhou, N. Zhang, and G. Das, "Leveraging history for faster sampling of online social networks," *Proc. VLDB Endowment*, vol. 8, no. 10, pp. 1034–1045, 2015.
[16] X. Xu, C.-H. Lee, and D. Y. Eun, "Challenging the limits: Sampling online social networks with cost constraints," in *Proc. INFOCOM*, May 2017, pp. 1–9.
[17] 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.
[18] J. Konc and D. Janežič, "An improved branch and bound algorithm for the maximum clique problem," *MATCH Commun. Math. Comput. Chem.*, vol. 58, pp. 569–590, Jun. 2007.
[19] H. Jiang, C.-M. Li, and F. Manyà, "An exact algorithm for the maximum weight clique problem in large graphs," in *Proc. AAAI*, 2017, pp. 830–838.
[20] L. Lovász, "Random walks on graphs: A survey," *Combinatorics, Paul Erdos Eighty*, vol. 2, no. 1, pp. 1–46, 1993.
[21] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*. Berlin, Germany: Springer, 2012.
[22] *Snap Datasets*. [Online]. Available: http://snap.stanford.edu/



LINGLING ZHANG is currently pursuing the Ph.D. degree in computer architecture with the Huazhong University of Science and Technology (HUST), Wuhan, China. Her research interests include big data, online social networks, and she has a paper published in ICDE'2019.

HONG JIANG received the B.Sc. degree in computer engineering from the Huazhong University of Science and Technology, China, in 1982, the M.A.Sc. degree in computer engineering from the University of Toronto, Canada, in 1987, and the Ph.D. degree in computer science from the Texas A&M University, USA, in 1991. Prior to joining UTA, he served as a Program Director at National Science Foundation, from January 2013 to August 2015, and he has been with the University of Nebraska-Lincoln, since 1991, where he was a Willa Cather Professor of computer science and engineering. He is currently the Chair and Wendell H. Nedderman Endowed Professor of Computer Science and Engineering Department, The University of Texas at Arlington. His current research interests include computer architecture, computer storage systems and parallel I/O, high performance computing, big data computing, cloud computing, and performance evaluation. He has over 200 publications in major journals and international conferences in these areas, including IEEE-TPDS, IEEE-TC, PROCEEDINGS OF THE IEEE, ACM-TACO, JPDC, ISCA, MICRO, USENIX ATC, FAST, EUROSYS, LISA, SIGMETRICS, ICDCS, IPDPS, MIDDLEWARE, OOPLAS, ECOOP, SC, ICS, HPDC, INFOCOM, ICPP, etc., and his research has been supported by NSF, DOD, the State of Texas, and the State of Nebraska. Dr. Jiang is a member of ACM and USENIX. He recently served as an Associate Editor for the IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS.

FANG WANG received the B.E. and master's degrees in computer science, in 1994 and 1997, respectively, and the Ph.D. degree in computer architecture from the Huazhong University of Science and Technology (HUST), China, in 2001, where she is currently a Professor of computer science and engineering. She has more than 50 publications in major journals and international conferences, including FGCS, ACM TACO, *Science China Information Sciences*, the *Chinese Journal of Computers* and HiPC, ICDCS, HPDC, and ICPP. Her research interests include distributed file systems, parallel I/O storage systems, and graph processing systems.

DAN FENG received the B.E., M.E., and Ph.D. degrees in computer science and technology from the Huazhong University of Science and Technology (HUST), China, in 1991, 1994, and 1997, respectively, where she is currently a Professor and the Vice Dean of the School of Computer Science and Technology. She has more than 100 publications in major journals and international conferences, including the IEEE-TC, IEEE-TPDS, ACM-TOS, JCST, FAST, USENIX ATC, ICDCS, HPDC, SC, ICS, IPDPS, and ICPP. She serves on the program committees of multiple international conferences, including SC 2011, 2013, and MSST 2012. Her research interests include computer architecture, massive storage systems, and parallel file systems. She is a member of ACM.