Knowledge and Information Systems
https://doi.org/10.1007/s10115-022-01691-8

REGULAR PAPER

Random walk on node cliques for high-quality samples to estimate large graphs with high accuracies and low costs

Lingling Zhang[1,2] · Fang Wang[2] · Hong Jiang[3] · Dan Feng[2] · Yanwen Xie[2] · Zhiwei Zhang[1] · Guoren Wang[1]

Received: 11 December 2018 / Revised: 18 May 2022 / Accepted: 21 May 2022

Abstract
Random-walk-based sampling is an efficient way to extract and analyze the properties of large and complex graphs representing social networks. However, it is almost impractical for existing random-walk-based sampling schemes to reach the desired node distribution because of the indeterministic sampling budget (i.e., the number of samples or sampling steps) required for doing so with large volumes of data in graphs. On the other hand, under a small sampling budget, these methods produce low-quality samples with many repeats and high correlations (i.e., many common attributes), which leads to a large deviation from the desired node distribution and large estimation errors. In this paper, we propose a new random-walk sampling scheme based on node cliques (a subset of cliques), called node-clique random walk, or NCRW, to strike a good balance between the estimation error and the sampling budget, by producing unique samples with low correlations. Meanwhile, both the deviation from the desired node distribution and the estimation errors under the constraint of the sampling budget are reduced both theoretically and experimentally. Thus, the sampling costs which are closely related to the sampling budget are reduced. Our extensive experimental evaluation driven by real-world datasets further confirms that NCRW significantly increases the quality of samples and accuracy of estimations with much lower costs than those of existing random-walk-based sampling schemes especially in estimating the higher-order node attributes.

Keywords Random walk · Sampling · Estimating · Graph mining

✉ Lingling Zhang
llzh@hust.edu.cn

Fang Wang
wangfang@hust.edu.cn

1 Beijing Institute of Technology, Beijing, China

2 Wuhan National Laboratory for Optoelectronics, Key Laboratory of Information Storage System Engineering Research Center of data storage systems and Technology, Ministry of Education of China, School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China

3 Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, USA

Published online: 23 June 2022

1 Introduction

The increasing volumes of data in large graphs representing social networks necessitate the use of node samples to estimate properties of the graphs efficiently. Existing sampling schemes focusing on acquiring samples can be roughly classified into three categories, namely *independent random sampling*, *traversal-based sampling* and *random-walk-based sampling*. Independent random sampling is inefficient when user-IDs of social networks are sparsely populated, resulting in low hit-to-miss ratios [34]. Traversal-based sampling, such as depth-first search (DFS) and breadth-first search (BFS), produces samples with unpredictable biases that result in inaccurate estimations [18]. Instead, random-walk-based sampling [34], [35], [24], [41], which is based on the Markov chain Monte Carlo (MCMC) approach, is widely used to acquire samples because of its simple implementation and high cost-efficiency. Therefore, this paper focuses on producing node samples by random-walk-based sampling.

Due to the huge volumes of networks, it is impossible to analyze each node in large graphs; high-quality node samples are powerful substitutes for supporting various tasks described below [45] [43].

1.1 Graph query

Since node samples have the ability to reflect the diversity of the properties in the original graphs, they are useful to query whether there are some users with some property in social networks. Querying node samples consumes much smaller costs than querying the whole graph and meanwhile the trustworthy results deriving from these nodes samples are important in network evolution and product advertisement [10] [46] [47].

1.2 Graph visualization

Analyzing a small set of nodes instead of the whole set is highly effective as huge resources in terms of computation, memory, and screen space are required to visualize a large graph. High-quality samples include important nodes, maintain reliable properties and quantify the key features so that the samples are very powerful tools for graph visualization at small cost [48] [49].

1.3 Graph representation

Node samples are effectively employed to estimate the degree distribution, the local cluster measurement and graph density. These parameters are important to accelerate the training process and further strengthen the scalability of machine learning algorithms for graph representation [33] [5].

However, the existing random-walk-based methods produce many repetitive and similar samples which are unable to reflect the diversity of the original network. Therefore, the quality of node samples produced by a random-walk-based scheme must be improved for supporting various applications. In general, three important criteria are used to measure the effectiveness of a random-walk-based sampling scheme, namely the quality of samples, the estimation error and the sampling cost. While the quality of samples determines the estimation errors, it is also affected by the cost of sampling as analyzed below.

First, high sample quality requires samples produced by random-walk-based sampling schemes to reach the desired node distribution (i.e., the stationary distribution) which is challenging. Furthermore, how many sampling steps are required by a typical random-walk-based sampling scheme to reach the desired node distribution is still a mystery because the whole topological characteristics of large graphs (i.e., online social networks) remain unknown. Consequently, it has been a common practice for existing random-walk-based sampling schemes to set a predetermined number of sampling steps or samples, also known as a sampling budget, and consider the desired node distribution reached when the budget is exhausted [34], [20], [41]. Under the constraint of the limited number of steps, the quality of samples is often very low, as a result of high ratios of repetitive and correlation samples, leading to large deviations from the desired node distribution and large estimation errors, as illustrated in Sect. 2.

Second, the quality of the samples in turn affects the sampling costs in terms of network communications, computation times and number of queries. The sampling cost is determined by the attributes themselves and the number of samples. For example, the sampling cost of estimating a basic attribute is different from that of estimating a higher-order attribute with the same sample set because it is enough to learn the neighbors of the samples to estimate a basic attribute (e.g., the degree distribution) while it is required to learn the relationships among these neighbors to characterize a higher-order attribute (e.g., the local cluster coefficient distribution). *The query cost* is referred to as the number of queries from the network while each time only one node and its neighbors are obtained by one query through the interface of online networks. The low quality of samples results in large quantities of the samples and further increasing the sampling cost.

There are generally three angles from which a random-walk-based sampling scheme attempt to improve the sample quality, including reducing the repeats, reducing the correlations (or shared attributes) among the samples, and reducing the deviation from the desired node distribution as described in Sect. 2. For example, non-backtracking random walk (NBRW) [20] and circulated neighbors random walk (CNRW) [51] attempt to reduce the repetitive samples by avoiding backtracking to the previously sampled nodes. Skipping random walk (SkipRW) [41] reduces the correlations among the samples by skipping some selected-for-sampling nodes without sampling. The study proposed in [7] sets the number of sampling steps required to reach the desired node distribution as the sampling budget, incurring huge sampling cost. However, the existing schemes usually improve the sample quality from only one of these three angles, failing to fundamentally improve the sample quality while significantly reducing the sampling cost.

Furthermore, the 'walker' in the random-walk-based sampling schemes, who *traverses* over a large graph during the sampling process to produce samples, can be in one of two states, namely the *residing state* and *the moving state*. The former is referred as the node the walker is currently on before traversing to another node while the latter is the set of candidate nodes selected as the next residing state by the walker. We observe that the root cause for the failures of these existing schemes [34], [23], [20], [51], [41] to fundamentally improve the sample quality lies in the fact that they share the essence of the simple random walk (SRW) scheme in which the two states of the walker are, respectively, the currently sampling node and its neighbors. Thus, many nodes in the moving state have high correlations with the node in which the walker is residing in and the consecutively moving states have many common nodes. The walker traversing a graph in this way, which in turn produces low-quality samples because of backtracking to the common nodes between the consecutively moving states, can be called *node-centric random-walk-based schemes.*

Inspired by the above analysis and observation, we improve the sample quality by changing the way the walker traverses a graph. More specifically, to reduce the chances of the walker backtracking to local and connected subgraphs, which is the main culprit for the existing schemes generating samples with repeats and correlations, we argue that, instead of a node, the walker should traverse from a node clique to one of its neighboring node cliques. *The clique of node, or node clique,* is defined as the biggest and completely connected subgraph corresponding to the node, while *the cliques of the graph* are defined as the completely connected subgraphs, meaning that the node cliques are a subset of the cliques in the graph. Nodes in the same clique are considered highly correlated as they share much more common neighbors than nodes in different cliques. In other words, a node clique reflects the maximum number of the common neighbors between the walker's two states.

In this paper, we propose a new random-walk-based sampling scheme, called node clique random walk or NCRW, to reduce the chances of the walker backtracking to cliques and further cut down the repetitive and similar samples. The main idea behind NCRW is to generate node samples by the walker traversing from one node clique to one of its neighboring node cliques based on the very important one-to-one relationship between a node and its node clique. During a specific sampling process, NCRW only needs to find the node clique of the currently sampled node, instead of finding all the node cliques of the nodes of a graph to reduce sampling costs. While NCRW incurs some cost for finding the node cliques for the sampled nodes on estimating the basic node attributes, this cost is more than compensated by the much smaller costs taken by NCRW on estimating the higher-order attributes. Furthermore, NCRW cuts down the sampling costs of estimating the node basic attributes because of a much smaller number of sampling steps required to produce high-quality samples than that of the other random-walk-based sampling methods.

In designing and analyzing NCRW, this paper makes the following contributions.

1 To the best of our knowledge, NCRW is the first scheme to explicitly increase the sample quality by simultaneously reducing the repeats and correlations among samples and the deviation from the desired node distribution. Node samples obtained by NCRW keep the diversity of the original graph.
2 The NCRW design rethinks the way the walker traverses a large graph to cut down the chances of the sampling process being trapped in cliques (described in detail in Sect. 3). For a prescribed sampling budget, theoretical analysis of NCRW ensures that it can increase the quality of samples, resulting in low sampling costs in terms of network communications, computation time and query costs and high estimation efficiency.
3 Under the constraint of the limited sampling budget, NCRW is able to reduce the sampling probabilities of the nodes with higher degrees while increasing those for the nodes with lower degrees, which helps significantly reduce the estimation errors and alleviate the constraints of the stationary distribution to some extent. Therefore, NCRW can estimate the distribution of key features accurately.
4 Experimental results driven by real-world graph datasets show that NCRW increases the quality of samples, with almost no repeats, fewer correlations among the samples and smaller deviation from the desired node distribution, compared with the existing random-walk-based sampling schemes. Thus, the estimations based on the samples produced by NCRW have higher accuracy with lower sampling cost than those on the samples produced by the existing random-walk-based schemes especially in estimating the higher-order attributes.

The rest of the paper is organized as follows. Section 2 describes background and motivation, which introduce the related work and further motivate the NCRW research. Section 3

Table 1 The notations used in this paper

$G = (V, E)$	An undirected graph without self-loops
V	Node set in G
E	Edge set in G
$N(\mu)$	Set of neighbors of the node μ
$deg(\mu)$	Degree of the node μ
$G_c = (NC, E_c)$	Graph G_g in terms of node cliques
NC	Node cliques associated with the node set V
E_c	Set of edges between node cliques
$subG(\mu)$	The subgraph related to μ
$NC(\mu)$	$\mu's$ node clique
$Nei(NC(\mu))$	Set of neighboring nodes of $NC(\mu)$
S	Sample set of G
t	Number of sampling steps of a single-run simulation
n	Simulation runs
T	Total number of sampling steps ($T = n \times t$)

introduces and analyzes the NCRW scheme. Evaluation results from experiments conducted on an NCRW prototype driven by a variety of graph datasets are presented in Sect. 4. Section 5 concludes our work.

2 Background and motivation

In this section, we describe and analyze the process of simple random walk (SRW) and its variations, which are most relevant to NCRW, to uncover reasons why they fail to produce high-quality samples. Furthermore, we describe the notion and importance of the desired node distribution and explain why samples with repeats and correlations increase the deviation between the measured node distribution and the desired node distribution. The observations and insight from this analysis motivate us to propose the NCRW scheme. A list of notations used in the rest of the paper is given in Table 1.

2.1 Existing sampling methods most relevant to NCRW

2.1.1 Simple random walk (SRW)

SRW first initializes a node randomly and then continues to select the next sample randomly from the neighbors of the current sampling node until the sampling budget (i.e., the number of the samples obtained or the number of the sampling steps required) is met. The transition probability from the current sampling node μ to the next sampling node ν is defined as

$$P_{(\mu,\nu)}^{\mathrm{SRW}} = \begin{cases} \frac{1}{\deg(\mu)} & if\ \nu\ is\ the\ neighbor\ of\ \mu, \\ 0 & \text{otherwise.} \end{cases}$$

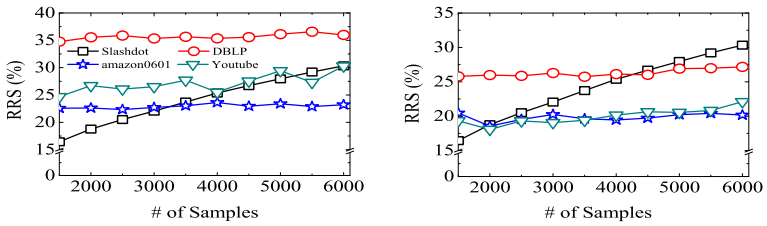


Fig. 1 The ratio of repetitive samples over the four datasets with SRW and NBRW, respectively. The number of the samples (labeled as # of Samples) is used as the sampling budget.

where deg (μ) is the node degree of μ. SRW's sampling procedure can be considered as a Markov chain-based process. Because of the nature of reversibility of the Markov chain, the sampling process can backtrack to the already sampled nodes, resulting in repetitive samples. RRS= $\frac{(B-U)}{B} \times 100$ is used to express the ratio of repetitive samples, where B is the total number of samples and U is the number of unique samples among B. Clearly, the higher the RRS value, the lower quality of obtained samples is because less useful information can be derived from the limited number of samples.

2.1.2 Non-backtracking random walk (NBRW)

Lee et al. [20] and circulated neighbors random walk (CNRW) [51] use the strategy of avoiding backtracking to the previously sampled nodes or the previously sampled paths. However, this strategy is limited to the narrowing sampling space for each step, which is formed of the very finite neighbors of the currently sampling node. When these neighbors have already been sampled, it is unavoidable to reselect from the sampled nodes or even re-initialize the sampling process. Take NBRW for example, Fig. 1 shows that SRW produces a ratio of repetitive samples between 16 and 36% while NBRW outputs that of repetitive samples ranging from 16 to 30% over the four datasets described in Sect. 4. Though NBRW employs the strategy of non-backtracking to the previous sampled nodes, it still generates significant numbers of repetitive samples. Therefore, the strategy of non-backtracking to the just previously sampled nodes alone will not be sufficiently effective.

Furthermore, the key step of SRW is to select the next sample randomly from the neighbors of the currently sampling node. Suppose that node μ is the current residing state of the walker and $\mu's$ entire set of neighbors labeled as $N(\mu) = \{\mu_1, \mu_2, ..., \mu_d\}$ denotes the walker's moving state where d is the number of the items in $N(\mu)$. NC $(\mu) = \{\mu, \mu_1, \mu_2, ..., \mu_c\}$ denotes the μ's clique where $c + 1$ is the total number of items in NC (μ), the nodes in which are completely connected. When the next sample (v) is selected from $N(\mu)$ randomly, it belongs to NC (μ) with the probability $p = \frac{c}{d}$. If c is sufficiently large, the two consecutively produced samples may be correlated as they have many common neighbors. In this paper, to measure the correlation among samples, if the node clique of the newly sampled node is different from the node cliques of the previously sampled nodes, the former is considered to have no correlations with the previously sampled nodes. RCS= $\frac{(B-C)}{B} \times 100$, called the correlation ratio, is used to quantify the percentage of correlated samples, where C is the number of the samples with no correlations among B.

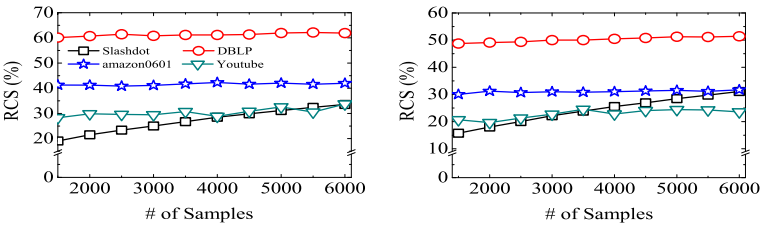


Fig. 2 The correlation ratios of samples over the four datasets with SRW and SkipRW, respectively, as a function of the number of the samples (# of Samples).

2.1.3 Skipping random walk (SkipRW)

Xu et al. [41] is proposed to reduce the correlations among the samples by means of bypassing some nodes without sampling according to a prescribed probability p. Though the walker of SkipRW traverses a large graph from one node to one of its neighbors, the same as that of SRW, SkipRW deviates from SRW in how samples are produced. Specifically, if SkipRW is currently residing in node μ (i.e., residing state being μ), μ is output as a sample with the probability p_{SkipRW}. Thus, SkipRW reduces the chances of the consecutively produced samples belonging to the same clique. However, because it cannot reduce the chances of the walker backtracking to the cliques, it does not fundamentally reduce the correlations among the samples. Furthermore, if SkipRW is required to produce the same number of samples as SRW does, the sampling steps taken by SkipRW will be much more than that by SRW. A lower value of p_{SkipRW} implies a larger number of sampling steps. Figure 2 shows that SkipRW produces a large fraction of samples with correlations, ranging from 15 to 51%, over the four datasets when $p_{\text{SkipRW}} = 0.5$ meaning the number of the sampling steps are almost twice of that of SRW, which produces the ratio of correlated samples ranging from 19 to 62%. Thus, SkipRW does not reduce the high correlations among the samples effectively while requiring larger number of sampling steps to produce the same number of samples as that of SRW.

2.2 The desired node distribution

We use $G = (V, E)$ to represent an undirected graph without self-loops where V denotes the set of nodes and E the set of edges between nodes. The ground-truth characteristics of the graph are derived as follows:

$$\text{Estimator}_{\pi_{\text{uniform}}}(f) = \sum_{\mu \in V} \frac{f(\mu)}{|V|}, \tag{1}$$

where $\pi_{\text{uniform}} = \frac{1}{|V|}$ denotes the uniform node distribution and $f : V \to R$ is a mapping function from any attribute of the nodes to a real value set R. For example, suppose f is used to characterize the degrees of the nodes, $f(\deg(\mu) = d) = 1$ $(d = 1, 2, .., |V| - 1)$ denotes $\mu's$ degree is d. Otherwise, $f(\deg(\mu) = d) = 0$. Then, $\frac{\sum_{\mu \in V} f(\deg(\mu)=d)}{|V|}$ denotes the degree distribution of the graph G.

Let S denote the set of node samples. Since the samples in S are biased as they are not sampled in the equal sampling probability, existing random-walk-based sampling methods

usually use the following estimator to weight the samples differently as a remedy to estimate properties of a graph accurately.

$$\text{Estimator}(f) = \frac{\sum_{\mu \in S} f(\mu) w(\mu)}{\sum_{\mu \in S} w(\mu)}, \tag{2}$$

where $w(\mu) = \frac{\pi_{\text{uniform}}}{\pi_{\text{desired}}}$ and π_{desired} is the desired node distribution which describes the desired probability of each node of the graph sampled by a random-walk-based scheme. The measured node distribution obtained by a sampling scheme reflects the practical probability of each node sampled by the method. It can be derived from the relative representation of each sampled node among all samples, i.e., the actual number of times a node is sampled (repeats) divided by the total number of samples produced. The estimator is unbiased when the measured node distribution is equal to the desired one [13], [24].

Specifically, most existing random-walk-based sampling schemes employ the stationary distribution of the SRW process as the desired node distribution to rectify the biases of the node samples [34], [20], [51]. If SRW's process reaches the stationary distribution, the sampling probability of the node μ converges to a fixed value: ${\pi_\mu}^{\text{SRW}} = \frac{\deg(\mu)}{2*|E|}$ [28]. In fact, since it is impossible for practical random-walk-based schemes to produce an infinite number of samples, and meanwhile it is always impractical to learn the exact number of sampling steps required to reach the desired node distribution for a large graph as explained in [9], [4], [7], the measured node distribution is often deviated from the desired node distribution. Instead, these schemes usually produce a pre-defined number of samples (i.e., $|S|$) [4], [29]. Consequently, there is a potential estimated error caused by the distance between the measured and the desired node distributions. This distance, denoted by (M-D)-distance, is evaluated as follows:

$$\begin{aligned}&(M\text{-}D) - \text{distance}\\ &= \sum_{\mu \in S} \left| \frac{\text{num}(\mu)}{|S|} - \frac{\deg(\mu)}{2|E|} \right| + \sum_{\mu \notin S, \mu \in V} \frac{\deg(\mu)}{2|E|},\end{aligned} \tag{3}$$

where $num(\mu)$ denotes the number of times μ appears in S. A smaller (M-D)-distance means that the measured node distribution is closer to the desired distribution and the errors of characterizing the node attributes estimated by the estimator described in Eq. (2) are smaller. For a fixed number of samples, Eq. (3) suggests that when the stationary distribution of SRW is considered as the desired node distribution, the samples with many repeats and high correlations increase the value of (M-D)-distance, as explained below.

1 The repetitive samples increase the value of (M-D)-distance by the large number of the nodes in the set $\{V - S = (\mu \in V, \mu \notin S)\}$.
2 The samples with high correlations (i.e., many common neighbors) tend to have their degrees cluttered in a small region, resulting in a severely skewed node degree distribution. This implies that the nodes in S just reflect only a small parts of the different attributes in a large graph. In other words, this imbalance among the samples with different attributes (degrees) leads to a large deviation from the desired node distribution.

2.3 Related work

Besides the studies mentioned above, there are many other works focusing on sampling schemes for estimating the properties of large graphs. According to the estimation goals, the sampling schemes can be divided into three groups. *The first* focus on estimating a large graph

from the perspective of the nodes. Zhong et al. [50], Stutzbach et al. [36], Ribeiro et al. [34] and [35], Gjoka et al. [12], Murai et al. [32] and Kutzkov et al. [19] propose various schemes to uncover the properties of the nodes, such as the degree distribution and the local cluster coefficient. Wang et al. [39] propose to uncover the nodes with large degrees. *The second* are designed to estimate the large graph from the perspective of the graphlets (small connected graphs formed by a given number of nodes). For example, Bhuiyan MA et al. [3], Chen et al. [6], and Wang et al. [37] develop schemes that focus on uncovering the distributions of the characteristics of the relationships among users (graphlets) for online social networks. As another example, Madhav Jha et al. [14], Jowhari et al. [15], Lim et al. [27], Lorenzo De Stefani et al. [11], Wang et al. [38], K. Ahmed [1] propose schemes that focus on counting the triangles (special graphlets) in the graph streams. *The third* are to uncover the properties of a large graph from the perspective of the content (i.e., music, books or videos stored by users in online social network), for example, the distributions of the contents (i.e., the videos) preserved in the large graphs [40], [44].

Since there are various properties in a large graph, it is infeasible to estimate all the characteristics of a large graph accurately by just using a single method. Kurant et al. [17] show that different estimation metrics may need different sampling schemes. This paper falls into the scope of the first group, i.e., discovering the properties of a large graph from the perspective of the nodes. Existing random-walk-based sampling methods in this space are required to reach their respective stationary distributions to reduce the estimation errors. Since the transition matrixes of many large graphs are unknown, Mislove et al. [30] and Mohaisen et al. [31] show that it is a mystery in many large graphs to know how many sampling steps are required to reach the stationary distribution. Although Zhou et al. [52] propose a method to cut down the required number of sampling steps by constructing a loosely connected graph, which results in additional cost, the specific number of the required sampling steps is unknown. The existing random-walk-based methods resort to setting an arbitrary value as the number of sampling steps, or budget, during a single-run simulation, resulting in huge sampling costs or estimation errors. In contrast to the existing random-walk-based sampling methods, this paper is designed to reduce the estimation errors by improving the quality of the samples while setting a small sampling budget to alleviate the constraints of the stationary distribution.

2.4 Motivation

As analyzed above, the walkers of existing random-walk-based schemes tend to backtrack to local and small subgraphs with relatively high probabilities, generating samples with many repeats and correlations and increasing deviation from the desired node distribution. Thus, under the constraint of a limited number of steps, these low-quality samples lead to large estimation errors. On the other hand, for the same sampling budget, a commonly-used strategy for existing random-walk-based schemes is to run the sampling procedures for many times to reduce the chances of the procedures being trapped in the same local and subgraphs. Therefore, the sampling costs in terms of network communications, computation time and query costs are increased when the existing random-walk-based methods are used to analyze the online networks. These problems of the existing schemes motivate us to propose a new sampling scheme, called node-clique random walk or NCRW, elaborated in the next section.

3 Design and analysis of NCRW

In this section, we first introduce the NCRW scheme with a formal description. Furthermore, we analyze and explain its improvements over the existing node-centric random-walk-based schemes given a pre-defined number of sampling steps.

3.1 Sampling scheme

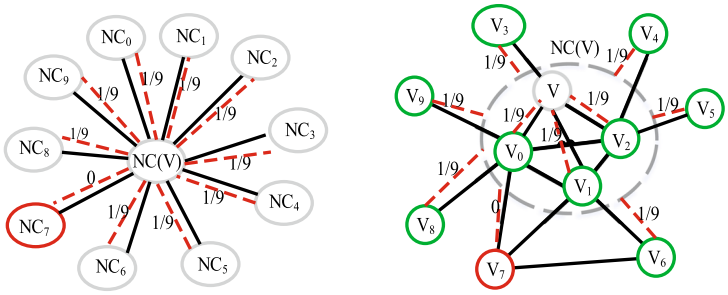
To improve the quality of the samples, we redefine the states of the walker by leveraging the structures of node cliques as defined in Sect. 1, to reduce the chances of the walker backtracking to the cliques. Each node in the graph G is associated with a unique node clique, which can be ensured by the method of finding the node clique (described later). Notice that if there are many cliques related to a node, which have the same number of the nodes, the firstly discovered one is called as its node clique. Hence, a node and its node clique have a one-to-one relationship. Thus, a graph G based on node set V can be transformed into a node-clique-based graph G_c defined as follows.

Definition 1 Graph in the form of node cliques. From the perspective of node cliques, G can be transformed into $G_c = (\mathrm{NC}, E_c)$, where NC denotes the set of node cliques and E_c denotes the set of edges connecting the node cliques. For example, NC_1, NC_2 are two neighboring node cliques of G_c, and an edge $(\mathrm{NC}_1, \mathrm{NC}_2) \in E_c$ means that there exists an edge $(\mu, \nu) \in E$, when $\mu \in \mathrm{NC}_1$ and $\nu \in \mathrm{NC}_2$.

In this paper, the node clique-based random walk scheme, namely NCRW, is proposed to reduce the chances of the process being trapped in small and connected subgraphs represented by cliques. NCRW selects the next sample from its not-already-sampled neighboring node cliques. Thus, NCRW produces the samples in the form of node cliques as illustrated in Fig. 3(a). However, it is necessary to convert the graph from G to Gc in advance to produce samples in the form of node cliques, incurring huge cost. To avoid this cost, the two states of NCRW's walker are redefined to produce samples in the form of nodes directly based on the one-to-one relationship between the node and its node clique. Specifically, NCRW redefines the residing and moving states of the walker to be, respectively, the node clique ($\mathrm{NC}(\bar{})$) of the latest sampled node (μ) and its not-already-sampled neighboring nodes (described below), denoted as $U\mathrm{Nei}(\mathrm{NC}(\bar{}))$, which makes it possible to avoid backtracking to the already-sampled-nodes and enlarge the sample selection space.

Definition 2 Neighboring nodes of the node clique $\mathrm{Nei}(\mathrm{NC}(\bar{}))$ are formed by the unique neighbors of the nodes in the node clique $\mathrm{NC}(\bar{})$, meaning that no matter whether node α is a neighbor of one or more member nodes in the node clique $\mathrm{NC}(\bar{})$, α appears in the set $\mathrm{Nei}(\mathrm{NC}(\bar{}))$ exactly once. $U\mathrm{Nei}(\mathrm{NC}(\bar{}))$ is consist of the nodes, which have not been sampled among the nodes in $\mathrm{Nei}(\mathrm{NC}(\bar{}))$.

As illustrated in Fig. 3b, the basic idea of NCRW is to walk randomly from the node clique of the currently sampled node to one of its neighboring nodes that have not been sampled. Notice that if all the neighboring nodes of the node clique have been sampled, the sampling process will be re-initialized. The probability of re-initialization decreases as the graph size increases. In our experiments, the maximum probability of re-initialization is 0.02% with different sampling budgets across the four datasets. Such a small probability of re-initialization will arguably not impact the effectiveness of NCRW noticeably as explained later.



(a) Sampling in the form of node cliques **(b)** Sampling in the form of nodes

Fig. 3 The sampling process of NCRW in two forms when NC(V) and V are the sampled node clique and node, respectively. While both forms can produce the same sample set, **b** is more cost-efficient. The node clique in **b** labeled as NC_i corresponds to $V_i, i \in \{0 \sim 9\}$. The grey circles denote the not-already-sampled neighboring node cliques while the green circles denotes the not-already-sampled neighboring nodes. The red one in (**a**) denotes the already-sampled node clique while in (**b**) denotes the already-sampled node. The values on the red dashed lines indicate the selection probabilities of the neighboring node cliques or nodes.

During the process of NCRW, it is necessary to find the node clique associated with the currently selected node to prepare the moving state for the next sampling step. This paper employs $MaxCliqueDyn(V(\mu), C(\mu))$ detailed in Algorithm 1 to find $\mu's$ clique, where $V(\mu)$ and $C(\mu)$ denote the node set in $subG(\mu)$ and its corresponding color set for $subG(\mu)$, respectively. The color set is determined by the algorithm of graph coloring problem [16]. Based on the above description, a pseudo code of the sampling scheme of NCRW is given in Algorithm 2, where the function of '$RandomSelect(UNei(NC(\mu_i)))$' is to select the next sample randomly from $UNei(NC(\mu_i))$.

Algorithm 1 MaxClique

Input: $V(\mu_i)$: the node set in the subgraph of μ_i; C : the color set of the nodes in $V(\mu_i)$
Output: Q_{max}: the maximum clique in $SubG(\mu_i)$;

```
while V(μ_i) ≠ ∅ do
    select a vertex v with the maximum color C(v) from V(μ_i);
    V(μ_i) ← V(μ_i) − v;
    if |Q| + C(v) > |Q_max| then
        Q ← Q ∪ v;
        if V(μ_i) ∩ V(v) ≠ ∅ then
            C' ← the color set of nodes in V(μ_i) ∩ V(v);
            MaxClique(V(μ_i) ∩ V(v), C');
        else|Q| > |Q_max|
            Q_max ← Q;
        end if
        Q ← Q − p;
    else
        Return;
    end if
end while
```

Intuitively, on the one hand, NCRW does not backtrack to the visited nodes and then has the strength of avoiding repetitive samples. On the other hand, for each sampling step, the sampling space is enlarged that is formed by neighbors of a node clique rather than the

Algorithm 2 NCRW

Input: t : the given number of sampling steps and $\mu_0 \in V$;
Output: S: the sample set;
1: **for** $i \leftarrow 0$ to t **do**
2: $S[i] \leftarrow \mu_i$;
3: $V(\mu_i) \leftarrow$ the node set in $subG(\mu_i)$;
4: $C(\mu_i) \leftarrow$ the color set of the nodes in $V(\mu_i)$;
5: $NC(\mu_i) \leftarrow MaxClique(V(\mu_i), C(\mu_i))$;
6: $\mu_i.hasSampled \leftarrow true$;
7: $UNei(NC(\mu_i)) \leftarrow$ the neighboring nodes of $NC(\mu_i)$ without $hasSampled = true$
8: $\mu_{i+1} \leftarrow RandomSelect(UNei(NC(\mu_i)))$;
9: **end for**

neighbors of a single node. Thus, the probability that the next sample is one of the direct neighbors of the current sample is reduced and the probability of similarity among the two samples is decreased. Therefore, NCRW has a larger chance to produce diverse samples and we will give formal analysis in Sect. 3.7.

3.2 Sampling costs

Sampling costs is measured from the following three angles when NCRW is used to estimate a property of a large graph.

1 *Network communications* are measured by the number of the total volumes of the data during the whole sampling process to be dealt with when NCRW is used to crawl the data from online networks. Assume that the cost of adding or obtaining a signal and obtaining a specific node from any graph (i.e., online social network) is set as $O(1)$. To produce a sample, NCRW needs to collect the neighbors of the nodes in the node clique of the currently sampled node with the signals to indicate whether the nodes are already sampled. Thus, for one step of sampling, the cost (labeled as Number(oneStep)) for estimating the basic node attributes is the same as that for estimating $\mu's$ higher-order attributes.

$$\text{Number(oneStep)} = \deg(\mu) + \sum_{v \in N(\mu)} \deg(v). \tag{4}$$

2 *Computation time* is mainly spent on obtaining the node cliques during the sampling procedures. To produce a sample set S, the total time of finding the node clique is $\sum_{\mu \in S} O(\deg(\mu) \times \deg(\mu))$, where $O(\deg(\mu) \times \deg(\mu))$ is the time complexity for learning a higher-order node attribute, including the node clique showed in Algorithm 1. Therefore, the time complexity for Algorithm 2 is $\sum_{\mu \in S} O(\deg(\mu) \times \deg(\mu))$.

3 Query costs are measured by the number of queries from the interfaces of online networks. For one-step sampling, the query cost (Query(oneStep)) for NCRW to produce a sample (μ) is given as follows.

$$\text{Query(oneStep)} = 1 + \deg(\mu). \tag{5}$$

NCRW needs to consume memory to preserve the samples which have already been sampled. The total volumes of the memory overheads can be labeled as $O(|S|)$. In this paper, we mainly focus on the sampling costs in terms of communication networks, computation time and query costs which are the dominant factor affecting NCRW's effectiveness, making the memory cost a negligible sampling cots.

3.3 Analysis

3.4 A formal description of NCRW

In essence, NCRW leverages node cliques to construct a higher-order Markov Chain by remembering already sampled nodes. For a pre-defined number of sampling steps t, the process of NCRW can be described as the $t-$ th-order Markov model as

$$\begin{aligned} &P(X_n = s_n | X_{n-1} = s_{n-1}, ..., X_1 = s_1) \\ &= P(X_n = s_n | X_{n-1} = s_{n-1}, ..., X_{n-t} = s_{n-t}), \end{aligned} \tag{6}$$

where i $(1 \leq i \leq t)$ is the order of the samples, X_i is the state of the Markov Chain, and s_i denotes the sampled node. Based on the sampling process of NCRW, the transition probability $P^{\text{NCRW}}_{(\mu,\nu)}$ from μ to ν is described as

$$P^{\text{NCRW}}_{(\mu,\nu)} = \begin{cases} \frac{1}{|U\text{Nei}(\text{NC}(\mu))|} & \nu \in U\text{Nei}(\text{NC}(\mu)), \\ 0 & \nu \notin U\text{Nei}(\text{NC}(\mu)). \end{cases} \tag{7}$$

3.5 Initialization

NCRW follows the basic idea of the initialization process of the existing random-walk-based sampling schemes, that is, the very first sample node is selected uniformly from the graph. Any node (μ) of a graph can be selected as the first sample node with the sampling probability $p(\mu) = \frac{1}{|V|}$. Suppose ν is a node in the set of not-already-sampled neighboring nodes $(U\text{Nei}(\text{NC}(\mu)))$ of node clique $\text{NC}(\mu)$, the transition probability from μ to ν consists of two parts. The first is that μ is selected as the initial sample while the second is that μ is not selected as the initialized sample. Thus, $P^{\text{NCRW}}_{(\mu,\nu)} = \frac{1}{|U\text{Nei}(\text{NC}(\mu))|} \times \frac{1}{|V|} + \frac{1}{|U\text{Nei}(\text{NC}(\mu))|} \times \frac{|V|-1}{|V|}$, $\mu \in UNei(NC(\nu))$. Thus, we have $P^{NCRW}_{(\mu,\nu)} = \frac{1}{|UNei}(NC(\mu))|$. Therefore, the randomness in initialization and a small probability of re-initialization does not affect NCRW's process fundamentally [2].

3.6 NCRW's desired node distribution

No matter whether the stationary distribution of the higher-order Markov chain constructed by the NCRW sampling process exist [8], [25], the goal of NCRW is to approximate SRW's stationary distribution given a very limited number of sampling steps. To estimate the properties accurately by using Eq. (2), random-walk-based sampling schemes, including NCRW, consider SRW's stationary distribution as their respective desired node distributions. Because NCRW produces samples with no repeat and much fewer correlations (explained below), NCRW is able to reduce the deviation from the desired node distribution.

As explained in Sec. 2, the random-walk-based sampling methods, including NCRW, employ a pre-defined and limited number of sampling steps or samples (i.e., $|S| << |V|$) to produce samples and control the sampling costs. This paper attempts to reduce the estimated errors from the ground-truth values with the limited number of sampling steps by using the estimator in Eq. (2). The following descriptions are to discuss the factors affecting the errors and efficiencies of the estimations.

Let $\text{Estimator}_{\text{NCRW}}(\text{f})$ denote the estimation value using Eq. (2) with the samples produced by NCRW. $\text{Estimator}_{\beta_{\text{uniform}}}(\text{f})$ denotes the ground-truth value. The mapping function f is

used to judge whether one node own the metric to be estimated or not. If the node indeed own the metric, $f(\mu) = 1$. Otherwise, $f(\mu) = 0$. Then, the distance among the estimated value and the ground-truth value, which determines the estimation errors as shown in Eq. (16), is given as:

$$\begin{aligned} &|\text{Estimator}(f) - \text{Estimator}_{\pi_{\text{uniform}}}(f)| \\ &= |\frac{\sum_{\mu \in S} f(\mu) \frac{1}{deg(\mu)}}{\sum_{\mu \in S} \frac{1}{deg(\mu)}} - \frac{\sum_{\mu \in V} f(\mu)}{|V|}| \end{aligned} \tag{8}$$

There are two cases for the distance:

1 $\text{Estimator}(f) > \text{Estimator}_{\pi_{\text{uniform}}}(f)$. In this case, the largest estimated error $Err^{(1)}_{\max}$ can be expressed as follows, where $d_{\max}$ denotes the maximum degree of the nodes in S, $d_{\min}$ the minimum degree and k is the maximum value in V with the function f. Because there is no repetitive samples in S, k is also the maximum value in S with the function f.

$$Err^{(1)}_{\max} \leq \left(\frac{k \times \frac{1}{d_{min}}}{\frac{|S|}{d_{max}}} - \frac{k}{|V|} \right) \geq 0 \tag{9}$$

Thus, we have the following relationship:

$$|S| \leq \frac{|V| \times d_{max}}{d_{min}} \tag{10}$$

2 $\text{Estimator(f)} < \text{Estimator}_{\text{ß}_{\text{uniform}}}\text{(f)}$. In this case, the largest estimated error $Err^{(2)}_{\max}$ can be expressed as follows.

$$Err^{(2)}_{\max} \leq \left(\frac{k}{|V|} - \frac{\frac{1}{d_{max}}}{\frac{|S|}{d_{min}}} \right) \geq 0 \tag{11}$$

Thus, we have the following relationship:

$$|S| \geq \frac{|V| \times d_{min}}{d_{max} \times k} \tag{12}$$

Thus, there are two angles to improve the accuracy of the estimations for a limited sampling budget.

(1) Reducing the estimation errors. It is a natural way to reduce the estimated errors by minimizing the values of $Err^{(2)}_{max}$ and $Err^{(2)}_{max}$. With a limited sampling budget (i.e., $\frac{|V| \times d_{min}}{d_{max} \times k} \leq |S| << |V|$), Eqs. (9) and (11) show that the estimated errors can be cut down by reducing the sampling probabilities of the nodes with large degrees but increasing the values of the nodes with small degrees.

(2) Improving the estimation efficiencies. If one property has different metrics, for example, the degrees of nodes in a graph have many different values, the value of $f(\mu), \mu \in S$ are required to reflect the different metrics of the property in V to improve the estimation efficiency about the property. When all of the samples in S do not have some metric, $\sum_{\mu \in S} f(\mu) = 0$. Thus, the samples are inefficiency in estimating this metric. In other words, with a limited sampling budget, it is to improve the number of unique samples in S, resulting in a small deviation from the desired node distribution as showed in Equation (3) (Sect. 2). The uniqueness is reflected from two angles: The first is that the samples are different from each other; The second is that the number of correlation samples should be small since the correlation samples in S may have the same metrics with lack of other metrics of a certain property.

Compared with the existing random-walk-based sampling methods, NCRW can obtain small estimation errors and high estimation efficiencies for its improvements as described below.

3.7 Improvements over existing random-walk-based schemes

NCRW consumes the sampling steps, or budgets, to produce unique samples, in contrast to existing random-walk-based schemes that consume a large fraction of the sampling steps to produce repetitive samples. Nevertheless, non-backtracking alone for existing node-centric random-walk-based sampling schemes is not enough to improve the quality of the samples and reduce the distance between the measured and desired node distributions. In addition to reducing the repeats, NCRW improves over the existing random-walk-based schemes in the following three important metrics.

First, with a given sampling budget, NCRW reduces the sampling probabilities of the nodes with large degrees, while increasing the sampling probabilities of the nodes with the low degrees to reduce estimation errors. Informally, in contrast to SRW, NCRW increases more ratios of the direct paths to reach the nodes with low degrees than that with high degrees, where the direct path is defined as the one-step transition of the walker to reach the goal node. On the whole, the higher sampling probability of the nodes is corresponding to the larger number of the direct paths from other nodes to itself. For example, from the perspective of cliques rather than that of the node cliques, suppose μ_1 and μ_2 share the same clique while $deg(\mu_1) < deg(\mu_2)$. In SRW, μ_1 sampling probability is smaller than μ_2's as the number of μ_1's direct paths is smaller than that of μ_2's. However, in NCRW, $NC(\mu_1)$ and $NC(\mu_2)$ have the same number (Num_{direct}) of the direct paths. Thus, $\frac{Num_{direct}-deg(\mu_1)}{deg(\mu_1)} > \frac{Num_{direct}-deg(\mu_2)}{deg(\mu_2)}$ means that NCRW indeed increases the sampling probability of the nodes with low degrees while reducing the sampling probability of the nodes with high degrees to reduce the estimation errors. Formally, suppose p_{ij}^T denote the transition probability from the node i to j while p_j denotes the sampling probability of the node j. Thus, p_j is given as follows after T steps.

$$p_j = \frac{1}{|V|}\sum_{i\in V}(p_{ij}^T) \tag{13}$$

After T steps, the transition probability from i to j is given as,

$$p_{ij}^T = \sum_{k\in V}(p_{ik} \times p_{ki})^T \tag{14}$$

Suppose j_1 denote the node with high degree while j_2 denote the node with the low degree while $NC(j_1)$ and $NC(j_2)$ have the same number of neighbors. After T steps, $p_{j_1}^{SRW}$, $p_{j_2}^{SRW}$ denote the sampling probabilities in the SRW's process while $p_{j_1}^{NCRW}$, $p_{j_2}^{NCRW}$ denote the sampling probabilities in the NCRW's process. Thus, based on Equation (14), we have $p_{j_1}^{NCRW} = p_{j_2}^{NCRW}$. Besides, because j_2 has fewer neighbors than j_1, there are many zero values in the set $\{p_{j_2k},\ k \in V\}$ and thus $p_{j_1}^{SRW} > p_{j_2}^{SRW}$. Therefore, compared with SRW, NCRW increases relatively $j_2's$ sampling probability while reducing $j_1's$ sampling probability. Thus, given with a limited sampling budget, NCRW can reduce the estimation errors.

Second, NCRW reduces the chances of producing the consequent samples which are highly correlated. The correlations between nodes μ and ν are evaluated by Jaccard coefficient $J(\mu, \nu)$ [26] as follows.

$$J(\mu, \nu) = \frac{|N(\mu) \cap N(\nu)|}{|N(\mu) \cup N(\nu)|} \tag{15}$$

Clearly, for any node pair (μ, ν) of a graph, a larger value of $|N(\mu) \cap N(\nu)|$ translates to a larger value of $J(\mu, \nu)$. Thus, if μ and ν participate in the same clique, which means that μ and ν have many common neighbors, $J(\mu, \nu)$ is significantly larger than that when μ and ν are not in the same clique. Thus, the correlations among the samples can be reflected by the number of samples participating in the same cliques.

Since the existing random-walk-based schemes have inherited how a walker traverses a large graph from SRW, SRW is considered as a baseline scheme for NCRW to analyze the chances of producing the consequent samples which are highly correlated. During each sampling step, NCRW has a much higher probability of selecting the nodes outside the clique than SRW does, as NCRW has a much larger sampling space by considering the neighboring nodes of the node clique. Furthermore, even if SRW also avoids backtracking to the already sampled nodes, the chance of the consecutive samples produced by these schemes is larger than those produced by NCRW. It is because that SRW selects the next sample just from the neighbors of the currently sampled node, while NCRW has the chance of selecting the next sample from the nodes outside the neighbors of the currently sampled node.

3.8 Third, NCRW reduces the sampling cost

A single-run simulation of a random-walk-based sampling scheme is referred to as a sampling procedure executed for only one time with a pre-defined number of sampling steps. The single-run simulation is usually simulated for many times to reduce the asymptotic variance of the estimations. Furthermore, as described in Section 1, the sampling costs can be analyzed in two cases from the comparisons of the costs of network communications, computation time and query costs: estimating for node basic attributes and estimating for higher-order attributes where the cost of finding the node clique is one of the necessary operations.

1 *Comparisons of network communications.* Let n_1 and t_1 refer the runs of the single-run simulation and the number of the sampling steps for NCRW's single-run simulation, respectively. Since the neighbors of the neighbors of the sampled nodes is required for NCRW to produce the next sample, the total volumes of the data to be dealt with for estimating the basic and the higher-order attributes are overlapping. Therefore the network communication costs for NCRW to estimate any attribute can be described as $\sum_{j=1}^{j=n_1} \sum_{i=1}^{i=t_1} (deg(\mu_i) + \sum_{v \in N(\mu_i)} deg(v))$. Let n_2 and t_2 refer as the runs of the single-run simulation and the number of the sampling steps for the existing random-walk-based schemes, respectively. For these schemes, the total number of the items to be dealt with to estimate the basic attributes is $\sum_{j=1}^{j=n_2} \sum_{i=1}^{i=t_2} deg(v_i)$ while that of estimating the higher-order attributes is $\sum_{j=1}^{j=n_2} \sum_{i=1}^{i=t_2} (deg(\mu_i) + \sum_{v \in N(\mu_i)} deg(v))$.
2 *Comparisons of computation time.* Since it is required to learn the node cliques during NCRW's sampling process, the computation time of estimating the basic node attributes is overlapping with that of estimating higher-order attributes. They are described as $O(\sum_{j=1}^{j=n_1} \sum_{i=1}^{i=t_1} (deg(\mu_i) \times deg(\mu_i)))$. On the other hand, for the other random-walk-based sampling methods, it is enough to learn the neighbors of the sampled nodes to estimate the basic node attributes. Thus, the computation time for estimating the basic node attributes is $O(\sum_{j=1}^{j=n_1} (|S_j|)$ where the time cost of learning the neighbors of one node is given as $O(1)$. When to estimating the higher-order node attributes, the time

costs for existing random-walk-based sampling methods is $O(\sum_{j=1}^{j=n_2} \sum_{i=1}^{i=t_2} (deg(\mu_i) \times deg(\mu_i)))$.

3 *Comparisons of query costs.* Similar to the costs of network communications and computation time, the total query costs of NCRW for the basic attributes are the same as that for the higher-order attributes is $\sum_{j=1}^{j=n_1} (S_j + \sum_{i=1}^{i=|S_j|} deg(\mu_i))$ because it is necessary to query the neighbors of the neighbors of the currently sampled node to obtain the next sample; For the existing random-walk-based sampling methods, the total query costs for the basic attributes are described as $O(|S|)$ while for the higher-order attributes $(\sum_{j=1}^{j=n_2} (S_j + \sum_{i=1}^{i=|S_j|} deg(\mu_i)))$.

Because the existing random-walk sampling schemes produce samples with many repeats and high correlations and meanwhile they have higher chances of being trapped in the local and small subgraphs, they need a larger number of sampling steps during a single-run simulation and lots of times of the single-run simulation to reduce the estimation errors, meaning that $t_1 < t_2$ and $n_1 < n_2$. As evaluated in Sect. 4, even if n_1 is the *one-hundredth* of n_2, NCRW still produces high-quality samples and obtains accurate estimations. Therefore, when to estimate the higher-order attributes, NCRW has the prominent improvements of reducing the costs. When to estimate the basic node attributes, NCRW incurs non-negligible cost of finding the node cliques. However, this cost can be partly made up by the cost of significantly larger number of steps ($n_2 \times t_2 >> n_1 \times t_1$) required by the existing random-walk-based sampling schemes and the higher estimation accuracy.

4 Evaluation

This section presents the evaluation of NCRW through simulation experiments conducted on a computer with Intel Xeon E5620 processors and 64-bit Ubuntu Linux OS. We choose five real-world datasets, which are summarized in Table 2 and used frequently in evaluating sampling schemes in recently published studies. In this paper, we evaluate the sampling schemes over the datasets by ignoring the directions of edges of the graphs, for ease of evaluation, though these schemes can be easily employed in the directed graphs. The five real-world datasets are described as follows.

- Slashdot was the dataset derived from the news website produced by specific user community, which include 77,360 users and their relevant 905,468 links between users.
- DBLP recorded information of research papers in computer science: if two authors coauthored at least one paper, then the two authors are considered as connected. DBLP used in this paper includes 317,080 nodes and 1,049,866 edges.
- amazon0601 was collected from Amazon as that: if one product was usually co-purchased with another product, then there was an edge between the two product. amazon0601 used in this paper includes 403,394 nodes and 3,387,388 edges.
- Youtube is a social network of video-sharing website. In Youtube, if a user share some movie with another user, there is an edge between the two users. Youtube used in this paper contains 1,134,890 users and 2,987,624 edges.
- WikiTalk collects information from website Wikipedia. Each registered user in Wikipedia has a talk page and the users can edit the information in Wikipedia. If two users edited a webpage collectively, the two users constructed an edge. WikiTalk used in this paper includes 2,394,385 nodes and 5,021,410 edges.

We select four existing state-of-the-art sampling schemes as the baseline algorithms for NCRW's evaluation, which are SRW, NBRW, CNRW and SkipRW as described in

Table 2 Summary of Graph Datasets, where $d_{\max}$ is the value of the maximum degree in the graph and $d_{\min}$ is the value of the minimum degree.

Graph	$\|V\|$	$\|E\|$	d_{max}	d_{min}
Slashdot [22]	77,360	905,468	2539	1
DBLP [42]	317,080	1,049,866	343	1
Amazon0601 [21]	403,394	3,387,388	2752	1
Youtube [42]	1,134,890	2,987,624	28754	1
WikiTalk [42]	2,394,385	5,021,410	100032	1

Sect. 2. SRW, NBRW, CNRW and NCRW take SRW's stationary distribution π_{SRW} as the desired node distribution while SkipRW considers $p \times \pi_{SRW}$ as its desired node distribution, where p=0.5 denotes the probability of skipping the nodes without sampling. The estimator described in Equation (3) is used to estimate the distributions of the degree and local cluster coefficient for the purpose of evaluating the five sampling algorithms. For each simulation experiment, SRW, NBRW, CNRW and SkipRW are simulated for 1000 times, the minimum number of simulation runs required by these sampling schemes over the four dataset with given budgets. As the process of NCRW cannot be trapped in small connected subgraphs and the samples are highly effective in estimating the structural properties of large graphs, 10 times of simulations of NCRW is implemented enough to estimate graphs. Furthermore, the number of the samples S is in an interval value of $\frac{|V|\times d_{min}}{d_{max}\times k} \leq |S| << |V|$.

4.1 Sample quality

4.1.1 Repeats

For a given sampling budget and scheme, the higher the RRS value (the ratio of repetitive samples as defined in Sect. 2), the lower quality of obtained samples is because less useful information can be derived from the limited number of unique samples. Measured in RRS, as a function of sampling budget, Fig. 4 shows that NCRW produces almost no repetitive samples with its RRS being smaller than 0.01% over Slashdot, DBLP, Youtube and WikiTalk. The nonzero RRS is caused by the re-initializations during NCRW's process. In contrast, the RRS of the four baseline schemes on Slashdot is between 12 and 38% while on DBLP is between 27 and 40%; furthermore, the ratios of the repetitive samples of the four baseline methods in Youtube and WikiTalk range from 18 to 30% and from 20 to 35%, respectively. Despite of the strategy of non-backtracking, neither NBRW nor CNRW can significantly cut down the repetitive samples because they narrow the sampling spaces without adding new optional nodes.

4.1.2 Correlations

Besides RRS, the ratio of correlation samples is also used to quantify the quality of the samples produced by random-walk-based sampling schemes. In this paper, if the cliques the newly sampled node participating in are different from the cliques the previously sampled nodes participating in, the former is considered to have no correlations with the previously sampled nodes when measuring the correlation among samples as explained in Sect. 3. The ratio of correlation ratio is used to quantify the percentage of correlated samples, where C is the sample set with no correlations among S. Figure 5 shows that RCS of SRW, NBRW,

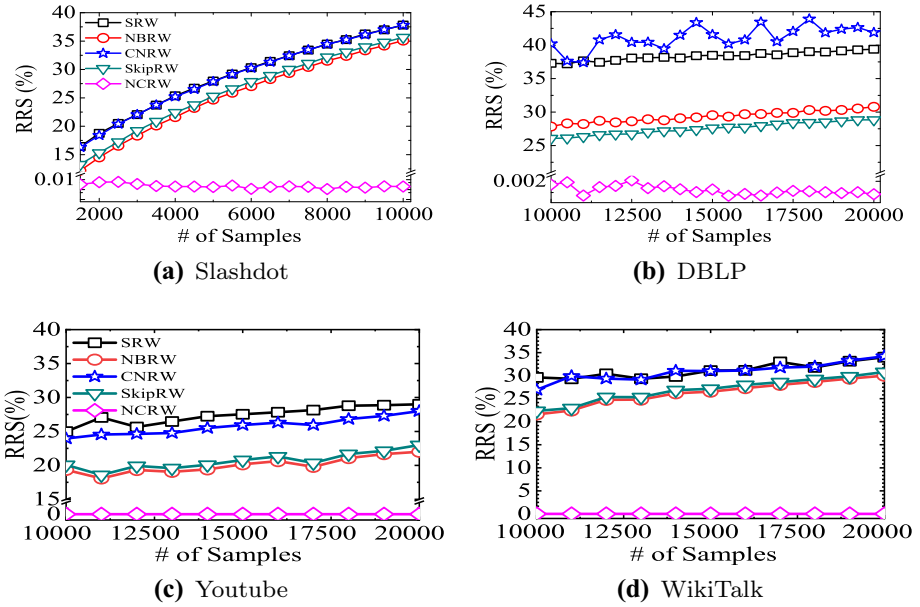


Fig. 4 The ratio of repetitive samples (RRS) of different schemes over Slashdot and DBLP as a function of the sampling budget (# of samples).

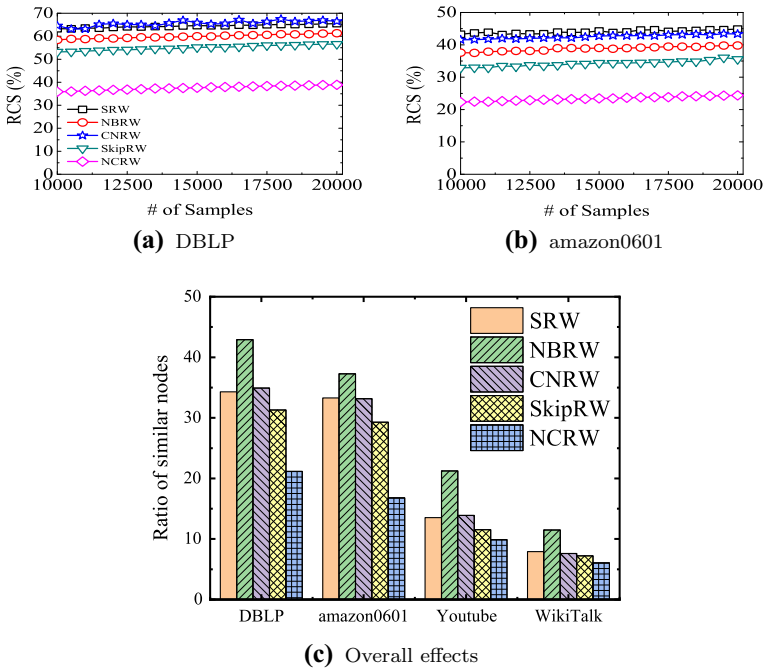
CNRW, and SkipRW ranges from 52 to 67% on DBLP and from 32 to 45% on amazon0601 as a function of sampling budget. In contrast, RCS of NCRW on DBLP is between 35 and 38% and that on amazon0601 is between 22 and 24%. Although the number of sampling steps of SkipRW is twice that of the other four schemes to produce the same number of samples by a single-run simulation, it still generates samples with many correlations with its RCS ranging from 32 to 56% over the two datasets. This is because SkipRW does not change the way the walker traverses the graph from one node to one of its neighboring nodes fundamentally. Furthermore, we do experiments over the four datasets with the sampling budget $B = 6000$ and Fig. 5 shows that NCRW produces much smaller samples than the other four random-walk sampling methods.

4.1.3 (M-D)-distance

The distance between the measured and desired node distributions is evaluated by Eq. (2) described in Sect. 2. Because NCRW reduces the repeats and correlations among the samples, Fig. 6 shows that, as a function of sampling budget, NCRW produces smaller values of (M-D)-distance than those produced by the other four schemes on DBLP and Youtube. A large value of *(M-D)-distance* indicates a large estimated error caused by using the unbiased estimator as explained in Sect. 2. Figure 6 shows (M-D)-distances over amazon0601 and WikiTalk, and Fig. 6 shows that the smaller (M-D)-distances will bring in smaller estimation errors.

4.2 Sample estimation quality

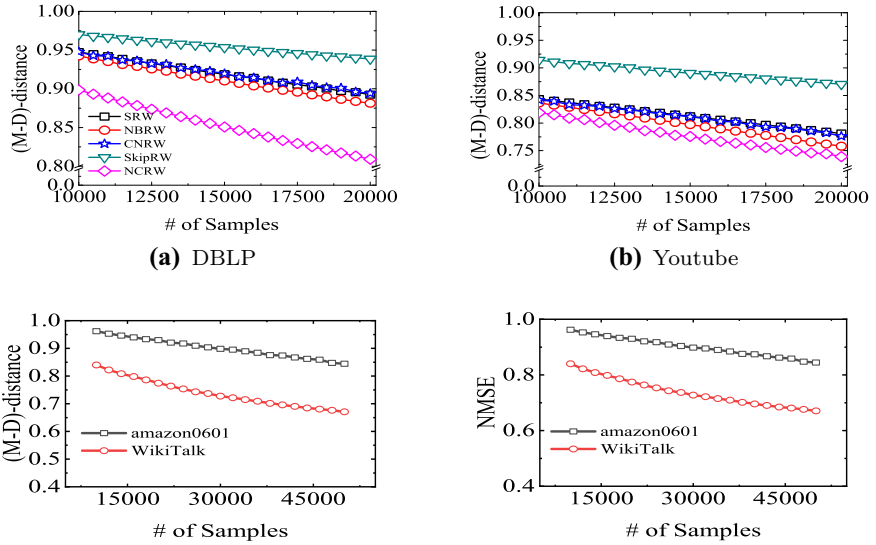
In addition to the sample quality with a pre-defined sampling budget, which we evaluated above, the quality of sample estimation can be quantified by three metrics: *(a)*. The estima-



(a) DBLP **(b)** amazon0601

(c) Overall effects

Fig. 5 The ratio of correlation samples of different schemes.



(a) DBLP **(b)** Youtube

(c) (M-D)-distances of amazon0601 and WikiTalk **(d)** Estimated errors of amazon0601 and WikiTalk

Fig. 6 The distances between the measured and desired node distributions of different schemes over DBLP and Youtube as a function of the sampling budget (# of samples).

tion error; *(b).* The efficiency of the samples; *(c).* The sampling cost in terms of network communications, computation time and query costs.

4.2.1 Estimation errors and efficiencies

To quantitatively describe the estimation errors, we adopt the measure of normalized mean square error (NMSE) defined below, with which the estimation error is considered acceptable if the corresponding NMSE value is smaller than one [34].

$$\mathrm{NMSE}(\tilde{\omega_k}) = \frac{\sqrt{E[(\tilde{\omega_k} - \omega_k)^2]}}{\omega_k}. \tag{16}$$

where ω_k and $\tilde{\omega_k}$ are the true and estimated values, respectively, of a graph characteristic labeled as k. $NMSE \geq 1$ means $\tilde{\omega_k}$ is ineffective in reflecting ω_k. The smaller NMSE means the more accurate estimated result. Suppose that a property of a graph has n different values and the samples produced by a random-walk-based scheme can effectively reflect m among the n values, then the effective ratio of these samples when used to estimate this graph property is $\mathrm{ERS} = \frac{\mathrm{m}}{\mathrm{n}} \times 100$. *The high accuracy of a typical random-walk-based sampling methods are reflected from the two angles: small estimation errors and high estimation efficiencies.*

Table 3 is an overall descriptions of the three metrics with a given sampling budget on Youtube. It shows that NCRW has higher quality of samples, lower network communication costs of estimating on the distributions of the two properties and smaller average estimated errors than SRW, NBRW, CNRW and SkipRW do when their respectively produced samples of Youtube are used to estimate the two attributes. Because the number of samples produced by NCRW is $\frac{1}{100}$ of those produced by SRW, the efficiency of NCRW is slightly smaller than SRW. Furthermore, although the computation time and query costs of NCRW for estimating the node degree distribution are smaller than that of the other four methods, NCRW incurs much smaller costs than the other four methods when estimating the local cluster coefficients. Therefore, although NBRW, CNRW and SkipRW can increase the quality of samples just from one aspect, which indeed improve another two aspects superficially, they cannot increase the efficiency and accuracy of estimations fundamentally.

4.2.2 Estimation errors (NMSE)

From Eq. (16), the smaller value of NMSE means the estimated value is closer to the ground truth with smaller estimation errors. For a given number of sampling steps (budget), the accuracy of the estimations on the distributions of degrees and local cluster coefficients is presented in terms of cumulative distribution function (CDF) of NMSE, which shows the probability of NMSE being smaller than a particular value between 0 and 1. If a value of CDF related to some value of NMSE f is equal to t, it means that the ratio of number of the estimated errors smaller than f in the total number of estimated errors is t. Therefore, with a given value of NMSE, the larger value of CDF means more properties which are estimated at smaller estimated errors. Figure 7 shows that, given the same number of sampling steps ($t = 8500$) over DBLP and Youtube, NCRW's CDF, as a function of small NMSE, is larger than the other four schemes, meaning that the estimations based on samples produced by NCRW have higher accuracy than those based on the other four schemes.

Table 3 Estimations on Youtube with the given number of sampling steps ($|S| = 8500$).

	SRW	NBRW	CNRW	SkipRW	NCRW
RRS(%)	31.16	28.42	28.53	21.47	**0**
RCS(%)	34.42	31.85	31.9	26.90	**4.42**
(M-D)-distance	0.853	0.852	0.850	0.949	**0.839**
ERS (degree)(%)	**88.14**	86.1	86.6	74.51	77.81
aveNMSE (degree)	0.56	0.58	0.58	0.63	**0.372**
ERS (local cluster coefficient)(%)	**8.98**	7.70	7.92	4.14	8.28
aveNMSE (local cluster coefficient)	0.74	0.73	0.74	0.73	**0.47**
NormNCost(degree)	4.25	4.69	4.63	4.40	**1**
NormNCost(local cluster coefficient)	129.60	141.17	166.71	64.06	**1**
NormTCost(degree)	**0.024**	0.026	0.031	0.053	1
NormTCost(local cluster coefficient)	109	113	109	111	**1**
NormQCost(degree)	**0.09**	**0.09**	**0.09**	0.17	1
NormQCost(local cluster coefficient)	434	479	448	457	**1**

RRS and RCS represent the ratio of repetitive and correlative samples, respectively, while M-D denotes the distance between the measured and desired distributions. ERS($*$) and aveNMSE($*$) denotes the effective estimations and average NMSE on the properties. NormNCost($*$) denotes the costs of network communications, NormTCost($*$) denotes the time of computation time when estimating the properties while NormQCost($*$) denotes the query costs when estimating the properties

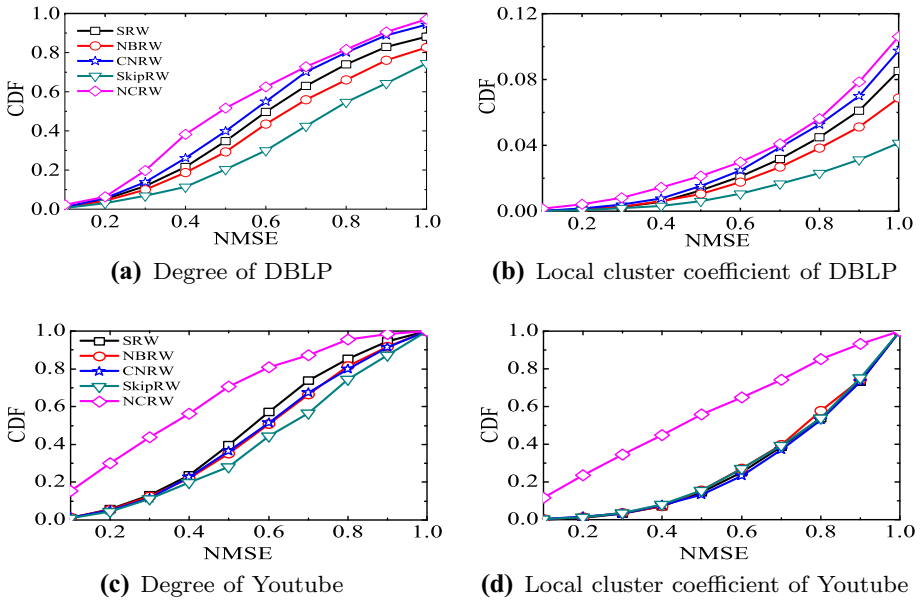
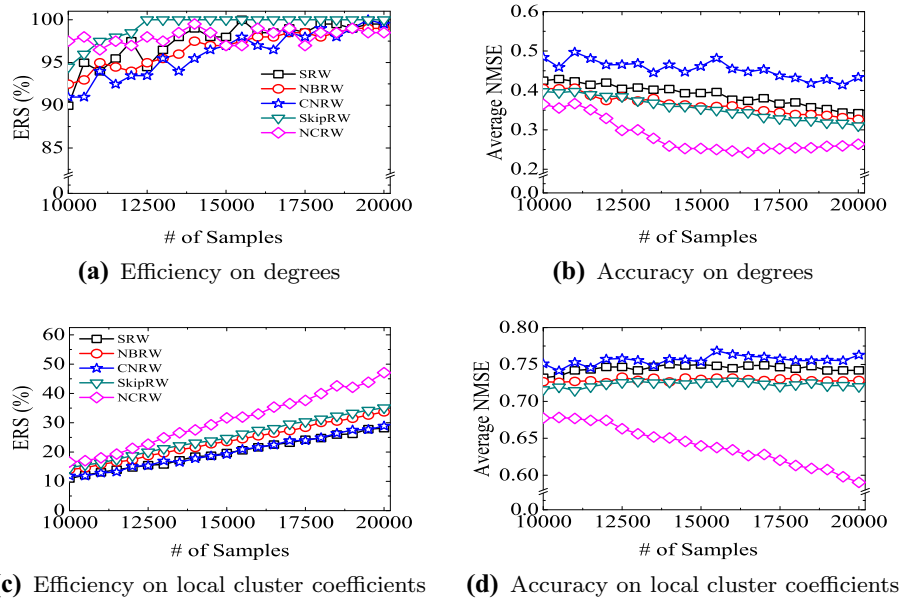


Fig. 7 The CDF of the estimated errors when the five different methods are used to estimate the distributions of the degree and local cluster coefficient with a give number of the sampling steps over DBLP and Youtube ($S = 8500$).



(a) Efficiency on degrees **(b)** Accuracy on degrees

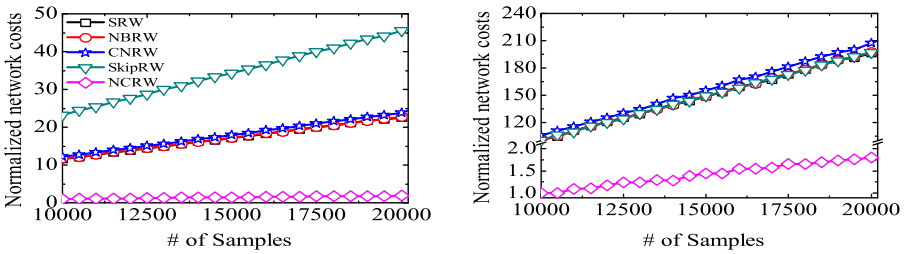
(c) Efficiency on local cluster coefficients **(d)** Accuracy on local cluster coefficients

Fig. 8 The estimation efficiency and accuracy on DBLP about the distributions on degree and local cluster coefficient with the samples produced by different methods.

4.2.3 Estimation efficiency (ERS)

Figure 8 shows that all the five schemes have very high ERS values on estimating the degree distribution, particularly with large sampling budgets, indicating that samples produced by them are effective in estimating degree distributions. Figure 8, on the other hand, shows that the average estimation errors caused by NCRW are much smaller than those by other four random-walk-based schemes. Furthermore, Fig. 8 shows that the samples produced by NCRW are more effective than those by the other four schemes in estimating the distribution of local cluster coefficient, while Fig. 8 shows much smaller average errors of estimation by NCRW than those by the other four schemes. Furthermore, Fig. 8 shows the estimated errors of the four baseline methods are decreased with the increase of the number of samples. This is because they obtain more low-quality samples in terms of repeats and similarity when the budgets are larger. These low-quality samples have similar local cluster coefficients which are short of diverse properties, resulting in low estimated errors. The experimental results confirm that the quality of samples has very high impact on estimation efficiency and accuracy, the higher the better.

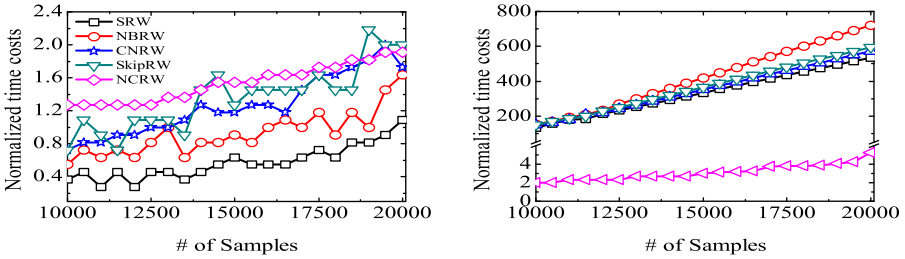
The costs of network communications are incurred when these sampling schemes are used to crawl online networks. Such costs are measured by the total volumes of data to be dealt with during the whole sampling process. When these sampling schemes are used to produce and analyze samples on amazon0601, Fig. 9 shows that NCRW incurs at least 11X and 112X lower communication cost, respectively, on estimating the distributions of degrees and local cluster coefficients than the other four sampling schemes. The percentage of NCRW's costs divided by the costs of the other random-walk-based sampling scheme is smaller than the percentage (one-hundredth) of the NCRW's runs divided by the runs of the other schemes. This is because that NCRW reduces the sampling probability of the nodes with higher degrees



(a) Costs for estimating degrees

(b) Costs for estimating local cluster coefficients

Fig. 9 The normalized network communication costs of estimating degrees and local cluster coefficients and on samples of different methods over amazon0601 as a function of the single-run sampling step.



(a) Costs for estimating degrees

(b) Costs for estimating local cluster coefficients

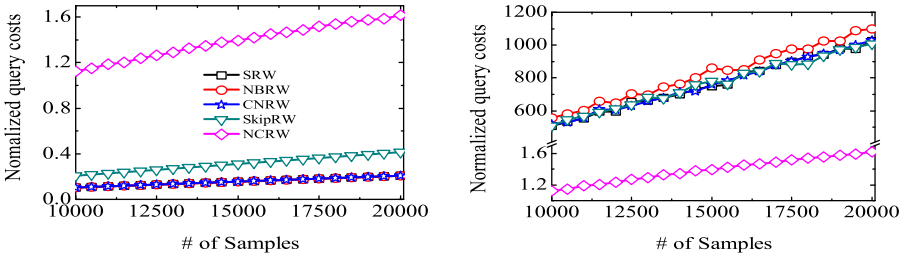
Fig. 10 The normalized computation time of estimating degrees and local cluster coefficients and on samples of different methods over com-DBLP as a function of the single-run sampling step.

which indeed increase the volumes of the data to be transmitted by networks. The estimation efficiencies and the accuracies for the two attributes on amazon0601 are similar to those on DBLP shown in Fig. 8. In addition, from Fig. 9, it can be seen that SkipRW incurs the highest sampling costs on estimating the node degree distribution because it requires almost twice the sampling steps of SRW, NBRW, CNRW and NCRW to produce the same number of samples by a single-run simulation.

4.2.4 The costs of computation time

When to estimate the degree distributions, Fig. 10 shows that NCRW consumes a bit more computation time that the other four schemes because it is required to find the node cliques of the sampled nodes. However, when to estimate the higher-order node attributes, Fig. 10 shows that NCRW consumes much more smaller (103X on average) smaller computation time than the other four random-walk-based schemes to estimate the local cluster coefficient.

Query costs are discussed into two parts. The first is to obtain the basic attributes while the second is to estimate the higher-order attributes. To illustrate the query costs among the several schemes, the query costs in the two parts are normalized by the same standards. Figure 11 shows that NCRW consumes a bit more (7.21X on average) query costs than the other four methods when to estimate the degree distributions or finish the sampling process, while Fig. 11 shows that the other four existing methods consumes much more (728X on average)



(a) Query costs for degrees **(b)** Query costs for local cluster coefficients

Fig. 11 The normalized query costs of different methods over Youtube as a function of the single-run sampling step.

query costs than the other four methods when to estimate the distributions of the local cluster coefficients.

5 Conclusions

This paper proposes a new random-walk-based sampling scheme called node clique-based random walk or NCRW, to produce high-quality samples. In particular, NCRW employs node cliques to construct a higher-order Markov chain model so that the sampling process has fewer chances of being trapped in small and local subgraphs than existing node-centric random-walk-based sampling methods. This is a first attempt at improving the quality of node samples by the means of rethinking and redesigning the way the walker traverses a large graph. We expect the design of NCRW to shed light on the design of more effective random-walk-based sampling schemes to substantially improve the quality of samples and the efficiency of sampling processes.

Acknowledgements We thanks to all the reviewers of this paper. Furthermore, this work is supported by NSFC No.61772216, 61832020,61821003, Wuhan application basic research project 2017010201010103, Fund from Science, Technology and Innovation Commission of Shenzhen Municipality(JCYJ20170307172248636).

References

1. Ahmed NK, Duffield N, Willke TL, Rossi RA (2017) On sampling from massive graph streams. VLDB 10(11):1430–1441
2. Avrachenkov K, Ribeiro B, Towsley D (2010) Improving random walk estimation accuracy with uniform restarts. In: Avrachenkov K et al (eds) Algorithms and models for the Web-Graph. Springer, Berlin, pp 98–109
3. Bhuiyan M. A, Rahman M, Rahman M, Al Hasan M.(2012) Guise: uniform sampling of graphlets for large graph analysis. In: 2012 IEEE 12th international conference on data mining, IEEE, pp 91–100
4. Chen F, Lovász L, Pak I.(1999) Lifting markov chains to speed up mixing. In: Proceedings of the thirty-first annual ACM symposium on Theory of computing, ACM, pp 275–281
5. Chen J, Gong Z, Mo J, Wang W, Wang C, Dong X, Liu W, Wu K (2021) Self-training enhanced: network embedding and overlapping community detection with adversarial learning. IEEE Trans Neural Netw Learn Syst. https://doi.org/10.1109/TNNLS.2021.3083318
6. Chen X, Li Y, Wang P, Lui J (2016) A general framework for estimating graphlet statistics via random walk. Proc VLDB Endow 10(3):253–264

7. Chiericetti F, Dasgupta A, Kumar R, Lattanzi S, Sarlós T (2016) On sampling nodes in a network. In: Proceedings of the 25th international conference on World Wide Web, international World Wide Web conferences steering committee, pp 471–481
8. Ching W-K, Ng MK, Fung ES (2008) Higher-order multivariate markov chains and their applications. Linear Algebra Appl 428(2–3):492–507
9. Cowles MK, Carlin BP (1996) Markov chain Monte Carlo convergence diagnostics: a comparative review. J Am Stat Assoc 91(434):883–904
10. Cui Y, Li X, Li J, Wang H, Chen X (2022) A survey of sampling method for social media embeddedness relationship. ACM Comput Surv. https://doi.org/10.1145/3524105
11. De Stefani L, Epasto A, Riondato M, Upfal E (2016) Trièst: Counting local and global triangles in fully-dynamic streams with fixed memory size. ACM Trans Knowl Discov Data (TKDD) 11:825–834
12. Gjoka M, Kurant M, Butts C. T, Markopoulou A (2010) Walking in facebook: A case study of unbiased sampling of osns. In: 2010 Proceedings IEEE Infocom, IEEE, PP 1–9
13. Gjoka M, Kurant M, Butts CT, Markopoulou A (2011) Practical recommendations on crawling online social networks. IEEE J Sel Areas Commun 29(9):1872–1892
14. Jha M, Seshadhri C, Pinar A (2013) A space efficient streaming algorithm for triangle counting using the birthday paradox. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp 589–597
15. Jowhari H, Ghodsi M (2005) New streaming algorithms for counting triangles in graphs. In: International computing and combinatorics conference, Springer, pp 710–716.
16. Konc J, Janezic D (2007) An improved branch and bound algorithm for the maximum clique problem. Proteins 4(5):590–596
17. Kurant M, Gjoka M, Butts C. T, Markopoulou A.(2011) Walking on a graph with a magnifying glass: stratified sampling via weighted random walks. In: Proceedings of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems, ACM, pp 281–292
18. Kurant M, Markopoulou A, Thiran P (2011) Towards unbiased bfs sampling. IEEE J Sel Areas Commun 29(9):1799–1809
19. Kutzkov K, Pagh R (2013) On the streaming complexity of computing local clustering coefficients. In: Proceedings of the sixth ACM international conference on Web search and data mining, ACM, pp 677–686
20. Lee C-H, Xu X, Eun DY (2012) Beyond random walk and metropolis-hastings samplers: why you should not backtrack for unbiased graph sampling. ACM SIGMETRICS Perform Eval Rev 40:319–330
21. Leskovec J, Adamic LA, Huberman BA (2007) The dynamics of viral marketing. ACM Trans Web (TWEB) 1(1):5
22. Leskovec J, Lang KJ, Dasgupta A, Mahoney MW (2009) Community structure in large networks: natural cluster sizes and the absence of large well-defined clusters. Internet Math 6(1):29–123
23. Li R.-H, Yu J. X, Huang X, Cheng H (2014) Random-walk domination in large graphs. In: 2014 IEEE 30th international conference on data engineering, IEEE, pp 736–747.
24. R.-H. Li, J. X. Yu, L. Qin, R. Mao, and T. Jin (2015) On random walk based graph sampling. In: 2015 IEEE 31st international conference on data engineering, IEEE, pp 927–938
25. Li W, Ng MK (2014) On the limiting probability distribution of a transition probability tensor. Linear Multili Algebra 62(3):362–385
26. Liben-Nowell D, Kleinberg J (2007) The link-prediction problem for social networks. J Assoc Inf Sci Technol 58(7):1019–1031
27. Lim Y, Kang U (2015) Mascot: memory-efficient and accurate sampling for counting local triangles in graph streams. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, ACM, pp 685–694
28. Lovász L (1993) Random walks on graphs: a survey. Combinatorics Paul Erdos Eighty 2(1):1–46
29. Lovász L,Winkler P (1995) Efficient stopping rules for markov chains. In: Proceedings of the twenty-seventh annual ACM symposium on theory of computing, ACM, pp 76–82
30. Mislove A, Marcon M, Gummadi KP, Druschel P, Bhattacharjee B (2007) Measurement and analysis of online social networks. In: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, ACM, pp 29–42
31. Mohaisen A, Yun A, Kim Y (2010) Measuring the mixing time of social graphs. In: Proceedings of the 10th ACM SIGCOMM conference on Internet measurement, ACM, pp 383–389
32. Murai F, Ribeiro B, Towsley D, Wang P (2013) On set size distribution estimation and the characterization of large networks via sampling. IEEE J Sel Areas Commun 31(6):1017–1025
33. Nakajima K, Shudo K (2021) Social graph restoration via random walk sampling. arXiv preprint arXiv:2111.11966,
34. Ribeiro B, Towsley D (2010) Estimating and sampling graphs with multidimensional random walks. In: Proceedings of the 10th ACM SIGCOMM conference on Internet measurement, ACM, pp 390–403

35. Ribeiro B, Wang P, Murai F, Towsley D (2012) Sampling directed graphs with random walks. In: 2012 Proceedings IEEE INFOCOM, IEEE, pp 1692–1700
36. Stutzbach D, Rejaie R, Duffield N, Sen S, Willinger W (2009) On unbiased sampling for unstructured peer-to-peer networks. IEEE/ACM Trans Netw (TON) 17(2):377–390
37. Wang P, Lui J, Ribeiro B, Towsley D, Zhao J, Guan X (2014) Efficiently estimating motif statistics of large networks. ACM Trans Know Discov Data (TKDD) 9(2):8
38. Wang P, Qi Y, Sun Y, Zhang X, Tao J, Guan X (2017) Approximately counting triangles in large graph streams including edge duplicates with a fixed memory usage. VLDB 11(2):162–175
39. Wang P, Ribeiro B, Zhao J, Lui J, Towsley D, Guan X (2013) Practical characterization of large networks using neighborhood information. arXiv preprint arXiv:1311.3037
40. Wang P, Zhao J, Lui JC, Towsley D, Guan X (2018) Fast crawling methods of exploring content distributed over large graphs. Know Inf Syst 59:1–26
41. Xu X, Lee CH et al (2017) Challenging the limits: sampling online social networks with cost constraints. In: IEEE INFOCOM 2017-IEEE conference on computer communications
42. Yang J, Leskovec J (2015) Defining and evaluating network communities based on ground-truth. Know Inf Syst 42(1):181–213
43. Yi P, Xie H, Li Y, Lui JC (2021) A bootstrapping approach to optimize random walk based statistical estimation over graphs. In: 2021 IEEE 37th International Conference on Data Engineering (ICDE), IEEE, pp 900–911
44. Zafar MB, Bhattacharya P, Ganguly N, Gummadi KP, Ghosh S (2015) Sampling content from online social networks: comparing random versus expert sampling of the twitter stream. ACM Trans Web (TWEB) 9(3):12
45. Zaykov AL, Vaganov DA, Guleva VY (2020) Diffusion dynamics prediction on networks using sub-graph motif distribution. In: International conference on complex networks and their applications, Springer, pp 482–493
46. Zhang L, Jiang H, Wang F, Feng D (2020) Draws: a dual random-walk based sampling method to efficiently estimate distributions of degree and clique size over social networks. Know-Based Syst 198:105891
47. Zhao J, Wang P, Lui J, Towsley D, Guan X (2019) Sampling online social networks by random walk with indirect jumps. Data Min Know Discov 33(1):24–57
48. Zhao Y, Jiang H, Qin Y, Xie H, Wu Y, Liu S, Zhou Z, Xia J, Zhou F et al (2020) Preserving minority structures in graph sampling. IEEE Trans Vis Comput Gr 27(2):1698–1708
49. Zhao Y, Shi J, Liu J, Zhao J, Zhou F, Zhang W, Chen K, Zhao X, Zhu C, Chen W (2021) Evaluating effects of background stories on graph perception. IEEE Trans Vis Comput Gr. https://doi.org/10.1109/TVCG.2021.3107297
50. Zhong M, Shen K (2006) Random walk based node sampling in self-organizing networks. SIGOPS 40(3):49–55
51. Zhou Z, Zhang N, Das G (2015) Leveraging history for faster sampling of online social networks. VLDB 8(10):1034–1045
52. Zhou Z, Zhang N, Gong Z, Das G (2016) Faster random walks by rewiring online social networks on-the-fly. ACM Trans Database Syst (TODS) 40(4):26