World Wide Web (2024) 27:18
https://doi.org/10.1007/s11280-024-01257-4

Efficiently estimating node influence through group sampling over large graphs

Lingling Zhang[1] · Zhiping Shi[1] · Zhiwei Zhang[2] · Ye Yuan[2] · Guoren Wang[2]

Received: 9 December 2023 / Revised: 8 February 2024 / Accepted: 19 February 2024 /
Published online: 29 February 2024

Abstract
The huge amount of graph data necessitates sampling methods to support graph-based analysis applications. Node influence is to count the influential nodes with a given node in large graphs that has wide applications including product promotion and information diffusion in social networks. However, existing sampling methods mainly consider node degree to compute the node influence while ignoring the important connections in terms of groups in which nodes participate, resulting in inaccuracy of influence estimations. To this end, this paper proposes group sampling, called GVRW, to count the groups along with node degrees to evaluate node influence in large graphs. Specifically, GVRW changes the way of random walker traversing a large graph from one node to a random neighbor node of the groups to enlarge the sampling space for the sake of characterizing the nodes and groups simultaneously. Furthermore, we carefully design the corresponding estimated method to employ the samples to estimate the specific distributions of groups and node degrees to compute the node influence. Experimental results on real-world graph datasets show that our proposed sampling and estimating methods can accurately obtain the properties and approximate the node influences closer to the real values than existing methods.

Keywords Graph sampling · Property distribution · Node influence

✉ Lingling Zhang
7089@cnu.edu.cn

Zhiping Shi
shizp@cnu.edu.cn

Zhiwei Zhang
zwzhang@bit.edu.cn

Ye Yuan
yuan-ye@bit.edu.cn

Guoren Wang
wanggr@bit.edu.cn

[1] Information Engineering College, Capital Normal University, Beijing, China

[2] School of Computer Science & Technology, Beijing Institute of Technology, Beijing, China

1 Introduction

With the penetration of social networks, they can be employed in many fields, such as viral marketing [2] and information diffusion [20]. Node influence is defined as the number of nodes it affects in the graphs. It can be used in estimating the number of ultimately influenced users when a set of users is randomly chose to promote a product. The node influence can also be employed to estimate the influence of the whole network. Many researchers [3, 8, 29] have studied influence of social networks in the way of finding a small set of influential nodes. The set brings in maximum affected nodes. The existing studies mainly employ a stochastic cascade model to search the set of nodes. With the model, greedy and heuristic algorithms are implemented [7, 28]. These studies address the problem of maximum influence with the whole datasets while they do not evaluate node influence in social networks. For example, in order to recommend products, a third party selects 100 users randomly from social networks. The existing studies can not estimate the ultimate affected users.

In this paper, we estimate the node influence from the perspective of the structures of graphs. Nodes are influenced with each other through the edges when social networks are modeled by graphs [17]. There are two ways in which the node influences other nodes in a large graph. One is that it affects its neighbors. The other is that it affects other nodes through its participated groups. We define a group in which nodes are completely connected. Compared to the relaxed definition of community in which nodes are densely connected, the definition of the group is rigorous. Obviously, the group is benefit to express node influence of graphs. Figure 1 shows the ways in which the node "V" affects other nodes. In Figure 1, the neighbors of "V" in Figure 1 are divided four groups. The node "V" influences the neighbors of groups.

Thus, the problem of estimating node influence can be divided into the tasks of characterizing the structures of graphs. According to the two propagating ways of a given node, two relevant structures can be used to evaluate node influence in a graph: one is the node degree distribution and the other is the characteristics of groups of the node. The groups of a node are divided into two types: the maximum group which contains the largest number of nodes

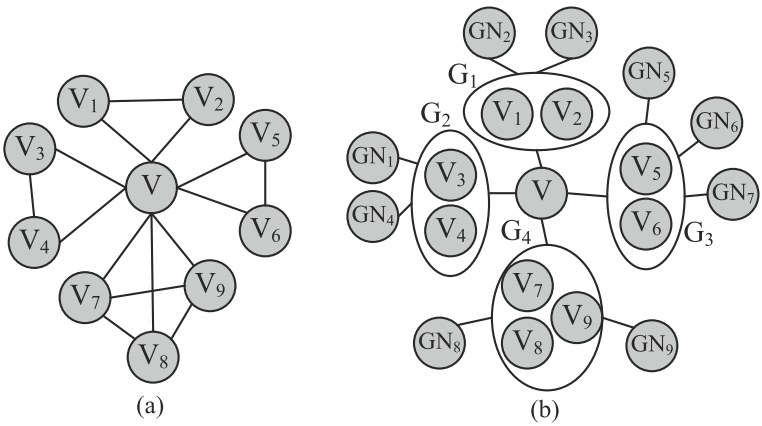

Figure 1 The node "V" affects others in two ways. It can influence the neighbors showed in (a) and V1,...,V8 denote the neighbors of "V". It also influences the neighbors of its participated groups presented in (b). In (b), G1,...,G4 denote the groups of "V", and GN1,...,GN9 denotes the neighbors of the groups

and remaining ones. In this paper, the maximum group of the node is referred as the node clique. Two properties of the groups are considered to evaluate the node influence: one is the number of groups of the node and the other is the size of the node clique.

Due to the great volume of data, how to characterize a large graph is a challenging problem. Worse still, the whole datasets of networks are usually unavailable to the third party. Thus, sampling techniques [5, 10, 15, 16] are popular to estimate the structures of graphs. Most of sampling techniques employ the way of a random walker as that: they select the next node randomly from the neighbors of the previous sampled one. However, these techniques mainly estimate the node degree distribution accurately while ignoring the characteristics of groups of the node. Even if they are employed to estimate the properties of groups by using the sampled nodes, the estimated results of the characteristics of groups are inaccurate. This is because these techniques ignore a lot of important connectives among the neighboring nodes when sampling and they tend to backtrack to the sampled nodes repetitively [10, 12].

In this paper, we propose a new method called group sampling to estimate node influence of large graphs by accurately estimating the structures of both groups and individual nodes. Differently from existing sampling techniques, group sampling chooses the next sampling node from the neighbors of groups of node. In this way, group sampling expands the range of selecting ranges and avoid lots of repetitive samples. Then, the samples can be employed to estimate node influence of graphs accurately. We employ group sampling to estimate three fundamental characteristics of graphs and then use these characteristics to estimate the node influence of graphs. The three characteristics are described as follows.

- The degree distribution. The degree of a given node is defined as the number of its neighbors. The degree distribution is one of the most common features obtained by previous graph sampling methods. We select it as a representative property of large graphs as it can be used to evaluate the influenced nodes by a given node. Group sampling is able to estimate the degree distribution in large graphs more accurately than the existing sampling methods.
- The distribution of the number of groups related to a vertex (GRV). The distribution of GRV in a large graph can be used to evaluate the connectivity of a node and then measure its influence through nodes in the groups.
- The distribution of the node clique size [1]. The node clique is referred as the largest group related to a given node [6]. Cliques are important components of large graphs [20]. By using information about the distribution of the clique size, the maximum nodes that can be influenced by the given node can be estimated.

The degree distribution is employed to estimate the node influence from the perspective of the direct connections of large graphs. The distributions of GRV and the node clique size are used to estimate the node influence from the perspective of the connections in the form of groups which reflect the indirect connections among nodes. The former is labeled as Individual-level-influence of a node while the latter is referred as group-level-influence of a node. Group sampling is able to obtain the above three distributions by traversing a large graph from one node to another node which has connections with the groups of the previously sampled node. In this paper, we make four contributions as follows.

1. To our best knowledge, we are the first to try to estimate node influence from the structures of graphs. Based on this idea, we propose a group sampling method named GVRW to obtain samples from large graphs by re-designing the traversal paths of the random walker over a large graph.

2. In order to obtain the distributions of GRV and the node clique size, a recursive algorithm is proposed to find the groups related to the node. Furthermore, to estimate three characteristics of graphs accurately, we design weighted estimator to employ these samples for accurate estimations on the above three distributions of a large graph.
3. We make extensive experimental evaluation on four datasets using different sampling methods. The results show that the node influence is estimated more accurately by GVRW compared with the existing sampling methods. The node influence of the graphs is close to the real values which are obtained through analyzing the whole datasets.

The rest of the paper is organized as follows. Section 2 describes preliminaries about group sampling. Section 3 introduces the algorithm of GVRW. Section 4 presents the experimental results in a variety of datasets while Section 5 introduces the related work. Section 6 concludes our work.

2 Preliminaries

In this section, we first introduce the definitions which are used in this paper. Then we introduce the definition of node influence followed by the introduction of popular sampling techniques. To estimate the properties accurately, we describe the frequently-used estimators which can be used to correct the sampling biases.

2.1 Definitions

We denote a undirected and no self-loop graph as $G = (V, E)$, where V denotes the set of nodes and E is set of edges between nodes. The set of neighbors of node μ is defined as $NEI(\mu)$. We define the degree of node μ as $D(\mu)$ and denote a subgraph related to node μ as $subGraph(\mu)$ in which the node set $V(\mu)$ contains itself and all of the neighbors while its edges set $ENEI(\mu)$ contains all of the edges(α,β) where $\alpha, \beta \in (\mu, NEI(\mu))$. The degree of node v in $subGraph(\mu)$ is denoted by $degS(v)$. From the perspective of groups in large graphs, we define the graph as $G_g = (C, E_g)$, where C is the set of groups, and E_g the set of edges between groups. If c_1, c_2 are two groups of G_g, there is an edge $(c_1, c_2) \in E_g$ when these conditions: $(u, v) \in E, u \in c_1$ and $v \in c_2$ are true. Nodes in large graphs usually take part in different groups. We define the groups related to node μ as $GRV(\mu)$. In this paper, we denote the number of the groups in which node μ takes part as $NG(\mu)$. For a given node, the node clique is defined as the maximum group of the node. We define the size of $\mu's$ clique as the number of nodes in the clique denoted as $sizeC(\mu)$. For a property c of graph, let function $F(c)$ define the value. These three $NG(\mu)$, $D(\mu)$, $sizeC(\mu)$ can be seen as graphs' properties. Let $\{\alpha_1,...,\alpha_k\}$, $\{\beta_1,...,\beta_k\}$, $\{\gamma_1,...,\gamma_k\}$ define the range of the functions $F(NG(\mu))$, $F(D(\mu))$, $F(sizeC(\mu))$. We propose a group sampling method to estimate distributions of three metrics denoted by $\omega^{(C(\mu))}=(\omega_1^{(C(\mu))},...,\omega_k^{(C(\mu))})$, $\omega^{(D(\mu))}=(\omega_1^{(D(\mu))},...,\omega_k^{(D(\mu))})$, $\omega^{(M(\mu))}=(\omega_1^{(M(\mu))},...,\omega_k^{(M(\mu))})$. where $\omega_k^{(C(\mu))}$, $\omega_k^{(D(\mu))}$, $\omega_k^{(M(\mu))}$ denote equations $F(NG(\mu)) = k$, $F(D(\mu)) = k$ and $F(sizeC(\mu)) = k$ respectively (Table 1).

2.2 Node influence

Different social networks can be ordered according to their node influences that provide valuable information for product marketing. For example, the market promoter should decide

Table 1 The definitions used in this paper

$G = (V, E)$	Graph G
$\|V\|$	Number of nodes in G
$\|E\|$	Number of edges in G
$NEI(\mu)$	Set of neighbors of the node μ
$D(\mu)$	Degree of the node μ
$NG(\mu)$	The number of $\mu's$ groups
$GRN(\mu)$	$\mu's$ groups
$sizeC(\mu)$	Number of the members in $C(\mu)$
$\omega(k)$	The distribution of a property k
B	Sampling budget

which social network (i.e., Facebook or Twitter) is a better choice in face of selecting 100 users randomly to promote the produce. Node influence is defined as the mean influential number of users through a random user which can be used to order the influence of social networks.

Existing methods study the node influence from two types. The first type of existing methods is to aim at finding a subset of users to maximize the spread of influence [28] while the second type of existing usually use the degree distribution to estimate the node influence of social networks [21]. They do not consider structural diversity which plays important role in social networks. Structural diversity aims at describing the the connected components among the neighbors of the individual users. Therefore, the structural characteristics of social networks in the form of cliques (groups), can be used to order node influence of social networks. In this paper, we not only obtain the three characteristics of large graphs, but also get the node influence which are denoted by IF. Let $FVMean$ denotes the mean number of groups related to a node in a large graph, and it is defined as follows:

$$FVMean = \sum_{\mu=1}^{|V|} \omega(F(NG(\mu))) \times F(NG(\mu))$$

Let $FMean$ denotes the mean scale of the clique related to a node in a large graph, and it is defined as follows:

$$FMean = \sum_{\mu=1}^{|V|} \omega(F(sizeC(\mu))) \times F(sizeC(\mu))$$

And then we measure IF by computing the maximum number of nodes that a node can influence through the function expressed ad $IF = FMean \times FVMean$. Figure 2 describes three example graphs which have different structural properties. Table 2 presents the node influence estimated by different methods. Table 2 shows that the node influences of the three graphs in Figure 2 have similar values when using the existing methods of maximizing the spread of each users and employing the existing methods based on degree distributions. However, the methods of using the structural characteristics can efficiently discriminate the node influences of the three graphs with significantly different structures.

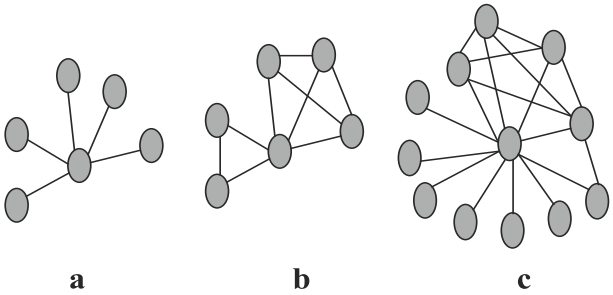

Figure 2 Different structures of graphs. a and b have the same nodes but with different number of edges. b and c have the same average degree of the nodes but with different number of nodes while they have different connections among neighboring nodes

2.3 Graph sampling

Existing studies mainly use the sampling techniques [16, 18] to estimate the structural properties of large graphs. These techniques are the variants of simple random walk (SRW) and Metropolis-Hastings random walk (MHRW) described below.

Simple random walk (SRW): The process of SRW works as that: a node is randomly selected, and then the next node is randomly selected from the neighbors from the previously sampled node iteratively. The process doesn't stop until it satisfies the sampling restrictions. The process of SRW can be seen as a Markov chain and its transition possibility p is defined as bellow.

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

The existing studies [13] have found that the process of SRW converges to a static probability as $\pi = \frac{D(\mu)}{2|E|}$, where $\mu \in V$. Thus, SRW shows biases to nodes with high degrees [4]. As the traversal paths of SRW are formed by the node pairs of nodes and their neighbors, the process of SRW can not take the groups of nodes into consideration and thus it will result in inaccurate estimations on the structural properties of groups in large graphs.

Metropolis-Hastings random walk (MHRW): The sampling process of MHRW is described as follows: First, an initial node is selected randomly i.e., (μ) and the state of the node is recorded, i.e., $\mu(t)$; Second, a node (ω) is randomly selected from the the neighbors of the previously sampled node ν and the state of the node is recorded (i.e., $\omega(t)$); Third, a transition probability $P(\mu(t), \omega(t))$ is to determine from which the sampling process can

Table 2 Ordering influence of the graphs in Figure 2

	aveMaxInflu	aveDegInflu	IF
a	1.08152	0.833	3.333
b	1.096365	3	7.58
c	1.14955	3	6.24

'aveMaxInflu' represents the first type of existing methods of maximizing the spread of each users. 'aveDegInflu' is the second type of existing methods to order influence of different graphs while 'IF' method uses the structural characteristics to order the node influences

select the next node from the neighbors of μ or ω. The latter two steps are iteratively executed until the MHRW process satisfies the sampling budget. Similar to SRW, MHRW can also be seen as a Markov chain. $P(\mu(t), \omega(t))$ is defined as follows:

$$P^{MHRW}_{(\mu(t),\omega(t))} = \begin{cases} \frac{1}{D(\mu)} \cdot min(1, \frac{D(\mu)}{D(\omega)}) & \text{if } \omega \neq \mu, \\ 1 - \sum_{\theta \neq \mu} P^{MHRW}_{(\mu,\theta)} & \text{if } \omega = \mu. \end{cases}$$

The method of MHRW converges to $\pi = \frac{1}{|V|}$ [5]. Compared with SRW, MHRW is an unbiased sampling method. Since MHRW can backtrack to the one node that has just been traversed, it leads to many repetitive samples.

2.4 Estimators and error metric

Suppose we denote the number of samples as B. We consider the property of a node as pro and the range of the property is $\{\alpha_1,...,\alpha_k\}$. We set weights for the sample node (i.e., μ) as $w(\mu)$. Then, Horvitz-Thompson estimator is usually used to estimate the distribution of pro through the following equation:

$$\tilde{\omega_k} = \frac{1}{W} \sum\nolimits_{\mu=1}^{|B|} \frac{1(F(pro(\mu) = \alpha_k)}{w(\mu)}$$

where $W = \sum_{\mu=1}^{|B|} \frac{1}{w(\mu)}$, $\mu \in V$.

For evaluating the estimation accuracy, we define error metric (NMSE) as bellow:

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

In this equation, $\tilde{\omega_k}$ is the estimated value through sampling methods while ω_k is the true value.

3 Group sampling

3.1 Algorithm of FGroup

To estimate the distribution of group numbers connected to vertexes, we should first discover the groups connected to a node. In general, a node in OSNs participates in more than one group depending on its social relationship which are connected through edges in large graphs. In this paper, we consider the minimum group is made up of two nodes which are connected with each other. We exclude the isolated nodes because it has no connectivity for spreading its influence. We adopt a recursive algorithm to discover groups of a node in this paper. Given a node, i.e., μ and $NEI(\mu)$=$\{\mu_1,...,\mu_k\}$, where k is the number of $\mu's$ neighbors. The groups of a node can be discovered from the connections among neighboring nodes by a recursive algorithm: we first order the neighboring nodes in descending order according to the number of nodes which they have connections with the nodes to be sorted; Then, we mark the nodes in the order using a bitmap as that: if a neighbor has been discovered to participate a clique, we store a bit of '1' in the corresponding position in the bitmap; Otherwise, a bit of '0' is stored. Thus, if the corresponding position of a node in the bitmap stores '0', its maximum clique among the neighboring nodes are found and the corresponding positions of the nodes in the

Algorithm 1 Algorithm of FGroup.

Require: the vertex μ and $subG(\mu)$;
Ensure: the groups $GRN(\mu)$;
1: $i \leftarrow 0$;
2: $k \leftarrow findNei(\mu)$;
3: $C(\mu) \leftarrow addToGroup(\mu)$;
4: **while** i <k **do**
5: **if** $flag(\mu_i) = false$ **then**
6: $C(\mu_i) \leftarrow addToGroup(\mu_i)$;
7: $flag(\mu_i) \leftarrow true$;
8: $NEI(\mu_i) \leftarrow findNei(subG(subG(\mu)))$
9: $GRN(\mu_i) \leftarrow FGroup(\mu_i, NEI(\mu_i)) \cup C(\mu_i)$;
10: **end if**
11: i←i+1;
12: **end while**
13: $GRN(\mu) = GRN(\mu_1) \cup GRN(\mu_2)... \cup GRN(\mu_k)$;

clique store the bits of '1'. The process is recursively executed until all the corresponding positions of the nodes in the bitmap are filled with the bits of '1'. We denote these cliques as $\mu's$ groups. The algorithm is depicted in the Algorithm 1.

Time and space complexity We assume that the computation complexity of the operation on justifying whether the two nodes μ_i and μ_j $(i, j = 1, ..., k)$ in a subgraph are connected is $O(1)$. Then finding all of the groups connected to the vertex is $O(K)$ in the best case if $subGraph(\mu)$ is a group. The worst case is $O(K \times K)$ when the neighbors of the vertex μ has no other neighbors except the vertex μ in $subGraph(\mu)$. The costs for finding groups related to a vertex are acceptable for modern computation ability.

3.2 Algorithm of GVRW

Since SRW, MHRW and their variants donot take the group structures into consideration during the sampling process, we propose a sampling method, called GVRW (Group-related-Vertex Random Walk), to traverse a large graph to estimate the distributions of GRVs, the sizes of the maximum node cliques and vertex degrees in OSNs. In other words, GVRW can simultaneously obtain node properties and groups structures in large graphs. GVRW employs the groups of nodes as the traversal units. GVRW transits from one node (i.e., μ) to another node (i.e., ν) in case that there are major differences between $G(\mu)$ and $G(\nu)$ to avoid repeated sample. Specifically, GVRW transits from one node μ to another node ν through the neighbors of $GRN(\mu)$. Take Figure 3 for example. The black nodes denote $V's$ subgraph which is comprised of three groups related to node V, labeled as $T1$, while the nodes in white color are the neighbors of $GRN(V)$. The next sample is selected randomly from the neighbors of the nodes in white color, labeled as $T2$. Compared to SRW which transits from one node to its neighbor node, GVRW enlarges the number of possible choices for selecting the next sample. Furthermore, we employ the idea of non-backtracking sampling for GVRW to sample large graphs, meaning that the previous sample will not be sampled again in the next sampling process. The process of GVRW is described as follows:

1. Initialize a node (i.e., μ) randomly;
2. Employ the algorithm of FGroup to obtain all of the groups related to the node $GRV(\mu)$ and record the numbers of groups in $GRN(\mu)$ denoted by $NG(\mu)$;

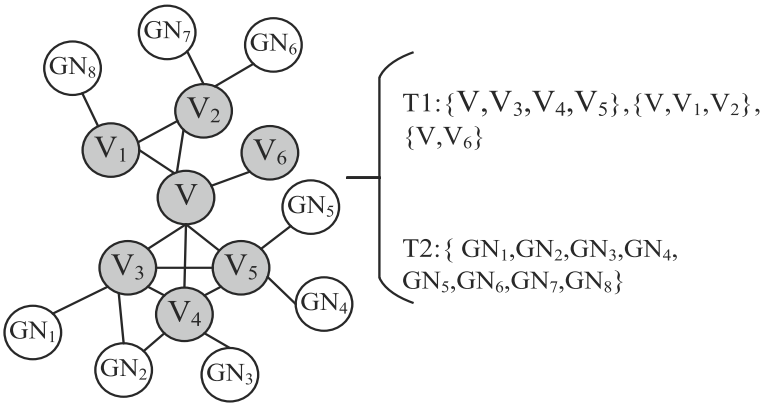

Figure 3 Example of GVRW including three groups related to node 'V' and the sampling spaces colored white for the next sampling node

3. Collect all of the neighbors denoted by $NEI(GRN(\mu)$ of the nodes in $GRV(\mu)$ which has not previously visited before and then we select the next sampled node randomly from $NEI(GRV(\mu)$.
4. Step 2 and 3 are executed recursively until the sampling budget is satisfied.

3.3 Estimator of GVRW

The process of GVRW is equivalent to a random walk traversing a graph in the form of G_{GRV} described in Section 2.1. Furthermore, the process of GVRW can be considered as a markov chain with its transition probability matrix $P = P(GRV(\mu), GRV(\nu))$. We set $m(\mu)$=$\sum_{\omega \in NEI(\mu)}(D(\omega) - SubD(\omega))$, where $SubD(\omega)$ denotes the number of edges which appear in $GRV(\mu)$. The transition probability P is defined as follows.

$$P(GRV(\mu), GRV(\nu)) = \begin{cases} \frac{1}{m(\mu)} & \text{if } \nu \in NEI(GRV(\mu)), \\ 0 & \text{otherwise.} \end{cases}$$

According to the knowledge of markov chain, the sampling process of GVRW converges to $\pi_{GVRW(\mu)} = \frac{m(\mu)}{\sum_{\nu \in V} m(\nu)}$. However, in the real application, it is complicated to compute $m(\mu)$ because it is computed by the total degrees of the nodes in $subGraph(\mu)$ subtracting the number of the edges of each node in $subGraph(\mu)$. Instead, we employ the value of $NG(\mu)$ for the substitution. We employ the equation described below to estimate the distribution of a estimated structure in a large graph denoted by C_μ.

$$\tilde{\omega_k} = \frac{1}{SUM} \sum_{\nu=1}^{B} \frac{1F(C_\nu = k))}{NG(\nu)}$$

where B is the number of total samples and $SUM = \sum_{\nu=1}^{B} \frac{1}{NG(\nu)}$.

Theorem 1 *If the graph G is non-bipartite and connected, then* $\tilde{\omega_k}$ *is an asymptotically unbiased estimator of* ω_k.

Proof We use the LEMMA 7.2 metioned in [18], we have

$$\lim_{B\to\infty}\frac{1}{B}\sum_{v=1}^{B}\frac{1(F(C_v=k))}{NG(v)}$$
$$\xrightarrow{a.s.}\frac{1}{\sum_{i=1}^{i=|V|}NG(i)}\sum_{\forall\mu\in V}\frac{1(F(C_v=k))}{NG(\mu)}\cdot\pi_{GVRW(v)}$$

Suppose a group (i.e., C_i) is selected from a graph (i.e., G) with the probability $p(c)=\frac{1}{totalG}$, and $totalG=\sum_{i=1}^{i=|V|}NG(i)$. The mean number of nodes of a group is denoted by $vmean$, the probability that we select a sample from the nodes in a group is described as $p(i)=\frac{1}{vmean}$. The number of the groups that one node (i.e.,v) participates in is denoted by $NG(\mu)$. Thus, the probability we sample node μ from the groups can be described as $p(\mu)=NG(\mu)\cdot p(c)\cdot p(i)$. Then we have

$$\frac{1}{totalG}\sum_{\forall\mu\in V}\frac{1(F(C_v=k))}{NG(\mu)}\cdot\pi_{GVRW(v)}$$
$$=\frac{1}{totalG}\sum_{\forall\mu\in V}\frac{1(F(C_v=k))}{NG(\mu)}\cdot\frac{NG(\mu)}{totalG\cdot vmean}$$
$$=\frac{1}{totalG}\cdot\frac{1}{|V|}\sum_{\forall\mu\in V}1(F(C_v=k)=k)$$
$$=\frac{1}{\sum_{i=1}^{i=|V|}NG(i)}\cdot\omega_k$$

Meanwhile, we use the Theorem 4.1 mentioned in [15] and also used in [18], we have

$$\lim_{B\to\infty}\frac{1}{B}\sum_{v=1}^{B}\frac{1}{NG(v)}\xrightarrow{a.s.}\frac{1}{\sum_{i=1}^{i=|V|}NG(i)}$$

Thus,

$$\tilde{\omega_k}=\frac{1}{SUM}\sum_{v=1}^{B}\frac{1(F(C_v=k))}{NG(v)}$$
$$\xrightarrow{a.s.}\frac{\sum_{i=1}^{i=|V|}NG(i)}{B}\sum_{v=1}^{B}\frac{1(F(C_v=k))}{NG(v)}$$
$$\xrightarrow{a.s.}\omega_k$$

□

Therefore, GVRW employs Theorem 1 to accurately estimate the characteristics of large graphs. The pseudo-code of GVRW sampling and estimating a large graph is depicted in Algorithm 2.

Time and space complexity In the process of GVRW, we require a dynamic array to preserve the neighbors of $subGraph(\mu)$. If the number of the neighbors of $subGraph(\mu)$ is K, then the dynamic array will consume $O(K)$ spaces. If the time complexity for finding the neighbors of a node is set as $O(1)$, the total time complexity for finding all the neighbors of $subGraph(\mu)$ is $O(K)$. Combined with the Algorithm FGroup, the time complexity of GVRW is $O(K)$ in the best case while the worst case is $O(K\times K)$. GVRW is to obtain three

Algorithm 2 Algorithm of GVRW.

Require: B and $\mu_0 \in E$;
Ensure: $GRV(\mu_1), GRV(\mu_2), ..., GRV(\mu_n)$;
1: $i \leftarrow 0$;
2: **while** $i < B$ **do**
3: $GRV(\mu_i) \leftarrow FGroup(\mu_i)$;
4: $m \leftarrow$ the number of groups related to μ_i;
5: $k \leftarrow$ the features of μ_i;
6: $w(\mu_i) \leftarrow w(\mu_i) + \frac{1}{m}$;
7: $totalW \leftarrow totalW + \frac{1}{m}$;
8: $\mu_i \leftarrow hasVisited$;
9: $neiG(\mu_i) \leftarrow neighbors(GRN(\mu_i))$;
10: i←i+1;
11: $\mu_i \leftarrow randomSelect(neiG(\mu_i))$;
12: **end while**
13: $\omega(1(F(chara(\mu) = k))) \leftarrow \frac{w(\mu_i)}{totalW}$

distributions of a vertex by discovering the connections among neighbors of the vertex. First, in practice, real-world networks follow skewed degree distributions such that the network contains many low-degree vertices and very few high-degree vertices. Thus, for many vertices, the specific value of K is small. Second, GVRW is a sampling method which just deals with a very proportion of vertices to obtain the distributions of the networks. Third, we can leverage the state-of-the-art techniques [1] to reduce the time complexity.

4 Evaluation

This section describes simulation experiments based on realworld graphs which are depicted in Table 3 to examine efficiencies of GVRW. The descriptions about the datasets are described below. The dataset of *DBLP* makes record about the lists of research papers in computer science. In *DBLP*, every author is considered as a node and if two authors publish at least one paper together, the two authors form an edge. The graph of *amazon0601* is collected by crawling Amazon website. In *amazon0601*, a product represents a node and an edge is formed if a product is frequently co-purchased with another product. *Youtube* is a social network about video-sharing. In *Youtube*, a user is seen as a node. The two users can be connected as an edge because of common interests or characteristics. The network of *WikiTalk* is a free encyclopedia written collaboratively by volunteers around the world. The pages in *WikiTalk* can be modified by the users of the network. Every user can be seen as a node. There is an edge between two users μ and ν if the user μ has at least once modified the page which is edited by the user ν. We conduct experiments on the four datasets while ignoring the directions of edges.

Table 3 Summary of Graph Data-sets, where $d_{\max}$ is the value of the maximum degree in the graph and $d_{\min}$ is the value of the minimum degree

Graph	$\|V\|$	$\|E\|$	$d_{\max}$	$d_{\min}$
DBLP [23]	317,080	1,049,866	343	1
amazon0601 [11]	403,394	3,387,388	2752	1
Youtube [23]	1,134,890	2,987,624	28754	1
WikiTalk [23]	2,394,385	5,021,410	100032	1

Evaluation methods We use the sophisticated sampling methods including SRW, $MHRW$ and FS to evaluate $GVRW$. All the methods are implemented in C to sample large graph datasets described in Table 3.

- ***RWM*** is the usage of SRW (simple random walk) to estimate the graphs from the properties of nodes and groups. Simple random walk uses a strategy of traversing a large graph from one node to its random neighbor node [15, 19]. RWM does not change the key step of SRW while using the algorithm of finding the groups of the sampled nodes to obtain the group properties for estimating the node influences.
- ***MHM*** is a variant of MHRW (Metropolis-Hastings random walk). Similar to RWM, we modify metropolis-Hastings random walk to obtain the properties of both nodes and groups. MHRW transits a large graph from one node to its random neighbor with a random walker which has a probability of staying on the just sampled node to sample its neighbor again. During the sampling process, MHM uses Algorithm 1 to obtain the group properties of graphs for estimating the node influences.
- ***FS*** (Frontier sampling) is proposed to accurately estimate large graphs in face of disconnected or loosely connected components while exhibiting the strengths of regular random walks. Similar to RWM and MHM, FS also uses Algorithm 1 to obtain the group properties of sampling nodes for estimating the node influences.
- ***GVRW*** (Group-related-Vertex Random Walk) is our proposed sampling method in this paper to estimate properties of nodes and groups for evaluating the node influences of large graphs accurately. RWM, MHM, and FS are used as baselines for evaluating the efficiency of GVRW.

4.1 Node degree distribution

Node degree distribution is one of the most important characteristics in large graphs. We use GVRW to evaluate the node degree distributions over com-youtube and wiki-Talk and we compare it with RWM, MHM and FS which aim at obtaining node properties of large graphs. Figure 4(a) and (c) show that GVRW estimated the node degree distributions approximately to the real values on com-youtube and wiki-talk with budgets $B = (0.001|S|, 0.005|S|, 0.01|S|)$. Figure 4(b) and (d) show that GVRW can increase the estimation accuracies with the increase of the sampling budgets. Figure 5 presents that GVRW shows smaller biases than other methods with the same sampling budget $B = 0.01|S|$ in contrast to the existing three sampling methods of RWM, MHM, and FS. The experimental results confirm that GVRW is able to estimate the properties of nodes in large graphs.

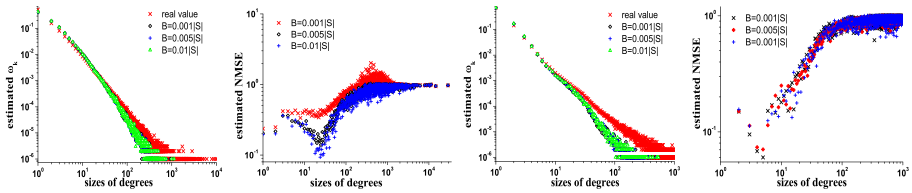

Figure 4 Estimated degree distribution and its NMSE with GVRW

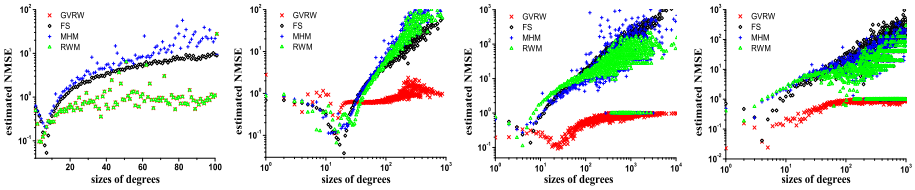

Figure 5 Estimated erroes of node degree distributions with different methods with B=0.01$|S|$

4.2 GRV distribution

We evaluate the performance of GVRW over com-dblp and amazon0601 to estimate the distributions of GRVs denoted by $\omega = (\omega_1, ..., \omega_k)$. Figure 6(a) and (c) show GVRW can estimate the distributions of GRVs accurately with the different sampling budgets labeled as $B = (0.001|S|, 0.005|S|, 0.01|S|)$ where $|S|$ denotes the number of items of the total sampling space. Figure 6(b) and (d) show that GCRW exhibits the smaller estimated NMSE with different budgets $B = (0.001|S|, 0.005|S|, 0.01|S|)$ and the larger number of samples obtained by GVRW has a smaller estimation error. We compare the methods of RWM, MHM and FS with GVRW showed in Figure 7 with the budget $B = 0.01|S|$. We find that in these four graphs, GVRW shows almost 10 times more accurate than the other three methods which present large estimated errors in estimating the GRV of graphs. From the Figure 7, we can infer that many nodes in these graphs take part in more than one group. If we know the properties of groups, we can easily obtain the differences between these groups. Thus, we can infer some nodes information which they can not offer in OSNs. These information can be used to many application such as recommending friends or products effectively.

4.3 Node clique size distribution

The cliques of nodes are the most important groups in graphs which can be used in wide applications of community detection and information spread. We employ GVRW to estimate the distributions of node cliques over two datasets: DBLP and WikiTalk. Figure 8(a) and (c) show the estimated distributions which are near to the real values with different sampling budgets as $B = (0.001|S|, 0.005|S|, 0.01|S|)$. Figure 8(b) and (d) describe the estimated NMSE over the two datasets with different sampling budgets. Since the estimated values are near to the real values and thus the values of estimated NMSE are smaller than one as shown in Figure 8(b) and (d). Figure 9 presents the comparisons of the four methods when estimating the distributions of node cliques with the same budget $B = 0.01|S|$. Figure 9 shows that GVRW is able to estimate the distributions of node cliques more accurately than RWM, MHM, and FS. This is because the sampling space of each sampling step of RWM,

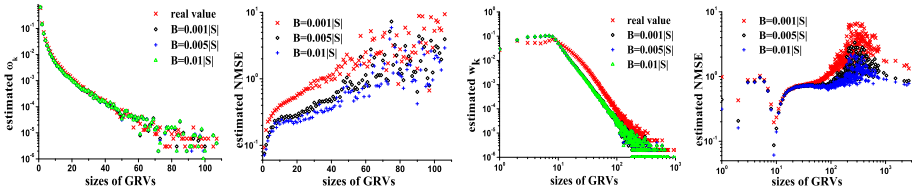

Figure 6 Estimated GRV distributions and the estimated errors with different sampling budgets

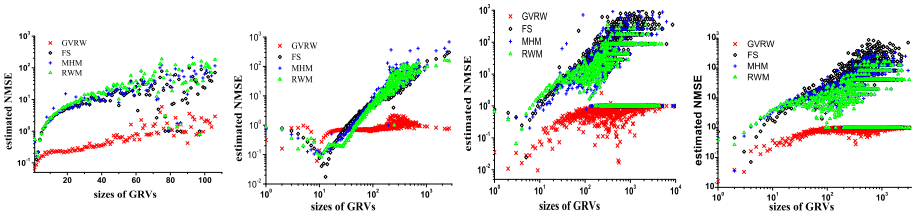

Figure 7 Estimated errors of the GRV distributions with different methods with B=0.01|S|

MHM, and FS is formed by the neighbors of one node, resulting in that the sampling processes have a large chance of trapping in the local subgraphs. Instead, GVRW enlarges the sampling space by considering the neighbors of nodes in the groups of a node as the possible selection of the next sample.

4.4 Sampling costs

To obtain the three distributions of a graph, a typical sampling process acquires two sampling sets along each sampling step, a node set containing potential samples to choose from in the next step based on the previously sampled node and an edge set containing node connection relationship of the node set. For the methods of GVRW, FS, MHW, and RWM, the sampling sets of each step can be saved in memory so that it is unnecessary to occupy the network bandwidth to collect them again when the sampled nodes are visited again. Meanwhile, the processing time for dealing with the repetitive samples can also be saved. Figure 10(a) and (b) show that GVRW is able to reduce the memory usage on DBLP and amazon0601. Computation time is mainly spent on obtaining the cliques during the sampling procedures. As it is shown in Figure 10(c) and (d), GVRW consumes a bit less processing time than the baseline sampling methods. This is because FS, RWM and MHM are biased to sample vertices with large degrees, meaning they should consume more time to discover the groups of these vertices and MHM should cost extra time to determine the next residence of the random walker.

4.5 Node influence estimation

As described in Section 2.2, the distribution of node degree is used to estimate the node influence from the property of nodes while the distributions of GRVs, degrees and node cliques to estimate the node influence of a large graph from the property of groups. Figure 5 confirms that GVRW is more capable of estimating the node influence more accurately than

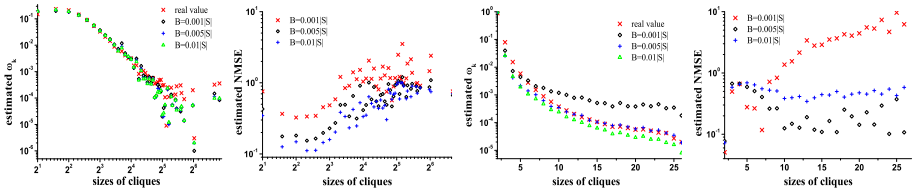

Figure 8 Estimated clique size distributions and the estimated NMSE with GVRW

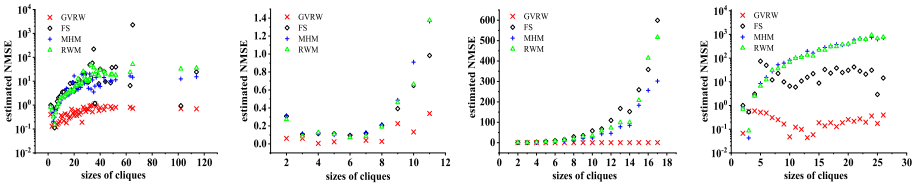

Figure 9 Estimated errors of the clique size distributions with different methods with B=0.01|S|

the existing methods. Then, the mean GRVs and clique sizes of nodes are used to infer the maximum influenced nodes through just a node from the perspective of its participated groups. For example, the mean GRV in DBLP is about 2.5 while the mean clique size is about 4.9. It means every author takes part in 2.5 groups averagely while the maximum researchers of the groups are 4.9 on average. Thus, the author registered in DBLP can influence about 12.6 authors.

We use IF-origin to denote the real values of node influences of graphs while employing IF-GS, IF-RW, IF-MH, and IF-FS denote the node influences estimated by the sampling methods of GVRW, RWM, MHM, and FS respectively. Figure 11 shows the estimated node influence through different methods about the four graphs with the same budget $B = 0.01|S|$. Because the sampling methods of RWM, MHM, and FS show great biases in estimating the GRV distribution and the node clique distribution so that they show large estimated errors in estimating the node influence. Figure 11 describes that the estimated node influences through GVRW approximate the real values in four datasets as GVRW is able to accurately estimate the properties of groups that nodes in graphs participate in. Therefore, GVRW can be used to estimate node influences from the perspectives of both nodes and groups.

5 Related work

Due to the huge volume of data in social networks, random walk based sampling techniques are popularly used for estimating the properties [14, 27, 30]. Besides the random walk based sampling methods described in Section 2, Li et al. [12] proposed a re-weighted random walk by redesigning the transition probability from one node to its neighbor. Xu et al. [22] proposed a skipping random walk by ignoring a number of nodes with small degrees which is benefit to accelerate the convergence of the sampling process. Zhang et al. [25, 26] proposed the random walk based on cliques rather than nodes to estimate the properties of the social networks. However, these existing random walk based sampling methods were designed to obtain the properties of nodes and can be used to estimate the node influences of large graphs from the perspective of groups.

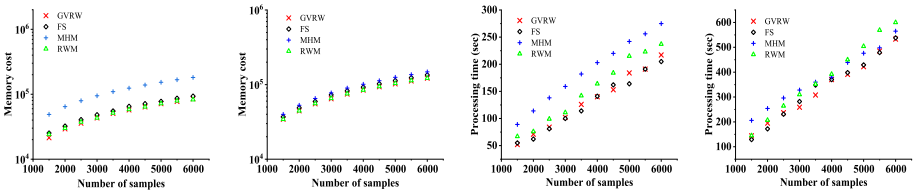

Figure 10 Sampling costs with different methods with B=0.01|S|

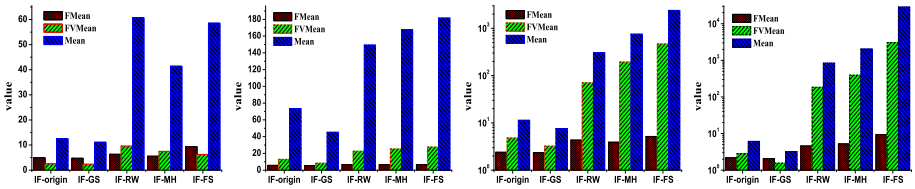

Figure 11 Node influences with different methods

On the other hand, many studies focus on node influences from the perspective of the influence maximization which aims at finding a static set of seed nodes to influence the most nodes over social networks. Different from using the sampling techniques to estimate the influences from the structural properties, the existing techniques on the influence maximization problem carefully designed the information transmission mode by analyzing the connectivity of nodes [9, 24]. Although Zhu et al. [30] employed the idea of a sampling technique to study the problem of influence maximization, they still employed the idea of the existing information transmission models while using the sampling technique to accelerate the discover of potential nodes for spreading information. This paper proposes group sampling to estimate the properties of both nodes and groups which can order the node influences in different social networks as described in Section 2.2. Our proposed method of estimating node influence from the perspective of groups is an important supplementary to the existing studies.

6 Conclusions

In this paper, we propose group sampling, called GVRW, to estimate the node and group properties for the sake of evaluating the node influences of large graphs. Specifically, we employ GVRW to accurately obtain the distributions of node degrees, GRVs, and node clique sizes to estimate the node influences of a large graph. Instead of traversing a large graph from one node to its neighbor node, we change the way of the random walker traversing from one node to one of the neighbors of the groups to a node. Furthermore, to improve the estimation accuracy, we carefully design an estimator for using the sampled nodes to estimate the properties required by computations on node influences. The experimental results on the four real-world datasets confirm the efficiency of our proposed methods.

Author Contributions Lingling Zhang: Conceptualization, Methodology, Software, Writing - original draft. Zhiping Shi: Conceptualization, Writing - original draft. Zhiwei Zhang and Ye Yuan: Supervision, Writing - review & editing. Guoren Wang: Writing - review & editing

Funding This work is supported by NSFC(Natural Science Foundation of China) 62302043.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

1. Abidi, A., Zhou, R., Chen, L., Liu, C.: Pivot-based maximal biclique enumeration. In: IJCAI, pp. 3558–3564 (2020)
2. Alspector, J., Kolcz, A., Karunanithi, N.: Comparing feature-based and clique-based user models for movie selection. In: Proceedings of the third ACM Conference on Digital Libraries, pp. 11–18. ACM (1998)
3. Chen, W., Wang, Y., Yang, S.: Efficient influence maximization in social networks. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 199–208 (2009)
4. Gjoka, M., Butts, C.T., Kurant, M., Markopoulou, A.: Multigraph sampling of online social networks. Sel. Areas Commun. **29**(9), 1893–1905 (2011)
5. Gjoka, M., Kurant, M., Butts, C.T., Markopoulou, A.: Walking in facebook: A case study of unbiased sampling of OSNs. In: INFOCOM, pp. 1–9. IEEE (2010)
6. Gjoka, M., Smith, E., Butts, C.: Estimating clique composition and size distributions from sampled network data. In: INFOCOM WKSHPS, pp. 837–842. IEEE (2014)
7. Guo, Q., Wang, S., Wei, Z., Chen, M.: Influence maximization revisited: Efficient reverse reachable set generation with bound tightened. In: Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data, pp. 2167–2181 (2020)
8. Han, M., Li, Y.: Influence analysis: A survey of the state-of-the-art. Math. Found. Comput. **1**(3), 201–253 (2018)
9. Huang, K., Tang, J., Xiao, X., Sun, A., Lim, A.: Efficient approximation algorithms for adaptive target profit maximization. In: 2020 IEEE 36th International Conference on Data Engineering (ICDE), pp. 649–660. IEEE (2020)
10. Lee, C.-H., Xu, X., Eun, D.Y.: Beyond random walk and metropolis-hastings samplers: Why you should not backtrack for unbiased graph sampling. In: SIGMETRICS, vol. 40, pp. 319–330. ACM (2012)
11. Leskovec, J., Adamic, L.A., Huberman, B.A.: The dynamics of viral marketing. ACM Trans. Web (TWEB) **1**(1), 5 (2007)
12. Li, R.-H., Yu, J.X., Qin, L., Mao, R., Jin, T.: On random walk based graph sampling. In: ICDE, pp. 927–938. IEEE (2015)
13. Lovász, L.: Random walks on graphs: A survey. Combinatorics, Paul Erdos is eighty, vol. 2(1), pp. 1–46 (1993)
14. Mo, S., Bao, Z., Zhang, P., Peng, Z.: Towards an efficient weighted random walk domination. Proc VLDB Endow **14**(4), 560–572 (2020)
15. Ribeiro, B., Towsley, D.: Estimating and sampling graphs with multidimensional random walks. In: SIGCOMM, pp. 390–403. ACM (2010)
16. Ribeiro, B., Wang, P., Murai, F., Towsley, D.: Sampling directed graphs with random walks. In: INFOCOM, pp. 1692–1700. IEEE (2012)
17. Strogatz, S.H.: Exploring complex networks. Nature **410**(6825), 268–276 (2001)
18. Wang, P., et al.: Efficiently estimating motif statistics of large networks. ACM Trans. Knowl. Discov. Data (TKDD) **9**(2), 8 (2014)
19. Wang, P., Ribeiro, B., Zhao, J., Lui, J., Towsley, D., Guan, X.: Practical characterization of large networks using neighborhood information. arXiv:1311.3037 (2013)
20. Wasserman, S., Faust, K.: Social network analysis: Methods and applications, vol. 8. Cambridge University Press (1994)
21. Xie, H., Yi, P., Li, Y., Lui, J.C.: Optimizing random walk based statistical estimation over graphs via bootstrapping. IEEE Trans. Knowl. Data Eng. (2021)
22. Xu, X., Lee, C.-H., et al.: Challenging the limits: Sampling online social networks with cost constraints. In: INFOCOM (2017)
23. Yang, J., Leskovec, J.: Defining and evaluating network communities based on ground-truth. Knowl. Inf. Syst. **42**(1), 181–213 (2015)
24. Zareie, A., Sakellariou, R.: Influence maximization in social networks: A survey of behaviour-aware methods. Soc. Netw. Anal. Min. **13**(1), 78 (2023)
25. Zhang, L., Jiang, H., Wang, F., Feng, D.: Draws: A dual random-walk based sampling method to efficiently estimate distributions of degree and clique size over social networks. Knowl.-Based Syst. **198**, 105891 (2020)
26. Zhang, L., Wang, F., Jiang, H., Feng, D., Xie, Y., Zhang, Z., Wang, G.: Random walk on node cliques for high-quality samples to estimate large graphs with high accuracies and low costs. Knowl. Inf. Syst. **64**(7), 1909–1935 (2022)

27. Zhang, L., Zhang, Z., Wang, G., Yuan, Y.: Efficiently sampling and estimating hypergraphs by hybrid random walk. In: 2023 IEEE 39th International Conference on Data Engineering (ICDE), pp. 1273–1285. IEEE (2023)
28. Zhang, Y., Li, Y., Bao, Z., Zheng, B., Jagadish, H.: Minimizing the regret of an influence provider. In: Proceedings of the 2021 International Conference on Management of Data, pp. 2115–2127 (2021)
29. Zhou, C., Zhang, P., Zang, W., Guo, L.: Maximizing the cumulative influence through a social network when repeat activation exists. Procedia Comput. Sci. **29**, 422–431 (2014)
30. Zhu, Y., Tang, J., Tang, X., Wang, S., Lim, A.: 2-hop+ sampling: Efficient and effective influence estimation. IEEE Trans. Knowl. Data Eng. (2021)