QMSampler: Joint Sampling of Multiple Networks with Quality Guarantee

Hong-Han Shuai  
Graduate Institute of Communication Engineering  
National Taiwan University  
d99942020@ntu.edu.tw

De-Nian Yang  
Institute of Information Science  
Academia Sinica  
dnyang@iis.sinica.edu.tw

Chih-Ya Shen  
Research Center for Info. Technology Innovation  
Academia Sinica  
chihya@citi.sinica.edu.tw

Philip S. Yu  
Department of Computer Science  
University of Illinois at Chicago  
psyu@cs.uic.edu

Ming-Syan Chen  
Department of Electrical Engineering  
National Taiwan University  
mschen@cc.ee.ntu.edu.tw

ABSTRACT

Because Online Social Networks (OSNs) have become increasingly important in the last decade, they have motivated a great deal of research on social network analysis (SNA). Currently, SNA algorithms are evaluated on real datasets obtained from large-scale OSNs, which are usually sampled by Breadth-First-Search (BFS), Random Walk (RW), or some variations of the latter. However, none of the released datasets provides any statistical guarantees on the difference between the crawled datasets and the ground truth. Moreover, all existing sampling algorithms only focus on crawling a single OSN, but each OSN is actually a sampling of a global offline social network. Hence, even if the whole dataset from a single OSN is crawled, the results may still be skewed and may not fully reflect the properties of the global offline social network. To address the above issues, we have made the first attempt to explore the joint sampling of multiple OSNs and propose an approach called Quality-guaranteed Multi-network Sampler (QMSampler) that can crawl and jointly sample multiple OSNs. QMSampler provides a statistical guarantee on the difference between the crawled real dataset and the ground truth (the perfect integration of all OSNs). Our experimental results demonstrate that the proposed approach generates a much smaller bias than any existing method. QMSampler has also been released as a free download.

1. INTRODUCTION

Online Social Networks (OSNs), such as Facebook, Twitter and Foursquare, have become increasingly important in the last decade and are now an integral part of many people’s everyday lives. Because OSNs contain abundant and varied information, they have generated a considerable amount of research on social network analysis (SNA). Currently, SNA algorithms and techniques are evaluated on real datasets obtained from large-scale real OSNs, which are usually crawled and sampled by Breadth-First-Search (BFS), Random Walk (RW), or uniform sampling. However, it has been shown that BFS and RW tend to introduce a bias toward high degree nodes [7] because such nodes, together with their neighbors, are more likely to be sampled. On the other hand, since no user IDs are provided by an OSN, uniform sampling by testing if a random number corresponding to a user ID [14] tends to involve a large number of trials [13] when the number of users is much smaller than the length of the user ID. For instance, Twitter and Facebook both use 64-bit user IDs which can represent $2^{64}$ users, but the number of users is much smaller (271 million $\approx 2^{28}$ users on Twitter[1] and 1.32 billion $\approx 2^{30}$ users on Facebook[2]). Therefore, a recent line of studies proposes effective schemes to approximate uniform sampling [13][23]. However, none of the released datasets provides any statistical guarantee and analysis of the difference between the crawled dataset (i.e., the social graph with only a small subset of sampled nodes and edges) and the ground truth (i.e., the social graph with all nodes and edges). As a result, it is difficult for researchers to evaluate the quality of existing real datasets.

In fact, an OSN can be considered as a sampling of the global offline social network because the friends of each person in the OSN are only a subset of that person’s friends in the world. Even if an SNA algorithm crawls all the nodes in a single OSN, the results may still be skewed, and may not fully reflect the properties of the global offline social network due to the skewed sampling provided by a single OSN. For example, the social influence model, which is widely used in viral marketing, advertisement targeting, and information diffusion, may not be precise if the set of edges incident to each node is incomplete as there is a tendency to underestimate the node’s activation probability of the social influence. Fortunately, after evolving for more than a decade, there are now several OSNs that meet the numerous demands of people in their everyday lives. Most people own multiple accounts in various OSNs, and each account may be used to contact different friends, such as company colleagues in LinkedIn and the friends with the same hobby in Pinterest. Therefore, it is possible that if multiple OSNs are crawled simultaneously and the crawled real datasets from different OSNs can be merged or integrated as a multi-layer network [12], the difference between the merged dataset and the global offline social network could be effectively reduced because the set of friends for each person would be more complete.

Table 1 presents a user study (detailed in Section 5) to compare the average degrees (AVGDEG) and local clustering coefficients (Local CC) for offline and online social networks of 484 people. Notice that the offline social network consists of only the 484 students, i.e., the friend of a user is not taken into the offline social network if she is not one of the 484 students, because the offline friends of other users cannot be acquired in the user study, and the edges in the offline social network include all online friends and the additional friends specified by each user. The average degree and clustering coefficient in the merged network of Facebook and
LinkedIn are close to those in the offline social network. By contrast, the average degree of the offline social network is higher than those in the two OSNs because many friends appear in only one of the two OSNs. Moreover, the clustering coefficient of the offline social network is much smaller than the online counterparts, indicating that current OSNs may overestimate the number of tightly knit groups.

In this paper, therefore, we propose a new framework for jointly sampling multiple OSNs. Our first goal is to provide statistical guarantees on the difference between the crawled (and then merged or integrated) real datasets and the ground truth. The ground truth in this paper is defined as the perfect integration of all the OSNs considered. That is, all the nodes and edges in the OSNs are included, and the nodes corresponding to the same person in different OSNs are correctly merged and integrated. The difference is the gap between the graph characteristic metrics of the sampled graph and those of the ground truth. Not surprisingly, for a single OSN, the difference increases (i.e., the quality of the crawled datasets degrades) when (1) a biased sampling strategy is employed; or (2) the number of nodes crawled is insufficient. It is more challenging to sample multiple networks simultaneously because both oversampling and undersampling of an OSN will degrade the results. In addition, the number of overlapping nodes in multiple OSNs is also a crucial factor when we try to ensure the quality of the output graph. In the following, we discuss the research challenges addressed in this paper and the opportunities they present.

Social Network Sampling with Quality Analysis. Compared with uniform sampling, BFS and RW are inclined to be biased toward high-degree nodes [13], and thus it appears that different sampling strategies require different numbers of nodes to meet the same quality. It is important to evaluate the quality of a sampled dataset based on the number of nodes crawled, especially when the dataset is large. Nevertheless, none of the current crawling schemes provides a statistical analysis of the difference between the sampled dataset and the OSN. It is, as a consequence, difficult for a researcher to ascertain how many nodes are sufficient to sample the OSN to meet the research requirements.

Joint Sampling of Multiple Networks. Although crawling more nodes would certainly increase the quality for a single network, during the sampling of multiple networks, the bias is correlated with the number of nodes crawled from each network. If an OSN is oversampled (i.e., if too many nodes are sampled), the graph may become more similar to the OSN and thus more skewed because each OSN is only a skewed sample of the ground truth. Multi-network sampling is therefore challenging because both undersampling and oversampling are inappropriate. Moreover, when merging the sampled networks, it is necessary to identify the same users in different social networks. Prior studies [10, 17, 18, 50] have achieved good results by utilizing link prediction approaches to analyze the similarity of users in different social networks, while machine learning techniques have also been employed to match the accounts of the same user in different social networks [13]. However, account matching errors may still appear; and an algorithm with a low matching accuracy will incur more biases when the number of crawled nodes increases since more accounts are not correctly matched, which biases the output graph. On the other hand, the number of overlapping nodes in multiple OSNs also has a strong impact on the quality of the output graph, as can be seen in the example of two networks having only a smaller number of overlapping nodes, where the quality of the output graph has a greater chance to be accepted even if no sophisticated matching algorithm is exploited.

Based on the above observations, given a set of OSNs, the matching accuracy, and a difference threshold, this paper aims to find the number of nodes to be crawled from each OSN. Our objective is to maximize the sampling quality, such that the probability that the difference between the crawled dataset and the ground truth is less than the difference threshold specified. Here, the difference means the difference between the sample mean and the mean of the ground truth of any graph characteristics after sampling $n$ nodes. Fundamental graph characteristic metrics include node attributes, such as the interest level and age. The metrics associated with edges, such as the node degrees and clustering coefficients, are also crucial. Since we notice that even when the number of nodes crawled from each network is given, deriving the above sampling quality directly is still very challenging, we examine the following three problems step-by-step. 1) Non-overlap sampling. First, we consider an impractical but simple case where every node in each sampled network belongs to a different person. In other words, each person has only one account, and it is thus not necessary to match the accounts across two social networks. We begin with this problem to determine the relation between the sampling quality and the number of nodes crawled from each OSN in order to identify oversampling and undersampling instances during multi-network crawling. 2) Overlap sampling with an accurate matching oracle. In this case, we assume that the networks overlap for some people, but there exists an accurate matching oracle that can identify all the accounts belonging to the same person. Nevertheless, even with the oracle, bias will occur when only a subset of accounts for one person is crawled. In other words, the friend list of the person is still incomplete in this case. In this problem, we try to find the correlation between the sampling quality and the proportion of nodes that overlap in multiple networks. 3) Overlap sampling with a practical matching oracle. This case explores the most generalized scenario and incorporates the matching accuracy of an existing matching oracle, such as [10, 17, 18, 50] into the analysis. Our objective is to determine how matching errors degrade the sampling quality and to accordingly adjust the number of nodes to be crawled from different networks. In each of the above cases, we examine two OSNs initially, and then extend the results to more OSNs.

Equipped with the above analytical results, it is now possible to determine the number of nodes to be sampled from each OSN in order to maximize the sampling quality. To this end, we propose a new crawler, called the Quality-guaranteed Multi-network Sampler (QMSampler), which can crawl and jointly integrate multiple OSNs. QMSampler is designed to support the following crawling scenarios. 1) Size-constrained sampling. The user specifies the difference threshold and the total number of nodes to be crawled from the target OSNs. QMSampler then outputs a graph with maximal quality by properly assigning the number of nodes to be sampled from each OSN. The sampling quality is returned as well. 2) Quality-constrained sampling. The user specifies the minimum required quality. QMSampler then crawls the minimal number of nodes and outputs a graph. 3) Time-constrained sampling. In many cases, the user prefers to generate a massive output graph and tends to crawl the OSNs until the deadline. However, because OSNs have different access policies and download bandwidths, some OSNs may be oversampled and yield biased results. To address this issue, QMSampler trims the oversampled networks according to our analysis above, and then generates a graph with the maximal number of nodes. It also outputs the quality to describe the property of the OSNs.
output dataset.

This paper makes several contributions.

• Most social network algorithms, e.g., community detection, and anomaly detection, are evaluated against crawled social network data. Improperly crawled datasets can lose or distort the original network characteristics and lead to inaccurate conclusions. This is especially the case when multiple networks are involved. It is thus critical to provide (1) robust metrics to evaluate the faithfulness of the crawled network as compared to the original network, and (2) an effective crawling method that can preserve the network characteristics. Here we introduce the concept of sampling quality and provide a sampling strategy that makes the crawled network dataset better maintain the original network characteristics.

• To the best of our knowledge, no performance metrics have ever been proposed to evaluate the quality of crawled datasets. To address this important issue, we define the concept of sampling quality, which is the probability that the difference between the crawled dataset and the ground truth is less than a user-specified difference threshold. We also analyze the sampling quality with different numbers of crawled nodes for a single OSN.

• All existing samplers are designed to crawl a single OSN. Because it is envisaged that the integration of a number of OSNs will be more similar to an offline global social network, in this paper, we make the first attempt to explore the joint sampling of multiple OSNs. We discover that both undersampling and oversampling of one of the OSNs undermine the overall sampling quality. In addition, the sampling quality is closely related to the proportion of overlapping nodes in different networks and the account matching error probability. All the above crucial factors are quantified in our derivation of the sampling quality for multiple OSNs.

• The proposed QMSampler is the first approach that jointly samples multiple OSNs. It supports three crawling scenarios: size-constrained sampling, quality-constrained sampling, and time-constrained sampling. In addition, QMSampler has been released as a free download.

The remainder of this paper is organized as follows. In Section 2, we review the related works on sampling strategies and account matching algorithms; and in Section 3 we analyze and derive the sampling quality in different scenarios. In Section 4 we describe the design of QMSampler based on the above theoretical analysis. In Section 5 we present the experimental results, and then in Section 6 offer our conclusions and directions for future work.

2. RELATED WORK

Network sampling is one of the cornerstones to provide real datasets for the analysis of social, information and biological networks due to the emergence of massive data in a variety of applications. Breadth-First-Search (BFS) and Random Walk (RW) (and its variations) have been adopted by many crawlers to generate released datasets for OSNs. Web, P2P networks, and other types of graphs. However, BFS and RW tend to introduce a bias to the scenario with multiple OSNs in Section 4.5. Moreover, we derive the sampling quality of graph characteristics from multiple OSNs. Fundamental graph metrics include node attributes, such as the interest level, age, and post number. The metrics associated with edges, such as the node degree, will also be analyzed in this paper. The analysis of clustering coefficients can be extended from the analysis of node degrees. In this paper, high sampling quality implies a small difference between the graph characteristics of the crawled dataset and those of the ground truth represented by the perfect integration of all the OSNs considered, where all the nodes and edges in the OSNs are included, and the nodes corresponding to the same person in different OSNs are correctly merged. To systematically solve the problem, we consider fundamental cases with two OSNs and then generalize our findings to the scenario with multiple OSNs in Section 4.5. Moreover, we derive the sampling quality of US first and then extend it to RW.

3. SAMPLING QUALITY

In the following, we derive the sampling quality of graph characteristics from multiple OSNs. Fundamental graph metrics include node attributes, such as the interest level, age, and post number. The metrics associated with edges, such as the node degree, will also be analyzed in this paper. The analysis of clustering coefficients can be extended from the analysis of node degrees. In this paper, high sampling quality implies a small difference between the graph characteristics of the crawled dataset and those of the ground truth represented by the perfect integration of all the OSNs considered, where all the nodes and edges in the OSNs are included, and the nodes corresponding to the same person in different OSNs are correctly merged. To systematically solve the problem, we consider fundamental cases with two OSNs and then generalize our findings to the scenario with multiple OSNs in Section 4.5. Moreover, we derive the sampling quality of US first and then extend it to RW.

3.1 Non-Overlap Sampling

Table 2 summarizes the notations used in this paper. Let \( \{a_1, a_2, \ldots, a_1, \ldots, a_n\} \) denote a sequence of \( n \) random samples of node attributes in the two networks, where \( n(N_{G_1}) \) and \( n(N_{G_2}) \) samples with \( n(G_1) + n(G_2) = n \) are drawn from \( G_1 \) and \( G_2 \), respectively. Let \( G \) denote the union of \( G_1 \) and \( G_2 \), i.e., \( G_1 \cup G_2 \), with size \( N \). Moreover, let \( N(G_1) \) and \( N(G_2) \) denote the size of network \( G_1 \) and \( G_2 \), respectively. The samples are represented by iid random variables with mean \( \mu_{A}(G) \) and variance \( \sigma^2_{A}(G) < \infty \). Here, \( \mu_{A}(G) \) represents the mean of the node attributes of all nodes in network \( G \), while \( \mu_{A}(G_1) \) and \( \mu_{A}(G_2) \) are the means in \( G_1 \) and \( G_2 \), respectively. Similarly, let \( \sigma_{A}(G_1) \) and \( \sigma_{A}(G_2) \) denote the standard deviations of the node attributes in \( G_1 \) and \( G_2 \), respectively. Moreover, let \( \bar{a}_n \) denote the sample mean of the node attributes of the \( n \) samples, i.e., \( \bar{a}_n = \frac{1}{n} \sum_{i=1}^{n} a_i \). Let \( Q_{\epsilon}^2 \) denote the probability that the difference between \( \bar{a}_n \) and \( \mu_{A}(G) \) is less than a threshold \( \epsilon \). The study of different quality settings, i.e., \( \epsilon \) and \( Q_{\epsilon}^2 \), will be discussed in Section 5.

Theorem 1. Given two non-overlapping networks \( G_1 \) with \( N(G_1) \) nodes and \( G_2 \) with \( N(G_2) \) nodes, together with \( n(G_1) \) and \( n(G_2) \) samples drawn from \( G_1 \) and \( G_2 \) respectively, \( Q_{\epsilon}^2 \) is not smaller than

\[
1 - \sum_{j=1}^{2} \frac{N(G_j) \sigma_{A}(G_j)}{\epsilon N(n(G_j))} \left(1 - \frac{n(N_{G_j}) - 1}{N(G_j) - 1}\right).
\]

Proof. Because the two networks do not overlap, each network can be regarded as a stratum. Therefore, we prove the theorem based on stratified random sampling (the main result is shown in Equation 3 of (24)). The variance of the sample mean \( V[\bar{a}_n] \) is
Theorem 2. $Q_d^D$ is no smaller than
\[ 1 - \left( \frac{N(G_i)}{N} \right) \frac{\sigma_{G_i}(\bar{\Delta}_i)}{\mu_{G_i}(\bar{\Delta}_i)} \left( \frac{N(G_i) - 1}{N(G_i) - 1} \right) \left[ \bar{\rho}(\bar{\Delta}_i) \right] + \left( \frac{n(G_i)}{N} \right) \frac{\sigma_{G_i}(\bar{\Delta}_i)}{\mu_{G_i}(\bar{\Delta}_i)} \left( \frac{N(G_i) - 1}{N(G_i) - 1} \right) \left[ \bar{\rho}(\bar{\Delta}_i) \right] + \mu^D(\bar{\Delta}_i) \left[ \bar{\rho}(\bar{\Delta}_i) \right], \]
where $\bar{\rho}(\bar{\Delta}_i)$ is the mean degree ratio of the $n(G_i)$ samples from network $G_i$.

Proof. For uniform sampling, $\rho_i$ and $\alpha_i$ are independent since the event that a node $v_i$ is sampled is independent to the event that any neighbors of $v_i$ are sampled. A random variable, denoted as $\bar{\rho}_i$, is defined as the average degree of the $n$ nodes in $G_i$, $\rho_i$. Moreover, as $\bar{\rho}_i$ is a random variable with mean $\mu_{G_i}(\bar{\Delta}_i)$ and variance $\sigma_{G_i}^2(\bar{\Delta}_i) < \infty$, we apply Chebyshev’s inequality on $\bar{\rho}_i$. For any real number $k > 0$, the probability
\[ P(|\bar{\rho}_i - \mu_{G_i}(\bar{\Delta}_i)| \geq k\sigma_{G_i}(\bar{\Delta}_i)) \leq \frac{1}{k^2}. \]

After setting $k$ as $\epsilon / \sigma_{G_i}(\bar{\Delta}_i)$ and replacing $\sigma_{G_i}^2(\bar{\Delta}_i)$ with $\sigma_{G_i}^2(\bar{\Delta}_i)$, we have
\[ P(|\bar{\rho}_i - \mu_{G_i}(\bar{\Delta}_i)| \geq \epsilon) \leq \frac{\sqrt{\bar{\rho}_i}}{\epsilon^2}. \]

Therefore, $P(|\bar{\rho}_i - \mu_{G_i}(\bar{\Delta}_i)| \leq \epsilon)$ is no smaller than
\[ 1 - \sum_{j=1}^{\frac{n(G_i)}{N}} \left( \frac{N(G_i)}{n(G_i)} \right)^2 \left( \frac{N(G_i) - 1}{N(G_i) - 1} \right), \]
Because $\mathbb{E}[\overline{v}_n^2] = \mathbb{E}[\overline{a}_n] + \mathbb{E}[\overline{a}_n]^2$, the above equation can be simplified as
\[
\overline{v}_n = (\mathbb{V}[\overline{a}_n] + \mathbb{E}[\overline{a}_n])^2(\mathbb{V}[\overline{a}_n] + \mathbb{E}[\overline{a}_n]) - (\mathbb{E}[\overline{a}_n])^2
\]
\[
= \mathbb{V}[\overline{a}_n]\mathbb{V}[\overline{a}_n] + 2\mathbb{V}[\overline{a}_n]\mathbb{E}[\overline{a}_n] + \mathbb{E}[\overline{a}_n]^2\mathbb{V}[\overline{a}_n].
\]
Let $D$ denote the graph density (i.e., $\frac{2\|E\|}{|V|(|V|-1)}$), which is assumed to be the same as $n$ increases for simplicity. To sample $n(G_j)$ nodes from network $G_j$ with size $N(G_j)$, the expected value of degree ratio between $\overline{d}_n(G_j)$ and $\overline{a}_n(G_j)$, i.e., $\mathbb{E}[\overline{a}_n(G_j)]$ is
\[
\frac{\overline{d}_n(G_j)}{\overline{a}_n(G_j)} = \frac{2\overline{d}_n(G_j)}{n(G_j)} = \frac{\overline{d}_n(G_j)\mathbb{E}[\overline{a}_n(G_j)]}{n(G_j)} = \frac{n(G_j) - 1}{N(G_j) - 1}.
\]
Therefore, we apply Chebyshev’s inequality on $\overline{d}_n$, and thus $P(|\overline{d}_n - \overline{a}_n(G_j)|/\overline{a}_n(G_j) \leq \epsilon)$ is no smaller than
\[
1 - \frac{\sum_{j=1}^{2N(G_j)} \mathbb{E}[\overline{a}_n(G_j)] - n(G_j)}{n(G_j)} = \left(1 - \frac{n(G_j) - 1}{N(G_j) - 1}\right)^2 + \left(\frac{n(G_j) - 1}{N(G_j) - 1}\right)^2.
\]
The theorem follows. \hfill \Box

The variance of $\overline{a}_n(G_j)$ is much smaller than $\overline{a}_n(G_j)$ here because $\overline{a}_n(G_j)$, i.e., set to zero. Equation (4) can be simplified as follows:
\[
1 - \frac{\sum_{j=1}^{2N(G_j)} \mathbb{E}[\overline{a}_n(G_j)] - n(G_j)}{n(G_j)} = \left(1 - \frac{n(G_j) - 1}{N(G_j) - 1}\right)^2 + \left(\frac{n(G_j) - 1}{N(G_j) - 1}\right)^2.
\]
To find the sampling quality of the clustering coefficients, let $\{c_1, c_2, \ldots, c_n\}$ denote a sequence of $n$ random samples of clustering coefficients in $G_{QM}$. Moreover, let $\mu(G)$ and $\sigma^2(G) < \infty$ be the mean and the variance of the clustering coefficients of $G$, respectively. Let $Q^2_c$ probability that the difference between $c_n$ and $\mu(G)$ is less than a threshold $\epsilon$. To sample $n(G_j)$ nodes from network $G_j$ with size $N(G_j)$, $\overline{a}_n(G_j)$ denotes the probability that two nodes are connected to one another in a network (i.e., $G_{QM}$) sampled from $G_j$ with size $n(G_j)$ and average degree $\mu(G)$.

**Lemma 1.** Given a network $G_j$ with $N(G_j)$ nodes, together with $n(G_j)$ samples drawn from $G_j$, $\overline{a}_n(G_j)$ equals $N(G_j)$.

**Proof.** From the theoretical results of Theorem 2
\[
P(\overline{a}_n(G_j) \geq \mu(G)) = \frac{n(G_j) - 1}{n(G_j)} \mu(G) = \frac{n(G_j) - 1}{n(G_j)} = \frac{n(G_j) - 1}{n(G_j) - 1}.
\]
The lemma follows. \hfill \Box

**Corollary 3.** $Q^2_c$ is no smaller than $1 - \sum_{j=1}^{2N(G_j)} \mathbb{E}[\overline{a}_n(G_j)]^2 / N(G_j) - 1$.

**Proof.** Let $|N(v_i)|$ and $|\overline{a}_N(v_i)|$ denote the number of edges of node $v_i$ respectively. From Lemma 4, the expected values of the clustering coefficient of node $v_i$ in a network with size $N(G_j)$ and $n(G_j)$ are the same, i.e.,
\[
P(\overline{a}_n(G_j) \geq \mu(G)) = \frac{n(G_j) - 1}{n(G_j) - 1} = \frac{n(G_j) - 1}{n(G_j) - 1} = \frac{n(G_j) - 1}{n(G_j) - 1}.
\]
The corollary follows. \hfill \Box

In the following, we explore a more practical case where multiple accounts in different OSNs belong to the same person.

### 3.2 Overlap Sampling with An Accurate Matching Oracle

The following theorem first analyzes the sampling quality of node attributes with an accurate matching oracle, i.e., the oracle can correctly match every two nodes from the same person in two OSNs. Two nodes in the overlapping network $G_1 \cap G_2$ are connected by an edge if the two nodes are neighbors in $G_1$ or $G_2$.

**Definition 1.** Given two networks $G_1$ and $G_2$, the difference graph is denoted as $G_1 \setminus G_2 = (V, E)$, where $V$ and $E$ contain the nodes and edges appearing in $G_1$ that are not in $G_2$, respectively.

In contrast to the scenario in Section 3.1, here $G_1$ overlaps with $G_2$. Although $N(G_1)$ and $N(G_2)$ can be estimated according to the corresponding released information in official websites, it is difficult to find out $N(G_1 \cap G_2)$. For instance, the numbers of accounts in Facebook and Twitter can be estimated by the statistics announced in their websites, but the number of matched accounts between Facebook and Twitter is unknown. A possible way is to regard $G_1 \cap G_2$, $G_2 \cap G_1$, and $G_1 \cap G_2$ as three strata and extend Equation (8) accordingly. Nevertheless, $N(G_1 \cap G_2)$, $N(G_1 \cap G_2)$, and $N(G_1 \cap G_2)$ for Equation (8) are not available here. Therefore, here one challenge for calculating the sampling quality is to correctly estimate $N(G_1 \cap G_2)$ since $N(G_1 \cap G_2)$ and $N(G_2 \cap G_1)$ can be obtained by $N(G_1) - N(G_1 \cap G_2)$ and $N(G_2) - N(G_2 \cap G_1)$ respectively.

If the ratio of $N(G_1 \cap G_2)/N$ is large, fewer samples are required for holding the solution quality as compared with the problem of non-overlap sampling because of the following reasons. (1) Fewer samples are required to find the correct mean for the network with a smaller size, and thus fewer samples are required for a graph with the size $N(G_1 \cap G_2)$ than the two non-overlap graphs with the total size $N(G_1) + N(G_2)$. (2) The node attributes of a node in $G_1 \cap G_2$ can be acquired by sampling either $G_1$ or $G_2$, i.e., the node attributes are the same in the two networks; otherwise, it is difficult for the two nodes to be matched. Therefore, if $G_1$ and $G_2$ are fully overlapped, it is necessary to sample either $G_1$ or $G_2$.

Nevertheless, it is more challenging to calculate the sampling quality from the perspective of node degrees because the node degree of a node $v_i$ in $G_1 \cap G_2$ changes if the corresponding node from another OSN is not sampled, i.e., the neighbors of $v_i$ in either $G_1$ or $G_2$ are absent. Therefore, the additional sampling error occurs as compared to the case of node attributes. The selection of $n(G_1)$ and $n(G_2)$ is more important in overlap sampling for preserving node degrees from the following two perspectives. (1) The ratios of $N(G_1 \cap G_2)$ to $N(G_1)$ and $N(G_2)$. For instance, if $N(G_1) \approx N(G_1 \cap G_2)$ and we set $n(G_1)$ very close to $n$, the sampling quality will be poor since the corresponding nodes in $G_2$ are not able to be sampled as well. (2) The difference between $|\mu(G_1 \cap G_2) - \mu(G_2)|$ and $|\mu(G_1) - \mu(G_2)|$. For instance, if $|\mu(G_1 \cap G_2) - \mu(G_2)|$ is larger than $|\mu(G_1) - \mu(G_2)|$, a larger error will be introduced when a sample in $G_1 \cap G_2$ drawn from $G_1$ is not matched with the corresponding node in $G_2$ (i.e., the expected value of error is $|\mu(G_1 \cap G_2) - \mu(G_1)|$).

Therefore, it is necessary to sample more nodes from $G_2$. Since $N(G_1 \cap G_2)$ plays an important role in determining the sampling quality, in the following, we show how to estimate $N(G_1 \cap G_2)$. Given two networks $G_1$ with $N(G_1)$ nodes and $G_2$ with $N(G_2)$ nodes, together with $n(G_1)$ and $n(G_2)$ samples drawn from $G_1$ and $G_2$, we only know that how many nodes are matched after sampling. Therefore, we use the number of matched nodes after sampling to estimate $N(G_1 \cap G_2)$. Let $\mathbb{E}[\overline{m}]$ denote the expected value of the number of matches in $G_1 \cap G_2$.
Given two networks $G_1$ and $G_2$ with $N(G_1)$ and $N(G_2)$ nodes, respectively, and $n(G_1)$ and $n(G_2)$ samples drawn from $G_1$ and $G_2$ respectively, $Q^N$ is no smaller than
\[
1 - ((\frac{N(G_1 \setminus G_2)}{eN})(\frac{\sigma_1(G_1 \setminus G_2)}{n(G_1 \setminus G_2)})^2(1 - \frac{n(G_1 \setminus G_2) - 1}{N(G_1 \setminus G_2) - 1}) + (\frac{N(G_2 \setminus G_1)}{eN})(\frac{\sigma_2(G_2 \setminus G_1)}{n(G_2 \setminus G_1)})^2(1 - \frac{n(G_2 \setminus G_1) - 1}{N(G_2 \setminus G_1) - 1}) + (\frac{N(G_1 \setminus G_2)}{eN})(\frac{\sigma_1(G_1 \setminus G_2)}{n(G_1 \setminus G_2)})^2(1 - \frac{n(G_1 \setminus G_2) - 1}{N(G_1 \setminus G_2) - 1})),
\]
where $\mathbb{E}(n(G_1 \setminus G_2))$, $\mathbb{E}(n(G_2 \setminus G_1))$, and $\mathbb{E}(n(G_1 \cap G_2))$ can be calculated by Lemma 2.

Proof: As $G_1 \setminus G_2$, $G_2 \setminus G_1$, and $G_1 \cap G_2$ are regarded as three strata, we can directly derive Equation 4 according to Theorem 1. Therefore, the expected number of samples in each strata can be calculated by Lemma 2. The theorem follows.

In the following, we analyze the sampling quality with an accurate matching oracle from the perspective of node degrees. Here, we model the sample mean of the node degrees in $G_1 \setminus G_2$ by letting $d_i = \rho_i(a_i - \eta_i)$ be a random sample of the node degrees, where $a_i$ is the node degree of the $i$-th sample in $G_1 \cup G_2$, and $\eta_i$ represents the sampling error of $a_i$. Here, the sampling error $\eta_i$ is the decrement of the node degree when node $v_i$ in $G_1 \cap G_2$ is sampled from only $G_1$ or $G_2$. The node degree will decrease to $a_i - \eta_i$, and the sampled node degree of $v_i$ will become $\rho_i(a_i - \eta_i)$, where $\rho_i$ is the degree ratio, as mentioned in Theorem 2. For example, for a node $v_i$ in $G_1 \cap G_2$ with 5 neighbors in $G_1$ and 6 neighbors in $G_2$, if $a_i$ is only sampled in $G_1$ with 3 neighbors and not sampled in $G_2$, $\eta_i$ is 6 and $\rho_i$ equals to 3/5 in this case, whereas $a_i$ is 11. Let $d_{\bar{a}}$ denote the average node degree of the $n$ nodes in $G_{QM}$ such that $d_{\bar{a}} = \frac{1}{n} \sum_{i=1}^{n} \rho_i(a_i - \eta_i) = \frac{1}{n} \sum_{i=1}^{n} \rho_i(a_i) - n \eta_i$. Because $a_i$ and $\eta_i$ are independent, $\text{Cov}(a_i, \eta_i)$ is zero, which implies that the variance of the average node degree of the $n$ nodes in $G_{QM}$, i.e., $\mathbb{V}[d_{\bar{a}}]$, is
\[
\mathbb{V}[d_{\bar{a}}] = (\mathbb{V}[a_i]\mathbb{E}[\rho_i])^2 + (\mathbb{V}[\eta_i]\mathbb{E}[\rho_i])^2.
\]
Therefore, $\mathbb{V}[d_{\bar{a}}]$ can be acquired from $\mathbb{V}[a_i]$, $\mathbb{V}[\eta_i]$, and $\mathbb{E}[\rho_i]$. Let $\sigma_2^2(G)$ and $\sigma_3^2(G)$ denote, respectively, the variance due to sampling $\mathbb{V}[a_i]\mathbb{E}[\rho_i]$ and the variance due to noise $\mathbb{V}[\eta_i]\mathbb{E}[\rho_i]$ of node degrees.

**Theorem 5.** Given two networks, $G_1$ with an average degree $\mu_1(G_1)$, $G_2$ with an average degree $\mu_2(G_2)$, and $G_1 \cap G_2$, $Q_1^N$ is no smaller than
\[
1 - (1 - \frac{\sigma_2^2(G)}{\sigma_3^2(G)} + \sigma_3^2(G)),
\]
where $\sigma_2^2(G)$ is the variance due to sampling $\mathbb{V}[a_i]\mathbb{E}[\rho_i]$ and $\sigma_3^2(G)$ is the variance due to noise.

Proof: The expected value of $n(G_1)$, $n(G_2)$, and $n(G_1 \cap G_2)$ are identical for node-only attributes and node degrees because each node has the same number of being selected during uniform sampling. However, noise is introduced by $G_1 \cap G_2$ for node degrees because not every two matched nodes in $G_1 \cap G_2$ will be sampled, i.e., probably only one of them is sampled. The variance due to sampling $\sigma_2^2(G)$ can be calculated by Theorems 3 and 4. By assuming that $\frac{n(G_1 \setminus G_2) - 1}{N(G_1 \setminus G_2) - 1} \approx n(G_1 \setminus G_2) \frac{N(G_2 \setminus G_1)}{N(G_1 \setminus G_2)}$, $\sigma_2^2(G)$ is
\[
\sigma_2^2(G) = (\frac{N(G_1 \setminus G_2)}{eN})(\frac{\sigma_1(G_1 \setminus G_2)}{n(G_1 \setminus G_2)})^2(1 - \frac{n(G_1 \setminus G_2) - 1}{N(G_1 \setminus G_2) - 1}) + (\frac{N(G_1 \setminus G_2)}{eN})(\frac{\sigma_1(G_1 \setminus G_2)}{n(G_1 \setminus G_2)})^2(1 - \frac{n(G_1 \setminus G_2) - 1}{N(G_1 \setminus G_2) - 1})),
\]
where $\mathbb{E}(n(G_1 \setminus G_2))$, $\mathbb{E}(n(G_2 \setminus G_1))$, and $\mathbb{E}(n(G_1 \cap G_2))$ can be calculated by Lemma 2. Therefore, the expected number of samples in each strata can be calculated by Lemma 2. The theorem follows.
which can be simplified as follows.

\[
\left(\frac{\sigma_{D}(G_1 \setminus G_2)}{\epsilon N}\right)^2 \left(1 - \frac{n(G_1 \setminus G_2)}{N(G_1 \setminus G_2)}\right) + \left(\frac{\sigma_{D}(G_2 \setminus G_1)}{\epsilon N}\right)^2\left(1 - \frac{n(G_2 \setminus G_1)}{N(G_2 \setminus G_1)}\right),
\]

\[\text{(9)}\]

Therefore, we focus on finding the variance due to noise \(\sigma^2_{\theta}(G)\). If a sample in \(G_1 \cap G_2\) does not have the corresponding sample in \(G_1\) for matching, the expected error is equal to \(-\mu_D(G_1)\). If a sample in \(G_1 \cap G_2\) does not have a corresponding sample in \(G_2\), the expected error is equal to \(-\mu_D(G_2)\). Let \(a\) and \(b\) denote \(n(G_1 \cap G_2)\) and \(n(G_1 \cap G_2)\), respectively. The variance is only introduced from \(G_1 \cap G_2\), and thus \(\sigma^2_{\theta}(G)\) can be approximated as follows. The average error in \(\sigma^2_{\theta}(G)\) is \(\frac{a \mu_D(G_2) - b \mu_D(G_1)}{n(G_1 \cap G_2)}\). We derive the variance due to noise as

\[
\sigma^2_{\theta}(G) = \frac{n(G_1 \cap G_2)}{N(G_1 \cap G_2)} \left(\frac{ab(a + b + 2E[\tilde{\eta}])}{n(G_1 \cap G_2) - 1} \right) \left(\frac{(\mu_D(G_1) - \mu_D(G_2))^2}{n(G_1 \cap G_2) \sigma^2_{\theta}(G)}\right)
\]

The theorem follows.

Thus, we analyze the sampling quality with an accurate matching oracle from the perspective of clustering coefficients with the error model, i.e., \(\{\alpha + \eta_i\}\).

**Corollary 6.** Given two networks \(G_1\) with average degree \(\mu_D(G_1)\) and average clustering coefficient \(\mu_C(G_1)\), \(G_2\) with average degree \(\mu_D(G_2)\) and average clustering coefficient \(\mu_C(G_2)\), and \(G_1 \cap G_2\), \(\sigma^2_{\theta}(G)\) is greater than \(\frac{1}{\epsilon N} \left(\sigma^2_{\theta}(G_1) + \sigma^2_{\theta}(G_2)\right)\).

**Proof.** The expected value of \(\tilde{\eta}\) is sampled is in the sampling error \(\eta_i\). The variance due to sampling \(\sigma^2_{\theta}(G)\) can be calculated by Theorem 4 and Corollary 3, i.e.,

\[
\sigma^2_{\theta}(G) = \frac{n(G_1 \cap G_2)}{N(G_1 \cap G_2)} \left(\frac{ab(a + b + 2E[\tilde{\eta}])}{n(G_1 \cap G_2) - 1} \right) \left(\frac{(\mu_D(G_1) - \mu_D(G_2))^2}{n(G_1 \cap G_2) \sigma^2_{\theta}(G)}\right)
\]

Theorem 7. Given two networks \(G_1\) and \(G_2\), and a practical matching algorithm with accuracy \(\alpha\), \(Q^2\) is no smaller than \(\frac{1}{\epsilon N} \left(\sigma^2_{\theta}(G_1) + \sigma^2_{\theta}(G_2)\right)\), where \(\sigma^2_{\theta}(G)\) can be found by replacing \(n(G_1 \cap G_2)\), \(n(G_2 \setminus G_1)\), and \(n(G_1 \cap G_2)\) with \(\alpha n(G_1 \cap G_2)\), and \(\alpha n(G_2 \setminus G_1)\), and \(\alpha n(G_1 \cap G_2)\) in Theorem 4, respectively, and the variance due to noise \(\sigma^2_{\theta}(G)\) is

\[
\frac{1}{n - 1} \left(\frac{n(G_1 \setminus G_2)\mu_D(G_1) - \mu_D(G_1))^2}{n(G_1 \setminus G_2)} + \frac{1}{2}((\mu_D(G_1) - \mu_D(G_1 \cap G_2))^2 + (\mu_D(G_2) - \mu_D(G_1 \cap G_2))^2 n(G_1 \cap G_2)).
\]

**Proof.** First, let \(a_n = a_i + \eta_i\) be a random sample, where \(\eta_i\) represents the sample noise of \(a_i\). Specifically, \(\eta_i\) is generated from mismatches, i.e., the matching algorithms are either not able to perfectly match the accounts belong to the same users from different networks or some accounts that do not belong to the same users are wrongly matched. \(\tilde{a}_n\) is defined as the average node attributes of the \(n\) nodes in \(G_{QM}\), i.e., \(\tilde{a}_n = \frac{1}{2} \sum_{i=1}^{n} (a_i + \eta_i)\). \(a_i\) and \(\eta_i\) are independent because we draw samples with uniform sampling from two networks independently, i.e., the event that \(v_i\) is sampled is independent to the event that the matched node in another network is sampled. Therefore, \(\tilde{a}_n\) can be represented as \(\sigma^2_{\theta}(G_1) + \sigma^2_{\theta}(G_2)\). However, the number of correct samples now correlates with accuracy \(\alpha\) of the oracle, and thus the number of correct samples becomes the number of samples in Theorem 4 multiplied by the accuracy \(\alpha\). With the number of correct samples, the variance due to sampling \(\sigma^2_{\theta}(G)\) can be found by Theorem 4. Then, we approximate the variance due to noise \(\sigma^2_{\theta}(G)\) by

\[
\sigma^2_{\theta}(G) = \frac{1}{n - 1} \sum_{i=1}^{n} (\eta_i - \bar{\eta})^2 = \frac{1}{n - 1} \sum_{i=1}^{n} (\eta_i - \bar{\eta})^2.
\]

where \(\bar{\eta}\) is the average error of the \(n\) nodes in \(G_{QM}\). Here, we assume that the node attribute of a node \(v_i\) will be different if \(v_i\) is
unmatched or mismatched. Specifically, for a sample \(a^n\),
\[
\eta^2 = \begin{cases} 
0 & \text{if } a^n \text{ is correct,} \\
(\mu_A(G_2) - \mu_A(G_1))^2 & \text{if } a^n \text{ is in } G_1 \setminus G_2 \text{ and matched to } G_2, \\
(\mu_A(G_1) - \mu_A(G_2))^2 & \text{if } a^n \text{ is in } G_2 \setminus G_1 \text{ and matched to } G_1, \\
(\mu_A(G_1) - \mu_A(G_1 \cap G_2))^2 & \text{if } a^n \text{ is in } G_1 \cap G_2 \text{ and not matched to } G_2, \\
(\mu_A(G_2) - \mu_A(G_1 \cap G_2))^2 & \text{if } a^n \text{ is also in } G_1 \cap G_2 \\
\text{and not matched to } G_1. 
\end{cases}
\]

As the ratio between the number of mismatched samples and unmatched samples depends on the account matching algorithm, to derive the closed form of \(\sigma_\lambda(G)\), we assume that \(\bar{\eta}\) approaches 0 to find \(\tilde{\sigma}_\lambda^2(G)\) is as follows.
\[
\bar{\eta}^2 = \frac{1}{n-1}((1 - \alpha)(n(G_1 \setminus G_2)(\mu_A(G_2) - \mu_A(G_1)))^2 \\
+ n(G_2 \setminus G_1)(\mu_A(G_1) - \mu_A(G_2))^2 + \frac{1}{2} n(G_1 \cap G_2) \\
((\mu_A(G_1) - \mu_A(G_1 \cap G_2))^2 + (\mu_A(G_2) - \mu_A(G_1 \cap G_2))^2)).
\]

The theorem follows. \(\square\)

Based on Theorems 3 and 2, we analyze the sampling quality from the perspective of node degrees in the following theorem.

**Theorem 8.** Given two networks \(G_1\) and \(G_2\), and a practical matching algorithm with accuracy \(\alpha\), \(Q^t\) is no smaller than 1 - \(\frac{1}{2}(\sigma_\mu^2(G) + \sigma_\nu^2(G))\), with the same \(n(G_1 \setminus G_2)\), \(n(G_2 \setminus G_1)\), and \(n(G_1 \cap G_2)\) as in Theorem 3.

**Proof.** For the same model in Theorem 2 i.e., \(d_i = \{a_i + \bar{\eta}_i\}\), the variance due to sampling \(\tilde{\sigma}_\lambda^2(G)\) can be modified from Equation 10 by replacing the number of samples \(n(G_1 \setminus G_2)\), \(n(G_2 \setminus G_1)\), and \(n(G_1 \cap G_2)\) with the number of accurate samples \(\alpha n(G_1)\), \(\alpha n(G_2)\), and \(\alpha n(G_1 \cap G_2)\),
\[
(\frac{\sigma_D(G_2) \cdot \sigma_G}{\epsilon N})^2(1 - \frac{\alpha n(G_1 \setminus G_2)}{N(G_1 \setminus G_2)}) + (\frac{\sigma_D(G_1) \cdot \sigma_G}{\epsilon N})^2(1 - \frac{\alpha n(G_2 \setminus G_1)}{N(G_2 \setminus G_1)}).
\]

The variance due to noise \(\tilde{\sigma}_\lambda^2(G)\) increases because of the practical matching oracle with accuracy \(\alpha\). The square of the expected error \(\bar{\eta}^2\) of the samples in \(G_1\) is as follows.
\[
\eta_i^2 = \begin{cases} 
0 & \text{if } c_i \text{ is correct,} \\
(\mu_D(G_2))^2 & \text{if } d_i \text{ is in } G_1 \setminus G_2 \text{ and matched to } G_2, \\
(-\mu_D(G_2))^2 & \text{if } d_i \text{ is in } G_1 \cap G_2 \text{ and not matched to } G_2.
\end{cases}
\]

Note that \(\tilde{\sigma}_\lambda^2(G)\) is comprised of two parts: the first is generated because not every two matched nodes in \(G_1 \cap G_2\) will be sampled, as mentioned in Corollary 6 and the second part contains the inaccurately matched nodes and unmatched nodes in the practical matching oracle.

For example, if the accuracy of the matching oracle is high, e.g., 91.38% in [30], and 89.8 ± 0.7% in [18], \(\alpha\) will be close to 1 and the noise term becomes negligible. Compared with Theorem 4 we need more than 1/s of the times of the samples to achieve the same sampling quality. If the matching algorithm is not accurate enough, since \(Q^t\) is related to \(\sigma_\lambda^2(G)\) and \(\tilde{\sigma}_\lambda^2(G)\), it becomes very important to reduce the variance due to noise, i.e., \(\tilde{\sigma}_\lambda^2(G)\), by setting different \(n(G_1)\) and \(n(G_2)\) for changing the expected values of \(\alpha n(G_1 \setminus G_2), n(G_2 \setminus G_1), \) and \(n(G_2 \cap G_1)\), because it is much larger than \(\sigma_\lambda^2(G)\) in this case.

**3.4 Random Walk Analysis**

In the following, by exploiting the theoretical results of uniform sampling (US), we analyze the sampling quality of RW. Specifically, the analysis of RW is based on the number of steps, denoted as \(t\), instead of the number of nodes in US since it is necessary for RW to carefully examine the state transitions (i.e., steps). Note that RW studied in this paper is Random Walk with Revisits [22], where nodes are allowed to be traversed multiple times. Nevertheless, it only needs to sample each node once, and the above traversing model is employed to identify the new nodes to be crawled. In other words, when a node is revisited in a step of RW, it is not necessary to sample the node again, and we move on to the next node accordingly.

Compared with US, it is envisaged that RW requires to sample more nodes in order to reach the same sampling quality in US, because RW tends to sample the nodes with high degrees, incurring
the problem that the sample mean will deviate from true mean when the number of samples is not sufficient. Based on the theoretical results of US, we derive the theoretical results of RW as follows.

**Definition 2.** Let $R^A_t$, $R^D_t$, and $R^C_t$ denote, respectively, the probability that the difference between the common mean and the sample mean of node-only attributes, node degrees, and clustering coefficients in RW is less than a threshold $\epsilon$.

**Definition 3.** Let $\delta_{LA}(G, t)$, $\delta_{LD}(G, t)$, and $\delta_{LC}(G, t)$ denote, respectively, the variance due to noise of the sample mean on node-only attributes, node degrees, and clustering coefficients in RW, which is related to the number of steps $t$.

**Theorem 10.** Given two networks $G_1$ and $G_2$, and a practical matching algorithm with accuracy $\alpha$, $R^A_t$ is greater than $1 - \frac{1}{\epsilon^2}(\delta_{LA}(G, t) + \delta_{LD}(G, t))$, $R^D_t$ is greater than $1 - \frac{1}{\epsilon^2}(\delta_{LA}(G, t) + \delta_{LD}(G, t))$, and $R^C_t$ is greater than $1 - \frac{1}{\epsilon^2}(\delta_{LA}(G, t) + \delta_{LD}(G, t))$.

**Proof.** Given an error threshold $\epsilon$, the number of nodes $n$ to be crawled from each network in US has been derived in the theorems early in this section. In the following, therefore, we exploit theoretical results of US to analyze $R^A_t$, $R^D_t$, and $R^C_t$ in RW. Specifically, let $N(G, d)$ and $N(GQM, d)$ denote the number of nodes with degree $d$ in the network $G$ and sampled graph $GQM$, respectively. Also, let $k_l$ and $k_h$ denote the lowest and highest degree in network $G$. First, we derive the expected number of nodes with degree $k$ in $GQM$ crawled by US as follows:

$$\mathbb{E}[N(GQM, k)] = n \frac{N(G, k_l)}{\sum_{d=k_l}^{k_h} N(G, d)}.$$  

Previous studies [11, 25] have pointed out that many graphs indeed exhibit a typical power-law degree distribution, i.e., the distribution $N(G, d)$ is proportional to $d^{-\alpha}$ for some constant $\alpha$ independent of the scale of the network. Therefore, we have

$$\mathbb{E}[N(GQM, k)] = n \frac{k_l^{-\alpha}}{\sum_{d=k_l}^{k_h} d^{-\alpha}}.$$  

For each node $v_i$, the literature [12, 13] demonstrates that the node selection probability $p(v_i)$ of RW is proportional to its node degree $d_{eg}(v_i)$,

$$p(v_i) = \frac{d_{eg}(v_i)}{|E|},$$

where $|E|$ is the number of edges. When we sample a sufficient number of nodes in RW such that the number of nodes with degree $k_i$ is identical to the one in US, the number of nodes with any degree higher than $k_i$ in RW is expected to exceed the one in US. The reason is that other nodes with higher degrees are more likely to be sampled according to the above equation. Therefore, to meet the sampling quality of US, let $t_{k_l}$ denote the number of steps for the nodes with degree $k_l$ to be visited by RW, such that the expected number of nodes with degree $N(GQM, k_l)$ is at least $\mathbb{E}[N(GQM, k_l)]$. We derive $t_{k_l}$ by the following inequality:

$$\sum_{i=1}^{t_{k_l}} \frac{C_{N(G, k_l)}^{N(GQM, k_l)} \cdot k_l^{N(GQM, k_l)} - N(G, k_l)}{N(G, k_l)^{k_l}} \geq \mathbb{E}[N(GQM, k_l)].$$

With $t_{k_l}$ derived from Equation (18), the total number of steps $t$ required by RW is:

$$t = \frac{t_{k_l}}{k_l} \sum_{d=k_l}^{k_h} d.$$  

Finally, we derive the relationship between the number steps and the number of sampled nodes in RW according to the analysis of the cover time in the literature, where the cover time represents the number of steps to traverse every node at least once. In the worst case, the cover time from any starting node in a graph with $n$ nodes is at most $(4/27 + o(1))n^2$ [9]. Therefore, by processing $t$ steps in RW, we are able to acquire a sampled graph with $n_{rw}$ nodes, where $n_{rw}$ follows

$$n_{rw} \geq \sqrt{\frac{4}{27}t}.$$  

In the sampled graph with $n_{rw}$ nodes, since the number of nodes with degree larger than $k_l$ is expected to exceed the one in US, we trim the oversampled graph to calibrate the bias with target degree distribution and clustering coefficient with an effective way in [27]. Let $f(n_{rw})$ denote the transfer function which transforms the number of nodes in RW $n_{rw}$ into an equivalent number of nodes in US. Therefore, $R^A_t$ is greater than $1 - \frac{1}{\epsilon^2}(\delta_{LA}(G, t) + \delta_{LD}(G, t))$

$$1 - \left(\frac{N(G_1 \setminus G_2) \sigma_{LA}(G_1 \setminus G_2)}{\epsilon N} \frac{1}{f(n(G_1 \setminus G_2))} \right)^2 \left(1 - \frac{1}{f(n(G_1 \setminus G_2))} \right)$$

$$+ \left(\frac{N(G_2 \setminus G_1) \sigma_{LD}(G_2 \setminus G_1)}{\epsilon N} \frac{1}{f(n(G_2 \setminus G_1))} \right)^2 \left(1 - \frac{1}{f(n(G_2 \setminus G_1))} \right)$$

$$+ \left(\frac{(N(G_1 \cap G_2) \sigma_{LC}(G_1 \cap G_2)}{\epsilon N} \frac{1}{f(n(G_1 \cap G_2))} \right)^2 \left(1 - \frac{1}{f(n(G_1 \cap G_2))} \right),$$

and $\delta_{LD}(G, t)$ is

$$\frac{1 - \alpha}{f(n)} \left(\frac{1}{f(n(G_1 \setminus G_2))} \right)^2 \left(1 - \frac{1}{f(n(G_1 \setminus G_2))} \right) + \left(\frac{1}{f(n(G_1 \cap G_2))} \right)^2 \left(1 - \frac{1}{f(n(G_1 \cap G_2))} \right).$$

Similarly, $R^D_t$ is greater than $1 - \frac{1}{\epsilon^2}(\delta_{LA}(G, t) + \delta_{LD}(G, t))$

$$\frac{1}{f(n)} \left(\frac{1}{f(n(G_1 \setminus G_2))} \right)^2 \left(1 - \frac{1}{f(n(G_1 \setminus G_2))} \right) + \left(\frac{1}{f(n(G_1 \cap G_2))} \right)^2 \left(1 - \frac{1}{f(n(G_1 \cap G_2))} \right).$$

Finally, $R^C_t$ is greater than $1 - \frac{1}{\epsilon^2}(\delta_{LA}(G, t) + \delta_{LD}(G, t))$

$$\frac{1}{f(n)} \left(\frac{1}{f(n(G_1 \setminus G_2))} \right)^2 \left(1 - \frac{1}{f(n(G_1 \setminus G_2))} \right) + \left(\frac{1}{f(n(G_1 \cap G_2))} \right)^2 \left(1 - \frac{1}{f(n(G_1 \cap G_2))} \right).$$

**□**

Although the sampling variance in RW is smaller than that in US, we trim the nodes crawled by RW to correct the bias and thus $n_{rw} \geq f(n_{rw})$. Therefore, compared with $Q^A_t$, $R^A_t$ requires to sample more nodes in order to reach the same sampling quality. It is worth noting that if the node attributes are independent of node degrees, $n_{rw} = f(n_{rw})$ and thus $Q^A_t = R^A_t$.

### 4. QMSAMPLER

For multiple OSNs, such as two OSNs $G_1$ and $G_2$ of sizes $N(G_1)$ and $N(G_2)$ respectively, a simple approach is to set the number of nodes to be crawled from each OSN as:

$$n(G_1) = \frac{N(G_1)}{N(G_1) + N(G_2)} n, \quad n(G_2) = \frac{N(G_2)}{N(G_1) + N(G_2)} n,$$

where $n$ is the target sample number. However, this approach does not examine the number of overlapping crawled nodes in the two OSNs. Moreover, it does not consider the variances of the two
sampled networks derived from the two OSNs. When the variance of the sampled values of the nodes in $G_1$ is larger than that in the other network, it is necessary to sample more nodes in $G_1$ because a small number of nodes is difficult to accurately reflect the properties of $G_1$. Therefore, based on the analytical results in Section 3 we propose the Quality-guaranteed Multi-network Sampler (QMSampler) for three sampling scenarios, i.e., size-constrained, quality-constrained, and time-constrained sampling. We introduce QMSampler in two OSNs for node attributes and then extend it to more OSNs.

4.1 Size-Constrained Sampling

In size-constrained sampling, given the user-specified error threshold $\epsilon$ and the number of crawled nodes $n$, QMSampler attempts to find $n(G_1)$ and $n(G_2)$ with $n(G_1) + n(G_2) = n$, in order to maximize the sampling quality $Q^A$. We calculate the first derivative of the sampling quality $\frac{\partial Q}{\partial n(G_1)}$ from Theorem 7 and replace $n(G_1)$ and $n(G_2)$ with $\frac{n(G_1) G_2}{N(G_1) G_2}$ and $n - \frac{n(G_1) G_2}{N(G_1) G_2}$, respectively. Assuming that $\frac{n(G_1) G_2}{N(G_1) G_2} < 1$, the first derivative can be calculated as follows:

$$\frac{\partial}{\partial n(G_1)}\left(\frac{n(G_1) G_2}{N(G_1) G_2}\right)^2 + \frac{\partial}{\partial n(G_1)}\left(\frac{n(G_2) G_1}{N(G_1) G_2}\right)^2$$

$$+ \frac{1 - \alpha}{n - 1} \frac{n(G_1) G_2}{N(G_1)} (\mu_A(G_2) - \mu_A(G_1))^2$$

$$+ \frac{n(G_2) G_1}{N(G_2)} (n(n - n(G_1))) (\mu_A(G_2) - \mu_A(G_1))^2).$$

The optimal number of samples of $G_1$ is obtained by setting the first derivative to 0. Therefore, let $u = \frac{n(G_1) G_2}{N(G_1) G_2}, v = \frac{n(G_2) G_1}{N(G_1) G_2}$. Then, we have

$$w^2 = \frac{n(G_1) G_2}{N(G_1) G_2} = \frac{n(G_2) G_1}{N(G_1) G_2} = w$$

We further simplify the equation as follows:

$$u(n(G_1) + n)^3 n(G_1)^3 - 2u(n(G_1) + n)^3 n(G_1) - n^3 = 0.$$ (22)

Note that Equation 22 is a sextic equation that can be solved by Kampe de Feriet functions [15]. Then, $n(G_2) + n(G_1)$ can be found by $n - n(G_1)$.

According to the above results, the maximal sampling quality is correlated with the standard deviations $\sigma_A(G_1), \sigma_A(G_2), \sigma_A(G_1 \cap G_2)$, the number of overlapping nodes $N(G_1 \cap G_2)$ in different OSNs, and the standard deviation due to noise $\sigma_A(G)$ (the third term in Equation 21). However, the above-mentioned parameters are unknown at the beginning. Therefore, QMSampler incrementally increases the number of samples to estimate the unknown parameters and approximates optimal allocation ratio until the number of samples equals $n$, which is illustrated in the following example. The pseudocode of QMSampler in the size-constrained scenario is presented in Algorithm 1.

**Algorithm 1 QMSampler in the size-constrained scenario**

**Input:** $m$ Graphs $G_1, G_2, G_3, ..., G_m$, social network size $N_1, N_2, ..., N_m$, matching accuracy $\tau$, error threshold $\epsilon$, and total sample size $n$

**Output:** An unbiased global graph

1. Initial sampling for parameter estimation on $G_1$
2. $G_e = G_1$
3. for $i = 2$ to $m - 1$
4. Initial sampling on $G_i$
5. User matching between $G_e$ and $G_i$
6. Estimate and record parameters between $G_e$ and $G_i$
7. $G_e = G_e \cup G_i$
8. Initial sampling on $G_m$
9. User matching between $G_e$ and $G_m$
10. Estimate and record parameters between $G_e$ and $G_m$
11. Set the remainder of total samples $R = n$
12. for $i = m$ to $2$
13. Compute the optimal number $n_i$ on $G_i$ and $G_e$, for sample size $R$
14. $G_e = G_e \cap G_i$
15. $R = R - n_i$
16. $n_m = R$
17. Output graph with the number of samples $n_i$ of network $G_i$ and the sampling quality $Q^A$

**Example 1** We illustrate QMSampler with an accurate matching oracle under the size-constrained scenario. Given two networks $G_1$ with 6000 nodes and $G_2$ with 12000 nodes, $N(G_1 \cap G_2)$ is 5900, which is unknown at the beginning, and the error threshold $\epsilon$ is set as 2, the goal is to crawl $n = 1800$ samples from the two networks. QMSampler first performs an initial sampling from the two networks with 20% of the samples, i.e., 360 samples. The numbers of samples from $G_1$ and $G_2$ follow the ratio of $N(G_1)$ and $N(G_2)$. Therefore, QMSampler draws 120 and 240 samples from $G_1$ and $G_2$, respectively, for estimating $\sigma_A(G_1), \sigma_A(G_2), \sigma_A(G_1 \cap G_2)$, and $N(G_1 \cap G_2)$. The estimation results in this example are $\sigma_A(G_1) = 64, \sigma_A(G_2) = 190$, and $\sigma_A(G_1 \cap G_2) = 88$. Moreover, QMSampler observes that the number of matched nodes in the initial sampling is 5. QMSampler also estimates the network size $N(G_1 \cap G_2)$ by substituting $E[\hat{n}]$ with 5, $n_1(G_1 \cap G_2)$ with $N(G_1 \cap G_2)(n_1(G_1)), i.e., N(G_1 \cap G_2) \hat{\tau} 240$ in Equation 21. Afterward, QMSampler solves Equation 21 and estimates $N(G_1 \cap G_2)$ as 5963.

Next, $\alpha$ is 1 in perfect matching oracle, and thus $w = 0$ in Equation 22. QMSampler solves Equation 22 and obtains the optimal solution $n(G_1) = 483$ and $n(G_2) = 1317$. In the next iteration, QMSampler samples 20% of total samples again. The number of samples for $G_1$ and $G_2$ are $\frac{483}{183, 183} = 720 - 120 = 71$ and $360 - 71 = 289$, respectively. It is worth noting that the samples in the previous iteration are kept and combined with new samples. After sampling 71 nodes from $G_1$ and 289 nodes from $G_2$, we have 191 and 529 nodes from $G_1$ and $G_2$, respectively, which is different from the ratio of $N(G_1)$ and $N(G_2)$. Afterward, QMSampler re-estimates $\sigma_A(G_1), \sigma_A(G_2), \sigma_A(G_1 \cap G_2)$, and $N(G_1 \cap G_2)$, and $N(G_1 \cap G_2) = 54, 187, 92, and 5948$, respectively, and solves Equation 22 for obtaining the optimal solution $n(G_1)$ and $n(G_2)$ again. The iteration repeats until 1800 nodes are sampled. Finally, $\sigma_A(G_1), \sigma_A(G_2), \sigma_A(G_1 \cap G_2)$, and $N(G_1 \cap G_2)$ as 52, 183, 98, and 5887. QMSampler outputs the sampled network and the sampling quality $Q^A$ is 95.7%.

4.2 Quality-Constrained Sampling

Size-constrained sampling can be extended to support quality-constrained sampling. More specifically, size-constrained sampling is given the number of crawled nodes $n$ to maximize the sampling quality $Q^A$. Quality-constrained sampling is given the minimum required quality $Q$ to minimize the number of crawled nodes $n$. The sampling quality $Q^A$ increases as $n$ grows; hence, quality-constrained sampling can be achieved by linearly increasing $n$ in size-constrained sampling in Equation 22 until the sampling quality just reaches the minimum required quality $Q$. A more efficient way is to exploit a binary search method; then, the minimal $n$ in quality-constrained sampling can be acquired in a much shorter time in this case. Note that quality-constrained sampling does not need to repeat
the crawling process of size-constrained sampling multiple times. It only repeats the calculation in Equation 22 with binary search to find the minimal n. The study of different quality settings will be presented in Section 5. Also, the pseudocode of QMSampler in the size-constrained scenario is presented in Algorithm 2.

4.3 Time-Constrained Sampling

Recall that \( \mathcal{Q} \) denotes the minimum required quality. Because neither n nor \( \mathcal{Q} \) are specified in time-constrained sampling, it is necessary to crawl all OSNs until the deadline has passed. According to Equation 13 since oversampling of an OSN \( G_1 \) will undermine the sampling quality, it is important for time-constrained sampling to identify the OSN that is oversampled before generating the output graph \( G_{QM} \). More specifically, let \( \overline{\mathcal{P}}(G_1) \) and \( \overline{\mathcal{P}}(G_2) \) denote, respectively, the numbers of nodes crawled from \( G_1 \) and \( G_2 \) before the deadline. In addition, let \( \overline{\mathcal{Q}} \) denote the sampling quality by including all crawled nodes in the output dataset. By inserting \( \overline{\mathcal{P}}(G_1) \) into Equation 21, QMSampler finds the optimal n(\( G_1 \)) in \( G_2 \) that yields the best sampling quality \( \overline{\mathcal{Q}} \), if nodes \( \overline{\mathcal{P}}(G_1) \) and n(\( G_2 \)) from \( G_1 \) and \( G_2 \) are included in the output dataset. In this case, \( G_2 \) will be oversampled if n(\( G_2 \)) < \( \overline{\mathcal{P}}(G_2) \), and thus \( \overline{\mathcal{Q}} > \mathcal{Q} \). On the other hand, by inserting \( \overline{\mathcal{P}}(G_2) \) into Equation 21, QMSampler finds the optimal n(\( G_1 \)) in \( G_1 \) that yields the best sampling quality \( \overline{\mathcal{Q}} \), if nodes n(\( G_1 \)) and \( \overline{\mathcal{P}}(G_2) \) from \( G_1 \) and \( G_2 \) are included in the output graph. In this case, \( G_1 \) will be oversampled if n(\( G_1 \)) < \( \overline{\mathcal{P}}(G_1) \) such that \( \overline{\mathcal{Q}} < \mathcal{Q} \). The following theorem proves that the optimal sampling quality can be acquired by comparing \( \overline{\mathcal{Q}}, \mathcal{Q} \), and \( \overline{\mathcal{Q}} \).

**Theorem 11.** The optimal quality in time-constrained sampling is the maximal one of \( \overline{\mathcal{Q}}, \mathcal{Q} \), and \( \overline{\mathcal{Q}} \).

**Algorithm 2 QMSampler in the quality-constrained scenario**

**Input:** m Graphs \( G_1, G_2, G_3, \ldots, G_m \), social network size \( N_1, N_2, \ldots, N_m \), matching accuracy \( \tau \), error threshold \( c \), node-size increment \( n_{inc} \), and the minimum required quality \( \mathcal{Q} \)

**Output:** The unbiased global graph with minimum size n

1. Initial sampling for parameter estimation on \( G_1 \)
2. for \( i = 0 \) and \( G_x \in G_1 \)
3. for \( i = 2 \) to \( m - 1 \)
4. Initialize sampling on \( G_i \)
5. User matching between \( G_x \) and \( G_i \)
6. Estimate and record parameters between \( G_x \) and \( G_i \)
7. \( G_x = G_x \cup G_i \)
8. Initial sampling on \( G_{mn} \)
9. User matching between \( G_x \) and \( G_{mn} \)
10. Estimate and record parameters between \( G_x \) and \( G_{mn} \)
11. Marker: repeat updates
12. \( n = n + n_{inc} \)
13. Set the remainder of total samples \( R = n_{inc} \)
14. for \( i = m \) to 2
15. Compute the optimal number \( n_i \) on \( G_x \) and \( G_i \) for sample size \( R \)
16. \( G_x = G_x \setminus G_i \)
17. \( R = R - n_i \)
18. \( n_i = R \)
19. Sample network \( G_i \) with the number of samples \( n_i \)
20. User matching between \( G_x \) and \( G_{mn} \)
21. Estimate and record parameters between \( G_x \) and \( G_{mn} \)
22. while \( \overline{\mathcal{Q}} < \mathcal{Q} \)
23. goto: repeat updates
24. Output graph with the minimum size n and the sampling quality \( \overline{\mathcal{Q}} \)

**Algorithm 3 QMSampler in the time-constrained scenario**

**Input:** m Graphs \( G_1, G_2, G_3, \ldots, G_m \), social network size \( N_1, N_2, \ldots, N_m \), matching accuracy \( \tau \), and total time \( T \)

**Output:** The unbiased global graph with maximum quality

1. while \( T \) does not run out do
2. Continue sampling and record \( n_i \)
3. Compute \( \overline{\mathcal{Q}} \)
4. for \( i = 1 \) to \( m \)
5. Compute the optimal number \( n_i \) with other \( n_j \), where \( j \neq i \)
6. if \( \overline{\mathcal{P}}(G_i) > \overline{\mathcal{P}}(G_j) \) then
7. \( n_i = \overline{\mathcal{P}}(G_i) \)
8. Compute \( \overline{\mathcal{Q}}_i \)
9. \( \overline{\mathcal{Q}} = \overline{\mathcal{Q}}_i \)
10. Output graph with the number of samples \( n_i \) of network \( G_i \) and the sampling quality \( \overline{\mathcal{Q}} \)

**Proof.** First, note that the case with both \( n(G_1) < \overline{\mathcal{P}}(G_1) \) and \( n(G_2) < \overline{\mathcal{P}}(G_2) \) does not exist because it is impossible to oversample both OSNs. Second, if \( n(G_2) > \overline{\mathcal{P}}(G_2) \) after inserting \( \overline{\mathcal{Q}}_i \) into Equation 21, it is not possible to sample more nodes from \( G_2 \) to improve \( \overline{\mathcal{Q}}_i \) since the deadline has passed. Similarly, if \( n(G_1) > \overline{\mathcal{P}}(G_1) \) after inserting \( \overline{\mathcal{Q}}_i \) into Equation 21, it is not possible to sample more nodes from \( G_1 \) to improve \( \overline{\mathcal{Q}}_i \). The theorem follows.

According to Theorem 11, we derive \( \overline{\mathcal{P}}(G_1) \) and \( \overline{\mathcal{P}}(G_2) \) as follows. For the sampling quality function shown in Equation 21, we find the first derivative of \( n(G_2) \) as:

\[
-2\left(\frac{\overline{\mathcal{P}}(G_2)\sigma^2_{\Delta}(G_2)}{\alpha N} \right) n(G_2)^{-3} + \overline{\mathcal{Q}}(G_1) - 1 \overline{\mathcal{P}}(G_1) - n(G_2) - 1)^2
\]

(23)

It can be simplified by letting \( D_2 = -2\left(\frac{\overline{\mathcal{P}}(G_2)\sigma^2_{\Delta}(G_2)}{\alpha N} \right) n(G_2)^{-3} + \overline{\mathcal{Q}}(G_1) - 1 \overline{\mathcal{P}}(G_1) - n(G_2) - 1)^2 \), and assuming that \( \overline{\mathcal{Q}}(G_1) - 1 \approx \overline{\mathcal{Q}}(G_1) \). By letting the first derivative be 0, we have

\[
\overline{\mathcal{Q}}(G_1)n(G_2)^3 + D_2 n(G_2)^2 + 2\overline{\mathcal{Q}}(G_1)D_2 n(G_2) + \overline{\mathcal{Q}}(G_1)^2 = 0.
\]

(24)

QMSampler finds the optimal n(\( G_2 \)) in \( G_2 \) that yields the best sampling quality \( \overline{\mathcal{Q}}_i \) by solving Equation 24. The optimal n(\( G_1 \)) in \( G_1 \) can be found in a similar manner. The pseudocode of QMSampler in the time-constrained scenario is presented in Algorithm 3.

4.4 Parameter Estimation

Because the sizes of many OSNs are specified on the corresponding websites or reported in the literature [5][13], QMSampler may not need to estimate parameters \( N(G_1) \) and \( N(G_2) \). Therefore, in the following, we first focus on finding the variance \( \sigma^2_{\Delta}(G) \) of the sample, the number of overlapping nodes \( N(G_1 \cap G_2) \) in different OSNs, and the variance \( \sigma^2_{\Delta}(G) \) from the noise. Specifically, QMSampler periodically evaluates \( \sigma^2_{\Delta}(G_1), \sigma^2_{\Delta}(G_2), \sigma^2_{\Delta}(G_1 \cap G_2), \mu_{\Delta}(G_1), \mu_{\Delta}(G_2), \mu_{\Delta}(G_1 \cap G_2), \mu_{\Delta}(G_1), \mu_{\Delta}(G_2), \) and \( \mu_{\Delta}(G_1 \cap G_2) \) based on the sampling variance of the n nodes. As the number of samples increases, it is expected that the sampling variance will become close to the exact value. Next, the number of overlapping nodes \( N(G_1 \cap G_2) \) can be estimated by the following equation.
In Equation (25), $E[n(G_1 \cap G_2)]$ is estimated by the number of sampled matches in multiple OSNs. The numbers $n_1(G_1 \cap G_2)$ and $n_2(G_1 \cap G_2)$ can be estimated by $N(G_1 \cap G_2)$, respectively. As the left-hand side of the above equation is a function of $N(G_1 \cap G_2)$, a binary search can be employed to solve the above equation and find $N(G_1 \cap G_2)$.

4.5 Three and More Networks

Although the theoretical results in Section 3 only cover two networks, they can be extended to support three or more networks by iteratively combining two OSNs. Let $g$ denote the number of network graphs. For the size-constrained scenario, QMSampler first estimates $\sigma^2_1(G_1 \cup G_2)$, $\sigma^2_2(G_1 \cup G_2)$, $\sigma^2_3(G_1 \cup G_2)$, $\mu_1(G_1 \cup G_2)$, $\mu_2(G_1 \cup G_2)$, $\mu_3(G_1 \cup G_2)$, and $N(G_1 \cup G_2)$. Then, it combines the samples of $G_1$ and $G_2$ and treat $G_1$ and $G_2$ as a new combined graph $G_c$, of size $N(G_c) = N(G_1) + N(G_2) - N(G_1 \cap G_2)$, to find the sample mean and variance $\sigma^2(G_c) = \sigma^2_1(G_1 \cup G_2) + \sigma^2_2(G_1 \cup G_2)$, where $\sigma^2_1(G_1 \cup G_2) = \sigma^2_2(G_1 \cup G_2) = \sigma^2_3(G_1 \cup G_2)$ and $\sigma^2_3(G_1 \cup G_2) = \sigma^2_3(G_1 \cup G_2)$ are calculated as shown in Theorem 4. Next, QMSampler combines $G_c$ and $G_3$ in a similar manner to find a new combined graph $G_c$. Notice that the nodes in $G_3$ are matched with the nodes in $G_c = G_1 \cup G_2$. The above procedure is repeated iteratively. Finally, for the last OSN $G_r$ and the merged graph $G_c = G_1 \cup G_2 \cup \ldots \cup G_r$, Equation (26) is used to derive the optimal number of samples $n(G_r)$ of $G_r$. After obtaining $n(G_r)$, QMSampler subtracts $n(G_r)$ from $n$ to find the number of samples $n(G_c)$. Then, $n(G_r)$ is divided into two parts similarly in order to find $n(G_{r-1})$. The above procedure is repeated until $n(G_1)$ is acquired.

**Theorem 12.** The time complexity of QMSampler is $O(g \cdot \Lambda(g \cdot |V_{max}|, |V_{max}|) + O(g^2 \cdot |E_{max}|)$, where $|V_{max}|$ and $|E_{max}|$ are the largest node and edge sizes among $G_1, \ldots, G_r$, respectively, and $\Lambda(x, y)$ is the estimated time required to match two graphs with node sizes $x$ and $y$, respectively.

**Proof.** Let $|V_i|$ and $|E_i|$ denote, respectively, the number of nodes and edges in graph $G_i$, and let $|V_{max}| = \max_{i \in [1, l]} |V_i|$, $|E_{max}| = \max_{i \in [1, l]} |E_i|$. In the worst case, when $n$ is very large, the initial sampling on $G_1$ needs to sample the whole graph of $G_1$. The cost of sampling the nodes and edges is $O(|E_{max}|)$. Then, QMSampler performs an initial sampling on each graph $G_i, 2 \leq i \leq g - 1$. Therefore, the total cost of initial sampling of $G_i$ is $O(g \cdot |E_{max}|)$. After initial sampling of each $G_i$, the matching time of $G_i$ and $G_i$ for each $i$ is estimated as $O(\Lambda(|V_i|, |V_i|))$, because the account matching algorithm in [20] computes the similarity of nodes based on the nodes’ attributes. As $G_c = \bigcup_{1 \leq i \leq g} G_i$, $|V_c| = \sum_{i=1}^{g} |V_i|$ in the worst case, the time complexity of matching $G_c$ and $G_i$ for each $i$ is $O(\Lambda(\sum_{j=1}^{g} |V_j|, |V_i|))$. Because the matching of $G_c$ and $G_i$ is performed once for each $i, i = 2, \ldots, (m - 1)$, the overall time complexity of matching $G_c$ and $G_i$ is $O(g \cdot \Lambda(g \cdot |V_{max}|, |V_{max}|))$. The parameter estimation of $G_i$ and $G_c$ is implemented in $O(1)$ time by looking up the parameters during each iteration of $i$. After the matching of $G_i$ and $G_c$, merging $G_c$ and $G_i$, i.e., $G_c = G_c \cup G_i$, can be implemented in $O(\min(|E_{max}|, |E_i|) + O(\sum_{j=1}^{g} |E_j|))$ time, where $|E_i| = \sum_{j=1}^{g} |E_j|$; therefore, for each $i, G_c = G_c \cup G_i$, costs $O(\min(|E_{max}|, |E_i|) \cdot |V_{max}|)$ time. Because $i$ ranges from $2$ to $m - 1$, the total time complexity of initial sampling of $G_i$, user matching of $G_i$ and $G_c$, and $G_c = G_c \cup G_i$, is $O(\sum_{i=2}^{m} (\min(|E_{max}|, |E_i|) \cdot |V_{max}|))$. Similarly, the initial sampling of $G_n$ and the matching of $G_n$ and $G_c$, costs $O(|V_{max}|)$ time and $O(\Lambda(|V_{max}|, |V_{max}|))$ time, respectively. Overall, the time complexity of the above procedures is $O(g \cdot \Lambda(g \cdot |V_{max}|, |V_{max}|) + O(g^2 \cdot |E_{max}|)$.

Estimating the parameters between $G_c$ and $G_i$ is implemented in $O(1)$ time. For each $i$, computing the optimal number $n_i$ on $G_c$ and $G_i$ for the sample size is implemented in $O(1)$ time. For computing $G_c = G_c \\backslash G_i$, it takes $O(|E_{max}|)$ time to remove the nodes and edges of $G_i$ from $G_c$. Since computing the optimal number $n_i$ on $G_c$ and $G_i$ and computing $G_c = G_c \\backslash G_i$ are performed once for each $G_i, i = g, \ldots, 2$, the time complexity of this step is $O(g \cdot |E_{max}|)$. In summary, the total time complexity of QMSampler is $O(g \cdot \Lambda(g \cdot |V_{max}|, |V_{max}|) + O(g^2 \cdot |E_{max}|)$.

5. EXPERIMENTAL RESULTS

We evaluate QMSampler on 1 user study dataset, 5 real datasets and 2 synthetic datasets. We first recruit 484 students in National Taiwan University for the user study. The ego networks of each user in Facebook and LinkedIn are crawled first. After merging the nodes in the two OSNs corresponding to the same person manually by the users, each user is required to specify additional friends not appearing in her online friend lists in the two OSNs. Afterward, we perform our experiments on the DBLP dataset [29] and the MS dataset from the KDD Cup 2010 [11]. DBLP contains 593,197 authors and 2,805,086 co-author relations with an average degree of 9.5, while the MS database is comprised of 1,731,675 authors and 2,805,086 co-author relations with an average degree of 107.3. The node attributes of each network (e.g., the author name, publication name, and publication date) are employed in relational clustering [10] to merge the two social networks into one network. In total, there are 12,864 pairs of authors matched across the DBLP and MS datasets. We also conduct experiments on two synthetic datasets, called SynDBLP and SynFlickr. The synthetic datasets are generated by the Kronecker graph generator [21] according to the degree distributions of DBLP and Flickr, respectively. SynDBLP and SynFlickr have 300,000 node with the average degrees of 9.6 and 13.4, respectively. In addition, we evaluate QMSampler in Flickr, Foursquare and Twitter for different sampling scenarios. More specifically, we first crawl the node (user id), edge (social relation), and node attributes (user profile) with the APIs provided by each OSN. Then, the nodes of those networks are matched according to [30] to produce a merged network. QMSampler with the user study and crawled datasets are available in [11].

We compare QMSampler with three widely adopted sampling approaches: RW (i.e., random walk), BFS (i.e., breadth-first-search) and MHWR [29] (i.e., an improved random walk with the results similar to the ones in uniform sampling) by evaluating the degree similarity [23], local clustering coefficient and degree distribution of the sampled graphs. We also evaluate the performance of QMSampler on preserving community structures and the global clustering coefficient. In our experiments, QMSampler derives the optimal...
network. With the matching information, it is envisaged that various matched nodes indeed play vital roles in producing a more realistic union of the two networks without matching is 82.3. Therefore, the ratio of the merged network is 95.6. By contrast, the average degree of the merged network because the nodes appearing in the two networks are not matched and the nodes appearing in the two networks are not matched. MHRW outperforms RW and BFS in a network, because MHRW does not evaluate the overlap of multiple networks and thus tends to oversample or undersample a network. The results indicate that it is difficult for traditional sampling schemes to generate the graphs similar to the target when proportions of sampled nodes are small. In contrast, the clustering coefficient of QMSampler is very close to that of DBLP_MS. The clustering coefficient of MHRW is higher than that of DBLP_MS because MHRW does not evaluate the overlap of multiple networks and adjust the number of nodes sampled from each network accordingly. RW and BFS produce much lower clustering coefficients even though they sample many high-degree nodes. For those high-degree nodes with many neighbors in different networks, the probability that most neighbors are connected to each other is inclined to be small.

The results for QMSampler under quality-constrained sampling are presented in Fig. 3. Fig. 3(a) compares the required number of sampled nodes with different quality, where quality is the probability that the difference between the mean of the sampled graph and the original graph is less than a threshold $\epsilon$. As the quality increases, it is necessary to sample more nodes because the high variance of MS forces QMSampler to crawl more nodes for preserving the corresponding properties. In contrast, when the threshold increases in Fig. 3(b) the number of nodes decreases because users are willing to accept a larger error.
As the overlapping ratio of multiple networks plays an important role in generating unbiased sampled networks from multiple networks, here we perform experiments on two synthetic datasets with different overlapping ratios to understand how overlapping ratios impact the performance of different sampling schemes. We generate two synthetic datasets, SynDBLP andSynFlickr, which follows the degree distributions of DBLP andFlickr, respectively. Each synthetic dataset has 300,000 nodes, and we randomly draw 10%, 20%, 30%, and 40% of nodes from each dataset as the overlapping nodes to compare QMSampler and traditional sampling schemes on different measures, i.e., communities (densely connected groups) and global clustering coefficient. Here, we evaluate QMSampler with the size constrained scenario.

Fig. 4(a) and 4(b) compare QMSampler with BFS and MHRW under different overlapping ratios with 30k sampled nodes. Fig. 4(a) shows the sample ratio in QMSampler for different networks. Due to the higher variance in SynFlickr, it is necessary for QMSampler to sample more nodes. As the overlapping ratio increases, the ratio of sampled nodes in SynFlickr grows due to the following reason. When the overlapping ratio increases, the numbers of matched and non-matched sampled nodes within (SynFlickr ∩ SynDBLP) both increase. However, if the total number of sampled nodes is small, e.g., 5% in our setting, the number of non-matched sampled nodes in $G_{QM}$ grows faster than that of the matched ones. Since the overlapping nodes can have edges in both networks, and the average degree of the nodes in SynFlickr is higher than that of SynDBLP, QMSampler draws more nodes from SynFlickr because, as compared to sampling more nodes from SynDBLP, a smaller error is likely to be introduced when an overlapping node in SynFlickr is sampled but not matched with its counterpart in SynDBLP.

According to recent reports, 42% percent of social media users have accounts in multiple OSNs [3], and on average, each user has two accounts in OSNs [4]. Therefore, it is envisaged that sampling multiple networks according to the corresponding network size is not a good way because this approach may deteriorate the sampling quality and cannot preserve the original network properties.

The degree similarities of the sampling approaches are shown in Fig. 4(b). QMSampler significantly outperforms BFS and MHRW because QMSampler constantly monitors the overlapping ratio when performing the sampling task. When the overlapping ratio increases, the degree similarities of BFS and MHRW drop because after matching, the degrees of the overlapping nodes increase significantly, but BFS and MHRW crawl the two networks according to the original degrees. By contrast, QMSampler examines Equation 7 to correctly estimate $N(G_{1} \cap G_{2})$ and adjusts the number of sampled nodes accordingly. Therefore, the degree similarity of QMSampler outperforms the other sampling schemes. It is also worth noting that when the ratio of overlapping nodes increases, the quality of QMSampler, i.e., $Q^{G}$, slightly drops as expected in Equation 7. However, the quality is still sufficiently high (above 0.97). This indicates that QMSampler is able to generate good results with a large number of overlapping nodes.

Fig. 4(c) and 4(d) demonstrate that QMSampler can effectively preserve community structure by presenting the results of community detection [21] on the two synthetic datasets under 30% overlapping ratio. Here, let SynDBLP_SynFlickr denote the merged network. As shown in Fig. 4(c), QMSampler significantly outperforms BFS and MHRW in terms of accuracy because QMSampler minimizes the bias among sampling multiple networks. Moreover, QMSampler in Fig. 4(d) also demonstrates that the number of communities is well preserved. This indicates that QMSampler can produce a small network while preserving communities structures. On the other hand, BFS is not able to effectively preserve the community structures because BFS tends to sample a large number of high-degree nodes. In this case, multiple communities in SynDBLP_SynFlickr tends to be merged into a larger community in BFS. In addition, the community structures are not well preserved by MHRW because it undersamples SynFlickr.

Fig. 4(e) and 4(f) present the results of global clustering coefficients, where the overlapping ratio is 30%. Fig. 4(e) shows the sampling ratios of QMSampler from each dataset. As the total

To measure how community structures are preserved, given the sampled network $G_{S}$, and the ground truth network $G_{T}$, the *truth positive number* $N_{TP}$ is the number of node pairs $(v_{i}, v_{j})$ such that if $(v_{i}, v_{j})$ are in the same community in $G_{S}$, they are also in the same community in $G_{T}$. Similarly, *truth negative number* $N_{TN}$ is the number of node pairs $(v_{i}, v_{j})$ such that if $(v_{i}, v_{j})$ are in different communities in $G_{S}$, they are also in different communities in $G_{T}$. Therefore, the accuracy is thus defined as $\frac{N_{TP} + N_{TN}}{2}$, where $|V_{S}|$ denotes the number of nodes in $G_{S}$.
UNION is the union of the sampled nodes in the profile. The order of the average degrees is Twitter ¿ Flickr count matching algorithm to identify Twitter-Foursquare account Foursquare and Twitter-Foursquare datasets. It is easier for the access coefficient of more inclined to come from Twitter. For node Twitter because QMSampler tries to optimize the sampling quality by QMSampler. Around 50% of sampled nodes mainly come from Twitter ¿ Foursquare ¿ Flickr.

The threshold \( \eta \) is sampled from only one network, the sampling from Twitter. The nodes appearing in all the three OSNs, if node \( i \) is sampled from only one network, the sampling error \( \eta_i \) is smaller when we sample the node from Twitter, instead of Foursquare or Flickr, as explained in Section 5.2. Therefore, QMSampler draws more nodes from Twitter.

Fig. 5(b) presents the degree distributions of QMSampler, MHRW, MHRW_UNION and RW when sampling 1.6 million nodes from all OSNs, where MHRW_UNION is the union of the sampled nodes from the three networks without matching. The degree distributions of the three OSNs are different from those in the DBLP and MS datasets, indicating that there are much more low-degree nodes in the three OSNs. Due to the lack of matching nodes, MHRW_UNION has more low-degree nodes. Compared with QMSampler, MHRW tends to sample more high-degree nodes because it is not aware of the overlap between the three OSNs, and thus does not adjust the number of sampled nodes from each network accordingly. By contrast, RW samples much more high-degree nodes because those nodes are inclined to be visited more frequently in RW.

Fig. 5(c) compares the average clustering coefficients with different sample sizes. As discussed in Section 5.2, RW has the lowest clustering coefficient due to a large number of high-degree nodes. When the number of sampled nodes increases, the clustering coefficients of RW and MHRW become closer to that of QMSampler. It is worth noting that the clustering coefficient of a single network tends to be larger than the one in the graph merged from multiple OSNs, since a node’s neighbors in different OSNs have low chances to become neighbors of each other. Therefore, the clustering coefficient of QMSampler is smaller than MHRW, because QMSampler tends to sample more nodes from the OSNs with more matched accounts, i.e., Twitter and Foursquare, in order to reduce the difference between the sample mean and the mean of the original graph, i.e., \( G \). It is worth noting that, in this case, increasing the number of matched nodes in QMSampler does not introduce bias because QMSampler increases the number of sampled matched nodes only when it can increase the quality.

Figs. 5(d), 5(e), and 5(f) compare QMSampler, RW, and MHRW under time-constrained sampling with 7 days. Based on the access policies and bandwidth requirements of Flickr, Twitter and Foursquare, we were able to extract, respectively, 1,200, 720 and 500 users and their profiles per hour. Fig. 5(d) presents the degree distributions of QMSampler, MHRW and RW, where RW is strongly biased toward high-degree nodes. It is worth noting that the clustering coefficient of a single network tends to be larger than the one in the graph merged from multiple OSNs, since a node’s neighbors in different OSNs have low chances to become neighbors of each other. Therefore, the clustering coefficient of QMSampler is smaller than MHRW, because it is impossible to oversample all three OSNs as proved in Theorem 11. However, after sampling 0.5 days, the networks become larger. QMSampler trims Flickr and Twitter because their high data rate leads to over sampling from these networks. Even though more nodes in Flickr are trimmed, the ratio of the number of trimmed nodes in each network becomes larger. QMSampler trims Flickr and Twitter because their high data rate leads to over sampling from these networks.

### 5.4 Crawling Flickr, Foursquare and Twitter

We also employ QMSampler, MHRW and RW to crawl 1.6 million nodes from Flickr, Twitter and Foursquare. In size-constrained sampling, RW and MHRW sample the same number of nodes from each OSN. The sampled nodes are then matched according to Table 3. The threshold \( \epsilon \) of QMSampler is set as 1.0. There are 9,684, 12,648 and 23,554 pairs of accounts matched across Flickr-Twitter, Flickr-Foursquare and Twitter-Foursquare datasets. It is easier for the account matching algorithm to identify Twitter-Foursquare account pairs because Foursquare allows users to show their Twitter accounts in the profile. The order of the average degrees is Twitter, Flickr, Foursquare, and the order of degree variances of these OSNs is Twitter ¿ Foursquare ¿ Twitter, Flickr.

Fig. 5(a) presents the ratio of nodes sampled from each OSN by QMSampler. Around 50% of sampled nodes mainly come from Twitter because QMSampler tries to optimize the sampling quality by increasing the number of nodes sampled from Twitter. The nodes in Twitter have a higher degree, and the edges of a matched node are more inclined to come from Twitter. For node \( i \) appearing in all the three OSNs, if node \( i \) is sampled from only one network, the sampling error \( \eta_i \) is smaller when we sample the node from Twitter,
need, we have proposed a crawler called the Quality-guaranteed Multi-network Sampler (QMSampler), to crawl and integrate multiple OSNs jointly and systematically. QMSampler is designed to support different crawling scenarios, namely, size-constrained sampling, quality-constrained sampling, and time-constrained sampling with quality guarantees. Our experiment results manifest that current sampling algorithms introduce much more biases than the proposed approach in crawling multiple networks. QMSampler is available as a free download. In our future work, we will provide statistical guarantees on global graph characteristics, such as the average shortest path length between nodes and the diameter.

7. REFERENCES

[1] QMSampler package. http://arbor.ee.ntu.edu.tw/~hhshuai/QMSampler.gz
[2] Stanford Large Network Dataset Collection. http://snap.stanford.edu/data
[3] Statistics from Hashtags.org. http://goo.gl/34Fwg1
[4] Statistics from Radicati Group. http://goo.gl/5t1m2Z
[5] Twitter Statistics. http://www.statisticbrain.com/twitter-statistics/
[6] UCSB Social Network Dataset Collection. http://current.cs.ucsb.edu/socialnets/
[7] N. K. Ahmed, J. Neville, and R. Kompella. Network sampling: from static to streaming graphs. TKDD, 2013.
[8] Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of topological characteristics of huge online social networking services. WWW, 2007.
[9] C. Avin and G. Ercal. On the cover time and mixing time of Random Geometric Graphs. Theor. Comput. Sci., 2007.
[10] I. Bhattacharya and L. Getoor. Collective entity resolution in relational data. TKDD, 2007.
[11] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. SIAM Rev., 2009.
[12] M. D. Domenico, A. Sol-Ribalta, E. Cozzo, M. Kivel, Y. Moreno, M. A. Porter, S. Gmez, and A. Arenas. Mathematical formulation of multilayer networks. Phys. Rev., 2013.
[13] M. Gjoka, M. Kurant, C. Butts, and A. Markopoulou. Walking in facebook: a case study of unbiased sampling of OSNs. Infocom, 2010.
[14] M. Gjoka, M. Sirivianis, A. Markopoulou, and X. Yang. Poking facebook: characterization of OSN applications. WOSN, 2008.
[15] T. Hagedorn. General formulas for solvable sextic equations. J. Algebra, 2000.
[16] M. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork. On near-uniform url sampling. WWW, 2000.
[17] P. Jain, P. Kumaraguru, and A. Joshi. @I seek ‘f.h.me’: identifying users across multiple online social networks. WWW, 2013.
[18] X. Kong, J. Zhang, and P. Yu. Inferring anchor links across heterogeneous social networks. CIKM, 2013.
[19] S. Lee, P. J. Kim, and H. Jeong. Statistical properties of sampled networks. Phys. Rev. E, 2006.
[20] J. Leskovec and C. Faloutsos. Sampling from large graphs. KDD, 2006.
[21] J. Leskovec and C. Faloutsos. Scalable modeling of real graphs using kronecker multiplication. ICML, 2007.
[22] L. Lovasz. Random walks on graphs: a survey. Combinatorics: Paul Erdos is 80, II, 1994.
[23] A. S. Maiya and T. Y. Berger-Wolf. Benefits of bias: towards better characterization of network sampling. KDD, 2011.
[24] O. Mathew, A. Sola, B. Oladiran, and A. Amos. Efficiency of neyman allocation procedure over other allocation procedures in stratified random sampling. AJTAS, 2(5):122-127, 2013.
[25] A. Mislove, M. Marcon, K. Gummad, P. Druschel, and S. Bhattacharjee. Measurement and analysis of online social networks. IMC, 2007.
[26] V. Pan. Solving a polynomial equation: some history and recent progress. SIAM Review, 1997.
[27] H. H. Shuai, D. N. Yang, P. S. Yu, Chih-Ya Shen, and M. S. Chen. Pattern based graph generator. ICDM, 2013.
[28] D. Stutzbach, R. Rejaie, N. Duffield, S. Sen, and W. Willinger. On unbiased sampling for unstructured peer-to-peer networks. IMC, 2006.
[29] J. Tang, A. Fong, B. Wang, and J. Zhang. A unified probabilistic framework for name disambiguation in digital library. TKDE, 2012.
[30] R. Zafarani and H. Liu. Connecting users across social media sites: a behavioral-modeling approach. KDD, 2013.