Estimating the Size and Average Degree of Online Social Networks at the Extreme

Emrah Cem and Kamil Sarac
Department of Computer Science, The University of Texas at Dallas
Richardson, Texas 75080–3021
Email: {emrah.cem, ksarac}@utdallas.edu

Abstract—Given the increasingly limiting nature of online social networks (OSNs), studying their structural characteristics under a limited data access model becomes important. In this study, we propose estimators for network size and average degree characteristics of OSNs. We sample an OSN graph using random neighbor API calls. A random neighbor API call returns only the id of a randomly selected neighbor of a given user. Although the existing estimators give good accuracy estimations for a given sample size, they are not applicable under the extremely limited data access model considered here. We conduct experiments on real world graphs to measure the performance of the proposed estimators.

I. I N T R O D U C T I O N

As the popularity of online social networks (OSNs) increases, estimating their structural properties such as network size and average degree has drawn a lot of attention. There is a conflicting trend on the assumptions in data access models between OSN providers and OSN researchers. Due to privacy and security concerns, OSN providers are imposing increasingly more restrictions on APIs to access data. On the other hand, in an effort to develop better and better estimators, OSN researchers are assuming access to more and more data. In this study, we consider an extremely limited data access model called random neighbor API (RNA) model. In this model, an RNA call returns only the id of a randomly selected neighbor of a given user and no information for other attributes are returned (see Fig. 1). Although the RNA model is more restrictive than the data access models used by current popular OSN applications, studying this model helps us understand the accuracy limitations of the estimators developed for such an extreme scenario.

Random walk sampling [1] is a practical sampling algorithm that is commonly used in social networks. Its theoretical properties are well-known from the theory of Markov chains as a random walk is a finite Markov chain that is time-reversible [1]. There are extensive studies [2], [3], [4], [5], [6], [7] on estimating the characteristics of large graphs from the data collected by a random walk. However, the common assumption in the proposed estimators is that the degrees of vertices visited by the random walk are observable. It is actually a reasonable assumption to be able to perform a random walk. Note that the building block of the random walk sampling is the selection of a random neighbor. In order to perform a random walk, one should either have all the neighbors of the currently visited vertex available so that one neighbor can be selected randomly or have a black-box that gives a random neighbor of the currently visited vertex. While all previous approaches fall into the former case, which implicitly corresponds to knowing vertex degrees of visited vertices, the RNA model falls into the latter case in which the API is the black box. In other words, a random walk can be performed under the RNA model, but the vertex degrees of visited vertices are not observable. Therefore, the proposed estimators are not directly applicable under the RNA model and the main challenge in the RNA model is to make estimation without observing the degrees of the sampled vertices.

In this paper, we study the estimation of structural properties of OSNs, specifically network size and average degree, under the RNA model. We do not aim to propose estimators that outperform random-walk based estimators having access to the degrees of visited vertices as they are not applicable under RNA model. We also develop the theory to establish the relationship between the number of repeated RNA calls needed for a vertex to estimate its degree with a given precision and confidence level, which is important in designing an efficient sampling design.

Our main finding is as follows: when compared to the estimation of network size, the estimation of average degree is expected to be a simpler problem. However, it is not clear whether average degree or network size can be estimated accurately under RNA model with a reasonable sample size. We have conducted experiments on 5 different real-world graphs and found that estimating average degree accurately with a reasonable sample size (<20%) is possible while estimating network size with a reasonable sample size is a hard problem under the RNA model.
II. PRELIMINARIES

A. Terminology

We denote the social network graph by $G = (V, E)$ where $V = \{1, 2, \ldots, N\}$ is the set of vertices/nodes and $E$ is the set of edges. We assume that $G$ is an undirected connected non-bipartite graph\(^1\). We use $d_i$ to denote the degree of vertex $i$. We use $\mu_d = 1/N \sum_{i=1}^{N} d_i$ to denote the average degree.

We use $S^n = \{S_1, S_2, \ldots, S_n\}$ to denote a sample of vertices with size $n$. Probability design of $S^n$ is given by a probability distribution over the set $V^n$ of all ordered sequences of vertices with size $n$ from $V$. We use $f = n/N$ to denote the sampling fraction. We use $M_i$ to denote the multiplicity of vertex $i$, i.e., $M_i = \sum_{j=1}^{N} I\{S_j = i\}$. We say that a collision occurs when $S_i = S_j$ for $1 \leq i < j \leq n$. We use

$$Y_{ij} = \begin{cases} 1 & \text{if } S_i = S_j \\ 0 & \text{otherwise} \end{cases}$$

(1)
as an indicator random variable for a collision between $S_i$ and $S_j$. Finally, we use $C = \sum_{i<j} Y_{ij}$ to denote the total number of collisions in $S^n$.

B. SRW Sampling

SRW can be defined as follows: Given a graph and a starting vertex $v_0 \in V$, we select a neighbor of $v_0$ uniformly at random, i.e., with probability $1/d_{v_0}$, and call it $v_1$. We move to $v_1$ and select a neighbor of $v_1$ uniformly at random and call it $v_2$, and so on. The sequence of selected vertices is called a simple random walk on $G$. Another well-known random walk sampling method is named Metropolis-Hastings Random Walk (MHRW) sampling, which requires the degree of the current vertex at each step as well as the degree of the next proposed vertex. In the RNA model, these are not available and estimating them is not very efficient especially when the proposed vertex is rejected (all RNA calls made to estimate the degree of the proposed vertex will be wasted).

In a random walk, there is a strong dependence on the consecutive draws. Thus, using estimators that assume independent vertex selection may yield inaccurate results if the sample obtained by a random walk is used directly. To overcome this problem, dependency reduction technique named safety margin can be used [8]. In safety margin (with parameter margin), the information obtained from the pairs of nodes $S_k$ and $S_l$ that are close, i.e., $|k - l| \leq \text{margin}$, are ignored and estimator is adjusted accordingly.

C. Network Size Estimation under SRW Sampling

We first introduce Hansen-Hurwitz (HH) estimator. It is a design unbiased estimator for any totals that can be represented as a sum of local values under unequal-probability sampling designs. More formally, let $y_k$ denote the value of interest at $S_k$ and $\pi_k$ denote the selection probability of $S_k$. Let $\tau_y = \frac{1}{\pi} \sum_{k=1}^{n} y_k$. If we let $y_k = 1$ for all vertices, $\hat{\tau}_y$ becomes the network size estimator. However, in order to use HH estimator, we also need $\pi_k$ values to be known. If we perform an SRW, we get $\hat{\pi}_k = d_k/|E|$. Given that both $d_k$ and $|E|$ are unknown, HH estimator is not directly applicable under the RNA model.

We also introduce mark-and-recapture estimator and discuss why it is not directly applicable. To use mark-and-recapture estimator, we first select $k_1$ units from the population uniformly at random without replacement, and then mark and release them to the population. Then, we select $k_2$ units from the same population uniformly at random without replacement. Let $c$ denote the number of marked units selected in the second sample. Then, mark-and-recapture estimator, which is asymptotically unbiased, is given by $\hat{\tau}_y = \frac{1}{n} \sum_{k=1}^{n} y_k / \pi_k$. However, selecting vertices uniformly at random without replacement could be done with MHRW in the RNA model but it is not efficient as stated in Section II-B.

In a recent study, Katzir et al. propose a VCB estimator for network size when the vertices are sampled independently from the stationary distribution of the graph $G$, denoted by $\pi(G)$ (see Algorithm 1). They show that the performance of the proposed estimator under this sampling design is superior to the collision-based estimator under the sampling design that selects vertices independently and uniformly at random. As Algorithm 1 assumes that degrees of sampled vertices are observable (up to a constant factor) and vertices can be sampled independently, it can not be directly used under the RNA model.

D. Average Degree Estimation under SRW Sampling

The generalized estimator [9] is one of the most popular estimators used to estimate average of any characteristic of interest given that the selection probability of each sampled unit is strictly positive and known up to a constant\(^2\). It is defined as

$$\hat{\mu}_y = \frac{\sum_{k=1}^{n} y_k / \pi_k}{\sum_{k=1}^{n} 1 / \pi_k},$$

(3)

\(^2\)An alternative is HH estimator which additionally requires the number of units in the population and the exact value of the selection probability of each sampled unit.
where \( y_k \) is the value of characteristic of interest at \( S_k \) and \( w_k \propto \pi_k \) is the weight of \( S_k \) that is proportional to the selection probability \( \pi_k \). As the selection probability of a vertex is proportional to its degree under SRW sampling, one can use the vertex degree instead of the selection probability \( \pi \) for network size. This estimator requires taking independent samples from the stationary distribution of the underlying graph. This estimator has two requirements: (a) vertices need to be sampled from the stationary distribution of the underlying graph independently and (b) vertex degrees need to be known or estimated up to a constant. We use a stationary SRW and apply the safety margin dependence reduction technique proposed in [8] to meet the first requirement. We use the vertex multiplicity as a vertex characteristic that is proportional to a vertex degree to meet the second requirement. The pseudocode is given in Algorithm 2. Even though the proposed algorithm has \( O(n^2) \) running time complexity, this complexity can easily be reduced to \( O(n) \) by using a similar idea as in [8].

### III. Estimation Under the RNA Model

#### A. Multiplicity-based Network size Estimator

Algorithm 1 is the current state of the art efficient estimator for network size. This estimator requires taking independent samples from the stationary distribution of the underlying graph. This estimator has two requirements: (a) vertices need to be sampled from the stationary distribution of the underlying graph independently and (b) vertex degrees need to be known or estimated up to a constant. We use a stationary SRW and apply the safety margin dependence reduction technique proposed in [8] to meet the first requirement. We use the vertex multiplicity as a vertex characteristic that is proportional to a vertex degree to meet the second requirement. The pseudocode is given in Algorithm 2. Even though the proposed algorithm has \( O(n^2) \) running time complexity, this complexity can easily be reduced to \( O(n) \) by using a similar idea as in [8].

#### B. Average Degree Estimation

Generalized estimator in (4) can not be directly used to estimate the average degree as the degrees of sampled vertices are not observable under the RNA model. However, it is possible to estimate the vertex degrees in the RNA model by repeated RNA calls for the same vertex. We still perform a SRW on the graph except that we make several RNA calls to estimate the degree of the current vertex at each step. After the degree of the current vertex is estimated, another RNA call is made to get the next vertex in the SRW and several RNA calls are made to estimate the degree of that vertex, and so on. We make a fixed number of RNA calls per sampled vertex (and denote it by \( \omega \)) to estimate its degree. If a vertex is selected multiple times, we simply average the estimations for that vertex. One can also perform a random walk with size \( \lceil n/\omega \rceil \) first, and then estimate the degree of each sampled vertex as shown in Algorithm 3. Note that there is a trade-off between the estimation accuracy of vertex degrees and number of vertices visited by SRW as the number of vertices in the SRW will be approximately \( n/\omega \). The larger the \( \omega \) is, the better the vertex degree estimations are, but the less the number of vertices in SRW sample are. We explore this trade-off using simulations in real-world graphs with known average degree in Section IV.

Although we use fixed number of RNA calls for each vertex degree estimation, one important question in this context is how many RNA calls are needed to get degree estimations with a given precision and confidence level. Once this is known and a corresponding sampling scheme under the RNA model is built, one can check if this strategy gives better estimation or not. Intuitively, we know that a vertex with high degree requires more API calls than a vertex with a small degree to get the same precision. In the rest of this section, we develop the theory to establish the relation between the number of RNA calls and the degree for a given precision and confidence level. Once this relation is known, the challenge will be to come up with a practical sampling scheme that utilizes this relation under the RNA model which we leave as a future work.

We make the analysis for an arbitrary vertex \( k \) without loss of generality. We use \( C_k \) to denote the number of collisions in \( \omega \) RNA calls for vertex \( k \). We assume that \( C_k \) follows \( Binomial\left(\binom{n}{\omega}, 1/d_k^2\right) \). The idea is that each ordered pair of sampled vertices have collusion probability of \( 1/d_k^2 \). It is known that there is no unbiased estimator for the reciprocal of a binomial distribution. Fortunately, we can derive an asymptotically unbiased estimator for \( d_k \) as follows:

**Lemma 1.** \( \hat{d}_k = \binom{n}{\omega}/C_k \) is an asymptotically unbiased estimator of \( d_k \).

**Proof.** Using the second order Taylor approximation \( g(X) - g(\mu) \approx g'(\mu)(X - \mu) + \frac{1}{2}g''(\mu)(X - \mu)^2 \) for \( g(x) = 1/x \) and taking the expectation of each side, we derive the approximate
bias of \( \hat{d}_k \) as

\[
B(\hat{d}_k) = \mathbb{E}[\hat{d}_k] - d_k \\
\approx \mathbb{E}[g'(1/d_k)(1 - 1/d_k)] + \frac{1}{2}\mathbb{E}[g''((1/d_k)(1 - 1/d_k)^2)] \\
= -d_k^2\mathbb{E}[1/d_k - 1/d_k^2] + d_k^2\text{Var}(1/d_k) \\
= d_k^2(\omega)^{-1}(1/d_k)(1 - 1/d_k) = 2d_k(d_k - 1)/(\omega(\omega - 1)).
\]

Note that \( \mathbb{E}[1/d_k - 1/d_k] = 0 \) since \( 1/d_k = C_k(\omega)^{-1} \) is an unbiased estimator for the binomial success probability \( 1/d_k \). As \( \lim_{\omega \to \infty} B(\hat{d}_k) = 0 \), \( \hat{d}_k \) is an asymptotically unbiased estimator for \( d_k \).

Following from Lemma 1 and skipping several intermediate steps (e.g., derivation of variance and confidence estimator for \( d_k \) and confidence interval for \( d_k \)), we derive the minimum required \( \omega \) to estimate \( d_k \) within a given relative error \( \epsilon \) and confidence level \( 100(1 - \alpha)\% \) as

\[
\omega \approx 1 + \frac{1}{\sqrt{1 + \frac{8z_{1-\alpha/2}^2(d_k - 1)}{\epsilon^2}}},
\]

(5)

where \( z_{1-\alpha/2} \) is the \( 100(1 - \alpha/2) \) percentile of the standard normal distribution. Although \( \hat{d}_k \) has a nice asymptotic property, it is biased for small sample sizes and is undefined when \( C_k = 0 \). To reduce the bias for small sample sizes, we instead use \( \hat{d}_k^{adj} = \left(\frac{\omega}{2}\right)(C + 1)^{-1} \), a slightly modified version of \( \hat{d}_k \), and call it adjusted estimator. The similar modification was also proposed for mark-and-recapture estimator by Chapman [10]. Note that asymptotic properties of \( \hat{d}_k \) and \( \hat{d}_k^{adj} \) are the same but we get better estimations for small sample sizes with \( \hat{d}_k^{adj} \).

Fig. 4a and 4b illustrate the relation between degree, precision, and confidence level obtained in (5). Fig. 4a illustrates the required number of RNA calls to get an estimate within a desired relative error with 90% confidence level. Fig. 4b illustrates the required number of API calls to get an estimate with a desired confidence level within \( \epsilon = 0.1 \) relative error.

The results support the intuition that as the degree of the vertex increases, the required number of RNA calls also increases to get the same precision and confidence level. Fig. 5 illustrates the mean and coefficient of variance of 10,000 independent degree estimations using fixed number of RNA calls for different vertex degrees. As the vertex degree increases, negative bias is observed for small sample sizes. This observation can be explained by the high likelihood of no collision when \( \omega \) is much less than the population degree of a vertex. When there is no collision, adjusted estimator gives \( \hat{d}_k \) and as \( \omega \) is small an underestimation is obtained. Fig. 5 also shows the normalized root mean square error (NRMSE) of 10,000 estimations for various degrees with varying \( \omega \).

IV. EXPERIMENTAL EVALUATIONS

In this section, we demonstrate the accuracy of our proposed estimators for network size and average degree using publicly available real-world graphs shown in Table I. We extract and use only the largest connected component of each graph. We run 500 experiments for each 90% Monte Carlo confidence intervals that we use to show the performance of the estimators. The y-axis in our figures shows the estimates normalized by the true parameter.

A. Network Size Estimation

For network size estimation, we plot Monte Carlo confidence intervals for various sampling fractions \((0.01 \leq f \leq 50)\). As other estimators are not directly applicable under the RNA model, we compare the accuracy of MB-VCB estimator with the VCB estimator using the sample obtained by SRW. Note that in this version of VCB, independent vertex selection is not possible due to the limitations of the RNA model. Note also that for a given \( f \) value, the VCB estimator is
expected to give much better performance as compared to MB-VCB estimator. Yet, this comparison demonstrates the effect of using vertex multiplicities when vertex degree information is not available.

Fig. 6 illustrates the estimation results for MB-VCB and VCB estimators. MB-VCB estimator underestimates the network size in all cases regardless of margin parameter for \( f < 10 \). The main reason of this observation is that multiplicity values of vertices do not become proportional to the vertex degree until we have a large sample size. For the margin = 1 case (top row in Fig. 6), small margin value may not also be enough to reduce the dependency of the data. When the margin parameter is increased, we observe that for most of the cases, confidence intervals cover higher values for small sample sizes even though the underestimation is still the case. This illustrates the trade-off between reducing the dependency and utilizing less data (bias-variance trade-off). If we ignore too much data, the precision becomes low but sample becomes more independent. On the other hand, if we do not ignore enough data, the precision becomes high, but sample becomes more dependent and, hence, estimation becomes biased. Another observation is that for networks with smaller density (e.g., GNUTELLA, CONDMAT), effect of margin parameter is more significant. This can be partially explained with the fact that the random walk selects nodes with less dependence as the network density increases. This becomes more clear when we consider a random walk in a complete graph: SRW becomes equivalent to the vertex sampling method selecting vertices independently and uniformly at random.

With VCB estimator, the bias-variance trade-off is also clear for small sample sizes. When margin = 1 (the top row in Fig. 6), we observe an underestimation with high precision for small sample sizes (\( f < 0.1 \)) for all graphs. When margin = 10 (the second row), we observe that estimations are not very precise but are around the true parameter. When margin = 100 (the bottom row), we observe much less precision due to ignoring too much data. In these graphs, margin = 10 seems to be better than others. We conjecture that this is directly related to the average degree of considered real-world graphs.

B. Average Degree Estimation

As the average degree estimation is more accurate for small sample sizes, we plot the confidence intervals only for sampling fractions \( f \leq 1 \) (see Fig. 7). We compare GEN-AD estimator with a generalized estimator (GEN) that works with actual degrees of the sampled nodes. To observe the effect of uncertainty in degree estimations, we use the same sample (of size \( \lfloor n/w \rfloor \)) returned by SRW for both estimators. Note that if we use the sample size \( n \) for GEN estimator, it will outperform GEN-AD but this case will not serve for the purpose of comparison. We use two different values for \( \omega \) as \( \omega = 10 \) and \( \omega = 100 \) to illustrate the trade-off between the accuracy of vertex degree estimations and the number of vertices obtained by SRW and to illustrate their effect on the average degree estimation.

According to results in Fig. 7, accuracy results for average degree estimation is better than that of network size estimation. One of the reasons for this result is that GEN estimator does not require independent vertex samples as opposed to the VCB estimator. Another observation from Fig. 7 is that sampling with \( \omega = 10 \) (top row) gives better results than the sampling with \( \omega = 100 \) (bottom row) in our experiments. According to Table I, the average degrees of vertices in the considered real-world graphs are no more than 26. Therefore, the case with \( \omega = 100 \) unnecessarily consumes large number of RNA calls on selected vertices resulting in fewer vertices to be sampled by SRW (recall that we have a fixed RNA call budget of \( n \)). This is also the reason for observing a similar performance between GEN and GEN-AD for \( \omega = 100 \).

C. Discussion

One of the assumptions of our estimators is that the sample is collected by a stationary random walk. In practice, this can
be achieved by ignoring a number of initial steps in random walk (known as burn-in period). Note that burn-in period is relevant for our SRW-based sample collection and is ignored in our experiments.

To estimate the network size, one can be tempted to save unique vertex ids visited by SRW until no new ids are obtained instead of using the multiplicity based estimator which requires high sample sizes ($f \approx 10$). However, this idea has much higher cost ($(1 - o(1))\log n < f < O(4m)$) for general undirected and non-bipartite graphs\(^4\) [1], therefore, we do not consider a complete topology discovery as a viable alternative.

V. RELATED WORK

Estimation of network size based on non-uniform samples has been the subject of recent studies [3], [5], [13]. Another recent study [8] has provided various estimators for network size under both independent samples and random walk-based samples as well as dependence reduction techniques for the random walk.

Estimation of average degree has been considered extensively in the literature [14], [15], [16]. A recent study [2] has given a comprehensive overview of different types of average degree estimators and proposed a novel estimator under a prescribed distribution. They also show how to adapt the estimator in a random walk setting.

The main difference between the above mentioned related work and the study presented here is that previous studies consider estimation of average degree and network size under various sampling designs but none of them are directly applicable under the RNA model as they require either vertex degree information or independent vertex selection.

VI. CONCLUSIONS AND FUTURE DIRECTIONS

In this paper, we have proposed estimators for a network size and average degree characteristics under the RNA model. Using simulations on real-world graphs, we have demonstrated that network size estimation requires high sampling fractions ($f > 10$) under the RNA model. However, the average degree can be estimated with practical sample sizes ($f < 0.3$).

\(^4\)Time to visit all vertices in a graph with a SRW is also known as the cover time of the graph.

In our future work, we plan to develop more efficient network size estimators under the RNA model. Estimating the number of edges in the graph and the sampled vertex degrees (as we did for average degree estimation) allow us to utilize the HH estimator in network size estimation. Depending on the estimation accuracy of number of edges and vertex degrees, HH estimator may outperform MB-VCB estimator. One of the key challenges in practice is to determine the sample size without knowing the true network size. Finally, we plan to study the estimation of other characteristics such as clustering coefficient.

REFERENCES