Estimating Sizes of Social Networks via Biased Sampling*

Liran Katzir[†], Edo Liberty[‡], Oren Somekh[§], Ioana A. Cosma[¶]

ABSTRACT

The paper presents algorithms for estimating the number of users in online social networks. While such networks sometimes publish such statistics, there are good reasons to validate their reports. The proposed schemes can also estimate the cardinality of network sub-populations. Since this information is seldom voluntarily divulged, such algorithms must operate only by interacting with the social networks' public APIs. No other external information can be assumed. Due to obvious traffic and privacy concerns, the number of such interactions is severely limited. We therefore focus on minimizing the number of API interactions needed for producing good size estimates.

We adopt the standard abstraction of social networks as undirected graphs and perform random walk based node sampling. By counting the number of collisions or non-unique nodes in the sample, we produce a size estimate. Then, we show analytically that the estimate error vanishes with high probability for fewer samples than those required by prior-art algorithms. Moreover, although provably correct for any graph, our algorithms excel when applied to social network-like graphs. The proposed algorithms were evaluated on synthetic and real social networks such as Facebook, IMDB, and DBLP. Our experiments corroborate the theoretical results, and demonstrate the effectiveness of the algorithms.

Keywords

Population Estimator, Sampling, Undirected Graph, Social Network

1. INTRODUCTION

Online social networks have become increasingly popular in the recent decade which gave rise to an increasing need in analyzing their properties and comparing them to one another. Many properties of online social networks are considered important. These include, for example: their user age distribution, net activity, connectivity, and many more. The literature on attaining such parameters is vast and any effort to reference all of it is bound to fail. We cite here only a handful of directly related works as examples. In [KNV06, HTV$^+$09] the authors present a way to detect proximity between two users. In [YSH$^+$07] the authors analyze the degree distribution, clustering property, degree correlation, and evolution over time for three networks. That said, the total number of users (or users in a certain demographic) seems to be one of the most crucial factors in deriving the worth and overall performance of social networks. These figures are also critical for business development issues, choosing between networks for advertisement campaigns or for launching social applications. Although countless sites, blogs, and other reports often present such numbers, these are usually based on reports of the social networks themselves or on traffic analysis and are not guaranteed to be accurate (e.g. [FBS]).[1] Moreover, since each network reports slightly different figures, it is almost impossible to compare between them. For example, Facebook reported lately on crossing the 500 million active users mark. However, it is unclear what constitutes an "active user". Thus, it is extremely important to be able to accurately and reliably estimate the size of social networks in a unified and unbiased way (without the networks' consent or control).

Since online social networks provide public interfaces, it is possible to traverse their members' network externally. By "crawling" the network, we can collect statistics on any of the above characteristics. In a similar spirit, in [BG07, BG08] the authors suggested ways to measure various parameters of search engines by interacting with their public search interface only. One approach, is to crawl the network extensively. In other words, since online social networks can be modeled as undirected unweighted graphs (where the users are nodes and edges are connections/friendships) we can perform a *breadth-first search* (BFS) on the graph.[2] This method is impractical when dealing with online social networks since the communication and computational burden of such an undertaking would probably be prohibitive.

The second approach is to sample users uniformly at random from the network. From a uniform sample, most statistics can be estimated. This includes estimating the size

[†]Microsoft Innovations Lab, Israel, lirank@microsoft.com. The work was done while the author was with Yahoo! Labs, Haifa, Israel

[‡]Yahoo! Labs, Haifa, Israel, edo.liberty@ymail.com

[§]Yahoo! Labs, Haifa, Israel, orens@yahoo-inc.com

[¶]Department of Mathematics and Statistics, University of Ottawa, icosma@uottawa.ca

*This work was presented in part in the 2012 World Wide Web Conference (WWW'12), Lyon, France.

[1]Moreover, the published statistics do not provide an estimate for connected sub-graphs, e.g., 20-30 year olds living at the US.

[2]Note that online social networks' public APIs provide lists of connected users for every user. Thus, acting like a neighbor list representation of the graph.

of the network using methods such as *mark-and-recapture*. These methods have the advantage of requiring only $O(\sqrt{n})$ users to be sampled to get a good estimate (where n is the overall number of users) [LP56]. Their disadvantage though, is that users must be sampled uniformly. Since online social networks interfaces usually do not provide this functionality, it must be simulated by other API queries which give for each user the list of their neighbors. Alas, producing a single uniformly chosen user might require many such queries. This is explained in the next section.

In this work we show that there is no need to sample users uniformly. In fact, by using the sampling bias we reduce the number of required samples dramatically. For example, for networks with a Zipfian-like degree distribution our algorithms require only $O(n^{1/4}\log(n))$ samples to converge. Moreover, since the bias does not have to be corrected, each such sample requires considerably less API queries. Surprisingly, our algorithm is extremely simple and gives provable error guarantees with high probability.

Experiments with our algorithm performed over a wide range of real and synthetic data corroborate that it converges significantly faster and to a more accurate estimate than uniform sampling based approaches.

As a side note, simple variants of our algorithm also give efficient sub-linear algorithm for estimating the size of transitive closures in graphs and for estimating the size of search-indexes as in [BG07]. However, this is a matter of farther research and is beyond the scope of this paper.

The rest of this work is organized as follows. Section 2 surveys related work. Our algorithms are presented and analyzed in Sections 3 and 4. In Section 5 we report our experimental results and conclude in Section 6. Various proofs and discussions are included in the Appendix.

2. BACKGROUND AND PRIOR WORK

From this point on we consider the general problem of estimating the size of undirected graphs. The graph representation of online social networks is the obvious one. Each node refers to one user and an edge is present between two nodes if their corresponding users are "friends" in the social network. Although our algorithms are correct for general graphs, they are especially suited for graphs which naturally occur in large social networks.

In [MV06] the authors provide a possible solution for another problem which could be used to solve the problem at hand as well. They present an algorithm for estimating the number of attributes in a database. The algorithm samples rows from the database uniformly at random. It then estimates the total number of attributes using the collected information of how many times each attribute was picked. This is identical to a well known problem in statistics called "estimating the number of unseen types". Their algorithm can be applied to estimating the size of graphs if the graph is represented as a database table containing two rows per edge, one for each adjacent node.[3] Clearly, the number of distinct attributes in this database is n, the number of nodes. The algorithm presented in [MV06] is guaranteed to take at most $r = O(n)$ samples. However, in graphs, unlike in databases, it is possible to sample nodes (analogously, attributes) uniformly at random which can dramatically decrease the number of samples.

In ecology, a method known as *mark and recapture* is used to estimate population sizes.[4] It relies on the same phenomenon as the so called "birthday paradox" effect. Informally, after sampling r nodes uniformly at random we expect to encounter $C \approx r^2/2n$ collisions (nodes already picked). Thus, n can be estimated by $r^2/2C$. Surprisingly, taking only $O(\sqrt{n})$ samples can guarantee that this estimate for n is rather accurate.[5] In [FTV98] the authors present a maximum likelihood estimator for this problem and show that their estimator converges almost surely when the number of samples increases. In [Cha87, HYCY03] the authors extend these methods to non-uniform, but known, distributions.

That said, to use these methods one must sample nodes uniformly at random from a graph, which is not straight forward. To see how this can be done we remind the reader a few basic facts from spectral graph theory. A random walk on an undirected graph with n nodes $\{v_1, \ldots, v_n\}$ is defined as such: start from an arbitrary node, then move to a neighboring node uniformly at random and repeat. After many such steps, the probability of being at any node v_i is close to $p_i = d_i/D$ where d_i is the degree of node v_i and $D = \sum_{i=1}^{n} d_i$ is the sum of all node degrees in the graph. This is called the stationary distribution of the random walk on the graph. The number of random walk steps needed for the stationary distribution to be reached depends on the mixing rate property of the graph (see a survey by Lovász [Lov93]). Fortunately, social network graphs and small world graphs are known to have good mixing rates. We therefore assume that nodes can be repeatedly sampled from the stationary distribution without much computational overhead.

Using these properties of random walks, one can sample nodes also uniformly at random by using, for example, rejection sampling. To be precise, a node v_i is first sampled according to the stationary distribution. Then, with probability $1/d_i$ it is kept. With probability $1 - 1/d_i$ it is rejected. Clearly the set of kept nodes is uniformly sampled. However, since we only expect to accept a node with probability n/D, to sample $\Omega(\sqrt{n})$ un-rejected nodes would require an expected $r = \Omega(D/\sqrt{n})$ biased samples. Several rejection sampling ideas and other methods for turning the node sampling distribution to uniform were suggested for specific graphs. Namely, the bipartite graph between search queries and search results [BB98, BG07].[6] Another approach was considered in [GKBM10]. The authors present a modified Metropolis-Hastings random walk which transitions from node v_i to an adjacent node v_j with probability $1/\max(d_i, d_j)$. With the remaining probability, it stays in v_i. Due the symmetry in the transition probabilities it can be shown that the stationary distribution of this walk is uniform on the nodes. However, the mixing rate of this walk can be significantly worse than that of the original graph, and so, it is unclear when it is expected to outperform rejection sampling, i.e., require fewer random walk steps.

An interesting tangentially related problem is known as the "German tank problem" [Joh]. It was supposedly used

[3]Sampling uniformly from this table is possible in our setting since random walks on graphs sample edges uniformly.

[4]Other names for this method or closely related ones, include capture-recapture, capture-mark-recapture, mark-recapture, and mark-release-recapture.

[5]We make a more general statement later in this paper.

[6]In fact, the authors try to compare between two different search services but their approach is suitable for this task as well.

during world-war II to estimate the number of German tanks based on manufacturing numbers found on those captured by the allied forces. In its mathematical formulation, elements with serial ID's are sampled uniformly without replacement and the objective is to provide an estimate for the total number of elements. This is not applicable to our scenario since the users do not have serially allocated and publicly available ID's.

Estimating the number of nodes in a graph was also studied. In [Knu74] the authors estimate the size of a tree. Their motivation was to estimate the running time of a backtracking program. Later [Pit87] extends their argument to acyclic graphs. Finally [MS89] extends this idea to general undirected graphs. However, the running time of the the latter is unbounded in the worst case and expected to be more than the number of nodes in the graph. Recently, in [YW10] the authors try to estimate the size of social networks in a setup very similar to ours. However, they either require that the users be sampled uniformly or use the algorithm from [MS89] which their experiments show is impractical.

Most similar to our approach is [HRH09] in which the authors exploit random walks properties to compute various network properties (such as, average clustering coefficient, degree distribution, degree correlation, and network size). This, by approximating node degree distributions and collision counting. While their approach is similar, it is less straight-forward and it does not provide exact approximation guarantees (algorithms are corroborated only by simulation results).

3. COLLISION COUNTING

In this section we present our graph size estimator. We start by taking r samples $\{x_1, \ldots, x_r\}$ independently from the stationary distribution of the graph, i.e., node i whose degree is d_i is sampled with probability $p_i = d_i/D$ where $D = \sum_{i=1}^{n} d_i$. More formally $\forall i, j \quad \Pr[x_j = i] = d_i/D$ and independently of all $x_{j'}$ for $j' \neq j$.

We define three variables that the algorithm keeps track of: (a) The sum of all sampled node degrees $\psi_1 \triangleq \sum_{j=1}^{r} d_{x_j}$; (b) The sum of reciprocal sampled degrees $\psi_{-1} \triangleq \sum_{j=1}^{r} 1/d_{x_j}$; and (c) Twice the number of collisions $C \triangleq \sum_{j \neq j'} Y_{j,j'}$ where $Y_{j,j'} = Y_{j',j} = 1$ if $x_j = x_{j'}$ and 0 else. Using those we define an auxiliary variable

$$R \triangleq \psi_1 \psi_{-1} - r \ .$$

The proposed estimator $\hat{n}$ for the number of nodes in the graph n is

$$\hat{n} \triangleq R/C \ . \tag{1}$$

To see why this makes sense we start by computing $\mathbb{E}[R]$ and $\mathbb{E}[C]$. Using linearity of expectation and the fact that the samples are taken independently we obtain that

$$\begin{aligned}
\mathbb{E}[R] &= \mathbb{E}\left[\sum_{j \neq j'} \frac{d_{x_j}}{d_{x_{j'}}}\right] = 2\binom{r}{2}\mathbb{E}_{j \neq j'}\left[\frac{d_{x_j}}{d_{x_{j'}}}\right] \\
&= 2\binom{r}{2}\mathbb{E}_j[d_{x_j}] \cdot \mathbb{E}_{j'}[1/d_{x_{j'}}] = 2\binom{r}{2} n \sum_{i=1}^{n} p_i^2 \\
\mathbb{E}[C] &= \mathbb{E}[\sum_{j \neq j'} Y_{j,j'}] = 2\binom{r}{2}\mathbb{E}[Y_{j,j'}] = 2\binom{r}{2}\sum_{i=1}^{n} p_i^2 \ .
\end{aligned}$$

Dividing these two expression, we get that the number of nodes is $n = \mathbb{E}[R]/\mathbb{E}[C]$. Thus, if both R and C are close to their expected values then intuitively $\hat{n}$ should also be close to n. The corollary below makes this exact.

COROLLARY 1. *For any degree distribution and C, R defined as above the estimator $\hat{n}$ guarantees for any $\varepsilon \in (0, 1]$ and $\delta \in (0, 1]$:*

$$\Pr(|\hat{n} - n| \geq \varepsilon n) \leq \delta$$

as long as the number of samples, r, satisfies:

$$r \geq r_c \triangleq 1 + \frac{63}{\varepsilon^2 \delta} \cdot \frac{D}{\min(\sqrt{\sum d_i^2}, n)}$$

PROOF. If it is the case that $|R - \mathbb{E}[R]| \leq \varepsilon \mathbb{E}[R]/3$ and that $|C - \mathbb{E}[C]| \leq \varepsilon \mathbb{E}[C]/3$ then:

$$(1-\varepsilon)n \leq \frac{(1-\varepsilon/3)}{(1+\varepsilon/3)}\frac{\mathbb{E}[R]}{\mathbb{E}[C]} \leq \frac{R}{C} \leq \frac{(1+\varepsilon/3)}{(1-\varepsilon/3)}\frac{\mathbb{E}[R]}{\mathbb{E}[C]} \leq (1+\varepsilon)n \ .$$

Since $\hat{n} = R/C$ this gives the desired result that $|\hat{n} - n| \leq \varepsilon n$. In order for both R and C to be close to their expectation with probability at least $1 - \delta$ it is sufficient invoke the union bound and require

$$\Pr(|R - \mathbb{E}[R]| \geq \varepsilon \mathbb{E}[R]/3) + \Pr(|C - \mathbb{E}[C]| \geq \varepsilon \mathbb{E}[C]/3) \leq \delta$$

Invoking Chebyshev's inequality (twice) these are obtained if the following holds

$$\frac{Var(R)}{\mathbb{E}^2[R]} + \frac{Var(C)}{\mathbb{E}^2[C]} \leq \varepsilon^2 \delta / 9 \ .$$

Since both of these quantities are decreasing in the number of samples r we seek the minimal number r which satisfies the conditions. In Appendix A we calculate the variances of both R and C and show that:

$$\begin{aligned}
\frac{Var(R)}{\mathbb{E}^2[R]} &\leq \frac{1}{r(r-1)} + \frac{ab}{r(r-1)} + \frac{a}{r} + \frac{b}{r} + \frac{2}{r} \\
\frac{Var(C)}{\mathbb{E}^2[C]} &\leq \frac{a^2}{r(r-1)} + \frac{2a}{r}
\end{aligned}$$

where $a = 1/\sqrt{\sum p_i^2}$ and $b = (\sum_{i=1}^{n} 1/p_i)/n^2$. It is easy to verify that if $r - 1 \geq 7c \max\{a, b\}$ then $\frac{Var(R)}{\mathbb{E}^2[R]} + \frac{Var(C)}{\mathbb{E}^2[C]} \leq 1/c$. Setting $c = 9/\varepsilon^2\delta$ completes the proof. □

To understand the bound better, notice that $a = 1/\sqrt{\sum p_i^2}$ $= D/\sqrt{\sum d_i^2} \leq \sqrt{n}$. This is tight when the node degrees are all equal to each other. In fact, when the node degree distribution is uniform, r_c is of order $O(\sqrt{n})$, which agrees with the bound derived from the maximum likelihood estimator of n under uniform sampling. However, when there is a heavy bias in node degree, as in social networks, this term is significantly smaller. Moreover, since we assume there are no zero degree nodes we have that $b = (\sum_{i=1}^{n} 1/p_i)/n^2 \leq D/n$. This is the mean degree of a nodes in the graph which is usually small in social network graphs. The reader should note that in most real life graphs and networks $\sqrt{\sum d_i^2} < n$. We give an example below.

3.1 Performance for online social networks

In order to argue that our algorithm is suitable for sizing social networks we have to assume something about their node degree distributions. In [MMG+07], [YSH+07] and in [GKBM10] the authors argue that, in several networks, the

nodes' degrees exhibits different kinds of heavy tail distributions. Mainly: Exponential, Double-Pareto and Zipfian. Here we analyze, as an example, the Zipfian distribution. Similar analyses can be performed for the other distributions as well.

If the nodes' degrees are distributed according to a Zipfian distribution with maximum degree of d_m and parameter $\alpha = 2$ we have:

$$Pr(d = j) = \frac{j^{-2}}{H} \quad ; \quad j = 1, \ldots, d_m \ ,$$

where $H = \sum_{j=1}^{d_m} j^{-2} \approx \frac{\pi^2}{6}$ and $1 \ll d_m = \Theta(\sqrt{n})$. The ℓ'th moment of the degree distribution is defined as $\mathcal{M}_\ell = \mathbb{E}[d^\ell]$ and the first few moments of the Zipfian distribution are:

$$\mathcal{M}_{-1} \leq \frac{1}{H}; \quad \mathcal{M}_1 \approx \frac{\log d_m}{H}; \quad \mathcal{M}_2 \approx \frac{d_m}{H} \ .$$

We also assume that the moments of the observed degree distribution are close to those of the generating distribution. This is true for large graphs by the strong law of large numbers (SLLN). This gives us that $\sum_{i=1}^{n} p_i^\ell \approx \frac{\mathcal{M}_\ell}{n^{\ell-1}(\mathcal{M}_1)^\ell}$. Substituting the above into Corollary 1 and using the fact that $d_m = \Theta(\sqrt{n})$ we get that $r_c \in O(n^{1/4}\log(n))$. Therefore, only $O(n^{1/4}\log(n))$ samples suffice for our estimator to be accurate. Note the significant reduction in the number of samples over the uniform distribution. For example, for $n = 10^9$, $\sqrt{n} \approx 30,000$ while $n^{1/4}\log(n) \approx 6000$.

3.2 Subgraph size estimation

One surprising aspect of this estimator is that it works for subgraphs as well. Let X' be the subset of samples X who are also in the subgraph. We perform the same random walk and compute the same parameters C', Ψ'_1, Ψ'_{-1}, and R', which are defined as above but for X' instead of X. The subgraph size is estimated by R'/C'. The proof provided above works for this case as well. The only change is that D is replaced by D' which is the sum of node degrees in the subgraph. But, since D (and therefore D' as well) cancels itself in the analysis our estimator remains unchanged.

That said, it is more efficient to first estimate the size of the entire graph and then estimate the subgraphs' relative size. Formally, $\Psi'_{-1} = \sum_{i=1}^{r} 1/d_{x_i} I_{x_i \in V'}$ where $I_{x_i \in V'}$ is equal to 1 if x_i is a node in the subgraph and 0 otherwise. From the above we have that $\mathbb{E}[\Psi_{-1}] = rn/D$, similarly for the subgraph, $\mathbb{E}[\Psi'_{-1}] = rn'/D$ (n' being the number of nodes in the subgraph). Isolating n' we get:

$$n' = n\frac{r}{r}\frac{D}{D}\frac{\mathbb{E}[\Psi'_{-1}]}{\mathbb{E}[\Psi_{-1}]} \approx n\frac{\Psi'_{-1}}{\Psi_{-1}}$$

If the number of samples is large enough, the last step is correct since $\mathbb{E}[\Psi'_{-1}] \approx \Psi'_{-1}$ and since $\mathbb{E}[\Psi_{-1}] \approx \Psi_{-1}$. Under most conditions, the ratio estimator requires only a constant number of samples to converge. Thus, the main computational effort is in estimating n, which is surprisingly lower than that of directly estimating n'. To see this, let r_c and r'_c be the number of samples needed to estimate the sizes of the graph and the subgraph respectively. Since, in a random walk we only hit a node in the subgraph with probability D'/D, we are expected to require Dr'_c/D' random samples to obtain r'_c samples from the subgraph. Thus, if $r'_c/D' \geq r_c/D$ the second method is preferable. Note that this holds in the natural situation that the nodes' degree distributions of the graph and subgraph are similar.

4. NON-UNIQUE ELEMENT COUNTING ESTIMATOR

In this section we present another estimator which is based on counting non-unique elements instead of collisions. On the one hand, it tends to be consistently, yet marginally, more accurate. On the other hand, its proof is much more involved. We thus choose to present the estimator along with its performance (in the experimental results section) without providing a proof of its correctness.

An element in the sample is considered non-unique if it was sampled at least once before. This is slightly different from counting collisions. For example, in the sequence $\{1, 2, 3, 1, 1\}$ there are two non-unique elements (the last two 1's) but three collisions ($x_1 = x_4$, $x_1 = x_5$, and $x_4 = x_5$). The intuition is that counting non-unique elements is less sensitive to errors in which a specific node is oversampled. This is because the non-unique count is linear in the number of times each item was sampled whereas the collision count is quadratic.

We estimate n by $\tilde{n}$ which is the unique solution to the following fixed point equation:

$$\tilde{n} = r - \tilde{C} + \frac{\tilde{n}}{\Psi_{-1}} \sum_{i=1}^{r} \frac{1}{d_{x_i}} \left(1 - \frac{d_{x_i}\Psi_{-1}}{\tilde{n}r}\right)^r \qquad (2)$$

Note that r, $\tilde{C}$, Ψ_{-1} and d_{x_i} are all observed quantities. To see why this is correct, first note that the expected number of non-unique elements is $\mathbb{E}[\tilde{C}] = r - n + \sum_{i=1}^{n}(1-p_i)^r$. Now, consider that

$$\sum_{i=1}^{n}(1-p_i)^r = \mathbb{E}[\frac{1}{p_i}(1-p_i)^r] \approx \frac{1}{r}\sum_{i=1}^{r}\frac{D}{d_{x_i}}(1-\frac{d_{x_i}}{D})^r \ .$$

Also, $D \approx \frac{rn}{\Psi_{-1}}$. Making these substitutions into the expectation expression gives the above fixed point equation. Intuitively, the size estimate $\tilde{n}$ is chosen such that the observed number of non-unique elements is equal to its expectation. As a remark, if the node distribution is uniform, this estimator is identical to the maximum likelihood estimator [FTV98].

5. EXPERIMENTAL EVALUATION

5.1 Networks of known sizes

In order to test our estimators' accuracy we first experimented with three networks whose exact sizes are known.

A synthetic network; a synthetically constructed graph consisting of 1 million nodes whose degree distribution is Zipfian with parameter $\alpha = 2$ and maximal degree $d_m = 1,000$.

The Digital Bibliography and Library Project; we used the Digital Bibliography and Library Project's (DBLP) entire database.[7] Edges in the graph were associated with co-authorship of at least one paper. The resulting graph included 845,211 nodes, each with at least one edge (authors with no co-authors were omitted).

The Internet Movie Database; we useed public Internet Movie Database's (IMDB) entire database.[8] Edge connec-

[7]The DBLP database can be found at http://dblp.uni-trier.de/xml/.

[8]The IMDB database can be found at ftp://ftp.fu-berlin.de/pub/misc/movies/database/.

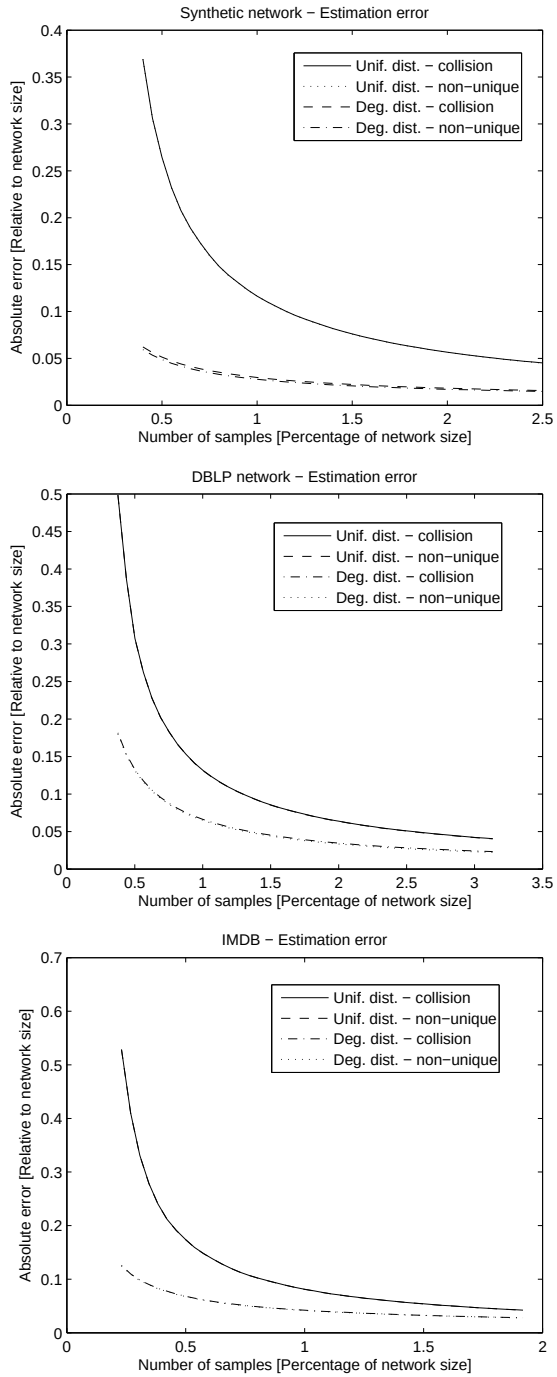

Figure 1: *Error curves* - absolute normalized size estimation errors vs. the percent of sampled nodes for three networks: a 1-million-node synthetic network (top), a network constructed from the Digital Bibliography and Library Project (DBLP) database (middle), and a network obtained from the Internet Movie Data Base (IMDB) database (bottom).

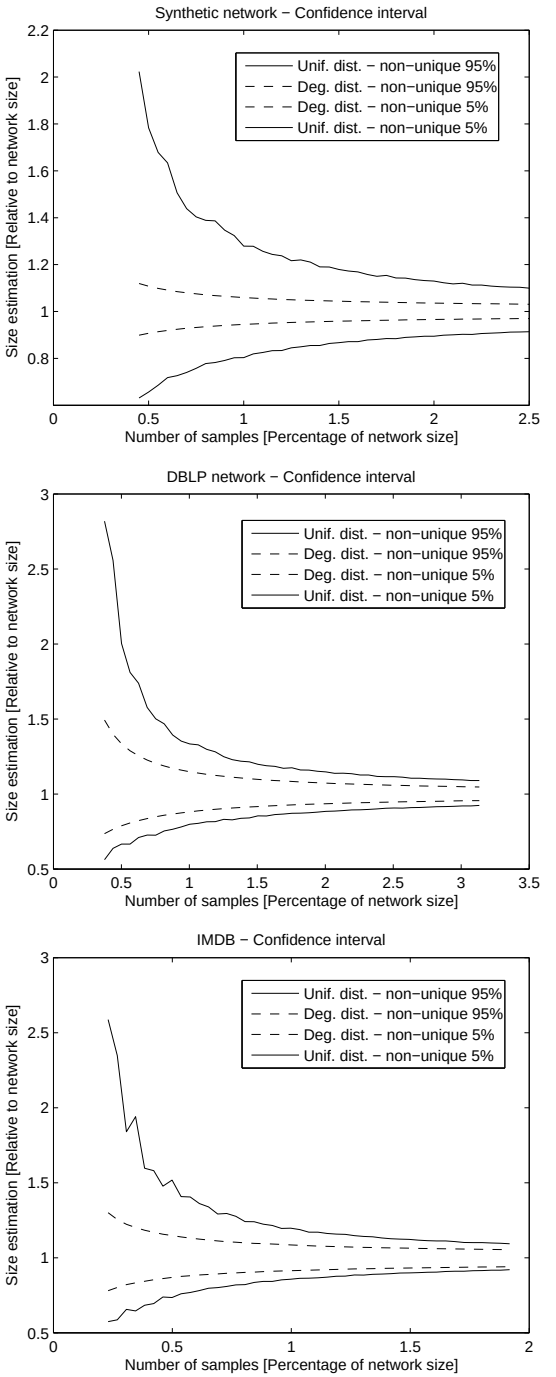

Figure 2: *Confidence interval curves* - relative estimated size vs. the percent of sampled nodes for three networks: a 1-million-node synthetic network (top), a network constructed from the Digital Bibliography and Library Project (DBLP) database (middle), and a network obtained from the Internet Movie Data Base (IMDB) database (bottom).

tions between actors were established according to joint participation in at least one movie or TV episode. The resulting graph included $1,955,508$ nodes.

We produced three types of curves. All three were plotted as functions of the percent of sampled nodes. *Error curves* present the normalized absolute estimation error, i.e.,

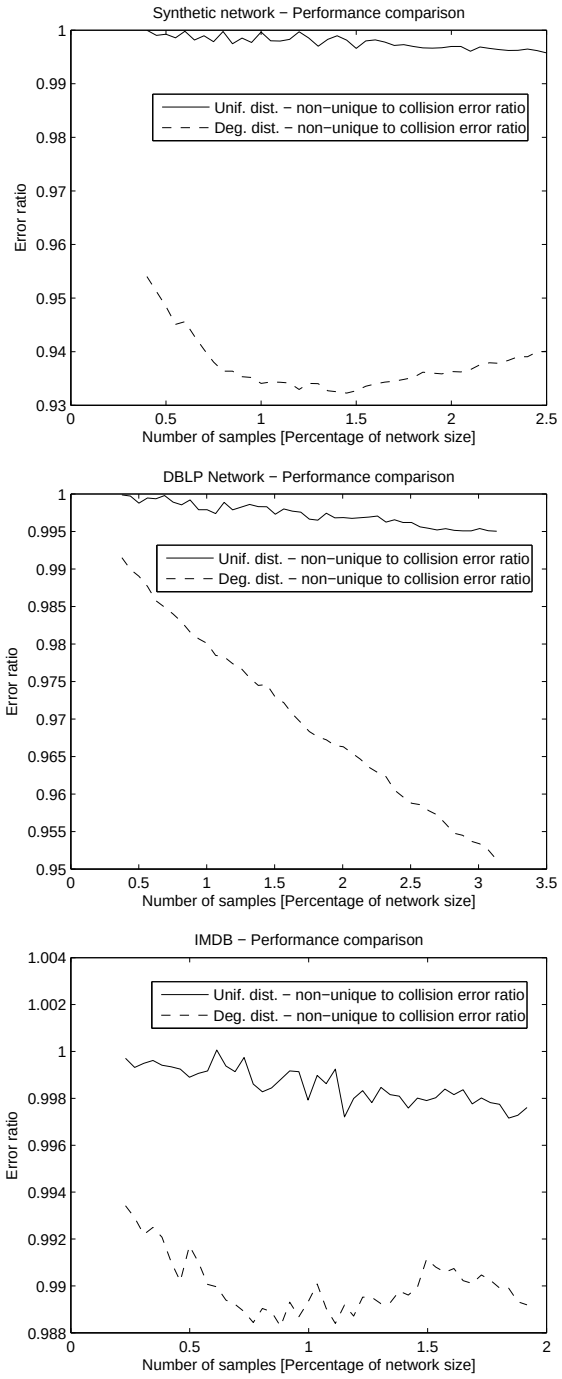

Figure 3: *Comparison curves* - absolute relative error ratio between the non-unique and collision estimators vs. the percent of sampled nodes for three networks: a 1-million-node synthetic network (top), a network constructed from the Digital Bibliography and Library Project (DBLP) database (middle), and a network obtained from the Internet Movie Data Base (IMDB) database (bottom).

$|n - \hat{n}|/n$ where n is the true size of the network and $\hat{n}$ is our estimate of it. *Confidence interval curves* give for each estimator the 5'th and 95'th percentile value from $10,000$ independent estimations. In other words, 90% of the estimated sizes fell between the lower and upper curves. Lastly, *comparison curves* present the ratio between the errors of the non-unique element estimator and that of the collision estimator. It is important to stress that this ratio is between the normalized absolute estimation errors and *not* between the estimated values. All presented plots were produced by averaging over $10,000$ independent experiments.

Examining the *error curves* depicted in Figure 1, the superiority of degree sampling estimation (both non-unique and collisions based) over uniform sampling estimation is well observed. In particular, for the synthetic network, uniform sampling estimation requires 5 times as many samples as required by degree sampling estimation (0.5% vs. 2.5%), to ensure a normalized absolute estimation error of less than 5%. Also, for the IMDB network, uniform sampling estimation required almost 3 times more samples than required by degree sampling estimation (0.3% vs. 0.8%) to ensure a normalized absolute error of less than 10%.

Similar observations regarding the estimation error are also notable examining the *confidence interval curves* depicted in Figure 2. These curves also demonstrate that there is an inherent asymmetrical bias towards size overestimation. This is probably because both estimators are inversely proportional to the number of collisions or non-unique elements. For example, a 50% discrepancy between the observed number of collisions and its expectation can result in 100% size overestimation but only in 35% size underestimation.

In all the curves above and for all three networks the non-unique elements estimator slightly outperformed the collisions based estimator. This phenomenon is visible when examining the *comparison curves* depicted in Figure 3. For example, for DBLP, the non-unique elements estimator provides a 5% reduced relative error over the collision based estimator when 3% of the network is sampled. The reader should note that the actual size estimates in this case differ by only 0.25%.

In all the aforementioned experiments we estimated the sizes of networks (whose sizes were already known) up to precision of a few single percents. We observed that both collision and non-unique based estimators perform well and that degree based sampling is significantly preferred to uniform sampling. In the next section we estimate the size of a subnetwork within Facebook and size of their entire network.

5.2 Facebook

We used two crawls performed on Facebook by the authors of [GKBM10]. The first crawl consisted of $984,830$ uniformly sampled users collected during April 2009.[9] The second crawl was performed during October 2010 and consisted of $988,116$ users. This crawl performed a simple random walk on the Facebook graph and therefore selected users with probability proportional to their degree.

5.2.1 Subnetwork size estimation

Since the actual size of Facebook is not known (other than Facebook's own reports) we first estimate the size of a subgraph whose size is known. We selected a random subset of $1,000,000$ Facebook users and tried to estimate the size of

[9]The Facebook uniformly sampled crawl can be found at http://odysseas.calit2.uci.edu/research/.

this sub-population using the first algorithm in Section 3.2. This is done for two reasons. First, to test the subgraph size estimation algorithm. Second, to make sure that Facebook's network topology and statistics are suitable for our estimators. We present an *error curve*, a *confidence interval curve*, and a *comparison curve* in Figure 4. Note that the x-axis here gives the percent of nodes sampled from the subnetwork and not the entire network as before.

These results corroborate that our subgraph size estimators behave almost identically to the complete graph estimators. This was expected since their analysis is essentially identical. A more important discovery is that the network topology and node degree distribution of Facebook are indeed suitable for our estimators to perform well.

5.2.2 Estimating the size of Facebook

We now estimate the size of the entire Facebook network. Presenting accuracy plots in this case is not possible since the true size of Facebook is not known. The uniform Facebook sample collected during April 2009, contains 2053 collisions and 2052 non-unique elements. Substituting these into Equations (1) and (2) yields estimates of $237,197,785$ and $236,984,623$ users respectively. The very same month, Facebook ([FBS]) reported of having "more than 200 million active users" and "more than 250 million active users" three months later. The crawl that was performed during October 2010 contained 4099 collisions and 4064 non-unique users. This gives estimates of $475,566,857$ and $475,864,724$ respectively. Facebook at the same time reported of having "more than 500 million active users". This is summarized in Table 1.

Table 1: Crawl details and consequent size estimates of the entire Facebook network for April 2009 and October 2010.

	April 2009	October 2010
Sampling distribution	uniform	degree
Number of samples	$0.98 \cdot 10^6$	$1 \cdot 10^6$
Number of collisions	2053	4099
Number of non-unique	2052	4064
Collision estimator	$237 \cdot 10^6$	$475 \cdot 10^6$
Nun-unique estimator	$236 \cdot 10^6$	$475 \cdot 10^6$
Facebook report	$200 - 250 \cdot 10^6$	$500 \cdot 10^6$

Notice that the second crawl sampled half as many users relative to the network size at the time it was made. Yet, it produced roughly twice as many collisions. This is because degree proportional sampling is expected to generate more collisions than uniform sampling. The discrepancy between our estimates and the official reports by Facebook stems from two main reasons. On the one hand, the crawler cannot distinguish between active and non active users. Thus our estimates include non-active users which causes an over estimation. On the other hand, our crawler cannot pass through users whose privacy settings hide their list of friends, which causes underestimation. Thus, our estimates and Facebook's reports give slightly different figures. While Facebook counts "active" users, we estimate the number of Facebook users whose list of friends is visible to the crawler regardless of their activity.

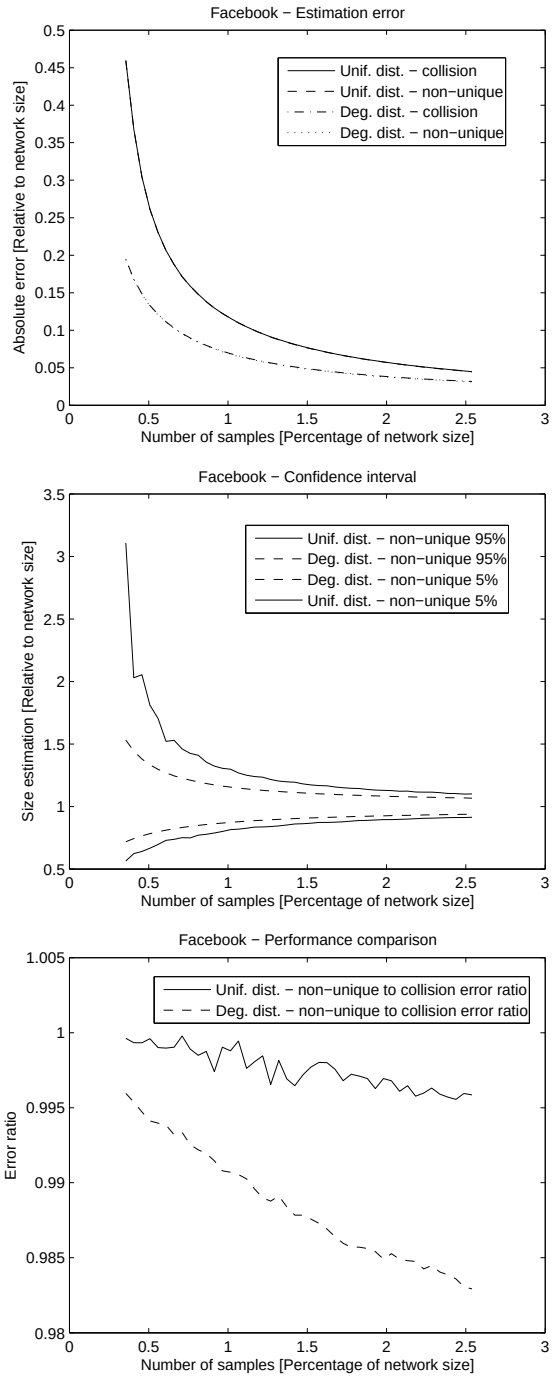

Figure 4: Absolute relative estimation error (top), confidence intervals (middle), and estimation relative error ratio (bottom) vs. the percentage of samples taken from a 1 million user subnetwork of Facebook.

Since the crawler has no indication of users' activity and since it is unclear what Facebook defines as "active", we cannot offset this effect. However, we can try to estimate the number of blocked users (those whose privacy settings block the crawler). This can be approximated using the fraction of such users in other users' friends lists which yields an

estimate of $650 \cdot 10^5$ users, active and non-active.

5.3 Synthetic network - large sample region

Interestingly, for a large enough number of samples (e.g., 30% of the network's size), uniform sampling estimation outperforms degree sampling estimation. This phenomenon repeats itself for all three known size networks we examined. We provide an *error curve*, a *confidence interval curve*, and a *comparison curve* in Figure 5 for the synthetic network only but this time extending the number of samples all the way to 100%.

5.4 Practical improvement of the algorithm

The algorithms presented here use random walks to sample r graph nodes *independently* from their stationary distribution. This requires a minimal number of random walk steps, say ℓ, between any two sampled nodes. Thus, to produce r independent samples, $r\ell$ random walk steps are required. While ℓ is small for rapid mixing graphs such as social networks, this is still rather wasteful since only a $1/\ell$ fraction of the encountered nodes are used to compute the estimator $\hat{n}$.

The first trivial improvement is to view the random walk of length $r\ell$ as ℓ disjoint and interleaved random walks. Producing ℓ different estimators $\hat{n}_1, \ldots, \hat{n}_\ell$ one could produce a better estimator which is either their mean or their median. Since $\hat{n}_1, \ldots, \hat{n}_\ell$ are *not* independent it is impossible to prove that the combined estimator exhibits better approximation guarantees but it does perform better in practice.

This can be viewed differently. Namely, modify the collision count estimator C to be $C' = \frac{1}{\ell}\sum_{t=1}^{\ell} C_{\ell'}$ where $C_{\ell'}$ is the number of collisions in random walk ℓ'. Unfortunately, $C_{\ell'}$ are dependent random variables and one cannot argue that $Var[C'] = Var[C]/\ell$. But it is still always true that $Var[C'] \leq Var[C]$ and that, practically, $Var[C']$ is significantly smaller than $Var[C]$. This reduces the required number of samples by roughly a factor of ℓ.

A better definition of C' can be $\sum_{|i-j|\geq\ell} Y_{i,j}$ where i and j range over $[r\ell]$. Intuitively, we consider any collision between two samples whose distance in the random walk is larger than ℓ. This insures that $\mathbb{E}[Y_{i,j}]$ is still $\sum_{i=1}^{n} p_i^2$. Notice that the number of such pairs is $\binom{r\ell}{2}(1 - O(\frac{1}{r}))$. Therefore, the number of expected collisions is roughly ℓ^2 times larger than C. Intuitively, this should also allow a reduction factor of ℓ^2 in the number of samples. While experimental results support this intuition, the same proof techniques cannot be used since $Y_{i,j}$ exhibit complex dependencies. For example, the probability that $Y_{i,j} = 1$ is significantly higher given that $Y_{i-1,j-1} = 1$. Nevertheless, this estimator was successfully used in [HRH09] and was shown to be practically useful.

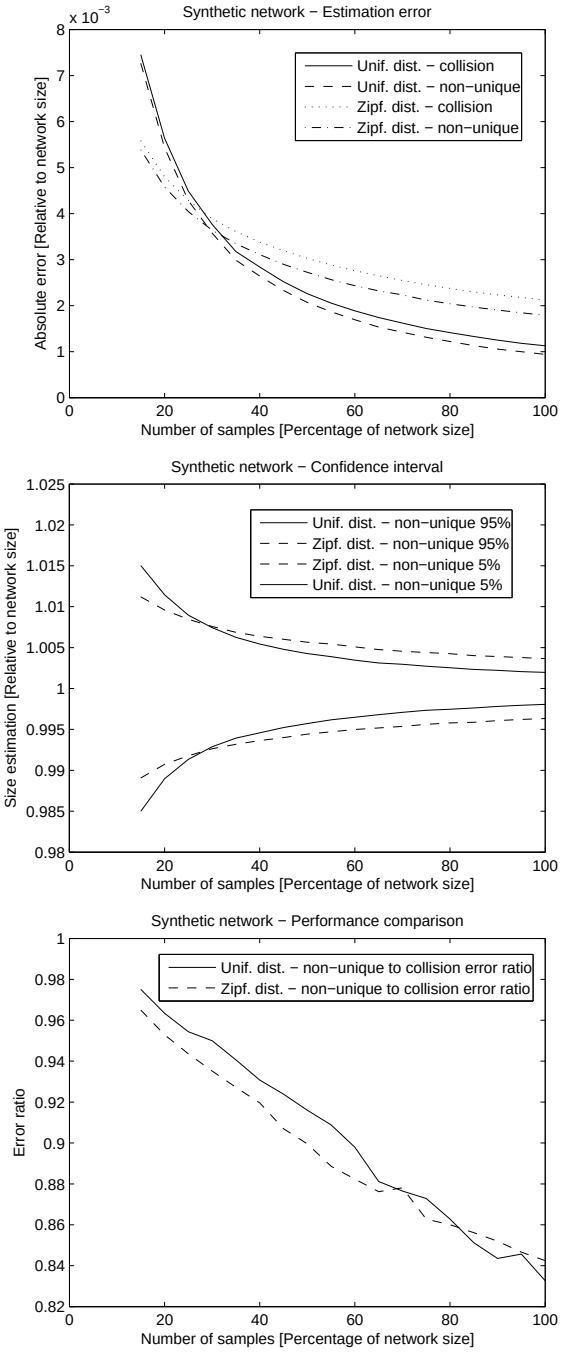

Figure 5: Absolute relative estimation error (top), confidence intervals (middle), and estimation relative error ratio (bottom) vs. the percent of sampled nodes for a 1 million node synthetic network whose node degree distribution is Zipfian. Note the large number of samples relative to the network size.

6. CONCLUSIONS

We presented two algorithms for estimating the size of graphs. Both algorithms rely on nodes being samples from the graph's stationary distribution. We showed both analytically and experimentally that, for social-networks and other small world graphs, these algorithms considerably outperform uniformly sampling nodes. They consistently provide more accurate estimates while using a smaller number of samples. This result is even more outstanding since uniformly sampling nodes is strictly harder than sampling them according to the stationary distribution.

7. ACKNOWLEDGMENT

We thank Minas Gjoka for being extremely generous and providing the Facebook data used for the simulations. We also thank Ronny Lampel, Yoelle Maarek and Ravi Kumar for their guidance and suggestions.

8. REFERENCES

[BB98] K. Bharat and A. Broder. A technique for measuring the relative size and overlap of public web search engines. *Comput. Netw. ISDN Syst.*, 30(1-7):379–388, 1998.

[BG07] Z. Bar-Yossef and M. Gurevich. Efficient search engine measurements. In *Proc. of the 16th international conference on World Wide Web (WWW'07)*, pages 401–410, Banff, Alberta, Canada, 2007.

[BG08] Z. Bar-Yossef and M. Gurevich. Random sampling from a search engine's index. *J. ACM*, 55(5), 2008.

[Cha87] A. Chao. Estimating the population size for capture-recapture data with unequal catchability. In *Biometrics*, Dec. 1987.

[FBS] Facebook statistics. http://www.facebook.com/press/info.php?statistics.

[FTV98] M. Finkelstein, H. G. Tucker, and J. A. Veeh. Confidence intervals for the number of unseen types. *Statistics & Probability Letters*, 37(4):423–430, Mar. 1998.

[GKBM10] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou. Walking in Facebook: A case study of unbiased sampling of OSNs. In *Proc. of IEEE INFOCOM '10*, San Diego, CA, Mar. 2010.

[HRH09] S. J. Hardiman, P. Richmond, and S. Hutzler. Calculating statistics of complex networks through random walks with an application to the on-line social network bebo. *The European Physical Journal B - Condensed Matter and Complex Systems*, 71:611–622, 2009.

[HTV+09] S. Han Hee, C. Tae Won, D. Vacha, Z. Yin, and Q.Lili. Scalable proximity estimation and link prediction in online social networks. In *Proc. of the 9th ACM SIGCOMM conference on Internet Measurement (IMC'09)*, pages 322–335, Chicago, IL, USA, 2009.

[HYCY03] R. Huggins, H.-C. Yang, A. Chao, and P. S. F. Yip. Population size estimation using local sample coverage for open populations. *Journal of Statistical Planning and Inference*, 113(2):699 – 714, 2003.

[Joh] Estimating the size of a population. http://www.rsscse.org.uk/ts/gtb/johnson.pdf.

[Knu74] D. E. Knuth. Estimating the efficiency of backtrack programs. Technical report, Stanford, CA, USA, 1974.

[KNV06] Y. Koren, S. C. North, and C. Volinsky. Measuring and extracting proximity in networks. In *Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD'06)*, pages 245–255, Philadelphia, PA, USA, 2006.

[Lov93] L. Lovasz. Random walks on graphs. a survey. *Combinatorics*, 1993.

[LP56] R. C. Lewontin and T. Prout. Estimation of the number of different classes in a population. *Biometrics*, 12(2):211–223, 1956.

[MMG+07] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proc. of the 5th ACM Conference on Internet Measurement (IMC'07)*, San Diego, CA, USA, 2007.

[MS89] A. Marchetti-Spaccamela. On the estimate of the size of a directed graph. In J. van Leeuwen, editor, *Graph-Theoretic Concepts in Computer Science*, volume 344 of *Lecture Notes in Computer Science*, pages 317–326. Springer Berlin / Heidelberg, 1989.

[MV06] R. Motwani and S. Vassilvitskii. Distinct values estimators for power law distributions. In *Proc. of the Third Workshop on Analytic Algorithmics and Combinatorics (ANALCO'06)*, Miami , FL, 2006.

[Pit87] L. Pitt. A note on extending Knuth's tree estimator to directed acyclic graphs. *Inf. Process. Lett.*, 24(3):203–206, 1987.

[YSH+07] A. Yong-Yeol, H. Seungyeop, K. Haewoon, M. Sue, and J. Hawoong. Analysis of topological characteristics of huge online social networking services. In *Proc. of the 16th international conference on World Wide Web (WWW'07)*, pages 835–844, Banff, Alberta, Canada, 2007.

[YW10] S. Ye and F. Wu. Estimating the size of online social networks. In *Proc. of the IEEE Second International Conference on Social Computing (SocialCom)*, pages 169–176, Aug. 2010.

APPENDIX

A. CONCENTRATION OF C AND R

The Proof of Corollary 1 required computing the values of $Var[C]/\mathbb{E}^2[C]$ and $Var[R]/\mathbb{E}^2[R]$. While the derivation is somewhat tedious it is straight forward nonetheless. The details are given below. For conciseness we define $P_\ell = \sum_{i=1}^n p_i^\ell$ where $p_i = d_i/D$.

In order to compute $Var[R]$ we need to calculate its first two moments. Starting with the first moment and recalling the samples are independent, we have that

$$\mathbb{E}[R] = \sum_{\substack{i,j=1 \\ j\neq i}}^{r} \mathbb{E}[\frac{d_{x_i}}{d_{x_j}}] = r(r-1)\sum_{i,j=1}^{n} p_i p_j \frac{d_i}{d_j} = r(r-1)nP_2 \tag{3}$$

Turning to the second moment of R, we have that

$$\mathbb{E}[R^2] = \sum_{\substack{i,j=1 \\ j\neq i}}^{r} \sum_{\substack{i',j'=1 \\ j'\neq i'}}^{r} \mathbb{E}[\frac{d_{x_i}}{d_{x_j}}\frac{d_{x_{i'}}}{d_{x_{j'}}}] \ . \tag{4}$$

The last summation can be divided into six cases.

1. There are $2!\binom{r}{2}$ occurrences where

$$\mathbb{E}[\frac{d_{x_i}}{d_{x_j}}\frac{d_{x_{i'}}}{d_{x_{j'}}}] = 1 \ ,$$

for $i = j', \ j = i'$.

2. There are $2!\binom{r}{2}$ occurrences where

$$\mathbb{E}[\frac{d_{x_i}}{d_{x_j}}\frac{d_{x_{i'}}}{d_{x_{j'}}}] = \mathbb{E}[\frac{d_{x_i}^2}{d_{x_j}^2}] = P_3 P_{-1} \ ,$$

for $i = i',\ j = j'$.

3. There are $3!\binom{r}{3}$ occurrences where

$$\mathbb{E}[\frac{d_{x_i}}{d_{x_j}}\frac{d_{x_{i'}}}{d_{x_{j'}}}] = \mathbb{E}[\frac{d_{x_i}^2}{d_{x_j}d_{x'_j}}] = n^2 P_3 \ ,$$

for $i = i',\ j \neq j'$.

4. There are $3!\binom{r}{3}$ occurrences where

$$\mathbb{E}[\frac{d_{x_i}}{d_{x_j}}\frac{d_{x_{i'}}}{d_{x_{j'}}}] = \mathbb{E}[\frac{d_{x_i}d_{x'_i}}{d_{x_j}^2}] = P_2^2 P_{-1} \ ,$$

for $i \neq i',\ j = j'$.

5. There are $2 \cdot 3!\binom{r}{3}$ occurrences where

$$\mathbb{E}[\frac{d_{x_i}}{d_{x_j}}\frac{d_{x_{i'}}}{d_{x_{j'}}}] = \mathbb{E}[\frac{d_{x_i}}{d_{x_j}}] = nP_2 \ ,$$

for $i = j',\ j \neq i'$ or $i \neq j',\ j = i'$.

6. There are $4!\binom{r}{4}$ occurrences where

$$\mathbb{E}[\frac{d_{x_i}}{d_{x_j}}\frac{d_{x_{i'}}}{d_{x_{j'}}}] = n^2 P_2^2 \ ,$$

for different $i,\ j,\ i',\ j'$. This term is at most $\mathbb{E}^2[R]$.

Combining the different cases above and using the fact that $nP_2 \geq 1$ it can be shown that for $r \geq 2$ the following inequality holds

$$\begin{aligned}\frac{Var(R)}{\mathbb{E}^2[R]} &\leq \frac{1}{r(r-1)}\frac{1}{n^2P_2^2} + \frac{1}{r(r-1)}\frac{P_3}{P_2^2}\frac{P_{-1}}{n^2} \\ &+\frac{r-2}{r(r-1)}\frac{P_3}{P_2^2} + \frac{r-2}{r(r-1)}\frac{P_{-1}}{n^2} + \frac{2(r-2)}{r(r-1)}\frac{1}{nP_2} \\ &\leq \frac{1}{r(r-1)} + \frac{ab}{r(r-1)} + \frac{a}{r} + \frac{b}{r} + \frac{2}{r}\end{aligned}$$

Where we define $a = 1/\sqrt{P_2}$ and $b = P_{-1}/n^2$. Note that this also requires the facts that $nP_2 \geq 1$, $(r-2)/(r-1) \leq 1$ and $P_3/P_2^2 \leq 1/P_2^{1/2}$. The last inequality is due the fact that $P_3^{1/3} \leq P_2^{1/2}$ from the monotonicity of ℓ_p norms.

We turn now to computing the variance of the number of collisions C. We remind the reader our notations. $Y_{i,j} = 1$ if $x_i = x_j$ and 0 else, where $\{x_1 \ldots, x_r\}$ are the r sampled nodes. Moreover, $C = \sum_{\substack{i,j=1 \\ j\neq i}}^{r} Y_{i,j}$. To compute the $Var(C)$ we need to calculate its first two moments. Starting with the first moment we have that

$$\mathbb{E}[C] = \sum_{\substack{i,j=1 \\ j\neq i}}^{r} \mathbb{E}[Y_{i,j}] = r(r-1)P_2 \ .$$

We compute $\mathbb{E}[C^2]$ using the linearity of the expectation.

$$\mathbb{E}[C^2] = \mathbb{E}[(\sum_{\substack{i,j=1 \\ j\neq i}}^{r} Y_{i,j})^2] = \sum_{\substack{i,j=1 \\ j\neq i}}^{r}\sum_{\substack{i,j=1 \\ j\neq i}}^{r} \mathbb{E}[Y_{i,j}Y_{i',j'}]$$

To calculate the last summation we divide it into three cases.

1. There are $2!\binom{r}{2}$ occurrences where

$$\mathbb{E}[Y_{i,j}Y_{i',j'}] = \mathbb{E}[Y_{i,j}] = P_2 \ ,$$

for $i = i',\ j = j'$.

2. There are $2 \cdot 3!\binom{r}{3}$ occurrences where

$$\mathbb{E}[Y_{i,j}Y_{i',j'}] = \mathbb{E}[Y_{i,j}Y_{i',j}]\mathbb{E}[Y_{i,j}Y_{i,j'}] = P_3 \ ,$$

for different $i \neq i',\ j = j'$ or $i = i',\ j \neq j'$.

3. There are $4!\binom{r}{4}$ occurrences where

$$\mathbb{E}[Y_{i,j}Y_{i',j'}] = (\mathbb{E}[Y_{i,j}])^2 = P_2^2 \ ,$$

for different $i,\ i',\ j,\ j'$. This term is at most $\mathbb{E}^2[C]$.

Combining the different cases above it can be easily shown that for $r \geq 2$, the following inequality holds

$$\frac{Var(C)}{\mathbb{E}^2[C]} \leq \frac{1}{r(r-1)P_2} + \frac{2(r-2)}{r(r-1)}\frac{P_3}{P_2^2} \leq \frac{a^2}{r(r-1)} + \frac{2a}{r}$$

where $a = 1/\sqrt{P_2}$ as before. This completes the derivation required for the proof.