Analysis of Degree Distribution with Respect to Collision-Based Estimator

Roman Vakhrushev, rv1057@nyu.edu

December 19, 2023

1 Motivation and Objectives

The paper "Estimating Sizes of Social Networks via Biased Sampling" (Katzir et. al., 2011) presented and described an estimator for the estimation of the size of a network via computing multiple statistics on the random walk. However, the authors' analysis was mostly focused on Zipfian distribution with particular parameters. To further understand the nature of this estimator, we decided to study the behavior of this estimator on different degree distributions. In particular, we study the required number of samples one needs to collect during a random walk to obtain the desired probability bounds. We further attempt to generalize this to an arbitrary degree distribution and show that the estimator requires $O(\sqrt{n})$ samples in the worst case.

2 Introduction and Background

This paper further investigates the properties of the estimator presented in the paper "Estimating Sizes of Social Networks via Biased Sampling". In the paper, the authors propose an estimator of the size of a social network by performing a random walk on it.

While performing a random walk on a graph of size n, they take r $\{x_1, ..., x_r\}$ samples independently, according to the stationary distribution of a graph $p_i = d_i/D$, where d_i is a degree of a node and $D = \sum_{i=1}^{n} d_i$.

They further define three variables C, Ψ_1, Ψ_{-1}. Define an indicator variable $Y_{i,j} = 1$ when two samples x_i and x_j are the same and 0 otherwise. Then, collision variable $C = \sum_{i<j} Y_{i,j}$. Ψ_1 is a sum of sampled degrees, $\Psi_1 = \sum_{i=1}^{d} d_i$ and Ψ_{-1} is a sum of reciprocals of sampled degrees $\Psi_{-1} = \sum_{i=1}^{d} 1/d_1$.

Using these variables, the estimator is defined as $\hat{n} = \Psi_1\Psi_{-1}/2C$. Most importantly, they prove an important theorem (stated as a Corollary), which serves as a basis for this paper:

Corollary 1. *For any degree distribution and C, Ψ_1, and Ψ_{-1}, the estimator $\hat{n} = \Psi_1\Psi_{-1}/2C$ with guarantees $\epsilon \leq 1/2$, $\delta \leq 1$:* $\Pr(n(1-\epsilon) \leq \hat{n} \leq n(1+\epsilon)) \geq 1-\delta$

As long as number of samples $r \in O\{\frac{1}{\epsilon\sqrt{\delta}\sqrt{\sum_{i=1}^{n} p_i^2}}, \frac{\sum_{i=1}^{n} p_i^3}{\epsilon^2\delta(\sum_{i=1}^{n} p_i^2)^2}, \frac{\sum_{i=1}^{n} 1/p_i}{\epsilon^2\delta n^2}\}$

Please, refer to the original paper for any additional details.

3 Moment-based version of the corollary

Corollary 1 is a very powerful statement that provides us with an easy way to estimate the desired sample size for any degree distribution. We only need to compute $\sum_{i=1}^{n} p_i^l$, for $l = -1, 2, 3$. Although this may be easy in certain scenarios (for example, uniform distribution), it is often hard to compute in other scenarios. Therefore, it becomes very convenient to convert this to a well-known study of moments.

Lemma 1. *Sum of probabilities to the l-th power can be approximated as follows:* $\sum_{i=1}^{n} p_i^l = \frac{\mathcal{M}_l}{n^{l-1}(\mathcal{M}_1)^l}$ *(Where $\mathcal{M}_l$ is l-th raw moment).*

Proof.

$$\begin{aligned}
\sum_{i=1}^{n} p_i^l &= \sum_{i=1}^{n} \frac{d_i^l}{D^l} && \text{(By the way we do the random walk)} \\
&= \frac{\sum_{i=1}^{n} d_i^l}{D^l} && \text{(D^l is independent of i)} \\
&= \frac{\sum_{i=1}^{n} d_i^l}{n^l(\frac{1}{n}D)^l} \\
&= \frac{\sum_{i=1}^{n} d_i^l}{n^l(\frac{1}{n}\sum_{i=1}^{n} d_i)^l} && \text{(By definition of D)} \\
&= \frac{\sum_{i=1}^{n} d_i^l}{n^l(\mathbb{E}[d])^l} \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n} d_i^l}{n^{l-1}(\mathbb{E}[d])^l} \\
&= \frac{\mathbb{E}[d^l]}{n^{l-1}(\mathbb{E}[d])^l} \\
&= \frac{\mathcal{M}_l}{n^{l-1}(\mathcal{M}_1)^l} && \text{(By the fact that, for any degree distribution, l-th moment: $\mathcal{M}_l = \mathbb{E}[d^l]$)}
\end{aligned}$$

□

Note: the last equality actually requires the Strong Law of Large Numbers, which guarantees convergence of moments to those of generating distribution. SLLN usually holds for big graphs.

Using this small Lemma (simply replacing $\sum_{i=1}^{n} p_i^l$ with $\frac{\mathcal{M}_l}{n^{l-1}(\mathcal{M}_1)^l}$), we can obtain a more convenient version of the Corollary 1.

Theorem 1. *For any degree distribution and C, Ψ_1, and Ψ_{-1}, the estimator $\hat{n} = \Psi_1\Psi_{-1}/2C$ with guarantees $\epsilon \le 1/2$, $\delta \le 1$: $\Pr(n(1-\epsilon) \le \hat{n} \le n(1+\epsilon)) \ge 1-\delta$*

As long as number of samples $r \in O\{\frac{\sqrt{n}\mathcal{M}_1}{\epsilon\sqrt{\delta}\sqrt{\mathcal{M}_2}}, \frac{\mathcal{M}_3\mathcal{M}_1}{\epsilon^2\delta(M_2)^2}, \frac{\mathcal{M}_{-1}\mathcal{M}_1}{\epsilon^2\delta}\}$

So, if we can compute the corresponding moments, we can actually compute r. The way to compute these moments is shown in the next section.

4 Analysis of Degree Distributions

4.1 Zipfian Distribution

4.1.1 Generalized Zipfian Distribution

Zipfian distribution is a parameterized probability distribution with the parameter α. The probability of degree i is given as $\Pr(d = i) = \frac{1}{Hi^{\alpha}}$ for $i = 1, ..., d_m$, where H is a normalization constant (different for each α).

Moment generating function of Zipfian distribution with parameter α and normalization constant H is given by $M(t) = \frac{1}{H}\sum_{i=1}^{d_m} \frac{e^{it}}{i^{\alpha}}$.

It is well-known (from Taylor series expansion of e^x), that for any probability distribution $\mathcal{M}_l = \mathbb{E}[d^l] = M^{(l)}(0)$, where $M^{(l)}$ is the l-th derivative of $M(t)$ with respect to t.

From there,

$\mathcal{M}_1 = \frac{1}{H}\sum_{i=1}^{d_m} \frac{i}{i^{\alpha}} = \frac{1}{H}\sum_{i=1}^{d_m} \frac{1}{i^{\alpha-1}}$.

$\mathcal{M}_2 = \frac{1}{H}\sum_{i=1}^{d_m} \frac{i^2}{i^{\alpha}} = \frac{1}{H}\sum_{i=1}^{d_m} \frac{1}{i^{\alpha-2}}$.

$\mathcal{M}_3 = \frac{1}{H}\sum_{i=1}^{d_m} \frac{i^3}{i^{\alpha}} = \frac{1}{H}\sum_{i=1}^{d_m} \frac{1}{i^{\alpha-3}}$.

$\mathcal{M}_{-1} = \frac{1}{H}\sum_{i=1}^{d_m} \frac{1}{i^{\alpha+1}}$.

Note that when $\frac{1}{i^a}$ for some $a > 0$, the term $\sum_{i=1}^{d_m} \frac{1}{i^a} = H_{d_m,n}$ (so called, generalized Harmonic number). This number is always bounded above by the Riemann function, $\zeta(m)$, which, for our purposes, is just a small number. Therefore, for the analysis of this distribution, the value of $\mathcal{M}_{-1}$ is irrelevant, as it is a constant.

4.1.2 Zipfian Distribution with Parameter $\alpha = 3$

The original paper discusses an important case of Zipfian distribution with $\alpha = 2$. Although graphs with Zipfian degree distribution and parameter $\alpha = 2$ often appear in the real world, many graphs also have Zipfian degree distribution, but with $\alpha \neq 2$. In particular, graphs often have Zipfian degree distribution with $2 \leq \alpha \leq 3$ (Yong-Yeol, et. al., 2007). Thus, we show the requirement number of samples for Zipfian distribution with $\alpha = 3$.

Using the results from the previous section, we have $\mathcal{M}_1 = \frac{1}{H}\sum_{i=1}^{d_m} \frac{1}{i^2} \approx \frac{\pi^2}{6H}$ (famous Basel problem), $\mathcal{M}_2 = \frac{1}{H}\sum_{i=1}^{d_m} \frac{1}{i} \approx \frac{\log d_m}{H}$ (Harmonic series), $\mathcal{M}_3 = \frac{1}{H}\sum_{i=1}^{d_m} 1 = \frac{d_m}{H}$. Plugging these into the Theorem 1, we have $r_c \in O\{\frac{\sqrt{n}}{\sqrt{\log d_m}}, \frac{d_m}{(\log d_m)^2}, 1\}$. Here, we also need to note that $d_m = \Theta(n^{1/3})$. So, $r \in O(\frac{\sqrt{n}}{\sqrt{\log n}})$.

4.2 Binomial Distribution

4.2.1 Generalized Binomial Distribution

Although binomial degree distribution is not often present in real-world graphs, it is an important case for some some artificially designed graphs (Boshra, Taylor, and Bogachev, 2023).

The binomial distribution is a famous probability distribution with the parameter p (probability of success). The probability of degree i is then given as $Pr(d = i) = c(n, p)p^i(1 - p)^{n-i}$.

The moment-generating function of this distribution is given as $(1 - p + pe^t)^n$.

From this we can compute the derivatives: $M^{(1)}(t) = npe^t(p(e^t - 1) + 1)^{n-1}$,

$M^{(2)}(t) = npe^t(p(e^t - 1) + 1)^{n-2}(p(ne^t - 1) + 1)$,

$M^{(3)}(t) = n^2p^2e^t(p(e^t - 1) + 1)^{n-2} + (n - 2)np^2e^t(p(ne^t - 1) + 1)(p(e^t - 1) + 1)^{n-3}$

$\mathcal{M}_1 = M^{(1)}(0) = np(1)^{n-1} = np$

$\mathcal{M}_2 = M^{(2)}(0) = np((n - 1)p + 1)$.

$\mathcal{M}_3 = M^{(3)}(0) = (-1 + n)np^2(2 + (-2 + n)p)$

Additionally, $M_{(-1)} = O(n^{-1})$. (Skorski, 2022)

As we can see from the equations above, the parameter p, unlike parameter α in the Zipfian distribution, does not play a role in the asymptotic analysis of this degree distribution. However, in practice, it is defined and can be any number. So, we treat p as constant and apply theorem 1, with the moments we computed.

$r \in O\{\frac{\sqrt{n}n}{n}, \frac{n^3n}{n^4}, 1\} = O(\sqrt{n})$.

4.3 Geometric and Other Infinite Distributions

Many real-world graphs follow degree distributions that are close to geometric (Mislove et. al., 2007) and (Gjoka et. al., 2010).

Geometric distribution is a well-known probability distribution with the parameter p (probability of success).

The probability of degree i is given as $\Pr(d = i) = (1 - p)^{k-1}p$ for $i = 1, 2, 3, ...$

Geometric distribution has the following moment generating function $M(t) = \frac{pe^t}{1-(1-p)e^t}$.

From this we can compute the derivatives: $M^{(1)}(t) = \frac{pe^t}{((p-1)e^t+1)^2}$, $M^{(2)}(t) = -\frac{pe^t((p-1)e^t-1)}{((p-1)e^t+1)^3}$,
$M^{(3)}(t) = \frac{pe^t((p-1)^2e^{2t}-4(p-1)e^t+1)}{((p-1)e^x+1)^4}$

While these may look a little complicated, we only need to evaluate them at $t = 0$ to obtain the moments.

Therefore,

$\mathcal{M}_1 = M^{(1)}(0) = \frac{p}{(p-1+1)^2} = \frac{1}{p}$.

$\mathcal{M}_2 = M^{(2)}(0) = -\frac{p(p-2)}{(p)^3} = \frac{2-p}{p^2}$.

$\mathcal{M}_3 = M^{(2)}(0) = \frac{(1-4(-1+p)+(-1+p)^2p}{(1-1+p)^4} = \frac{p^2-6p+6}{p^3}$.

The analysis for these kinds of distributions should be different from the other distributions, as these distributions are "infinite" (in the sense that n goes to infinity). Even though these distributions are discrete, they do not depend on n. If we simply plug these results into Theorem 1, we will not obtain any meaningful results as we are mostly focused on big-O analysis with respect to n. Therefore, it might be more useful to study truncated versions of these distributions, but sadly this requires exact definitions and, most importantly, moment-generating functions, which are not known to us.

This also applies to other important "infinite" distributions, including Poisson distribution.

4.4 Uniform Distribution

Uniform distribution is an important case for degree distributions. It represents connections in the regular graphs, where each vertex has the same number of neighbors. However, these kinds of graphs are rarely found in the real world. Nonetheless, it is an important case for understanding the theoretical aspects of the collision-based estimators. Some of these theoretical aspects are discussed in the next section.

In this case, it is actually much easier to compute the desired number of samples directly from Corollary 1. We substitute probability $p = \frac{1}{n}$ into Corollary 1 to obtain:

$r \in O\{\sqrt{n}, \frac{n^2}{n^2}, 1\} = O(\sqrt{n})$.

4.5 Almost Degenerate Distribution

Another special case is the following distribution of n nodes: $p_1 = \frac{1}{2}, p_i = \frac{1}{2(n-1)}$ for any $i \neq 1$.

This represents a graph with a big hub (with degree $n-1$) and every other node is connected to this hub via a single edge. This distribution is similar to degenerate distribution, however, the probability for all nodes (except for the hub) is small but non-zero. The reason, we cannot make all of these probabilities equal to zero is because all of the nodes in the graph have to be connected (this is a requirement for random walk). Thus, this distribution, in some sense, is the worst distribution we could obtain. The significance of this degree distribution will be realized later.

Let's compute the desired number of samples using Corollary 1.

$r \in O(2\sqrt{(n-1)}, \frac{n^4}{(n^2)^2}), \frac{n^2}{n^2}) = O(\sqrt{n})$.

5 General Bound for the Number of Samples for any Degree Distribution

We attempt to find the general bound for the number of samples. We will refer to Corollary 1 multiple times and define condition 1 as $\frac{1}{\epsilon\sqrt{\delta}\sqrt{\sum_{i=1}^{n} p_i^2}}$, condition 2 as $\frac{\sum_{i=1}^{n} p_i^3}{\epsilon^2\delta(\sum_{i=1}^{n} p_i^2)^2}$ and condition 3 as $\frac{\sum_{i=1}^{n} 1/p_i}{\epsilon^2\delta n^2}\}$.

5.1 Redundancy of the Second Condition

We can show that the second condition in the Corollary 1 (and thus in Theorem 1) is redundant.

To see this, compare condition 1: $\frac{1}{\epsilon\sqrt{\delta}\sqrt{\sum_{i=1}^{n} p_i^2}}$, and condition 2: $\frac{\sum_{i=1}^{n} p_i^3}{\epsilon^2\delta(\sum_{i=1}^{n} p_i^2)^2}$

Note that

$$\begin{aligned}
\frac{\sum_{i=1}^{n} p_i^3}{\epsilon^2\delta(\sum_{i=1}^{n} p_i^2)^2}\} &\leq \frac{\sum_{i=1}^{n} p_i^2}{\epsilon^2\delta(\sum_{i=1}^{n} p_i^2)^2} && \text{(Each } p_i \text{ is a number between 0 and 1)} \\
&= \frac{1}{\epsilon^2\delta(\sum_{i=1}^{n} p_i^2)} \\
&\leq \frac{1}{\epsilon\sqrt{\delta}\sqrt{\sum_{i=1}^{n} p_i^2}} && (\frac{1}{\sqrt{x}} > \frac{1}{x} \text{ for all } x > 1)
\end{aligned}$$

Therefore, we actually only have two conditions $1, 3$ and both the Corollary 1 and the Theorem 1 can exclude the second condition from their formulation.

5.2 Simple Algorithm for Analysis of the conditions

We only need to analyze condition 1 and 3. To do these we define a simple operation and an algorithm.

Operation *flip*(p_1,p_2,μ) is defined as follows: for two probabilities p_1 and p_2 and $0 < \mu < \min(p_1, p_2)$ if $p_1 \leq p_2$, subtract μ from p_1 and add to p_2. Now $p_1' = p_1 - \mu$ and $p_2' = p_2 + \mu$. If it is not the case that $p_1 \leq p_2$, swap them and do the same operation. As a result of this operation, the sum in $p_1 + p_2$ does not change, but the actual values of p_1 and p_2 change by μ.

Algorithm *Alg1*(D_1, D_2) is defined as follows: given two finite discrete distributions D_1 and D_2 (each defined as n probabilities with the corresponding nodes) convert D_1 to D_2 as follows: using *flip* operation with appropriate μ convert one of the probabilities of D_1 to D_2, then repeat until cannot convert any probability.

Claim 1 Given some degree distribtution D (defined as n probabilities with the corresponding nodes), use of *flip* operation on any probabilities in this distribution always increases the sum of squares of probabilities.

Proof. Apply *flip* to any two probabilities p_1 and p_2 (assume $p_1 \leq p_2$). Note that the squared sum of all the other probabilities does not change. The squared sum of the two probabilities was $p_1^2+p_2$ and after flip it becomes $(p_1-\mu)^2+(p_2+\mu)^2 = p_1^2-2p_1\mu+\mu^2+p_2^2-2p_1\mu+\mu^2 = p_1^2+p_2^2+2\mu(p_2-p_1)+2\mu^2$. Since $p_2 - p_1$ is greater ot equal than 0, and $\mu > 0$, we have $(p_1 - \mu)^2 + (p_2 + \mu)^2 > p_1^2 + p_2^2$. □

Claim 2 Given some degree distribtution D (defined as n probabilities with the corresponding nodes), use of *flip* operation on any probabilities in this distribution always increases the sum of reciprocals of the probabilities.

Proof. Apply *flip* to any two probabilities p_1 and p_2 (assume $p_1 \leq p_2$). Note that the sum of reciprocals of all the other probabilities does not change. Consider values of $\frac{1}{p_1-\mu}$ and $\frac{1}{p_2+\mu}$. Note that since $p_1 \leq p_2$, $\frac{1}{p_1} \geq \frac{1}{p_2}$. And then, by the properties of $1/x$, $\frac{1}{p_1-\mu} \geq \frac{1}{p_2+\mu}$. □

Claim 3 Starting from uniform distribution, by using algorithm *Alg1* we can obtain any finite discrete distribution.

Proof. Note that any other distribution except for uniform has the following property: at least one of the probabilities is greater than the other. While running the algorithm, we can set $n/2$ of probabilities of D_2 just from probabilities uniform probabilities $1/n$ and then repeat the process. We can always achieve this, since uniform distribution $\mu = 1/n$, which is the largest value possible for this operation. □

Claim 4 Uniform degree distribution provides us with the largest possible bound for condition 1.

Proof. By Claim 3, any other degree distribtuion can be represented by applying *Alg1* on uniform degree distribution. Since *Alg1* only uses *flip* operations, by Claim 1, the value of sum of squares of probabilities always increases. Therefore, any other degree distribution has a higher sum of squares of probabilities. Since in condition 1, we have sum of squares (its square root) in the denominator, the condition 1 is largest for uniform degree distribution. □

Claim 5 Almost degenerate distribution provides us with the largest possible bound for condition 3.

Proof. Starting from uniform degree distribution, arrive at almost degenerate distribution using *Alg1* (this is always possible by Claim 3). Note: by using *flip* operations on this distribution, we cannot obtain any other distribution (this is true since all nodes need to be connected - so they have the smallest probability, so we cannot do *flip* on them). Also note that we can always convert to this distribution from any other distribution (always possible to decrease one probability by increasing the other). Since we are only using *flip* operations, by Claim 2, the value of sum of reciprocals of probabilities always increases. Since condition 3 has this sum in the numerator, almost degenerate distribution, the condition 3 is the largest for this distribution. □

5.3 General Bound for the Number of Samples

By Claim 4 and Claim 5, we know that either uniform distribution or almost degenerate distribution should result in the largest value of r (since they maximize Condition 1 or 3). However, both of them require $r \in O(\sqrt{n})$ samples (as shown in section 3). Therefore, the algorithm always requires at most $O(\sqrt{n})$ samples for any degree distribution.

6 References

Katzir, Liran & Liberty, Edo & Somekh, Oren. (2011). Estimating Sizes of Social Networks via Biased Sampling. Internet Mathematics. 10. 597-606. 10.1145/1963405.1963489.

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

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

M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou. Walking in Facebook: A Case Study of Unbiased Sampling of OSNs. In Proceedings of IEEE INFOCOM '10, San Diego, CA, March 2010

Maciej Skorski. (2022). Handy Formulas for Binomial Moments.

Boshra Alarfaj, Charles Taylor, & Leonid Bogachev. (2023). The joint node degree distribution in the Erdős-Renyi network.