GRAPH SAMPLING: ESTIMATION OF DEGREE DISTRIBUTIONS

Joya A. Deri and José M. F. Moura

Carnegie Mellon University
Department of Electrical and Computer Engineering
Pittsburgh, PA 15213 USA
E-mail: {jderi,moura}@ece.cmu.edu

ABSTRACT

Online social networks and the World Wide Web lead to large underlying graphs that might not be completely known because of their size. To compute reliable statistics, we have to resort to sampling the network. In this paper, we investigate four network sampling methods to estimate the network degree distribution and the so-called biased degree distribution of a 3.7 million wireless subscriber network. We measure the quality of our estimates of the degree distributions by using the Kolmogorov-Smirnov statistic. Among all four sampling methods, node sampling yields Pareto optimal sample sizes in terms of the Kolomogorov-Smirnov statistic for the degree distribution, while node-by-edge sampling yields optimal sample sizes for the biased distribution. We also find that random walk sampling performs better than the Metropolis-Hastings random walk.

Index Terms— Graph sampling, Markov Chain Monte Carlo (MCMC) sampling, Pareto optimality, large-scale networks

1. INTRODUCTION

Online social networks such as Facebook and Twitter can have millions of nodes, and for networks such as the World Wide Web, knowledge of the entire network may be impossible. Such large networks have motivated the study of graph sampling with the goal of obtaining smaller, manageable subgraphs that are representative of the original network (see, e.g., [1, 2, 3]). In this paper we ask which sampling method is optimal with respect to a metric (see next paragraph). We apply four standard sampling methods: node, node-by-edge, random walk and Metropolis-Hastings random walk sampling [1, 3].

A common method for determining the optimality of a sampling scheme is to identify the minimum sample size at which we obtain a subgraph with the desired properties. We try a multiobjective approach based on Pareto optimality [4], where we want simultaneously to minimize the metric statistic and the sample size. We perform this analysis by computing metrics over a range of sample sizes and locating the knee, or point of maximum curvature, as a suitable compromise between metric value and sample size. The sampling method that attains both the lowest sample size and the lowest metric at its knee is optimal.

We seek sample sizes that yield accurate estimation of degree distributions. For an undirected graph $G = G(V, E)$ with node set V and edge set E, the degree distribution p is defined as

$$p_k = \frac{n_k}{N}, \tag{1}$$

where n_k is the number of nodes with degree k and N is the total number of nodes. We also consider a so-called biased degree distribution $\widetilde{p}_k$ that is given by

$$\widetilde{p}_k = \frac{k p_k}{\langle k \rangle}, \tag{2}$$

where k is the degree, p_k is the degree distribution value for degree k, and $\langle k \rangle = \sum_k k p_k$ is the mean degree [5]. Previous work such as [1, 6, 7] have focused on attaining estimates of p_k. In this paper we investigate which of the four network sampling methods leads to minimum sample sizes such that representative subgraphs yield accurate estimates of both the degree distribution and the biased degree distribution.

These distributions and their generating functions can be used to estimate the diameter and giant component size if the network obeys properties of the configuration model - namely, node degrees are independent and the probability of cycles negligible [5, 8]. Even when such conditions do not hold, such as for real-world networks, the biased degree distribution defines the probability of traversing the neighbors of a node's neighbors [5]; thus, the two degree distributions together give us information about the connectivity of 2-hop neighborhoods. We believe the biased degree distribution is a meaningful statistic to study for these reasons.

The remainder of the paper is as follows. Section 2 presents the sampling methods we consider and derives the expected sample degree distributions. Section 3 discusses the expected unbiased and biased degree distribution maximum

This work has been supported by AFOSR grant FA95501010291 and by NSF grants CCF1011903 and CCF1018509.

likelihood estimates for each sampling method. Section 4 presents empirical results, focusing on the knees of the metric curves obtained from both estimates. We discuss key results in Section 5, and we summarize and outline future work in Section 6.

2. SAMPLING METHODS

We describe the sampling methods and for each method specify the probability of choosing a node with fixed degree. We use these probabilities to define the expected sample degree distributions in Section 2.5.

2.1. Node Sampling (NS)

Node sampling is a standard sampling technique [1, 3]. We sample a node $v \in V$ uniformly at random with probability $1/N$. We repeat until we collect n nodes. The probability of choosing a node v of degree $k_v = k$ is

$$\pi_k = \mathbb{P}\left(v \mid k_v = k\right) = \frac{1}{N}.$$

2.2. Node-by-Edge Sampling (NES)

Node-by-edge sampling is another standard sampling technique [3]. We randomly select an edge $e \in E$ with probability $p_{\mathrm{e}} = 1/M$, where M is the number of elements in E. Then, one of the nodes connected to e is chosen with equal probability. We repeat until we collect n nodes.

We show that the probability of selecting a node of fixed degree is linear with degree for small p_{e}. A node v with degree k will not be chosen in two cases: (i) none of its edges are picked; or (ii) its edges are picked but the other connected nodes are selected. The probability of not picking v via its first edge is $(1 - p_{\mathrm{e}}) + p_{\mathrm{e}}/2 = 1 - p_{\mathrm{e}}/2$. Thus the probability of sampling a node of degree k is

$$\pi_k = 1 - \left(1 - \frac{p_{\mathrm{e}}}{2}\right)^k. \tag{3}$$

We fix degree k and form the first-order Taylor series around $p_{\mathrm{e}} = 0$, which yields

$$\pi_k\left(p_{\mathrm{e}}\right) \approx \frac{1}{2} k p_{\mathrm{e}}. \tag{4}$$

This approximation holds for small $p_{\mathrm{e}} = 1/M$. Assuming M grows at least as fast as $\mathcal{O}(N)$ as N increases, we see that, for large enough N, $\pi_k = \frac{k}{2M}$.

2.3. Random Walk (RW)

We select a seed node $v \in V$ uniformly at random. As in [1], we select the next node $w \in V$ conditional on the current node v with probability

$$\mathbb{P}\left(w \,|\, v\right) = \begin{cases} \frac{1}{k_v} & \text{if } (v, w) \in E \\ 0 & \text{otherwise} \end{cases}, \tag{5}$$

where k_v is the degree of node v. For a connected, aperiodic network, the probability of selecting a node given its degree k converges (as $N \to \infty$) to the steady-state probability distribution $\pi_k = k/2M$ [1].

2.4. Metropolis-Hastings Random Walk (MHRW)

The procedure follows as in [1]: we select a seed node v and select the next node $w \in V$ with probability

$$\mathbb{P}\left(w \,|\, v\right) = \begin{cases} \frac{1}{k_v} \min\left(1, \frac{k_v}{k_w}\right) & \text{if } (v, w) \in E,\ w \neq v \\ 1 - \sum_{i, i \neq v} \mathbb{P}\left(i \,|\, v\right) & \text{if } v = w \\ 0 & \text{otherwise} \end{cases}. \tag{6}$$

This method weights the transition probabilities so that nodes with low degree are visited more frequently than in the random walk. The resulting steady-state distribution for the Metropolis-Hastings walk (on a connected, aperiodic network) is $\pi_k = 1/N$ as in node sampling [1].

2.5. Expected Sample Degree Distributions

We denote by $\widehat{p}_k$ the sample (empirical) estimate of the degree distribution. The expected sample degree distributions for the sampling methods in this paper arise directly from the general expected distribution derived in [9] (see also [8, 10, 11]). The expected sample degree distribution is given by

$$\mathbb{E}\left[\widehat{p}_k\right] = \frac{\sum_{l \geq k}^{\infty} p\left(k \,|\, l\right) \pi_l p_l}{\sum_{l=0}^{\infty} \pi_l p_l}, \tag{7}$$

where: p_k is the probability of choosing a node of degree k in the original graph; $\pi_k\left(v\right)$ is the probability of choosing a node $v \in V$ given it has degree k in the original graph; $p\left(k \,|\, l\right)$ is the conditional probability of choosing a node that has degree k in the sampled network given that its original degree is l; and the sums are over all degrees in the network. We assume that the sampled nodes retain their original degrees, which is known as unlabeled star sampling [12], so the conditional probability becomes an indicator function of degree and we get

$$\mathbb{E}\left[\widehat{p}_k\right] = \frac{\sum_{l \geq k}^{\infty} \delta\left(l - k\right) \pi_l p_l}{\sum_{l=0}^{\infty} \pi_l p_l} = \frac{\pi_k p_k}{\sum_{l=0}^{\infty} \pi_l p_l}. \tag{8}$$

The π_k in (8) depends on the sampling method. We consider the two cases that apply here: constant π_k and linear π_k.

Suppose the sampling method selects nodes uniformly at random with some probability p, i.e., $\pi_k = p$. Then we apply (8):

$$\mathbb{E}\left[\widehat{p}_k\right] = \frac{p \cdot p_k}{\sum_{l=0}^{\infty} p \cdot p_l} = \frac{p_k}{\sum_{l=0}^{\infty} p_l} = p_k, \tag{9}$$

so we expect to recover an unbiased estimate of the original degree distribution when π_k is constant.

Now suppose the sampling method selects nodes propor-

tional to their degree, i.e., $\pi_k = Ck$ for some constant C. Then we apply (8):

$$\mathbb{E}\left[\widehat{p}_k\right] = \frac{Ckp_k}{\sum_{l=0}^{\infty} Clp_l} = \frac{kp_k}{\sum_{l=0}^{\infty} lp_l} = \frac{kp_k}{\langle k \rangle}, \tag{10}$$

so we expect to recover an unbiased estimate of $\widetilde{p}$. By inspection of π_k derived in Sections 2.1–2.4, we see that the expected NS and MHRW estimates are the degree distribution, while the expected NES and RW estimates are the biased degree distribution.

3. MAXIMUM LIKELIHOOD ESTIMATORS

The maximum likelihood degree distribution estimates for NS and NES before re-weighting are derived in [13] by estimating p_k, the probability of choosing a node with degree k, as parameters of a multinomial distribution. The maximum likelihood estimates are

$$\widehat{p}_{k,\mathrm{NS}} = \widehat{\widetilde{p}}_{k,\mathrm{NES}} = \frac{n_k}{n}, \tag{11}$$

where the estimate for NES is weighted as seen in Section 2.5.

For MCMC methods RW and MHRW, we assume the multinomial steady-state model [14], which yields the same maximum likelihood estimates as in (11):

$$\widehat{p}_{k,\mathrm{MHRW}} = \widehat{\widetilde{p}}_{k,\mathrm{RW}} = \frac{n_k}{n}. \tag{12}$$

We next convert the $\widetilde{p}_k$ estimates to p_k estimates. We use the Horvitz-Thompson estimator, which is used to compute population totals when elements are randomly chosen with probability proportional to their size [15]. The Horvitz-Thompson estimate of the mean degree $\langle k \rangle$ is

$$\widehat{\langle k \rangle}_{\mathrm{H}} = \frac{n}{\sum_{v \in V_S} 1/d_v}, \tag{13}$$

where V_S is the set of sampled nodes, $n = |V_S|$ is the sample size, and d_v is the degree of node v. We can show that this estimator is also the maximum likelihood degree estimator for NES and RW and is asymptotically unbiased (not shown for space). By the functional invariance of maximum likelihood estimates, the degree distributions are given by

$$\widehat{p}_{k,\mathrm{NES}} = \widehat{p}_{k,\ \mathrm{RW}} = \frac{\widehat{\langle k \rangle}_{\mathrm{HH}} \widehat{\widetilde{p}}_k}{k}. \tag{14}$$

We similarly find the maximum likelihood biased estimates for NS and MHRW:

$$\widehat{\widetilde{p}}_{k,\mathrm{NS}} = \widehat{\widetilde{p}}_{k,\mathrm{MHRW}} = \frac{kp_k}{\widehat{\langle k \rangle}}, \tag{15}$$

where $\widehat{\langle k \rangle} = \sum_k k\widehat{p}_k$ is the maximum likelihood estimate of the mean degree.

We use the Kolmogorov-Smirnov statistic (KS) to compare the degree distribution estimates. It is defined as $\mathrm{KS}\left(p \| \widehat{p}\right) = \max_k \left| F\left(k\right) - \widehat{F}\left(k\right) \right|$ where F and $\widehat{F}$ are the cumulative distributions of the probability distributions p and its estimate $\widehat{p}$ respectively. The KS statistic compares the shape of the distributions without accounting for scaling (see, e.g., [3]).

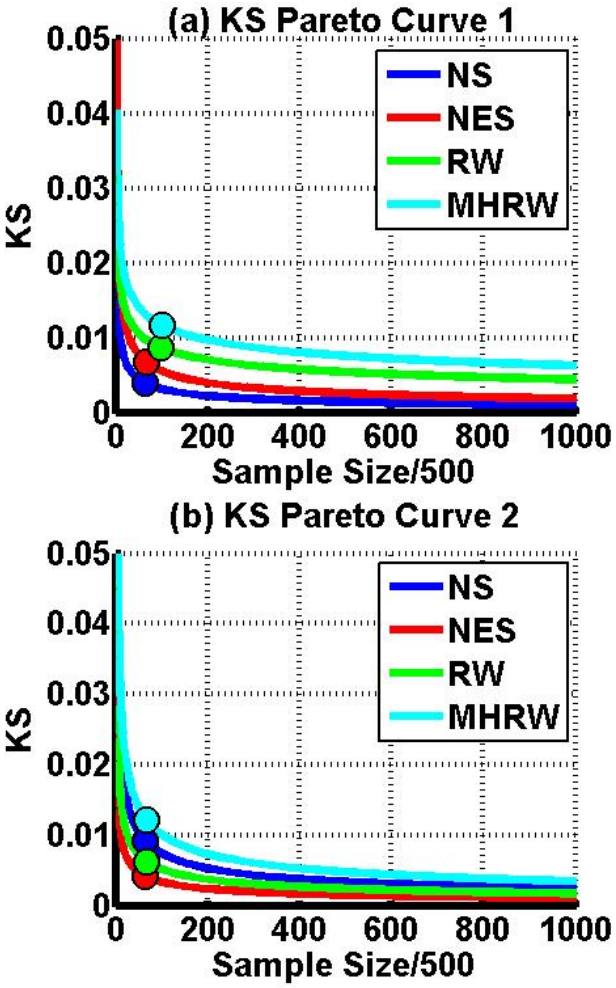

Fig. 1. KS Pareto curves for p_k (a) and $\widetilde{p}_k$ (b). The circles mark the knee points (τ, μ) and $(\widetilde{\tau}, \widetilde{\mu})$.

4. EMPIRICAL RESULTS

We sample a caller network of 3.7 million wireless subscribers. We model the data as a network of active subscribers, where an active subscriber makes at least one in-network call in a single month period. The maximum degree is > 3000 and the average degree is 12.

For the MCMC methods we would like to ensure that the random walks have converged before we collect our samples.

	Knee Points for KS Metric Curves			
Sampling Method	τ	μ	$\widetilde{\tau}$	$\widetilde{\mu}$
NS	$\mathbf{32,500}$	$\mathbf{0.0053}$	$\mathbf{32,500}$	0.0084
NES	$33,500$	0.0068	$\mathbf{32,500}$	$\mathbf{0.0052}$
RW	$48,500$	0.0130	$33,500$	0.0094
MHRW	$50,500$	0.0081	$34,000$	0.0086

Table 1. Unbiased and biased knee points for the KS metric curves. The minimum points in each column are in boldface.

To do this, we implement a burn-in, which is an initial walk on the network that is not included in the sample [1]. For our purposes we use a burn-in of 800 for both MCMC methods. We calculate the KS statistic for sample sizes up to 1000 and average over 500 Monte Carlo iterations, which yields an effective maximum sample size of $500,000$.

Determining the optimal sampling size is a multiobjective optimization. We would like to minimize: (i) the metric KS because a lower value indicates a better estimate, and (ii) the sample size. This multiobjection leads to a Pareto optimality problem [4]. We denote the optimal sample size/metric pair by (τ, μ) for p_k and by $(\widetilde{\tau}, \widetilde{\mu})$ for $\widetilde{p}_k$.

We plot the Pareto curves that show the KS statistics versus sample size in Fig. 1. Every point on this curve represents a good compromise between KS and the sample size. We choose the knee of this curve, as the point closest to the origin. To detect the knee, we iteratively partition the metric curves over a range of sample sizes and compute the left-hand and right-hand linear regressions; the knee occurs at the sample size that minimizes the root mean-square error of the regressions as explained in [16]. From Fig. 1 it is clear that NS sampling minimizes (τ, μ) for p_k and that NES sampling minimizes $(\widetilde{\tau}, \widetilde{\mu})$ for $\widetilde{p}_k$. If we only consider the MCMC methods, we see that RW sampling yields the minimum pairs for both p_k and $\widetilde{p}_k$.

We verify the optimality of τ and $\widetilde{\tau}$ by computing the average KS values for these sample sizes. The values are listed in Table 1. We see that NS sampling minimizes both τ and μ and that NES sampling minimizes $\widetilde{\tau}$ and $\widetilde{\mu}$ as we expect. On the other hand, if we only consider the MCMC methods, we see that RW minimizes τ and $\widetilde{\tau}$, but MHRW minimizes μ and $\widetilde{\mu}$, so we cannot claim either as Pareto optimal based on the local average KS values. However, the Pareto curve allows us to see that RW is indeed optimal compared to MHRW.

We also computed τ and $\widetilde{\tau}$ values for Jensen-Shannon divergence [17] and the root mean-square error of the degree distributions (not shown). None of the four sampling methods yielded Pareto optimal points for these metrics.

5. DISCUSSION

Our results in Fig. 1 and Table 1 show that node sampling yields the Pareto optimal sample size $\tau = 32,500$ for p, while node-by-edge sampling yields the Pareto optimal sample size $\widetilde{\tau} = 32,500$ for $\widetilde{p}$. For both methods the optimal sample size is $32,500$, or about 0.88% of the total network size.

We note that there is no Pareto optimal method for estimating both p and $\widetilde{p}$. Instead we must weigh the trade-offs from the Pareto analysis to determine which method to use to estimate both p and $\widetilde{p}$. For example, Table 1 shows that $\widetilde{p}$ can be estimated under node sampling with a 62% increase in $\widetilde{\mu}$, and that p can be estimated that under node sampling with a 28% increase in μ and an additional 1000 samples. If we cannot afford to increase the sample size, then the optimal sampling method would be node sampling, for example.

As discussed in [7], the sampling method of choice for networks of unknown structure is determined in part by the available information. In particular, very large networks may not be stored as a complete list (sampling frame) of nodes and edges, in which case we must resort to methods that explore local neighborhoods such as RW and MHRW. For our goal of estimating both p_k and $\widetilde{p}_k$ for very large networks, our results suggest that RW is preferable, although we cannot yet infer Pareto optimality for general networks.

6. CONCLUSION

We evaluate two simple random sampling methods and two Markov chain Monte Carlo methods based on estimates of the degree distribution p and the biased degree distribution $\widetilde{p}$. We generate Pareto curves for the KS statistic as a function of sample size and use these curves to identify Pareto optimal sampling size/metric pairs and the corresponding optimal sampling methods. For the 3.7 million caller network we find that node sampling is optimal for estimating p and that node-by-edge sampling is optimal for estimating $\widetilde{p}$. In addition, we see that RW is optimal for estimating both degree distributions with respect to MHRW. Future work includes studying how Pareto optimality can be extended to general networks and also investigating a wider range of sampling methods, including respondent-driven sampling.

7. ACKNOWLEDGEMENTS

We would like to thank Pavel V. Krivitsky and Pedro Ferreira from Carnegie Mellon University for their discussions. We would also like to acknowledge i-Lab of Carnegie Mellon University for providing access to the mobile caller dataset.

8. REFERENCES

[1] M. Gjoka, M. Kurant, C.T. Butts, and A. Markopoulou, "Walking in Facebook: A case study of unbiased sampling of OSNs," in *29th IEEE Conf. on Computer Communications (INFOCOM)*, March 2010, pp. 1–9.

[2] A.S. Maiya and T.Y. Berger-Wolf, "Sampling community structure," in *19th International Conf. on World Wide Web*, April 2010, pp. 701–710.

[3] J. Leskovec and C. Faloutsos, "Sampling from large graphs," in *12th ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD)*, 2006, pp. 631–636.

[4] D. Fudenberg and J. Tirole, *Game Theory*, MIT Press, 2000.

[5] M.O. Jackson, *Random-Graph Models of Networks: Properties of Random Networks*, Princeton Review Press, 2008.

[6] D. Stutzbach and R. Rejaie, "Sampling techniques for large, dynamic graphs," in *25th IEEE International Conf. on Computer Communications (INFOCOM)*, Apr. 2006, pp. 1–6.

[7] M. Kurant, M. Gjoka, C.T. Butts, and A. Markopoulou, "Walking on a graph with a magnifying glass: stratified sampling via weighted random walks," *SIGMETRICS Perform. Eval. Rev.*, vol. 39, no. 1, pp. 241–252, Jun. 2011.

[8] M.E.J. Newman, "The structure and function of complex networks," *Society for Industrial and Applied Mathematics (SIAM) Review*, vol. 45, no. 2, pp. 167–256, Jun. 2003.

[9] M.P.H. Stumpf and C. Wiuf, "Sampling properties of random graphs: The degree distribution," *Physical Review E*, vol. 72, pp. 036118, Sep. 2005.

[10] M.J. Salganik and D.D. Heckathorn, "Sampling and estimation in hidden populations using respondent-driven sampling," *Sociological Methodology*, vol. 34, pp. 193–239, 2004.

[11] M. Kurant, A. Markopoulou, and P. Thiran, "On the bias of BFS," *Arxiv e-print (arXiv:1004.1729v1)*, Apr. 2010.

[12] E.D. Kolaczyk, *Sampling and Estimation in Network Graphs*, Springer Series in Statistics, 2009.

[13] C.M. Bishop, *Multinomial Variables*, Springer, 2006.

[14] T.W. Anderson and L.A. Goodman, "Statistical inference about Markov chains," *Ann. Math. Statist.*, vol. 28, no. 1, pp. 89–110, 1957.

[15] D.G. Horvitz and D.J. Thompson, "A generalization of sampling without replacement from a finite universe," *Journal of the American Statistical Association*, vol. 47, pp. 663–685, 1952.

[16] S. Salvador and P. Chan, "Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms," in *Proc. of 16th IEEE Intl. Conf. on Tools with Artificial Intelligence (ICTAI)*, Nov. 2004, pp. 576–584.

[17] I. Grosse, P. Bernaola-Galván, P. Carpena, R. Román-Roldán, J. Oliver, and H.E. Stanley, "Analysis of symbolic sequences using the Jensen-Shannon divergence," *Physical Review E*, vol. 65, pp. 041905, 2002.