Respondent-driven Sampling for Characterizing Unstructured Overlays

Amir H. Rasti, Mojtaba Torkjazi, Reza Rejaie[‡]
University of Oregon
{amir, moji, reza}@cs.uoregon.edu

Nick Duffield, Walter Willinger
AT&T Labs—Research
{duffield, walter}@research.att.com

Daniel Stutzbach
Stutzbach Enterprises
agthorr@stutzbachenterprises.com

Abstract—**This paper presents Respondent-Driven Sampling (RDS) as a promising technique to derive unbiased estimates of node properties in unstructured overlay networks such as Gnutella. Using RDS and a previously proposed technique, namely Metropolized Random Walk (MRW) sampling, we examine the efficiency of estimating node properties in unstructured overlays and identify some of the key factors that determine the accuracy of sampling techniques. We evaluate the RDS and MRW techniques using simulation over a wide range of static and dynamic graphs as well as experiments over a widely deployed Gnutella network. Our study sheds light on how the connectivity structure among nodes and its dynamics affect the accuracy and efficiency of the two sampling techniques. Both techniques exhibit a rather similar performance over a wide range of scenarios. However, RDS significantly outperforms MRW when the overlay structure exhibits a combination of highly skewed node degrees and highly skewed (local) clustering coefficients.**

I. INTRODUCTION

During the past few years, unstructured Peer-to-Peer (P2P) systems such as Gnutella and BitTorrent have become very popular and have significantly contributed to the total traffic over the Internet. This has motivated researchers to characterize the basic properties of these systems through measurement. Such characterizations can be leveraged to address several key issues about these systems including: *(i)* understanding the properties and dynamics of these systems, and use these findings to improve their performance and scalability, and *(ii)* assessing the impact of these systems on the Internet.

To characterize unstructured P2P systems, one needs to capture accurate "snapshots" of the connectivity structure. Examining individual snapshots reveals the connectivity structure at a particular point of time whereas comparing consecutive snapshots over time illustrates the evolution of the connectivity structure. Such snapshots are typically captured by a crawler that queries a set of known nodes to learn about their neighbors and progressively discovers the connectivity structure. Capturing accurate snapshots of the connectivity structure for large-scale unstructured overlays is challenging because such systems may significantly evolve during the time required to capture a full snapshot. Therefore, captured snapshots are likely to be *distorted* and this could significantly degrade the accuracy of any results derived from such snapshots. In our earlier work [16], we have shown that commonly used sampling techniques in prior empirical studies on P2P systems (*e.g.*, [14]) can easily lead to significant bias towards short-lived or high degree peers due to the dynamics of peer participation or the heterogeneity of peer degrees, respectively.

This paper presents *Respondent-Driven Sampling (RDS)* as a promising technique for sampling unstructured P2P overlays. This allows one to accurately estimate the distribution of a desired peer property without capturing the entire overlay structure. RDS is a variant of snowball sampling that has been proposed and used in the social sciences to characterize hidden population in a society [9], [13]. We apply the RDS technique to unstructured P2P network and evaluate its performance over a wide range of static and dynamic graphs as well as a widely deployed P2P system. Throughout our evaluation, we compare and contrast the performance of the RDS technique with another sampling technique, namely *Metropolized Random Walk (MRW)*, that we developed in our earlier work [16].

Our main findings can be summarized as follows: First, RDS outperforms MRW across all scenarios. In particular, RDS exhibits a significantly better performance than MRW when the overlay structure exhibits a combination of highly skewed node degrees and highly skewed (local) clustering coefficients. Second, our simulation and empirical evaluations reveal that both the RDS and MRW techniques can accurately estimate key peer properties over dynamic unstructured overlays. Third, our empirical evaluations suggest that the efficiency of the two sampling techniques in practice is lower than in our simulations involving synthetic graphs. We attribute this to our inability to capture accurate reference snapshots.

The rest of the paper is organized as follows: Section II presents an overview of both the RDS and MRW techniques, and sketches our evaluation methodology. We examine both techniques over variety of static and dynamic graphs in Section III and IV, respectively. Section V presents the empirical evaluation of the two sampling techniques over Gnutella network.

II. GRAPH SAMPLING TECHNIQUES

An unstructured overlay can be represented as an evolving undirected graph G, with vertices V and edges E. The vertices and edges of G represent the peers and pairwise connections between them, respectively. An accurate snapshot of the full graph (the overlay) is not available, however we can query any known peer for a list of adjacent peers in order to progressively discover portions of the overlay. Some fraction of discovered peers are selected as samples and the distribution of the desired peer property (number of neighbors, number of files, access link bandwidth or session time) among the samples provides

an estimate for that property among all peers. The efficiency of sampling can be quantified by the ratio of sampled peers to the total number of peers queried. To provide an accurate estimate, sampled peers should be selected uniformly at random. This is challenging because the overlay topology and peer dynamics introduce bias towards discovery and thus selection of peers with large degrees and short session times, respectively [16].

Random walk is a promising technique for sampling. In an ordinary random walk, the sampler begins at a node, x, and chooses a new node, y, uniformly at random from x's neighbors. The walk transitions to the neighbor and then chooses a new node from y's neighbors. Formally, the ordinary random walk has a transition function, $P(x, y)$, defined as follows:

$$P(x, y) = \begin{cases} \frac{1}{\text{degree}(x)} & y \text{ is a neighbor of x,} \\ 0 & \text{otherwise} \end{cases}$$

The *stationary distribution*, $\pi(x)$, of the walk defines the probability of being at any particular node x. For an ordinary random walk, graph theory [10] proves $\pi(x) \propto \text{degree}(x)$. That is, the fraction of time spent at a node is directly proportional to the node's degree. Thus, the ordinary random walk is inherently biased towards nodes with higher degree.

A. Respondent Driven Sampling

Respondent Driven Sampling (RDS) is a development of Snowball Sampling (SBS) [9], a group of related sampling techniques proposed in the social sciences to sample hidden populations. Salganik [13] defines a population as "hidden" when there is no central directory of all population members, such that samples may only be gathered through iterative referrals from existing samples.

RDS is a variant of SBS [9], which forms asymptotically unbiased estimators by appropriate re-weighting of estimators to take account of topological biases [13]. The special case where each respondent recruits only one individual maps exactly onto the case of a random walk on a graph. This in turn can be recast as a Monte Carlo Markov Chain (MCMC) problem [5] The problem of estimating peer properties in unstructured overlays is analogous to the sampling of hidden population in the social sciences. We wish to estimate the distribution of a node property X; specifically, consider any partition $\{R_1, \ldots, R_m\}$ of the range of possible values of X. We partition the node set V accordingly into groups of nodes $\{V_1, \ldots, V_m\}$, i.e., $V_i = \{v \in V : X(v) \in R_i\}$. A simple example is when X is positive integer value and we group by value: $V_i = \{v \in V : X(v) = i\}$.

The RDS approach is to estimate the proportion p_i of nodes that are in group i from observed node degree and group memberships of nodes traversed in the random walk. Specifically, consider the n-step walk that visits the set of nodes $T = \{t_1, t_2, \ldots, t_n\}$ where individual nodes may be visited more than once. Let $T_i = T \cap V_i$ denote the visited nodes that lie in group i. For any node property X, the Hansen-Hurwitz [7] estimator $\hat{S}(X) := n^{-1} \sum_{v \in T} \frac{X(v)}{\pi(v)}$ is an unbiased and consistent estimator of the sum $S(X) := \sum_{v \in V} X(v)$ when T

is drawn from a stationary random walk, *i.e.*, one that evolves from an initial node that is randomly selected according to the stationary distribution. Consider two special cases. When $X = I_{V_i}$ is the indicator of a node being in group i, *i.e.*, $I_{V_i}(v) = 1$ if $v \in V_i$ and 0 otherwise, then $\hat{S}(I_{V_i})$ estimates the total number of nodes in V_i. When $X = 1$ then $\hat{S}(1)$ estimates the total number of nodes $|V|$ in the graph. Thus we can estimate the proportion p_i by

$$\hat{p}_i = \frac{\hat{S}(I_{V_i})}{\hat{S}(1)} = \frac{\sum_{v \in T_i} \frac{1}{\text{degree}(v)}}{\sum_{u \in T} \frac{1}{\text{degree}(u)}}$$

where $\text{degree}(v)$ is the degree of the node v. $\hat{p}_i$ is consistent—it converges to the true value p_i—as the number n of visited nodes grows. The RDS estimator can be recognized as an importance sampling estimator weighted by the stationary distribution π, applied to the MCMC of the random walk on the vertex set V.

B. Metropolized Random Walk

Our earlier work [16] evaluates the use of Metropolized Random Walks (MRW) for gathering unbiased samples from unstructured P2P networks. The Metropolis–Hastings technique [4], [8], [11] provides a way to alter the next-hop selection to produce any desired stationary distribution, $\pi(x)$. In [16], we choose the next-hop appropriately to produce the uniform distribution, $\pi(x) = \frac{1}{|V|}$, as follows:

$$Q(x, y) = \begin{cases} P(x, y) \min \left(\frac{\text{degree}(x)}{\text{degree}(y)}, 1 \right) & \text{if } x \neq y, \\ 1 - \sum_{z \neq x} Q(x, z) & \text{if } x = y \end{cases}$$

Essentially, the walk tentatively selects a neighbor of x uniformly at random ($P(x, y)$) and then accept the transition randomly with probability $\min \left(\frac{\text{degree}(x)}{\text{degree}(y)}, 1 \right)$. Otherwise $(1 - \sum_{z \neq x} Q(x, z))$, the walk remains at the current node, effectively taking a self-edge. Put simply, the bias toward higher degree nodes is removed by reducing the probability of transitioning to higher degree nodes at each step.

We note that RDS is complementary to the MRW approach in the following way. In MRW, we seek to modify the random walk in order to have an equal probability of visiting each node and hence derive unbiased estimates. In RDS, the walk is unmodified; however, we reweight the sampled values to obtain an unbiased estimate of the group proportions p_i.

C. Evaluation Methodology

To evaluate the RDS technique, first we simulate the sampling techniques over a wide range of static and dynamic graphs where the accurate distribution of the sampled property (ground truth) is known. Simulation over synthetic graphs not only offers an opportunity for accurate evaluation of the sampling techniques, but also allows us to identify the separate effects of graph properties and graph dynamics on the accuracy and efficiency of these techniques. Second, we empirically evaluate both techniques over Gnutella P2P overlay.

Performance Metric: To quantify the accuracy of a sampling technique in each scenario, we compare the sampled and true

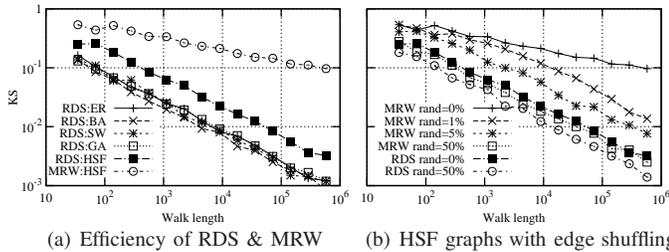(a) Efficiency of RDS & MRW   (b) HSF graphs with edge shuffling

Fig. 1. Efficiency of RDS and MRW techniques in estimating degree distribution over different graph types

distributions of a desired peer property using the Kolmogorov-Smirnov (KS) statistic, D. If we plot the estimated and true CDFs of a desired property, D is the maximum vertical distance between the plots of the two functions with a range of $[0, 1]$. For example, a value of $D \leq 0.01$ corresponds to no more than a one percentage point difference between CDFs and is excellent for most measurement purposes. Due to the limited space, we only present a subset of our results that illustrate our main findings in Sections III and IV. Complete results are available in the related technical report [12].

III. EVALUATION OVER STATIC GRAPHS

In this section, we examine how the connectivity structure of a graph affects the accuracy and efficiency of the RDS and MRW sampling techniques using the following candidate graph types: *(i) Erdös-Rényi Random graphs* (ER) [3], *(ii) Small-world graphs* (SW) [18], *(iii) Barabási-Albert graphs* (BA) [1]: Scale-free graphs of the preferential attachment-type, *(iv) Hierarchical Scale-Free graphs* (HSF) [2]: A class of (deterministic) graphs generated by an iterative algorithm to produce heterogeneous node degree and heterogeneous node clustering coefficients. More specifically, node degree distribution follows power-law while clustering coefficients at individual nodes is inversely proportional to node degree, independent of graph size. *(v) Gnutella graphs (GA)*: A snapshots of the Gnutella ultrapeer topology, captured on 05/15/2008 using *cruiser* [16].

Figure 1(a) shows the KS error for the degree distribution from samples collected by the RDS and MRW techniques as a function of the number of samples over the different graph types. To make the results comparable, the number of vertices ($|V| = 390,625$) and edges ($|E| = 1,769,110$) are similar across the different graph types. Figure 1(a) illustrates the following two important points. First, the accuracy of the RDS technique rapidly improves with the number of samples. The rate of improvement in accuracy across all graph types (*i.e.*, slope of the line) is similar. The overall accuracy of the MRW technique follows a trend similar to RDS for all graphs (except the HSF graph) but on average slightly ($\approx 2 \cdot 10^{-3}$) lower than the RDS technique. Given this similarity, the results for MRW are not shown in Figure 1(a) except for the HSF graph. For the HSF graph, MRW sampling not only exhibits a significantly lower accuracy compared to the other graph types, but the rate of improvement in accuracy with the sample size (*i.e.*, slope) is much worse. Second, for a given number of samples,

while both techniques exhibit a lower accuracy for the HSF graph, the impact on the MRW technique is significantly more pronounced, *i.e.*, the rate of improvement in accuracy with the sample size (*i.e.*, slope) for the MRW technique is much worse than RDS.

Focusing on the HSF graph, the reported differences in the accuracy of RDS and MRW and their observed lower performance can be attributed to the following phenomenon. In HSF graphs, at each level of their hierarchical structure, there are groups of well inter-connected low degree nodes which form pronounced *clusters*. The only way for a random walker to leave these clusters is via a much higher degree node that resides outside these clusters, *i.e.*, the walker has to traverse an edge from a low degree node within such a cluster to a much higher degree node outside this cluster. As described in Section II, for the MRW technique, the probability of moving along such an edge is proportional to the ratio of the (low) degree of the node within the cluster to the (very high) degree of the node outside this cluster which is very small. Therefore, when an MRW walker ends up in one of these clusters, it keeps collecting samples from low degree nodes within these clusters for a disproportionally long time. This in turn degrades the accuracy of sampling especially among high degree nodes. The impact of clusters on the RDS technique is significantly lower because the probability of selecting the next node in RDS does not depend on node degree.

Figure 1(b) shows the accuracy of the MRW sampling technique over the same HSF graph when 0%, 1%, 5%, and 50% of its edges are randomly shuffled (*i.e.*, rewired) while preserving the degree of individual nodes. Increasing the percentage of randomly shuffled edges gradually removes the explicit hierarchical structure of HSF graphs and enforces a more homogeneous clustering behavior across the graph structure as compared to the original HSF graph. For comparison, we also present the results for the RDS technique over the graphs when 0% and 50% of edges are shuffled. The figure demonstrates that even a small percentage of shuffled edges dramatically improves the accuracy of the MRW technique. These results suggest that the main reason for the degraded performance of the two sampling techniques over HSF graphs is a combination of highly skewed node degrees and highly skewed node clustering coefficients.

IV. EVALUATION OVER DYNAMIC GRAPHS

In this section, we use our session-level simulator [15], called *psim*, to examine the behavior of the RDS technique over dynamic graphs. *psim* simulates peer arrivals, departures, pairwise latencies, per discovery and neighbor connections. The latencies between peers are randomly selected from the King data set [6]. Peers use the following popular boot-strapping mechanisms for peer discovery [16]: *Oracle, FIFO, HeartBeat* and *History*. Individual peers try to maintain the number of their connections (*i.e.*, their degree) between a given minimum ($MinDeg$) and maximum ($MaxDeg$) degree. When the number of connections for a peer drops below
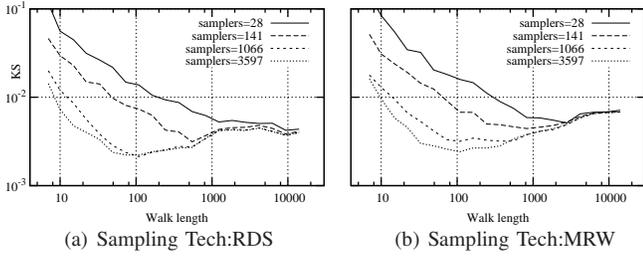
(a) Sampling Tech:RDS   (b) Sampling Tech:MRW

Fig. 2. Effect of number of walkers and walk length in sampling node degree. Churn Model: Weibull with $k = 0.59$, median sess. len.=21min, MinDeg=30

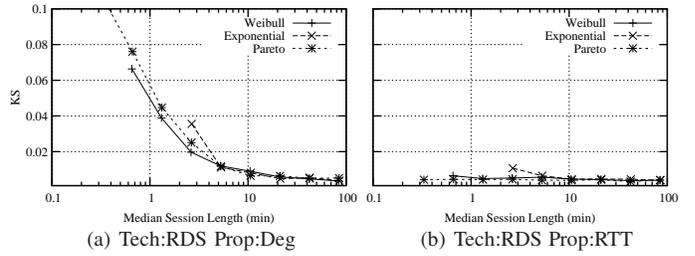

(a) Tech:RDS Prop:Deg   (b) Tech:RDS Prop:RTT

Fig. 3. Sampling error for two peer properties as a function of *median session length* for different churn models. Bootstrapping=FIFO, MinDeg=30, sampled by 1066 parallel samplers, each taking 49 hops

$MinDeg$, it uses the discovery mechanism to establish additional connections and reach $MinDeg$. A peer neither accepts nor initiates any new connections once its degree reaches $MaxDeg$. To query a peer for a list of neighbors, the sampling node must establish a TCP connection, submit its query, and receive a response. *psim* simulates churn by controlling the distributions of peer inter-arrival intervals and peer session lengths. New peers arrive according to a Poisson process, where the mean peer arrival rate combined with the session length distribution yield a desired mean population size in steady state. We use the following models for session length distribution that we derived in our earlier empirical study of churn [15] in P2P networks: Weibull, Pareto and Exponential. We run each simulation for a warm-up period until it reaches steady state with 100,000 concurrent peers before gathering samples.

Impact of Parallel Sampling: A desired number of samples from a dynamic overlay can be collected by a number of parallel (RDS or MRW) walkers that start from the same nodes. Increasing the number of parallel samplers has two conflicting effects and thus introduces an interesting tradeoff. Increasing the number of parallel walkers, reduces the required walk length to collect a desired number of samples. This in turn decreases the time to collect the samples and thus reduces the error that occurs due to the evolution (*i.e.*, churn) in the overlay. However, increasing the number of samplers leads to redundant sampling of nodes around the starting point and degrades sampling accuracy. Figure 2 demonstrates this tradeoff and depicts the accuracy of the RDS and MRW techniques as a function of walk length for different number of parallel samplers.

Clearly, the accuracy of the RDS and MRW techniques in estimating a peer property is not affected by overlay dynamics if the desired peer property does not interact with the walk. Therefore, to evaluate these techniques over dynamic graphs, we only consider the following peer properties that may interact with the walk: *(i) Node Degree (DEG)*: The degree of an individual node in the graph determines the probability that a node is visited. *(ii) Session length or Uptime (UT)*: The dynamics of peer participation drives the evolution of the graph with time and affect the probability of visit for individual peers. *(iii) Query latency (RTT)*: In a dynamic overlay, each step requires querying a peer. Since the query latency for individual peers depends on their relative round-trip time, this could lead to a bias correlated with the query latency.

A. Effect of Churn

Figure 3(a) depicts the accuracy of the RDS technique in estimating the distribution of node degree as a function of median session length (*i.e.*, churn rate) for different churn models. The results for sampling session length is very similar to Figure 3(a) and thus is not shown. Figure 3(a) shows that the median session length is the primary factor that affects the accuracy of sampling techniques. To explain this behavior, we note that the median session length is a rough measure of the level of overlay dynamics. When the churn rate is high (*i.e.*, median session length is less than 5 minutes), the overlay significantly evolves during the sampling period which in turn leads to larger error. Earlier empirical studies suggest that the median session length in actual P2P systems is rarely below 10 minutes for which the sampling error is below 0.01. Figure 3(b) presents the accuracy of the RDS technique in estimating the distribution of *query latency*. Since the query latency between pairs of nodes are selected from the fixed King data set, its distribution among samples is less sensitive to the dynamics of peer participation.

B. Effect of Target Node Degree

Figure 4(a) presents the accuracy of the RDS sampling technique in estimating the distribution of node degree as a function of minimum node degree ($MinDeg$). This figure reveals that when $MinDeg$ is larger than a threshold of about five, the accuracy does not change with the the minimum node degree (except for the *History* bootstrapping mechanism). Figure 4(b) shows that the accuracy of the RDS technique in estimating query latency follows a similar pattern. The rapid degradation of accuracy for lower node degrees is mainly due to the fragmentation of the overlay which makes some parts of the graph inaccessible to the random walkers. In real P2P systems, such a fragmentation does not occur since the peer degree is often larger than five.

To explain the abnormal behavior of the *History* bootstrapping mechanism in Figure 4, we note that in this mechanism each peer relies on the list of its neighbors in previous sessions. This leads to a number of isolated peers since all their neighbors from previous sessions have departed. While the number of isolated peers is not large at any given time, and they eventually get connected to the overlay by contacting bootstrapping node, their extended isolation time have an

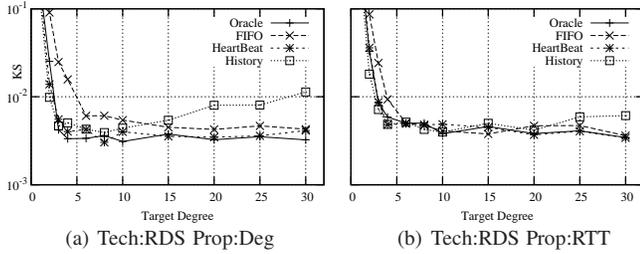(a) Tech:RDS Prop:Deg   (b) Tech:RDS Prop:RTT

Fig. 4. Sampling error for two peer properties as a function of *MinDeg* using different bootstrapping mechanisms. Churn Model: Weibull $k = 0.59$, Med. sess. len.=21min, 1066 parallel samplers, each taking 49 hops

(a) Degree distribution   (b) Node degree sampling error

Fig. 5. Results of sampling experiment over Gnutella

impact on the reachability of these nodes and thus on the accuracy of sampling techniques.

V. EVALUATION OVER GNUTELLA

To empirically evaluate our sampling techniques, we use them to estimate properties of ultrapeers in the Gnutella network. We incorporate both sampling techniques into our sampling tool called *ion-sampler* [16]. We concurrently start 1000 RDS and 1000 MRW samplers, where each sampler takes a 500-step walk to sample the degree of Gnutella ultrapeers. At the same time, we use *cruiser* [17] to collect complete back-to-back snapshots of the top-level Gnutella overlay roughly every seven minutes.

Figure 5(a) presents the distribution of node degree from collected samples by the RDS and MRW techniques as well as full snapshots collected by the crawler. Figure 5(a) shows that all three distributions of node degree are almost indistinguishable, *i.e.*, both sampling techniques exhibit similar performance. To further investigate the variability of observed accuracy for the sampling techniques, we repeat each sampling experiment with different walk length for six times. Figure 5(b) presents the average KS error and associated error bars as a function of walk length for both sampling techniques. Figure 5(b) indicates that increasing the walk length beyond about 30 hops quickly decreases the KS error because of the larger number of collected samples. However, the rate (*i.e.*, slope) of improvement in accuracy is diminishing beyond a certain walk length. Collecting more samples through longer walks does not improve the fidelity of samples due to major changes in the system during the sampling period.

VI. CONCLUSION AND FUTURE WORK

In this paper, we presented RDS as a powerful technique for sampling unstructured P2P networks. While RDS has been developed in the social sciences for sampling static graphs, we adopted this technique to the networking domain and explored its applicability in the context of dynamic connectivity structures. Through simulations involving a variety of synthetically generated static and dynamic graphs and experiments over the Gnutella network, we examined the performance of the RDS technique and compared its performance with another graph sampling technique, MRW. Our study demonstrates how the connectivity structure among nodes and its dynamics affect the accuracy of both sampling techniques. We showed that RDS 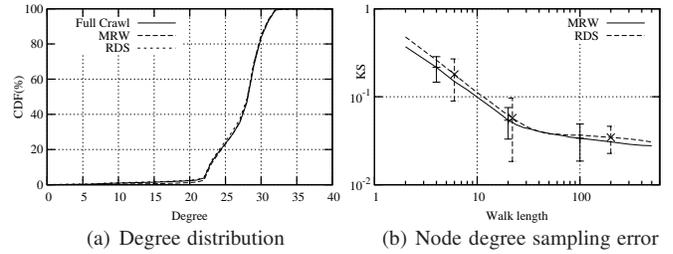generally performs as good or better than MRW. In particular, RDS achieves a significantly better performance than MRW when the overlay structure exhibits a combination of highly skewed node degrees and highly skewed node clustering coefficients.

We are currently investigating the performance of RDS when sampling the connectivity of the undirected friendship graph in large Online Social Networks (OSN) such as MySpace and LiveJournal. We are also examining in more detail the advantages of parallel sampling. Finally, we plan to extend our sampling technique to "directed" graphs that are common in OSNs.

REFERENCES

[1] A.-L. Barabási and R. Albert. Emergence of Scaling in Random Networks. *Science*, 286:509–512, Oct. 1999.

[2] A.-L. Barabasi, Z. Dezso, E. Ravasz, S.-H. Yook, and Z. Oltvai. Scale-free and Hierarchical Structures in Complex Networks. In *Seventh Granada Lectures*, 2002.

[3] B. Bollobás. A probabilistic Proof of an Asymptotic Formula for the Number of Labelled Regular Graphs. *European Journal of Combinatorics*, 1:311–316, 1980.

[4] S. Chib and E. Greenberg. Understanding the Metropolis–Hastings Algorithm. *The Americian Statistician*, 49(4):327–335, Nov. 1995.

[5] S. Goel and M. J. Sagalnik. Respondent-driven sampling as markov chain monte carlo. Manuscript, 2007.

[6] K. P. Gummadi, S. Saroiu, and S. D. Gribble. King: Estimating Latency between Arbitrary Internet End Hosts. In *Internet Measurement Workshop*, Nov. 2002.

[7] M. Hansen and W. Hurwitz. On the Theory of Sampling from Finite Populations. *Annals of Mathematical Statistics*, 14(4):333–362, 1943.

[8] W. Hastings. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57:97–109, 1970.

[9] D. Heckathorn. Respondent-Driven Sampling II: Deriving Valid Population Estimates from Chain-Referral Samples of Hidden Populations. *Social Problems*, 49(1):11–34, 2002.

[10] L. Lovász. Random walks on graphs: A survey. *Combinatorics: Paul Erdös is Eighty*, 2:1–46, 1993.

[11] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equations of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

[12] A. Rasti, R. Rejaie, N. Duffield, D. Stutzbach, and W. Willinger. Evaluating Sampling Techniques for Large Dynamic Graphs. Technical report, University of Oregon, Mar. 2008.

[13] M. Salganik and D. Heckathorn. Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology*, 34:193–239, 2004.

[14] S. Saroiu, P. K. Gummadi, and S. D. Gribble. A Measurement Study of Peer-to-Peer File Sharing Systems. In *MMCN*, Jan. 2002.

[15] D. Stutzbach and R. Rejaie. Understanding Churn in Peer-to-Peer Networks. In *Internet Measurement Conference*, Oct. 2006.

[16] D. Stutzbach, R. Rejaie, N. Duffield, S. Sen, and W. Willinger. On Unbiased Sampling for Unstructured Peer-to-Peer Networks. *TON*, 2008.

[17] D. Stutzbach, R. Rejaie, and S. Sen. Characterizing Unstructured Overlay Topologies in Modern P2P File-Sharing Systems. *TON*, 16(2), Apr. 2008.

[18] D. J. Watts. Six Degrees. In *The Essence of a Connected Edge*. ACM Press, 2003.