Walking in Facebook: A Case Study of Unbiased Sampling of OSNs

Minas Gjoka
Networked Systems
UC Irvine
mgjoka@uci.edu

Maciej Kurant
School of Comp.& Comm. Sciences
EPFL, Lausanne
maciej.kurant@epfl.ch

Carter T. Butts
Sociology Dept
UC Irvine
buttsc@uci.edu

Athina Markopoulou
EECS Dept
UC Irvine
athina@uci.edu

***Abstract*—With more than 250 million active users [1], Facebook (FB) is currently one of the most important online social networks. Our goal in this paper is to obtain a representative (unbiased) sample of Facebook users by crawling its social graph. In this quest, we consider and implement several candidate techniques. Two approaches that are found to perform well are the Metropolis-Hasting random walk (MHRW) and a re-weighted random walk (RWRW). Both have pros and cons, which we demonstrate through a comparison to each other as well as to the "ground-truth" (UNI - obtained through true uniform sampling of FB userIDs). In contrast, the traditional Breadth-First-Search and Random Walk (without re-weighting) perform quite poorly, producing substantially biased results. In addition to offline performance assessment, we introduce online formal convergence diagnostics to assess sample quality during the data collection process. We show how these can be used to effectively determine when a random walk sample is of adequate size and quality for subsequent use (i.e., when it is safe to cease sampling). Using these methods, we collect the first to the best of our knowledge unbiased sample of Facebook. Finally, we use one of our representative datasets, collected through MHRW, to characterize several key properties of Facebook.**

***Index Terms*—Measurements, online social networks, Facebook, graph sampling, crawling, bias.**

I. Introduction

The popularity of online social networks (OSNs) in recent years is continuously increasing. Facebook (FB), in particular, is one of the most important online social networks (OSNs) today. It has the highest number of active users (at least 250M [1]) with more than half active FB users returning daily and the largest number of visitors among OSNs according to Comscore [2] (295M unique worldwide Internet users in March 2009). This success has generated interest within the networking community and has given rise to a number of measurement and characterization studies. Some studies are based on complete datasets of specific Facebook networks [3], [4]. However, the complete dataset is typically not available and, as most OSNs, Facebook is unwilling to share their company's data. Therefore, a relatively small but representative sample is desirable in order to study properties and test algorithms for these OSNs. A number of studies have already crawled social networks and Facebook, using mostly BFS-like techniques, which are known to introduce bias.

Minas Gjoka and Athina Markopoulou were partially supported by NSF CAREER grant 0747110. Maciej Kurant was visiting UC Irvine during the period that this work was conducted and was supported by grant ManCom 2110 of the Hasler Foundation, Bern, Switzerland. Carter T. Butts was supported by DOD ONR award N00014-08-1-1015.

Our primary goal in this paper is to explore the utility of various graph-crawling algorithms for producing a representative sample of Facebook users. We crawl Facebook's web front-end, which can be challenging in practice. A second goal of this paper is to introduce the use of formal convergence diagnostics to assess sample quality in an online fashion. These methods allow us to determine, in the absence of a ground truth, when a sample is adequate for subsequent use, and hence when it is safe to stop sampling, which is a critical issue in implementation. In the process of applying these methods to Facebook, we hope to illuminate more general characteristics of crawling methods that can be used to achieve asymptotically unbiased sampling of Facebook and other OSNs.

In terms of methodology, we consider several candidate crawling techniques. First, we consider Breadth-First-Search (BFS) - the heretofore most widely used technique for measurements of OSNs [5], [6] and FB [7]. BFS is well known to introduce bias towards high degree nodes; moreover, this bias is not formally characterized. Second, we consider Random Walk (RW) sampling, which also leads to bias towards high degree nodes, but at least its bias can be quantified by Markov Chain analysis and thus can be potentially corrected via re-weighting of the estimators (RWRW). Third, we consider the Metropolis-Hastings Random Walk (MHRW) that directly achieves the goal, *i.e.,* yields a uniform stationary distribution of nodes (users). This technique has been used in the past for P2P sampling [8], recently for a few OSNs [9], [10], but not for Facebook. Finally, we also collect a sample that represents the "ground truth" (UNI) *i.e.,* a truly uniform sample of Facebook userIDs, selected by a rejection sampling procedure from the system's 32-bit ID space. Such ground truth is in general unavailable, and our ability to use it as a basis of comparison is therefore a valuable asset of this study. We compare all sampling methods in terms of their bias and convergence properties. We also provide recommendations for their use in practice: *e.g.,* we implement online formal convergence diagnostic tests and parallel walks for improved speed; we also discuss pros and cons of MHRW vs. RWRW in practice.

In terms of results, we show that MHRW and RWRW work remarkably well in practice. We demonstrate their aggregate statistical properties, validating them against the known uniform sample, and show how our formal diagnostics can be used to identify convergence during the sampling process. In contrast, we find that the more traditional methods - BFS and RW - lead to significant bias in the case of FB. Finally, using

one of our validated samples (MHRW), we also characterize some key properties of Facebook; we find some of them to be substantively different from what was previously believed based on biased samples. The collected datasets are made publicly available for use by the research community at [11].

The structure of the paper is as follows. Section II discusses related work. Section III describes the sampling techniques and convergence diagnostics. Section IV summarizes the data collection process and the data sets. Section V evaluates and compares all sampling techniques in terms of convergence of various node properties and quality (lack of bias) of the obtained sample. Section VI provides a characterization of some key Facebook properties, based on the MHRW sample. Section VII concludes the paper.

II. Related Work

Crawling techniques can be roughly classified into two categories (a) graph traversal techniques and (b) random walks. In *graph traversal techniques*, each node in the connected component is visited exactly once, if we let the process run until completion. These methods vary in the order in which they visit the nodes; examples include Breadth-Search-First (BFS), Depth-First Search (DFS), Forest Fire (FF) and Snowball Sampling (SBS). BFS, in particular, is a basic technique that has been used extensively for sampling OSNs in past research [5]–[7]. One reason for this popularity is that an (even incomplete) BFS sample collects a full view (all nodes and edges) of some particular region in the graph, which is sometimes believed to be representative of the entire graph [7]. However, BFS leads to bias towards high degree nodes [12], [13]. Furthermore, this bias has not been analyzed so far for arbitrary graphs. In order to remove the bias, effort is usually put on completing the BFS, *i.e.,* on collecting all or most of the nodes in the graph.

Random walks allow node re-visiting and have well-known properties - see [14] for an excellent survey. They have been used for sampling the Web [15], P2P networks [8], [16], [17], and other large graphs [18]. The application of random walks to OSNs, such as Twitter [10] and Friendster [9], is very recent; to the best of our knowledge we are the first to apply these techniques to Facebook sampling [19]. Random walks can be biased but their bias can be analyzed using classic results from Markov Chains and corrected by re-weighting the estimators. This has been demonstrated in the context of P2P sampling [16], where the re-weighted random walk is considered as a special case of Respondent-Driven Sampling (RDS) [20] (if revisiting nodes is allowed and exactly one neighbor is selected in every step [21]). Alternatively, the random walk can be modified using the Metropolis filter so as to achieve, by design, *any* desired stationary distribution [22], [23]. In our case, this distribution is the uniform, because it has no sampling bias. This algorithm, known as Metropolis-Hasting Random Walk (MHRW) has been applied to P2P networks [8], modified to deal with peer churn (Metropolized Random Walk with Backtracking) and recently compared against re-weighted random walk (or RDS in the terminology of [9], [16]).

Compared to the aforementioned sampling techniques, our work is mostly related to the random walk techniques, as we obtain unbiased estimators using MHRW and RWRW; BFS and RW (without re-weighting) are used mainly as baselines for comparison. We accompany the basic crawling techniques with formal, *online convergence diagnostic tests* using several node properties, which, to the best of our knowledge, has not been done before in measurements of such systems. We also implement *multiple parallel chains*, which have also been recently used in [16] but started at the same node (while we start from different nodes, thus better utilizing the multiple chains). In terms of application, we perform unbiased sampling of *Facebook* for the first time. A unique asset of our study is a true uniform sample through sampling of userIDs, which can serve as *ground truth* to evaluate the crawling technique.

Other Measurements of Facebook. The work by Wilson et al. [7] measures social and user interaction graphs in Facebook between March and May 2008. Their sampling methodology is a region-constrained BFS. Such Region-Constrained BFS might be appropriate to study particular regions, but it does not provide Facebook-wide information, which is the goal of our study; furthermore, and unlike random walks, the bias of BFS has not been formally analyzed for arbitrary graphs. In [24] the authors examine the usage of privacy settings in Myspace and Facebook and the potential privacy leakage. In our previous work in [25], we characterized the popularity and user reach of Facebook applications. Finally, there are also two complete and publicly available datasets corresponding to two university networks from Facebook, namely Harvard [3] and Caltech [4]. In contrast, we sample the global Facebook social graph. To the best of our knowledge, compared to previous measurements this paper provides the first unbiased sample of Facebook.

III. Sampling Methodology

A. Scope and Assumptions

The FB social graph can be modeled as an undirected graph $G = (V, E)$, where V is a set of nodes (users) and E is a set of edges (mutual friendship relationships). Let k_v be the degree of node v. In this paper: (i) we are interested only in the publicly declared friends, which, under default privacy settings, are available to any logged-in user; (ii) we are not interested in isolated users, *i.e.,* users without any declared friends; (iii) we consider that the FB graph remains *static* during our crawling. We justify and discuss in detail assumption (iii) in Section IV.

B. Sampling Methods

The crawling of the social graph starts from an initial node and proceeds iteratively. In every operation, we visit a node and discover all its neighbors. There are many ways, depending on the particular sampling method, in which we can proceed. In this section, we describe the sampling methods we implemented in this paper. Our ultimate goal is to obtain a uniform random sample of users in Facebook.

1) Breadth First Search (BFS): At each new iteration the earliest explored but not-yet-visited node is selected next. As this method discovers all nodes within some distance from the starting point, an incomplete BFS is likely to densely cover only some specific region of the graph.

2) Random Walk (RW): In the classic random walk [14], the next-hop node w is chosen uniformly at random among the neighbors of the current node v. *I.e.,* the probability of moving from v to w is

$$P_{v,w}^{RW} = \begin{cases} \frac{1}{k_v} & \text{if } w \text{ is a neighbor of } v, \\ 0 & \text{otherwise.} \end{cases}$$

The random walk is [14] inherently biased. In a connected and aperiodic graph, the probability of being at the particular node v converges to the stationary distribution $\pi_v^{RW} = \frac{k_v}{2 \cdot |E|}$, *i.e.* the classic RW samples nodes w.p. $\pi_v^{RW} \sim k_v$. This is clearly biased towards high degree nodes; *e.g.,* a node with twice the degree will be visited by RW twice more often. In Section V, we show that several other node properties are correlated with the node degree and thus estimated with bias by RW sampling.

3) Re-Weighted Random Walk (RWRW): A natural next step is to crawl the network using RW, but correct for the bias of the estimator by re-weighting at the end. This can be done using the Hansen-Hurwitz estimator [26] as first shown in [21], [27] for random walks and also later used in [16]. Consider a stationary random walk that has visited $V = v_1, ...v_n$ unique nodes. Each node can belong to one of m groups with respect to a property of interest A, which might be the degree, network size or any other discrete-valued node property. Let $(A_1, A_2, .., A_m)$ be all possible values of A and corresponding groups; $\cup_1^m A_i = V$. *E.g.,* if the property of interest is the node degree, A_i contains all nodes u that have degree $k_u = i$. To estimate the probability distribution of A, we need to estimate the proportion of nodes with value $A_i,\ i = 1, ..m$:

$$\hat{p}(A_i) = \frac{\sum_{u \in A_i} 1/k_u}{\sum_{u \in V} 1/k_u}$$

Estimators for continuous properties can be obtained using related methods, e.g. kernel density estimators.

4) Metropolis-Hastings Random Walk (MHRW): Instead of correcting the bias after the walk, one can appropriately modify the transition probabilities so that it converges to the desired uniform distribution. The Metropolis-Hastings algorithm [22] is a general Markov Chain Monte Carlo (MCMC) technique [23] for sampling from a probability distribution μ that is difficult to sample from directly. In our case, we would like to sample nodes from the uniform distribution $\mu_v = \frac{1}{|V|}$. This can be achieved by the following transition probability:

$$P_{v,w}^{MH} = \begin{cases} \frac{1}{k_v} \cdot \min(1, \frac{k_v}{k_w}) & \text{if } w \text{ is a neighbor of } v, \\ 1 - \sum_{y \neq v} P_{v,y}^{MH} & \text{if } w = v, \\ 0 & \text{otherwise.} \end{cases}$$

It can be shown that the resulting stationary distribution is $\pi_v^{MH} = \frac{1}{|V|}$, which is exactly the uniform distribution we are looking for. $P_{v,w}^{MH}$ implies the following algorithm, which we refer to simply as MHRW in the rest of the paper:

$v \leftarrow$ initial node.
while stopping criterion not met **do**
  Select node w uniformly at random from neighbors of v.
  Generate uniformly at random a number $0 \leq p \leq 1$.
  if $p \leq \frac{k_v}{k_w}$ **then**
    $v \leftarrow w$.
  else
    Stay at v
  end if
end while

In every iteration of MHRW, at the current node v we randomly select a neighbor w and move there w.p. $\min(1, \frac{k_v}{k_w})$. We always accept the move towards a node of smaller degree, and reject some of the moves towards higher degree nodes. This eliminates the bias towards high degree nodes.

C. Convergence

1) Using Multiple Parallel Walks: Multiple parallel walks are used in the MCMC literature [23] to improve convergence. Intuitively, if we only have one walk, we might get trapped in a certain region of the graph and that may erroneously declare convergence. Having multiple parallel chains reduces the probability of this happening and allows for more accurate convergence diagnostics. An additional advantage of multiple parallel walks, is that it is amenable to parallel implementation from different machines or different threads in the same machine; in both cases, this reduces the duration of the crawl.

We implemented each of the considered crawling algorithms with several parallel MHRW walks. Each walk starts from a different node in $V_0 \subset V$, $|V_0| \geq 1$ ($|V_0| = 28$ in our case) and proceeds independently of the others. The initial nodes V_0 are chosen randomly. For a fair comparison, we compare multiple MHRWs to multiple RWs and multiple BFSs, all starting from the same set of nodes V_0.

2) Detecting Convergence with Online Diagnostics: Inferences from MCMC assume that the samples are derived from the equilibrium distribution, which is true asymptotically. To correctly diagnose when convergence occurs, we use online diagnostic tests developed within the MCMC literature [23], for the first time in the OSN sampling context.

One type of convergence has to do with losing dependence from the starting point. A standard approach is to run the sampling long enough and to discard a number of initial 'burn-in' iterations. This comes at a cost, which in the case of FB is the consumed bandwidth (in the order TB) and measurement time (days or weeks). It is therefore crucial to assess the convergence of our MCMC sampling, and to decide on appropriate settings of burn-in and total running time. The burn-in can be decided by using intra-chain and inter-chain diagnostics. In particular, we use two standard convergence tests, widely accepted and well documented in the MCMC literature, Geweke [28] and Gelman-Rubin [29], described below. We outline the rationale of these tests and we refer the interested reader to the references for more details. In Section V-A3, we apply these tests on several node properties, such as the node degree, privacy settings, network ID and membership.

Geweke Diagnostic. The Geweke diagnostic [28] detects the convergence of a single Markov chain. Let X be a single sequence of samples of our metric of interest. Geweke considers two subsequences of X, its beginning X_a (typically the first 10%), and its end X_b (typically the last 50%). Based on X_a and X_b, we compute the z-statistic: $z = \frac{E(X_a) - E(X_b)}{\sqrt{Var(X_a) + Var(X_b)}}$

With increasing number of iterations, X_a and X_b move further apart, which limits the correlation between them. As they measure the same metric, they should be identically distributed when converged and, according to the law of large

numbers, the z values become normally distributed with mean 0 and variance 1. We can declare convergence when most values fall in the $[-1, 1]$ interval.

Gelman-Rubin Diagnostic. Monitoring one long sequence has some disadvantages. *E.g.,* if our chain stays long enough in some non-representative region of the parameter space, we might erroneously declare convergence. For this reason, Gelman and Rubin [29] proposed to monitor $m > 1$ sequences. Intuitively speaking, the Gelman-Rubin diagnostic compares the empirical distributions of individual chains with the empirical distribution of all sequences together: if these two are similar, we declare convergence. The test outputs a single value R that is a function of means and variances of all chains. With time, R approaches 1, and convergence is declared typically for values smaller than 1.02.

We note that even after the burn-in period, strong correlation of consecutive samples in the chain may affect sequential analysis. This is typically addressed by thinning, *i.e.,* keeping only one every r samples. Instead of thinning, we do subsampling of nodes, which has essentially the same effect.

D. Ground Truth: Uniform Sample (UNI)

Assessing the quality of any sampling method on an unknown graph is a challenging task. In order to have a "ground truth" to compare against, the performance of such methods is typically tested on artificial graphs (using models such as Erdös-Rényi, Watts-Strogatz or Barabási-Albert, etc.). This has the disadvantage that one can never be sure that the results can be generalized to real networks that do not follow the simulated graph models and parameters.

Fortunately, Facebook was an exception at the time we performed our crawling. It allowed us to obtain a truly uniform sample of Facebook nodes by generating uniformly random 32-bit userIDs, and by polling Facebook about their existence. If the ID exists, we keep it, otherwise we discard it. This simple method is a textbook technique known as *rejection sampling* [30] and in general it allows to sample from any distribution of interest, which in our case is the uniform. In particular, it guarantees to select uniformly random userIDs from the existing FB users regardless of their actual distribution in the userID space, *i.e., even* if though the userIDs are *not* allocated sequentially or evenly across the userID space. For completeness, we derive this property of UNI sampling in the Appendix. We refer to this method as 'UNI', and use it as a ground-truth uniform sampler.

Although UNI sampling currently solves the problem of uniform node sampling in Facebook and is a valuable asset of this study, it is not a general solution for sampling OSNs. First, the ID space must not be sparse for this operation to be efficient.[1] Second, such an operation must be supported by the system, which is not the case in many OSNs. FB currently allows to verify the existence of an arbitrary userID and retrieve her list of friends. However, soon after we collected the UNI sample, FB moved from using numbers to using names as user IDs. In the near future, it is possible that FB may remove access to userIDs through the web-front interface.

[1]The number of Facebook users at the time of our study (2.0e8) was comparable to the size of the userID space (4.3e9), resulting in about one user retrieved per 22 attempts on average. If the userID was 64bits long [2] or consisting of strings of arbitrary length, UNI would be infeasible. *E.g.*, Orkut has a 64bit userID and hi5 uses a concatenation of userID+Name.

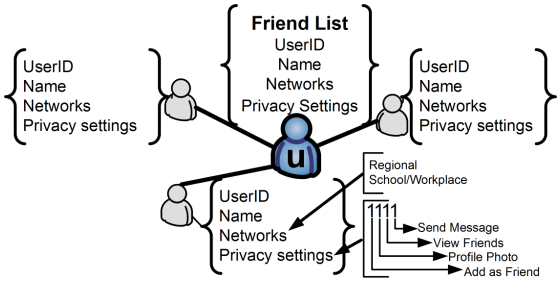

Fig. 1. **Basic node information** collected when visiting user u. **(1) Friends list:** this is a core feature of any OSN. In FB, friendship is always mutual thus leading to undirected edges. **(1) UserID and Name:** each user is uniquely defined by her userID, which is a 32-bit number, and provides her presumably real name. **(2) Networks.** Facebook groups its users into networks of two types: regional (geographical) and workplace/school. **(4) Privacy settings Q_u.** Each user u can restrict the amount of information or interaction with any non-friend node w. These are captured by four basic binary privacy attributes: 1 (Add as friend), 2 (Photo), 3 (View Friends), 4(Send message). We refer to the resulting 4-bit number as privacy settings Q_u of node u. By default, Facebook sets $Q_v = 1111$ (allow all).

In summary, we were fortunate to have obtained uniform sampling of userIDs and thus be able to evaluate the different sampling methods against "ground truth". However, crawling friendship relations is a fundamental primitive available in all OSNs and, we believe, the right building block for designing sampling techniques in OSNs in the long run.

IV. Data Collection

In this paper, we focus on open/publicly available basic information and do not study detailed user profiles that are more privacy-sensitive.

One node view. Fig. 1 shows the information collected when visiting the "show friends" webpage of a given user u, which we refer to as *basic node information*. We should emphasize here that when we visit user u, we collect network and privacy information for all her friends.

Invalid nodes. There are two types of nodes that we declare *invalid*. First, if a user u decides to hide her friends and to set the privacy settings to $Q_u = **0*$, the crawl cannot continue. We address this problem by backtracking to the previous node and continuing the crawl from there, as if u was never selected. Second, there exist nodes with degree $k_u = 0$; these are not reachable by any crawls, but we stumble upon them during the UNI sampling of the userID space. Discarding both types of nodes is consistent with our problem statement, where we declared that we exclude such nodes (either not publicly available or isolated) from the graph under study.

Implementation Details. In Section III-C1, we mentioned that we ran $|V_0| = 28$ independent crawls for each algorithm, namely MHRW, BFS and RW, all seeded at the same initial, randomly selected nodes V_0. The number of independent crawls comes from the number of different machines used. We let each independent crawl continue until exactly 81K samples are collected. In addition to the 28×3 crawls (BFS, RW and MHRW), we ran the UNI sampling until we collected 982K valid users, which is comparable to the 957K unique users collected with MHRW.

	MHRW	RW	BFS	UNI
# of valid users	28×81K	28×81K	28×81K	982K
# of *unique* users	957K	2.19M	2.20M	982K
# of *unique* neighbors	72.2M	120.1M	96.6M	58.3M
Crawling period	04/18-04/23	05/03-05/08	04/30-05/03	04/22-04/30
Avg Degree	95.2	338	323.9	94.1
Median Degree	40	234	208	38

TABLE I
COLLECTED DATASETS BY DIFFERENT ALGORITHMS. THE CRAWLING ALGORITHMS (MHRW, RW AND BFS) CONSIST OF 28 PARALLEL WALKS EACH, WITH THE SAME 28 RANDOMLY SELECTED STARTING POINTS. UNI IS THE UNIFORM SAMPLE OF USERIDS.

A crawler does HTML scraping to extract the basic node information (Fig. 1) of each visited node u. A server coordinates the crawls so as to avoid downloading duplicate information of previously visited users. This coordination brings many benefits: it takes advantage of the parallel chains to speed up the process, avoids overloading the FB platform with duplicate requests, and the crawling process continues in a faster pace since each request to FB servers returns new information.

Ego Networks. Elaborate topological measures, such as clustering coefficient and assortativity, cannot be estimated based purely on a single-node view. For this reason, after finishing the BFS, RW, MHRW crawls, we also collected a number of *ego nets* for a sub-sample of the MHRW dataset only (which is a representative one). The ego net is defined in the social networks literature [31], as follows: full information (edges and node properties) about a user and all its one-hop neighbors. This requires visiting 100 nodes per node (ego) on average, which is impossible to do for all visited nodes. For this reason, during 04/24-05/01 we collect the ego-nets of $\sim$ 37K nodes, randomly selected from all nodes in MHRW.

Data sets description. The datasets collected for this paper are summarized in Table I. This information refers to all sampled nodes, before discarding any "burn-in". The MHRW dataset contains 957K unique nodes, which is less than the $28 \times 81K = 2.26M$ iterations in all 28 random walks; this is because MHRW may repeat the same node in a walk. The number of rejected nodes in the MHRW process, without repetitions, adds up to 645K nodes.

For the UNI sampling, we checked 18.53M user IDs picked uniformly at random from $[0, 2^{32} - 1]$. Among them, only 1216K users existed, the rest were discarded. Also 228K valid userIDs had zero friends; we discarded these isolated users to be consistent with our problem statement. This results in a set of 985K valid users with at least one friend each. Considering that the percentage of zero degree nodes is unusually high, we manually confirmed that 200 of the discarded users have indeed zero friends.

Finally, we collected $\sim$ 37K egonets, a randomly chosen sub-sample of the $\sim$ 1M MHRW sample, which contain basic node information (see Fig 1) for 5.83M unique neighbors.

Overall, we crawled 11.6M unique nodes with basic node information. However, the total number of unique users for which we have basic privacy and network membership information (which includes the sampled nodes and their neighbors) is immense: we have such data for $\sim$ 172M unique Facebook users. This is a significant sample by itself given that Facebook had close to 200M active users at the time of the measurements.

Timescale of crawls. We treat the FB graph as static during the execution of our crawls, despite the fact that Facebook is growing. We believe that this assumption is a valid approximation in practice for several reasons. First, the FB characteristics change in longer timescales than the duration of our walks. During the period that we did our crawls, (April 18, 2009 - May 08, 2009, see table I), Facebook was growing at a rate of $450K/day$ as reported by websites such as [1], [32]. With a population of $\sim$200M users during that period, this translates to a growth of 0.22% of users/day. Each of our crawls lasted around 4-7 days (during which, the total FB growth was 0.9%-1.5%); in fact, our convergence analysis shows that the process converged even faster, *i.e.,* in only one day. Therefore, the growth of Facebook was negligible during our crawls. Second, the FB social (not interaction) graph is much more static than P2P systems that are known to have high churn; in the latter case, dealing with dynamic graphs becomes important [8], [33]. Third, we obtained empirical evidence by comparing our metrics of interest between the UNI sample of Table I and a similarly sized UNI sample obtained 45 days later. The distributions we obtained were virtually identical; we omit more details due to lack of space. Thus, while issues of dynamics are important to consider when sampling changing graphs, they appear not to be problematic for this particular study.

V. EVALUATION OF SAMPLING TECHNIQUES

In this section, we evaluate all candidate methodologies, namely BFS, RW and RWRW, MHRW, in terms of convergence and estimation bias. First, in Section V-A, we study in detail the convergence of the random walk methods, with respect to several properties of interest. We find a burn-in period of 6K samples, which we exclude from each independent crawl. The remaining 75K x 28 sampled nodes is our main sample dataset; for a fair comparison we also exclude the same number of burn-in samples from all datasets. Second, in Section V-B we examine the quality of the estimation based on each sample. Finally, in Section V-C, we summarize our our findings and provide recommendations for the use of sampling methods in practice.

A. Convergence Analysis

There are several crucial parameters that affect the convergence of MCMC, which apply to the random walk methods under study (but not to BFS).

1) How to count: Counting samples in BFS is trivial since nodes are visited at most once. However, in the random walks, nodes can be revisited and repetitions *must* be included in the sample in order to ensure the desired statistical properties. For RW the same node cannot be immediately visited twice, but non-consecutive repetitions are possible. In practice, that happens infrequently in the RW sample (as can be seen from the number of unique nodes given in table I). On the other hand, MHRW repeatedly samples some (typically low degree) nodes, a property which is intrinsic to its operation. For instance, if some node v_l has only one neighbor v_h, then the chain stays at (repeatedly samples) v_l for an average of k_{v_h} iterations (k_v is the degree of node v). Where k_{v_h} is large (e.g., $\mathcal{O}(10^2)$ or more), the number of repetitions may be locally large. While counterintuitive, this behavior is essential for convergence to the uniform distribution. In our MHRW

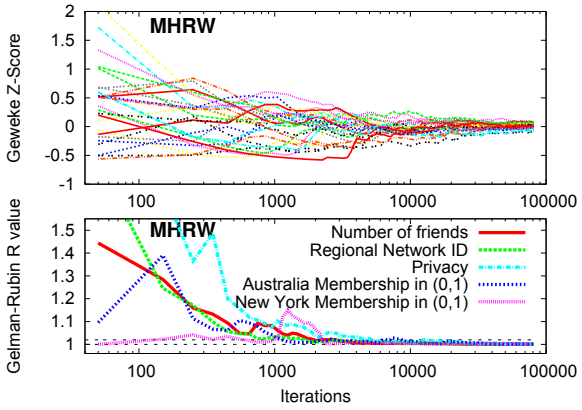

Fig. 2. **Convergence of the MHRW techniques.** (Top): Geweke z score for node degree. Each line shows the Geweke score for each of the 28 parallel chains. (Bottom) Gelman-Rubin score for four different metrics.

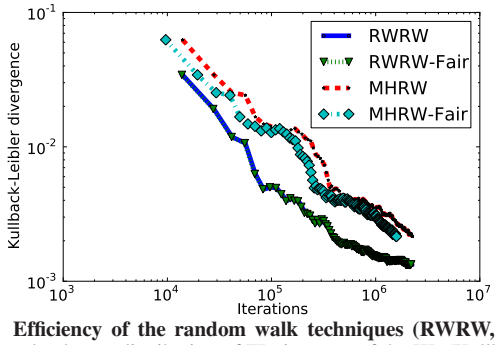

Fig. 3. **Efficiency of the random walk techniques (RWRW, MHRW)** in estimating the degree distribution of FB, in terms of the KL (Kullback-Leibler) divergence. We observe that (i) RWRW converges faster than MHRW and approximates UNI slightly better at the end (0.0021 for MHRW vs 0.0013 for RWRW) (ii) RWRW-Fair is also more efficient than MHRW-Fair. The "Fair" versions of the algorithms count the real bandwidth cost of contacting a previously unseen neighbor, either for sampling (in RW) or to learn its degree (in MHRW), based on our measurements.

sample, roughly 45% of the proposed moves are accepted (the *acceptance rate* in MCMC terms). As a result, a typical MHRW visits fewer unique nodes than a RW or BFS sequence of the same length. This raises the question: what is a fair way to compare the results of MHRW with RW and BFS? Since queries are only made for new nodes, if $k_{v_l} = 1$ and MHRW stays at v_l for some $\ell > 1$ iterations when crawling an OSN, the bandwidth consumed is equal in cost to one iteration (assuming that we cached the visited neighbor of v_l). This suggests that an appropriate practical comparison should be based not on the total number of iterations, but rather on the number of visited unique nodes. In our subsequent comparisons, we will denote RW and MHRW indices as "RW-Fair" and "MHRW-Fair" when we compare using the number of visited unique nodes, as this represents the methods in terms of equivalent bandwidth costs.

2) Convergence Tests: A decision we have to make is about the number of iterations for which we run the algorithms. This length should be appropriately long to ensure that we are at equilibrium (in the case of random walks).

The iterations taken before reaching (approximate) equilibrium are known as "burn-in" draws, and should be discarded to remove bias due to the choice of initial seed node. We ran the Geweke and Gelman-Rubin diagnostics on RW, RWRW and MHRW to determine the burn-in period. The Geweke diagnostic was run separately on each of the 28 chains for the metric of node degree. Fig. 2(top) presents the results for the convergence of the average node degree in the MHRW sample. We declare convergence when all 28 values fall in the $[-1, 1]$ interval, which happens at roughly iteration 500. In contrast, the Gelman-Rubin diagnostic analyzes all the 28 chains at once. In Fig 2 we plot the R score for four different metrics in the MHRW sample, namely (i) node degree (ii) regional network iii) privacy settings (iv) membership in specific regional networks. After 3000 iterations all the R scores drop below 1.02, the typical target value used for convergence indicator. We omit the plots for RW and RWRW since results look similar.

We declare convergence when all tests have detected it. The Gelman-Rubin test converges around 3K nodes. In each independent chain we conservatively discard 6K nodes, out of 81K total. In the remainder of the paper, we work only with the remaining 75K nodes per independent chain for RW, RWRW and MHRW.

In addition, we compared the random walk techniques in terms of their distance from the true uniform (UNI) distribution as a function of the iterations. In Fig.3, we show the distance of the estimated distribution from the ground truth in terms of the KL (Kullback-Leibler) metric that captures the distance of the 2 distributions accounting for the bulk of the distributions. Similar results hold for the Kolmogorov-Smirnov (KS) statistic that captures the maximum vertical distance of two distributions; we omit them due to lack of space. We should note here that the usage of distance metrics such as KL and KS cannot replace the role of the formal diagnostics which are able to determine convergence online and most importantly in the absence of the ground truth.

3) The choice of metric matters: MCMC is typically used to estimate some feature/metric, *i.e.,* a function of the underlying random variable. The choice of this metric can greatly affect the convergence time. The choice of metrics used in the online diagnostics in figure 2 was guided by the following principles. We chose the *node degree* because it is one of the metrics we want to estimate; therefore we need to ensure that the MCMC has converged at least with respect to it. The distribution of the node degree is also typically heavy tailed, and thus slow to converge. We also used several additional metrics (*e.g., network ID, privacy and network membership*), which are uncorrelated to the node degree and to each other, and thus provide additional assurance for convergence.

Let us focus on two of these metrics of interest, namely *node degree* and *sizes of geographical network* and study their convergence in more detail. The results for both metrics and all four methods are shown in Fig.4. We expected node degrees to not depend strongly on geography, while the relative size of geographical networks to strongly depend on geography. If our expectation is right, then (i) the degree distribution will converge fast to a good uniform sample even if the chain has poor mixing and stays in the same region for a long time; (ii) a chain that mixes poorly will take long time to barely reach the networks of interests, not to mention producing a reliable network size estimate. The results presented in the bottom part of Fig. 4 confirm our expectations. *E.g.* MHRW

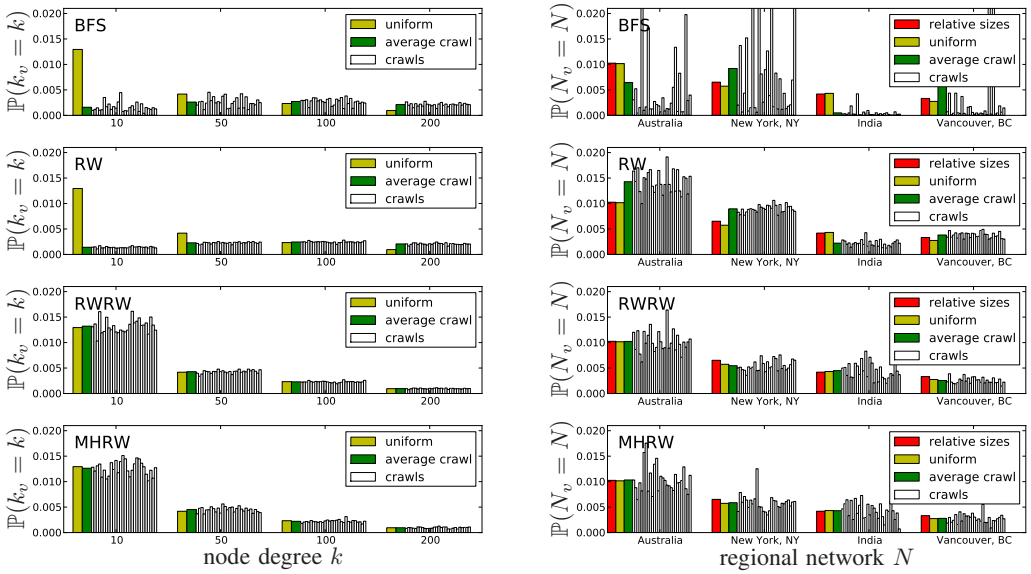

Fig. 4. Histograms of visits at node of a specific degree (left) and in a specific regional network (right). We consider four sampling techniques: BFS, RW, RWRW and MHRW. We present how often a specific type of node is visited by the 28 crawlers ('crawls'), and by the uniform UNI sampler ('uniform'). We also plot the visit frequency averaged over all the 28 crawlers ('average crawl'). Finally, 'size' represents the real size of each regional network normalized by the the total FB size. We used all the 81K nodes visited by each crawl, except the first 6k burn-in nodes. The metrics of interest cover roughly the same number of nodes (about 0.1% to 1%), which allows for a fair comparison.

performs much better when estimating the probability of a node having a given degree, than the probability of a node belonging to a specific regional network. One MHRW crawl overestimates the size of 'New York, NY' by roughly 100%. The probability that a perfect uniform sampling makes such an error (or larger) is $\sum_{i=i_0}^{\infty} \binom{i}{n} p^i (1-p)^i \simeq 4.3 \cdot 10^{-13}$, where $i_0 = 1k$, $n = 81K$ and $p = 0.006$. Even given such single-chain deviations, however, the multiple-chain average for the MHRW and RWRW crawls provides an excellent estimate of the true population size.

B. Unbiased Estimation

This section presents the main results of this paper. First, the MHRW and RWRW methods perform very well: they estimate two distributions of interest (namely node degree, regional network size) essentially identically to the UNI sampler. Second, the baseline algorithms (BFS and RW) deviate substantively from the truth and lead to misleading estimates.

1) Node degree distribution: In Fig. 5 we present the degree distributions estimated by BFS, RW, RWRW and MHRW. The average MHRW crawl's pdf, shown in Fig.5(a) is virtually identical to UNI. Moreover, the degree distribution found by each of the 28 chains separately are almost perfect. In contrast, RW and BFS shown in Fig.5(b) and (c) introduce a strong bias towards the high degree nodes. For example, the low-degree nodes are under-represented by two orders of magnitude. As a result, the estimated average node degree is $\overline{k}_v \simeq 95$ for MHRW and UNI, and $\overline{k}_v \simeq 330$ for BFS and RW. Interestingly, this bias is almost the same in the case of BFS and RW, but BFS is characterized by a much higher variance. Notice that that BFS and RW estimate wrong not only the parameters but also the shape of the degree distribution, thus leading to wrong information. Re-weighting the simple RW corrects for the bias results to RWRW, which performs almost identical to UNI, as shown in 5(b). As a side observation we can also see that the true degree distribution clearly *does not* follow a power-law.

2) Regional networks: Let us now consider a geography-dependent sensitive metric, *i.e.,* the relative size of regional networks. The results are presented in Fig. 4 (right). BFS performs very poorly, which is expected due to its local coverage. RW also produces biased results, possibly because of a slight positive correlation that we observed between network size and average node degree. In contrast, MHRW and RWRW perform very well.

C. Findings and Practical Recommendations

Choosing between methods. First and most important, the above comparison demonstrates that both MHRW and RWRW succeed in estimating several Facebook properties of interest virtually identically to UNI. In contrast, commonly used baseline methods (BFS and simple RW) fail, *i.e.,* deviate significantly from the truth and lead to substantively erroneous estimates. Moreover, the bias of BFS and RW shows up not only when estimating directly node degrees (which was expected), but also when we consider other metrics seemingly uncorrelated metrics (such as the size of regional network), which end up being correlated to node degree. This makes the case for moving from "1st generation" traversal methods such as BFS, which have been predominantly used in the measurements community so far [5]–[7], to more principled, "2nd generation", sampling techniques whose bias can be analyzed and/or corrected for. The random walks considered in this paper, RW, RWRW and MHRW, are well-known in the field of Monte Carlo Markov Chains (MCMC). We apply and adapt these methods to Facebook, for the first time, and we demonstrate that, when appropriately used, they perform remarkably well on real-world OSNs.

Adding convergence diagnostics and parallel crawls. A key ingredient of our implementation - not previously employed in network sampling - was the use of formal *online* convergence diagnostic tests. We tested these on several metrics

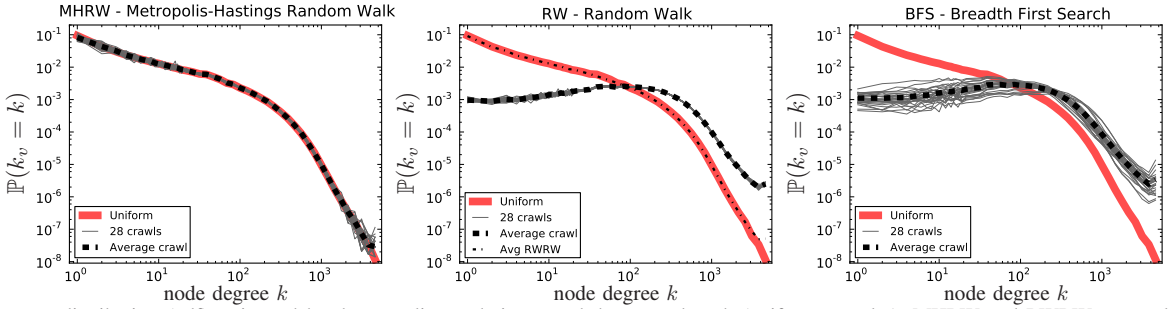

Fig. 5. Degree distribution (pdf) estimated by the sampling techniques and the ground truth (uniform sampler). MHRW and RWRW agree almost perfectly with the UNI sample; while BFS and RW deviate significantly.

of interest within and across chains, showing that convergence was obtained within a reasonable number of iterations. We believe that such tests can and should be used in field implementations of walk-based sampling methods to ensure that samples are adequate for subsequent analysis. Another key ingredient of our implementation, which we recommend, was the use of parallel crawlers/chains (started from several random independent starting points, unlike [9], [16] who use a single starting point), which both improved convergence and decreased the duration of the crawls.

MHRW vs. RWRW. Both MHRW and RWRW performed excellently in practice. When comparing the two, RWRW is slightly more efficient for the applications considered here, consistent with the findings in [9], [16]; this appears to be due to faster mixing in the latter Markov chain, which (unlike the former) does not require large numbers of rejections during the initial sampling process. However, when choosing between the two methods there are additional trade-offs to consider. First, MHRW yields an asymptotically uniform sample, which requires no additional processing for subsequent analysis. By contrast, RWRW samples are heavily biased towards high-degree nodes, and require use of appropriate re-weighting procedures to generate correct results. For the creation of large data sets intended for general distribution (as in the case of our Facebook sample), this "ready to use" aspect of MHRW has obvious merit; for example our released data sets are intended to be used by people that are not necessarily experts in the re-weighting methods, for whom the potential for erroneous misuse is high. A second advantage of MHRW is the ease of online testing for convergence to the desired target (uniform) distribution. In contrast, in RWRW, we test for convergence on a different distribution and then re-weight, which can introduce distortion. It is in principle possible to diagnose convergence on re-weighted statistics with RWRW. However, this requires appropriate use of re-weighting during the convergence evaluation process, which can increase the complexity of implementation. Finally, simple re-weighting is difficult or impossible to apply in the context of many purely data-analytic procedures such as multidimensional scaling or hierarchical clustering. Simulated Importance Resampling [34] provides a useful alternative for RWRW samples, but suffers from well-known problems of asymptotic bias (see [35] for a discussion and some palliatives). This is of less concern for applications such as moment estimation, for which re-weighting is both simple and effective.

Ultimately, the choice of RWRW versus MHRW is thus a trade-off between efficiency during the initial sampling process (which often favors RWRW) and simplicity/versatility of use for the resulting data set (which often favors MHRW). For our present purposes, these trade-offs favor MHRW, and we employ it here for producing a uniform ready-to-use sample of users. However, both approaches are viable alternatives in many settings, thus we present and analyze both in this paper.

VI. Facebook Characterization

In this section, we use the unbiased sample of 1M nodes, collected through MHRW, and the subsample of 37K egonets to study some features of Facebook. In contrast to previous work, which focused on particular regions [3], [4] or used larger but potentially biased samples [5], [7], our results are representative of the entire FB graph. Due to lack of space, we outline observations about topological characteristics only and refer the interested reader to our tech. report [19] for additional details as well as other features (e.g. privacy) omitted here.

Degree distribution. In Fig. 5, we present the node degree distribution of Facebook. Interestingly, and unlike previous studies of crawled datasets in online social networks in [5]–[7], we observe that node degree *is not* a power law. Instead, we can identify two regimes, roughly $1 \leq k < 300$ and $300 \leq k \leq 5000$, each roughly approximable by a power law with exponents $\alpha_{k<300} = 1.32$ and $\alpha_{k\geq 300} = 3.38$, respectively. We note, however, that the regime $300 \leq k \leq 5000$ covers only slightly more than one decade.

Assortativity. Depending on the type of complex network, nodes may tend to connect to similar or different nodes. For example, in many social networks high degree nodes tend to connect to other high degree nodes [36]. Such networks are called *assortative*. In contrast, biological and technological networks are typically *disassortative*, *i.e.,* they exhibit more high-degree than low-degree connections. In the plot of the node degree vs. the degrees of its neighbors (omitted due to lack of space), we observe a positive correlation, which implies assortative mixing and is in agreement with previous studies of similar social networks. We can also summarize this plot by calculating the Pearson correlation coefficient, or *assortativity coefficient* which is $r = 0.233$. This value is higher than $r' = 0.17$ reported in [7]. A possible explanation is that the Region-Constrained BFS used in [7] stops at regional network boundaries and thus misses many connections to, typically high-degree, nodes outside the network.

Clustering coefficient. In social networks, it is likely that two friends of a user are also friends to each other. The intensity of this phenomenon can be captured by the *clustering coefficient* C_v of a node v, defined as the relative number of

connections between the nearest neighbors of v. The clustering coefficient of a network is just an average value C over all nodes. We find the average clustering coefficient of Facebook to be $C = 0.16$, similar to that reported in [7]. Since the clustering coefficient tends to depend strongly on the node's degree k_v, we looked at C_v as a function of k_v (graph omitted due to lack of space) and we found a larger range in the degree-dependent clustering coefficient ([0.05, 0.35]) than what was found in [7] ([0.05, 0.18]).

VII. Conclusion

In this paper, we obtained for the first time representative (i.e., approximately uniform) samples of 1M Facebook users, using multiple methods. These samples were validated against a true uniform sample, as well as via formal convergence diagnostics, and shown to have good statistical properties; the datasets are accessible at [11]. To perform this task, we implemented and compared several crawling methods. We demonstrated that two principled approaches (MHRW and RWRW) perform remarkably well (almost identical to the ground truth) while the more traditional methods (BFW, RW) lead to substantial bias. We also give practical recommendations about the use of these methods for sampling OSNs in practice, including online convergence diagnostics and the proper use of multiple chains. Finally, using one of our representative samples, we were able to provide an accurate characterization of some key features of Facebook.

References

[1] "Fb statistics," http://facebook.com/press/info.php?statistics, July 2009.

[2] Comscore, http://meta.wikimedia.org/wiki/User:Stu/comScore_data_on_Wikimedia, 2009.

[3] K. Lewis, J. Kaufman, M. Gonzalez, A. Wimmer, and N. Christakis, "Tastes, ties, and time: A new social network dataset using Facebook.com," Social Networks, 2008.

[4] A. L. Traud, E. D. Kelsic, P. J. Mucha, and M. A. Porter, "Community structure in online collegiate social networks," 2008, arXiv:0809.0960.

[5] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and S. Bhattacharjee, "Measurement and Analysis of Online Social Networks," in *Proc. of IMC*, 2007.

[6] Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong, "Analysis of Topological Characteristics of Huge Online Social Networking Services," in *Proc. of WWW*, 2007.

[7] C. Wilson, B. Boe, A. Sala, K. P. Puttaswamy, and B. Y. Zhao, "User interactions in social networks and their implications," in *Proc. of EuroSys*, 2009.

[8] D. Stutzbach, R. Rejaie, N. Duffield, S. Sen, and W. Willinger, "On unbiased sampling for unstructured peer-to-peer networks," in *Proc. of IMC*, 2006.

[9] A. Rasti, M. Torkjazi, R. Rejaie, N. Duffield, W. Willinger, and D. Stutzbach, "Evaluating sampling techniques for large dynamic graphs," University of Oregon, Tech. Rep., Sept 2008.

[10] B. Krishnamurthy, P. Gill, and M. Arlitt, "A few chirps about twitter," in *Proc. of WOSN*, 2008.

[11] "Uniform sampling of facebook users: Publicly available datasets." http://odysseas.calit2.uci.edu/fb/, 2009.

[12] S. H. Lee, P.-J. Kim, and H. Jeong, "Statistical properties of sampled networks," *Phys. Rev. E*, vol. 73, p. 016102, 2006.

[13] L.Becchetti, C.Castillo, D.Donato, and A.Fazzone, "A comparison of sampling techniques for web graph characterization," in *LinkKDD*, 2006.

[14] L. Lovasz, "Random walks on graphs. a survey," in *Combinatorics*, 1993.

[15] M. R. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork, "On near-uniform url sampling," in *Proc. of WWW*, 2000.

[16] A. Rasti, M. Torkjazi, R. Rejaie, N. Duffield, W. Willinger, and D. Stutzbach, "Respondent-driven sampling for characterizing unstructured overlays," in *INFOCOM Mini-Conference*, April 2009.

[17] C. Gkantsidis, M. Mihail, and A. Saberi, "Random walks in peer-to-peer networks," in *Proc. of Infocom*, 2004.

[18] J. Leskovec and C. Faloutsos, "Sampling from large graphs," in *Proc. of ACM SIGKDD*, 2006.

[19] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, "A walk in facebook: Uniform sampling of users in online social networks," *http://arxiv.org/abs/0906.0060*, 2009.

[20] D. Heckathorn, "Respondent-driven sampling: A new approach to the study of hidden populations," *Social Problems*, vol. 44, p. 174199, 1997.

[21] M. Salganik and D. Heckathorn, "Sampling and estimation in hidden populations using respondent-driven sampling," *Sociological Methodology*, vol. 34, p. 193239, 2004.

[22] N. Metropolis, M. Rosenblut, A. Rosenbluth, A. Teller, and E. Teller, "Equation of state calculation by fast computing machines," *J. Chem. Physics*, vol. 21, pp. 1087–1092, 1953.

[23] W. Gilks, S. Richardson, and D. Spiegelhalter, *Markov Chain Monte Carlo in Practice*. Chapman and Hall/CRC, 1996.

[24] B. Krishnamurthy and C. E. Wills, "Characterizing privacy in online social networks," in *Proc. of WOSN*, 2008.

[25] M. Gjoka, M. Sirivianos, A. Markopoulou, and X. Yang, "Poking facebook: characterization of osn applications," in *Proc. of WOSN*, 2008.

[26] M.Hansen and W.Hurwitz, "On the theory of sampling from finite populations," *Annals of Mathematical Statistics*, vol. 14, 1943.

[27] E. Volz and D. D. Heckathorn, "Probability based estimation theory for respondent-driven sampling," *Journal of Official Statistics*, 2008.

[28] J. Geweke, "Evaluating the accuracy of sampling-based approaches to calculating posterior moments," in *Bayesian Statist. 4*, 1992.

[29] A. Gelman and D. Rubin, "Inference from iterative simulation using multiple sequences," in *Statist. Sci. Volume 7*, 1992.

[30] A. Leon-Garcia, *Probability, Statistics, and Random Processes For Electrical Engineering*. Prentice Hall, 2008.

[31] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[32] "Inside facebook," http://www.insidefacebook.com/2009/07/02/facebook-now-growing-by-over-700000-users-a-day-updated-engagement-stats/.

[33] W. Willinger, R. Rejaie, M. Torkjazi, M. Valafar, and M. Maggioni, "Osn research: Time to face the real challenges," in *HotMetrics*, 2009.

[34] D. B. Rubin, "Using the SIR algorithm to simulate posterior distributions," in *Bayesian Statistics 3*, 1988.

[35] O. Skare, "Improved sampling-importance resampling and reduced bias importance sampling." Scandin. journ. of stat., theory and apps, 2003.

[36] M. Newman, "Assortative mixing in networks," in *Phys. Rev. Lett. 89*.

Appendix: Correctness of UNI Sampling

Proposition: UNI (defined in Section III-D, as uniform sampling of 32-bit IDs and discarding the non existing ones) yields a uniform sample of the *existing (allocated)* user IDs in Facebook for *any allocation policy* (*e.g.,* even if the userIDs are not evenly allocated in the 32-bit address space).

Proof. Denote by U the set of all possible user IDs, *i.e.,* the set of all integers in $[0, 2^{32}-1]$. Let $A \subset U$ be the set of allocated user IDs in Facebook. We would like to sample the elements in A uniformly, *i.e.,* with pdf $f_A(x) = \frac{1}{|A|}\sum_{y\in A}\delta(y)$, where $\delta(y)$ is the Dirac delta. The difficulty is that we do not know the allocated IDs A beforehand. However, we are able to verify whether id x exists ($x \in A$) or not, for any x.

To achieve this goal, we apply rejection sampling [30] as follows. Choose uniformly at random an element from U (which is easy), *i.e.,* with pdf $f_U(x) = \frac{1}{|U|}\sum_{y\in U}\delta(y)$. Let $K = \frac{|U|}{|A|}$ s.t. $f_A(x) \le K \cdot f_U(x)$ for any x. Now, draw x from $f_U(x)$ and accept it with probability $\frac{f_A(x)}{K\cdot f_U(x)} = 1_{x\in A}$, *i.e.*, always if $x \in A$ (ID x exists/is allocated) and never if $x \notin A$ (ID x is not allocated). The resulting sample follows the distribution $f_A(x)$, *i.e.,* is taken uniformly at random from A (the set of *allocated* user IDs). □

The above is just a special case of *rejection sampling* [30], when the distribution of interest is uniform. It is presented here for completeness, given the importance of UNI sampling as "ground truth" in the paper.