Unbiased Sampling of Facebook

Minas Gjoka
Networked Systems
UC Irvine
mgjoka@uci.edu

Maciej Kurant
School of ICS
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

The popularity of online social networks (OSNs) has given rise to a number of measurements studies that provide a first step towards their understanding. So far, such studies have been based either on complete data sets provided directly by the OSN itself or on Breadth-First-Search (BFS) crawling of the social graph, which does not guarantee good statistical properties of the collected sample. In this paper, we crawl the publicly available social graph and present the first unbiased sampling of Facebook (FB) users using a Metropolis-Hastings random walk with multiple chains. We study the convergence properties of the walk and demonstrate the uniformity of the collected sample with respect to multiple metrics of interest. We provide a comparison of our crawling technique to baseline algorithms, namely BFS and simple random walk, as well as to the "ground truth" obtained through truly uniform sampling of userIDs. Our contributions lie both in the measurement methodology and in the collected sample. With regards to the methodology, our measurement technique (i) applies and combines known results from random walk sampling specifically in the OSN context and (ii) addresses system implementation aspects that have made the measurement of Facebook challenging so far. With respect to the collected sample: (i) it is the first representative sample of FB users and we plan to make it publicly available; (ii) we perform a characterization of several key properties of the data set, and find that some of them are substantially different from what was previously believed based on non-representative OSN samples.

1. INTRODUCTION

In recent years, the popularity of online social networks (OSNs) is continuously increasing: in May 2009, the total number of users in the top five OSNs combined (Myspace, Facebook, hi5, Friendster and Orkut) was 791M people. Facebook (FB) is one of the most important OSNs today. Indeed, it is the first OSN in terms of the number of active users (at least 200M [1]) and the first in terms of web visitors according to Comscore [4] (222M unique worldwide Internet users monthly), with more than half active FB users returning daily. It is also the fourth website on the Internet, according to Alexa's traffic rank in May 2009. In addition to its popularity, Facebook is also rich in functionality thanks to its open platform to third-party application developers. Clearly, OSNs in general and Facebook in particular have become an important phenomenon on the Internet, which is worth studying.

This success has generated interest within the networking community and has given rise to a number of measurements and characterization studies, which provide a first step towards the understanding of OSNs. Some of the studies are based on complete datasets provided by the OSN companies, such as Cyworld in [2]; or on complete datasets of specific networks within an OSN, typically university networks such as the Harvard [18] and Caltech [26] networks in Facebook. However, the complete dataset is typically not available to researchers, as most OSNs, including Facebook, are unwilling to share their company's data. In practice, a relatively small but representative sample may be a sufficient input for studies of OSN properties themselves or for algorithms that use OSN information to improve systems design. Therefore, it is important to develop techniques for obtaining small but representative OSN samples. A number of studies already exist that crawl social graphs, typically using BFS-type of graph traversal techniques, such as [2, 22, 29].

Our goal in this paper is to obtain a *representative sample of Facebook users by crawling the social graph*. We make the following assumptions in our problem statement: (i) we are interested only in the publicly declared lists of 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 also assume that the FB graph is static, which is valid if the FB characteristics change much slower than the duration of our crawl. To collect our sample, we crawl the Facebook's web front-end, which can be challenging in practice.[1] Beyond the imple-

[1]Measuring the entire Facebook is not a trivial task. Facebook has more than 200M users, each encoded by a 4B=32 bits long userID. A FB user has on average 100 friends which requires fetching on average an HTML page of 220KBytes to retrieve her friend list. Therefore, the raw topological data alone, without any node attributes, amounts to $200M \times 100 \times 32bit \simeq 80GB$. More importantly, the crawling overhead is tremendous: in order to collect 80GB, one would have to download about $200M \times 220KB = 44TB$ of HTML data.

mentation details, and more importantly, we are interested in designing the crawling in such a way that we collect a uniform sample of Facebook users, which is therefore representative of all FB users and appropriate for further statistical analysis.

In terms of methodology, we use multiple independent Metropolis-Hastings random walks (MHRW) and we perform formal convergence diagnostics. Our approach combines and applies known techniques from the Markov Chain Monte Carlo (MCMC) literature [7], for the first time, in the Facebook context. Parts of these techniques have been used recently in our community, although with some methodological differences (*i.e.,* without the multiple chains or the formal convergence diagnostics) and in different context (for P2P networks [27] and Twitter [12], but not for Facebook); for a detailed comparison please see Section 2. We compare our sampling methodology to popular alternatives, namely Breadth-First-Search (BFS) and simple random walk (RW), and we show that their results are substantially biased compared to ours. We also compare our sampling technique to the "ground truth", *i.e.,* a truly uniform sample of Facebook userIDs, randomly selected from the 32-bit ID space; we find that our results agree perfectly with the ground truth, which confirms the validity of our approach. We note, however, that such ground truth is in general unavailable or inefficient to obtain, as discussed in Section 3.3; in contrast, crawling friendship relations is a fundamental primitive available in OSNs and, we believe, the right building block for designing sampling techniques for OSNs. Therefore, we believe that out proposed approach is applicable to any OSN.

In terms of results, we obtain the first provably representative sample of Facebook users and we thoroughly demonstrate its good statistical properties. We plan to properly anonymize and make it publicly available. We also characterize some key properties of our sample, namely the degree distribution, assortativity, clustering and privacy features. We find that some of these properties are substantially different from what was previously believed based on biased sampling methods, such as BFS, even with an order of magnitude more samples than our technique. *E.g.,* we demonstrate that the degree distribution is clearly *not* a power-law distribution.

The structure of the paper is as follows. Section 2 discusses related work. Section 3 describes our sampling methodology, convergence diagnostics, and the alternative algorithms used as baselines for comparison. Section 4 describes the data collection process and summarizes the data set. Section 5 evaluates our methodology in terms of (i) convergence of various node properties and (ii) uniformity of the obtained sample as compared to alternative techniques as well as to the ground truth. Section 6 provides a characterization of some key Facebook properties, based on our representative sample, including topological properties of the social graph and user privacy features. Section 7 concludes the paper.

2. RELATED WORK

Broadly speaking, there are two types of work most closely related to this paper: (i) crawling *techniques*, focusing on the quality of the sampling technique itself and (ii) *characterization studies*, focusing on the properties of online social networks. These two categories are not necessarily disjoint.

First, in terms of sampling through crawling techniques, these can be roughly classified into BFS-based and random walks. Incomplete BFS sampling and its variants, such as snowball [28], are known to result in bias towards high degree nodes [16] in various artificial and real world topologies; we also confirmed this in the context of Facebook. Despite this well-known fact, BFS is still widely used for measuring OSNs, *e.g.,* in [22, 29] to mention a few examples; in order to remove the known bias, effort is usually put on completing the BFS, *i.e.,* on collecting all or most of the nodes in the graph. Interestingly, in our study we saw that the size of the sample does not in itself guarantee good properties.[2] It is also worth noting that BFS and its variants lead to samples that not only are biased but also do not have provable statistical properties.

Random walks may also lead to bias, but the stationary distribution is well-known and one could correct for it after the fact. Despite the possible bias, simple random walks have often been used in practice to achieve near-uniform sampling of P2P networks [10] and the web [11]. Gkantsidis et al. [10] simulate various P2P topologies and show that random walks outperform flooding (BFS) with regards to searching for two cases of practical interest. They also argue that random walks simulate uniform sampling well with a comparable number of samples. In [11], a random walk with jumps is used to achieve near-uniform sampling of URLs in the WWW. Their setting is different since the URL graph is directed and random jumps are needed to avoid entrapment in a region of the web. Leskovec et al. in [17] explore several sampling methods and compare them in terms of various graph metrics; their evaluations in static and dynamic graphs show that random walks perform the best.

The closest to our paper is the work by Stutzbach et al. in [27]: they use a Metropolized Random Walk with Backtracking (MRWB) to select a representative sample of peers in a P2P network and demonstrate its effectiveness through simulations over artificially generated graphs as well as with measurements of the Gnutella network. They also address the issue of sampling dynamic graphs, which is out of the scope here. Our work is different in two ways. In terms of methodology: (i) we use the *basic* Metropolis Random walk (ii) with *multiple* parallel chains and (iii) we extensively evaluate the *convergence* using several node properties and formal diagnostics. In terms of application, we ap-

[2] *E.g.,* We will see later that the union of all our datasets include ∼171M unique users, *i.e.,* a large portion of the Facebook population. Despite the large size, this aggregate dataset turns out to be biased and leads to wrong statistics. In contrast, our sample consists of ∼1M nodes but is representative.

ply our technique to *online social*, instead of peer-to-peer, networks, and we study characteristics specific to that context (*e.g.*, properties of egonets, the node degree, which we find not to follow a power-law, etc. We are also fortunate to be able to obtain a true uniform sample, which can serve as *ground truth* to validate our crawling technique. Finally in [12], Krishnamurthy et al. ran a single Metropolis Random Walk, inspired by [27], on Twitter as a way to verify the lack of bias in their main crawl used throughout the paper; the metropolis algorithm was not the main focus of their paper.

Second, in terms of studies that measure and characterize pure online social networks, other than Facebook, there have been several papers, including [2, 3, 21, 22]. Ahn et. al. in [2] analyze three online social networks; one complete social graph of Cyworld obtained from the Cyworld provider, and two small samples from Orkut and MySpace crawled with BFS. Interestingly, in our MHRW sample we observe a multi-scaling behavior in the degree distribution, similarly with the complete Cyworld dataset. In contrast, the crawled datasets from Orkut and MySpace in the same paper were reported to have simple scaling behavior. We believe that the discrepancy is due to the bias of the BFS-sampling they used. In [22] and [21] Mislove et al. studied the properties of the social graph in four popular OSNs: Flickr, LiveJournal, Orkut, and YouTube. Their approach was to collect the large Weakly Connected Component, also using BFS; their study concludes that OSNs are structurally different from other complex networks.

The work by Wilson et al. [29] is closely related to our study as it also studies Facebook. They collect and analyze social graphs and user interaction graphs in Facebook between March and May 2008. In terms of methodology, their approach differs from previous work in that they use what we call here a Region-Constrained BFS. They exhaustively collect all "open" user profiles and their list of friends in the 22 largest regional networks (out of the 507 available). First, such Region-Constrained BFS might be appropriate to study particular regions, but it does not provide any general Facebook-wide information, which is the goal of our study. Second, it seems that the percentage of users in the social graph retrieved in [29] is 30%-60% less than the maximum possible in each network.[3] In terms of results, the main conclusion in [29] is that the interaction graph should be preferred over social graphs in the analysis of online social networks, since it exhibits more pronounced small-world clustering. In our work, we collect a representative sample of the social graph. This sample can also allow us to fetch a representative sample of user profiles Facebook-wide in the future. In terms of findings, some noteworthy differences from [29] are that we find larger values of the degree-dependent clustering coefficient as well as a higher assortativity coefficient.

Other works that have measured properties of Facebook include [13] and [9]. In [13] the authors examine the usage of privacy settings in Myspace and Facebook, and the potential privacy leakage in OSNs. Compared to that work, we have only one common privacy attribute, "View friends", for which we observe similar results using our unbiased sample. But we also have additional privacy settings and a view of the social graph, which allows us to analyze user properties conditioned on their privacy awareness. In our previous work in [9], 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 [18] and Caltech [26]. In contrast, we collect a sample of the global Facebook social graph.

Finally, other recent works on OSNs include [14] by Kumar et al., which studied the structure and evolution of Flickr and Yahoo! 360, provided by their corresponding operators, and discovered a giant well-connected core in both of them. Liben-Nowell *et al.* [19] studied the LiveJournal online community and showed a strong relationship between friendship and geography in social networks. Girvan et al. [8] considered the property of community structure and proposed a method to detect such a property in OSNs.

3. SAMPLING METHODOLOGY

Facebook can be modeled as an undirected graph $G = (V, E)$, where V is a set of nodes (Facebook users) and E is a set of edges (Facebook friendship relationships). Let k_v be the degree of node v. We assume the following in our problem statement: (i) we are interested only in the publicly declared lists of 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 also assume that the FB graph is static, which is valid if the FB characteristics change much slower than the duration of our crawl (a few days).

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 first describe sampling methods commonly used in previous measurements of online social networks and are known to potentially introduce a significant bias to the results. Then we propose to use a technique that is provably asymptotically unbiased.

3.1 Previous sampling methods

3.1.1 Breadth First Search (BFS)

BFS is a classic graph traversal algorithm which starts

[3] More specifically, we believe that, for the collection of the social graph, their BFS crawler does not follow users that have their "view profile" privacy setting closed and "view friends" privacy setting open. We infer that by comparing the discrepancy in the percentage of users for those settings as reported in a Facebook privacy study conducted during the same time in [13] *i.e.*, in networks New York, London, Australia, Turkey.

from a seed node and progressively explores neighboring nodes. 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. BFS is known to be biased towards high degree nodes [15, 23] and no statistical properties can be proven for it. Nevertheless, BFS-based crawling and its variants, such as snowball, are widely used techniques for network measurements.

3.1.2 Random Walk (RW)

Another classic sampling technique is the classic random walk [20]. In this case, the next-hop node v is chosen uniformly at random among the neighbors of the current node u. Therefore, the probability of moving from u to v is

$$P^{RW}_{u,w} = \begin{cases} \frac{1}{k_u} & \text{if } w \text{ is a neighbor of } u, \\ 0 & \text{otherwise.} \end{cases}$$

Random walk has been deeply studied; *e.g.*, see [20] for an excellent survey. It is simple and there are analytical results on its stationary distribution and convergence time. Unfortunately, it is also inherently biased. Indeed, in a connected graph, the probability of being at the particular node u converges with time to:

$$\pi^{RW}_u = \frac{k_u}{2 \cdot |E|}$$

which is the stationary distribution of the random walk. *E.g.*, a node with twice the degree will be visited by RW two times more often. Moreover, we show later that many other node properties in OSNs are correlated with the node degree; these include, for example, the privacy settings, clustering coefficient, network membership, or even the 32 bit user ID. As a result of this correlation, all these metrics are inherently badly estimated by RW sampling.

3.2 Our sampling method

Our goal is to eliminate the biases of methods mentioned above and *obtain a uniformly distributed random sample of nodes* in Facebook. We can achieve a uniform stationary distribution by appropriately modifying the transition probabilities of the random walk, as follows.

3.2.1 Metropolis-Hastings Random Walk (MHRW)

The Metropolis-Hastings algorithm is a general Markov Chain Monte Carlo (MCMC) technique [7] for sampling from a probability distribution μ that is difficult to sample from directly. In our case, by performing the classic RW we can easily sample nodes from the non-uniform distribution π^{RW}, where $\pi^{RW}_u \sim k_u$. However, we would like to sample nodes from the uniform distribution μ, with $\mu_u = \frac{1}{|V|}$. This can be achieved by the following transition matrix:

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

It can be easily shown that the resulting stationary distribution of $P^{MH}_{u,w}$ is $\pi^{MH}_u = \frac{1}{|V|}$, which is exactly the uniform distribution we are looking for. The transition matrix $P^{MH}_{u,w}$ implies the following sampling procedure that we call Metropolis-Hastings Random Walk (MHRW):

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

In other words, in every iteration of MHRW, at the current node u we randomly select a neighbor w and move there with probability $\min(1, \frac{k_u}{k_w})$. We always accept the move towards a node of smaller degree, and reject some of the moves towards higher degree nodes. As a result, we eliminate the bias of RW towards high degree nodes.

3.2.2 Multiple Parallel Walks

Multiple parallel walks are used in the MCMC literature [7] to improve convergence. Intuitively, if we only have one walk, we might run into a scenario where it is trapped in a certain region while exploring the graph and that may lead to erroneous diagnosis of convergence. Having multiple parallel chains reduces the probability of this happening and allows for more accurate convergence diagnostics.[4] An additional advantage of multiple parallel walks, from an implementation point of view, is that it is amenable to parallel implementation from different machines or different threads in the same machine. Some coordination is then required to increase efficiency by not downloading information about nodes that have already been visited by independent walks.

Our proposed crawling technique consists of 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 randomly chosen in different networks. For a fair comparison, we compare our approach (multiple MHRWs) to multiple RWs and multiple BFSs, all starting from the same set of initial nodes V_0.

3.2.3 Convergence Tests

[4] We note that the advantage of multiple random walks is achieved when there is no fixed budget in the number of samples that would lead to many short chains; this is true in our case.

Valid inferences from MCMC are based on the assumption that the samples are derived from the equilibrium distribution, which is true asymptotically. In order to correctly diagnose when convergence occurs, we use standard diagnostic tests developed within the MCMC literature [7].

One type of convergence has to do with losing dependence from the starting point. A standard approach to achieve this is to run the sampling long enough and to discard a number of initial 'burn-in' iterations. From a practical point of view, the "burnt-in" comes at a cost. In the case of Facebook, it is the consumed bandwidth (in the order of terabytes) 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. From a theoretical point of view, 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 [6] and Gelman-Rubin [5], described below. In Section 5, we apply these tests on several node properties, including the node degree, userID, network ID and membership; please see Section 5.1.4 for details. Below, we briefly outline the rationale of these tests and we refer the interested reader to the references for more details.

Geweke Diagnostic. The Geweke diagnostic [6] 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 [5] 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 they are similar enough, we can declare convergence. This is captured by a single value R that is a function of means and variances of all chains (taken separately and together). With time, R approaches 1, and convergence is declared typically for values smaller than 1.02.

Finally, 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. In our approach, instead of thinning, we do sub-sampling of nodes after burn-in, which has essentially the same effect.

3.3 Ground Truth: Uniform Sample (UNI)

Assessing the quality of any graph sampling method on an unknown graph, as it is the case when measuring real systems, 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 is an exception (for the moment): there is a unique opportunity 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, known as rejection sampling, guarantees to select uniformly random userIDS from the existing FB users regardless of their actual distribution in the userID space. 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, we believe that our methodology (and results) remain important. There are two necessary conditions for UNI to work. First, the ID space must not be sparse for this operation to be efficient. The number of Facebook (2.0e8) users today is 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 [5] 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. Second, such an operation must be supported by the system. Facebook currently allows to verify the existence of an arbitrary userID and retrieve her list of friends; however, FB may remove this option in the future, *e.g.*, for security reasons.

In summary, we were fortunate to be able to obtain the ground truth, through uniform sampling of userIDs. This allowed us to demonstrate that our results perfectly agree with it. 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.

4. DATA COLLECTION

4.1 Collecting user properties of interest

[5]That is probable in the future either for security reasons i.e. to hinder efforts of data collection; or to allocate more userID space. See part 5.2.3 for current userID space usage

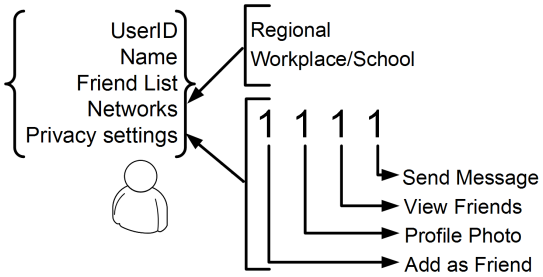

Figure 1: Information that we obtain about a user.

bit	attribute	explanation
1	Add as friend	=1 if w can propose to 'friend' u
2	Photo	=1 if w can see the profile photo of u
3	View friends	=1 if w can see the friends of u
4	Send message	=1 if w can send a message to u

Table 1: Basic privacy settings of a user u with respect to her non-friend w.

Fig. 1 summarizes the information that we obtain about each user that we visit during our crawls.

Name and userID. Each user is uniquely defined by its userID, which is a 32-bit number. Each user presumably provides her real name. The names do not have to be unique.

Friends list. A core idea in social networks is the possibility to declare friendship between users. In Facebook, friendship is always mutual and must be accepted by both sides. Thus the social network is undirected.

Networks. Facebook uses the notion of networks to organize its users. There are two types of networks. The first type is *regional* (geographical) networks. There are 507 predefined regional networks that correspond to cities and countries around the world. A user can freely join any regional network but can be a member of only one regional network at a time. Changes are allowed, but no more than two every 6 months. Roughly 62% of users belong to no regional network. The second type of networks indicates workplaces or schools and has a stricter membership: it requires a valid email account from the corresponding domain. On the other hand, a user can belong to many such networks.

Privacy settings Q_v. Each user u can restrict the amount of information revealed to any non-friend node w, as well as the possibility of interaction with w. These are captured by four basic binary privacy attributes, as described in Table 1. We refer to the resulting 4-bit number as privacy settings Q_v of node v. By default, Facebook sets $Q_v = 1111$ (allow all).

Profiles. Much more information about a user can potentially be obtained by viewing her profile. Unless restricted by the user, the profile can be displayed by her friends and users from the same network.In this paper, we do not collect any profile, even if it is open/publicly available. We study only the basic information mentioned above.

4.2 Collection Process

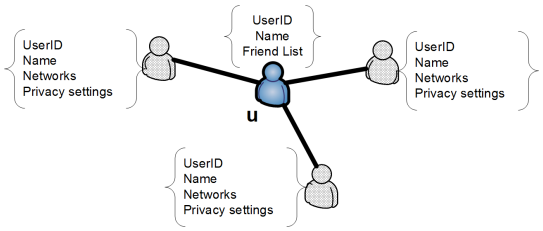

Figure 2: Basic node information collected when visiting a given user.

Crawling FB to collect this information faces several challenges, which we describe below, along with our solutions.

One node view. Fig. 2 shows the information collected when visiting the "show friends" webpage of a given user u, which we refer to as *basic node information*. Because the Network and Privacy information of u are not directly visible, we collect it indirectly by visiting one of u's friends and using the "show friends" feature.

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 $k_v = 0$; 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 already declared that we exclude such nodes (either not publicly available or isolated) from the graph we want to sample.

Implementation Details about the Crawls. In Section 3.2.2, we discussed the advantages of using multiple parallel chains both in terms of convergence and implementation. We ran $|V_0| = 28$ different independent crawls for each algorithm, namely MHRW, BFS and RW, all seeded at the same initial, randomly selected nodes V_0. We let each independent crawl continue until exactly 81K samples are collected.[6] 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.

In terms of implementation, we developed a multi-threaded crawler in Python and used a cluster of 56 machines. A crawler does HTML scraping to extract the basic node information (Fig. 2) of each visited node. We also have a server that coordinates the crawls so as to avoid downloading duplicate information of previously visited users. This coordination brings many benefits: we take advantage of the parallel chains in the sampling methodology to speed up the process, we do not overload the FB platform with duplicate

[6]We count towards this value all repetitions, such as the self-transitions of MHRW, and returning to an already visited state (RW and MHRW). As a result, the total number of unique nodes visited by each MHRW crawl is significantly smaller than 81K.

	MHRW	RW	BFS	Uniform
Total number of valid users	28×81K	28×81K	28×81K	982K
Total number of *unique* users	957K	2.19M	2.20M	982K
Total number 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

	Num of overlap. users
MHRW ∩ RW	16.2K
MHRW ∩ BFS	15.1K
MHRW ∩ Uniform	4.1K
RW ∩ BFS	64.2K
RW ∩ Uniform	9.3K
BFS ∩ Uniform	15.1K

Table 2: (Left:) Collected datasets by different algorithms. The crawling algorithms (MHRW, RW and BFS) consist of 28 parallel walks each, with the same 28 starting points. UNI is the uniform sample of userIDs. (Right:) The overlap between different datasets is small.

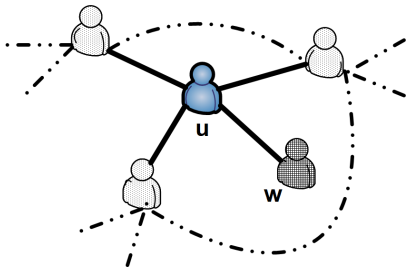

Figure 3: The ego network of a user u. (Invalid neighbor w, whose privacy settings $Q_w = **0*$, do not allow friend listing is discarded.)

Number of egonets	55K
Number of neighbors	9.28M
Number of unique neighbors	5.83M
Crawling period	04/24-05/01
Avg Clustering coefficient	0.16
Avg Assortativity	0.233

Table 3: Ego networks collected for 55K nodes, randomly selected from the users in the MHRW dataset.

requests, and the crawling process continues in a faster pace since each request to FB servers returns new information.

Ego Networks. The sample of nodes collected by our method enables us to study many features of FB users in a statistically unbiased manner. However, more 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 decided to also collect a number of *ego nets* for a sub-sample of the MHRW dataset only (because this is the only representative one). The ego net is defined in the social networks literature [28], and shown in Fig. 3, as 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, we collect the ego-nets only for 55K nodes, randomly selected from all nodes in MHRW (considering all 28 chains, after the 6000 'burn-in' period. This sub-sampling has the side advantage that it eliminates the correlation of consecutive nodes in the same crawl, as discussed in Section 3.2.3.

4.3 Data sets description

Information about the datasets we collected for this paper is summarized in Table 2 and Table 3. 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.[7] In the BFS crawl, we observe that the overlap of nodes between the 28 different BFS instances is very small: 97% of the nodes are unique, which also confirms that the random seeding chose different areas of Facebook. In the RW crawl, there is still repetition of nodes but is much smaller compared to the MHRW crawl, as expected. Again, unique nodes represent 97% of the RW dataset. Table 2 (right) shows that the common users between the MHRW, RW, BFS and Uniform datasets are a very small persentage, as expected. The largest observed, but still objectively small, overlap is between RW and BFS and is probably due to the common starting points selected.

During the Uniform userID sampling, we checked 18.53M user IDs picked uniformly at random from $[1, 2^{32}]$. Out of them, only 1216K users [8] existed. Among them, 228K users had zero friends; we discarded these isolated users to be consistent with our problems 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.

Also, we collected 55.5K egonets that contain basic node information (see Fig 2) for 5.83M unique neighbors. A summary of the egonets dataset, which includes properties that we analyze in Section 6, is summarized in Table.3.

Finally, as a result of (i) the multiple crawlings, namely BFS, random Walks, Metropolis random walks, uniform,

[7]Note that in order to obtain an unbiased sample, we also discard 6K burnt-in nodes from each of the 28 MHRW independent walks.

[8]In the set of 1216K existing users retrieved using uniform userID sampling, we find a peculiar subset that contains 37K users. To be exact, all users with $userID > 1.78 \cdot 10^9$ have zero friends and the name field is empty in the list of friends HTML page. This might be an indication that these accounts do not correspond to real people. Part 5.2.3 contains more information about the overall observed userID space.

neighbors of uniform users and (ii) the ego networks of a sub-sample of the Metropolis walk, we are able to collect 11.6 million unique nodes with basic node information. As a result, the total number of unique users (including the sampled nodes and the neighbors in their egonets) for which we have basic privacy and network membership information becomes immense. In particular, we have such data for 171.82 million[9] unique Facebook users. This is a significant sample by itself given that Facebook is reported to have close to 200million users as of this moment. Interestingly, during our analysis we have seen that this set of 171.82M (of sampled + egonet) nodes is a large but not representative set of FB. In contrast, the MHRW sample is much smaller (less than 1M) but representative, which makes the case for the value of unbiased sampling vs. exhaustive measurements.

5. EVALUATION OF OUR METHODOLOGY

In this section, we evaluate our methodology (multiple MHRW) both in terms of convergence and in terms of the representativeness of the sample. First, in Section 5.1, we study in detail the convergence of the proposed algorithm, with respect to several properties of interest. We find a burn-in period of 6K samples, which we exclude from each independent MHRW crawl. The remaining 75K x 28 sampled nodes from the MHRW method is our sample dataset. Section 5.2 presents essentially the main result of this paper. It demonstrates that the sample collected by our method is indeed uniform: it estimates several properties of interest perfectly, *i.e.* identically to those estimated by the true UNI sample. In contrast, the baseline methods (BFS and RW) deviate significantly from the truth and lead to substantively erroneous estimates.

5.1 MHRW convergence analysis

5.1.1 Typical MHRW evolution

To get more understanding of MHRW, let us first have a look at the typical chain evolution. At every iteration MHRW may either remain at the current user, or move to one of its neighbors. An example outcome from a simulation is: ...1, 1, 1, 1, 17, 1, 3, 3, 3, 1, 1, 1, 1, 2, 1, 1, 1, 2, 3, 9, 1..., where each number represents the number of consecutive rounds the chain remained at a given node. We note that a corresponding outcome for RW would consist only of ones. In our runs, roughly 45% of the proposed moves are accepted, which is also denoted in the literature as the acceptance rate. Note that MHRW stays at some nodes for relatively long time (e.g., 17 iterations in the example above). This happens usually at some low degree node v_l, and can be easily explained by the nature of MHRW. For example, in the extreme case, if v_l has only one neighbor v_h, then the chain stays at v_l on average for k_{v_h} iterations (k_v is a degree of node v), which often reaches hundreds. This behavior is required to make the walk converge to the uniform distribution.

[9]Interestingly, $\sim$ 800 out of 171.82M users had $userID > 32bit$ (or $5 \cdot 10^{-4}\%$), in the form of $1000000000xxxxx$ with only the last five digits used. We suspect that these userIDs are special assignments.

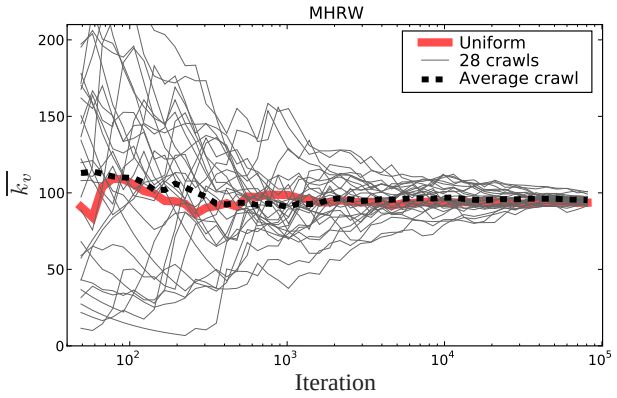

Figure 4: Average node degree $\overline{k_v}$ observed by the MHRW chains and by UNI, as a function of the number of iterations.

As a result, a typical MHRW visits fewer unique nodes than RW or BFS of the same length. This raises the question: what is a fair way to compare the results of MHRW with RW and BFS? Indeed, when crawling OSN, if $k_{v_l} = 1$ and MHRW stays at v_l for say 17 iterations, its bandwidth cost is equal to that of one iteration (assuming that we cache the visited neighbor of v_l). This suggests, that in our comparisons it might be fair to fix not the total number of iterations, but the number of visited unique nodes. However, we decided to follow the conservative iteration-based comparison approach, which favors the alternatives rather than our algorithm. This also simplifies the explanation.

5.1.2 Chain length and Thinning

One decision we have to make is about the number of iterations for which we run MHRW, or the *chain length*. This length should be appropriately long to ensure that we are at equilibrium. Consider the results presented in Fig. 4. In order to estimate the average node degree $\overline{k_v}$ based on a single MHRW only, we should take at least 10K iterations to be likely to get within $\pm 10\%$ off the real value. In contrast, averaging over all 28 chains seems to provide similar confidence after fewer than 1K iterations. Fig. 5 studies the frequency of visits at nodes with specific degrees, rather than the average over the entire distribution. Again, a chain length of 81K (top) results in much smaller estimation variance than taking 5K consecutive iterations (middle).

Another effect that is revealed in Fig.5 is the correlation between consecutive samples, even after equilibrium has been reached. It is sometimes reasonable to break this correlation, by considering every ith sample, a process which is called *thinning*, as discussed at the end of Section 3.2.3. The bottom plot in Fig. 5 is created by taking 5K iterations per chain with a thinning factor of $i = 10$. It performs much

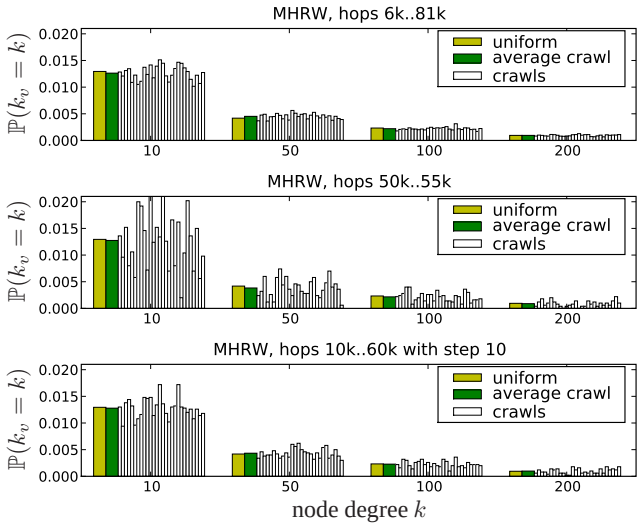

Figure 5: The effect of chain length and thinning on the results. We present histograms of visits at nodes with a specific degree $k \in \{10, 50, 100, 200\}$, generated under three conditions. (top): All nodes visited after the first 6k burn-in nodes. (middle): 5k consecutive nodes, from hop 50k to hop 55k. This represents a short chain length. (bottom): 5k nodes by taking every 10th sample (thinning).

better than the middle plot, despite the same total number of samples. In addition, thinning in MCMC samplings has the side advantage of saving space instead of storing all collected samples. However, in the case of crawling OSNs, the main bottleneck is the time and bandwidth necessary to perform a single hop, rather than storage and post-processing of the extracted information. Therefore we did not apply thinning to our basic crawls.

However, we applied another idea (sub-sampling) that had a similar effect with thinning, when collecting the second part of our data - the egonets. Indeed, in order to collect the information on a single egonet, our crawler had to visit the user and all its friends, an average ~ 100 nodes. Due to bandwidth and time constraints, we could fetch only 55K egonets. In order to avoid correlations between consecutive egonets, we collected a random sub-sample of the MHRW (post burn-in) sample, which essentially introduced spacing among sub-sampled nodes.

5.1.3 *Burn-in and Diagnostics*

As discussed on Section 3.2.3, the iterations before reaching equilibrium, known as "burn-in period" should be discarded. The Geweke and Gelman-Rubin diagnostics are designed to detect this burn-in period within each independent chain an across chains, respectively. Here we apply these diagnostics to several node properties of the nodes collected by our method and choose the maximum period from all tests.

The Geweke diagnostic was run separately on each of the 28 chains for the metric of node degree. Fig. 7 presents the

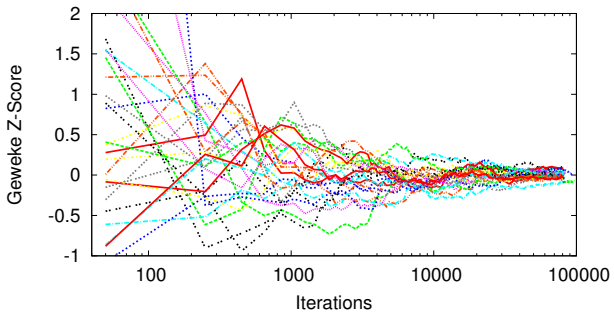

Figure 7: Geweke z score for node degree. We declare convergence when all values fall in the $[-1, 1]$ interval. Each line shows the Geweke score for a different MHRW chain, out of the 28 parallel ones. For metrics other than node degree, the plots look similar.

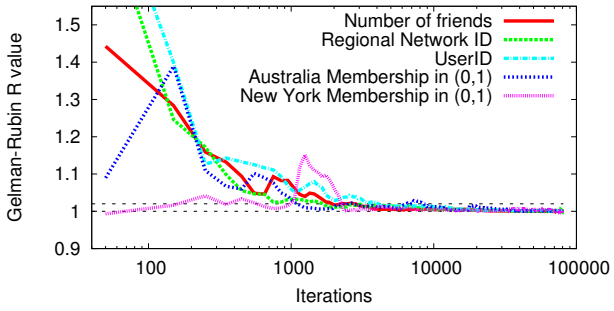

Figure 8: Gelman-Rubin R score for five different metrics. Values below 1.02 are typically used to declare convergence.

results for the convergence of average node degree. 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 8 we plot the R score for five different metrics, namely (i) node degree (ii) networkID (or regional network) (iii) user ID (iv) and (v) membership in specific regional networks (a binary variable indicating whether the user belongs to that network). After 3000 iterations all the R scores drop below 1.02, the typical target value used for convergence indicator.

We declare convergence when all tests have detected it. The Gelman-Rubin test is the last one at 3K nodes. To be even safer, in each independent chain we conservatively discard 6K nodes, out of 81K nodes total. For the remainder of the paper, we work only with the remaining 75K nodes per independent chain.

5.1.4 *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

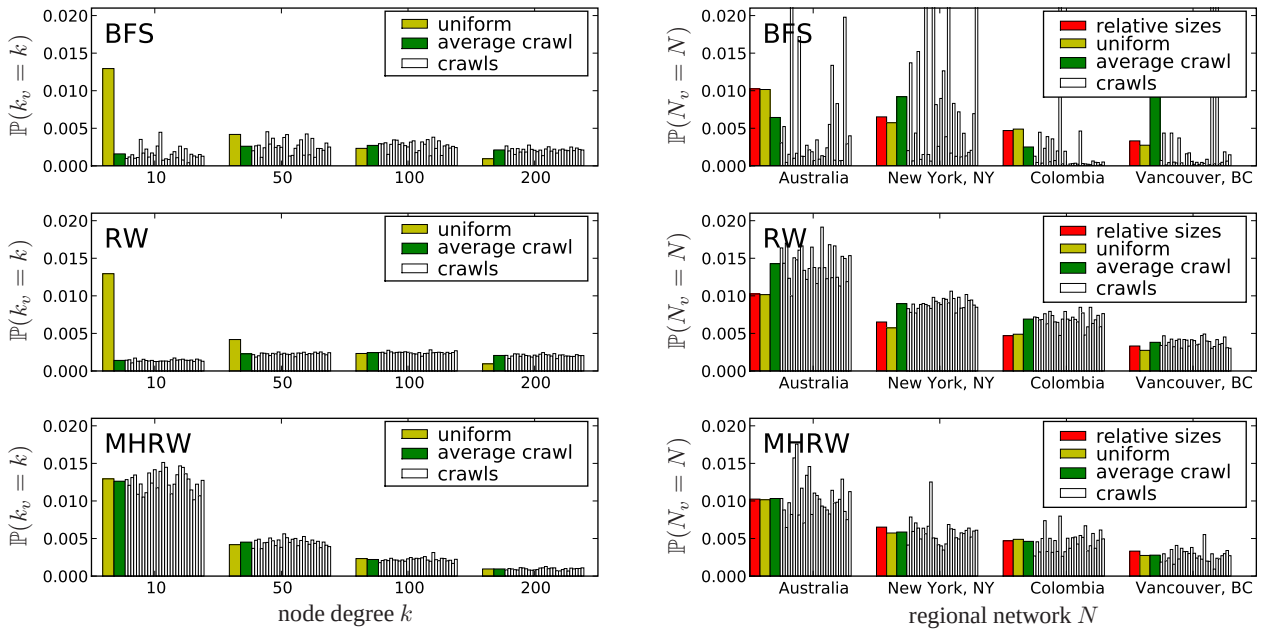

Figure 6: Histograms of visits at node of a specific degree (left) and in a specific regional network (right). We consider three sampling techniques: BFS (top), RW (middle) and MHRW (bottom). We present how often a specific type of nodes is visited by the 28 crawlers ('crawls'), and by the uniform UNI 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 facebook 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.

of this metric can greatly affect the convergence time. The choice of metrics used in the diagnostics of the previous section 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 not easy to converge.
- We also used several additional metrics (e.g. network ID, user ID and membership to specific networks), 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 three methods are shown in Fig.6. 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. In the latter case, MHRW will need a large number of iterations before collecting a representative sample.

The results presented in Fig. 6 (bottom) indeed confirm our expectations. MHRW 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. For example, 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 we took $i_0 = 1k$, $n = 81K$ and $p = 0.006$.

5.2 Comparison to other sampling techniques

This section presents essentially the main result of this paper. It demonstrates that our method collects a truly uniform sample. It estimates three distributions of interest, namely those of node degree, regional network size and userID, perfectly, *i.e.*, identically to the UNI sampler. In contrast, the baseline algorithms (BFS and RW) deviate substantively from the truth and lead to misleading estimates and behavior. This was expected for the degree distribution, which is known to be biased in the BFS and RW cases, but it is surprising in the case of the other two metrics.

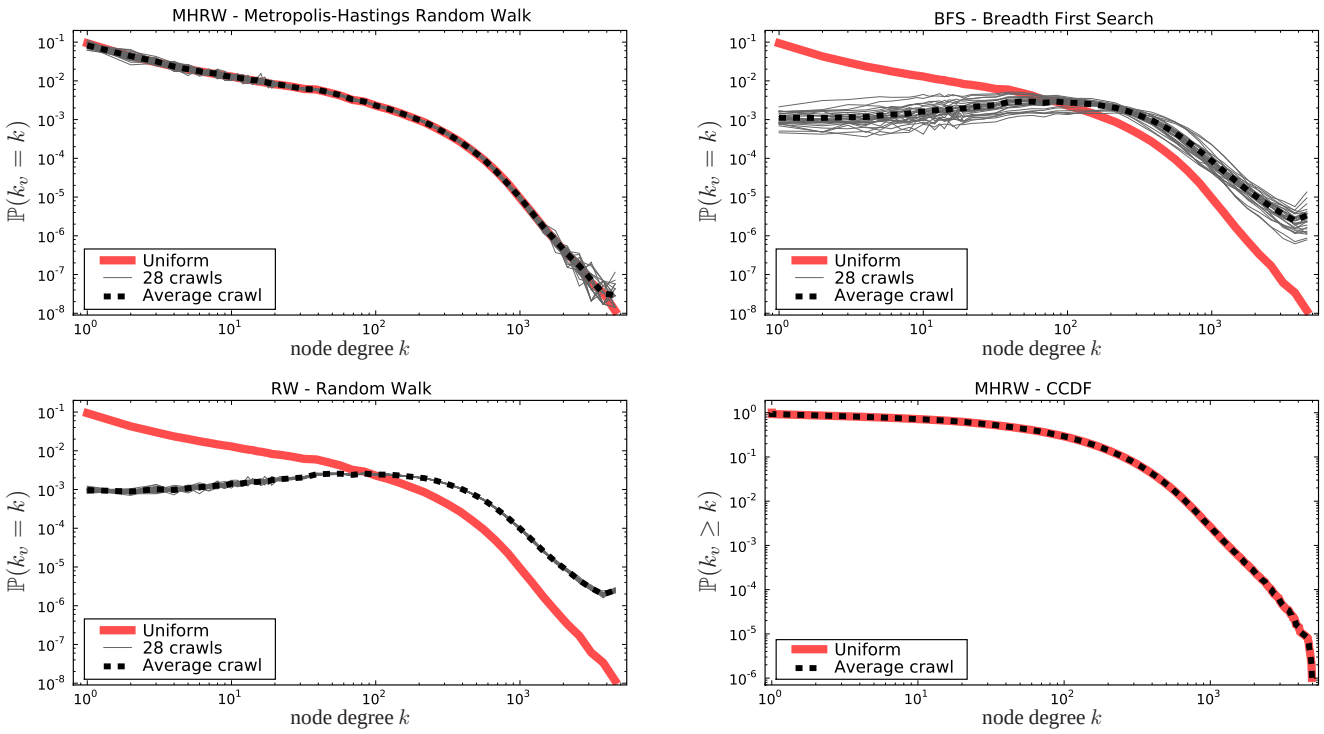

Figure 9: Degree distribution estimated by the crawls and the uniform sampler. All plots use log-log scale. For the first three (pdf) plots we used logarithmic binning of data; the last plot is a ccdf.

5.2.1 *Node degree distribution*

In Figure 9 we present the degree distributions based on the BFS, RW and MHRW samples. The average MHRW crawl's pdf and ccdf, shown in Fig.9(a) and (d) respectively, are virtually identical with UNI. Moreover, the degree distribution found by each of the 28 chains separately are almost perfect. In contrast, BFS and RW 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. These results are consistent with the distributions of specific degrees presented in Figure 6 (left).

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. As a side observation we can also see that the true degree distribution clearly *does not* follow a power-law.

5.2.2 *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. 6 (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 performs very well albeit with higher variance, as already discussed in Section 5.1.4.

5.2.3 *The userID space*

Finally, we look at the distribution of a property that is completely uncorrelated from the topology of FB, namely the user ID. When a new user joins Facebook, it is automatically assigned a 32-bit number, called userID. It happens before the user specifies its profile, joins networks or adds friends, and therefore one could expect no correlations between userID and these features. In other words, the degree bias of BFS and RW should not affect the usage of userID space. Therefore, at first sight we were very surprised to find big differences in the usage of userID space discovered by BFS, RW and MHRW. We present the results in Fig 10. BFS and RW are clearly shifted towards lower userIDs.

The origin of this shift is probably historical. The sharp steps at $2^{29} \simeq 0.5$e9 and at $2^{30} \simeq 1.0$e9 suggest that FB was first using only 29 bit of userIDs, then 30, and now 31. As a result, users that joined earlier have the smaller userIDs. At the same time, older users should have higher degrees on average. If our reasoning is correct, userIDs should be negatively correlated with node degrees. This is indeed the case, as we show in the inset of Fig 10. This, together with the degree bias of BFS and RW, explains the shifts of userIDs distributions observed in the main plot in Fig 10.

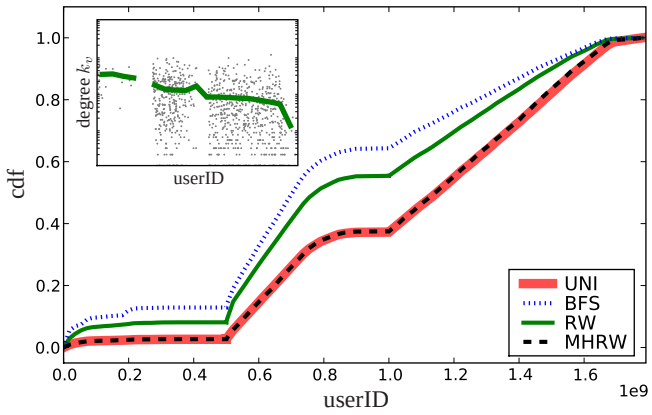

Figure 10: User ID space usage discovered by BFS, RW, MHRW and UNI. Each user is assigned a 32 bit long userID. Although this results in numbers up to $4.3e9$, the values above $1.8e9$ almost never occur. Inset: The average node degree (in log scale) as a function of userID.

Needless to mention, that in contrast to BFS and RW, our MHRW performed perfectly with respect to the userID metric.

5.3 Conclusion

We have demonstrated that MHRW converges and performs remarkably well, virtually undistinguishable from UNI. In contrast, the two alternative sampling techniques, RW and BFS, are strongly biased. Moreover, this bias shows up not only when estimating directly node degrees (which was expected), but also when we consider other metrics such as the size of regional network, or the seemingly independent userID. This is because these and many other metrics correlate, positively or negatively, with the node degree.

6. FACEBOOK CHARACTERIZATION

In the previous sections, we sampled Facebook and demonstrated convergence and a true uniform sample of about 1M Facebook nodes. In this section, we use this unbiased sample and the egonets dataset to study some topological and non-topological features of Facebook. In contrast to previous works, which focused on some particular regions [18,26] or used biased sampling methods [22,29], our results are representative of the entire Facebook graph.

6.1 Topological characteristics

We first focus on purely topological aspects of the graph of Facebook.

6.1.1 Degree distributions

In Fig. 9, we present the true node degree distributions of Facebook, the pdf (upper left) and the corresponding ccdf (lower right). Interestingly, and unlike previous studies of crawled datasets in online social networks in [2, 22, 29], we

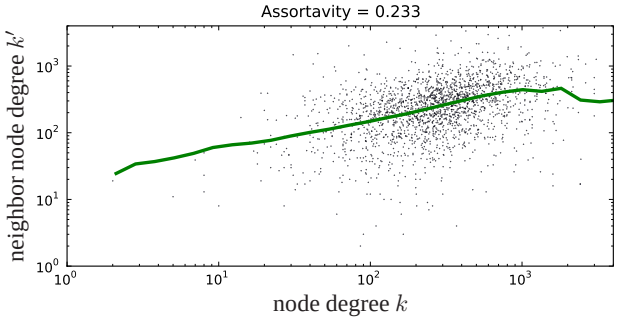

Figure 11: Assortativity - correlation between degrees of neighboring nodes. The dots represent the degree-degree pairs (randomly subsampled for visualization only). The line uses log-binning and takes the average degree of all nodes that fall in a corresponding bin.

conclude that the node degree distribution of Facebook *does not* follow a power law distribution. Instead, we can identify two regimes, roughly $1 \leq k < 300$ and $300 \leq k \leq 5000$, each following a power law with exponent $\alpha_{k<300} = 1.32$ and $\alpha_{k\geq 300} = 3.38$, respectively.[10] We note, however, that the regime $300 \leq k \leq 5000$ covers only slightly more than one decade.

6.1.2 Assortativity

Depending on the type of complex network, nodes may tend to connect to similar or different nodes. For example, in most social networks high degree nodes tend to connect to other high degree nodes [24]. Such networks are called *assortative*. In contrast, biological and technological networks are typically *disassortative*, *i.e.*, they exhibit significantly more high-degree than low-degree connections.

In Fig. 11 we plot the node degree vs. the degrees of its neighbors. We observe a positive correlation, which implies assortative mixing and is in agreement with previous studies of social networks. We can also summarize this plot by calculating the Pearson correlation coefficient, or *assortativity coefficient* r. The assortativity coefficient of Facebook is $r = 0.233$. This value is higher than $r' = 0.17$ reported by Wilson et al in [29]. A possible explanation of this difference is that [29] uses the Region-Constrained BFS to sample Facebook. It stops at a boundary of a regional network and thus misses many connections to, typically high-degree, nodes outside the network.

6.1.3 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 formally captured by the *clustering coefficient* C_v of a node v, defined as the relative number of connections between the nearest neighbors of v. In other words, $C_v = \frac{2m_v}{k_v(k_v-1)}$, where m_v is the total number of edges be-

[10] Exponents were computed with the help of formula (5) in [25].

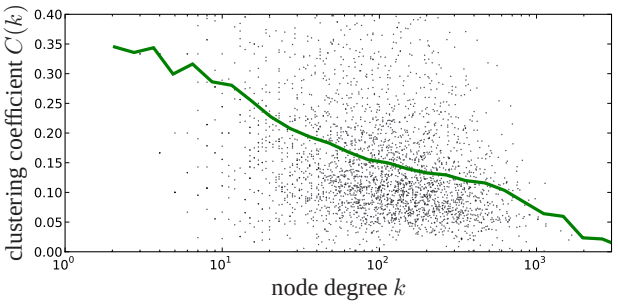

Figure 12: Clustering coefficient of Facebook users as function of their degree.

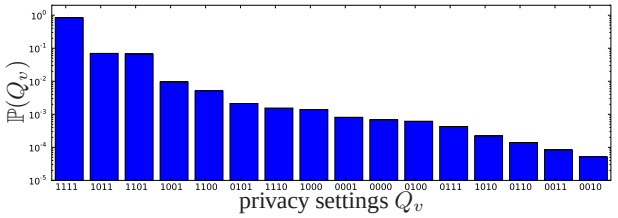

Figure 13: The distribution of the privacy settings among $\sim$ 171.8M Facebook users. Value $Q_v = 1111$ corresponds to default settings (privacy not restricted) and covers 84% of all users.

tween the nearest neighbors of v, and k_v is the degree of node v. The clustering coefficient of a network is just an average value $C = \frac{1}{n}\sum_v C_v$, where n is the number of nodes in the network. We find the average clustering coefficient of Facebook to be $C = 0.16$, similar to that reported in [29].

Since the clustering coefficient tends to depend strongly on the node's degree k_v, it makes sense to study its average value $C(k)$ conditioned on k_v. We plot C_v as a function of k_v in Fig. 12. Comparing with [29], we find a larger range in the degree-dependent clustering coefficient ([0.05, 0.35] instead of [0.05, 0.18]).

6.2 Privacy awareness

Recall from Section 4 that our crawls collected, among other properties, the privacy settings Q_v for each node v. Q_v consists of four bits, each corresponding to one privacy attribute. By default, Facebook sets these attributes to 'allow', i.e., $Q_v = 1111$ for a new node v. Users can freely change these default settings of Q_v. We refer to the users that choose to do so as 'privacy aware' users, and we denote by PA the level of privacy awareness of a user v, *i.e.*, privacy aware users have $PA = \mathbb{P}(Q_v \neq 1111)$.

In Fig. 13, we present the distribution of privacy settings among Facebook users. About 84% of users leave the settings unchanged, *i.e.*, $\mathbb{P}(Q_v{=}1111) \simeq 0.84$. The remaining 16% of users modified the default settings, yielding $PA = 0.16$ across the entire Facebook. The two most popular modifications are $Q_v = 1011$ ('hide my photo') and $Q_v = 1101$ ('hide my friends'), each applied by about 7% of users.

PA	Network n	PA	Network n
0.08	Iceland	...	...
0.11	Denmark	0.22	Bangladesh
0.11	Provo, UT	0.23	Hamilton, ON
0.11	Ogden, UT	0.23	Calgary, AB
0.11	Slovakia	0.23	Iran
0.11	Plymouth	0.23	India
0.11	Eastern Idaho, ID	0.23	Egypt
0.11	Indonesia	0.24	United Arab Emirates
0.11	Western Colorado, CO	0.24	Palestine
0.11	Quebec City, QC	0.25	Vancouver, BC
0.11	Salt Lake City, UT	0.26	Lebanon
0.12	Northern Colorado, CO	0.27	Turkey
0.12	Lancaster, PA	0.27	Toronto, ON
0.12	Boise, ID	0.28	Kuwait
0.12	Portsmouth	0.29	Jordan
...	...	0.30	Saudi Arabia

Table 4: Regional networks with respect to their privacy awareness $PA = \mathbb{P}(Q_v \neq 1111 \,|v \in n)$ among $\sim$ 171.8M Facebook users. Only regions with at least 50K users are considered.

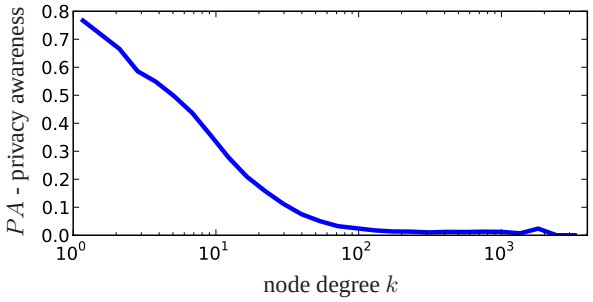

Figure 14: Privacy awareness as a function of node degree in the egonets dataset. We consider only the nodes with privacy settings set to '1*', because only these nodes allow us to see their friends and thus degree. So here $PA = \mathbb{P}(Q_v \neq 1111 \mid k_v = k,\ Q_v = **1*)$.**

The privacy awareness PA of Facebook users depends on many factors, such as the geographical location, node degree or the privacy awareness of friends. For example, in Table 4 we classify the regional networks with respect to PA of their members. Note the different types of countries in the two extreme ends of the spectrum. In particular, many FB users in the Middle East seem to be highly concerned about privacy. Interestingly, Canada regions show up at both ends, clearly splitting into English and French speaking parts.

Another factor that affects the privacy settings of a user is the node degree. We present the results in Fig. 14. Low degree nodes tend to be very concerned about privacy, whereas high degree nodes hardly ever bother to modify the default settings. This clear trend makes sense in practice. Indeed, to protect her privacy, a privacy concerned user would carefully select her Facebook friends, *e.g.*, by avoiding linking to strangers. At the other extreme, there are users who prefer to have as many 'friends' as possible, which is much easier

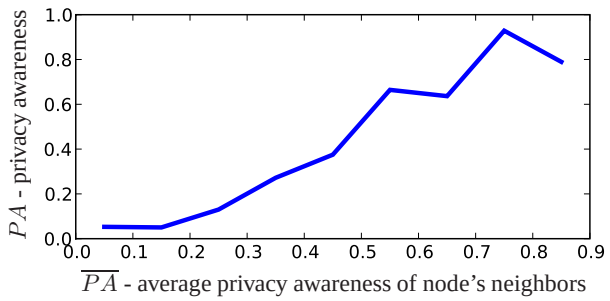

Figure 15: Privacy awareness as a function of privacy awareness of node's neighbors in the egonets dataset. We consider only the nodes with privacy settings set to '1*', because only these nodes allow us to see their friends, so $PA = \mathbb{P}(Q_v \neq 1111 \mid \overline{PA},\ Q_v = **1*)$.**

with unrestricted privacy attributes.

Finally, in Fig. 15 we show how privacy awareness of a user depends on the privacy awareness of her friends. We observe a clear positive correlation.

7. CONCLUSION

In this paper, first, we proposed a method for sampling Facebook in a principled way so as to obtain a uniform sample of Facebook users. Our approach consists of (i) running multiple chains in parallel, each of which performs a Multiple Hasting Random Walk and (ii) ensuring convergence using appropriate diagnostics run on several metrics of interest. We demonstrate that, for all practical purposes, our method achieves a perfectly random sample of 1M nodes (a small sample size), while traditional alternative techniques (BFS and RW) introduce significant bias on degree distribution and other metrics, even with a significantly number of samples. Second, and based on our unbiased sample and on a sub-sample of egonets, we also studied some interesting properties of Facebook. Some of our findings agree with previous studies, some disagree and reveal a substantive bias of prior sampling techniques, and some are new to the best of our knowledge. The sampling approach we described in this paper is principled, effective, and applicable to any OSN (as it is based on crawling the friendship relation which is a fundamental primitive in any OSN).

8. REFERENCES

[1] Facebook statistics. http://www.facebook.com/press/info.php?statistics, 2009.

[2] 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.

[3] H. Chun, H. Kwak, Y.-H. Eom, Y.-Y. Ahn, S. Moon, and H. Jeong. Comparison of online social relations in volume vs interaction: a case study of cyworld. In *Proc. of IMC*, 2008.

[4] Comscore. Comscore press release. http://ir.comscore.com/releasedetail.cfm?ReleaseID=361041, 2009.

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

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

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

[8] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. In *Proc. of the National Academy of Sciences*, 2002.

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

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

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

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

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

[14] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *Proc. of ACM SIGKDD*, 2006.

[15] S. H. Lee, P.-J. Kim, and H. Jeong. Statistical properties of sampled networks.

[16] S. H. Lee, P.-J. Kim, and H. Jeong. Statistical properties of sampled networks, 2006.

[17] J. Leskovec and C. Faloutsos. Sampling from large graphs. In *Proc. of ACM SIGKDD*, 2006.

[18] 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.

[19] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. In *Proc. of the National Academy of Sciences*, 2005.

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

[21] A. Mislove, H. S. Koppula, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Growth of the flickr social network. In *Proc. of WOSN*, 2008.

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

[23] M. Najork and J. L. Wiener. Breadth-first search crawling yields high-quality pages. In *Proc. of WWW*, 2001.

[24] M. Newman. Assortative mixing in networks. In *Phys. Rev. Lett. 89*, 2002.

[25] M. Newman. Power laws, pareto distributions and zipf's law. In *Contemporary Physics 46*, 2005.

[26] M. Porter. Facebook5 data. http://www.insna.org/software/data.html, 2008.

[27] 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.

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

[29] 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.