Walking on a Graph with a Magnifying Glass:
Stratified Sampling via Weighted Random Walks

Maciej Kurant
CalIT2
UC Irvine
mkurant@uci.edu

Minas Gjoka
CalIT2
UC Irvine
mgjoka@uci.edu

Carter T. Butts
Sociology Dept, CalIT2
UC Irvine
buttsc@uci.edu

Athina Markopoulou
EECS, CalIT2, CPCC
UC Irvine
athina@uci.edu

ABSTRACT

Our objective is to sample the node set of a large unknown graph via crawling, to accurately estimate a given metric of interest. We design a random walk on an appropriately defined weighted graph that achieves high efficiency by preferentially crawling those nodes and edges that convey greater information regarding the target metric. Our approach begins by employing the theory of stratification to find optimal node weights, for a given estimation problem, under an independence sampler. While optimal under independence sampling, these weights may be impractical under graph crawling due to constraints arising from the structure of the graph. Therefore, the edge weights for our random walk should be chosen so as to lead to an equilibrium distribution that strikes a balance between approximating the optimal weights under an independence sampler and achieving fast convergence. We propose a heuristic approach (stratified weighted random walk, or S-WRW) that achieves this goal, while using only limited information about the graph structure and the node properties. We evaluate our technique in simulation, and experimentally, by collecting a sample of Facebook college users. We show that S-WRW requires 13-15 times fewer samples than the simple re-weighted random walk (RW) to achieve the same estimation accuracy for a range of metrics.

Categories and Subject Descriptors

C.4 [**Performance of Systems**]: Measurement techniques

General Terms

Measurement, Algorithms

Keywords

Graph Sampling, Random Walks on Weighted Graphs, Stratified Sampling, Online Social Networks.

* This work was supported by SNF grant PBELP2-130871, Switzerland, and by the NSF CDI Award 1028394, USA.

SIGMETRICS'11, June 7–11, 2011, San Jose, California, USA.

1. INTRODUCTION

Many types of online networks, such as online social networks (OSNs), Peer-to-Peer (P2P) networks, or the World Wide Web (WWW), are measured and studied today via sampling techniques. This is due to several reasons. First, such graphs are typically too large to measure in their entirety, and it is desirable to be able to study them based on a small but representative sample. Second, the information pertaining to these networks is often hard to obtain. For example, OSN service providers have access to all information in their user base, but rarely make this information publicly available.

There are many ways a graph can be sampled, *e.g.,* by sampling nodes, edges, paths, or other substructures [23, 28]. Depending on our measurement goal, the elements with different properties may have *different importance* and should be sampled with a different probability. For example, Fig. 1(a) depicts the world's population, with residents of China (1.3B people) represented by blue nodes, of the Vatican (800 people) by black nodes, and all other nationalities represented by white nodes. Assume that we want to compare the median income in China and Vatican. Taking a uniform sample of size 100 from the entire world's population is ineffective, because most of the samples will come from countries other than China and Vatican. Even restricting our sample to the union of China and Vatican will not help much, as our sample is unlikely to include any Vatican resident. In contrast, uniformly sampling 50 Chinese and 50 Vaticanese residents would be much more accurate with the same sampling budget.

This type of problem has been widely studied in the statistical and survey sampling literature. A commonly used approach is *stratified sampling* [12,29,35], where nodes (*e.g.,* people) are partitioned into a set of non-overlapping *categories* (or strata). The objective is then to decide how many independent draws to take from each category, so as to minimize the uncertainty of the resulting measurement. This effect can be achieved in expectation by a weighted independence sampler (WIS) with appropriately chosen sampling probabilities π^{WIS}. In our example, WIS samples Vatican residents with much higher probabilities than Chinese ones, and avoids completely the rest of the world, as illustrated in Fig. 1(b).

However, WIS, as every independence sampler, requires a sampling frame, *i.e.,* a list of all elements we can sample from (*e.g.,* a list of all Facebook users). This information is typically not available in today's online networks. A feasible alternative is *crawling* (also known as exploration or link-

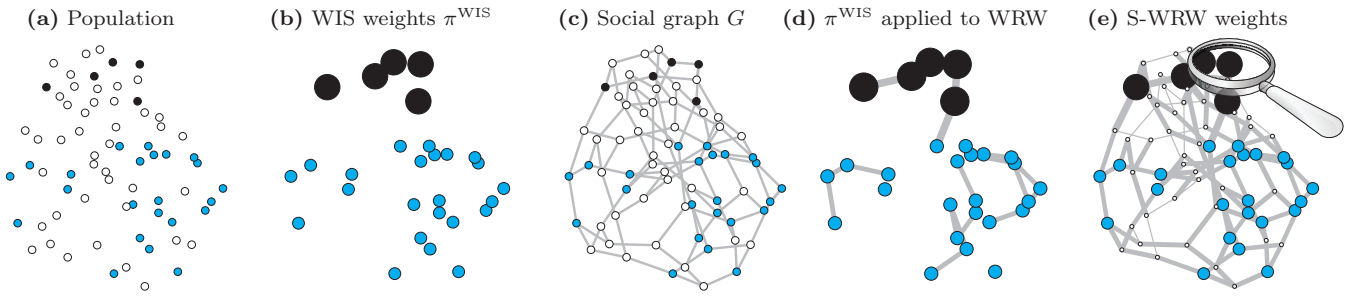


Figure 1: Illustrative example. Our goal is to compare the blue and black subpopulations (*e.g.,* with respect to their median income) in population (a). Optimal independence sampler, WIS (b), over-samples the black nodes, under-samples the blue nodes, and completely skips the white nodes. A naive crawling approach, RW (c), samples many irrelevant white nodes. WRW that enforces WIS-optimal probabilities may result in poor or no convergence (d). S-WRW (e) strikes a balance between the optimality of WIS and fast convergence.

trace sampling). It is a graph sampling technique in which we can see the neighbors of already sampled users and make a decision on which users to visit next.

In this paper, we study how to perform stratified sampling through graph crawling. We illustrate the key idea and some of the challenges in Fig. 1. Fig. 1(c) depicts a social network that connects the world's population. A simple random walk (RW) visits every node with frequency proportional to its degree, which is reflected by the node size. In this particular example, for a simplicity of illustration, all nodes have the same degree equal to 3. As a result, RW is equivalent to the uniform sample of the world's population, and faces exactly the same problems of wasting resources, by sampling all nodes with the same probability.

We address these problems by appropriately setting the edge weights and then performing a random walk on the weighted graph, which we refer to as *weighted random walk* (WRW). One goal in setting the weights is to mimic the WIS-optimal sampling probabilities π^{WIS} shown in Fig. 1(b). However, such a WRW might perform poorly due to potentially slow mixing. In our example, it will not even converge because the underlying weighted graph is disconnected, as shown in Fig. 1(d). Therefore, the edge weights under WRW (which determine the equilibrium distribution π^{WRW}) should be chosen in a way that strikes a balance between the optimality of π^{WIS} and fast convergence.

We propose *Stratified Weighted Random Walk* (S-WRW), a practical heuristic that effectively strikes such a balance. We refer to our approach as "walking on the graph with a magnifying glass", because S-WRW over-samples more relevant parts of the graph and under-samples less relevant ones. In our example, S-WRW results in the graph presented in Fig. 1(e). The only information required by S-WRW are the categories of neighbors of every visited node, which is typically available in crawlable online networks, such as Facebook. S-WRW uses two natural and easy-to-interpret parameters, namely: (i) $\tilde{f}_{\ominus}$, which controls the fraction of samples from irrelevant categories and (ii) γ, which is the maximal resolution of our magnifying glass, with respect to the largest relevant category.

The main contributions of this paper are the following.

- We propose to improve the efficiency of crawling-based graph sampling methods, by performing a stratified weighted random walk that takes into account not only the graph structure but also the node properties that are relevant to the measurement goal.
- We design and evaluate S-WRW, a practical heuristic that sets the edge weights and operates with limited information.
- As a case study, we apply S-WRW to sample Facebook and estimate the sizes of colleges. We show that S-WRW requires 13-15 times fewer samples than a simple random walk for the same estimation accuracy.

The outline of the rest of the paper is as follows. Section 2 summarizes the graph sampling techniques. Section 3 introduces S-WRW that combines stratified sampling with graph crawling. Section 4 presents simulation results. Section 5 presents an implementation of S-WRW for the problem of sampling college users on Facebook. Section 6 reviews related work. Section 7 concludes the paper.

2. SAMPLING TECHNIQUES

2.1 Notation

We consider an undirected, static,[1] graph $G = (V, E)$, with $N = |V|$ nodes and $|E|$ edges. For a node $v \in V$, denote by $\deg(v)$ its degree, and by $\mathcal{N}(v) \subset V$ the list of neighbors of v. A graph G can be weighted. We denote by $\mathrm{w}(u, v)$ the weight of edge $\{u, v\} \in E$, and by

$$\mathrm{w}(u) = \sum_{v \in \mathcal{N}(u)} \mathrm{w}(u, v) \tag{1}$$

the weight of node $u \in V$. For any set of nodes $A \subseteq V$, we define its volume $\mathrm{vol}(A)$ and weight $\mathrm{w}(A)$, respectively, as

$$\mathrm{vol}(A) = \sum_{v \in A} \deg(v) \quad \text{and} \quad \mathrm{w}(A) = \sum_{v \in A} \mathrm{w}(v). \tag{2}$$

We will often use

$$f_A = \frac{|A|}{|V|} \quad \text{and} \quad f_A^{\text{vol}} = \frac{\mathrm{vol}(A)}{\mathrm{vol}(V)} \tag{3}$$

to denote the relative size of A in terms of the number of nodes and the volumes, respectively.

Sampling. We collect a sample $S \subseteq V$ of $n = |S|$ nodes. S may contain multiple copies of the same node, *i.e.,* the sampling is with replacement. In this section, we briefly review the techniques for sampling nodes from graph G. We also present the weighted random walk (WRW) which is the basic building block for our approach.

[1]Sampling dynamic graphs is currently an active research area [36,41,43], but out of the scope of this paper.

2.2 Independence Sampling

Uniform Independence Sampling (UIS) samples the nodes directly from the set V, with replacements, uniformly and independently at random, *i.e.,* with probability

$$\pi^{\text{UIS}}(v) \;=\; \frac{1}{N} \qquad \text{for every } v \in V. \tag{4}$$

Weighted Independence Sampling (WIS) is a weighted version of UIS. WIS samples the nodes directly from the set V, with replacements, independently at random, but with probabilities proportional to node weights $\mathrm{w}(v)$:

$$\pi^{\text{WIS}}(v) \;=\; \frac{\mathrm{w}(v)}{\sum_{u\in V}\mathrm{w}(u)}. \tag{5}$$

In general, UIS and WIS are not possible in online networks because of the lack of sampling frame. For example, the list of all user IDs may not be publicly available, or the user ID space may be too sparsely allocated. Nevertheless, we present them as baseline for comparison with the random walks.

2.3 Sampling via Crawling

In contrast to independence sampling, the crawling techniques are possible in many online networks, and are therefore the main focus of this paper.

Simple Random Walk (RW) [30] selects the next-hop node v uniformly at random among the neighbors of the current node u. In a connected and aperiodic graph, the probability of being at the particular node v converges to the stationary distribution

$$\pi^{\text{RW}}(v) \;=\; \frac{\deg(v)}{2\cdot|E|}. \tag{6}$$

Metropolis-Hastings Random Walk (MHRW) is an application of the Metropolis-Hastings algorithm [31] that modifies the transition probabilities to converge to a desired stationary distribution. For example, we can achieve the uniform stationary distribution

$$\pi^{\text{MHRW}}(v) \;=\; \frac{1}{N} \tag{7}$$

by randomly selecting a neighbor v of the current node u and moving there with probability $\min(1, \frac{\deg(u)}{\deg(v)})$. However, it was shown in [18,36] that RW (after re-weighting, as in Section 2.4) outperforms MHRW for most applications. We therefore restrict our attention to comparing against RW.

Weighted Random Walk (WRW) is RW on a weighted graph [4]. At node u, WRW chooses the edge $\{u,v\}$ to follow with probability $P_{u,v}$ proportional to the weight $\mathrm{w}(u,v)\geq 0$ of this edge, *i.e.,*

$$P_{u,v} = \frac{\mathrm{w}(u,v)}{\sum_{v'\in\mathcal{N}(u)}\mathrm{w}(u,v')}. \tag{8}$$

The stationary distribution of WRW is:

$$\pi^{\text{WRW}}(v) \;=\; \frac{\mathrm{w}(v)}{\sum_{u\in V}\mathrm{w}(u)}. \tag{9}$$

WRW is the basic building block of our design. In the next sections, we show how to choose weights for a specific estimation problem.

Graph Traversals (BFS, DFS, RDS, ...) is a family of crawling techniques where no node is sampled more than once. Because traversals introduce a generally unknown bias (see Section 6), we do not consider them in this paper.

2.4 Correcting the bias

RW, WRW, and WIS all produce biased (nonuniform) node samples. But their bias is known and therefore can be corrected by an appropriate re-weighting of the measured values. This can be done using the Hansen-Hurwitz estimator [20] as first shown in [40,42] for random walks and also used in [36]. Let every node $v\in V$ carry a value $x(v)$. We can estimate the population total $x_{\text{tot}}=\sum_v x(v)$ by

$$\hat{x}_{\text{tot}} = \frac{1}{n}\sum_{v\in S}\frac{x(v)}{\pi(v)}, \tag{10}$$

where $\pi(v)$ is the sampling probability of node v in the stationary distribution. In practice, we usually know $\pi(v)$, and thus $\hat{x}_{\text{tot}}$, only up to a constant, *i.e.,* we know the (non-normalized) weights $\mathrm{w}(v)$. This problem disappears when we estimate the population mean $x_{\text{av}}=\sum_v x(v)/N$ as

$$\hat{x}_{\text{av}} \;=\; \frac{\sum_{v\in S}\frac{x(v)}{\pi(v)}}{\sum_{v\in S}\frac{1}{\pi(v)}} \;=\; \frac{\sum_{v\in S}\frac{x(v)}{\mathrm{w}(v)}}{\sum_{v\in S}\frac{1}{\mathrm{w}(v)}}. \tag{11}$$

For example, for $x(v)=1$ if $\deg(v)=k$ (and $x(v)=0$ otherwise), $\hat{x}_{\text{av}}(k)$ estimates the node degree distribution in G.

All the results in this paper are presented *after this re-weighting* step, whenever necessary.

3. STRATIFIED WRW

In this section we introduce Stratified Weighted Random Walk (S-WRW). S-WRW builds on stratification under the optimal independence sampler, and additionally addresses practical challenges arising in graph crawling.

3.1 Stratified Independence Sampling

In Section 1, we argued that in order to compare the median income of residents of China and Vatican we should take 50 random samples from each of these two countries, rather than taking 100 UIS samples from China and Vatican together (or, even worse, from the world's population). This problem naturally arises in the field of survey sampling. A common solution is *stratified sampling* [12,29,35], where nodes V are partitioned into a set $\mathcal{C}=\{C_1,C_2,\ldots,C_{|\mathcal{C}|}\}$ of non-overlapping node categories (or "strata"), with union $\bigcup_{C\in\mathcal{C}} C = V$. Next, we select uniformly at random n_i nodes from category C_i. We are free to choose the allocation $(n_1,n_2,\ldots,n_{|\mathcal{C}|})$, as long as we respect the total budget of samples $n=\sum_i n_i$.

There are many possible allocations n_i. We are interested in the optimal allocation n_i^{opt}, that minimizes the measurement error with respect to our measurement objective. In [25] we show how to calculate n_i^{opt} for various measurement scenarios. One obvious hint is to set $n_{\ominus}^{\text{opt}}=0$ for the *irrelevant category* $C_{\ominus}\in\mathcal{C}$ that groups all nodes not relevant to our measurement objective. For example, in Fig. 1, $C_{\ominus}$ consists of all white nodes.

We also show in [25] how to allocate the samples between the relevant categories. If (i) we are interested in comparing the node categories with respect to some properties (*e.g.,* average node degree, category size) rather than estimating a

property across the entire population, and (ii) no additional information is available (such as property variances - rarely known in practice), then we should take an equal number of samples from every relevant category, *i.e.,* use

$$n_i^{\text{opt}} \;=\; \frac{n}{|\mathcal{C}\setminus\{C_\ominus\}|} \qquad \text{for every } C_i \neq C_\ominus. \tag{12}$$

Stratification in Expectation with WIS. Ideally, we would like to enforce strictly stratified sampling and collect exactly n_i^{opt} samples from category C_i. However, when we use crawling, strict stratification is possible only by discarding observations. It is thus more natural to frame the problem in terms of the probability mass placed on each category, with the goal of collecting n_i^{opt} samples from category C_i *in expectation*. Under WIS, this is achieved by enforcing that (see Appendix C for the derivation):

$$\mathrm{w}^{\text{WIS}}(C_i) \;\propto\; n_i^{\text{opt}}, \tag{13}$$

where $\mathrm{w}^{\text{WIS}}(C_i) = \sum_{v\in C_i}\mathrm{w}^{\text{WIS}}(v)$ is the weight of category C_i. In strictly stratified sampling, the individual node sampling probabilities $\mathrm{w}^{\text{WIS}}(v)$ are equal across the category C_i. Achieving it by setting the edge weights (as we do in crawling) would require the knowledge of entire graph G before we start sampling, which is, of course, impractical. Instead, we show below that we are able to effectively obtain the necessary information at the category-level granularity, which allows us to control the aggregated weight $\mathrm{w}^{\text{WIS}}(C_i)$.

3.2 Stratified Crawling

We have argued earlier that, due to lack of a sampling frame, the independence sampling (including its stratified version) is typically infeasible in online networks, making crawling the only practical alternative. In this section, we show how to perform a weighted random walk (WRW) which approximates the stratified sampling. The general problem can be stated as follows:

Given a category-related measurement objective, an error metric and a sampling budget $|S|=n$, set the edge weights in graph G such that WRW on this graph achieves a minimal estimation error.

Although we are able to solve this problem analytically for some specific and fully known topologies, it is not obvious how to address it in general, especially under a limited knowledge of G. Instead, in this paper, we propose S-WRW, a heuristic to set the edge weights. S-WRW starts from a solution that is optimal under WIS, and takes into account practical issues that arise in graph crawling. Once the edge weights in G are set, we simply perform a WRW as described in Section 2.3 and we collect node samples.

3.3 Our practical solution: S-WRW

As the main (but not the only) guideline, S-WRW tries to realize the category weights that are optimal under WIS, *i.e.,* to achieve

$$\mathrm{w}^{\text{WRW}}(C_i) \;=\; \mathrm{w}^{\text{WIS}}(C_i). \tag{14}$$

There are many edge weight settings in G that satisfy it. We give preference to equal weights, as follows. First, note that if every edge incident on nodes of C_i carries the same weight $\mathrm{w}_e(C_i)$, then $\mathrm{w}^{\text{WRW}}(C_i) = \mathrm{w}_e(C_i)\cdot\mathrm{vol}(C_i)$. Consequently, we can achieve Eq.(14) by setting

$$\mathrm{w}_e(C_i) \;=\; \frac{\mathrm{w}^{\text{WIS}}(C_i)}{\mathrm{vol}(C_i)}.$$

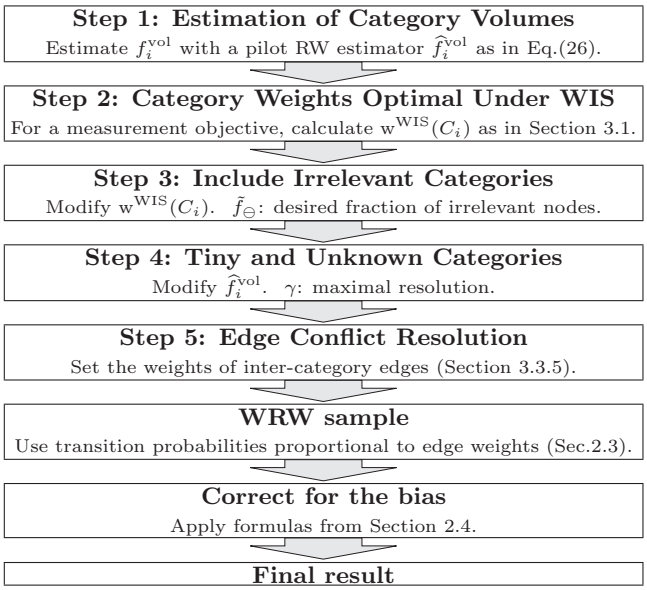


Figure 2: Overview of S-WRW.

In fact, we need to know $\mathrm{w}_e(C_i)$ only up to a constant factor, because these factors cancel out in the calculation of transition probabilities of WRW in Eq.(8). Therefore, exactly the same WRW can be obtained by setting

$$\mathrm{w}_e(C_i) \;=\; \frac{\mathrm{w}^{\text{WIS}}(C_i)}{f_i^{\text{vol}}}. \tag{15}$$

This formulation replaces the absolute volume $\mathrm{vol}(C_i)$ of category C_i by its relative version f_i^{vol} that is much easier to estimate (similarly to x_{tot} and x_{av} in Section 2.4).

Eq.(15) is central to the S-WRW heuristic. But in order to apply it, we first have to calculate or estimate its terms f_i^{vol} and $\mathrm{w}^{\text{WIS}}(C_i)$. Below, we show how to do that in Steps 1 and 2, respectively. Next, in Steps 3-5, we show how to modify these terms to account for practical problems arising from the underlying graph structure.

3.3.1 Step 1: Estimation of Category Volumes

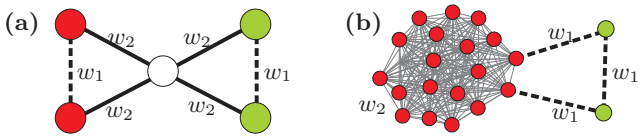
In general, we have no prior information about G. Fortunately, it is easy and inexpensive to estimate the relative category volumes f_i^{vol}, which is the first piece of information we need in Eq.(15). Indeed, it is enough to run a relatively short *pilot* RW, and use the collected sample S in Eq.(26) derived in Appendix B, and repeated here for convenience:

$$\widehat{f}_i^{\text{vol}} \;=\; \frac{1}{n}\sum_{u\in S}\left(\frac{1}{\deg(u)}\sum_{v\in\mathcal{N}(u)}1_{\{v\in C_i\}}\right).$$

3.3.2 Step 2: Category Weights Optimal Under WIS

In order to find the optimal WIS category weights $\mathrm{w}^{\text{WIS}}(C_i)$ in Eq.(15), we first count all the categories discovered by the pilot RW in Step 1, and use it as an estimator of the real number $|\mathcal{C}|$ of existing categories. Next, we calculate n_i^{opt} as shown in Section 3.1, and we plug it in Eq.(13), *e.g.,* by setting $\mathrm{w}^{\text{WIS}}(C_i) = n_i^{\text{opt}}$.

In particular, in the case where all relevant categories are equally important (which is rather common), we apply Eq.(12) and Eq.(13). This boils down to assigning the



WIS: $w_1 > 0,\ w_2 = 0$ WIS: $w_1 = 190\, w_2$
WRW: $w_1 = 0,\ w_2 > 0$ WRW: $w_1 \cong 60\, w_2$ for n=50
$w_1 \cong 100\, w_2$ for n=500
$w_1 = 190\, w_2$ for $n \to \infty$

Figure 3: Optimal edge weights: WIS vs WRW. The objective is to compare the sizes of red (dark) and green (light) categories.

same weight to every category C_i, *e.g.,* $\mathrm{w}^{\mathrm{WIS}}(C_i) = 1$, with no need of exploiting the pilot RW.

3.3.3 Step 3: Irrelevant Categories

Problem: Potentially poor or no convergence. Trying to achieve the optimal category weights may lead to poor or no convergence of the random walk. We already discussed this problem in Fig. 1. As another illustrative example, consider the toy example in Fig. 3(a) and assume that we are interested in finding the relative sizes of red (dark) and green (light) categories. The white node in the middle is irrelevant for our measurement objective. Due to symmetry, we distinguish between two types of edges with weights w_1 and w_2. Under WIS, the optimal weights are $w_1 > 0$ and $w_2 = 0$ (see [25]), *i.e.,* WIS samples every non-white node with the same probability and never samples the white one. However, under WRW with these weights, relevant nodes get disconnected into two components and WRW does not converge.

Guideline: Occasionally visit irrelevant nodes. We show in [25] that the optimal WRW weights in Fig. 3(a) are $w_1 = 0$ and $w_2 > 0$. In that case, half of the samples are due to visits in the white (irrelevant) node. In other words, WRW may benefit from allocating small weight $\mathrm{w}(C_\ominus) > 0$ to category $C_\ominus$ that groups all (if any) categories irrelevant to our estimation. The intuition is that irrelevant nodes may not contribute to estimation but may be needed for connectivity or fast mixing.

Implementation in S-WRW. In S-WRW, we achieve this goal by replacing in Eq.(15) the term $\mathrm{w}^{\mathrm{WIS}}(C_i)$ with

$$\tilde{\mathrm{w}}^{\mathrm{WIS}}(C_i) = \begin{cases} \mathrm{w}^{\mathrm{WIS}}(C_i) & \text{if } C_i \neq C_\ominus \\ \tilde{f}_\ominus \cdot \sum_{C \neq C_\ominus} \mathrm{w}^{\mathrm{WIS}}(C) & \text{if } C_i = C_\ominus. \end{cases} \quad (16)$$

The parameter $0 \leq \tilde{f}_\ominus \ll 1$ controls the desired fraction of visits in $C_\ominus$.

3.3.4 Step 4: Tiny and Unknown Categories

Problem: "black holes". Every optical system has a fundamental magnification limit due to diffraction and our "graph magnifying glass" is no exception. Consider the toy graph in Fig. 3(b): it consists of a big clique C_{big} of 20 red nodes with edge weights w_2, and a green category C_{tiny} with two nodes only and edge weights w_1. WIS optimally estimates the relative sizes of red and green categories for $\mathrm{w}(C_{\mathrm{big}}) = \mathrm{w}(C_{\mathrm{tiny}})$, *i.e.,* for $w_1 = 190\, w_2$ (see [25]). However, for such large values of w_1, the two green nodes behave as a "black hole" for a WRW of finite length, thus increasing the variance of the category size estimation.

Guideline: limit edge weights of tiny categories. In Fig. 3(b), the setting $w_1 \simeq 60\, w_2$ ($\ll 190 w_2$) is optimal for WRW of length $n = 50$ (simulation results). In other words, although WIS suggests to over-sample small categories, WRW should "under-over-sample" very small categories to avoid black holes.

Implementation in S-WRW. In S-WRW, we achieve this goal by replacing f_i^{vol} in Eq.(15) with

$$\tilde{f}_i^{\mathrm{vol}} = \max\left\{\widehat{f}_i^{\mathrm{vol}},\ f_{\min}^{\mathrm{vol}}\right\}, \quad \text{where} \quad (17)$$

$$f_{\min}^{\mathrm{vol}} = \frac{1}{\gamma} \cdot \max_{C_i \neq C_\ominus} \{\widehat{f}_i^{\mathrm{vol}}\}. \quad (18)$$

Moreover, this formulation takes care of every category C_i that was not discovered by the pilot RW in Section 3.3.1, by setting $\tilde{f}_i^{\mathrm{vol}} = f_{\min}^{\mathrm{vol}}$.

3.3.5 Step 5: Edge Conflict Resolution

Problem: Conflicting desired edge weights. With the above modifications, our target edge weights defined in Eq.(15) can be rewritten as

$$\tilde{\mathrm{w}}_e(C_i) = \frac{\tilde{\mathrm{w}}^{\mathrm{WIS}}(C_i)}{\tilde{f}_i^{\mathrm{vol}}}. \quad (19)$$

Denote by $C(v)$ the category of node v. We can directly set the weight $\mathrm{w}(u,v) = \tilde{\mathrm{w}}_e(C(u)) = \tilde{\mathrm{w}}_e(C(v))$ for every intra-category edge $\{u,v\}$. But for every inter-category edge, we may have conflicting weights $\tilde{\mathrm{w}}_e(C(u)) \neq \tilde{\mathrm{w}}_e(C(v))$ desired at the two ends of the edge in the two different categories.

Fortunately, we show in Appendix A that we can achieve any target category weights by setting edge weights (under a mild assumption that there exists at least one intra-category link within each category - this link is the required self-loop). However, the construction therein is likely to result in high weights on intra-category edges and small weights on inter-category edges, making WRW stay in small categories C_{tiny} for a long time.

Guideline: prefer inter-category edges. In order to improve the mixing time, we should do exactly the opposite, *i.e.,* assign relatively high weights to inter-category edges (connecting relevant categories). As a result, WRW will enter C_{tiny} more often, but will stay there for a short time. This intuition is motivated by Monte Carlo variance reduction techniques such as the use of *antithetic variates* [15], which seek to induce negative correlation between consecutive draws so as to reduce the variance of the resulting estimator.

Implementation in S-WRW. We assign an edge weight $\tilde{\mathrm{w}}_e$ that is in between $\tilde{\mathrm{w}}_e(C(u))$ and $\tilde{\mathrm{w}}_e(C(v))$. We consider several choices for combining the two conflicting weights.

$$\begin{aligned} \mathrm{w}^{\mathrm{ar}}(u,v) &= \frac{\tilde{\mathrm{w}}_e(C(u)) + \tilde{\mathrm{w}}_e(C(v))}{2} \\ \mathrm{w}^{\mathrm{ge}}(u,v) &= \sqrt[2]{\tilde{\mathrm{w}}_e(C(u)) \cdot \tilde{\mathrm{w}}_e(C(v))} \\ \mathrm{w}^{\max}(u,v) &= \max\{\tilde{\mathrm{w}}_e(C(u)), \tilde{\mathrm{w}}_e(C(v))\} \\ \mathrm{w}^{\mathrm{hy}}(u,v) &= \begin{cases} \mathrm{w}^{\mathrm{ge}}(u,v) & \text{if } C_\ominus \in \{C(u), C(v)\} \\ \mathrm{w}^{\max}(u,v) & \text{otherwise.} \end{cases} \end{aligned}$$

w^{ar} and w^{ge} are the arithmetic and geometric means, respectively. $\mathrm{w}^{\max}$ should improve mixing, but could assign high weight to irrelevant nodes. We avoid this undesired effect in a *hybrid* solution w^{hy}.

We found that the hybrid edge assignment works best in practice; see Section 5.

3.4 Discussion

3.4.1 Information needed about the neighbors

In the pilot RW (Section 3.3.1) as well as in the main WRW, we assume that by sampling a node v we also learn the category $C(u)$ of each of its neighbors $u \in \mathcal{N}(v)$. Fortunately, such information is typically available in most online graphs at no additional cost, especially when scraping HTML pages, as we do. For example, when sampling colleges in Facebook in Section 5, we use the college membership information of all v's neighbors, which is available at v together with the friends list.

Our approach could potentially further benefit from the knowledge of the degree of v's neighbors. However, this information is rarely available without sampling these neighbors, which is costly and thus not required by S-WRW.

3.4.2 Cost of pilot RW

The pilot RW volume estimator described in Section 3.3.1 considers the categories not only of the sampled nodes, but also of their neighbors. As a result, it achieves high efficiency, as we show in simulations (Section 4.2.1) and Facebook measurements (Section 5.1). Given that, and the high robustness of S-WRW to estimation errors (Section 4.2.5), the pilot RW should be only a small fraction of main S-WRW. For example, this is equal to 6.5% in our Facebook measurements in Section 5.

3.4.3 Setting the parameters

S-WRW sets the edge weights trying to achieve roughly $\mathrm{w}^{\mathrm{WIS}}(C_i)$. We slightly modify $\mathrm{w}^{\mathrm{WIS}}(C_i)$ to avoid black holes and improve mixing, which is controlled by two natural and easy-to-interpret parameters, $\tilde{f}_\ominus$ and γ.

Visits to irrelevant nodes $\tilde{f}_\ominus$. Parameter $0 \le \tilde{f}_\ominus \ll 1$ controls the desired fraction of visits in $C_\ominus$. When setting $\tilde{f}_\ominus$, we should exploit the information provided by the pilot RW. If the relevant categories appear poorly interconnected and often separated by irrelevant nodes, we should set $\tilde{f}_\ominus$ relatively high. We have seen an extreme case in Fig. 3(a), with disconnected relevant categories and optimal $\tilde{f}_\ominus = 0.5$. In contrast, when the relevant categories are strongly interconnected, we should use much smaller $\tilde{f}_\ominus$. However, because we can never be sure that the graph induced on relevant nodes is connected, we recommend using $\tilde{f}_\ominus > 0$. For example, when measuring Facebook in Section 5, we set $\tilde{f}_\ominus = 1\%$.

Maximal resolution γ. The parameter $\gamma \ge 1$ can be interpreted as the maximal resolution of our "graph magnifying glass", with respect to the largest relevant category C_{big}. S-WRW will typically sample well all categories whose size is at least equal to $|C_{\mathrm{big}}|/\gamma$.[2] All categories smaller than that are relatively under-sampled (see Section 5.2.4). In the extreme case, for $\gamma \to \infty$, S-WRW tries to cover every category, no matter how small, which may cause the "black hole" problem discussed in Section 3.3.4. In the other extreme, for $\gamma = 1$, and for identical $\mathrm{w}^{\mathrm{WIS}}(C_i)$ for all categories, S-WRW reduces to RW. We recommend always setting $1 < \gamma < \infty$. Ideally, we know $|C_{\mathrm{smallest}}|$ - the smallest category size that is still relevant to us. In that case we should set $\gamma = |C_{\mathrm{big}}|/|C_{\mathrm{smallest}}|$. For example, in Section 5 the categories are US colleges; we set $\gamma = 1000$, because colleges with size smaller than 1/1000th of the largest one (*i.e.,* with a few tens of students) seem irrelevant to our measurement. As another rule of thumb, we should try to set smaller γ for relatively small sample sizes and in graphs with tight community structure (see Section 4.2.5).

[2]Strictly speaking, γ is related to volumes $\mathrm{vol}(C_i)$ rather than sizes $|C_i|$. They are equivalent when category volume is proportional to its size.

3.4.4 Conservative approach

Note that a reasonable setting of these parameters (*i.e.,* $\tilde{f}_\ominus > 0$ and $1 < \gamma < \infty$, and any conflict resolution discussed in the paper), increases the weights of large categories (including $C_\ominus$) and decreases the weight of small categories, compared to $\mathrm{w}^{\mathrm{WIS}}(C_i)$. This makes S-WRW allocate category weights between the two extremes: RW and WIS. In this sense, S-WRW can be considered conservative.

3.4.5 S-WRW is unbiased

It is also important to note that the collected WRW sample is eventually corrected with the actual sampling weights as described in Section 2.4. Consequently, the S-WRW estimation process is unbiased, regardless of the choice of weights, as long as convergence is attained. In contrast, suboptimal weights (*e.g.,* due to estimation error of $\widehat{f}_C^{\mathrm{vol}}$) can increase the WRW mixing time and/or the variance of the resulting estimator. However, our simulations (Section 4) and empirical experiments on Facebook (Section 5) show that S-WRW is robust to suboptimal choice of weights.

4. SIMULATION RESULTS

The gain of our approach compared to RW comes from two main factors. First, S-WRW avoids, to a large extent or completely, the nodes in $C_\ominus$ that are irrelevant to our measurement. This fact alone can bring an arbitrarily large improvement ($\frac{N}{N-|C_\ominus|}$ under WIS), especially when $C_\ominus$ is large compared to N. We demonstrate this in the Facebook measurements in Section 5. Second, we can better allocate samples among the relevant categories. This factor is observable in our Facebook measurements as well, but it is more difficult to evaluate due to the lack of ground-truth therein. In this section, we evaluate the optimal allocation gain in a controlled simulation that illustrates some key insights.

4.1 Setup

4.1.1 Topology

We consider a graph G with 101K nodes and 505.5K edges organized in two densely connected communities[3] as shown in Fig. 4(h). The inter- and intra-community edges are chosen at random.

The nodes in G are partitioned into two node categories: C_{tiny} with 1K nodes (dark red), and C_{big} with 100K nodes (light yellow). We consider two extreme scenarios of such a partition. The "Random" scenario uses a purely random partition, as shown in Fig. 4(a). In contrast, under "Clustered", categories C_{tiny} and C_{big} coincide with the existing communities in G, as shown in Fig. 4(h). Clustered is arguably the worst case scenario for graph sampling by exploration.

[3]The term "community" refers to cluster and is defined purely based on topology. The term "category" is a property of a node and is independent of topology.

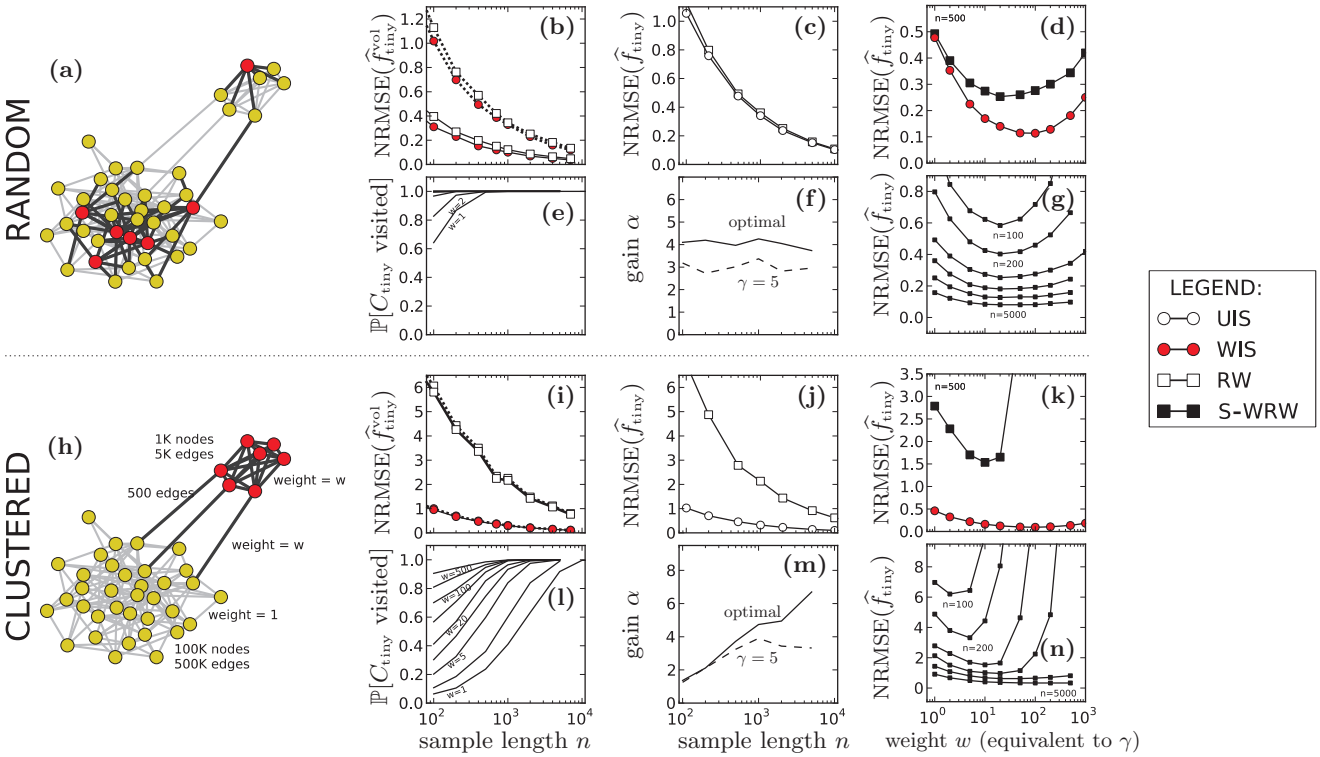


Figure 4: RW and S-WRW under two scenarios: Random (a-g) and Clustered (h-n). In (b,i), we show error of two volume estimators: naive Eq.(23) (dotted) and neighbor-based Eq.(26) (plain). Next, we show error of size estimator as a function of n (c,j) and w (d,g,k,n); in the latter, UIS and RW correspond to WIS and S-WRW for $w=1$. In (e,l), we show the empirical probability that S-WRW visits C_{tiny} at least once. Finally, (f,m) is gain α of S-WRW over RW under the optimal choice of w (plain), and for fixed $\gamma=w=5$ (dashed).

We fix the edge weights of all internal edges in C_{big} to 1. All the remaining edges, *i.e.,* all edges incident on nodes in category C_{tiny}, have weight w each, where $w \geq 1$ is a parameter. Note that this is equivalent to setting $\tilde{w}_e(C_{\text{big}})=1$, $\tilde{w}_e(C_{\text{tiny}})=w$, and max or hybrid conflict resolution.

4.1.2 Objective and performance metrics

We are interested in measuring the relative sizes f_{tiny} and f_{big} (see Eq.(3)) of categories C_{tiny} and C_{big}, respectively. We use Normalized Root Mean Square Error (NRMSE) to assess the estimation error, defined in [38] as:

$$\text{NRMSE}(\widehat{x}) = \frac{\sqrt{\mathbb{E}\left[(\widehat{x}-x)^2\right]}}{x}, \qquad (20)$$

where x is the real value and $\widehat{x}$ is the estimated one.

In order to simplify the practical interpretation of the results, we also show how NRMSE translates into sample length. We define as *gain* α of S-WRW over RW the number of times RW must be longer than S-WRW in order to achieve the same error NRMSE, *i.e.,*

$$\text{gain } \alpha = \frac{n^{\text{RW}}}{n^{\text{S-WRW}}},$$

subject to $\text{NRMSE}^{\text{RW}} = \text{NRMSE}^{\text{S-WRW}}$.

4.2 Results

4.2.1 Estimating volumes is usually cheap

The first step in S-WRW is obtaining category volume estimates $\widehat{f}_i^{\text{vol}}$. We achieve it by running a short pilot RW and applying the estimator Eq.(26). We show $\text{NRMSE}(\widehat{f}_{\text{tiny}}^{\text{vol}})$ as plain curves in Fig. 4(b). This estimator takes advantage of the knowledge of the categories of the neighboring nodes, which makes it much more efficient than the naive estimator Eq.(23) shown by dashed curves. Moreover, the advantage of Eq.(26) over Eq.(23) grows with the graph density and the skewness of its degree distribution (not shown here).

Note that under Random, RW and WIS (with the sampling probabilities of RW) are almost equally efficient. However, on the other extreme, *i.e.,* under Clustered, the performance of RW becomes much worse and the advantage of Eq.(26) over Eq.(23) diminishes. This is because essentially all neighbors of a node from category C_i are in C_i too, which reduces formula Eq.(26) to Eq.(23). Nevertheless, we show in Section 4.2.5 that even several fold volume estimation errors are likely not to affect significantly the results.

4.2.2 Visiting the tiny category

Fig. 4(e,l) presents the empirical probability $\mathbb{P}[C_{\text{tiny}}$ visited$]$ that our walk visits at least one node from C_{tiny}. Of course, this probability grows with the sample length. However, the choice of weight w also affects it. Indeed, WRW with $w > 1$ is more likely to visit C_{tiny} than RW ($w = 1$, bottom line). This demonstrates the first advantage of introducing edge weights and WRW.

4.2.3 Optimal w and γ

Let us now focus on the estimation error as a function of w, shown in Fig. 4(d,k). Interestingly, this error does not

drop monotonically with w but follows a U-shaped function with a clear optimal value w^{opt}.

Under WIS, we have $w^{\text{opt}} \simeq 100$, which confirms our findings discussed in Section 3.1. In particular, we achieve the optimal solution for the same number of samples $n^{\text{opt}}_{\text{tiny}} = n^{\text{opt}}_{\text{big}}$, which translates to $\text{w}^{\text{WIS}}(C_{\text{tiny}}) = \text{w}^{\text{WIS}}(C_{\text{big}})$. By plugging this and $f^{\text{vol}}_{\text{big}} = 100 \cdot f^{\text{vol}}_{\text{tiny}}$ to Eq.(15), we finally obtain the WIS-optimal edge weights in C_{tiny}, *i.e.,* $w^{\text{opt}} = \text{w}_e(C_{\text{tiny}}) = 100 \cdot \text{w}_e(C_{\text{big}}) = 100$.[4]

In contrast, WRW is optimized for $w < 100$. For the sample length $n = 500$ as in Fig. 4(d,k), the error is minimized already for $w^{\text{opt}} \simeq 20$ and increases for higher weights. This demonstrates the "black hole" effect discussed in Section 3.3.4. It is much more pronounced under Clustered, confirming our intuition that black-holes become a problem only in the presence of relatively isolated, tight communities. Of course, the black hole effect diminishes with the sample length n (and vanishes for $n \to \infty$), which can be observed in Fig. 4(g,n), especially in (n).

In other words, the optimal assignment of edge weights (in relevant categories) under WRW lies somewhere between RW (all weights equal) and WIS. In S-WRW, we control it by parameter γ. In this example, we have $\gamma \equiv w$ for $\gamma \leq 100$. Indeed, by combining Eq.(15), Eq.(17), Eq.(18) and $\text{w}^{\text{WIS}}(C_{\text{tiny}}) = \text{w}^{\text{WIS}}(C_{\text{big}})$, we obtain

$$w \;=\; \frac{w_e(C_{\text{tiny}})}{w_e(C_{\text{big}})} \;=\; \frac{\frac{\text{w}^{\text{wis}}(C_{\text{tiny}})}{\tilde{f}^{\text{vol}}_{\text{tiny}}}}{\frac{\text{w}^{\text{wis}}(C_{\text{big}})}{\tilde{f}^{\text{vol}}_{\text{big}}}} = \frac{\tilde{f}^{\text{vol}}_{\text{big}}}{\tilde{f}^{\text{vol}}_{\text{tiny}}} \;=\; \frac{f^{\text{vol}}_{\text{big}}}{\frac{1}{\gamma} f^{\text{vol}}_{\text{big}}} \;=\; \gamma.$$

Consequently, the optimal setting of γ is the same as w^{opt}.

4.2.4 Gain α

A comparison of Fig. 4(c) and Fig. 4(d) reveals that a 500 hop-long WRW with $w \simeq 20$ yields roughly the same error NRMSE $\simeq 0.3$ as a 2000 hop-long RW. This means that WRW reduces the sampling cost by a factor of $\alpha \simeq 4$. Fig. 4(f) shows that this gain does not vary much with the sampling length. Under Clustered, both RW and WRW perform much worse. Nevertheless, Fig. 4(m) shows that WRW may significantly reduce the sampling cost in this scenario as well, especially for longer samples.

It is worth noting that WRW can significantly outperform UIS. This is the case in Fig. 4(d), where UIS is equivalent to WIS with $w = 1$. Because no walk can mix faster than UIS (that is independent and thus has perfect mixing), improving the mixing time alone [5,10,38,39] cannot achieve the potential gains of stratification, in general.

So far we focused on the smaller set C_{tiny} only. When estimating the size of C_{big}, all errors are much smaller, but we observe similar gain α.

4.2.5 Robustness to γ and volume estimation

The gain α shown above is calculated for the optimal choice of w, or, equivalently, γ. Of course, in practice it might be impossible to analytically obtain this value. Fortunately, S-WRW is relatively robust to the choice of parameters. The dashed lines in Fig. 4(f,m) are calculated for γ fixed to $\gamma = 5$, rather than optimized. Note that this value is often drastically smaller than the optimal one (*e.g.,* $w^{\text{opt}} \simeq 50$ for $n = 5000$). Nevertheless, although the performance somewhat drops, S-WRW still reduces the sampling cost about three-fold.

This observation also illustrates the robustness to the category volume estimation errors (see Section 3.3.1). Indeed, setting $\gamma = 5$ means that every category C_i with volume estimated at $\widehat{f}^{\text{vol}}_i \leq \frac{1}{5}\widehat{f}^{\text{vol}}_{\text{big}}$ is treated the same. In Fig. 4(f), the volume of C_{tiny} would have to be overestimated by more than 20 times in order to affect the edge weight setting and thus the results. We have seen in Section 4.2.1 that this is very unlikely, even under smallest sample lengths and most adversarial scenarios.

4.3 Summary

S-WRW brings two types of benefits: (i) it avoids irrelevant nodes $C_\ominus$ and (ii) it carefully allocates samples between relevant categories of different sizes. Even for $C_\ominus = \emptyset$, *i.e.,* the scenario studied in this section, S-WRW can still reduce the sampling cost by 75%. This second benefit is more difficult to achieve when the categories form strong and tight communities, which may lead to the black hole effect. We should then choose smaller, more conservative values of γ in S-WRW, which translate into smaller w in our example. In contrast, under a looser community structure this problem disappears and S-WRW is closer to WIS.

5. IMPLEMENTATION IN FACEBOOK

As a concrete application, we apply S-WRW to measure the Facebook social graph. This is an undirected graph and can also be considered a static graph, for all practical purposes in this study.[5] In Facebook, every user may declare herself a member of a college[6] he/she attends. We interpret the college affiliation as a user's category. This information is publicly available by default and allows us to answer some interesting questions. For example, how do the college networks (or "colleges" for short) compare with respect to their sizes? What is the college-to-college friendship graph? In order to answer these questions, one needs to collect many college user samples, preferably evenly distributed across colleges. This is the main goal of this section.

5.1 Measurement Setup

By default, the publicly available information for every Facebook user includes the name, photo, and a list of friends together with their college memberships (if any). We developed a high performance multi-threaded crawler to explore Facebook's social graph by scraping this web interface.

To make an informed decision about the parameters of S-WRW, we first ran a short pilot RW (see Section 3.3.1) with a total of $65K$ samples (which is only 6.5% of the length of the main S-WRW sample). Although our pilot RW visited only 2000 colleges, it estimated the relative volumes f^{vol}_i for about 9500 colleges discovered among friends of sampled users, as discussed in Section 3.4.2. In Fig. 6(a), we show that the neighbor-based estimator Eq.(26) greatly outperforms the naive estimator Eq.(23). These volumes cover

[4]For simplicity, we ignored in this calculation the conflicts on the 500 edges between C_{big} and C_{tiny}.

[5]The Facebook characteristics do change but in time scales much longer than the 3-day duration of our crawls. Websites such as Facebook statistics, Alexa etc show that the number of Facebook users is growing with rate 0.1-0.2% per day.

[6]There also exist categories other than colleges, namely "work" and "high school". Facebook requires a valid category-specific email for verification.

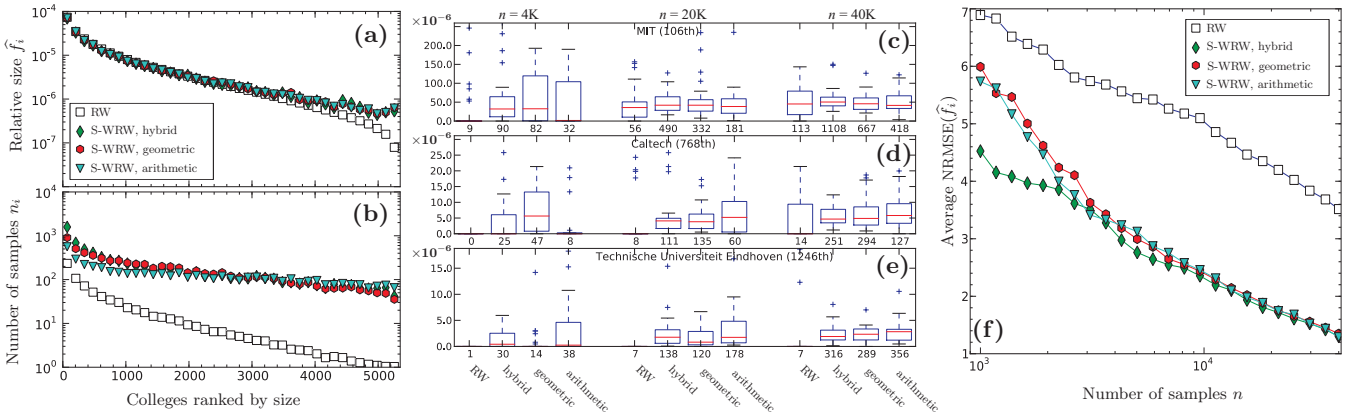


Figure 5: 5331 colleges discovered and ranked by RW. (a) Estimated relative college sizes $\widehat{f}_i$. (b) Absolute number of user samples per college. (c-e) 25 estimates of size $\widehat{f}_i$ for three different colleges and sample lengths n. (f) Average NRMSE of college size estimation. Results in (a,b,f) are binned.

	RW	S-WRW		
		Hybrid	Geometric	Arithmetic
Unique samples	1,000K	1,000K	1,000K	1,000K
Total samples	1,016K	1,263K	1,228K	1,237K
College samples	9%	86%	79%	58%
Unique Colleges	5,331	9,014	8,994	10,439

Table 1: Overview of collected Facebook datasets.

several decades. We set the maximal resolution to $\gamma = 1000$, which means that we target colleges with at least a few tens of users (see the discussion in Section 3.4.3).

We also used the information collected by the pilot RW to set the desired fraction $\tilde{f}_\ominus$ of irrelevant nodes. We found that a typical college user visited by pilot RW (without correcting for the degree bias) has on average 733 friends: 103 in the same college, 141 in a different college, and 489 without any college affiliation. Such a high number of inter-college links should generally result in a good S-WRW mixing even with no visits to the irrelevant (non-college) nodes, *i.e.,* for $\tilde{f}_\ominus = 0$. However, in order to account for rare but possible cases with a college user(s) surrounded exclusively by non-college friends, we chose a small but positive parameter $\tilde{f}_\ominus = 1\%$.

In the main measurement phase, we perform three S-WRW crawls, each with different edge weight conflict resolution (hybrid, geometric, and arithmetic), and one simple RW crawl as a baseline for comparison (Table 1). For each crawl type we collected 1 million unique users. Some of them are sampled multiple times (at no additional cost), which results in higher total number of samples in the second row of Table 1. All the results presented here would look almost the same for 1 million total (rather than unique) samples. Our crawls were performed on Oct. 16-19 2010, and the datasets are available at [1].

5.2 Results: RW vs. S-WRW

5.2.1 *Avoiding irrelevant categories*

Only 9% of the RW's samples come from colleges, which means that the vast majority of sampling effort is wasted. In contrast, the S-WRW crawls achieved 6-10 times better efficiency, collecting 86% (hybrid), 79% (geometric) and 58% (arithmetic) samples from colleges. Note that these values are significantly lower than the target 99% suggested by our choice of $\tilde{f}_\ominus = 1\%$, and that S-WRW hybrid reaches the highest number. This is in agreement with our discussion in Section 3.3.5. We also note that S-WRW crawls discovered $1.6 - 1.9$ times more unique colleges than RW.

At first, it seems surprising that RW samples colleges in 9% of cases while only 3.5% of Facebook users belong to colleges. This is because the college users have on average 422 Facebook friends - much higher than the global average 144. Consequently, the college users attract RW approximately three times more often than average users.

5.2.2 *Stratification*

The advantage of S-WRW over RW does not lie exclusively in avoiding the nodes in the irrelevant category $C_\ominus$. S-WRW can also over-sample small categories (here colleges) at the cost of under-sampling large ones (which are well sampled anyway). This feature becomes important especially when the category sizes differ significantly, which is the case in Facebook. Indeed, Fig. 5(a) shows that college sizes exhibit great heterogeneity. For a fair comparison, we only include the 5,331 colleges discovered by RW. (This filtering actually gives preference to RW. S-WRW crawls discovered many more colleges that we do not show in this figure.) They span more than two orders of magnitude and follow a heavily skewed distribution.

Fig. 5(b) confirms that S-WRW successfully oversamples the small colleges. Indeed, the number of S-WRW samples per college is almost constant (roughly around 100). In contrast, the number of RW samples follows closely the college size, which results in a 100-fold difference between RW and S-WRW for smaller colleges.

5.2.3 *College size estimation*

With more samples per college, we naturally expect a better estimation accuracy under S-WRW. We demonstrate it for three colleges of different sizes (in terms of the number of Facebook users): MIT (large), Caltech (medium), and Eindhoven University of Technology (small). Each boxplot in Fig. 5(c-e) is generated based on 25 independent college size estimates $\widehat{f}_i$ that come from walks of length $n = 4K$ (left), 20K (middle), and 40K (right) samples each. For the three studied colleges, RW fails to produce reliable estimates in all cases except for MIT's (largest college) two longest crawls. Similar results hold for the overwhelming

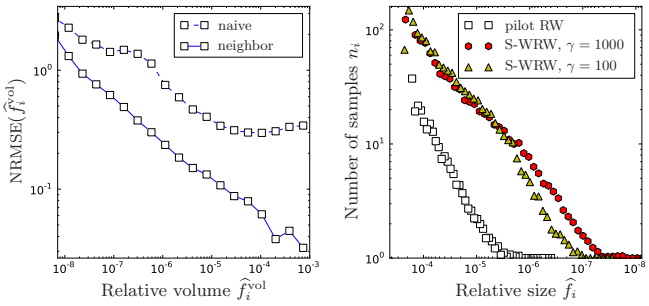


Figure 6: Facebook: Pilot RW and some S-WRW walks of length $n = 65K$. (a) The performance of the neighbor-based volume estimator Eq.(26) (plain line) and the naive one Eq.(23) (dashed line). As ground-truth we used $f_i^{\rm vol}$ calculated for all 4×1M collected samples. (b) The effect of the choice of γ.

majority of medium-sized and small colleges. The underlying reason is the very small number of samples collected by RW in these colleges, averaging at below 1 sample per walk. In contrast, the three S-WRW crawls contain typically 5-50 times more samples than RW (in agreement with Fig. 5(b)), and produce much more reliable estimates.

Finally, we aggregate the results over all colleges and compute the gain α of S-WRW over RW. We calculate the error NRMSE($\widehat{f}_i$) by taking as our "ground truth" f_i the grand average of $\widehat{f}_i$ values over all samples collected via all full-length walks and crawl types. Fig. 5(f) presents NRMSE($\widehat{f}_i$) averaged over all 5,331 colleges discovered by RW, as a function of walk length n. As expected, for all crawl types the error decreases with n. However, there is a consistently large gap between RW and all three versions of S-WRW. RW needs $\alpha = 13-15$ times more samples than S-WRW in order to achieve the same error.

5.2.4 *The effect of the choice of γ*

In the S-WRW results described above, we used the resolution $\gamma = 1000$. In order to check how sensitive the results are to the choice of this parameter, we also tried a (shorter) S-WRW run with $\gamma = 100$. In Fig. 6(b), we see that the number of samples collected in the smallest colleges is smaller under $\gamma = 100$ than under $\gamma = 1000$. In fact, the two curves diverge for colleges about 100 times smaller than the biggest college, *i.e.,* exactly at the maximal resolution $\gamma = 100$.

Both settings of γ perform orders of magnitude better than RW of the same length, which confirms the robustness of S-WRW to the choice of γ.

5.3 Summary

Only about 3.5% of 500M Facebook users are college members. There are more than 10K colleges and they greatly vary in size, ranging from 50 (or fewer) to 50K members (we consider students, alumni and staff). In this setting, state-of-the-art sampling methods such as RW (and its variants) are bound to perform poorly. Indeed, UIS (*i.e.,* an idealized version of RW) with as many as 1M samples would collect only one sample from size-500 college, on average. Even if we could sample directly from colleges only, we would typically collect fewer than 30 samples per size-500 college.

S-WRW solves these problems. We showed that S-WRW of the same length (1M) collects typically about 100 samples per size-500 college. As a result, S-WRW outperforms RW by $\alpha = 13-15$ times or $\alpha = 12-14$ times if we also consider the 6.5% overhead from the initial pilot RW. This gain can be decomposed into two factors, say $\alpha = \alpha_1 \cdot \alpha_2$. Factor $\alpha_1 \simeq 8$ can be attributed to about 8 times higher fraction of college samples in S-WRW compared to RW. Factor $\alpha_2 \simeq 1.5$ is due to over-sampling smaller networks, *i.e.,* by applying stratification to relevant categories.

Another important observation is that S-WRW is robust to the way we resolve target edge weight conflicts in Section 3.3.5. The differences between the three S-WRW implementations are minor - it is the application of Eq.(19) that brings most of the benefit.

6. RELATED WORK

Graph Sampling by Crawling. Early crawling of P2P, OSN and WWW typically used graph traversal techniques, mainly Breath-First-Search (BFS) [3,32–34,44] and its variants. However, incomplete BFS introduces bias towards high-degree nodes that is unknown and thus impossible to correct in general graphs [2,8,18,26,27].

Other studies followed a more principled approach based on random walks (RW) [4,30]. The Metropolis-Hasting RW (MHRW) [16,31] removes the bias during the walk; it has been used to sample P2P networks [36,41] and OSNs [18]. Alternatively, one can use RW, whose bias is known and can be corrected for [21,40], thus leading to a re-weighted RW [18,36]. RW was also used to sample Web [22], P2P networks [19,36,41], OSNs [18,24,34,37], and other large graphs [28]. It was empirically shown in [18,36] that RW outperforms MHRW in real-life topologies. RW has also been used to sample *dynamic graphs* [36,41,43], which are outside the scope of this paper.

Fast Mixing Markov Chains. The mixing time of RW in many OSNs was found larger than commonly believed [34]. There exist many approaches that try to minimize the mixing time of random walks, such as multiple dependent random walks [38], multigraph sampling [17], or the addition of random jumps [5,28,39]. Given the knowledge of the entire graph, [10] proposes an optimal solution by explicitly minimizing the second largest eigenvalue modulus (SLEM) of the transition probability matrix.

All the above methods try to minimize mixing time towards a given target stationary distribution, (*e.g.,* treating all nodes with equal importance). Therefore, they are complementary to our technique that primarily aims at finding the right distribution for a given category-related measurement objective, while also maintaining fast mixing.

Stratified Sampling. Our approach builds on *stratified sampling* [35], a widely used technique in statistics; see [12, 29] for a good introduction. A related work in a different networking problem is [14], where threshold sampling is used to vary sampling probabilities of network traffic flows and estimate their volume.

Weighted Random Walks for Sampling. Random walks on graphs with weighted edges [4,30], are well studied and heavily used in Monte Carlo Markov Chain simulations [16] to sample a state space with a specified probability distribution. However, to the best of our knowledge, WRWs had not been used for measurements of real online systems with a goal other than improving mixing (discussed above).

Recent applications of WRW in online social networks

include [6,7]. In both these papers, the goal is to predict/extract something from a known graph. In contrast, we use WRW to estimate features of an unknown graph.

In the context of World Wide Web crawling, *focused crawling* techniques [11,13] have been introduced to follow web pages of specified interest and to avoid the irrelevant pages. This is achieved by performing a BFS type of sampling, except that instead of FIFO queue they use a priority queue weighted by the page relevancy. In our context, such an approach suffers from the same problems as regular BFS: (i) collected samples strongly depend on the starting point, and (ii) we are not able to analytically correct for the bias.

7. CONCLUSION

We have introduced Stratified Weighted Random Walk (S-WRW) - an efficient heuristic for sampling large, static, undirected graphs via crawling and using minimal information about node categories. S-WRW performs a random walk on a graph whose edge weights are set taking into account the estimation objective. We apply S-WRW to measure the Facebook social graph, and we show that it brings a very significant gain.

In future work, we plan to combine S-WRW with existing orthogonal techniques, some of which have been reviewed in the related work, to further improve performance.

8. REFERENCES

[1] Weighted Random Walks of the Facebook social graph: http://odysseas.calit2.uci.edu/osn, 2011.

[2] D. Achlioptas, A. Clauset, D. Kempe, and C. On the bias of traceroute sampling: or, power-law degree distributions in regular graphs. *Journal of the ACM*, 2009.

[3] Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of topological characteristics of huge online social networking services. In *WWW*, pages 835–844, 2007.

[4] D. Aldous and J. A. Fill. *Reversible Markov Chains and Random Walks on Graphs*. In preparation.

[5] K. Avrachenkov, B. Ribeiro, and D. Towsley. Improving Random Walk Estimation Accuracy with Uniform Restarts. In *I7th Workshop on Algorithms and Models for the Web Graph*, 2010.

[6] L. Backstrom and J. Kleinberg. Network Bucket Testing. In *WWW*, 2011.

[7] L. Backstrom and J. Leskovec. Supervised Random Walks: Predicting and Recommending Links in Social Networks. In *ACM International Conference on Web Search and Data Minig (WSDM)*, 2011.

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

[9] H. R. Bernard, T. Hallett, A. Iovita, E. C. Johnsen, R. Lyerla, C. McCarty, M. Mahy, M. J. Salganik, T. Saliuk, O. Scutelniciuc, G. a. Shelley, P. Sirinirund, S. Weir, and D. F. Stroup. Counting hard-to-count populations: the network scale-up method for public health. *Sexually Transmitted Infections*, 86(Suppl 2):ii11–ii15, Nov. 2010.

[10] S. Boyd, P. Diaconis, and L. Xiao. Fastest mixing Markov chain on a graph. *SIAM review*, 46(4):667–689, 2004.

[11] S. Chakrabarti. Focused crawling: a new approach to topic-specific Web resource discovery. *Computer Networks*, 31(11-16):1623–1640, May 1999.

[12] W. G. Cochran. *Sampling Techniques*, volume 20 of *McGraw-Hil Series in Probability and Statistics*. Wiley, 1977.

[13] M. Diligenti, F. Coetzee, S. Lawrence, C. Giles, and M. Gori. Focused crawling using context graphs. In *Proceedings of the 26th International Conference on Very Large Data Bases*, pages 527–534, 2000.

[14] N. Duffield, C. Lund, and M. Thorup. Learn more, sample less: control of volume and variance in network measurement. *IEEE Transactions on Information Theory*, 51(5):1756–1775, May 2005.

[15] J. Gentle. *Random number generation and Monte Carlo methods*. Springer Verlag, 2003.

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

[17] M. Gjoka, C. Butts, M. Kurant, and A. Markopoulou. Multigraph Sampling of Online Social Networks. *arXiv*, (arXiv:1008.2565v1):1–10, 2010.

[18] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou. Walking in Facebook: A Case Study of Unbiased Sampling of OSNs. In *INFOCOM*, 2010.

[19] C. Gkantsidis, M. Mihail, and A. Saberi. Random walks in peer-to-peer networks. In *INFOCOM*, 2004.

[20] M. Hansen and W. Hurwitz. On the Theory of Sampling from Finite Populations. *Annals of Mathematical Statistics*, 14(3), 1943.

[21] D. D. Heckathorn. Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations. *Social Problems*, 44:174–199, 1997.

[22] M. R. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork. On near-uniform URL sampling. In *WWW*, 2000.

[23] E. D. Kolaczyk. *Statistical Analysis of Network Data*, volume 69 of *Springer Series in Statistics*. Springer New York, 2009.

[24] B. Krishnamurthy, P. Gill, and M. Arlitt. A few chirps about Twitter. In *WOSN*, 2008.

[25] M. Kurant, M. Gjoka, C. Butts, and A. Markopoulou. Walking on a Graph with a Magnifying Glass. *Arxiv preprint arXiv:1101.5463*, 2011.

[26] M. Kurant, A. Markopoulou, and P. Thiran. On the bias of BFS (Breadth First Search). In *ITC, also in arXiv:1004.1729*, 2010.

[27] S. H. Lee, P.-J. Kim, and H. Jeong. Statistical properties of Sampled Networks. *Phys. Rev. E*, 73:16102, 2006.

[28] J. Leskovec and C. Faloutsos. Sampling from large graphs. In *KDD*, pages 631–636, 2006.

[29] S. Lohr. *Sampling: design and analysis*. Brooks/Cole, second edition, 2009.

[30] L. Lovász. Random walks on graphs: A survey. *Combinatorics, Paul Erdos is Eighty*, 2(1):1–46, 1993.

[31] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculation by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

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

[33] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *IMC*, pages 29–42, 2007.

[34] A. Mohaisen, A. Yun, and Y. Kim. Measuring the mixing time of social graphs. *IMC*, 2010.

[35] J. Neyman. On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, 97(4):558, 1934.

[36] A. Rasti, M. Torkjazi, R. Rejaie, N. Duffield, W. Willinger, and D. Stutzbach. Respondent-driven sampling for characterizing unstructured overlays. In *Infocom Mini-conference*, pages 2701–2705, 2009.

[37] A. H. Rasti, M. Torkjazi, R. Rejaie, and D. Stutzbach. Evaluating Sampling Techniques for Large Dynamic Graphs. In *Technical Report*, volume 1, 2008.

[38] B. Ribeiro and D. Towsley. Estimating and sampling

graphs with multidimensional random walks. In *IMC*, volume 011, 2010.
[39] B. Ribeiro, P. Wang, and D. Towsley. On Estimating Degree Distributions of Directed Graphs through Sampling. *UMass Technical Report*, 2010.
[40] M. Salganik and D. D. Heckathorn. Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology*, 34(1):193–240, 2004.
[41] D. Stutzbach, R. Rejaie, N. Duffield, S. Sen, and W. Willinger. On unbiased sampling for unstructured peer-to-peer networks. In *IMC*, 2006.
[42] E. Volz and D. D. Heckathorn. Probability based estimation theory for respondent driven sampling. *Journal of Official Statistics*, 24(1):79–97, 2008.
[43] W. Willinger, R. Rejaie, M. Torkjazi, M. Valafar, and M. Maggioni. OSN Research: Time to Face the Real Challenges. In *HotMetrics*, 2009.
[44] C. Wilson, B. Boe, A. Sala, K. P. N. Puttaswamy, and B. Y. Zhao. User interactions in social networks and their implications. In *EuroSys*, 2009.

Appendix A: Achieving Arbitrary Node Weights

Achieving arbitrary node weights by setting the edge weights in a graph $G=(V,E)$ is sometimes impossible. For example, for a graph that is a path consisting of two nodes (v_1-v_2), it is impossible to achieve $\mathrm{w}(v_1)\neq\mathrm{w}(v_2)$. However, it is always possible to do so, if there are self loops in each node.

OBSERVATION 1. *For any undirected graph $G=(V,E)$ with a self-loop $\{v,v\}$ at every node $v\in V$, we can achieve an arbitrary distribution of node weights $w(v)>0,\ v\in V$, by appropriate choice of edge weights $w(u,v)>0,\ \{u,v\}\in E$.*

PROOF. Let $\mathrm{w}_{\min}$ be the smallest of all target node weights $\mathrm{w}(v)$. Set $\mathrm{w}(u,v)=\mathrm{w}_{\min}/N$ for all non self-loop edges (i.e., where $u\neq v$). Now, for every self-loop $\{v,v\}\in E$ set

$$\mathrm{w}(v,v) \;=\; \frac{1}{2}\left(\mathrm{w}(v)-\frac{\mathrm{w}_{\min}}{N}\cdot(\deg(v)-2)\right).$$

It is easy to check that, because there are exactly $\deg(v)-2$ non self-loop edges incident on v, every node $v\in V$ will achieve the target weight $\mathrm{w}(v)$. Moreover, the definition of $\mathrm{w}_{\min}$ guarantees that $\mathrm{w}(v,v)>0$ for every $v\in V$. □

Appendix B: Estimating Category Volumes

In this section, we derive efficient estimators of the relative volume $\widehat{f}_C^{\mathrm{vol}}=\frac{\mathrm{vol}(C)}{\mathrm{vol}(V)}$. Recall that $S\subset V$ denotes an independent sample of nodes in G, with replacement.

Node sampling
If S is a uniform sample UIS, then we can write

$$\widehat{f}_C^{\mathrm{vol}} \;=\; \frac{\sum_{v\in S}\deg(v)\cdot 1_{\{v\in C\}}}{\sum_{v\in S}\deg(v)}, \qquad (21)$$

which is a straightforward application of the classic ratio estimator [29].

In the more general case, when S is selected using WIS, then we have to correct for the linear bias towards nodes of higher weights w(), as follows:

$$\widehat{f}_C^{\mathrm{vol}} \;=\; \frac{\sum_{v\in S}\deg(v)\cdot 1_{\{v\in C\}}/\mathrm{w}(v)}{\sum_{v\in S}\deg(v)/\mathrm{w}(v)}. \qquad (22)$$

In particular, if $\mathrm{w}(v)\sim\deg(v)$, then

$$\widehat{f}_C^{\mathrm{vol}} \;=\; \frac{1}{n}\cdot\sum_{v\in S}1_{\{v\in C\}}. \qquad (23)$$

Star sampling
Another approach is to focus on the set of all neighbors $\mathcal{N}(S)$ of sampled nodes (with repetitions) rather than on S itself, *i.e.,* to use 'star sampling' [23]. The probability that a node v is a neighbor of a node sampled from V by UIS is

$$\sum_{u\in V}\frac{1}{N}\cdot 1_{\{v\in\mathcal{N}(u)\}} \;=\; \frac{\deg(v)}{N}.$$

Consequently, the nodes in $\mathcal{N}(S)$ are asymptotically equivalent to nodes drawn with probabilities linearly proportional to node degrees. By applying Eq.(23) to $\mathcal{N}(S)$, we obtain[7]

$$\widehat{f}_C^{\mathrm{vol}} \;=\; \frac{1}{\mathrm{vol}(S)}\sum_{u\in S}\sum_{v\in\mathcal{N}(u)}1_{\{v\in C\}}, \qquad (24)$$

where we used $|\mathcal{N}(S)|=\sum_{u\in S}\deg(u)=\mathrm{vol}(S)$.

In the more general case, when S is selected using WIS, then we correct for the linear bias towards nodes of higher weights w(), as follows:

$$\widehat{f}_C^{\mathrm{vol}} \;=\; \frac{1}{\sum_{u\in S}\frac{\deg(u)}{\mathrm{w}(u)}}\sum_{u\in S}\left(\frac{1}{\mathrm{w}(u)}\sum_{v\in\mathcal{N}(u)}1_{\{v\in C\}}\right). \qquad (25)$$

In particular, if $\mathrm{w}(v)\sim\deg(v)$, then

$$\widehat{f}_C^{\mathrm{vol}} \;=\; \frac{1}{n}\sum_{u\in S}\left(\frac{1}{\deg(u)}\sum_{v\in\mathcal{N}(u)}1_{\{v\in C\}}\right). \qquad (26)$$

Note that for every sampled node $v\in S$, the formulas Eq.(24-26) exploit all the $\deg(v)$ neighbors of v, whereas Eq.(21-23) rely on one node per sample only. Not surprisingly, Eq.(24-26) performed much better in all our simulations and implementations.

Appendix C: Stratification in Expectation

In this section, we show the correctness of Eq.(13) in Section 3.1. Recall from Eq.(5) that under WIS, at every iteration, the probability $\pi^{\mathrm{WIS}}(v)$ of sampling node v is proportional to its weight $\mathrm{w}^{\mathrm{WIS}}(v)$. So the probability $\pi^{\mathrm{WIS}}(C_i)$ of sampling a node from category C_i is proportional to the weight $\mathrm{w}^{\mathrm{WIS}}(C_i)$ of C_i, *i.e.,*

$$\pi^{\mathrm{WIS}}(C_i) \;\propto\; \mathrm{w}^{\mathrm{WIS}}(C_i).$$

This, together with Eq.(13), imply

$$\pi^{\mathrm{WIS}}(C_i) \;\propto\; n_i^{\mathrm{opt}}.$$

We can now use $\sum_i\pi^{\mathrm{WIS}}(C_i)=1$ and $\sum_i n_i^{\mathrm{opt}}=n$ to rewrite the above formula as the following equation

$$\pi^{\mathrm{WIS}}(C_i)=n_i^{\mathrm{opt}}/n.$$

Consequently, under WIS

$$\mathbb{E}[n_i] \;=\; \mathrm{Binom}(n,\pi^{\mathrm{WIS}}(C_i))=n\cdot\pi^{\mathrm{WIS}}(C_i)=n_i^{\mathrm{opt}}.$$

[7] As a side note, observe that Eq.(24) generalizes the "scale-up method" [9] used in social sciences to estimate the size (here $|C|$) of hidden populations (*e.g.,* of drug addicts). Indeed, if we assume that the average node degree in V is the same as in C, then $f_C^{\mathrm{vol}}=\mathrm{vol}(C)/\mathrm{vol}(V)=|C|/|V|$, which reduces Eq.(23) to the core formula of the scale-up method.