Towards Unbiased BFS Sampling

Maciej Kurant, Athina Markopoulou, *IEEE Member*, and Patrick Thiran, *IEEE Member*

***Abstract*—Breadth First Search (BFS) is a widely used approach for sampling large graphs. However, it has been empirically observed that BFS sampling is biased toward high-degree nodes, which may strongly affect the measurement results. In this paper, we quantify and correct the degree bias of BFS.**

First, we consider a random graph $RG(p_k)$ with an arbitrary degree distribution p_k. For this model, we calculate the node degree distribution expected to be observed by BFS as a function of the fraction f of covered nodes. We also show that, for $RG(p_k)$, all commonly used graph traversal techniques (BFS, DFS, Forest Fire, Snowball Sampling, RDS) have exactly the same bias. Next, we propose a practical BFS-bias correction procedure that takes as input a collected BFS sample together with the fraction f. Our correction technique is exact (*i.e.*, leads to unbiased estimation) for $RG(p_k)$. Furthermore, it performs well when applied to a broad range of Internet topologies and to two large BFS samples of Facebook and Orkut networks.

***Index Terms*—Breadth-First-Search (BFS), network topology, sampling methods, bias, estimation, online social networks.**

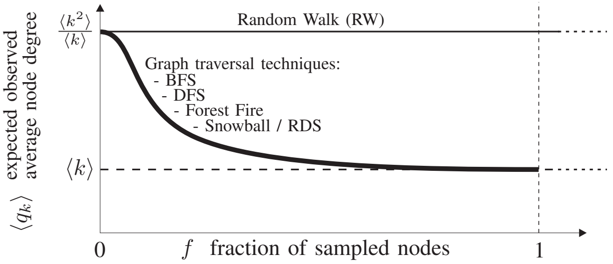

Fig. 1. **Overview of results.** In this paper, we calculate the node degree distribution q_k expected to be observed by BFS in a random graph $RG(p_k)$ with a given degree distribution p_k, as a function of the fraction of sampled nodes f. In this plot, we show only the average $\langle q_k\rangle$. We show the RW as a reference. $\langle k\rangle = \langle p_k\rangle$ is the real average node degree, and $\langle k^2\rangle$ is the real average squared node degree. *Observations:* (i) For a small sample size, BFS has the same bias as RW; with increasing f, the bias decreases; a complete BFS ($f = 1$) is unbiased. (ii) All graph traversal techniques (that use sampling without replacement) lead to the same bias in $RG(p_k)$. (iii) The shape of the BFS curve depends on the graph (the real node degree distribution p_k), but it is always monotonically decreasing; we calculate it precisely in this paper. (iv) We also correct for the bias and compute the original distribution p_k based on the sampled q_k and f (not shown here).

I. Introduction

A LARGE body of work in the networking community focuses on Internet topology measurements at various levels, including the IP or AS connectivity, the Web (WWW), peer-to-peer (P2P) and online social networks (OSN). The size of these networks and other restrictions often make measuring the entire graph impossible. For example, learning only the topology of Facebook social graph would require downloading more than 115TB of HTML data [2], which is impractical. Instead, researchers typically collect and study a small but representative sample of the graph of interest.

In this paper, we are particularly interested in sampling networks that naturally allow to explore the neighbors of a given node, as it is the case in WWW, P2P and OSN. A number of graph exploration techniques use this basic operation for sampling. They can be roughly classified in two categories: (i) random walks, and (ii) graph traversals.

In the first category, *random walks*, nodes can be revisited. This category includes the classic Random Walk (RW) [3] and its variations [4]–[10]. They are used for sampling of nodes on the Web [4], P2P networks [5]–[7], OSNs [11,12] and large graphs in general [13]. Although most random walks introduce a bias towards high-degree nodes [3], it can be easily corrected for [14]–[17]. In this paper, we use RW as baseline for comparison only; for more details on random walks please refer to our companion paper [2] in this issue.

In the second category, which we refer to as *graph traversals*, sampling is without replacement: each node is visited at most once (or exactly once, when the process runs until completion and the graph is connected). These methods vary in the order in which they visit the nodes; examples include BFS, Depth-First Search (DFS), Forest Fire (FF) [13], Snowball Sampling (SBS) [18] and Respondent-Driven Sampling (RDS) [19]. Graph traversals, especially BFS, are very popular and widely used for sampling Internet topologies, *e.g.*, in WWW [20] or OSNs [21]–[23]. Reasons for its popularity include (i) its simplicity and efficiency and (ii) the fact that a BFS sample reveals the topology (all the nodes and edges) around the starting point. This allows to characterize the topological characteristics (*e.g.*, shortest path lengths, clustering coefficients, community structure) in that part of the graph, which is an advantage of BFS over random walks.

However, a BFS sample may or may not be representative of the entire graph. For example, a BFS sample of a lattice is a lattice. Unfortunately, this is not true in general. It has been observed empirically that BFS introduces a bias towards high-degree nodes [20,24]–[26]. We also confirmed this fact in a recent measurement of Facebook [2,11], where our BFS crawler found the average node degree equal to 324, while the real value is only 94; in other words BFS overestimated the average node degree by about 250%. This bias clearly affects the inferences based on a BFS sample, whether node attributes

Manuscript received 15 December 2010; revised 2 June 2011. This is a revised and extended version of [1]. This work has been supported by the following grants: SNF grant PBELP2-130871, Switzerland; ManCom 2110 of the Hasler Foundation, Switzerland; NSF CDI award 1028394, AFOSR FA9550-10-1-0310, AFOSR MURI Prime Award FA9550- 09-0643.

M. Kurant and A. Markopoulou are with the California Institute for Telecommunications and Information Technology (CalIT2) at the University of California, Irvine, CA 92697-2800 (e-mail: maciej.kurant@gmail.com, athina@uci.edu). Athina Markopoulou is also with the EECS Dept and the Center for Pervasive Communications and Computing (CPCC) at UC Irvine.

P. Thiran is with the School of Computer & Comm. Sciences, EPFL, Lausanne, Switzerland (e-mail: patrick.thiran@epfl.ch).

Digital Object Identifier 10.1109/JSAC.2011.111005.

or topological properties are of interest. Despite the popularity of BFS on the one hand, and its bias on the other hand, we still know relatively little about the statistical properties of node sequences returned by BFS. The analysis is challenging because BFS, and more generally sampling without replacement, introduces complex dependencies between the sampled nodes, which make it difficult to deal with mathematically.

Our work is a step towards understanding and correcting the bias of BFS sampling. We make the following main contributions. First, we consider a random graph $RG(p_k)$ with a given (and arbitrary) degree distribution p_k. We calculate precisely the node degree distribution q_k expected to be observed by BFS as a function of the fraction f of sampled nodes. We illustrate this and related results in Fig. 1. In our analysis, we use arguments following the lines of those used by Achlioptas et al. [27] to analyze the bias of traceroute sampling. However, we analyze a different sampling design: in traceroute sampling all nodes are visited and some edges are missing, whereas in BFS sampling some nodes are sampled but all edges incident to them are seen.

Second, we propose a practical BFS-bias correction procedure. It takes as input a collected BFS sample together with the fraction f of covered nodes, and estimates the distribution of an arbitrary function $x(v)$ defined on graph nodes. The correction procedure is exact (*i.e.*, leads to unbiased estimation) for $RG(p_k)$ graphs, and also turns out to be a good heuristic when applied to a broad range of Internet topologies, as well as to two large BFS samples of Facebook and Orkut networks. We make its ready-to-use `python` implementation publicly available at [28].

The outline of the rest of paper is as follows. Section II discusses related work. Section III presents BFS and other graph traversal algorithms under study. Section IV presents the random graph $RG(p_k)$ model used in this paper. Section V analyzes the degree bias of BFS. Section VI shows how to correct for this bias in $RG(p_k)$. Section VII provides simulation results and evaluation in real world networks. Section VIII gives some practical recommendations, and Section IX concludes the paper.

II. Related Work

A. BFS used in practice

BFS is widely used today for exploring large networks, such as OSNs. In [21], Ahn et al. used BFS to sample Orkut and MySpace. In [22] and [29], Mislove et al. used BFS to crawl the social graph in four popular OSNs: Flickr, LiveJournal, Orkut, and YouTube. In [23], Wilson et al. measured the social graph and the user interaction graph of Facebook using several BFSs, each BFS constrained in one of the largest 22 regional Facebook networks. In our recent work [2,11], we have also crawled Facebook using various sampling techniques, including BFS.

B. BFS bias

It has been empirically observed that incomplete BFS and its variants introduce bias towards high-degree nodes [20] [24]–[26]. We confirmed this in Facebook [2,11], which, in fact, inspired and motivated this paper. Analogous bias has been observed in the field of social science, for sampling techniques closely related to BFS, *i.e.*, Snowball Sampling and RDS [16,18,19] (see Section III-B5).

C. Analyzing BFS

To the best of our knowledge, the sampling bias of BFS has not been analyzed so far. [30] and [27] are the closest related papers in terms of methodology. The original paper by Kim [30] analyzes the size of the largest connected component in classic Erdös-Rényi random graph by essentially applying the configuration model [31] with node degrees chosen from a Poisson distribution. To match the stubs (or "clones" in [30]) uniformly at random in a tractable way, Kim proposes a "cut-off line" algorithm. He first assigns each stub a random index from $[0, np]$, and next progressively scans this interval. Achlioptas et al. used this powerful idea in [27] to study the bias of traceroute sampling in random graphs $RG(p_k)$ with a given degree distribution p_k. The basic operation in [27] is traceroute (*i.e.*, "discover a path") and is performed from a single node to all other nodes in the graph. The union of the observed paths forms a "BFS-tree", which includes all nodes but misses some edges (*e.g.*, those between nodes at the same depth in the tree). In contrast, the basic operation in the traversal methods presented in our paper is to discover all neighbors of a node, and it is applied to all nodes in increasing distance from the origin. Another important difference is that [27] studies a completed BFS-tree, whereas we study the sampling process when it has visited only a fraction $f < 1$ of nodes. Indeed, a completed BFS ($f{=}1$) is trivial in our case: it has no bias, as all nodes are covered.

In the field of social science, a significant effort was put to correct for the bias of BFS's close cousin - Snowball Sampling (SBS) [18]. SBS together with a bias correction procedure is called Respondent-Driven Sampling (RDS) [19]. The currently used correction technique [16,17] assumes that nodes can be revisited, which essentially approximates SBS by Random Walk. In this paper, we formally show that this approximation is valid if the fraction f of sampled nodes is small. However, as [32] points out, the current RDS methodology is systematically biased for larger f. Consequently, [33] proposed an SBS bias correction method based on the random graph $RG(p_k)$. This is essentially the same basic starting idea as used in our original paper published independently [1]. However, the two papers fundamentally differ in the final solution: [33] proposes a simulation-aided approach, whereas we solve the problem analytically.

Another recent and related paper is [34]. The authors propose and evaluate a heuristic approach to correct the degree bias in the ith generation of SBS, based on the values measured in the generation $i-1$. In practice, this generation-based scheme may be challenging to implement, because the number of nodes per generation may grow close to exponential with i. Consequently, we are likely to face a situation where collecting the next generation is prohibitively expensive, while the current generation has much fewer nodes than our sampling capabilities allow for.

D. Probability Proportional to Size Without Replacement

At a closer look, our $RG(p_k)$-based approach reduces BFS (and other graph traversals) to a classic sampling design called Probability Proportional to Size Without Replacement (PPSWOR) [35]–[42]. Unfortunately, to the best of our knowledge, none of the existing results is directly applicable to our problem. This is because, speaking in the terms used later in this paper, the available results either (i) require the knowledge of $q_k(f)$ (expected, not sampled) as an input, (ii) propose how to calculate $q_k(f)$ for the first two nodes only, or (iii) calculate $q_k(f)$ as an average of many simulated traversals of the known graph (in contrast, we only have one run on unknown graph) [42].

III. Graph exploration techniques

Let $G = (V, E)$ be a connected graph with the set of vertices V, and a set of undirected edges E. Initially, G is unknown, except for one (or few) seed node(s). When sampling through graph exploration, we begin at the seed node, and we recursively visit (one, some or all) its neighbors. We distinguish two main categories of exploration techniques: random walks and graph traversals.

A. Random walks (baseline)

Random walks allow revisiting the same node many times. They come in many flavors [3]–[10]; see our companion paper [2] in this issue for more details. Because in this paper we use random walks merely as a useful reference, we include only the classic simple *Random Walk (RW)* [3]. RW selects the next-hop node uniformly at random among the neighbors of the current node.

B. Graph traversals

In contrast, graph traversals never revisits the same node. At the end of the process, and assuming that the graph is connected, all nodes are visited. However, when using graph traversals for sampling, we terminate after having collected a fraction $f < 1$ (usually $f \ll 1$) of graph nodes.

1) Breadth First Search (BFS): BFS is a classic graph traversal algorithm that starts from the seed and progressively explores all neighbors. At each new iteration the earliest explored but not-yet-visited node is selected next. Consequently, BFS discovers first the nodes closest to the seed.

2) Depth First Search (DFS): This technique is similar to BFS, except that at each iteration we select the latest explored but not-yet-visited node. As a result, DFS explores first the nodes that are faraway (in the number of hops) from the seed.

3) Forest Fire (FF): FF is a randomized version of BFS, where for every neighbor v of the current node, we flip a coin, with probability of success p, to decide if we explore v. FF reduces to BFS for $p=1$. It is possible that this process dies out before it covers all nodes. In this case, in order to make FF comparable with other techniques, we revive the process from a random node already in the sample. Forest Fire is inspired by the graph growing model of the same name proposed in [43] and is used as a graph sampling technique in [13].

TABLE I
NOTATION SUMMARY.

$G = (V, E)$	graph G with nodes V and edges E
k_v	degree of node v
$p_k \ = \frac{1}{\lvert V\rvert} \sum_{v \in V} 1_{k_v = k}$	degree distribution in G
$\langle k \rangle \ = \ \langle p_k \rangle \ = \sum_k k\, p_k$	average node degree in G
q_k	expected sampled degree distribution
$\langle q_k \rangle \ = \sum_k k\, q_k$	expected sampled average node degree
$\widehat{q}_k$	sampled degree distribution
$\widehat{p}_k$	estimated original degree distribution in G
f	fraction of nodes covered by the sample

4) Snowball Sampling (SBS): According to a classic definition by Goodman [18], an n-name Snowball Sampling is similar to BFS, but at every node v, not all k_v, but exactly n neighbors are chosen randomly out of all k_v neighbors of v. These n neighbors are scheduled to visit, but only if they have not been visited before.

5) Respondent-Driven Sampling (RDS): Respondent-Driven Sampling (RDS) [16,17,19] adopts SBS to penetrate hidden populations (such as that of drug addicts) in social surveys. RDS is essentially SBS equipped with some bias correction procedure (omitted in Fig. 1). In Section II, we comment on these techniques.

IV. Graph model $RG(p_k)$

A basic, yet very important property of every graph is its node degree distribution p_k, *i.e.*, the fraction of nodes with degree equal to k, for all $k \geq 0$.[1] Depending on the network, the degree distribution can vary, ranging from constant-degree (in regular graphs), a distribution concentrated around the average value (*e.g.*, in Erdös-Rényi random graphs or in well-balanced P2P networks), to heavily right-skewed distributions with k covering several decades (as this is the case in WWW, unstructured P2P, Internet at the IP and Autonomous System level, OSNs). We handle all these cases by assuming that we are given *any* fixed node degree distribution p_k. Other than that, the graph G is drawn uniformly at random from the set of all graphs multigraphs[2] with degree distribution p_k. We denote this model by $RG(p_k)$.

Because $RG(p_k)$ mimics an arbitrary node degree distribution p_k, it can be considered a "first-order approximation" of real-life graphs. Of course, there are many graph properties other than p_k that are not captured by $RG(p_k)$. However, we show later that, with respect to the BFS sampling bias, $RG(p_k)$ approximates the real Internet topologies surprisingly well.

We use a classic technique to generate $RG(p_k)$, called the *configuration model* [31]: each node v is given k_v "stubs" or "edges-to-be". Next, all these $\sum_{v \in V} k_v \ = \ 2|E|$ stubs are randomly matched in pairs, until all stubs are exhausted (and $|E|$ edges are created). In Fig. 2 (ignore the rectangular interval [0,1] for now), we present four nodes with their stubs (left) and an example of their random matching (right).

[1] As we define p_k as a 'fraction', not the 'probability', p_k determines the degree sequence in the graph, and vice versa.

[2] A multigraph is a graph that accepts multiple edges and self-loops. An alternative approach is to connect nodes v and w independently with probability $\frac{k_w \cdot k_v}{z}$ [44]. By construction, this procedure avoids multiple edges and self-loops. However, it achieves the desired degree sequence not in every realization, but in expectation only. More importantly, it imposes a limit $k_v \leq \sqrt{z}$ on node degrees.

V. Analyzing the Node Degree Bias

In this section, we study the node degree bias observed when the graph exploration techniques of Section III are run on the random graph $RG(p_k)$ of Section IV. In particular, we are interested in the node degree distribution q_k expected to be observed in the raw sample. Typically, the observed distribution is different from the original one, $q_k \neq p_k$, with higher average value $\langle q_k \rangle > \langle p_k \rangle$ (*i.e.*, average sampled and observed node degree, respectively). Below, we derive q_k as a function of p_k and, in the case of BFS, of the fraction of sampled nodes f.

A. Random walks (baseline)

Under RW, in any given connected and aperiodic graph, the probability of being at a particular node v converges at equilibrium to the stationary distribution $\pi_v^{\text{RW}}{=}\frac{k_v}{2|E|}$. (*I.e.*, the sampling probability of a node v is proportional to its degree k_v.) Therefore, the expected observed degree distribution q_k^{RW} is (after [14]–[17])

$$q_k^{\text{RW}} \;=\; \sum_v \pi_v^{\text{RW}} \cdot 1_{\{k_v=k\}} \;=\frac{k}{2|E|}\,p_k\,|V| \;=\; \frac{k\,p_k}{\langle k \rangle}, \qquad (1)$$

where $\langle k \rangle$ is the average node degree in G. Consequently, the expected observed average node degree is

$$\langle q_k^{\text{RW}} \rangle \;=\; \sum_k k\, q_k^{\text{RW}} \;=\; \frac{\sum_k k^2\, p_k}{\langle k \rangle} \;=\; \frac{\langle k^2 \rangle}{\langle k \rangle}, \qquad (2)$$

where $\langle k^2 \rangle$ is the average squared node degree in G. We show this value $\frac{\langle k^2 \rangle}{\langle k \rangle}$ in Fig. 1.

B. Graph traversals (Main Result)

In RW, nodes can be revisited. So the state of the system at iteration $i+1$ depends only on iteration i, which makes it possible to analyze with Markov Chain techniques. In contrast, graph traversals do not allow for node revisits, which introduces crucial dependencies between all the iterations and significantly complicates the analysis. To handle these dependencies, we adopt an elegant technique recently introduced in [30] (to study the size of the largest connected component) and extended in [27] (to study the bias of traceroute sampling). We develop our arguments following the lines of [27]. However, this work differs in many aspects from both [30] and [27], on which we comment in detail in Section II.

1) Exploration without replacement at the stub level: We begin by defining Algorithm 1 (below) - a general graph traversal technique that collects a sequence of nodes S, without replacements. To be compatible with the configuration model (see Section IV), we are interested in the process *at the stub level*, where we consider one stub at a time, rather than one node at a time. An integral part of the algorithm is a queue Q that keeps the discovered, but still not-yet-followed stubs. First, we enqueue on Q all the stubs of some initial node v_1, and by setting $S \leftarrow [v_1]$. Next, at every iteration, we dequeue one stub from Q, call it a, and follow it to discover its partner-stub b, and b's owner $v(b)$. If node $v(b)$ is not yet discovered, *i.e.*, if $v(b) \notin S$, then we append $v(b)$ to S and we enqueue on Q all the other stubs of $v(b)$.

Algorithm 1 Stub-Level Graph Traversal

```
1: $S \leftarrow [v_1]$  and  $Q \leftarrow$ [all stubs of $v_1$]
2: while $Q$ is nonempty do
3:    Dequeue $a$ from $Q$
4:    Discover $a$'s partner $b$
5:    if $v(b) \notin S$ then
6:       Append $v(b)$ to $S$
7:       Enqueue on $Q$ all stubs of $v(b)$ except $b$
8:    else
9:       Remove $b$ from $Q$
10:   end if
11: end while
```

Depending on the scheduling discipline for the elements in Q (line 3), Algorithm 1 implements BFS (for a first-in first out scheduling), DFS (last-in first-out) or Forest Fire (first-in first-out with randomized stub losses). Line 9 guarantees that the algorithm never tracebacks the edges, *i.e.*, that stub a dequeued from Q in line 3 never belongs to an edge that has already been traversed in the opposite direction.

2) Discovery on-the-fly: In line 4 of Algorithm 1, we follow stub a to discover its partner b. In a fixed graph G, this step is deterministic. In the configuration model $RG(p_k)$, a fixed graph G is obtained by matching all the stubs uniformly at random. Next, we can sample this fixed graph and average the result over the space of all the random graphs $RG(p_k)$ that have just been constructed. Unfortunately, this space grows exponentially with the number of nodes $|V|$, making the problem untractable. Therefore, we adopt an alternative construction of G - by iteratively selecting b on-the-fly (*i.e.*, every time line 4 is executed), uniformly at random from all still unmatched stubs. By the principle of deferred decisions [45], these two approaches are equivalent.

With the help of the on-the-fly approach, we are able to write down the equations we need. Indeed, let us denote by $X_i \in V$ the ith selected node, and let $\mathbb{P}(X_1 = u)$ be the probability that node $u \in V$ is chosen as a starting node. It is easy to show that with $z{=}2|E|$ we have

$$\mathbb{P}(X_2{=}v) \;=\; \sum_{u \neq v} \frac{k_v}{z - k_u} \cdot \mathbb{P}(X_1{=}u) \qquad (3)$$

$$\mathbb{P}(X_3{=}w) \;=\; \sum_{v \neq w} \sum_{u \neq w,v} \frac{k_w}{z - k_v - k_u} \cdot \frac{k_v}{z - k_u} \cdot \mathbb{P}(X_1{=}u), \qquad (4)$$

and so on. Theoretically, these equations allow us to calculate the expected node degree at any iteration, and thus the degree bias of BFS.

3) Breaking the dependencies: There is still one problem with the equations above. Due to the increasing number of nested sums, the results can be calculated in practice for a first few iterations only. This is because we select stub b uniformly and independently at random from all the *unmatched* stubs. So the stub selected at iteration i depends on the stubs selected at iterations $1 \ldots i-1$, which results in the nested sums. We remedy this problem by implementing the on-the-fly approach

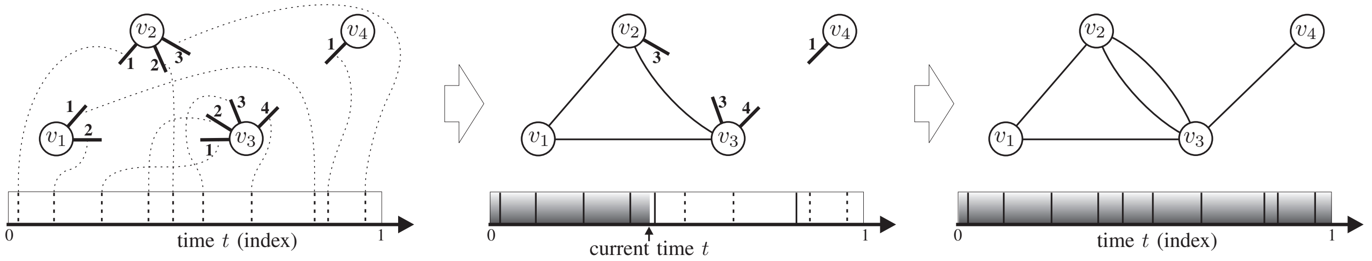

Fig. 2. An illustration of the stub-level, on-the-fly graph exploration without replacements. In this particular example, we show an execution of BFS starting at node v_1. **Left:** Initially, each node v has k_v stubs, where k_v is a given target degree of v. Each of these stubs is assigned a real-valued number drawn uniformly at random from the interval $[0,1]$ shown below the graph (these numbers remain unchanged throughout the entire process). Next, we follow Algorithm 1 with a starting node v_1. The numbers next to the stubs of every node v indicate the order in which these stubs are enqueued on Q. The first in Q is stub 1 of node v_1. We discover its partner (stub 1 of node v_2) by scanning the interval $[0,1]$ from left to right. Because v_2 has not been sampled yet, we "sample" it now (Step 6) and we enqueue its stubs 2 and 3 on Q (Step 7). The next in Q is stub 2 of node v_1; we continue scanning to discover its partner (stub 1 of node v_3). And so on. **Center:** The state of the system at time t. All stubs in $[0,t]$ have already been matched (the indices of matched stubs are set in plain line). All unmatched stubs are distributed uniformly at random on $(t,1]$. This interval can contain also some (here two) already matched stubs. **Right:** The final result is a realization of a random graph G with a given node degree sequence (*i.e.*, of the configuration model). G may contain self-loops and multiedges.

as follows. First, we assign each stub a real-valued index t drawn uniformly at random from the interval $[0,1]$. Then, every time we process line 4, we pick b as the unmatched stub with the smallest index. We can interpret this as a continuous-time process, where we determine progressively the partners of stubs dequeued from Q, by scanning the interval from time $t=0$ to $t=1$ in a search of unmatched stubs. Because the indices chosen by the stubs are independent from each other, the above trick breaks the dependence between the stubs, which is crucial for making this approach tractable.

In Fig. 2, we present an example execution of Algorithm 1, where line 4 is implemented as described above.

4) Expected sampled degree distribution q_k^{BFS}: Now we are ready to derive the expected observed degree distribution q_k. Recall that all the stub indices are chosen independently and uniformly from $[0,1]$. A vertex v with degree k is not sampled yet at time t if the indices of all its k stubs are larger than t, which happens with probability $(1-t)^k$. So the probability that v is sampled before time t is $1-(1-t)^k$. Therefore, the expected fraction of vertices of degree k sampled before t is

$$f_k(t) = p_k(1-(1-t)^k). \tag{5}$$

By normalizing Eq.(5), we obtain the expected observed (*i.e.*, sampled) degree distribution at time t:

$$q_k^{\text{BFS}}(t) \;=\; \frac{f_k(t)}{\sum_l f_l(t)} \;=\; \frac{p_k(1-(1-t)^k)}{\sum_l p_l(1-(1-t)^l)}. \tag{6}$$

Unfortunately, it is difficult to interpret $q_k^{\text{BFS}}(t)$ directly, because t is proportional neither to the number of matched edges nor to the number of discovered nodes. Recall that our primary goal is to express q_k^{BFS} as a function of fraction f of covered nodes. We achieve this by calculating $f(t)$ - the expected fraction of nodes, of any degree, visited before time t

$$f(t) = \sum_k f_k(t) = 1 - \sum_k p_k(1-t)^k \;. \tag{7}$$

Because $p_k \geq 0$, and $p_k > 0$ for at least one $k > 0$, the term $\sum_k p_k(1-t)^k$ is continuous and strictly decreasing from 1 to 0 with t growing from 0 to 1. Thus, for $f \in [0,1]$ there exists a well defined $t{=}t(f)$ that satisfies Eq.(7), *i.e.*, the inverse of $f(t)$. Although we cannot compute $t(f)$ analytically (except in some special cases such as for $k \leq 4$), it is straightforward to find it numerically. Now, we can rewrite Eq. (6) as

$$q_k^{\text{BFS}}(f) \;=\; \frac{p_k(1-(1-t(f))^k)}{\sum_l p_l(1-(1-t(f))^l)}, \tag{8}$$

which is the expected observed degree distribution after covering fraction f of nodes of graph G. Consequently, the expected observed average degree is

$$\langle q_k^{\text{BFS}} \rangle(f) \;=\; \sum_k k \cdot q_k^{\text{BFS}}(f). \tag{9}$$

In other words, Eq.(8) and Eq.(9) describe the bias of BFS sampling under $RG(p_k)$, which was our first goal in this paper. Below, we further analyze these equations to get more insights in the nature of BFS bias.

5) Equivalence of traversal techniques under $RW(p_k)$: An interesting observation is that, under the random graph model $RW(p_k)$, all common traversal techniques (BFS, DFS, FF, SBS, etc) are subject to exactly the same bias. The explanation is that the sampled node sequence S is fully determined by the choice of stub indices on $[0,1]$, independently of the way we manage the elements in Q.

6) Equivalence of traversals to weighted sampling without replacement: Consider a node v with a degree k_v. The probability that v is discovered before time t, given that it has not been discovered before $t_0 \leq t$, is

$$\mathbb{P}(v \text{ before time } t \mid v \text{ not before } t_0) = 1 - \left(\frac{1-t}{1-t_0}\right)^{k_v} \tag{10}$$

We now take the derivative of the above equation with respect to t, which results in the conditional probability density function $k_v(\frac{1-t}{1-t_0})^{k_v-1}$. Setting $t \to t_0$ (but keeping $t > t_0$), reduces it to k_v, which is the probability density that v is sampled at t_0, given that it has not been sampled before. This means that at every point in time, out of all nodes that have not yet been selected, the probability of selecting v is proportional to its degree k_v. Therefore, this scheme is equivalent to node sampling weighted by degree, without replacements.

7) Equivalence of traversals with $f\to 0$ to RW: Finally, for $f\to 0$ (and thus $t\to 0$), we have $1-(1-t)^k \simeq kt$, and Eq. (6) simplifies to Eq. (1). This means that in the beginning of the sampling process, every traversal technique is equivalent to RW, as shown in Fig. 1 for $f\to 0$.

8) $\langle q_k^{\text{BFS}}\rangle$ is decreasing in f: As in Section V-B2, let $X_i \in V$ be the ith selected node, and let $z{=}2|E|$. We have shown above that our procedure is equivalent to weighted sampling without replacements, thus we can write $\mathbb{P}(X_1 = u) = \frac{k_u}{z}$. Now, it follows from Eq. (3) that $\mathbb{P}(X_2 = w) = \frac{k_w}{z} \cdot \alpha_w$, where $\alpha_w = \sum_{u\neq w} \frac{k_u}{z-k_u}$. Because for any two nodes a and b, we have $\alpha_b - \alpha_a = z(k_a-k_b)/((z-k_a)(z-k_b))$, α_w strictly decreases with growing k_w. As a result, $\mathbb{P}(X_2)$ is more concentrated around nodes with smaller degrees than is $\mathbb{P}(X_1)$, implying that $\mathbb{E}[k_{X_2}] < \mathbb{E}[k_{X_1}]$. We can use an analogous argument at every iteration $i \le |V|$, which allows us to say that $\mathbb{E}[k_{X_i}] < \mathbb{E}[k_{X_{i-1}}]$. In other words, $\langle q_k^{\text{BFS}}\rangle(f)$ is a decreasing function of f.

A practical consequence is that many short traversals are more biased than a long one, with the same total number of samples.

9) Comments on the graph connectivity: Note that the configuration model $RG(p_k)$ might result in a graph G that is not connected. In this case, every exploration technique covers only the component C in which it was initiated; consequently, the process described in Section V-B3 stops once C is covered.

In practice, it is also possible to efficiently generate a simple and connected random graph with a given degree sequence [46].

VI. Correcting for Node Degree Bias

In the previous section, we derived the expected observed degree distribution q_k as a function of the original degree distribution p_k. The distribution q_k is usually biased towards high-degree nodes, *i.e.*, $\langle q_k\rangle > \langle p_k\rangle$. Moreover, because many node properties are correlated with the node degree [2], their estimates are also potentially biased. For example, let $x(v)$ be an arbitrary function defined on graph nodes V (*e.g.*, node age) and let its mean value

$$x_{\text{av}} = \frac{1}{|V|}\sum_{v\in V} x(v) \tag{11}$$

be the value we are trying to estimate. If $x(v)$ is correlated with node degree k_v, then the straightforward estimator $\widehat{x}_{\text{av}}^{naive} = 1/|S| \cdot \sum_{v\in S} x(v)$ is subject to the same bias as is $\langle q_k\rangle$. In this section, we derive estimators $\widehat{x}_{\text{av}}$ of x_{av}.

Note that with this approach we can estimate not only the mean values, but also the entire distributions. For example, if $x(v){=}1_{\{k_v=k\}}$ then $\widehat{x}_{\text{av}}$ estimates the proportion of nodes with degree equal to k, *i.e.*, the original node degree distribution p_k. We derive $\widehat{p}_k$ for all cases below.

Let $S \subset V$ be a sequence of vertices that we sampled. Based on S, we can estimate q_k as

$$\widehat{q}_k \;=\; \frac{\text{number of nodes in } S \text{ with degree } k}{|S|}. \tag{12}$$

A. Random walks (baseline)

Under RW, a straightforward application of the Hansen-Hurwitz estimator [47] leads to (after [14]–[17])

$$\widehat{x}_{\text{av}}^{\text{RW}} \;=\; \frac{\sum_{v\in S} x(v)/k_v}{\sum_{v\in S} 1/k_v}. \tag{13}$$

By plugging $x(v){=}1_{\{k_v=k\}}$, we can estimate the original node degree distribution as

$$\widehat{p}_k^{\text{RW}} \;=\; \frac{\widehat{q}_k}{k} \;\cdot\; \left(\sum_l \frac{\widehat{q}_l}{l}\right)^{-1} \tag{14}$$

where we used the fact that $\sum_{v\in S} 1_{\{k_v=k\}} = |V|\cdot\widehat{q}_k$. From Eq.(14), we can estimate the average node degree as

$$\langle\widehat{p}_k^{\text{RW}}\rangle \;=\; \sum_k k\,\widehat{p}_k^{\text{RW}} \;=\; 1\cdot\left(\sum_l \frac{\widehat{q}_l}{l}\right)^{-1} = \frac{|S|}{\sum_{v\in S}\frac{1}{k_v}}, \tag{15}$$

where the last equation follows from Eq.(12).

B. Graph traversals

Under BFS and other traversals, the inclusion probability π_v^{BFS} (*i.e.*, the probability of node v being included in sample S) of node $v \in V$ is proportional to

$$\pi_v^{\text{BFS}} \;\sim\; \frac{q_{k_v}^{\text{BFS}}}{p_{k_v}} \;\sim\; 1-(1-t(f))^{k_v},$$

where the second relation originates from Eq.(8). Consequently, an application of the Horvitz-Thompson estimator [49], designed typically for sampling without replacement, leads to

$$\widehat{x}_{\text{av}}^{\text{BFS}} \;=\; \left(\sum_{v\in S}\frac{x(v)}{1-(1-t(f))^{k_v}}\right)\cdot\left(\sum_{v\in S}\frac{1}{1-(1-t(f))^{k_v}}\right)^{-1}. \tag{16}$$

Now, similarly to the analysis of RW (above), we obtain

$$\widehat{p}_k^{\text{BFS}} \;=\; \frac{\widehat{q}_k}{1-(1-t(f))^k} \;\cdot\; \left(\sum_l \frac{\widehat{q}_l}{1-(1-t(f))^l}\right)^{-1} \tag{17}$$

$$\langle\widehat{p}_k^{\text{BFS}}\rangle \;=\; \sum_k k\,\widehat{p}_k^{\text{BFS}}. \tag{18}$$

However, in order to evaluate these expressions, we need to evaluate $t(f)$, that, in turn, requires p_k. We can solve this chicken-and-egg problem iteratively, if we know the real fraction f^{real} of covered nodes, or equivalently the graph size $|V|$. First, we evaluate Eq.(17) for some values of t and feed the resulting $\widehat{p}_k$'s into Eq. (7) to obtain the corresponding f's. By repeating this process, we can efficiently drive the values of f arbitrarily close to f^{real}, and thus find the desired $\widehat{p}_k$.

In summary, for BFS, we showed how to estimate the mean x_{av} of an arbitrary function $x(v)$ defined on graph nodes, with the estimator of the original degree distribution p_k as a special case. Note that our approach is feasible, as it requires only the sample S (with value $x(v)$ and degree k_v for every node $v\in S$) and the fraction f of sampled nodes. In [28], we make a `python` implementation of all the above estimators publicly available.

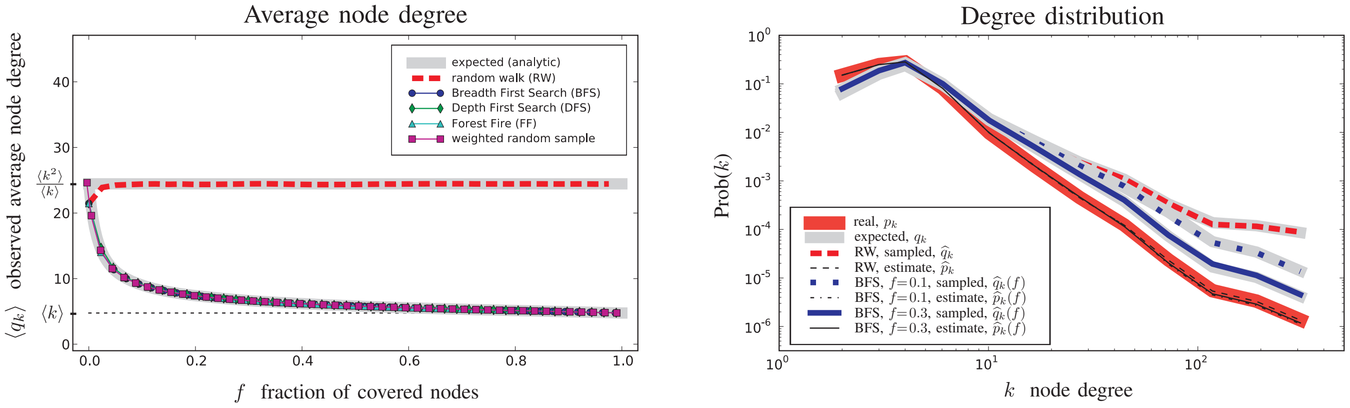

Fig. 3. **Comparison of sampling techniques in theory and in simulation. Left:** Observed (sampled) average node degree $\langle q_k \rangle$ as a function of the fraction f of sampled nodes, for various sampling techniques. The results are averaged over 1000 graphs with 10000 nodes each, generated by the configuration model with a fixed heavy-tailed degree distribution p_k (shown on the right). **Right:** Real, expected, and estimated (corrected) degree distributions for selected techniques and values of f (other techniques behave analogously). We obtained analogous results for other degree distributions and graph sizes $|V|$. The term $\langle k \rangle$ is the real average node degree, and $\langle k^2 \rangle$ is the real average squared node degree.

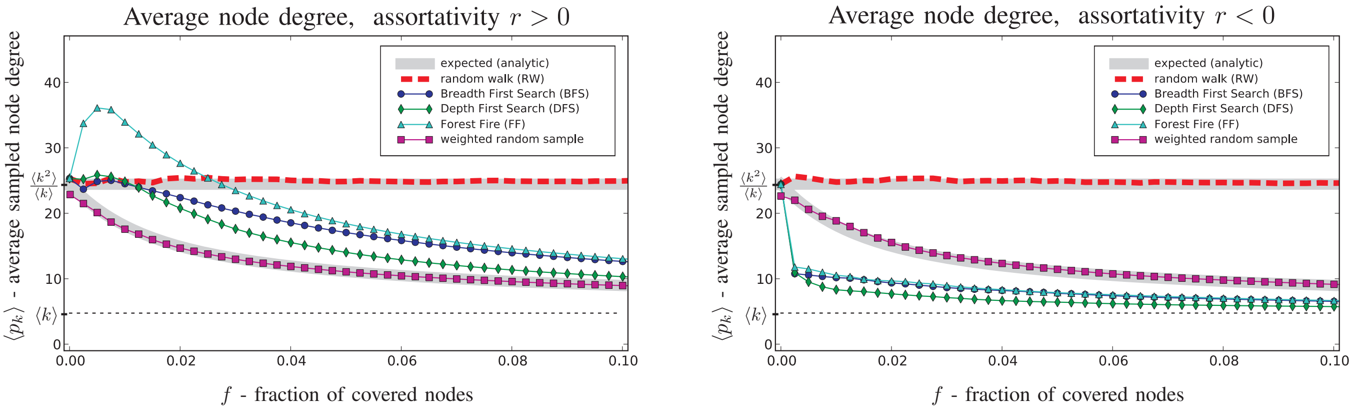

Fig. 4. **The effect of assortativity** r **on the results**. First, we use the configuration model with the same degree distribution p_k as in Fig. 3 (and the same number of nodes $|V| = 10000$) to generate a graph G. Next, we apply the pairwise edge rewiring technique [48] to change the assortativity r of G without changing node degrees. This technique iteratively takes two random edges $\{v_1, w_1\}$ and $\{v_2, w_2\}$, and rewires them as $\{v_1, w_2\}$ and $\{v_2, w_1\}$ only if it brings us closer to the desired value of assortativity r. As a result, we obtain graphs with a positive (left) and negative (right) assortativity r. Note that for a better readability, we present only the values of $f \in [0, 0.1]$, *i.e.*, ten times smaller than in Fig. 3.

VII. Simulation Results

In this section, we evaluate our theoretical findings on random and real-life graphs.

A. Random graphs

Fig. 3 verifies all the formulae derived in this paper, for the random graph $RG(p_k)$ with a given degree distribution. The analytical expectations are plotted in thick plain lines in the background and the averaged simulation results are plotted in thinner lines lying on top of them. We observe almost a perfect match between theory and simulation in estimating the sampled degree distribution q_k (Fig. 3, right) and its mean $\langle q_k \rangle$ (Fig. 3, left). Indeed, all traversal techniques follow the same curve (as predicted in Section V-B5), which initially coincides with that of RW (see Section V-B7) and is monotonically decreasing in f (see Section V-B8). We also show that degree-weighted node sampling without replacements exhibits exactly the same bias (see Section V-B6). Finally, applying the estimators $\widehat{p}_k$ derived in Section VI perfectly corrects for the bias of q_k.

Of course, real-life networks are substantially different from $RG(p_k)$. For example, depending on the graph type, nodes may tend to connect to similar or different nodes. Indeed, in most social networks high-degree nodes tend to connect to other high-degree nodes [55]. Such networks are called *assortative*. In contrast, biological and technological networks are typically *disassortative*, *i.e.*, they exhibit significantly more high-degree-to-low-degree connections. This observation can be quantified by calculating the *assortativity coefficient* r [55], which is the correlation coefficient computed over all edges (*i.e.*, degree-degree pairs) in the graph. Values $r<0$, $r>0$ and $r = 0$ indicate disassortative, assortative and purely random graphs, respectively.

For the same initial parameters as in Fig. 3 (p_k, $|V|$), we simulated different levels of assortativity. Fig. 4 shows the results. Graph assortativity r strongly affects the first iterations of traversal techniques. Indeed, for assortativity $r > 0$ (Fig. 4, left), the degree bias is even stronger than for $r = 0$ (Fig. 3, left). This is because the high-degree nodes are now interconnected more densely than in a purely random graph, and are thus easier to discover by sampling techniques

TABLE II

INTERNET TOPOLOGIES USED IN SIMULATIONS. ALL GRAPHS ARE CONNECTED AND UNDIRECTED (WHICH REQUIRED PREPROCESSING IN SOME CASES).

Dataset	# nodes	# edges	$\langle k\rangle = \langle p_k\rangle$	$\frac{\langle k^2\rangle}{\langle k\rangle}$	Description
ca-CondMat	21 363	91 341	8.6	22.5	Collaboration network of Arxiv Condensed Matter [50]
email-EuAll	224 832	340 794	3.0	567.9	Email network of a large European Research Institution [50]
Facebook-New-Orleans	63 392	816 885	25.8	88.1	Facebook New Orleans network [51]
wiki-Talk	2 388 953	4 656 681	3.9	2705.4	Wikipedia talk (communication) network [52]
p2p-Gnutella31	62 561	147 877	4.7	11.6	Gnutella peer to peer network from August 31 2002 [50]
soc-Epinions1	75 877	405 738	10.7	183.9	Who-trusts-whom network of Epinions.com [53]
soc-Slashdot0811	77 360	546 486	14.1	129.9	Slashdot social network from November 2008 [54]
as-caida20071105	26 475	53 380	4.0	280.2	CAIDA AS Relationships Datasets, from November 2007
web-Google	855 802	4 291 351	10.0	170.4	Web graph from Google [54]

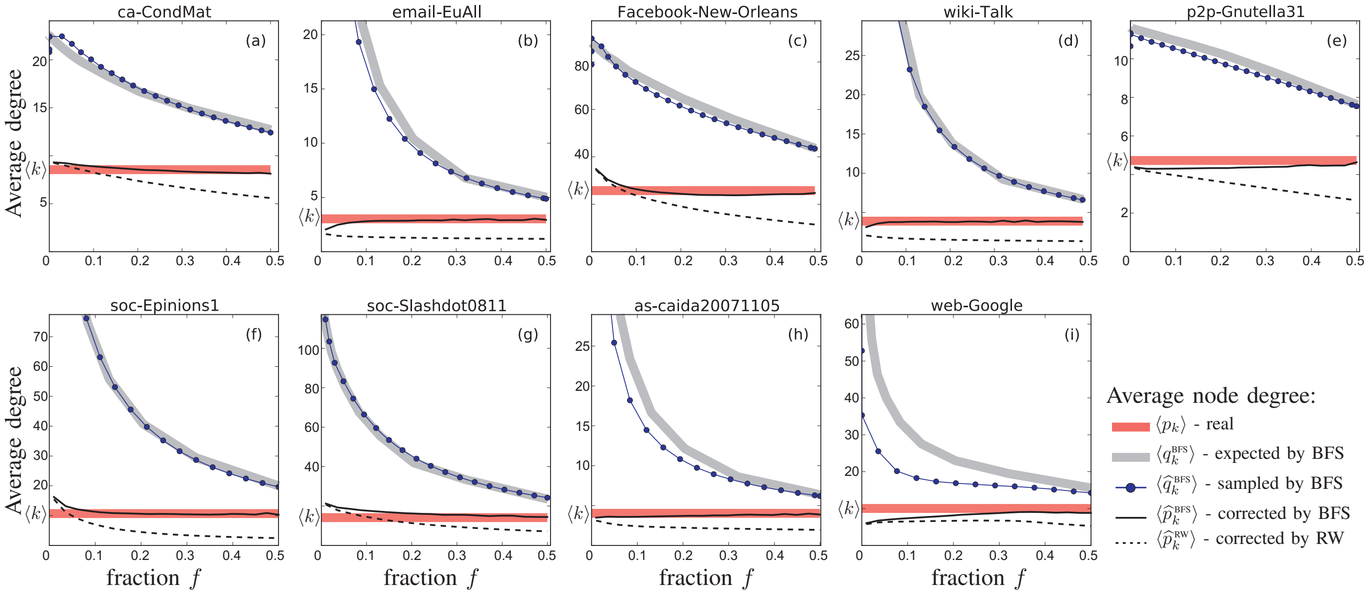

Fig. 5. **BFS in real-life (fully known) Internet topologies described in Table VII-A.** The blue circles represent the average node degree $\langle \widehat{q}_k^{\text{BFS}}\rangle$ sampled by BFS, as the function of the fraction of covered nodes f. The thin lines are the corrected values $\langle \widehat{p}_k^{\text{BFS}}\rangle$ resulting from the BFS estimator Eq.(18) (plain line) and the RW estimator Eq.(15) (dashed). Results are averaged over 1000 randomly seeded BFS samples. The thick lines are the analytical expectations assuming the random graph model $RG(p_k)$. Thick red line (top) is the expectation of $\langle q_k^{\text{BFS}}\rangle$, calculated with Eq.(9) given the knowledge of the true node degree distribution p_k. Thick gray line (bottom) is the expectation of corrected $\langle \widehat{p}_k^{\text{BFS}}\rangle$, Eq.(18), *i.e.*, precisely $\langle k\rangle$.

that are inherently biased towards high-degree nodes. Interestingly, Forest Fire is by far the most affected. A possible explanation is that under Forest Fire, low-degree nodes are likely to be completely skipped by the first sampling wave. Not surprisingly, a negative assortativity $r < 0$ has the opposite effect: every high-degree node tends to connect to low-degree nodes, which significantly slows down the discovery of the former.

In contrast, random walk RW is not affected by the changes in assortativity. This is expected, because their stationary distributions hold for *any* fixed (connected and aperiodic) graph regardless of its topological properties.

B. Real-life fully known topologies

Recall, that our analysis is based on the random graph model $RG(p_k)$ (see Section IV), which is only an approximation of a typical real-life network G. Indeed, $RG(p_k)$ follows the node degree distribution of G, but is likely to miss other important properties such as assortativity [55], whose effect on the BFS process we have just demonstrated. For this reason, one may expect that the technique based on $RG(p_k)$ performs poorly on real-life graphs. Surprisingly, this is not the case.

We evaluated our approach on a broad range of large, real-life, fully known Internet topologies. As our main source of data we use SNAP Graph Library [56]; Table VII-A overviews these datasets. We present the results in Fig. 5. Interestingly, in most cases the sampled average node degree $\langle \widehat{q}_k^{\text{BFS}}\rangle$ closely matches the prediction $\langle q_k^{\text{BFS}}\rangle$ of the random graph model $RG(p_k)$. More importantly, applying our BFS estimator $\langle \widehat{p}_k^{\text{BFS}}\rangle$ of real average node degree corrects for the bias of $\langle \widehat{q}_k^{\text{BFS}}\rangle$ surprisingly well. Some significant differences are visible only for $f \rightarrow 0$ and for some specific topologies (the last two in Fig. 5), which is exactly because the real-life graphs are not fully captured by graph model $RG(p_k)$.

Finally, we also study the RW estimator Eq.(15), as a simpler alternative to the BFS one Eq.(18). Although they coincide for $f \rightarrow 0$, the RW estimator systematically and significantly underestimates the average node degree $\langle k\rangle$ for larger values of f.

C. Sampling Facebook and Orkut

In this section, we apply and test the previous ideas in sampling real-life, large-scale, and not fully known online social networks: Facebook and Orkut.

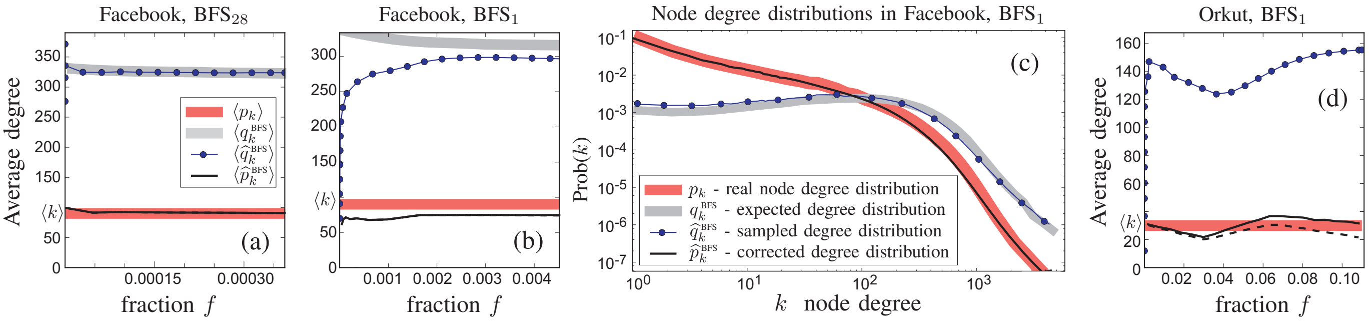

Fig. 6. **BFS in on-line (not fully known) topologies.** As in Fig. 5, except that the plots are based on BFS samples taken in Facebook with 28 (random) seeds (a) and one seed (b), as well as in Orkut with one seed (d). Additionally, we show in (c) the full node degree distributions for Facebook. Because we do not have the true degree distribution p_k of Orkut, we cannot calculate its analytical curve $\langle q_k^{\text{BFS}}\rangle$. Nevertheless, we show in (d) our best guess of Orkut's average node degree $\langle k\rangle$ learned by other means, as explained in Footnote 2.

1) Facebook: We have implemented a set of crawlers to collect the samples of Facebook (FB) following the BFS and RW techniques. The data sets are summarized in Table VII-C2. BFS$_{28}$ consists of 28 small BFS-es initiated at 28 different nodes, which allowed us to easily parallelize the process. Moreover, at the time of data collection, we (naively) thought that this would reduce the BFS bias. After gaining more insight (which, nota bene, motivated this paper), we collected a single large BFS$_1$. UNI represents the ground truth. The details of our implementation are described in [2,11].

Results. We present the Facebook sampling results in Fig. 6(a-c) and in Table VII-C2. First, we observe that under BFS$_{28}$, our estimators q_k^{BFS} and $\widehat{p}_k^{\text{BFS}}$ perform very well. For example, we obtain $\langle \widehat{p}_k^{\text{BFS}}\rangle = 85.4$ compared with the true value $\langle k\rangle = 94.1$. In contrast, BFS$_1$ yields $\langle \widehat{p}_k^{\text{BFS}}\rangle = 72.7$ only. Most probably, this is because BFS$_1$ consists of a single BFS run that happens to begin in a relatively sparse part of Facebook. Indeed, note that this run starts at $\widehat{q}_k^{\text{BFS}} = 50$ for $f = 0$, and systematically grows with f instead of falling.

Finally, note that both BFS$_{28}$ and BFS$_1$ are very short compared to the Facebook size, with $f < 1\%$ in both cases. For this reason, we observe almost no drop in the sampled average node degre $\langle q_k^{\text{BFS}}\rangle$ in Fig. 6(a,b). For the same reason, both the BFS and RW estimators yield almost identical results.

All the above observations hold also for the *entire* degree distribution, which is shown in Fig. 6(c).

2) Orkut: Finally, we apply our methodology to a single BFS sample of Orkut collected in 2006 and described in [22]. It contains $|S| = 3072K$ nodes, which accounts for $f = 11.3\%$ of entire Orkut size.

We show the results in Fig. 6(d). Similarly to Facebook BFS$_1$, the sampled average node degree $\langle \widehat{q}_k^{\text{BFS}}\rangle$ does not decrease monotonically in f. Again, the underlying reason might be the arbitrary choice of the starting node (in sparsely connected India in this case). Nevertheless, the estimator $\langle \widehat{p}_k^{\text{BFS}}\rangle$ approximates the average node degree[3] relatively well.

TABLE III
FACEBOOK AND ORKUT DATA SETS AND MEASUREMENTS.

Facebook	UNI	RW	BFS$_{28}$	BFS$_1$
$\vert S\vert$	982K	2.26M	28×81K	1.19M
f	0.44%	1.03%	28×0.04%	0.54%
$\langle \widehat{q}_k\rangle$	94.1	338.0	323.9	285.9
$\langle q_k\rangle$	-	329.8	329.1	328.7
$\langle \widehat{p}_k\rangle$	-	93.9	85.4	72.7
Orkut				
$\vert S\vert$	-	-	-	3.07M
f	-	-	-	11.3%
$\langle \widehat{p}_k\rangle$	30 [2]			33.1

VIII. PRACTICAL RECOMMENDATIONS

In order to sample *node properties*, we recommend using RW and its variants. RW is simple, unbiased for arbitrary topologies (assuming that we use correction procedures summarized in Section VI), and practically unaffected by the starting point.

In contrast, RW is not useful when sampling *non-local graph properties*, such as the graph diameter or the average shortest path length. In this case, BFS seems very attractive, because it produces a full view of a particular region in the graph, which is usually a plausible graph for which the non-local properties can be easily calculated. However, all such results should be interpreted very carefully, as they may be also strongly affected by the bias of BFS. For example, the graph diameter drops significantly with growing average node degree of a network. Whenever possible, it is a good practice to restrict BFS to some well defined community in the sampled graph. If the community is small enough, we may be able to exhaust it (at least its largest connected component), which automatically makes our BFS sample representative of this community. For example, [23,51] collected full samples of several Facebook regional networks, and [54,58] completely covered the WWW graph restricted to one or few domains. When such communities are not available (*e.g.*, regional networks are not accessible anymore in Facebook), we are left with a regular unconstrained BFS sample. In that case, we recommend applying the $RG(p_k)$-based correction procedure presented in this paper to quantify the node degree bias, which may help us evaluate the bias introduced in the topological metrics.

[3] Unfortunately, according to our personal communication with Orkut administrators, there is no ground truth value of the Orkut's average node degree $\langle k\rangle$ for October 2006, *i.e.*, the period when the BFS sample of [22] was collected. However, many hints point to a number close to $\langle k\rangle = 30$, *e.g.*, [21] reports $\langle k\rangle = 30.2$ in June-September 2006, and [57] reports $\langle k\rangle = 19$ in late 2004 (which is in agreement with the densification law [43,50]). But, as these studies may potentially be subject to various biases, we cannot take these numbers for granted.

IX. Conclusion

To the best of our knowledge, our work is the first to quantify the node-degree bias of BFS sampling. In particular, we calculated the node degree distribution q_k expected to be observed by BFS as a function of the fraction f of covered nodes, in a random graph $RG(p_k)$ with a given degree distribution p_k. We found that for a small sample size, $f \rightarrow 0$, BFS has the same bias as the simple Random Walk, and with increasing f, the bias monotonically decreases. Based on our theoretical analysis, we proposed a practical $RG(p_k)$-based procedure to correct for this bias when calculating any node statistics. Our technique performed well on a broad range of Internet topologies. Its ready-to-use implementation can be downloaded from [28].

In this paper, we used our $RG(p_k)$-based correction procedure to estimate local graph properties, such as node statistics. A direction for future work is to exploit the node degree-biases calculated here to develop estimators of non-local graph properties, such as graph diameter.

Acknowledgments

We would like to thank Bruno Ribeiro for insightful discussions; Alan Mislove for providing the Orkut BFS sample; and Minas Gjoka for collecting the Facebook BFS sample.

References

[1] M. Kurant, A. Markopoulou, and P. Thiran, "On the bias of BFS (Breadth First Search)," in *Proc. 22nd Int. Teletraffic Congr., also in arXiv:1004.1729*, 2010.

[2] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, "Practical Recommendations on Sampling OSN Users by Crawling the Social Graph," *To appear in IEEE J. Sel. Areas Commun. on Measurement of Internet Topologies*, 2011.

[3] L. Lovász, "Random walks on graphs: A survey," *Combinatorics, Paul Erdos is Eighty*, vol. 2, no. 1, pp. 1–46, 1993.

[4] M. R. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork, "On near-uniform URL sampling," in *Proc. 9th Int. Conf. on World Wide Web*, Amsterdam, Netherlands, 2000.

[5] C. Gkantsidis, M. Mihail, and A. Saberi, "Random walks in peer-to-peer networks," in *Proc. IEEE INFOCOM*, Hong Kong, China, 2004.

[6] D. Stutzbach, R. Rejaie, N. Duffield, S. Sen, and W. Willinger, "On unbiased sampling for unstructured peer-to-peer networks," in *Proc. 6th ACM SIGCOMM Conf. on Internet measurement*, Rio de Janeiro, Brazil, 2006.

[7] A. Rasti, M. Torkjazi, R. Rejaie, N. Duffield, W. Willinger, and D. Stutzbach, "Respondent-driven sampling for characterizing unstructured overlays," in *Proc. IEEE INFOCOM Mini-conference*, Rio de Janeiro, Brazil, 2009, pp. 2701–2705.

[8] B. Ribeiro and D. Towsley, "Estimating and sampling graphs with multidimensional random walks," in *Proc. 10th ACM SIGCOMM Conf. on Internet measurement*, Melbourne, Australia, 2010.

[9] 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.

[10] M. Kurant, M. Gjoka, C. T. Butts, and A. Markopoulou, "Walking on a Graph with a Magnifying Glass: Stratified Sampling via Weighted Random Walks," in *Proc. ACM SIGMETRICS*, 2011.

[11] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, "Walking in Facebook: A Case Study of Unbiased Sampling of OSNs," in *Proc. IEEE INFOCOM*, San Diego, CA, 2010.

[12] B. Krishnamurthy, P. Gill, and M. Arlitt, "A few chirps about twitter," in *Proc. 1st workshop on Online social networks*, Seattle, WA, 2008, pp. 19–24.

[13] J. Leskovec and C. Faloutsos, "Sampling from large graphs," in *Proc. 12th ACM SIGKDD Int. Conf. on Knowledge discovery and data mining*, Philadelphia, PA, 2006, pp. 631–636.

[14] S. L. Feld, "Why Your Friends Have More Friends Than You Do," *American Journal of Sociology*, vol. 96, no. 6, p. 1464, May 1991.

[15] M. Newman, "Ego-centered networks and the ripple effect," *Social Networks*, vol. 25, pp. 83–95, 2003.

[16] M. Salganik and D. D. Heckathorn, "Sampling and estimation in hidden populations using respondent-driven sampling," *Sociological Methodology*, vol. 34, no. 1, pp. 193–240, 2004.

[17] E. Volz and D. D. Heckathorn, "Probability based estimation theory for respondent driven sampling," *J. Official Statistics*, vol. 24, no. 1, pp. 79–97, 2008.

[18] L. A. Goodman, "Snowball sampling," *Annals of Mathematical Statistics*, vol. 32, pp. 148–170, 1961.

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

[20] M. Najork and J. L. Wiener, "Breadth-first search crawling yields high-quality pages," in *WWW*, 2001.

[21] Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong, "Analysis of topological characteristics of huge online social networking services," in *Proc. 16th Int. Conf. on World Wide Web*, Banff, Alberta, Canada, 2007, pp. 835–844.

[22] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proc. 7th ACM SIGCOMM Conf. on Internet measurement*, San Diego, CA, 2007, pp. 29–42.

[23] C. Wilson, B. Boe, A. Sala, K. Puttaswamy, and B. Zhao, "User interactions in social networks and their implications," in *Proc. 4th ACM European Conf. on Computer systems*, Nuremberg, Germany, 2009, pp. 205–218.

[24] S. H. Lee, P.-J. Kim, and H. Jeong, "Statistical properties of sampled networks," *Physical Review E*, vol. 73, p. 16102, 2006.

[25] L. Becchetti, C. Castillo, D. Donato, A. Fazzone, and I. Rome, "A comparison of sampling techniques for web graph characterization," in *Proc. Workshop on Link Analysis*, Philadelphia, PA, 2006.

[26] S. Ye, J. Lang, and F. Wu, "Crawling online social graphs," in *Proc. 12th Asia-Pacific Web Conference*, Busan, Korea, 2010, pp. 236–242.

[27] D. Achlioptas, A. Clauset, D. Kempe, and C. Moore, "On the bias of traceroute sampling: or, power-law degree distributions in regular graphs," in *Proc. 37th Annu. ACM Symp. on Theory of computing*, 2005, pp. 694–703.

[28] M. Kurant, "Python scripts for BFS sampling and bias correction: http://mkurant.com/maciej/publications/papers/traversals.zip."

[29] A. Mislove, H. Koppula, K. Gummadi, P. Druschel, and B. Bhattacharjee, "Growth of the flickr social network," in *Proc. 1st workshop on Online social networks*, Seattle, WA, 2008, pp. 25–30.

[30] J. H. Kim, "Poisson cloning model for random graphs," in *International Congress of Mathematicians (ICM)*, 2006.

[31] M. Molloy and B. Reed, "A critical point for random graphs with a given degree sequence," *Random structures and algorithms*, vol. 6, no. 2-3, pp. 161–180, 1995.

[32] K. Gile and M. Handcock, "Respondent-driven sampling: An assessment of current methodology," *To appear in Sociological Methodology*, 2011.

[33] K. Gile, "Improved Inference for Respondent-Driven Sampling Data with Application to HIV Prevalence Estimation," *arXiv:1006.4837*, 2010.

[34] J. Illenberger, G. Flötteröd, and N. Kai, "An approach to correct bias induced by snowball sampling," in *Sunbelt Social Networks Conference*, 2009.

[35] F. Yates and P. Grundy, "Selection without replacement from within strata with probability proportional to size," *J. Royal Statistical Society. Series B (Methodological)*, vol. 15, no. 2, pp. 253–261, 1953.

[36] D. Raj, "Some estimators in sampling with varying probabilities without replacement," *J. American Statistical Association*, pp. 269–284, 1956.

[37] M. Murthy, "Ordered and unordered estimators in sampling without replacement," *Sankhyà: The Indian Journal of Statistics*, vol. 18, no. 3, pp. 379–390, 1957.

[38] H. Hartley and J. Rao, "Sampling with unequal probabilities and without replacement," *The Annals of Mathematical Statistics*, 1962.

[39] G. Andreatta and G. Kaufman, "Estimation of finite population properties when sampling is without replacement and proportional to magnitude," *J. American Statistical Association*, vol. 81, no. 395, pp. 657–666, 1986.

[40] T. J. Rao, S. Sengupta, and B. K. Sinha, "Some Order Relations Between Selection and Inclusion Probabilities for PPSWOR Sampling Scheme," *Metrika*, vol. 38, no. 1, pp. 335–343, Dec. 1991.

[41] S. Kochar and R. Korwar, "On random sampling without replacement from a finite population," *Annals of the Institute of Statistical Mathematics*, vol. 53, no. 3, pp. 631–646, 2001.

[42] L. Fattorini, "Applying the Horvitz-Thompson criterion in complex designs: A computer-intensive perspective for estimating inclusion probabilities," *Biometrika*, vol. 93, no. 2, pp. 269–278, Jun. 2006.

[43] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graphs over time: densification laws, shrinking diameters and possible explanations," in *KDD*, 2005.

[44] F. Chung and L. Lu, "Connected components in random graphs with given degree sequences," *Annals of Combinatorics*, vol. 6, p. 125145, 2002.

[45] R. Motwani and P. Raghavan, *Randomized Algorithms*. Cambridge University Press, 1990.

[46] F. Viger and M. Latapy, "Efficient and simple generation of random simple connected graphs with prescribed degree sequence," *LNCS Computing and Combinatorics*, vol. 3595, pp. 440–449, 2005.

[47] M. Hansen and W. Hurwitz, "On the Theory of Sampling from Finite Populations," *Annals of Mathematical Statistics*, vol. 14, no. 3, 1943.

[48] S. Maslov and K. Sneppen, "Specificity and stability in topology of protein networks," *Science*, vol. 296, no. 5569, p. 910, 2002.

[49] D. Horvitz and D. Thompson, "A generalization of sampling without replacement from a finite universe," *J. American Statistical Association*, vol. 47, no. 260, pp. 663–685, 1952.

[50] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graph evolution: Densification and shrinking diameters," *ACM Trans. Knowledge Discovery from Data*, vol. 1, no. 1, p. 2, Mar. 2007.

[51] B. Viswanath, A. Mislove, M. Cha, and K. Gummadi, "On the evolution of user interaction in facebook," in *Proc. 2nd workshop on Online social networks*, Barcelona, Spain, 2009, pp. 37–42.

[52] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Predicting positive and negative links in online social networks," in *WWW*, New York, New York, USA, 2010, p. 641.

[53] M. Richardson, R. Agrawal, and P. Domingos, "Trust management for the semantic web," *The SemanticWeb-ISWC 2003*, pp. 351–368, 2003.

[54] J. Leskovec, K. Lang, A. Dasgupta, and M. Mahoney, "Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters," *Internet Mathematics*, vol. 6, no. 1, pp. 29–123, 2009.

[55] M. Newman, "Assortative mixing in networks," *Physical Review Letters*, vol. 89, no. 20, p. 208701, 2002.

[56] "SNAP Graph Library." [Online]. Available: http://snap.stanford.edu/data/

[57] Z. Anwar, W. Yurcik, V. Pandey, A. Shankar, I. Gupta, and R. Campbell, "Leveraging Social-Network Infrastructure to Improve Peer-to-Peer Overlay Performance: Results from Orkut," *Arxiv preprint cs/0509095*, 2005.

[58] R. Albert, H. Jeong, and A. Barabási, "Diameter of the world-wide web," *Nature*, vol. 401, no. 6749, pp. 130–131, 1999.

Maciej Kurant received a M.Sc. degree from Gdansk University of Technology, Poland, in 2002, and a Ph.D. degree from EPFL, Lausanne, Switzerland, in 2009. Currently, he is a postdoc at University of California, Irvine. His main areas of research interest include sampling and inference from large-scale networks (such as Internet topologies or the human brain), multipath routing with FEC, and survivability in WDM networks. He is also a founder of AcronymCreator.net - a tool that helps creating new, meaningful acronyms.

Athina Markopoulou (SM '98, M'02) is an assistant professor in the EECS Dept. at the University of California, Irvine. She received the Diploma degree in Electrical and Computer Engineering from the National Technical University of Athens, Greece, in 1996, and the M.S. and Ph.D. degrees, both in Electrical Engineering, from Stanford University in 1998 and 2003, respectively. She has been a postdoctoral fellow at Sprint Labs (2003) and at Stanford University (2004-2005), and a member of the technical staff at Arastra Inc. (2005). Her research interests include network coding, network measurements and security, media streaming and online social networks. She received the NSF CAREER award in 2008.

Patrick Thiran received the electrical engineering degree from the Université Catholique de Louvain, Louvain-la-Neuve, Belgium, in 1989, the M.S. degree in electrical engineering from the University of California at Berkeley, USA, in 1990, and the Ph.D. degree from EPFL, in 1996. He is an Associate Professor at EPFL. He became an Adjunct Professor in 1998, an Assistant Professor in 2002 and an Associate Professor in 2006. From 2000 to 2001, he was with Sprint Advanced Technology Labs, Burlingame, CA. His research interests include communication networks, performance analysis, dynamical systems, and stochastic models. He is currently active in the analysis and design of wireless multihop networks and in network monitoring. Dr. Thiran served as an Associate Editor for the IEEE Transactions on Circuits and Systems in 1997-1999, and he is currently an Associate Editor for the IEEE/ACM Transactions on Networking. He was the recipient of the 1996 EPFL Ph.D. award and of the 2008 Crédit Suisse Teaching Award.