Characterizing Directed and Undirected Networks via Multidimensional Walks with Jumps

FABRICIO MURAI, University of Massachusetts Amherst
BRUNO RIBEIRO, Purdue University
DON TOWSLEY, University of Massachusetts Amherst
PINGHUI WANG, Xi'an Jiaotong University

Estimating distributions of labels associated with nodes (e.g., number of connections or citizenship of users in a social network) in large graphs via sampling is a vital part of the study of complex networks. Due to their low cost, sampling via random walks (RWs) has been proposed as an attractive solution to this task. Most RW methods assume either that the network is undirected or that walkers can traverse edges regardless of their direction. Some RW methods have been designed for directed networks where edges coming into a node are not directly observable. In this work, we propose Directed Unbiased Frontier Sampling (DUFS), a sampling method based on a large number of coordinated walkers, each starting from a node chosen uniformly at random. It is applicable to directed networks with invisible incoming edges because it constructs, in real-time, an undirected graph consistent with the walkers trajectories, and due to the use of random jumps which prevent walkers from being trapped. DUFS generalizes previous RW methods and is suited for undirected networks and to directed networks regardless of in-edges visibility. We also propose an improved estimator of vertex label distributions which combines information from the initial walker locations with subsequent RW observations. We evaluate DUFS, comparing it against other RW methods, investigating the impact of its parameters on estimation accuracy and providing practical guidelines for choosing them. In estimating out-degree distributions, DUFS yields significantly better estimates of the head than other methods, while matching or exceeding estimation accuracy of the tail. Last, we show that DUFS outperforms VS when estimating distributions of node labels of the top 10% largest degree nodes, even when uniform vertex sampling has the same cost as RW steps.

CCS Concepts: •**Mathematics of computing** → *Maximum likelihood estimation; Multivariate statistics;* Density estimation;

Additional Key Words and Phrases: complex networks, directed networks, graph sampling, random walks

ACM Reference format:
Fabricio Murai, Bruno Ribeiro, Don Towsley, and Pinghui Wang. 2016. Characterizing Directed and Undirected Networks via Multidimensional Walks with Jumps. *ACM Trans. Knowl. Discov. Data.* 1, 1, Article 1 (January 2016), 32 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

This work is supported by the Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053, the Army Research Office under MURI W911NF-08-1-0233 and the CNPq, National Council for Scientific and Technological Development - Brazil.

Author's addresses: F. Murai and D. Towsley, College of Computer and Information Sciences, University of Massachusetts Amherst; B. Ribeiro, Purdue University; Pinghui Wang, Xi'an Jiaotong University.

 1556-4681/2016/1-ART1 $15.00
DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

A number of studies [4, 5, 7, 8, 14, 16, 18, 22, 24, 25, 32] are dedicated to the characterization of complex networks. A complex network is a network with topological features that do not occur in simple networks such as lattices or random networks. Examples of such networks include the Internet, the Web, social, business, and biological networks. Characterizing a network consists of computing or estimating a set of statistics that describe the network. In this work we model a complex network as a directed or undirected graph with labeled vertices. A label can be, for instance, the degree of a vertex or, in a social network setting, someone's hometown. Label statistics (e.g., average, distribution) are often used to characterize a network.

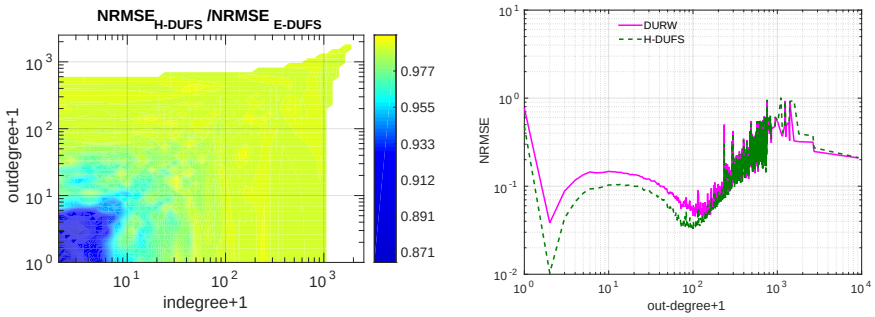
Characterizing a network with respect to its labels requires querying vertices and/or edges; associated with each query is a resource cost (time, bandwidth, money). For example, information about web pages must be obtained by querying web servers while subject to a maximum query rate. Characterizing a large network by querying the whole graph is often too costly. Even if the network is stored in disk it may constitute several terabytes of data. As a result, researchers have turned their attention to estimation of network characteristics based on incomplete (sampled) data.

Simple strategies such as uniform vertex and uniform edge sampling possess desirable statistical properties: the former yields unbiased samples of the population and the bias introduced by the latter is easily removed. However, these strategies are often rendered unfeasible because they require either a directory containing the list of all vertex (edge) ids, or an API that allows uniform sampling from the vertex (edge) space. Even when the space of possible vertex (edge) ids is known, its occupancy is usually so low that querying randomly generated ids is expensive. An alternate, cheaper, way to sample a network is via a random walk (RW). A RW samples a network by moving a particle (walker) from a vertex to a neighboring vertex. It is applicable to any network where we can query the edges connected to a given vertex. Furthermore, RWs share some of the desirable properties of uniform edge sampling (i.e., easy bias removal, accurate estimation of characteristics such as the tail of the degree distribution).

On one hand, a great deal of research has focused on designing sampling methods for *undirected networks* using RWs [9, 24]. Ribeiro and Towsley proposed Frontier Sampling (FS), an n-dimensional random walk that uses n *coupled* random walkers. This method yields more accurate estimates than the uniform RW and also outperforms the use of n independent walkers. In the presence of disconnected or loosely connected components, FS is even better suited than the uniform RW and independent RWs to sample the tail of the degree distribution of the graph. On the other hand, few works have focused on developing tools for characterizing *directed networks* in the wild. A network is said to be directed when edges are not necessarily reciprocated. Characterizing directed networks through crawling becomes challenging when only outgoing edges from a node are visible (incoming edges are hidden): unless all vertices have a directed path to all other vertices, a walker will eventually be restricted to a (strongly connected) component of the graph. Furthermore, classic RWs incur biases that can only be removed by conditioning on the entire graph structure. In [27],we addressed these issues by proposing Directed Unbiased Random Walk (DURW), a random walk sampling technique that performs degree-proportional jumps to obtain asymptotically unbiased estimates of the distribution of vertex labels on a directed graph.

In this work[1], we propose the Directed Unbiased Frontier Sampling (DUFS) method, that generalizes the FS and the DURW algorithms. Building on the ideas in [27], we extend Frontier Sampling to allow the characterization of networks regardless of whether they are undirected, directed with observable incoming edges, or directed with unobservable incoming edges. DUFS matches or

[1]Parts of this work are based on previous papers from the authors: [28] and [27].



(a) soc-Slashdot0902 (b) livejournal-links

Fig. 1. Comparison between proposed method (DUFS) and previous state-of-the-art respectively for visible and for invisible incoming edges scenarios; (a) NRMSE ratios between DUFS ($w = 1, b = 10$) and FS ($b = 10$) of the estimated joint in- and out-degree distribution on the soc-Slashdot0902 dataset; (b) NRMSEs associated with DUFS and DURW of the estimated out-degree distribution on the livejournal-links dataset.

exceeds the performances of FS and DURW[2]. This is illustrated in Figure 1. Methods' parameters (w and b), simulation setup, datasets and the error metric (normalized root mean square error) will be described in Section 5.1.

Contributions. Our main contributions are as follows:

(1) *Directed Unbiased Frontier Sampling (DUFS)*: we propose a new algorithm based on multiple coordinated random walks that extends Frontier Sampling (FS) to directed networks. DUFS generalizes FS and DURW.
(2) *More accurate estimator for vertex label distribution*: when the number of walkers is a large fraction of the number of random walk steps (e.g., 10%), we lose a considerable source of information by not accounting for the walkers initial locations as observations. We introduce a new estimator that combines these observations with those made during the walks to produce better estimates.
(3) *Practical recommendations*: we investigate the impact of the number of walkers and the probability of jumping to an uniformly chosen vertex (controlled via a parameter called random jump weight) on DUFS estimation errors, given a fixed budget. By increasing the number of walkers the sequence of sampled edges approaches the uniform distribution faster, but this also increases the fraction of the budget spent to place the walkers in their initial locations. Moreover, increasing the random jump weight favors sampling vertex labels with large probability masses, which translates into more accurate estimates for these labels, but worse estimates for those in the tail. We study these trade-offs through simulations and propose guidelines for choosing DUFS parameters.
(4) *Comprehensive evaluation*: we compare DUFS against other random walk-based methods w.r.t. estimation errors when applied on directed networks, both when incoming edges are directly observable and when they are not. In the first scenario, in addition to some graph properties evaluated in previous works, we evaluate DUFS performance on estimating joint in- and out-degree distributions, and on estimating distribution of group memberships among the 10% largest degree nodes.

[2]The software and all results presented in this work are available at http://bitbucket.com/after-acceptance.

(5) *Theoretical analysis:* we derive expressions for the normalized mean squared error associated with uniform vertex and uniform edge sampling on power law networks and show that in both cases the error behaves asymptotically as a power law function of the observed degree. This helps explain our evaluation results.

Outline. Definitions are presented in Section 2. In Section 3, we review FS and DURW methods. In Section 4, we propose the Directed Unbiased Frontier Sampling (DUFS) algorithm (along with some estimators), which generalizes the previous methods. We investigate the impact of DUFS parameters on estimation accuracy of degree distributions and vertex label distributions respectively in Sections 5 and 6, providing practical guidelines on how to set them. A comparison against other random walk techniques is also provided. Section 7 discusses the performance of DUFS when the uniform vertex sampling mechanism is faulty. We present some related work and present our conclusions in Sections 8 and 9, respectively.

2 DEFINITIONS

In what follows we present some definitions. Let $G_d = (V, E_d)$ be a labeled directed graph representing the network graph, where V is a set of vertices and E_d is a set of ordered pairs of vertices (u, v) representing a connection from u to v (a.k.a. edges). We refer to an edge (u, v) as an *in-edge* with respect to v and an *out-edge* with respect to u. The *in-degree* and *out-degree* of a vertex u in G_d are the number of distinct edges respectively into and out of u. We assume that each vertex in G_d has at least one edge (either an in-edge or an out-edge). Some networks can be modeled as undirected graphs. In this case, G_d is a symmetric directed graph, i.e., $(u, v) \in E_d$ iff $(v, u) \in E_d$.

Let $\mathcal{L}_v$ be a finite set of vertex labels. We associate a set of labels (possibly empty) to each vertex, $\mathcal{L}_v(v) \subseteq \mathcal{L}_v$, $\forall v \in V$.

Input scenarios

When performing a random walk, we assume that a walker retrieves the out-edges of node where it resides by performing a query and that vertices are distinguishable. We define two scenarios depending on whether the walker can also retrieve in-edges. In the *first scenario* (both out- and in-edges can be retrieved) it is possible to move the walker over any edge regardless of the edge direction (if the edge is $(u, v) \in E_d$ a walker can move from u to v and vice versa). In this case, the walker can be seen as moving over $G = (V, E)$, an undirected version of G_d, i.e., $E = \{(u, v) : (u, v) \in E_d \vee (v, u) \in E_d\}$. Define $\deg(v) = |\{(u, v) : (u, v) \in E\}|$. Let $\mathrm{vol}(S) = \sum_{\forall v \in S} \deg(v)$, $\forall S \subseteq V$, denote the volume of the set of vertices in $S \subseteq V$.

In the *second scenario* (only out-edges are directly observable), we can build on-the-fly an undirected graph G_u based on the out-edges that have been sampled. Note that G_u is not an undirected version of G_d as some of the in-edges of a node may not have been observed. By moving the walker over G_u – possibly traversing edges in G_d in the opposite direction – we can compute its stationary behavior and thus, remove the bias by accounting for the probability that each observation appears in the sample.

While this has been mostly overlooked by other works in the literature, we emphasize that, in either scenario, it is useful to keep track of some variant of the observed graph during the sampling process. Storing information about visited nodes in memory saves resources that would be consumed to query those nodes in subsequent visits – i.e., revisiting a node has no cost. The specific variant of the observed graph to be stored will be described in the context of two random walk-based methods in the following section.

ALGORITHM 1: Frontier Sampling (FS)

Input: budget per walker b
$n \leftarrow B/(c+b)$ {c is the cost of uniform vertex sampling};
$i \leftarrow 0$ {i is step counter};
Initialize $L = (v_1, \ldots, v_n)$ with n randomly chosen vertices (uniformly);
repeat
  Select $u \in L$ with probability $\deg(u) / \sum_{\forall v \in L} \deg(v)$;
  Select an edge (u, v), uniformly at random;
  Replace u by v in L and add (u, v) to sequence of sampled edges;
  $i \leftarrow i + 1$;
until $i \geq B - nc$;

3 BACKGROUND

In what follows, we review a representative random-walk based method proposed for each of the two scenarios proposed in Section 2. First, we describe the Frontier Sampling algorithm proposed in [28], an n-dimensional random walk that benefits from starting its walkers at uniformly sampled vertices. This technique can be applied to undirected graphs and to directed graphs provided that the edges coming into a node are observable. Then, we describe the Directed Unbiased Random Walk algorithm we proposed in [27], that adapts a single random walk to directed graphs when incoming edges are not directly observable. The goal of these methods is to obtain samples from a graph, which are then used for inferring graph characteristics via an estimator. An *estimator* is a function that takes a sequence of observations (sampled data) as input and outputs an estimate of an unknown population parameter (graph characteristic).

3.1 Frontier Sampling: a multidimensional random walk for undirected networks

In essence, *Frontier Sampling* (FS) is a random walk-based algorithm for sampling and estimating characteristics of an undirected graph. FS performs n *coordinated* random walks on the graph. One of the advantages of using multiple walkers is that they can cover multiple connected components (when they exist), while a single walker is restricted to one component in the absence of a random jump or restart mechanism. However, when random walks are independent (not coordinated) the number of samples obtained from a component is proportional to the number of walkers in that component. Therefore, the probability of sampling an edge in steady state will differ for different components, unless the number of walkers in each component is set to be proportional to its volume. Unfortunately, initializing the walkers in such a way requires knowing the component volumes in advance, which cannot be done in practice. By coordinating multiple random walkers, FS is able to sample edges uniformly at random in steady state regardless of how the walkers are initially placed.

Algorithm 1 describes FS. There are three parameters, the total sampling budget B, the initial cost of placing a walker $c \geq 1$ and the average number of new nodes sampled by a walker b. The initial walker locations are chosen uniformly at random over the vertex set. Note that the number of walkers is taken to be $n = \lfloor B/(c+b) \rfloor$, that the cost of taking a random walk step is one (except for previously sampled nodes) and that the cost of initially placing a walker, c, can be greater than one because uniform vertex sampling is often expensive. FS maintains a list L of n vertices representing the locations of the n walkers. At each step, a walker is chosen from L in proportion to the degree of the node where it is currently located. The walker then moves from u to an adjacent vertex v.

The Frontier sampling process is equivalent to the sampling process of a single random walker over the n-th Cartesian power of G, $G^n = (V^n, E_n)$, where

$$V^n = \{(v_1, \ldots, v_n) \,|\, v_1 \in V \wedge \cdots \wedge v_n \in V\}$$

is the n-th Cartesian power of V. For all $\mathbf{v}, \mathbf{u} \in V^n$, $(\mathbf{v}, \mathbf{u}) \in E_n$ if there exists an index $i \in \{1, \ldots, n\}$ such that $(v_i, u_i) \in E$ and $u_j = v_j$ for $j \in \{1, \ldots, n\}/\{i\}$ [28, Lemma 5.1]. For this reason, Frontier Sampling can be thought of as an n-dimensional random walk (see Fig. 2).

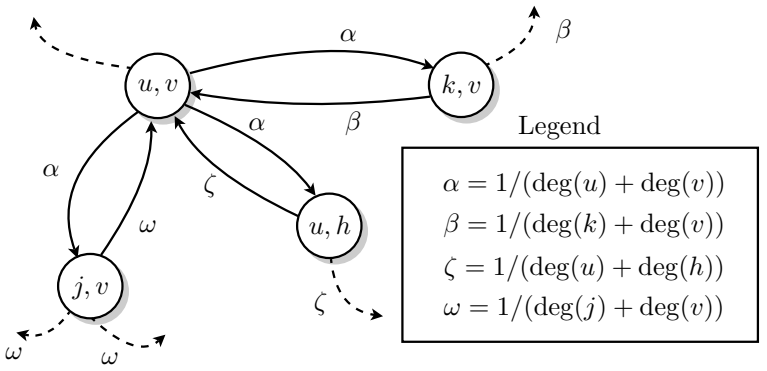


Fig. 2. Illustration of the Markov chain associated to FS with dimension $n = 2$.

Let $L_t = (v_1, \ldots, v_n)$ denote the state of FS before the t-th step, $t = 1, \ldots$. Theorem 3.1 establishes key statistical properties of Frontier Sampling. A more complete version of this theorem is presented and proved in [28, Theorem 5.2].

Theorem 3.1. *Recall that G is an undirected graph. If G is connected and non-bipartite, then the stationary behavior FS exhibits the following properties:*

(I) *sampled edges form a stationary sequence and their marginal distribution is uniform on E,*
(II) *$L_\infty = (v_1, \ldots, v_n)$ has the unique distribution*

$$\pi_{\mathbf{v}} = \frac{\sum_{i=1}^{n} \deg(v_i)}{n|V|^{n-1}\mathrm{vol}(V)}, \quad \textit{for } \mathbf{v} \in V^n.$$

Using FS samples to estimate vertex label distributions is simple when the input corresponds to the first scenario described in Section 2. The probability of sampling a given node is proportional to its undirected degree in G. Hence, each sample must be weighted inversely proportional to the respective node's undirected degree. Storing the undirected version of the observed graph along with labels associated with sampled nodes allows the sampler to avoid having to pay the cost of revisiting a node.

Conversely, when incoming edges are not observed, there is a less straightforward way to adapt Frontier Sampling, which we propose in Section 4.

3.2 Directed Unbiased Random Walk: a random walk adapted for directed networks with unobservable in-edges

The presence of hidden incoming edges but observable outgoing edges makes characterizing large directed graphs through crawling challenging. Edge (u, v) is a hidden incoming edge of node v if (u, v) can only be observed from node u. For instance, in Wikipedia we cannot observe the

edge ("Columbia Records", "Thomas Edison") from Thomas Edison's wiki entry (but this edge is observable if we access the Columbia Records's wiki entry).

These hidden incoming edges make it impossible to remove the bias incurred by walking on the observed graph, unless we crawl the entire graph. Moreover, there may not even be a directed path from a given node to all other nodes. Graphs with hidden outgoing edges but observable incoming edges exhibit essentially the same problem. In [27], we proposed the Directed Unbiased Random Walk (DURW) algorithm, which obtains asymptotically unbiased estimates of vertex label densities on a directed graph with unobservable incoming edges. Our random walk algorithm resorts to two main principles to achieve unbiased samples and reduce variance:

- *Backward edge traversals*: in real-time we construct an undirected graph G_u using the nodes that are sampled by the walker on the directed graph G_d. The role of the undirected graph is to guarantee that, at the end of the sampling process, we can approximate the probability of sampling a node, even though in-edges are not observed. The random walk proceeds in such a way that its trajectory on G_d is consistent with that of a random walk on G_u. The walker is allowed to traverse some of the edges in G_d in a reverse direction. However, we prevent some of the observed edges to be traversed in the reverse direction by not including them in G_u. More precisely, once a node v is visited at the i-th step, no in-edges to v observed at step $j > i$ (by visiting nodes w such that $(w, v) \in E_d$) are added to G_u. This is important to reduce the random walk transient and thus, reduce estimation errors.
- *Degree-proportional jumps*: the walker makes a limited number of random jumps to guarantee that different parts of the directed graph are explored. In DURW, the probability of randomly jumping out of a node v, $\forall v \in V$, is $w/(w + \deg(v))$, $w > 0$. This modification is based on the following observation: let G_u be a weighted undirected graph formed by adding a *virtual* node σ such that σ is connected to all nodes in V with edges having weight w. All remaining edges have unit weight. In a weighted graph a walker transverses a given edge with probability proportional to the weight of this edge. The steady state probability of visiting a node v on G_u is $(w + \deg(v))/(\mathrm{vol}(V) + w|V|)$. Similar to the cost of placing a FS walker through uniform vertex sampling, we assume that each random jump incurs cost $c \geq 1$.

The DURW algorithm. DURW is a random walk over a *weighted undirected connected graph* $G_u = (V, E_u)$, which is built on-the-fly. We build an undirected graph using the underlying directed graph G_d and the ability to perform random jumps. Let $G^{(i)} = (V, E^{(i)})$ denote the undirected graph constructed by DURW at step i, where V is the node set and $E^{(i)}$ is the edge set. Denote by $G_u \equiv \lim_{i\to\infty} G^{(i)}$. In what follows we describe the construction of $G^{(i)}$.

Let $\mathcal{N}(v)$ denote the set of out-edges of a node v in G_d. To simplify our exposition, we include a virtual node σ in the constructed graph, which represents a random jump. Let $\mathcal{S}^{(i)} = \{s_1, \ldots, s_i\}$ be the set of nodes from $V \cup \{\sigma\}$ sampled by the random walk up to step i, where s_j denotes the node on which the walker resides at step j. The walker starts at node $s_1 \in V$. We initialize $G^{(1)} = (V, E^{(1)})$, where $E^{(1)} = \mathcal{N}(s_1) \cup \{(u, \sigma) : \forall u \in V\}$, where $\{(u, \sigma) : \forall u \in V\}$ is the set of all undirected virtual edges to node σ. Let

$$W(u, v) = \begin{cases} w & \text{if } u = \sigma \text{ or } v = \sigma \\ 1 & \text{otherwise} \end{cases}$$

denote the weight of edge (u, v), $\forall (u, v) \in E^{(i)}$. The next node, s_{i+1}, is selected from $E^{(i)}$ with probability $W(s_i, s_{i+1}) / \sum_{\forall (s_i, v) \in E^{(i)}} W(s_i, v)$. Upon selecting s_{i+1} we update $G^{(i+1)} = (V, E^{(i+1)})$, where

$$E^{(i+1)} = E^{(i)} \cup \mathcal{N}'(s_{i+1}), \tag{1}$$

and

$$\mathcal{N}'(s_{i+1}) = \{(s_{i+1}, v) : \forall (s_{i+1}, v) \in \mathcal{N}(s_{i+1}) \text{ s.t. } v \notin \mathcal{S}^{(i)}\}$$

is the set of all edges (u, v) in $\mathcal{N}(s_{i+1})$ where node $v \notin \mathcal{S}^{(i)}$. Note that $\mathcal{N}'(s_{i+1}) \subseteq \mathcal{N}(s_{i+1})$. By using $\mathcal{N}'(s_{i+1})$ instead of $\mathcal{N}(s_{i+1})$ in equation (1) we guarantee that no node in $\mathcal{S}^{(i)}$ changes its degree, i.e., $\forall v \in \mathcal{S}^{(i)}$ the degree of v in $G^{(i)}$ is also the degree of v in G_u. Thus, we comply with the requirement that once a node v, $\forall v \in V$, is visited by the RW no edge can be added to G_u with v as an endpoint.

In the actual implementation, it is only necessary to keep track of nodes in $\mathcal{S}^{(i)} \cup \bigcup_{v \in \mathcal{S}^{(i)} \setminus \{\sigma\}} \mathcal{N}(v)$ and the edges in E_d leaving each node $v \in \mathcal{S}^{(i)} \setminus \{\sigma\}$. In fact, while the virtual node σ is connected to all nodes in V, the sampler does not have access to the identities of nodes other than the ones that were already observed. In order to estimate vertex label distributions from DURW observations, we weight samples in proportion to the reciprocal of the probability that the corresponding vertices are visited by a random walk in G_u, in steady state. Storing the labels associated with nodes in $\mathcal{S}^{(i)} \setminus \{\sigma\}$ saves the cost of querying repeated nodes.

4 GENERALIZING FS: A NEW METHOD APPLICABLE REGARDLESS OF IN-EDGES VISIBILITY

This section is divided into two parts. In Section 4.1 we propose the Directed Unbiased Frontier Sampling (DUFS) method, which generalizes FS to allow estimation also on directed graphs with unobservable in-edges (second scenario described in Section 2). DUFS also generalizes DURW: the latter is a special case of DUFS where the number of walkers is one. Next, in Section 4.2, we describe two ways to estimate node label distributions from DUFS samples. The first uses only on the observations collected when moving the walkers. The second is a new estimator we propose to leverage information contained in the initial walker locations in addition to the walker subsequent steps.

4.1 Directed Unbiased Frontier Sampling

Like FS, the Directed Unbiased Frontier Sampling (DUFS) samples a network through n coordinated walkers. At each step, it selects a walker in proportion to the degree of the node where it currently resides.

Similarly to the Directed Unbiased Random Walk, it constructs an undirected graph $G_u = (V, E_u)$ in real-time that allows *backward edge traversals*. Denote by $G^{(i)} = (V, E^{(i)})$ the undirected graph constructed by DUFS at step i. DUFS does not include edges in $G^{(i)}$ that would cause walkers to have a view of the graph that is inconsistent with the view at a previous point in time. In other words, when node u is visited for the first time at step i, u is inserted in $G^{(i)}$ along with all edges $(u, v) \in E_d$ such that v has not been sampled. Thus, the degree of u is fixed in $G^{(j)}$, for all $j \geq i$.

It may seem that there is no need to include *degree-proportional jumps* to visit different graph components when a large number of walkers are initially spread throughout the graph (e.g., on vertices chosen uniformly at random). However, including degree-proportional jumps in DUFS is extremely beneficial because it prevents walkers from being trapped when initially located on vertices whose out-degree is zero. More generally, it allows walkers to move from small volume to large volume components and, hence, obtain more samples among large degree nodes.

Algorithm 2 gives a high-level pseudo-code description of DUFS. At each step i, DUFS needs to keep track of $G^{(i)}$ for $i = 1, \ldots, B - nc$. In the extreme case where $n = B/c$, walkers are initialized but no budget is left to perform steps (i.e., $b = 0$). Thus, DUFS degenerates to uniform vertex sampling. When the underlying graph is symmetric and the jump weight is $w = 0$, it becomes FS.

ALGORITHM 2: Directed Unbiased Frontier Sampling (DUFS)

Input: budget per walker b, random jump weight w
$n \leftarrow B/(c+b)$ {n is the number of walkers};
$i \leftarrow 0$ {i is the current number of steps};
Initialize $L = \{v_1, \ldots, v_n\}$ with n randomly chosen vertices (uniformly);
repeat
  Select $v \in L$ with probability $\frac{w+\deg(v)}{nw+\sum_{\forall v_j \in L} \deg(v_j)}$;
  Sample $p \sim \text{Uniform}(0,1)$;
  if $p < \frac{w}{w+\deg(v)}$ **then**
    Select a vertex $v \in V$ uniformly at random;
  else
    Select an outgoing edge of v, (v, v'), uniformly at random;
  end
  Replace v by v' in L and add (v, v') to sequence of sampled edges;
  $i \leftarrow i + 1$;
until $i \geq B - nc$;

4.2 Estimation

In this section we describe two estimators of vertex label distributions from samples obtained by DUFS. These estimators generalize estimators proposed for FS and DURW. For a description of estimators of edge label distribution and other graph characteristics, please refer to [28].

4.2.1 Vertex Label Distribution: random edge-based estimator. Let s_i denote the i-th node visited by DUFS, $i = 1, \ldots, t$, $t \leq B - nc$. Let θ_ℓ be the fraction of nodes in V with label $\ell \in \mathcal{L}_v$. Let $\pi(v)$ be the steady state probability of sampling node v in G_u, $\forall v \in V$. The vertex label distribution is estimated at step t as

$$\hat{\theta}_\ell = \frac{1}{n} \sum_{i=1}^{t} \frac{\mathbb{1}\{\ell \in \mathcal{L}_v(v)\}}{\hat{\pi}(s_i)}, \qquad \ell \in \mathcal{L}_v,\ t = 1, \ldots, B - nc, \tag{2}$$

where $\mathbb{1}\{P\}$ takes value one if predicate P is true and zero otherwise, and $\hat{\pi}(s_i)$ is an estimate of $\pi(s_i)$: $\hat{\pi}(s_i) = (w + \deg(s_i))S$. Here $\deg(v)$ is the degree of v in $G^{(\infty)}$ and

$$S = \frac{1}{t} \sum_{i=1}^{t} \frac{1}{w + \deg(s_i)} . \tag{3}$$

The following theorem states that $\hat{\pi}(s_i)$ is asymptotically unbiased.

Theorem 4.1. *$\hat{\pi}(s_i)$ is an asymptotically unbiased estimator of $\pi(s_i)$.*

Proof. To show that $\hat{\pi}(s_i)$ is asymptotically unbiased, we first note that the limit $\lim_{t\to\infty} E^{(t)} = E^{(\infty)}$ exists, since after visiting all vertices we do not include any additional edges. We then invoke Theorem 4.1 of [28], yielding $\lim_{t\to\infty} S = |V|/(|E^{(\infty)}| + |V|w)$ almost surely. Thus, $\lim_{t\to\infty} \hat{\pi}(s_i) = \pi(s_i)$ almost surely. Taking the expectation of (2) in the limit as $t \to \infty$ yields $E[\lim_{t\to\infty} \hat{\theta}_\ell] = \theta_\ell$, which concludes our proof. □

4.2.2 Vertex Label Distribution: leveraging information from walkers' initial locations. Note that the estimator presented in (2) does not make use of information associated with the initial set of nodes on which the walkers are placed. When the number of walkers is large this results in the

loss of a considerable amount of statistical information. However, including these observations is challenging because subsequent observations from random walk steps are not independent of the initial observations. Moreover, the normalizing constant for the random walk observations is no longer given by (3), since the degree distribution estimates also depend on the information contained in the random vertex samples.

In this section, we derive a new estimator that circumvents these problems by approximating the likelihood of random walk samples by that associated with random edge sampling. We call it the *hybrid estimator* because it combines observations from initial walker locations and from random walks steps. The hybrid estimator significantly improves the estimation accuracy for labels associated with large probability masses.

Let us index the vertex labels $\mathcal{L}_v$ from 1 to W, where $W = |\mathcal{L}_v|$. We refer to the sum $\deg(v) + w$ in DUFS as the *random walk bias* for vertex $v \in V$. To keep the notation simple, we assume that each vertex has exactly one label and that random walk biases take on integer values in $[1, \ldots, Z]$. Denote the vertex label distribution as $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_W)$. Let n_i denote the number of walkers starting on label i nodes and $m_{i,j}$ the number of subsequent observations of label i and bias j nodes. We approximate random walk samples in DUFS by uniform edge samples from G_u. Experience from previous studies shows us that this approximation works very well in practice. Hence, the likelihood function given the samples $\mathbf{n} = \{n_i : i = 1, \ldots, W\}$ and $\mathbf{m} = \{m_{i,j} : i = 1, \ldots, W \text{ and } j = 1, \ldots, Z\}$ is expressed as

$$L(\boldsymbol{\theta}|\mathbf{n}, \mathbf{m}) = \frac{\prod_i \theta_i^{n_i} \prod_k (k\theta_{i,k})^{m_{i,k}}}{\left(\sum_{s,t} t\theta_{s,t}\right)^M}. \tag{4}$$

The maximum likelihood estimator $\theta^\star$ is the value of θ that maximizes (4) subject to $0 \le \theta_i \le 1$ and $\sum_i \theta_i = 1$. This defines a non-convex optimization problem with constraints. However, we can turn this optimization problem into an unconstrained problem using the reparameterization $\theta_i = e^{\beta_i} / \sum_k e^{\beta_k}$ for $i = 1, \ldots, W$. As shown in Appendix A, the partial derivatives of the resulting objective function are given by

$$\frac{\partial \mathcal{L}(\boldsymbol{\beta}|\mathbf{n}, \mathbf{m})}{\partial \beta_i} = n_i + m_i - \frac{N e^{\beta_i}}{\sum_j e^{\beta_j}} - \frac{M e^{\beta_i} m_i/\mu_i}{\sum_s e^{\beta_s} m_s/\mu_s}, \qquad i = 1, \ldots, W, \tag{5}$$

where $m_i = \sum_k m_{i,k}$ and $\mu_i = \sum_k m_{i,k}/k$. Setting one of the variables to a constant (say, $\beta_W = 1$) for identifiability and then using the gradient descent procedure to change the remaining variables according to (5) is guaranteed to converge provided that we make small enough steps.

An interesting interpretation of (5) is obtained by setting the derivatives to zero and substituting back $\theta_i = e^{\beta_i} / \sum_k e^{\beta_k}$:

$$\theta_i^\star = \frac{n_i + m_i}{N + M \frac{m_i/\mu_i}{\sum_s \theta_s^\star m_s/\mu_s}}, \qquad i = 1, \ldots, W. \tag{6}$$

According to (6), the estimated fraction of nodes with label i is the total number of times label i was observed (i.e., $n_i + m_i$) normalized by sum of (i) the number of random vertex samples and (ii) the number of random edge samples weighted by the probability of sampling label i from one random edge sample. In the limit as N and M go to infinity, we can show that $\boldsymbol{\theta}^\star = \boldsymbol{\theta}$ is a solution, but we cannot prove that it is unique or that $\boldsymbol{\theta}^\star$ converges to $\boldsymbol{\theta}$. Hence, we cannot prove that $\boldsymbol{\theta}^\star$ is asymptotically unbiased.

The system of non-linear equations determined by (6) cannot be solved directly, but can be tackled by Expectation Maximization (EM). In this case, the term $\sum_s \theta_s^\star m_s/\mu_s$ in the denominator is replaced by its expected value given θ_i's from the previous iteration. Based on the same idea, if

we replace $\sum_s \theta_s^\star m_s/\mu_s$ with an edge sampled-based estimator $\hat{d}$ for the average degree in G_u, we obtain the following non-recursive variant of the hybrid estimator,

$$\hat{\theta}_i = \frac{n_i + m_i}{N + M\frac{m_i}{\mu_i \hat{d}}}, \qquad i = 1, \ldots, W, \tag{7}$$

where $\hat{d} = M/(\sum_i \mu_i)$. Theorem 4.1 below states the conditions under which $\hat{\theta}_i$ is asymptotically unbiased (see appendix for proof). In practice, we find no significant difference between $\theta_i^\star$ and $\hat{\theta}_i$, except when the number of walkers N is very large and the jump weight w is very small. For those cases, $\theta_i^\star$ tends to be slightly more accurate than $\hat{\theta}_i$ for small values of i, which in some applications may justify the additional computational cost of executing gradient descent or EM.

THEOREM 4.1. *Let $N = \alpha B$ and $M = (1-\alpha)B$, for some $0 < \alpha < 1$. In the limit as $B \to \infty$, the estimator $\hat{\theta}_i$ is an unbiased estimator of θ_i.*

In the special case where the label is the undirected degree itself, we have $\mu_i = m_i/i$. Hence, eq. (7) reduces to

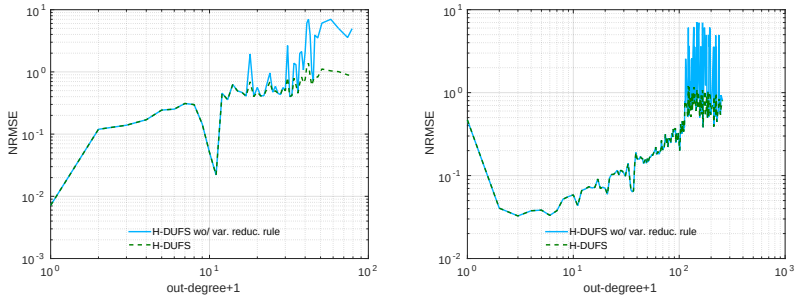
$$\bar{\theta}_i = \frac{n_i + m_i}{N + Mi/\hat{d}}, \tag{8}$$

where $\hat{d}$ is the estimated average degree. When the average degree is known, we can show that $\bar{\theta}_i$ is unbiased and, moreover, the minimum variance unbiased estimator (MVUE) of θ_i (see appendix for proof).

When $n_i > 0$ but $m_i = 0$, the estimator in eq. (7) reduces to $\hat{\theta}_i = n_i/N$, which is essentially the MLE for random vertex sampling. It is well known that this estimator is not nearly as accurate as a random walk based estimator for large out-degree values with small probability mass. In some sense, the estimator $\hat{\theta}_i = n_i/N$ does not account for the fact that the number of random walk samples is zero. As a result, mass estimates for large out-degrees tend to have very large variance when no random walk samples are observed. Fortunately, we find that the following heuristic rule can drastically reduce the estimator variance in these cases.

Variance reduction rule. If no random edge samples are observed for out-degree i, we set the estimate $\hat{\theta}_i = 0$. This implies that we ignore any random vertex samples seen of nodes that have out-degree i. While this clearly results in a biased estimate, as the budget per walker b goes to infinity, the probability of invoking this rule goes to zero. Hence, it produces an asymptotically unbiased estimate. This rule can be interpreted as a combination of vertex-based and random edge-based estimates in proportion to the reciprocals of their estimated variances. That is, when no random edge samples are observed for a given out-degree, the corresponding estimated variance is zero and hence, random vertex samples should be ignored. We note that the converse rule (i.e., set $\hat{\theta}_i = 0$ if no random vertex samples were observed) would not perform well, as the probability of sampling large out-degrees with random vertex sampling is very small.

We simulate DUFS on several datasets and compare the results obtained with the hybrid estimator when the rule is used and when it is not. Simulation details, datasets and the error metric (normalized root mean square error) will be described in Section 5.1. Figures 3(a-b) show representative results of the impact of the rule when estimating out-degree distributions using DUFS in conjunction with the hybrid estimator on two network datasets (averaged over 1000 runs). The results show that the rule consistently reduces estimation error in the distribution tail without affecting estimation quality for small values of i.



(a) p2p-Gnutella31 (b) web-Stanford

Fig. 3. **(visible in-edges)** Effect of variance reduction rule on NRMSE, when $B = 0.1|V|$ and $c = 1$. Using information contained in uniform vertex samples can increase variance for large out-degree estimates. However, the proposed rule effectively controls for that effect without decreasing head estimates accuracy.

In-degree distribution: impossibility result. The fact that long random walks are often approximated by random edge sampling brings up the question of whether they can be used to estimate in-degree distributions when the in-degree is not observed directly. Under random edge sampling, the number of observed edges pointing to a node is binomially distributed and a maximum likelihood estimator can be derived for estimating the in-degree distribution. This problem is related to the set size distribution estimation problem, where elements are randomly sampled from a collection of non-overlapping sets and the goal is to recover the original set size distribution from samples. In addition to in-degree distribution in large graphs, this problem is related to the uncovering of TCP/IP flow size distributions on the Internet.

In [23], we derive error bounds for the set size distribution estimation problem from an information-theoretic perspective. The recoverability of original set size distributions presents a sharp threshold with respect to the fraction of elements sampled from the sets. If this fraction lies below the threshold, typically half of the elements in power-law and heavier-than-exponential-tailed distributions, then the original set size distribution is unrecoverable. Please refer to [23, Theorem 2] for details.

5 RESULTS ON DEGREE DISTRIBUTION ESTIMATION

Here we focus on the estimation of degree distributions on directed networks. This section is divided in four parts. In Section 5.1, we investigate the impact of DUFS parameters on estimation accuracy. We then compare DUFS against other random walk-based methods when both outgoing and incoming edges are visible in Section 5.2. In Section 5.3, we perform a similar comparison when only out-edges are visible. Last, in Section 5.4 we provide some analysis to explain the relationship observed between the NRMSE and the out-degree (in-degree) in the results. We will refer to the edge-based estimator defined in (2) and the hybrid estimator defined in (6) as E-DUFS and H-DUFS respectively.

The 15 directed network datasets in our evaluation were obtained from Stanford's SNAP [17]. These datasets describe the topology of a variety of social networks, communication networks, web graphs, one Internet peer-to-peer networks and one product co-purchasing networks. We found it informative to extract the largest strongly connected component of each directed network and to apply our methods to the resulting datasets – hereby referred to as LCC datasets – as well as to the original datasets. Figure 4 shows the out-degree probability mass function (p.m.f.) for each network, along with the out-degree p.m.f. for the corresponding LCC dataset. We opt to show the p.m.f.

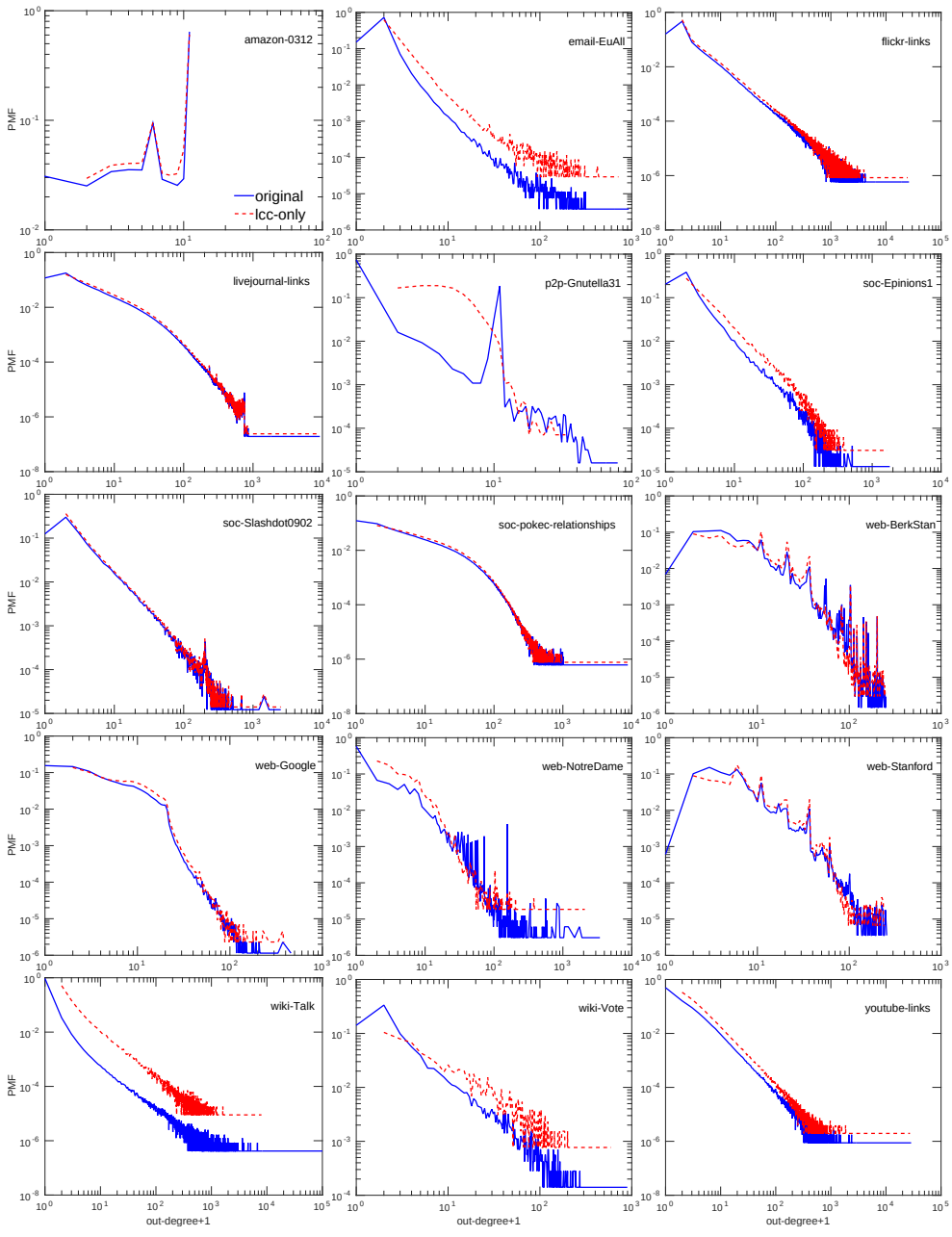


Fig. 4. Out-degree probability mass function (p.m.f.) for each network and its largest strongly connected component (LCC). A large difference between these p.m.f.s suggests it is beneficial to use multiple walkers and/or random jumps.

instead of the complementary cumulative distribution function (CCDF) because the estimation task in this work is defined in terms of the p.m.f.'s. Defining the estimation task in terms of the CCDF would give H-DUFS an unfair advantage, as we will explain in Section 5.2.

Simulations consist of sampling the network until a budget $B = 0.1|V|$ (i.e., 10% of the number of vertices) is depleted. Note that budget is decremented when walkers are initially placed and each time one of them moves to a vertex and when they perform random jumps. We construct an undirected graph in the background throughout each simulation. As a result, we assume that the cost to revisit a vertex is zero, even if this visit occurs due to a random jump[3].

When both outgoing and incoming edges are observable, random walks disregard edge direction, and move as if the network is undirected. In this scenario, we focus either on the estimation of the marginal out- and in-degree distributions or the joint distribution. The methods we investigate here can be used to estimate other node label distributions. For instance, if the underlying network is undirected, we can estimate the (undirected) degree distribution or even non-topological properties, such as the distribution of user nationalities in a social network. In the light of the impossibility results described in the end of Section 4.2, we focus on out-degree distribution estimation when incoming edges are not directly observable.

Let $\boldsymbol{\theta} = \{\theta_i\}_{\forall i \in \mathcal{L}}$ denote the vertex label distribution, where θ_ℓ is the fraction of vertices with label ℓ. Denote by $\hat{\theta}_\ell$ the estimate for θ_ℓ. We use the normalized root mean square error (NRMSE) of $\hat{\theta}_\ell$ as the error metric, which is a normalized measure of the dispersion of the estimates, given by

$$\text{NRMSE}(\ell) = \frac{\sqrt{E[(\hat{\theta}_\ell - \theta_\ell)^2]}}{\theta_\ell}. \tag{9}$$

In the case of marginal in-degree (out-degree) distribution, we refer to in-degrees (out-degrees) smaller than the average as the *head* of the distribution. We refer to the top 1% largest in- (out-degree) values as the *tail* of the distribution.

5.1 Impact of DUFS parameters and practical guidelines

To provide some intuition on how the random jump weight w and the budget per walker b affect the accuracy of DUFS estimates, assume for now that we replace samples collected via random walks by uniform edge samples from the weighted undirected graph G_u. In this hypothetical scenario, the budget B is used to collect $n \geq 1$ uniform vertex samples and $B - nc$ uniform edge samples. Clearly, when the edge-based estimator defined in (2) is used, the most accurate vertex label distribution estimates are obtained by setting $n = 1$, or equivalently, $b = B - c$. Therefore, we focus on the case where the hybrid-estimator defined in (6) is used. In particular, consider estimation of the out-degree distribution.

For a given value of b, the number of uniform vertex samples will be $B/(c + b)$. For each of the remaining $B - B/(c + b)$ samples, a vertex v is sampled in proportion to $\deg(v) + w$, where $\deg(v)$ is the undirected degree of v in G_u. The choice of w and b impose, individually, a trade-off between estimation accuracy of the head and of tail of the distribution. For a fixed value of w, smaller values of b translate into better estimates of the head (and worse estimates of the tail) because we collect more (less) information about that region of the distribution from uniform vertex samples. For a fixed value of b, larger values of w also translate into more (less) accurate estimates of the head (tail), because random jumps are more likely to move a node to low in- and out-degree nodes (as they tend to occur more frequently).

[3]Note that the alternative, i.e. always taking c units off the budget per random jump, is unlikely to impact results significantly when $B = 0.1|V|$, since the vast majority of random jumps will find a non-visited node.

Table 1. Practical guidelines on setting H-DUFS parameters to obtain accurate head or tail estimates depending on in-edge visibility and vertex sampling cost c.

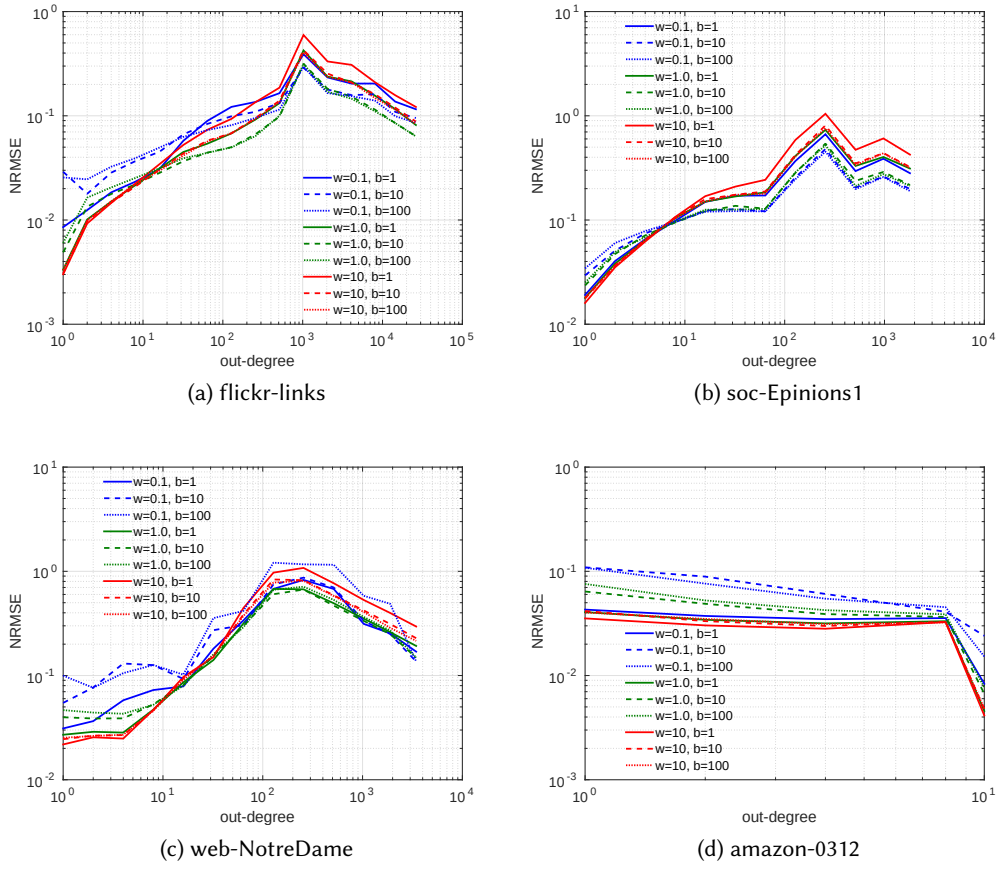
	uniform vertex sampling cost			
	$c = 1$		$c = 10$	
in-edges	visible	not visible	visible	not visible
most accurate for small out-degrees	$w = 10$ $b = 1$	$w = 10$ $b = 1$	$w = 1$ $b = 10^2$	$w = 10$ $b = 1$
most accurate for large out-degrees	$w = 1$ $b = 10$	$w = 1$ $b = 10, 10^2, 10^3$	$w = 0.1$ $b = 10^3$	$w = 0.1$ $b = 10, 10^2, 10^3$

In what follows, we observe through simulations that *despite the uniform edge sampling approximation, the previous intuition holds for H-DUFS head estimates, but not always for tail estimates.* In many cases, as we increase the number of walkers (i.e., decrease b) or increase w, we still obtain good estimates of the tail. This occurs because varying w or b changes the transition probability matrix that governs the sampling process, and thus, the sample distribution.

We simulate DUFS on each original network dataset for combinations of random jump weight $w \in \{0.1, 1, 10\}$ and budget per walker $b \in \{1, 10, 10^2, 10^3\}$ (1000 runs each). Values of w much smaller and much larger than these would be approximately equivalent to H-DUFS without jumps and random vertex sampling, respectively. Larger values of b would approximately correspond to DURW. We consider four scenarios that correspond to whether the incoming edges are directly observable or not and to two different costs of independent vertex sampling $c = 1$ or $c = 10$. Evaluating these parameter combinations is useful to establish practical guidelines for choosing H-DUFS parameters, which we summarize in Table 1. We observe that the estimation accuracy tends to be lower for extreme values of these parameters, suggesting that combinations other than the ones investigated here would not provide large accuracy gains (if any).

Visible in-edges, $c = 1$. Figures 5(a-c) show typical results when varying w and b. To avoid clutter, we show only estimates for powers of two (or the closest out-degree values) and omit results for $b = 10^3$ as they are similar to those for $b = 10^2$. Figure 5(d) shows similar results for amazon-0312, the dataset with the smallest maximum out-degree (max. is 10). Similarly to our intuition for uniform edge sampling, the NRMSE associated with the head increases with b and decreases with w, on virtually all datasets[4]. Also as expected, for a fixed values of w, $b = 1$ yields larger errors in the tail than $b \in \{10, 100\}$ (except for amazon-0312). However, contrary to the intuition for uniform edge sampling, $w = 1$ matches or outperforms $w = 0.1$ for (except for $b = 1$). This is best visualized in Figure 5d. This happens because setting $w = 1$ allows DUFS to sample regions with large probability mass (in this case, the head) and, at the same time, allows the sampler to move walkers from low volume to high volume components more often than $w = 0.1$. We also observe that $b = 10$ outperforms $b \in \{10^2, 10^3\}$ for $w \in \{0.1, 1\}$. Dataset amazon-0312 is the only dataset where $(w = 10, b = 1)$ obtained the best results over the entire out-degree distribution. As a side note, we observe that for most datasets used here, in log-log scale, the NRMSE grows approximately linearly on the out-degree up to a certain point and then starts to decrease, roughly linearly too. In Section 5.4 we explain why this is the case.

[4]For simplicity, the observations regarding the distribution head (tail) are based on the single smallest (largest) out-degree on each dataset. Similar conclusions are obtained when combining NRMSEs associated with several of the smallest (largest) out-degrees.

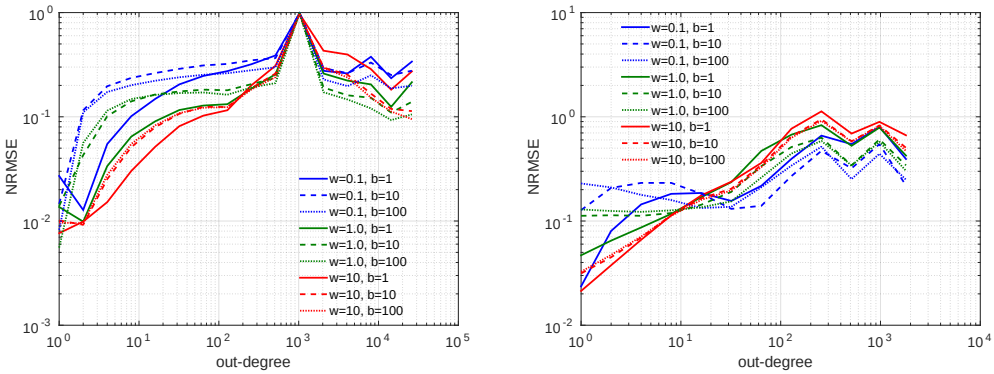


(a) flickr-links (b) soc-Epinions1

(c) web-NotreDame (d) amazon-0312

Fig. 5. (visible in-edges) Effect of DUFS parameters on datasets with many connected components, when $B = 0.1|V|$ and $c = 1$. Legend shows the average budget per walker (b) and jump weight (w). Trade-off shows that configurations that result in many random vertex samples, such as ($w = 10, b = 1$), yield accurate head estimates, whereas configurations such as ($w = 1, b = 10$) yield accurate tail estimates.

Invisible in-edges, $c = 1$. The results we obtained are similar to those obtained for the visible in-edge scenario, but NRMSEs tend to be larger. Figures 6(a,b) show typical results for different DUFS parameters, represented by two datasets (also shown in the previous figure). Once again, the intuition for uniform edge sampling holds for the distribution head: decreasing b and increasing w yield more accurate estimates for the smallest out-degrees. While $b = 1$ results in poor estimates for the largest out-degrees, our intuition regarding w does not hold true for the tail. More precisely, in most cases $w = 1$ outperforms $w = 0.1$ (one exception being dataset soc-Epinions1). As opposed to the visible in-edge scenario, increasing b tends to provide more accurate tail estimates for $w = 1$. We investigate this effect in Section 5.3. We find that, for a fixed w, larger values of b make the random walks jump more often, moving them from small volume components to large volume components, yielding better tail estimates.



(a) flickr-links (b) soc-Epinions1

Fig. 6. (invisible in-edges) Effect of DUFS parameters on datasets with many connected components, when $B = 0.1|V|$ and $c = 1$. Legend shows the average budget per walker (b) and jump weight (w). Configurations that result in many walkers which jump too often, such as ($w \geq 10, b = 1$) yield accurate head estimates, whereas configurations such as ($w = 1, b = 10^3$), yield accurate tail estimates.

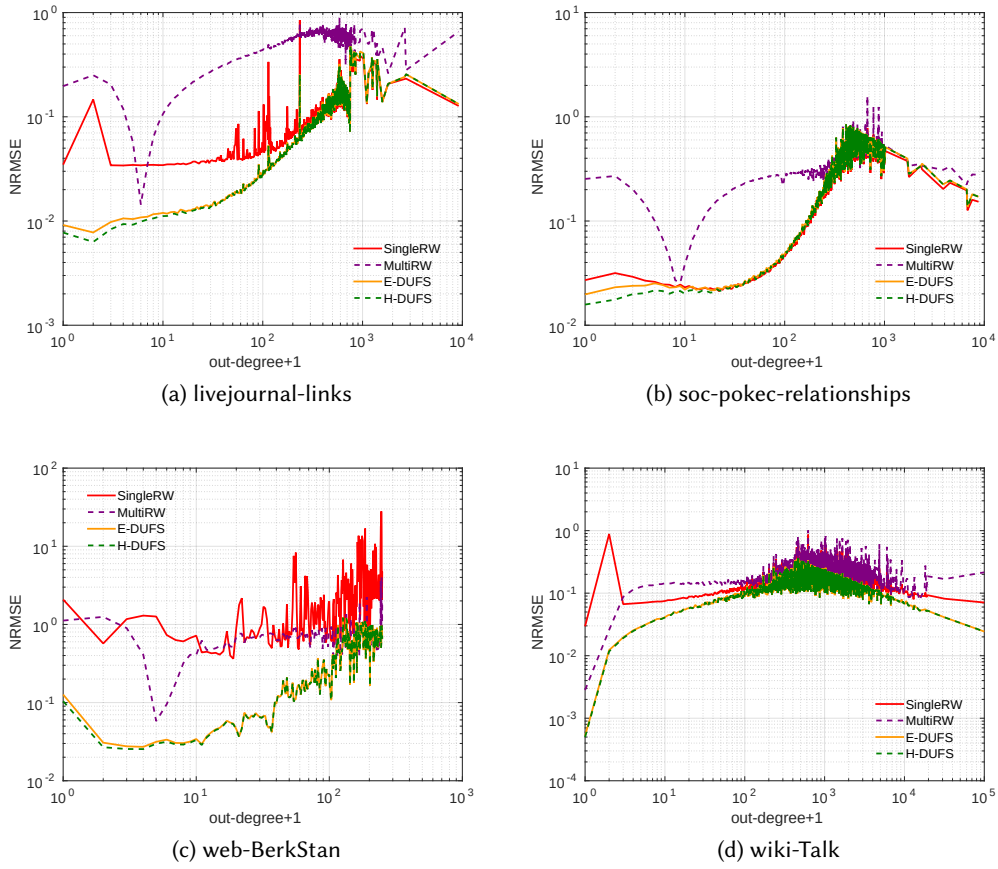
Visible in-edges, $c = 10$. Consider the case where the cost of obtaining independent vertex samples is large, more precisely, 10 times larger than the cost of moving a walker. It is no longer clear that using many walkers and frequent random jumps achieves the most accurate head estimates, as this could rapidly deplete the budget. In fact, we observe that setting $w = 10$ or $b = 1$ yields poor estimates for both the smallest and largest out-degrees. While increasing the jump weight w or decreasing b sometimes improves estimates in the head, it rarely does so in the tail. The best results for the smallest out-degrees are often observed when setting $w = 1$ and $b = 10$ or 10^2. On the other hand, setting ($w = 0.1, b = 10^3$) or ($w = 1, b = 10^2$) usually achieves relatively small NRMSEs for the largest out-degree estimates.

Invisible in-edges, $c = 10$. Unlike the scenario with visible in-edges, setting $w = 10$ and $b = 1$ often produces the most accurate estimates for the smallest out-degrees. This is because many of the datasets have nodes with no out-edges; these nodes can only be reached through a neighbor or through random vertex sampling. Conversely, the general trends for tail estimates are similar to those observed for the visible in-edges case: large values of w and small values of b yield less accurate estimates for the largest out-degree values. For $w = 1$, however, $b = 10^2$ often outperforms $b = 10^3$. On the other hand, for $w = 0.1$ there is little difference in the estimates for different values of b.

5.2 Evaluation of DUFS in the visible in-edges scenario

In this section we compare two variants of Directed Unbiased Frontier Sampling: E-DUFS, which uses the edge-based estimator and H-DUFS, which uses the hybrid estimator, against a single random walk (SingleRW) and multiple independent random walks (MultiRW).

5.2.1 Out-degree and in-degree distribution estimates. Here we focus on estimating the marginal in- and out-degree distributions. Each simulation consists of 1000 runs used to compute the empirical NRMSE. For MultiRW, E-DUFS and H-DUFS we set the average budget per walker to be $b = 10$. For conciseness, we only show a few representative results.

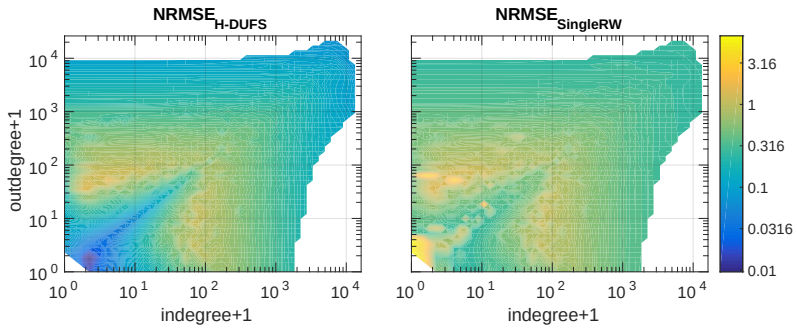


(a) livejournal-links (b) soc-pokec-relationships

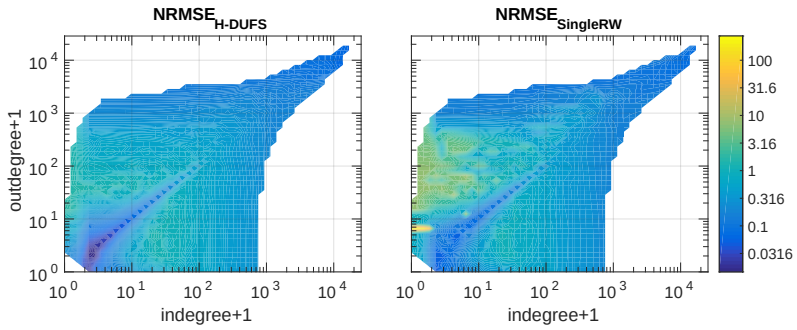
(c) web-BerkStan (d) wiki-Talk

Fig. 7. Comparison of single random walk (SingleRW), multiple independent random walks (MultiRW), DUFS with edge-based estimator (E-DUFS) and with hybrid estimator (H-DUFS). MultiRW yields the worst results, as the edge sampling probability is not the same across different connected components. Both DUFS variants outperform SingleRW, but H-DUFS is slightly more accurate in the head.

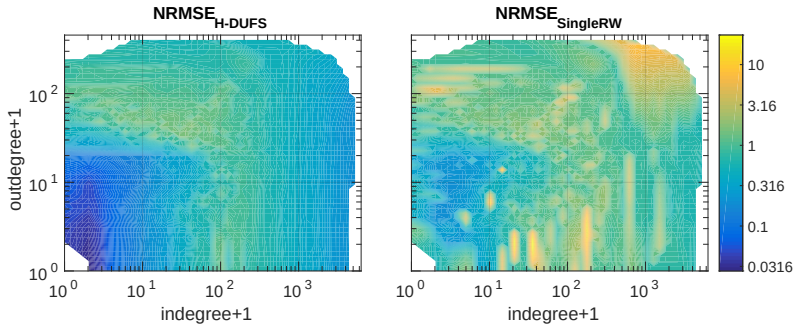
Figure 7 shows typical results obtained when using SingleRW, MultiRW, E-DUFS and H-DUFS to estimate out-degree distributions on the datasets. In 8 out of 15 datasets, MultiRW yields much larger NRMSEs than does the SingleRW. As pointed out in [28, Section 4.5], this is due to the fact that the estimator in (2) assumes that all edges are sampled with the same probability. This assumption is violated by MultiRW because the stationary sampling probability depends on the size of the connected component within which each walker is located. E-DUFS estimates are consistently more accurate than those of MultiRW and SingleRW, except on datasets where the original graph and its LCC have similar out-degree distributions. In some of these cases SingleRW slightly outperforms E-DUFS in the tail (see Fig. 7b). H-DUFS, in turn, outperforms E-DUFS in the head of the out-degree distribution and has similar performance when estimating other out-degree values. For this reason, defining the estimation task in terms of the CCDF would give H-DUFS an unfair advantage.



(a) flickr-links



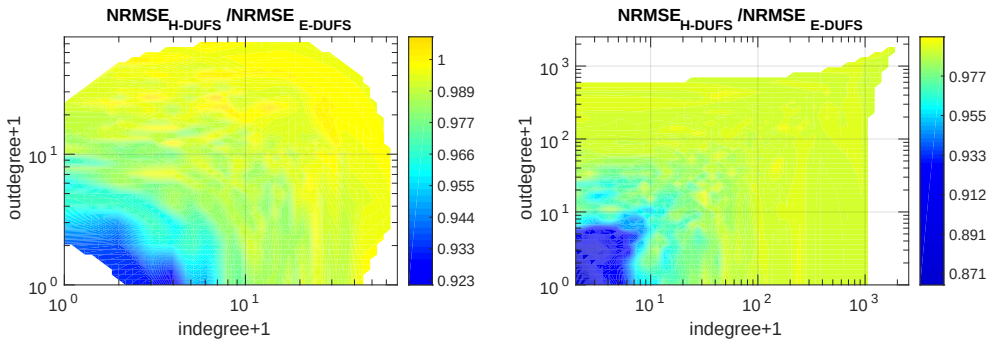
(b) youtube-links



(c) web-Google

Fig. 8. Comparison between H-DUFS and SingleRW w.r.t. NRMSE when estimating the joint in- and out-degree distribution. In most cases SingleRW will exhibit "hot spots" (regions with large NRMSE), which are mitigated by H-DUFS.

When restricted to the largest connected component, the performance differences between SingleRW and E-DUFS and those between SingleRW and H-DUFS become smaller, for $B = 0.1|V|$. Results for in-degree distribution estimation are qualitatively similar and are omitted.



(a) p2p-Gnutella31 (b) soc-Slashdot0902

Fig. 9. NRMSE ratios between H-DUFS and E-DUFS of the estimated joint in- and out-degree distribution for two datasets. H-DUFS is typically better than H-DUFS at low in and out-degree regions (**left**), but in social network graphs presented improvements over most of the joint distribution (**right**).

5.2.2 Joint in- and out-degree distributions. We compare the NRMSEs associated with H-DUFS and SingleRW for the estimates of the joint in- and out-degree distribution. We observe that H-DUFS consistently outperforms SingleRW on all datasets. On 10 out of 15 datasets, the estimates corresponding to low in-degree and low out-degree exhibit much smaller errors when using H-DUFS than when using SingleRW. Furthermore, H-DUFS also achieves smaller estimation errors for most of the remaining points of the joint distribution in 11 out of 15 datasets. Figures 8(a-b) show heatmaps corresponding to typical NRMSE results for H-DUFS and SingleRW. Interestingly, we note that on the web graph datasets and on the email-EuAll dataset, H-DUFS outperforms SingleRW by one or two orders of magnitude, as illustrated by Figure 8(c), which shows the heatmap comparison for dataset web-Google. Although the NRMSE exhibited by SingleRW applied to the LCC datasets is much smaller, the comparison between H-DUFS and SingleRW is qualitatively similar and is, therefore, omitted.

We then investigated the performance gains obtained by using the hybrid estimator instead of the original estimator. Figures 9(a-b) show the ratios between the NRMSEs obtained with H-DUFS (hybrid) to those obtained with the E-DUFS (original) for two networks. We chose to use the NRMSE ratio (or equivalently, the root MSE ratio) to make it easier to visualize the differences. We observe that H-DUFS consistently outperforms E-DUFS on all datasets. More precisely, the error ratio is rarely above one and, for points corresponding to small in- and out-degrees, it often lies below 0.9. Results on most datasets are similar to that depicted in Figure 9(a), but results on social networks datasets are closer to that shown in Figure 9(b), where large in- and out-degrees also seem to benefit from the information contained in the walkers' initial locations. Results for the LCC datasets are qualitatively similar, with accuracy gains from the hybrid estimator slightly larger on these datasets than on the original datasets.

5.3 Evaluation of DUFS in the invisible in-edges scenario

In this section, we compare the NRMSEs associated with DUFS and Directed Unbiased Random Walk (DURW) method when estimating out-degree distributions in the case where in-edges are not directly observable. We note that DURW is known to outperform a reference method for this scenario proposed in [3]. For a comparison between DURW and this reference method, please refer to [27].

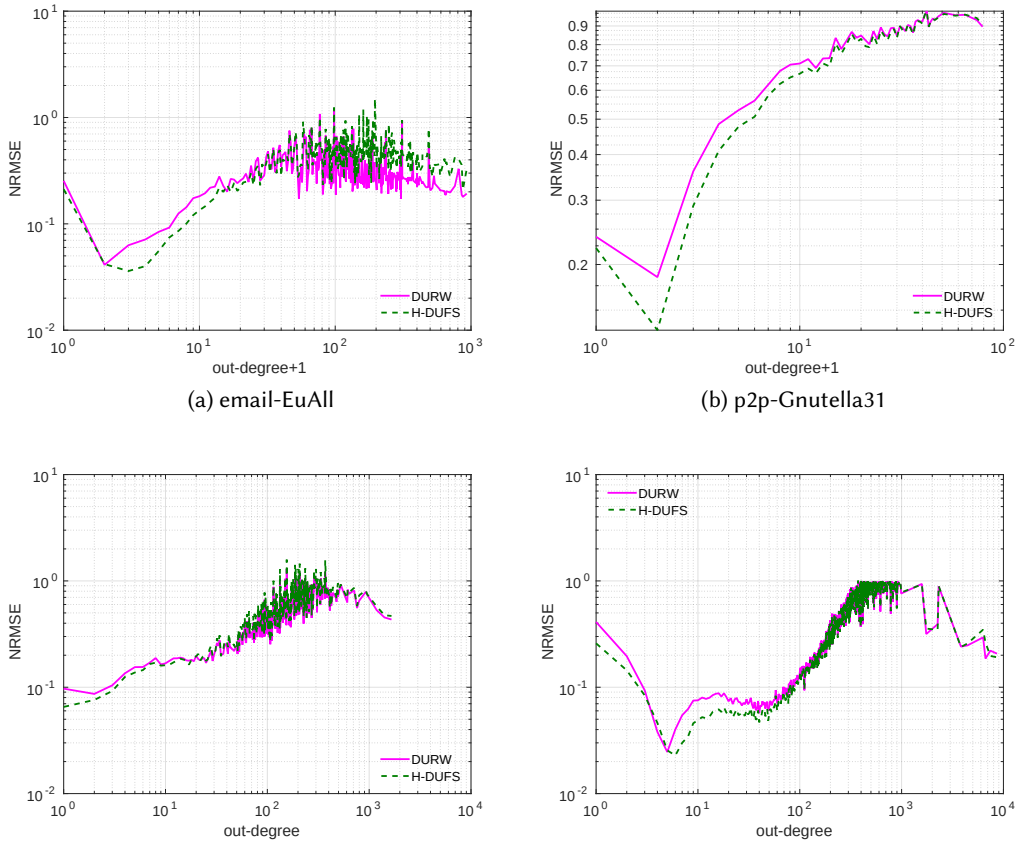


Fig. 10. NRMSEs associated with DUFS ($b = 10$, $w = 1$) and DURW (w' chosen to match average number of vertex samples) when estimating out-degree distribution. DURW performs more random jumps, thus better avoiding small volume components. On the original datasets, this improves DURW results in the tail, but often results in lower accuracy in the head (**top left**). In one third of the original datasets, DUFS yielded similar or better results than DURW over most out-degree points (**top right**). On most LCC datasets, DUFS outperforms DURW in the head and matches DURW's performance in the tail (**bottom**).

As we mentioned in Section 5.1, DURW results are similar to those obtained with DUFS when the budget per walker b is large, since DURW is a special case of DUFS where $b = B - c$. Therefore, we focus on comparing DUFS for small values of b and DURW, when the total number of uniform vertex samples collected by each method is roughly the same. More precisely, we simulate DUFS for $b = 10$ and $w = 1$ and set the DURW parameter w so that the number of vertex samples differs by at most 1% (averaged over 1000 runs). This aims to provide a fair comparison between these methods.

We find that neither of the two methods consistently outperforms the other over all datasets. The extra random jumps performed by DURW will prevent the walker from spending much of the budget in small volume components. As a result, DURW tends to exhibit larger errors in

the head but smaller errors in the tail of the out-degree distribution than DUFS. Figures 10(a,b) show typical results for $w = 1$ and $b = 10$. DUFS exhibited lower estimation errors in the head of the distribution on 11 datasets, being outperformed by DURW on one dataset and displaying comparable performance on the others. In 6 out of 15 datasets, DURW had better performance in the tail, while DUFS yielded better results on other five datasets. Results for $w = 1$ and $b \in \{10^2, 10^3\}$ are similar and are, therefore, omitted. As b increases, differences between DUFS and DURW start to vanish.

To better understand the impact of multiple connected components in DUFS and DURW performances, we simulate each method on the largest strongly connected component of each dataset (i.e., on the LCC datasets). Figures 10(c,d) show typical results among the LCC datasets. In most networks, DUFS yields smaller NRMSE than DURW in the head and yield similar results in the tail. Once again, for larger b the performances of DUFS and DURW become equivalent.

5.4 Relationship between NRMSE and out-degree distribution

Throughout Section 5 we observed that the NRMSE associated with RW-based methods tends to increase with out-degree up to a certain out-degree and to decrease after that. Moreover, for some out-degree ranges the log NRMSE seems to vary linearly with the log out-degree. Figure 5). For simplicity, we discuss the undirected graph case, but the extension to directed graphs is straightforward. The RW methods discussed here combine uniform vertex sampling with RW sampling approximated as uniform edge sampling. For simplicity, we analyze below the accuracy of uniform vertex and uniform edge sampling. We assume that each sampled edge results in exactly one observation, obtained by retrieving the set of labels associated with one of the adjacent vertices chosen equiprobably. Therefore both vertex sampling and edge sampling will collect vertex labels.

Let $\mathbb{S} = \{s_1, \ldots, s_B\}$ be the sequence of sampled vertices. For uniform vertex sampling, the probability of observing a given label ℓ in $\mathcal{L}(s_i)$ is θ_ℓ, for any $i = 1, \ldots, B$. The minimum variance unbiased estimator of θ_ℓ is

$$T^\ell_{\text{vs}}(\mathbb{S}) = \frac{1}{B} \sum_{i=1}^{B} \mathbb{1}\{\ell \in \mathcal{L}(s_i)\}. \tag{10}$$

Note that the summation in (10) is binomially distributed with parameters B and θ_ℓ. It follows that the mean squared error (MSE) of $T^\ell_{\text{vs}}(\mathbb{S})$ is given by

$$\begin{aligned} \text{MSE}(T^\ell_{\text{vs}}(\mathbb{S})) &= E[(T^\ell_{\text{vs}}(\mathbb{S}) - \theta_\ell)^2], \\ &= \frac{\theta_\ell(1-\theta_\ell)}{B}. \end{aligned} \tag{11}$$

For uniform edge sampling, the probability of observing a given label $\ell \in \mathcal{L}$ in the sample $\mathcal{L}(s_i)$ for $i = 1, \ldots, B$, is equal to

$$\pi_\ell = \frac{\sum_{v \in V} \mathbb{1}\{\ell \in \mathcal{L}(v)\} \deg(v)}{\sum_{u \in V} \deg(u)}.$$

In that case, the following estimator can be shown to be asymptotically unbiased

$$T^\ell_{\text{es}}(\mathbb{S}) = \frac{1}{B} \frac{\sum_{k=1}^{B} \mathbb{1}\{\ell \in \mathcal{L}(s_k)\} \deg^{-1}(s_k)}{\sum_{j=1}^{B} \deg^{-1}(s_j)}. \tag{12}$$

In particular, when vertex labels are the undirected degrees of each node, the probability of observing a given degree d becomes $\pi_d = d\theta_d/\bar{d}$, where $\bar{d}$ is the average undirected degree. The estimator for $B = 1$ reduces to $T^d_{\text{es}}(\mathcal{S}_1) = \mathbb{1}\{s_1 = d\}$, which is a random variable distributed according

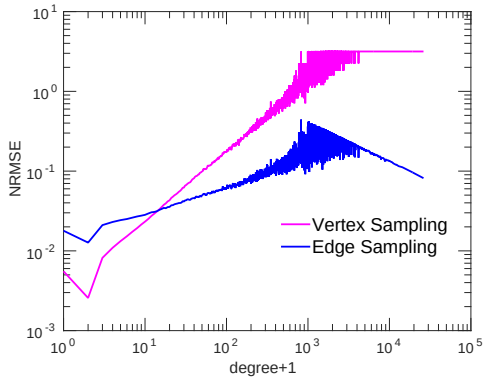


Fig. 11. NRMSE associated with uniform vertex sampling and uniform edge sampling when estimating degree distribution of the Flickr dataset (for $B = 0.1|V|$).

to a Bernoulli with parameter π_d. As a result, the MSE for $B > 1$ independent samples is given by

$$\mathrm{MSE}(T_{\mathrm{es}}^d(\mathbb{S})) \quad = \quad \frac{\pi_d(1-\pi_d)}{B}. \tag{13}$$

Equations (11) and (13) characterize the conditions under which each sampling model is more accurate. More precisely, for all i such that $\theta_d > \pi_d$ (or equivalently, $d < \bar{d}$), uniform vertex sampling yields better estimates than uniform edge sampling. This dichotomy is illustrated in Figure 11, which shows the NRMSE associated with degree distribution estimates resulting from each sampling model on the flickr-links dataset, for $B = 0.1|V|$.

Note that in log-log scale, both curves resemble a straight line for $d = 2, \ldots, 10^3$, which indicates a power law. For degrees larger than 5×10^3, the NRMSE associated with vertex sampling is constant, while the NRMSE associated with edge sampling decreases linearly with the degree. We show that these observations are direct consequences of the fact that the degree distribution in this network (as well as many other real networks) approximately follows a power law distribution. However, the degree distribution of a finite network cannot be an exact power law distribution because the tail is truncated. As a result, most of the largest degree values are observed exactly once. This can be seen in Figure 4 by noticing that on the flickr-links (and many other datasets) the p.m.f. is constant for the largest out-degrees. Assume, for instance, that the degree distribution can be modeled as

$$\theta_d = \begin{cases} d^{-\beta}/Z\,, & 1 \le d \le \tau \\ 1/|V|\,, & d > \tau, \end{cases}$$

for some $\beta \ge 1$ and some normalizing constant Z.

From (11), we have for uniform vertex sampling,

$$\mathrm{NRMSE}(T_{\mathrm{vs}}^d(\mathbb{S})) = \sqrt{(1/\theta_d - 1)/B}. \tag{14}$$

For $\theta_l \ll 1$ (true for large degrees), this implies

$$\mathrm{NRMSE}(T_{\mathrm{vs}}^d(\mathbb{S})) \approx \begin{cases} \sqrt{Zd^{\beta}/B}\,, & 1 \le d \le \tau \\ \sqrt{|V|/B}\,, & d > \tau. \end{cases}$$

For $d > \tau$, the NRMSE is constant. Otherwise, taking the log on both sides yields

$$\log(\text{NRMSE}(T^d_{\text{vs}}(\mathbb{S}))) \approx \frac{\beta}{2}\log d + \frac{1}{2}(\log Z - \log B),\ 1 \leq d \leq \tau, \tag{15}$$

which explains the relationship observed for uniform vertex sampling in Fig. 11.

From (13), we have for uniform edge sampling,

$$\text{NRMSE}(T^d_{\text{es}}(\mathbb{S})) = \sqrt{(1/\pi_d - 1)/B}. \tag{16}$$

For $\theta_d \ll 1$ (true for large degrees), this implies

$$\text{NRMSE}(T^d_{\text{es}}(\mathbb{S})) \approx \begin{cases} \sqrt{Z\bar{d}d^{\beta-1}/B}, & 1 \leq d \leq \tau \\ \sqrt{\frac{|E|}{d}/B}, & d > \tau. \end{cases}$$

Taking the log on both sides, it follows that

$$\log(\text{NRMSE}(T^d_{\text{es}}(\mathbb{S}))) \approx \begin{cases} \frac{\beta-1}{2}\log d + \frac{1}{2}(\log Z + \log \bar{d} - \log B), & 1 \leq d \leq \tau \\ -\frac{1}{2}(\log d - \log|E| - \log B), & d > \tau, \end{cases} \tag{17}$$

which explains the linear increase followed by the linear decrease observed in Fig. 11. Although RW-based methods can include uniform vertex sampling mechanisms, for large degrees NRMSE trends are better described by (17) than by (15), since most of the information about these degrees comes from RW samples.

6 RESULTS ON VERTEX LABEL DISTRIBUTIONS ESTIMATION

This section focuses on network datasets which possess (non-topological) node labels. Using these datasets, all of which represent undirected networks, we investigate which combinations of DUFS parameters outperform uniform vertex sampling when estimating node label distributions of the top 10% largest degree nodes. These nodes often represent the most important objects in a network.

Two of the four undirected attribute-rich datasets we use are social networks (DBLP and LiveJournal) obtained from Stanford SNAP, while two are information networks (DBpedia and Wikipedia) obtained from CMU's Auton Lab GitHub repository `active-search-gp-sopt` [19]. In these datasets, node labels correspond to some type of group membership and a node is allowed to be part of multiple groups simultaneously.

We simulate H-DUFS on each undirected network for all combinations of random jump weight $w \in \{0.1, 1, 10\}$ and budget per walker $b \in \{1, 10, 10^2\}$, and perform 1000 runs. Figure 13 compares the NRMSE associated with H-DUFS for different parameter combinations against uniform vertex sampling. Uniform vertex sampling results are obtained analytically using eq. (14). On DBpedia, Wikipedia and DBLP, almost all H-DUFS configurations outperform uniform vertex sampling. On LiveJournal, vertex sampling outperforms H-DUFS for attributes associated with large probability masses, but underperforms H-DUFS for attributes with smaller masses. In summary, we observe that H-DUFS with $w \in \{0.1, 1.0\}$ and $b \in \{10, 10^2\}$ yields superior accuracy than uniform vertex sampling when estimating node label distributions among the top 10% largest degree nodes.

7 DISCUSSION: DUFS PERFORMANCE IN THE ABSENCE OF UNIFORM VERTEX SAMPLING

In this section, we investigate the estimation accuracy of {E,H}-DUFS when random walkers are *not* initialized uniformly over V. We consider two simple non-uniform distributions over V to determine the initial walker locations walker positions:

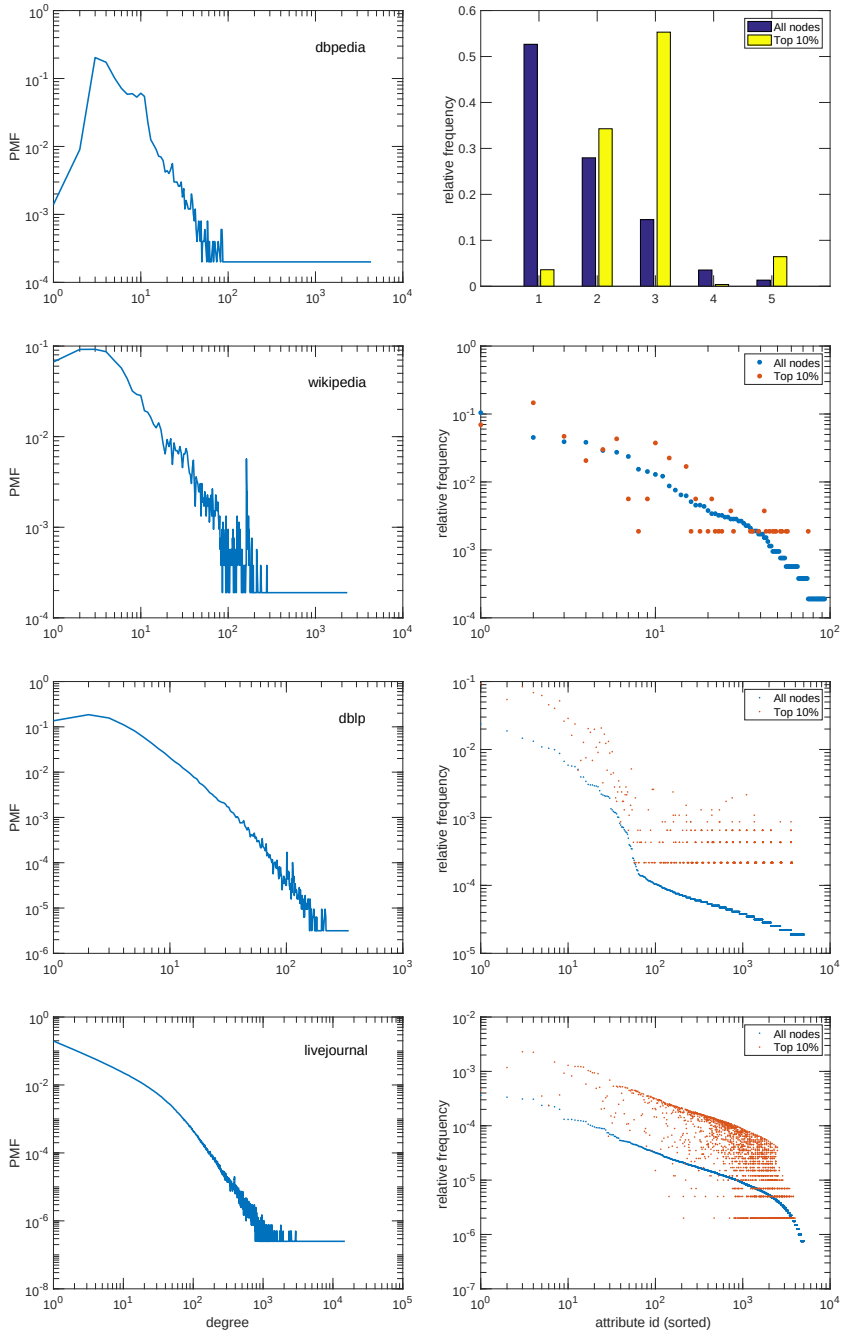
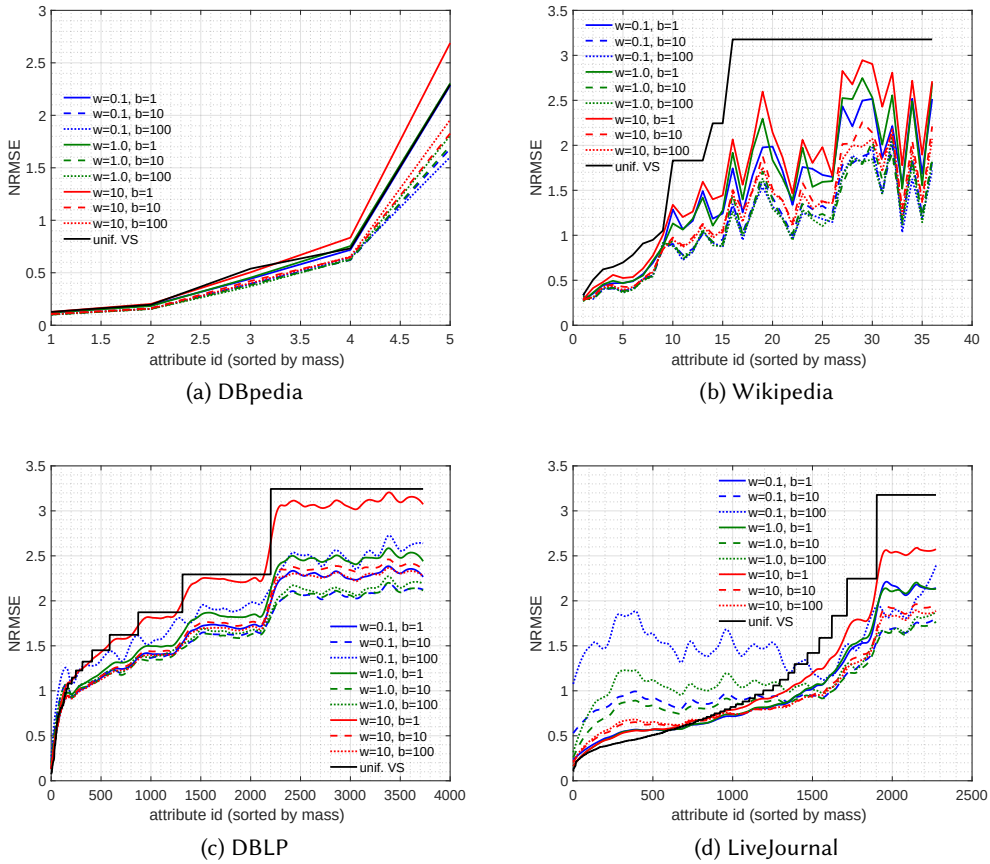


Fig. 12. Degree and node attribute distribution for undirected attribute-rich networks.



(a) DBpedia (b) Wikipedia

(c) DBLP (d) LiveJournal

Fig. 13. Comparison of hybrid estimator (H-DUFS) with uniform vertex sampling. H-DUFS curves on DBLP plot are smoothed by a local regression using weighted linear least squares and a second degree polynomial model to avoid clutter. H-DUFS with $w \in \{0.1, 1.0\}$ and $b \in \{10, 10^2\}$ yields comparable or superior accuracy than uniform vertex sampling.

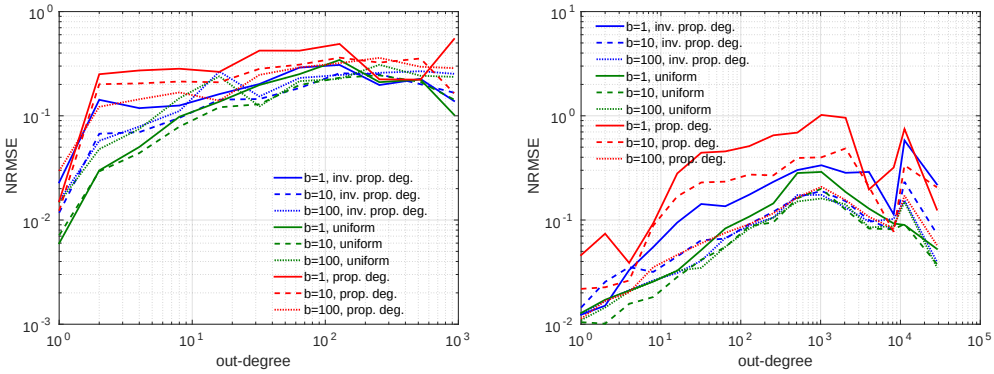
- Distribution Prop: proportional to the undirected degree, that is,

$$P(\text{initial walker location is } v) = \frac{\deg(v)}{\sum_{u \in V} \deg(u)}; \tag{18}$$

- Distribution Inv: proportional to the reciprocal of the undirected degree, that is,

$$P(\text{initial walker location is } v) = \frac{\deg^{-1}(v)}{\sum_{u \in V} \deg^{1}(u)}. \tag{19}$$

We simulate E-DUFS and H-DUFS on each network dataset setting the budget per walker to $b \in \{1, 10, 10^2, B-1\}$ in a scenario where in-edges are visible, performing 100 runs. Note that $b = B - 1$ corresponds to the case of a single random walker. Since we assume uniform vertex sampling (VS) is not available, we must set the random jump weight to $w = 0$. We include, however, results obtained when the initial walker locations are determined via VS for comparison.



(a) email-EuAll (b) youtube-links

Fig. 14. Effect of initializing walkers non-uniformly over V on E-DUFS accuracy. NRMSE decreases with budget per walker until $b = 10^2$.

Figures 14(a,b) show typical values of NRMSE associated with E-DUFS out-degree distribution estimates. We observe that NRMSE decreases with the budget per walker until $b = 10^2$, both for Prop and Inv. Simulations for $b = B - 1$ (i.e., using a single walker) yielded poor results and are omitted.

Intuitively, using the hybrid estimator when the initial walker locations come from some non-uniform distribution can incur unknown – and potentially large – biases. We conducted a set of simulations with H-DUFS, which corroborated this intuition. These results are omitted for conciseness. In summary, our results indicate that when the initial walker locations are determined according to some unknown distribution, a practitioner should use E-DUFS with moderately large b (e.g., 10^2).

8 RELATED WORK

Crawling methods for exploring undirected graphs: A number of papers investigate crawling methods (e.g., breadth-first search, random walks, etc.) for generating subgraphs with similar topological properties as the underlying network [12, 16]. On the other hand, [20] empirically investigates the performance of such methods w.r.t. specific measures of representativeness that can be useful in the context of specific applications (e.g., finding high-degree nodes for outbreak detection). However, these works focus on techniques that yield biased samples of the network and do not possess any accuracy guarantees. [1, 14] demonstrate that Breadth-First-Search (BFS) introduces a large bias towards high degree nodes, and it is difficult to remove these biases in general, although it can be reduced if the network in question is almost random [14]. Random walk (RW) is biased to sample high degree nodes, however its bias is known and can be easily corrected [28]. Random walks in the form of Respondent Driven Sampling (RDS) [10, 30] have been used to estimate population densities using snowball samples of sociological studies. The Metropolis-Hasting RW (MHRW) [31] modifies the RW procedure, aimed at sampling nodes with equal probability to estimation errors introduced by sampling. [5, 26] analytically prove that MHRW degree distribution estimates perform poorly in comparison to RWs. Empirically, the accuracy of RW and MHRW has been compared in [8, 24] and, as predicted by the theoretical results, RW is consistently more accurate than MHRW.

Reducing the mixing time of a regular RW is one way of improving the performance of RW based crawling methods. [2] proves that random jumps increase the spectral gap of the random walk, which in turn, leads to faster convergence to the steady state distribution. [13] assigns weights to nodes that are computed using their neighborhood information, and develop a weighted RW-based method to perform stratified sampling on social networks. They conduct experiments on Facebook and show that their stratified sampling technique achieves higher estimation accuracy than other methods. However, the neighborhood information in their method is limited to helping find random walk weights and is not used in estimators of graph statistics of interest. To solve this problem, [6] randomly samples nodes (either uniformly or with a known bias) and then uses neighborhood information to improve its unbiased estimator. [33] modifies the regular random walk by "rewiring" the network of interest on-the-fly in order to reduce the mixing time of the walk.

Crawling methods for exploring directed graphs: Estimating observable characteristics by sampling a directed graph (in this case, the Web graph) has been the subject of [3] and [11], which transform the directed graph of web-links into an undirected graph by adding reverse links, and then use a MHRW to sample webpages uniformly. The DURW method we propose in [27] adapts the "backward edge traversal" of [3] to work with a pure random walk and random jumps. Both of these Metropolis-Hastings RWs ([3] and [11]) are designed to sample directed graphs and do not allow random jumps. However, the ability to perform random jumps (even if jumps are rare) makes DURW and DUFS more efficient and accurate than the MetropolisHastings RW algorithm. Random walks with PageRank-style jumps are used in [16] to sample large graphs. In [16], however, there is no technique to remove the large biases induced by the random walk and the random jumps, which makes this method unfit for estimation purposes. More recently, another method based on PageRank was proposed in [29], but it assumes that obtaining uniform vertex samples is not feasible. In the presence of multiple strongly connected components, this method offers no accuracy guarantees.

In the last decade, there has been a growing interest in graph sketching for processing massive networks. A sketch is a compact representation of data. Unlike a sample, a sketch is computed over the entire graph, that is observed as a data stream. For a survey on graph sketching techniques, please refer to [21].

9 CONCLUSION

In this paper, we proposed the Directed Unbiased Frontier Sampling (DUFS) method for characterizing networks. DUFS generalizes the Frontier Sampling (FS) and the Directed Unbiased Random Walk (DURW) methods. In some sense, DUFS extends FS to make it applicable to directed networks when incoming edges are not directly observable by building on ideas from DURW. Like DURW, DUFS can also be applied to undirected networks without any modification.

We also proposed a novel estimator for vertex label distribution that can account for FS and DUFS walkers initial locations – or more generally, random vertex samples – and a heuristic that can reduce the variance incurred by vertex samples that happen to sample nodes whose labels have extremely low probability masses. When the proposed estimator is used in combination with the heuristic, we showed that estimation errors can be significantly reduced in the distribution head when compared with the estimator proposed in [28], regardless of whether we are estimating out-degree, in-degree or joint in- and out-degree distributions.

We conducted an empirical study on the impact of DUFS parameters (namely, budget per walker and random jump weight) on the estimation of out-degree and in-degree distributions using a large variety of datasets. We considered four scenarios, corresponding to whether incoming edges are directly observable or not and whether random vertex sampling has a similar or larger cost than

moving random walkers on the graph. This study allowed us to provide practical guidelines on setting DUFS parameters to obtain accurate head estimates or accurate tail estimates. When the goal is a balance between the two objectives, intermediate configurations can be chosen.

Last, we compared DUFS against random walk-based methods designed for undirected and directed networks. In our simulations for the scenario where in-edges are visible, DUFS yielded much lower estimation errors than a single random walk or multiple independent random walks. We also observed that DUFS consistently outperforms FS due to the degree proportional jumps mechanism implemented by the former. In the scenario where in-edges are unobservable, DUFS outperformed DURW when estimating the probability mass associated with the smallest out-degree values (for equivalent parameter settings). In addition, more often than not, DUFS slightly outperformed DURW when estimating the mass associated to the largest out-degrees. In the presence of multiple strongly connected components, DURW tends to move from small to largest components more often than DUFS, sometimes exhibiting lower estimation errors in the distribution tail. However, when restricting the estimation to the largest component, DUFS outperforms DURW in virtually all datasets used in our simulations.

A HYBRID ESTIMATOR AND ITS STATISTICAL PROPERTIES

In this appendix, we derive the recursive variant of the hybrid estimator. From that we derive its non-recursive variant. Next, we show that the non-recursive variant is asymptotically unbiased. In the case of undirected networks where the average degree is given, we show that the resulting hybrid estimator of the undirected degree mass is the minimum variance unbiased estimator (MVUE).

Let us recall variables and constants used in the definition of the hybrid estimator:

n_i	number of vertex samples with label i
$\theta_{i,j}$	fraction of nodes in $G^{(t)}$ with label i and undirected degree j
$m_{i,j}$	number of edge samples with label i and bias j
$m_i = \sum_j m_{i,j}$	total number of edge samples with label i
$N = \sum_i n_i$	total number of vertex samples
$M = \sum_i m_i$	total number of edge samples
$B = N + M$	total budget

We approximate random walk samples in DUFS by uniform edge samples from G_u. Experience from previous papers shows us that this approximation works very well in practice. This yields the following likelihood function

$$L(\boldsymbol{\theta}|\mathbf{n},\mathbf{m}) = \frac{\prod_i \theta_i^{n_i} \prod_k (k\theta_{i,k})^{m_{i,k}}}{\left(\sum_{s,t} t\theta_{s,t}\right)^M}. \tag{20}$$

The key idea in our derivation is that we can bypass the numerical estimation of the $\theta_{i,j}$'s by noticing that $\theta_{i,j} \propto \theta_i$, $\theta_{i,j} \propto m_{i,j}$ and $\theta_{i,j} \propto 1/j$. Hence, the maximum likelihood estimator of $\theta_{i,j}$ for $j = 1, \ldots, Z$ is the Horvitz-Thompson estimator

$$\hat{\theta}_{i,j} = \frac{\theta_i m_{i,j}}{j\mu_i}, \tag{21}$$

where $\mu_i = \sum_k m_{i,k}/k$.

Substituting (21) in (20) yields

$$L(\boldsymbol{\theta}|\mathbf{n},\mathbf{m}) = \frac{\prod_i \theta_i^{n_i} \prod_k (\theta_i m_{i,k}/\mu_i)^{m_{i,k}}}{\left(\sum_s \theta_s \sum_z m_{s,z}/\mu_s\right)^M}. \tag{22}$$

The log-likelihood approximation is then given by

$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{n},\mathbf{m}) = -M\log\left(\sum_s \theta_s \sum_z \frac{m_{s,z}}{\mu_s}\right) + \sum_i n_i \log\theta_i + \sum_k m_{i,k}(\log\theta_i + \log m_{i,k} - \log\mu_i). \quad (23)$$

We rewrite θ_i as $e^{\beta_i}/\sum_j e^{\beta_j}$ to account for the distribution constraints $\sum_i \theta_i = 1$ and $\theta_i \in [0,1]$. Hence, we have

$$\mathcal{L}(\boldsymbol{\beta}|\mathbf{n},\mathbf{m}) = -M\log\left(\sum_s \frac{e^{\beta_s} m_s}{\mu_s}\right) + \sum_i (n_i + m_i)\beta_i - N\log\left(\sum_j e^{\beta_j}\right) + C, \quad (24)$$

where $m_i = \sum_k m_{i,k}$ and C is a constant that does not depend on β.

The partial derivative w.r.t. β_i is given by

$$\frac{\partial\mathcal{L}(\boldsymbol{\beta}|\mathbf{n},\mathbf{m})}{\partial\beta_i} = -\frac{Me^{\beta_i} m_i/\mu_i}{\sum_s e^{\beta_s} m_s/\mu_s} + n_i + m_i - \frac{Ne^{\beta_i}}{\sum_j e^{\beta_j}}. \quad (25)$$

Setting $\partial\mathcal{L}(\boldsymbol{\beta}|\mathbf{n},\mathbf{m})/\partial\beta_i = 0$ and substituting back θ_i yields

$$\theta_i^\star = \frac{n_i + m_i}{N + M\frac{m_i/\mu_i}{\sum_s \theta_s^\star m_s/\mu_s}}. \quad (26)$$

Theorem A.1. *Let $N = cB$ and $M = (1-c)B$, for some $0 < c < 1$. The estimator*

$$\hat{\theta}_i = \frac{n_i + m_i}{N + M\frac{m_i}{\mu_i \hat{d}}}, \quad (27)$$

where $\mu_i = \sum_k m_{i,k}/k$ and $\hat{d} = M/\sum_i \mu_i$, is an asymptotically unbiased estimator of θ_i.

Proof. In the limit as $B \to \infty$, we have

$$E[n_i] = N\theta_i, \qquad E[m_{i,k}] = M\frac{k\theta_{i,k}}{\sum_{s,l} l\theta_{sl}}, \qquad E[m_i] = M\frac{\sum_k k\theta_{i,k}}{\sum_{s,l} l\theta_{s,l}},$$

and thus,

$$E[\mu_i] = M\frac{\sum_k k\theta_{i,k}/k}{\sum_{s,l} l\theta_{sl}} = M\frac{\theta_i}{\sum_{s,l} l\theta_{sl}} \quad \text{and} \quad E\left[\frac{m_i}{\mu_i}\right] = \frac{\sum_k k\theta_{i,k}}{\theta_i}.$$

It follows that

$$\lim_{B\to\infty} E[\hat{d}] = \frac{M}{M\frac{\sum_i \theta_i}{\sum_{s,l} l\theta_{sl}}} = \sum_{s,l} l\theta_{sl}.$$

Substituting the above in eq. (27), we have

$$\lim_{B\to\infty} E[\theta_i^\star] = \frac{N\theta_i + M\frac{\sum_k k\theta_{i,k}}{\sum_{s,l} l\theta_{s,l}}}{N + M\frac{\sum_k k\theta_{i,k}/\theta_i}{\sum_{s,l} l\theta_{s,l}}} = \theta_i.$$

This concludes the proof. □

In Section 4.2.2 we mentioned a special case of the previous estimator, where the vertex label is the undirected degree itself. We prove that this estimator, denoted by $\hat{\theta}_i$ is the minimum variance unbiased estimator (MVUE) of θ_i.

THEOREM A.2. *The estimator*

$$\bar{\theta}_i = \frac{n_i + m_i}{N + Mi/\bar{\mu}},$$

where $\bar{\mu} = \sum_j j\theta_j$, *is an unbiased estimator of* θ_i.

PROOF. We know that $n_i \sim \text{Binomial}(N, \theta_i)$ and $m_i \sim \text{Binomial}(M, i\theta_i/\bar{\mu})$. Hence,

$$\begin{aligned}
E[\hat{\theta}_i] &= \sum_{n_i, m_i} \frac{n_i + m_i}{N + Mi/\bar{\mu}} \overbrace{\binom{N}{n_i}\theta_i^{n_i}(1-\theta_i)^{N-n_i}}^{A(n_i)} \overbrace{\binom{M}{m_i}\left(\frac{i\theta_i}{\bar{\mu}}\right)^{m_i}\left(1-\frac{i\theta_i}{\bar{\mu}}\right)^{M-m_i}}^{B(m_i)} \\
&= \frac{1}{N + Mi/\bar{\mu}}\left(\sum_{n_i} n_i A(n_i) \sum_{m_i} B(m_i) + \sum_{m_i} m_i B(m_i) \sum_{n_i} A(n_i)\right) \\
&= \frac{1}{N + Mi/\bar{\mu}}\left(\sum_{n_i} n_i A(n_i) + \sum_{m_i} m_i B(m_i)\right) \\
&= \frac{1}{N + Mi/\bar{\mu}}\left(N\theta_i + Mi\theta_i/\bar{\mu}\right) \\
&= \theta_i.
\end{aligned}$$

□

Having proved that $\hat{\theta}_i$ is unbiased, we are now ready to show that it is also the minimum variance unbiased estimator (MVUE). In order to do so, we prove Lemmas A.1 and A.3 that show respectively that $n_i + m_i$ is a sufficient and complete statistic of θ_i.

LEMMA A.1. *The statistic* $n_i + m_i$ *is a sufficient statistic with respect to the likelihood of* θ_i.

PROOF. The log-likelihood equation for estimator (8) is given by

$$\begin{aligned}
L(\boldsymbol{\theta}|\mathbf{n}, \mathbf{m}) &= \frac{\prod_i \theta_i^{n_i} \prod_j (j\theta_j)^{m_j}}{\hat{\mu}^M} \\
&= \frac{\prod_j j^{m_j}}{\hat{\mu}^M} \prod_i \theta_i^{n_i + m_i}. \qquad (28)
\end{aligned}$$

We can see from eq. (28) that the likelihood function $L(\boldsymbol{\theta}|\mathbf{n}, \mathbf{m})$ can be factored into a product such that one factor, $\prod_j j^{m_j}/\hat{\mu}^M$, does not depend on θ_i and the other factor, which does depend on θ_i, depends on $\mathbf{n}$ and $\mathbf{m}$ only through $n_i + m_i$. From the Fisher-Neyman factorization Theorem [15], we conclude that $n_i + m_i$ is a sufficient statistic for the distribution of the sample. □

We now prove that $n_i + m_i$ is also a complete statistic for the distribution of the sample.

Definition A.2. Let X be a random variable whose probability distribution belongs to a parametric family of probability distributions P_θ parametrized by θ. The statistic s is said to be complete for the distribution of X if for every measurable function g (which must be independent of θ) the following implication holds:

$$E(g(s(X))) = 0 \text{ for all } \theta \Rightarrow P_\theta(g(s(X)) = 0) = 1 \text{ for all } \theta.$$

LEMMA A.3. *The statistic* $n_i + m_i$ *is a complete statistic w.r.t. the likelihood of* θ_i.

Proof.

$$\begin{aligned} E[g(n_i + m_i)] &= 0 \\ \sum_{n_i, m_i} g(n_i + m_i) P_\theta(n_i, m_i) &= 0 \\ \sum_{n_i, m_i} g(n_i + m_i) A(n_i) B(m_i) &= 0 \end{aligned} \tag{29}$$

The LHS of (29) is a polynomial of degree $M + N$ on θ_i. Hence, it can be written as

$$C_0 + C_1\theta_i + C_2\theta_i^2 + \ldots + C_{N+M}\theta_i^{N+M} = 0. \tag{30}$$

We prove that $P_\theta(g(s(X)) = 0) = 1$ for all θ by contradiction. Suppose that there is a θ such that $P_\theta(g(s(X)) \neq 0) > 0$. In order to have $E(g(s(X))) = 0$, there must be terms for which $g(.)$ is strictly positive and terms for which $g(.)$ is strictly negative. Let $g(h_1)$ be the smallest h_1 such that $g(h_1) > 0$. Let $g(h_2)$ be the smallest h_2 such that $g(h_2) < 0$. Let $h = \min(h_1, h_2)$.

Expanding $A(n_i)B(m_i)$ in eq. (29) we note that the degree of the resulting polynomial is at least $n_i + m_i$ on θ_i. Therefore, the coefficient C_h in eq. (30) associated with θ_i^h cannot have terms of $g(.)$ larger than h. Then C_h can only be zero if $h_1 = h_2$ which is a contradiction. □

Theorem A.3. *The estimator $\bar{\theta}_i$ is the minimum variance unbiased estimator (MVUE) of θ_i.*

Proof. According to the Lehmann-Scheffe Theorem [15], if $T(\mathbb{S})$ is a complete sufficient statistic, there is at most one unbiased estimator that is a function of $T(\mathbb{S})$. From Lemmas A.1 and A.3, we have that $n_i + m_i$ is a complete sufficient statistic of θ_i. Clearly, the unbiased estimator $\hat{\boldsymbol{\theta}}$ in eq. (27) is a function $n_i + m_i$. Therefore, $\hat{\theta}_i$ must be the MVUE. □

Alternatively, we can prove Theorem A.3 from Lemmas A.1 and A.3 by showing that applying the Rao-Blackwell Theorem to the unbiased estimator $\hat{\theta}_i$ using the complete sufficient statistic $n_i + m_i$ yields exactly the same estimator:

$$\begin{aligned} \theta_i' &= E\left[\hat{\theta}_i | n_i + m_i\right] \\ &= \sum_{t_j} t_j P(\hat{\theta}_i = t_j | n_i + m_i) \\ &= \sum_{t_j} t_j 1\left\{\frac{n_i + m_i}{N + Mi/\bar{\mu}} = t_j\right\} \\ &= \frac{n_i + m_i}{N + Mi/\bar{\mu}}. \end{aligned}$$

REFERENCES

[1] Dimitris Achlioptas, Aaron Clauset, David Kempe, and Cristopher Moore. 2009. On the Bias of Traceroute Sampling: Or, Power-law Degree Distributions in Regular Graphs. *J. ACM* 56, 4, Article 21 (July 2009), 28 pages.

[2] Konstantin Avrachenkov, Bruno Ribeiro, and Don Towsley. 2010. *Improving Random Walk Estimation Accuracy with Uniform Restarts.* Springer Berlin Heidelberg, Berlin, Heidelberg, 98–109.

[3] Ziv Bar-Yossef and Maxim Gurevich. 2008. Random sampling from a search engine's index. *J. ACM* 55, 5 (2008), 1–74.

[4] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. 2006. Complex networks: Structure and dynamics. *Physics Reports* 424, 4-5 (2006), 175–308.

[5] Flavio Chiericetti, Anirban Dasgupta, Ravi Kumar, Silvio Lattanzi, and Tamás Sarlós. 2016. On Sampling Nodes in a Network. In *Proceedings of the 25th International Conference on World Wide Web (WWW '16).* International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 471–481.

[6] Anirban Dasgupta, Ravi Kumar, and D. Sivakumar. 2012. Social Sampling. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '12).* ACM, New York, NY, USA, 235–243.

[7] Nathan Eagle, Alex (Sandy) Pentland, and David Lazer. 2009. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences* 106, 36 (2009), 15274–15278.

[8] Minas Gjoka, Carter T. Butts, Maciej Kurant, and Athina Markopoulou. 2010. Walking in Facebook: A Case Study of Unbiased Sampling of OSNs. In *Proceedings of IEEE INFOCOM 2010*. 1–9.

[9] Douglas D. Heckathorn. 1997. Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations. *Social Problems* 44, 2 (1997), 174–199.

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

[11] Monika R. Henzinger, Allan Heydon, Michael Mitzenmacher, and Marc Najork. 2000. On near-uniform URL sampling. *Computer Networks* 33, 1-6 (2000), 295 – 308.

[12] Christian Hubler, H-P Kriegel, Karsten Borgwardt, and Zoubin Ghahramani. 2008. Metropolis Algorithms for Representative Subgraph Sampling. In *2008 Eighth IEEE International Conference on Data Mining*. 283–292.

[13] Maciej Kurant, Minas Gjoka, Carter T. Butts, and Athina Markopoulou. 2011a. Walking on a Graph with a Magnifying Glass: Stratified Sampling via Weighted Random Walks. In *ACM SIGMETRICS 2011*. ACM, New York, NY, USA, 281–292.

[14] Maciej Kurant, Athina Markopoulou, and Patrick Thiran. 2011b. Towards Unbiased BFS Sampling. *IEEE Journal on Selected Areas in Communications* 29, 9 (September 2011), 1799–1809.

[15] Erich Leo Lehmann, George Casella, and George Casella. 1991. *Theory of point estimation*. Wadsworth & Brooks/Cole Advanced Books & Software.

[16] Jure Leskovec and Christos Faloutsos. 2006. Sampling from Large Graphs. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '06)*. ACM, New York, NY, USA, 631–636.

[17] Jure Leskovec and Andrej Krevl. 2014. SNAP Datasets: Stanford Large Network Dataset Collection. http://snap.stanford.edu/data. (June 2014).

[18] Jure Leskovec, Kevin J. Lang, Anirban Dasgupta, and Michael W. Mahoney. 2008. Statistical Properties of Community Structure in Large Social and Information Networks. In *Proceedings of the 17th International Conference on World Wide Web (WWW '08)*. ACM, New York, NY, USA, 695–704.

[19] Yifei Ma, Tzu-Kuo Huang, and Jeff G Schneider. 2015. Active Search and Bandits on Graphs using Sigma-Optimality.. In *Conference on Uncertainty in Artificial Intelligence*. 542–551.

[20] Arun S. Maiya and Tanya Y. Berger-Wolf. 2011. Benefits of Bias: Towards Better Characterization of Network Sampling. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '11)*. ACM, New York, NY, USA, 105–113.

[21] Andrew McGregor. 2014. Graph stream algorithms: a survey. *ACM SIGMOD Record* 43, 1 (2014), 9–20.

[22] Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. 2007. Measurement and Analysis of Online Social Networks. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement (IMC '07)*. ACM, New York, NY, USA, 29–42.

[23] Fabricio Murai, Bruno Ribeiro, Don Towsley, and Pinghui Wang. 2013. On Set Size Distribution Estimation and the Characterization of Large Networks via Sampling. *IEEE Journal on Selected Areas in Communications* 31, 6 (June 2013), 1017–1025.

[24] Amir H. Rasti, Mojtaba Torkjazi, Reza Rejaie, Nick Duffield, Walter Willinger, and Daniel Stutzbach. 2009. Respondent-Driven Sampling for Characterizing Unstructured Overlays. In *Proceedings of the IEEE INFOCOM 2009*. 2701–2705.

[25] Bruno Ribeiro, William Gauvin, Benyuan Liu, and Don Towsley. 2010. On MySpace Account Spans and Double Pareto-Like Distribution of Friends. In *INFOCOM IEEE Conference on Computer Communications Workshops , 2010*. 1–6.

[26] Bruno Ribeiro and Don Towsley. 2012. On the estimation accuracy of degree distributions from graph sampling. In *51st IEEE Conference on Decision and Control (CDC 2012)*. 5240–5247.

[27] Bruno Ribeiro, Pinghui Wang, Fabricio Murai, and Don Towsley. 2012. Sampling directed graphs with random walks. In *Proceedings of IEEE INFOCOM 2012*. 1692–1700.

[28] Bruno F. Ribeiro and Donald F. Towsley. 2010. Estimating and Sampling Graphs with Multidimensional Random Walks. *CoRR* abs/1002.1751 (2010). http://arxiv.org/abs/1002.1751

[29] M. Salehi and H. R. Rabiee. 2013. A Measurement Framework for Directed Networks. *IEEE Journal on Selected Areas in Communications* 31, 6 (June 2013), 1007–1016.

[30] Matthew J. Salganik and Douglas D. Heckathorn. 2004. Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology* 34 (2004), 193–239.

[31] Daniel Stutzbach, Rea Rejaie, Nick Duffield, Subhabrata Sen, and Walter Willinger. 2009. On unbiased sampling for unstructured peer-to-peer networks. *IEEE/ACM Transactions on Networking* 17, 2 (April 2009), 377–390.

[32] Erik Volz and Douglas D. Heckathorn. 2008. Probability Based Estimation Theory for Respondent Driven Sampling. *Journal of Official Statistics* 24, 1 (03 2008), 79.

[33] Zhuojie Zhou, Nan Zhang, Zhiguo Gong, and Gautam Das. 2016. Faster Random Walks by Rewiring Online Social Networks On-the-Fly. *ACM Trans. Database Syst.* 40, 4, Article 26 (Jan. 2016), 36 pages.