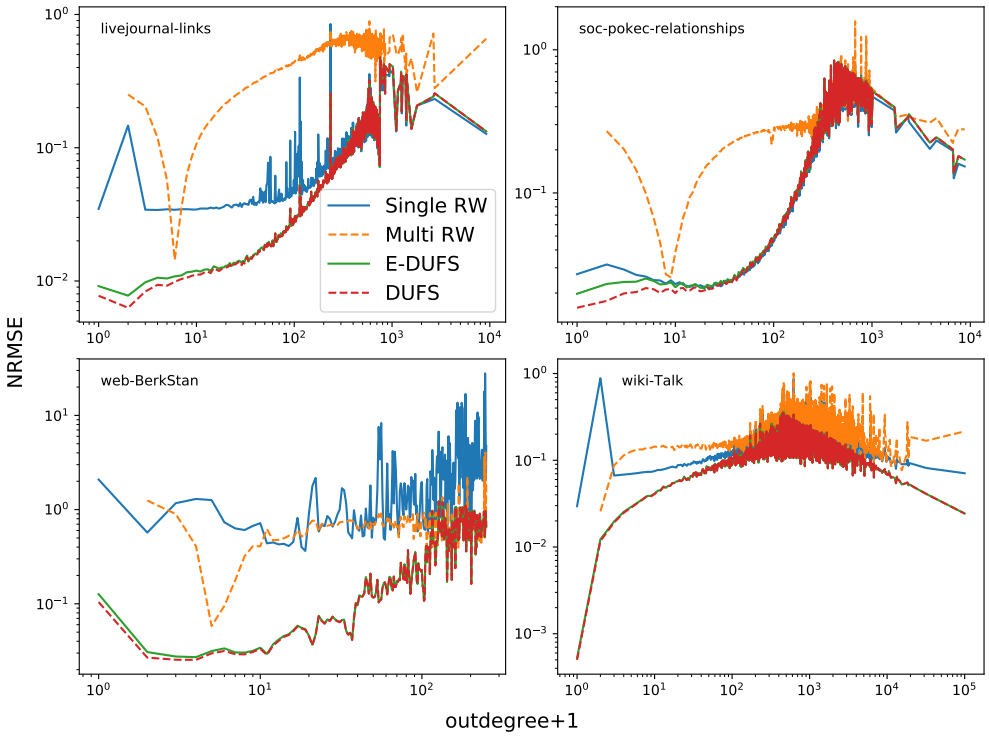



Fig. 7: Comparison of single random walk (SingleRW), multiple independent random walks (MultiRW), DUFS with edge-based estimator (E-DUFS) and with hybrid estimator (DUFS). MultiRW yields the worst results, as the edge sampling probability is not the same across different connected components. Both DUFS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

### 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 DUFS, which uses the hybrid estimator, to each other and to a single random walk (SingleRW) and multiple independent random walks (MultiRW). We do not include Frontier Sampling in the comparison as it is a special case of DUFS where $w = 0$ and we know from Section 5.1 that allowing random jumps effectively reduce estimation errors.

*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 from which we compute the empirical NRMSE. For MultiRW, E-DUFS and DUFS we set the average budget per walker to be $b = 10$. For conciseness, we only show a few representative results.

Figure 7 shows typical results obtained when using SingleRW, MultiRW, E-DUFS and 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 [Ribeiro and Towsley 2010, Section 4.5], this is due to the fact that the estimator in (1) 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 top-right fig.). 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 DUFS an unfair advantage.

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

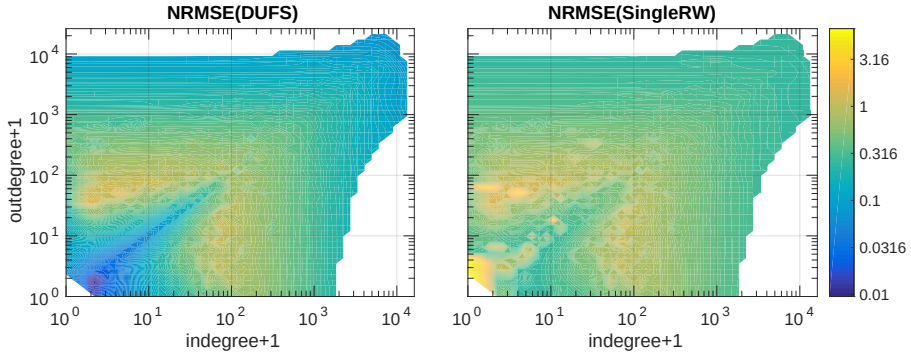
*5.2.2. Joint in- and out-degree distributions.* We compare the NRMSEs associated with DUFS and SingleRW for the estimates of the joint in- and out-degree distribution. We observe that DUFS consistently outperforms SingleRW on all datasets. On 10 out of 15 datasets, the estimates corresponding to low in-degree and low out-degree exhibit much smaller errors when using DUFS than when using SingleRW. Furthermore, 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 DUFS and SingleRW. Interestingly, we note that on the web graph datasets and on the email-EuAll dataset, 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 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 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 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.

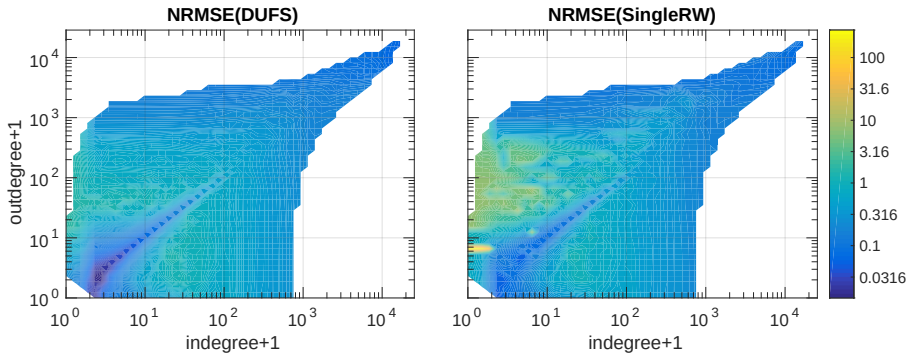
### 5.3. Evaluation of DUFS in the invisible in-edges scenario

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 [Bar-Yossef and Gurevich 2008]. For a comparison between DURW and this reference method, please refer to [Ribeiro et al. 2012].

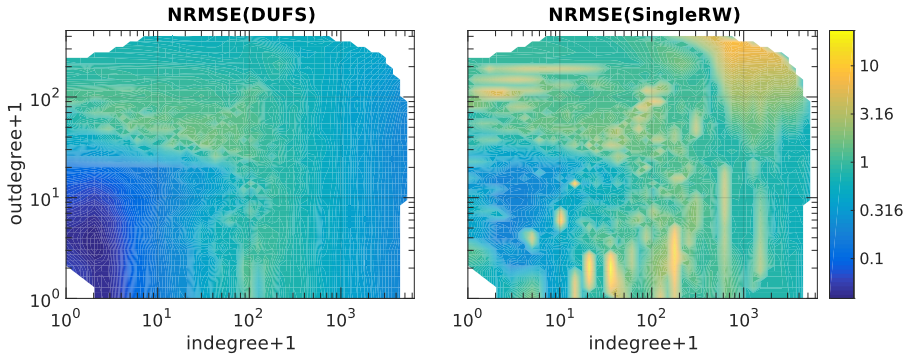
As we mentioned in Section 5.1, DURW results are similar to those obtained with DUFS when the budget per walker $b$ is large, since DURW is a special case of DUFS where $b = B - c$. Therefore, we focus on comparing DUFS for small values of $b$ and DURW, when the total number of uniform node samples collected 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 node samples differs by at most




(a) flickr-links

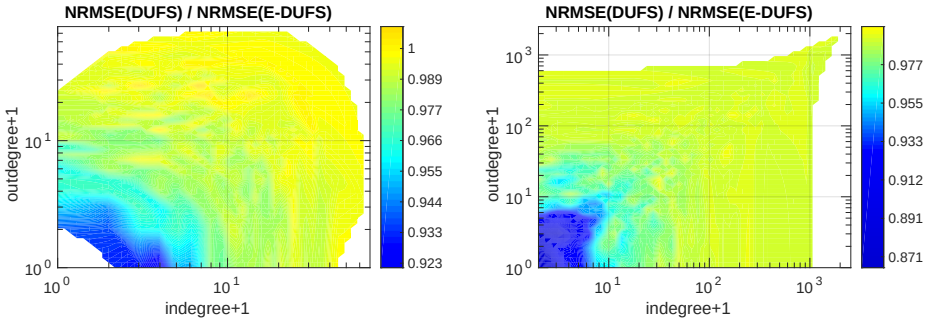



(b) youtube-links




(c) web-Google

Fig. 8: Comparison between 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 DUFS.




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

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

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. Figure 10 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). Figure 11 shows 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 node sampling with RW sampling approximated as uniform edge sampling. For simplicity, we analyze below the accuracy of uniform node and uniform edge sampling. We assume that each sampled edge produces one observation, obtained by retrieving the set of labels associated with one of the adjacent vertices chosen equiprobably. Therefore both node sampling and edge sampling collect node labels.

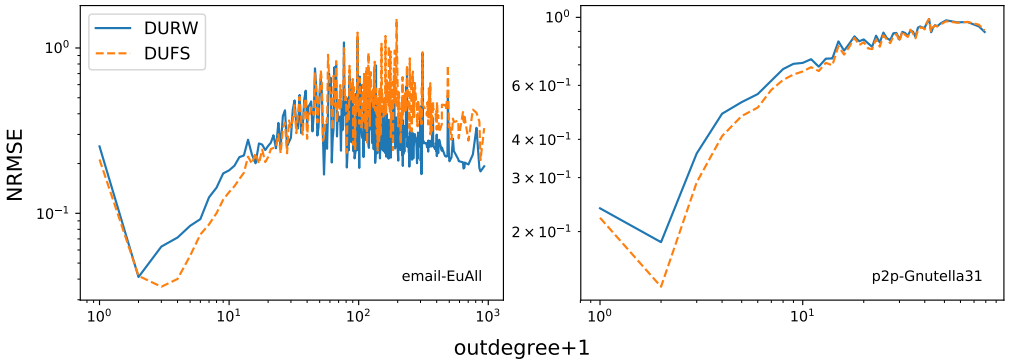



Fig. 10: NRMSEs associated with DUFS ($w = 1$, $b = 10$) and DURW ($w'$ chosen to match average number of node samples) when estimating out-degree distribution. DURW performs more random jumps, thus better avoiding small volume components. This improves DURW results in the tail, but often results in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

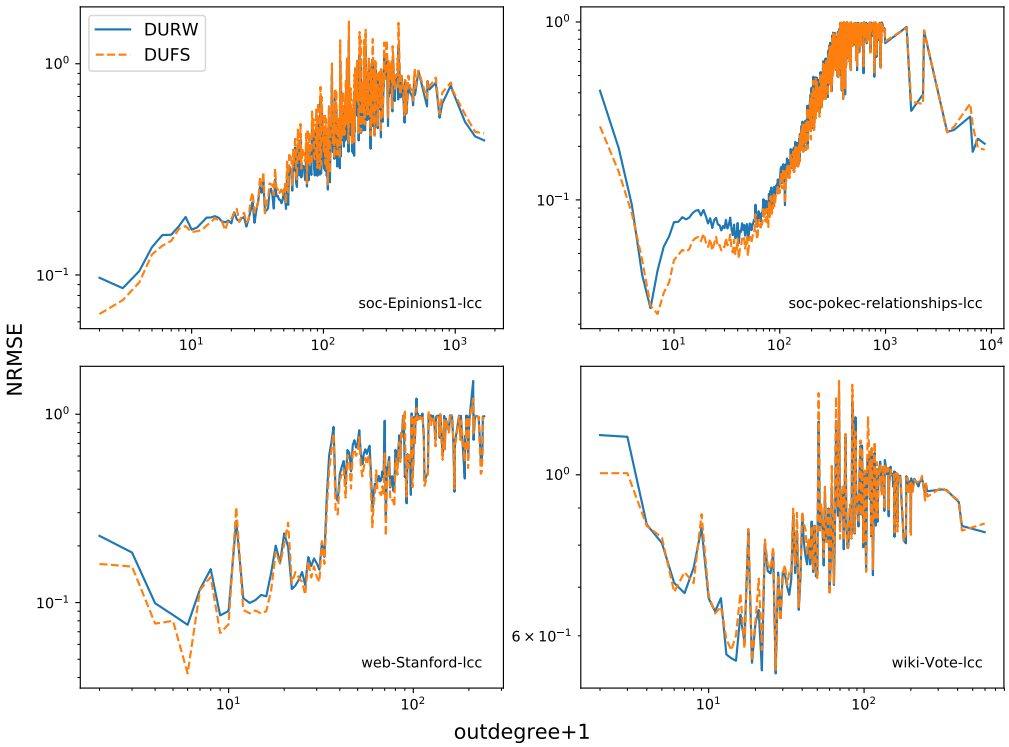



Fig. 11: NRMSEs associated with DUFS ($w = 1$, $b = 10$) and DURW ($w'$ chosen to match average number of node samples) when estimating out-degree distribution.

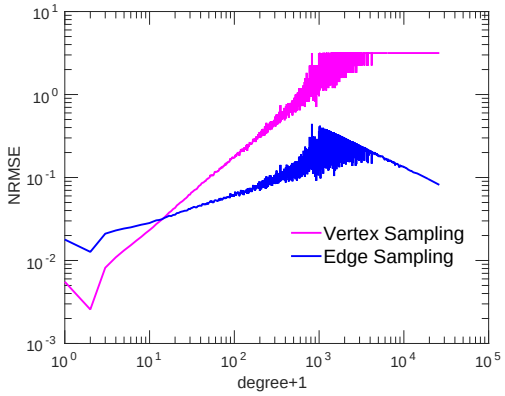



Fig. 12: NRMSE associated with uniform node sampling and uniform edge sampling when estimating degree distribution of the Flickr dataset (for $B = 0.1|V|$).

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

$$T^\ell_{\text{vs}}(\mathbb{S}) = \frac{1}{B}\sum_{i=1}^{B} \mathbb{1}\{\ell \in \mathcal{L}(s_i)\}. \tag{9}$$

Note that the summation in (9) is binomially distributed with parameters $B$ and $\theta_\ell$. It follows that the mean squared error (MSE) of $T^\ell_{\text{vs}}(\mathbb{S})$ is

$$\begin{aligned}\text{MSE}(T^\ell_{\text{vs}}(\mathbb{S})) &= E[(T^\ell_{\text{vs}}(\mathbb{S}) - \theta_\ell)^2], \\ &= \frac{1}{B}\theta_\ell(1-\theta_\ell). \end{aligned} \tag{10}$$

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

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

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

$$T^\ell_{\text{es}}(\mathbb{S}) = \frac{1}{B}\frac{\sum_{k=1}^{B}\mathbb{1}\{\ell \in \mathcal{L}(s_k)\}\deg^{-1}(s_k)}{\sum_{j=1}^{B}\deg^{-1}(s_j)}. \tag{11}$$

In particular, when node 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 to a Bernoulli with parameter $\pi_d$. As a result, the MSE for $B > 1$ independent samples is

$$\text{MSE}(T^d_{\text{es}}(\mathbb{S})) \;=\; \frac{1}{B}\pi_d(1-\pi_d) = \frac{1}{B}\frac{d\theta_d}{\bar{d}}\left(1 - \frac{d\theta_d}{\bar{d}}\right) \tag{12}$$

Equations (10) and (12) 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 node sampling yields better estimates than uniform edge sampling. This dichotomy is illustrated in Figure 12, 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 $5 \times 10^3$, the NRMSE associated with node 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

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

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

From (10), we have for uniform node sampling,

$$\text{NRMSE}(T^d_{\text{vs}}(\mathbb{S})) = \sqrt{(1/\theta_d - 1)/B}. \tag{13}$$

For $\theta_d \ll 1$, this implies

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

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

$$\log(\text{NRMSE}(T^d_{\text{vs}}(\mathbb{S}))) \approx \frac{\beta}{2}\log d + \frac{1}{2}(\log Z - \log B),\ 1 \leq d \leq \tau, \tag{14}$$

which explains the relationship observed for uniform node sampling in Fig. 12.

From (12), we have for uniform edge sampling,

$$\text{NRMSE}(T^d_{\text{es}}(\mathbb{S})) = \sqrt{(1/\pi_d - 1)/B}. \tag{15}$$

For $\theta_d \ll 1$, 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

$$\log(\text{NRMSE}(T^d_{\text{es}}(\mathbb{S}))) \approx \begin{cases} \frac{\beta-1}{2}\log d + \frac{1}{2}(\log Z + \log \bar{d} - \log B)\,, & 1 \leq d \leq \tau \\ -\frac{1}{2}\left(\log d - \log|E| - \log B\right)\,, & d > \tau, \end{cases} \tag{16}$$

which explains the linear increase followed by the linear decrease observed in Fig. 12. Although some RW-based methods can collect uniform node samples (e.g., via random jumps), NRMSE trends for large degrees are better described by (16) than by (14), since most of the information about these degrees comes from RW samples.

## 6. RESULTS ON NODE 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 node 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` [Ma et al. 2015]. In these datasets, node labels correspond to some type of group membership and a node is allowed to be part of multiple groups simultaneously. Figure 13 shows, on the left, the degree distribution for each network. On the right, it displays the relative frequency of each attribute in decreasing order (blue bars/dots) along with attribute frequency among the top 10% largest degree nodes (red bars/dots).

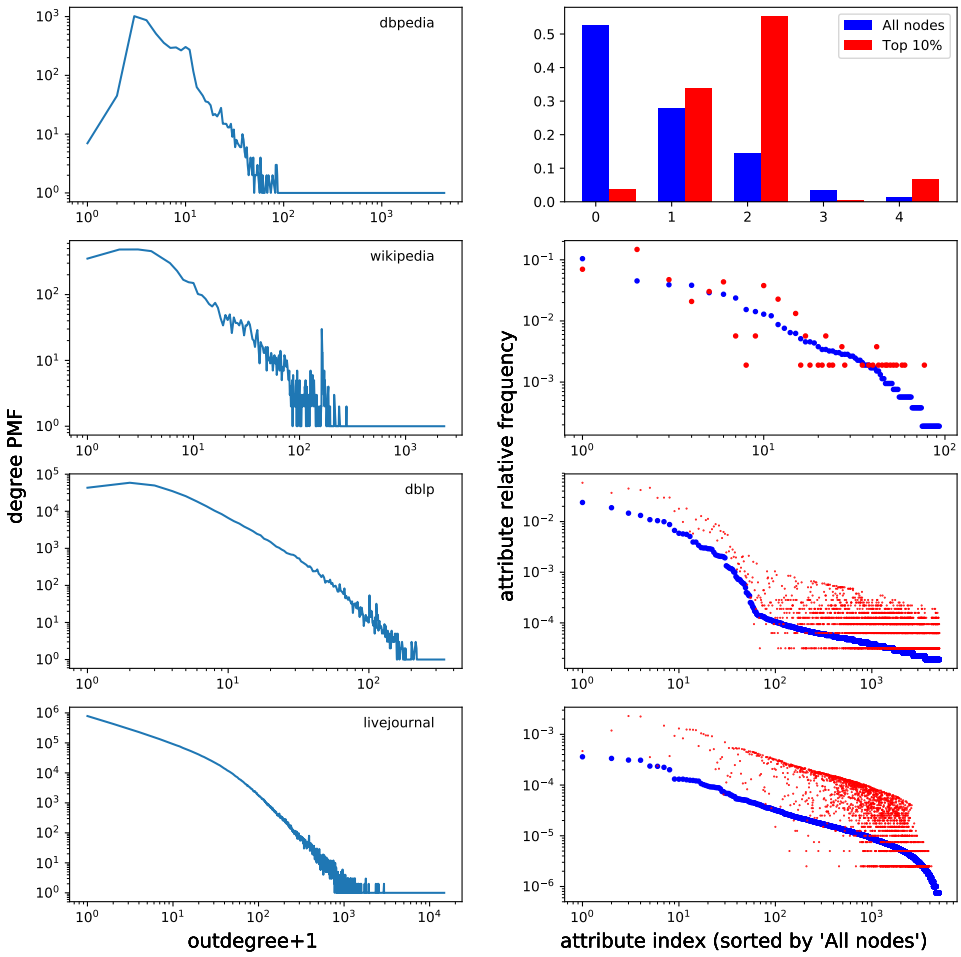
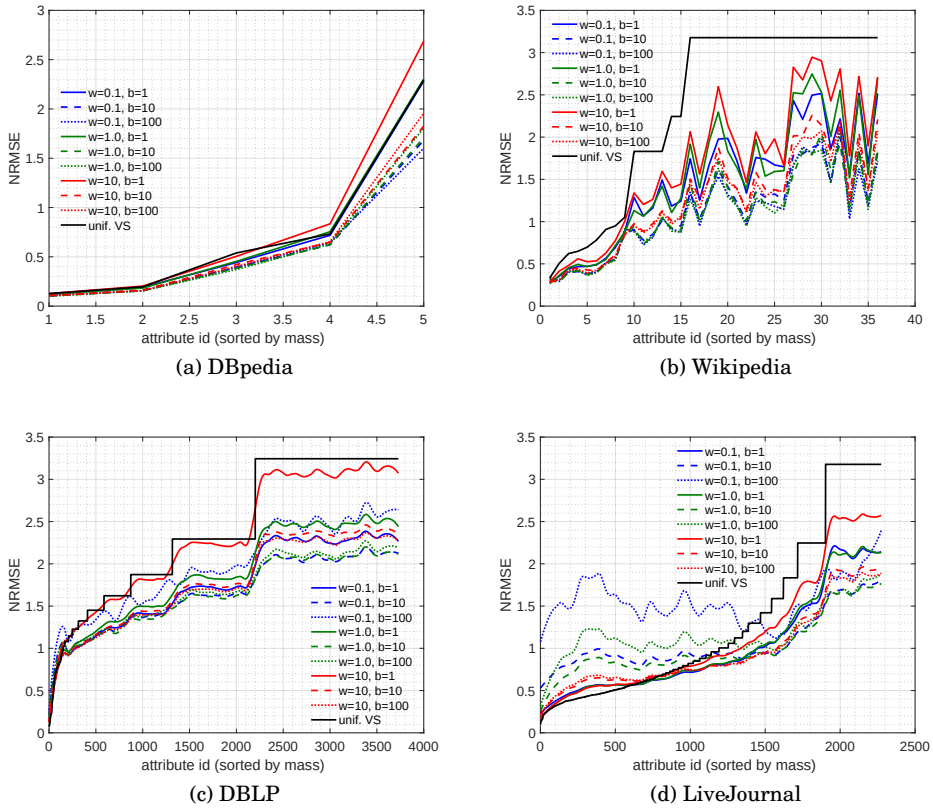



Fig. 13: Degree and node attribute distribution for undirected attribute-rich networks.




(a) DBpedia (b) Wikipedia

(c) DBLP (d) LiveJournal

Fig. 14: Comparison of hybrid estimator (DUFS) with uniform node sampling. 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. DUFS with $w \in \{0.1, 1.0\}$ and $b \in \{10, 10^2\}$ yields comparable or superior accuracy than uniform node sampling.

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

## 7. DISCUSSION: DUFS PERFORMANCE IN THE ABSENCE OF UNIFORM NODE 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:

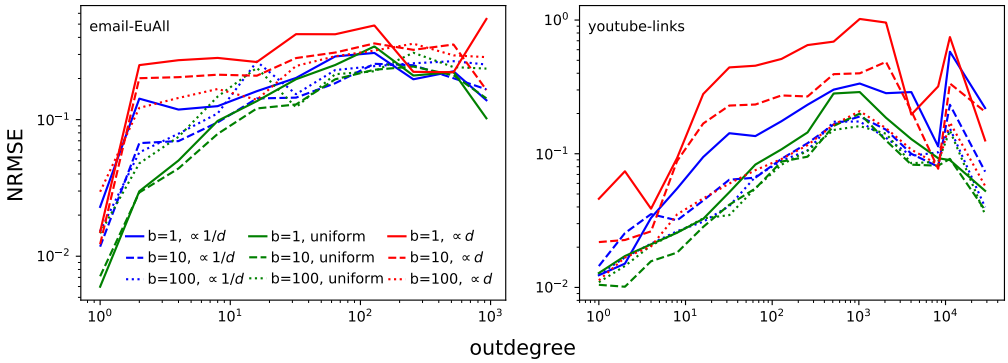



Fig. 15: Effect of initializing walkers non-uniformly over $V$ on E-DUFS accuracy. NRMSE decreases with budget per walker until $b = 10^2$.

— Distribution PROP: proportional to the undirected degree, that is,

$$P(\text{initial walker location is } v) = \frac{\deg(v)}{\sum_{u \in V} \deg(u)}; \tag{17}$$

— Distribution INV: proportional to the reciprocal of the undirected degree, that is,

$$P(\text{initial walker location is } v) = \frac{\deg^{-1}(v)}{\sum_{u \in V} \deg^{1}(u)}. \tag{18}$$

We simulate E-DUFS and 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 node 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. Figure 15 shows 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 the case of a single walker ($b = B - 1$) 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 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 [Leskovec and Faloutsos 2006; Hubler et al. 2008]. On the other hand, [Maiya and Berger-Wolf 2011] 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. [Achlioptas et al. 2009; Kurant et al. 2011b] demonstrate that Breadth-First-Search (BFS) introduces a large bias towards high degree nodes, and that is difficult to remove these biases in general, although they can be reduced if the network in question is almost random [Kurant et al. 2011b]. Random walk (RW) is biased to sample high degree nodes, however its bias is known and can be easily corrected [Ribeiro and Towsley 2010]. Random walks in the form of Respondent Driven Sampling (RDS) [Heckathorn 2002; Salganik and Heckathorn 2004] have been used to estimate population densities using snowball samples of sociological studies. The Metropolis-Hasting RW (MHRW) [Stutzbach et al. 2009] modifies the RW procedure to adjust for degree bias, in order to obtain uniform node samples. [Ribeiro and Towsley 2012; Chiericetti et al. 2016] analytically prove that MHRW degree distribution estimates perform poorly in comparison to RWs. Empirically, the accuracy of RW and MHRW has been compared in [Rasti et al. 2009; Gjoka et al. 2010] 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. [Avrachenkov et al. 2010] proves that random jumps increase the spectral gap of the random walk, which in turn, leads to faster convergence to the steady state distribution. [Kurant et al. 2011a] 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 the estimation of graph statistics of interest. To solve this problem, [Dasgupta et al. 2012] randomly samples nodes (either uniformly or with a known bias) and then uses neighborhood information to improve its unbiased estimator. [Zhou et al. 2016] 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 [Bar-Yossef and Gurevich 2008] and [Henzinger et al. 2000], 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 proposed in [Ribeiro et al. 2012] adapts the "backward edge traversal" of [Bar-Yossef and Gurevich 2008] to work with a pure random walk and random jumps. Both of these Metropolis-Hastings RWs ([Bar-Yossef and Gurevich 2008] and [Henzinger et al. 2000]) 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 [Leskovec and Faloutsos 2006] to sample large graphs. In [Leskovec and Faloutsos 2006], however, no technique is proposed 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 [Salehi and Rabiee 2013], but it assumes that obtaining uniform node 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 [McGregor 2014].

## 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. DUFS extends FS to make it applicable to directed networks when incoming edges are not directly observable by building on ideas from DURW. DUFS adapts DURW to use multiple coordinated walkers. Like DURW, DUFS can also be applied to undirected networks without any modification.

We also proposed a novel estimator for node label distribution that can account for FS and DUFS walkers initial locations – or more generally, uniform node samples – and a heuristic that can reduce the variance incurred by node 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 [Ribeiro and Towsley 2010], 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 uniform node 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 with 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 random jumps and use of the improved estimator. 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.

## Appendices

## A. HYBRID ESTIMATOR AND ITS STATISTICAL PROPERTIES

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

| | |
|---|---|
| $n_i$ | number of node 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 node samples |
| $M = \sum_i m_i$ | total number of edge samples |
| $B = N + M$ | total budget |

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).

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

$$L(\boldsymbol{\theta}|\mathbf{n},\mathbf{m}) = \frac{\prod_i \theta_i^{n_i} \prod_k (k\theta_{i,k})^{m_{i,k}}}{\left(\sum_{s,t} t\theta_{s,t}\right)^M}. \tag{19}$$

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

$$\hat{\theta}_{i,j} = \frac{\theta_i m_{i,j}}{j\mu_i}, \tag{20}$$

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

Substituting (20) in (19) yields

$$L(\boldsymbol{\theta}|\mathbf{n},\mathbf{m}) = \frac{\prod_i \theta_i^{n_i} \prod_k (\theta_i m_{i,k}/\mu_i)^{m_{i,k}}}{(\sum_s \theta_s \sum_z m_{s,z}/\mu_s)^M}. \tag{21}$$

The log-likelihood approximation is then given by

$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{n},\mathbf{m}) = -M \log \left( \sum_s \theta_s \sum_z \frac{m_{s,z}}{\mu_s} \right) + \sum_i \left( n_i \log \theta_i + \sum_k m_{i,k} (\log \theta_i + \log m_{i,k} - \log \mu_i) \right). \tag{22}$$

We rewrite $\theta_i$ 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

$$\mathcal{L}(\boldsymbol{\beta}|\mathbf{n},\mathbf{m}) = -M \log \left( \sum_s \frac{e^{\beta_s} m_s}{\mu_s} \right) + \sum_i (n_i + m_i)\beta_i - N \log \left( \sum_j e^{\beta_j} \right) + C, \tag{23}$$

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

The partial derivative w.r.t. $\beta_i$ is given by

$$\frac{\partial \mathcal{L}(\boldsymbol{\beta}|\mathbf{n},\mathbf{m})}{\partial \beta_i} = -\frac{M e^{\beta_i} m_i/\mu_i}{\sum_s e^{\beta_s} m_s/\mu_s} + n_i + m_i - \frac{N e^{\beta_i}}{\sum_j e^{\beta_j}}. \tag{24}$$

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

$$\theta_i^\star = \frac{n_i + m_i}{N + M \frac{m_i/\mu_i}{\sum_s \theta_s^\star m_s/\mu_s}}. \tag{25}$$

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

$$\hat{\theta}_i = \frac{n_i + m_i}{N + M \frac{m_i}{\mu_i \hat{d}}}, \tag{26}$$

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

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. (26), 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 node label is the undirected degree itself. We prove that this estimator, denoted by $\hat{\theta}_i$ is the minimum variance unbiased estimator (MVUE) of $\theta_i$.

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* $\theta_i$*.*

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 $\theta_i$.

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

PROOF. The log-likelihood equation for estimator (7) is given by

$$\begin{aligned} L(\boldsymbol{\theta}|\mathbf{n},\mathbf{m}) &= \frac{\prod_i \theta_i^{n_i} \prod_j (j\theta_j)^{m_j}}{\hat{\mu}^M} \\ &= \frac{\prod_j j^{m_j}}{\hat{\mu}^M} \prod_i \theta_i^{n_i+m_i}. \end{aligned} \tag{27}$$

We can see from eq. (27) 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 $\theta_i$ and the other factor, which does depend on $\theta_i$, depends on $\mathbf{n}$ and $\mathbf{m}$ only through $n_i + m_i$. From the Fisher-Neyman factorization Theorem [Lehmann et al. 1991], 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 $P_\theta$ parametrized by $\theta$. The statistic $s$ is said to be complete for the distribution of $X$ if for every measurable function $g$ (which must be independent of $\theta$) 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 $\theta_i$.*

PROOF.

$$\begin{aligned} E[g(n_i + m_i)] &= 0 \\ \sum_{n_i,m_i} g(n_i + m_i) P_\theta(n_i, m_i) &= 0 \\ \sum_{n_i,m_i} g(n_i + m_i) A(n_i) B(m_i) &= 0 \end{aligned} \tag{28}$$

The LHS of (28) is a polynomial of degree $M + N$ on $\theta_i$. Hence, it can be written as

$$C_0 + C_1\theta_i + C_2\theta_i^2 + \ldots + C_{N+M}\theta_i^{N+M} = 0. \tag{29}$$

We prove that $P_\theta(g(s(X)) = 0) = 1$ for all $\theta$ by contradiction. Suppose that there is a $\theta$ 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. (28) we note that the degree of the resulting polynomial is at least $n_i + m_i$ on $\theta_i$. Therefore, the coefficient $C_h$ in eq. (29) associated with $\theta_i^h$ 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 unbiased estimator $\bar{\theta}_i$ is the minimum variance unbiased estimator (MVUE) of $\theta_i$.*

PROOF. According to the Lehmann-Scheffe Theorem [Lehmann et al. 1991], 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 $\theta_i$. Clearly, the unbiased estimator $\hat{\boldsymbol{\theta}}$ in eq. (26) 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:

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

## ELECTRONIC APPENDIX

The electronic appendix for this article can be accessed in the ACM Digital Library.

## REFERENCES


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. DOI: http://dx.doi.org/10.1145/1538902.1538905

Konstantin Avrachenkov, Bruno Ribeiro, and Don Towsley. 2010. *Improving Random Walk Estimation Accuracy with Uniform Restarts*. Springer Berlin Heidelberg, Berlin, Heidelberg, 98–109. DOI: http://dx.doi.org/10.1007/978-3-642-18009-5_10

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

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. DOI: http://dx.doi.org/10.1016/j.physrep.2005.10.009

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. DOI: http://dx.doi.org/10.1145/2872427.2883045

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. DOI: http://dx.doi.org/10.1145/2339530.2339572

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. DOI: http://dx.doi.org/10.1073/pnas.0900282106

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. DOI: http://dx.doi.org/10.1109/INFCOM.2010.5462078

Douglas D. Heckathorn. 1997. Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations. *Social Problems* 44, 2 (1997), 174–199. DOI: http://dx.doi.org/10.2307/3096941

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. DOI: http://dx.doi.org/10.1525/sp.2002.49.1.11

Monika R. Henzinger, Allan Heydon, Michael Mitzenmacher, and Marc Najork. 2000. On near-uniform URL sampling. *Computer Networks* 33, 1-6 (2000), 295 – 308. DOI: http://dx.doi.org/10.1016/S1389-1286(00)00055-4

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. DOI: http://dx.doi.org/10.1109/ICDM.2008.124

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. DOI: http://dx.doi.org/10.1145/1993744.1993773

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.

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

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. DOI:http://dx.doi.org/10.1145/1150402.1150479

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

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. DOI:http://dx.doi.org/10.1145/1367497.1367591

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.

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. DOI:http://dx.doi.org/10.1145/2020408.2020431

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

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. DOI:http://dx.doi.org/10.1145/1298306.1298311

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. DOI:http://dx.doi.org/10.1109/JSAC.2013.130604

Fabricio Murai, Bruno Ribeiro, Don Towsley, and Pinghui Wang. 2018. *Characterizing Directed and Undirected Networks via Multidimensional Walks with Jumps*. Technical Report arXiv:1703.08252.

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. DOI:http://dx.doi.org/10.1109/INFCOM.2009.5062215

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. DOI:http://dx.doi.org/10.1109/INFCOMW.2010.5466698

Bruno Ribeiro and Don Towsley. 2010. Estimating and Sampling Graphs with Multidimensional Random Walks. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement (IMC '10)*. ACM, New York, NY, USA, 390–403. DOI:http://dx.doi.org/10.1145/1879141.1879192

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. DOI:http://dx.doi.org/10.1109/CDC.2012.6425857

Bruno Ribeiro, Pinghui Wang, Fabricio Murai, and Don Towsley. 2012. Sampling directed graphs with random walks. In *Proceedings of IEEE INFOCOM 2012*. 1692–1700. DOI:http://dx.doi.org/10.1109/INFCOM.2012.6195540

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. DOI:http://dx.doi.org/10.1109/JSAC.2013.130603

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

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.

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

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. DOI:http://dx.doi.org/10.1145/2847526