Non-backtracking random walks in network sampling algorithms

Sofija Stefanović *

*University of Oxford

Submitted to Proceedings of the National Academy of Sciences of the United States of America

Network sampling is a key area of network science applicable to problems ranging from surveying hard-to-reach populations, to web crawling and immunisation methods. In this project we explore the use of both simple random walks and biased random walks in the context of network sampling with a focus on the structural properties of the network. The main contribution of the project is the implementation of an integrated sampling method which combines the explicit encoding of non-backtracking in the sampling process, with biases based either on non-backtracking or eigenvector centrality. We find that the introduction of non-backtracking into biased random walks based on eigenvector and non-backtracking centrality performs at least as good as the centrality-based methods alone, and in some cases shows signs of slight advantage for measures of sample representativeness employed.

Non-backtracking | Sampling | Centrality

Significance Statement

The results on sampling methods presented in the project can be useful for understanding which network sampling strategies could benefit from random walks exploiting sampling with non-backtracking and non-backtracking centrality bias.

Introduction and Motivation

Network sampling first emerged as a means of surveying 'hidden' or hard-to-reach populations. The traditional problem of finding samples representative of a target population in unknown networks led to the development of network sampling strategies such as respondent-driven sampling (RDS) [1], which can be modelled by a random walk along the links in the network. In the context of population surveying, a sampling strategy relies on the local network of each participant in the survey to obtain a statistically significant estimate of a characteristic of interest in the target population through a sample (e.g. average age of people who are homeless). Aside from surveying, strategies based on random walks can be used for studying the structural properties of an unknown network rather than the attributes associated with the nodes. Moreover, such approaches are also useful when it is practically impossible to analyse or access the network in its entirety.

A distinction can be drawn between the *scale-down* and *back-in-time* sampling goals [2], whereby the former aim to obtain a sample which captures the structural properties of the full network, while the latter are concerned with reconstructing the state of the network at some point in the past. Random walks are found to be among the top performing methods for the scale-down goal, since they are biased towards high degree nodes and give sampled graphs that are connected [2]. In addition to simple random walks which have uniform probability for visiting a neighbouring node, advanced strategies based on random walks are employed to address various sampling challenges, such as web crawls, immunisation strategies [3], or studying influence maximisation [4].

Since the nodes in most real-world networks have different roles [5], making the transition probability to more important neighbours is often preferable for obtaining unbiased samples (e.g. in the cases of crawling a social network graph [6]). Moreover, in some contexts it may be preferable to introduce a constraint that no step can immediately go backwards (backtrack). While the motivation for such non-backtracking random walks has clear real-world interpretation in the context of surveying, the reasons can be more subtle in the context of sampling networks for network's sake. For instance, when non-backtracking is included the random walker does not visit a node more than $(1 + o(1))\frac{\log n}{\log \log n}$ times, as opposed to a simple random walk on the same graph which is likely to visit a node at least $\Omega(\log n)$ times [7], with n being the number of nodes in the network. Hence, a non-backtracking random walk is theoretically less likely to sample the same nodes multiple times, which could be an advantage for some sampling strategies.

In this project we consider how non-backtracking random walks can be used in random walk-based network sampling techniques. We employ a recently developed non-backtracking centrality-based random walk (NBCRW), and refine it by introducing the non-backtracking explicitly into the sampling procedure. Furthermore, the non-backtracking is explicitly introduced into a version of the eigenvector centrality-based random walk, and the performance of the two algorithms is compared alongside that of the unbiased random walk, and the biased random walk based on eigenvector centrality. The different sampling techniques are evaluated using a measure for similarity between degree distributions of the network and sample, a measure of hub inclusion into the sample, and a measure for how much of the network is discovered by the sampling procedure.

Background

Graphs and Random walks.

Networks and Sampling Concepts

Let $G\{V, E\}$ be a finite connected undirected network with the node set $V = \{1, 2, ..., N\}$ of N nodes, and edge set $E = \{(i, j)|i, j \in V\}$ of E edges. The links between the nodes

Reserved for Publication Footnotes

are given by the adjacency matrix $\mathbf{A} = (a_{ij})_{N\times N}$, with the element $a_{ij} = 1$ if $(i, j) \in E$, and $a_{ij} = 0$ otherwise. We denote the neighbours of node i with N_i, and the degree of node i with $d_i = |N_i| = \sum_{j=1}^{N} a_{ij}$. The diagonal matrix populated by node degrees is denoted as $\mathbf{D} = diag(d_1, d_2, ..., d_N)$.

A sample S is defined as a subset of nodes, $S \subset V$, and the neighbourhood of the sample is defined as $N(S) = \{j \in V - S, i \in S \text{ s.t. } (i, j) \in E\}$. The induced subgraph of G based on sample S is defined as $G_S = (S, E_S)$, where $E_S = (S \times S) \cap E$.

Simple Random Walk

A random walk can be defined as a succession of adjacent nodes from a start node to an end node. More formally, a discrete time nearest neighbour random walk on a graph G is a finite Markov chain characterised by the transition probability matrix $\mathbf{P} = \mathbf{D}^{-1}\mathbf{A}$ [8]. The elements of the matrix representing the transition probability for a walker at node i to one of its neighbours j are given by:

$$p_{ij} = \frac{a_{ij}}{d_i} \qquad [1]$$

Given the matrix $\mathbf{P}$ the stationary distribution of the random walk is defined as an N-dimensional vector π which satisfies $\mathbf{P}\pi = \pi$ and $\sum_{i=1}^{N} \pi_i = 1$. The hitting time from node i to node j for a random walk on G is defined as the expected number of steps required for the walker starting from source node i to arrive at the target node j for the first time [14]. The mean hitting time of a random walk is directly related to the stationary probability of residing at any node i. Hence, if the goal of the sampling process is to collect all the highest degree nodes in the sample, then intentionally biasing the random walk using centrality measures can be a strategy for achieving the desired result.

Maximum Entropy Biased Random Walks

One of the commonly used centrality measures, the eigenvector centrality, is given by the dominant eigenvector of the adjacency matrix $\mathbf{A}$. A random walk which uses eigenvector centrality as a source of bias is sometimes called the Maximal Entropy Random Walk (MERW) [10]. If $\alpha_1 > \alpha_2 > ... > \alpha_N$ are the real eigenvalues of $\mathbf{A}$, and $\psi_1, \psi_2, ...\psi_N$ the corresponding mutually orthogonal eigenvectors, where $\psi_i = (\psi_{i1}, \psi_{i2}, ...\psi_{iN})^T$ for $i = 1, 2, ..., N$, then the transition probability from node i to node j in MERW is defined by [11]:

$$p_{ij} = \frac{a_{ij}}{\lambda_1}\frac{\psi_{1j}}{\psi_{1i}} \qquad [2]$$

While eigenvector centrality is a widely used centrality measure, it suffers from a localisation transition in heterogeneous networks due to the presence of high degree nodes (hubs) [9]. More specifically, as the centrality of a node depends on the centrality of the neighbouring nodes, it can become artificially high for the hubs and their neighbourhoods at the expense of nodes with lower degrees in the rest of the network [12].

Non-backtracking Random Walks

A non-backtracking random walk on G is a random walk in which the step is not allowed to go back to the immediately previous state. In terms of the walk on the nodes, the walk can be defined as a sequence of nodes $(i_1, i_2, ..., i_k)$ with the node i_{m+1} chosen uniformly from the neighbours of i_m subject to the constraint $i_{m-1} \neq i_{m+1}$ [13]. Introducing the non-backtracking condition means the random walk between the nodes is no longer well-defined as a Markov process, as the node chosen at time step $t + 1$ will now depend on the nodes visited at time step t and $t - 1$. However, it is possible to define the non-backtracking or Hashimoto matrix $\mathbf{B}$ for the directed edges of the graph [13], such that the 'Markovianity' of the process is preserved. For any undirected network G we can replace each edge (i, j) with two directed edges $i \to j$ and $j \to i$, and define a 2E x 2E non-backtracking edge adjacency matrix $\mathbf{B}$, which describes the relation between the directed edges on G. The elements of B are given as:

$$B_{i\to j,k\to l} = \begin{cases} 1, & \text{if } j = k \text{ and } i \neq l. \\ 0, & \text{otherwise.} \end{cases} \qquad [3]$$

Given that all entries of $\mathbf{B}$ are non-negative real numbers, the Perron-Frobenius theorem states that the leading eigenvalue is real and non-negative, and guarantees the existence of a corresponding leading eigenvector with real non-negative entries. The non-backtracking matrix $\mathbf{B}$ will have positive eigenvalues when the graph G is not a tree, and zero when G is a tree. To ensure that the entries of the leading eigenvector can be chosen to be positive we assume that the graph G is not a tree in what follows.

Non-backtracking centrality was introduced as an attempt to avoid some of the deficiencies associated with eigenvector centrality. To avoid the feedback mechanism in the neighbourhood of a hub the centrality of edge $k \to l$ is calculated as the sum of the centralities of all edges incident on $k \to l$, excluding the edge $l \to k$ [15]. If κ is the leading eigenvalue of $\mathbf{B}$, and $v_{i\to j}$ the leading eigenvector entry corresponding to the directed edge $i \to j$, then the outgoing centrality is defined as [10]:

$$x_i = \sum_{j\in N_i} v_{i\to j} \qquad [4]$$

and corresponds to the non-backtracking centrality.

By lemma 3.2 in [10] there exists a relation between the leading eigenvalue and eigenvector of the non-backtracking matrix $\mathbf{B}$ and the corresponding values of a $2N \times 2N$ matrix:

$$\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{I - D} \\ \mathbf{I} & \mathbf{0} \end{pmatrix} \qquad [5]$$

with $\mathbf{I}$ the $N \times N$ identity matrix. The sampling strategies implemented below exploit this relation to compute the leading eigenvector of matrix $\mathbf{M}$ and extract the non-backtracking centrality x_i for node i from its i^{th} entry. The reduction in matrix size is a particularly significant factor for dense networks, where high edge counts would make it prohibitively computationally expensive to calculate the non-backtracking centrality.

Non-backtracking centrality based random walk. The random walk based on non-backtracking centrality biases the transition from node i to its neighbouring nodes using the values for the non-backtracking centrality of the neighbouring nodes. In other words, the probability for the walker to transition from node i to node j is proportional to the non-backtracking centrality of node j. The elements of the transition matrix for row i and column j are then:

$$p_{ij} = \frac{a_{ij}x_j}{\sum_{k=1}^{N} a_{ik}x_k} \qquad [6]$$

Furthermore it can be shown that the non-backtracking centrality based random walk (NBCRW) is equivalent to the simple random walk on the weighted network W with edges whose weight corresponds to the non-backtracking centrality [14].

Non-backtracking NBCRW and non-backtracking MERW

In the numerical simulations outlined in the following sections we modified the procedure for NBCRW and MERW such that both the bias towards high non-backtracking and eigenvector centrality is encoded, respectively, and the walker is explicitly prevented from backtracking during the sampling process whenever possible. When non-backtracking is explicitly encoded as part of the sampling, it is necessary to account for the dangling nodes in the network. This is to prevent the walker from getting 'stuck' upon reaching a node with a single neighbour, and the sampling process being interrupted before the desired sample size is reached. The routines implemented allow for the choice of whether the walker should backtrack only when it reaches a dangling node, or if these should be removed prior to the start of sampling. Given that dangling nodes do not form a substantial percentage of the networks analysed (see Table 1), they are removed for most of the evaluations presented. However, if the methods were to be used on networks with a significant proportion of dangling nodes, the effects of dangling removal and occasional backtrack on the quality of the sample obtained should be accounted for.

The choice to allow for backtracking during a sampling procedure has been applied for instance in estimating the clustering coefficient in social networks [17], yet the authors of this work performed the non-backtracking random walk with a uniform probability to visit the neighbours remaining after the previously visited node is removed. The approach taken in the project can be seen as having elements of the NBCRW defined in [10], and the straightforward implementation of non-backtracking for a concrete sampling walk.

Evaluating Sampling Techniques

In this section, we outline the different measures used to evaluate the sampling techniques. Since the aim is to understand which types of sampling problems the implemented techniques are suitable for, we focus on three measures that could serve as indicators for a variety of potential applications.

Degree distribution similarity

The degree distribution of a network is one of the fundamental properties that can be used to evaluate how representative a sample S is of the full network. Reproducing the degree distribution of the network in the sample would be desirable for scale-down goals, or inferring the degree distribution of a lagre network based on the distribution of the sample. Following earlier work on exploiting bias in sampling strategies [3] [2], we use the Kolmogorov-Smirvnov test to measure the distance between the degree distribution obtained from the degree sequence of the sample, and the degree sequence of the whole network. The D-statistic is defined as

$$D = \max_x \{|F(x) - F_S(x)|\} \qquad [7]$$

where x is the range of node degrees and $F(x)$ and $F_S(x)$ are the cumulative degree distributions for G and G_S. The similarity between the degree distributions is then calculated as $1 - D$ and the average over the sampling runs is plotted against the percentage size of the sample.

Hub inclusion

The hub inclusion measure is used as an indicator for how quickly the hubs are collected in the sampling process, which is a property of interest for instance for practical outbreak detection strategies [3]. A higher hub inclusion value for small sample sizes implies that the sampling strategy is more appropriate for identifying the most connected nodes in the network when the full network is either unknown, or not practically accessible. Our method identifies the hubs in the network and calculates the proportion of the sample made up of the nodes which belong to this set.

Discovery quotient

The *discovery quotient* is taken as measure of *network reach* [3], or roughly how much of the network is one step away from the sample obtained. The set of discovered nodes is defined as the nodes in the sample and their neighbours: $|S \cup (S)|$, and the *discovery quotient* is this value normalised by the number of nodes in the network: $|S \cup (S)|/|V|$. Obtaining a small sample with a high discovery quotient could be useful for instance for applications which require spreading information via social media efficiently and effectively.

Numerical results for model and real networks

In this section, we implement sampling strategies based on the simple non-backtracking walk (NBRW), the non-backtracking NBCRW (nbNBCRW), and non-backtracking MERW (nbMERW) on model networks and real networks, and compare the results of the degree distribution similarity, discovery quotient and hub inclusion with those of the SRW and MERW. Since samples of a given size are not unique, all results are taken as the average over many sampling runs for a given sample size.

Models and Datasets. To model heterogenous networks we work with the preferential attachment model of Barabaśi and Albert, which is used to explain the observed power-law degree sequence in a variety of real-world networks [5]. The power law behaviour $P(d) = d^{-\gamma}$ is observed in the high degree limit of the Barabaśi-Albert graph, with $\gamma = 3$ [5], which belongs to the $\gamma > 5/2$ regime in which localisation on the hub of the eigenvector centrality should occur [12]. Table 1 summarizes some of the properties of the empirical networks used to evaluate the sampling strategies.

All networks are treated as undirected and unweighted.

Table 1. Empirical networks

Network	Nodes	Edges	Density	% dangling
Dolphins	62	159	0.147	0
Jazz musicians	198	2,742	0.147	0
Facebook	4039	88234	0.011	0.019

Degree distribution similarity. The degree distribution similarity is compute for the Dolphins and Jazz networks. Figure 1 shows that at 10 percent of the total network size the nbNCRW and other centrality-based methods show around 70% similarity between the sample distribution and the original distribution, while with the sample size of approximately 20 percent of the total network size all methods lie in the $70\% - 85\%$ range (figure 1). In both cases, the SRW and nbRW are seen to perform better as the sample size increases to approximately 25% of the sample, with the difference between the two methods and the rest being more pronounced in the Dolphin network.

The choice of Kolmogorov-Smirnov test might not be the best choice for all distributions, as it puts more emphasis on

the bulk of the distribution, rather than the tails. An alternative measure could be based on some version of the Anderson-Darling test, which gives more weight to the tails than the KS test [19].

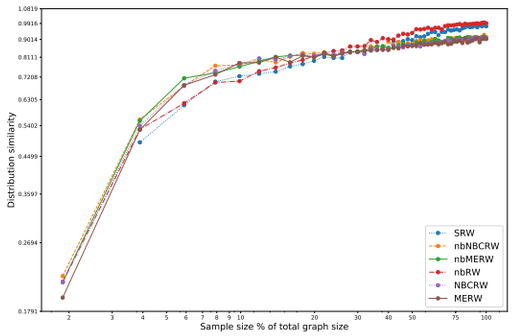

Fig. 1. Distribution similarity for the Dolphins network for 50 sample sizes, and averaged over 50 sampling runs.

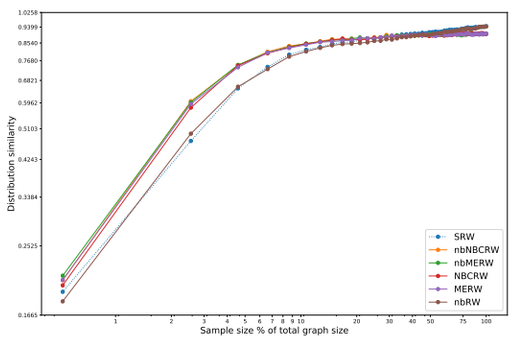

Fig. 2. Distribution similarity for the Jazz network for 50 sample sizes, and averaged over 50 sampling runs.

Discovery Quotient. The discovery quotient for the BA graph with a 1000 nodes and the number of nodes connecting to every incoming node $m = 4$ (see figure 3) features a gap between the centrality-based random walks and the SRW and nbRW for sample sizes up to approximately 10 percent of the size of the network. The nbRW discovers more of the network in smaller samples as compared to SRW, and this trend is visible amongst the centrality-based RWs as well. Introducing the non-backtracking into the centrality-based sampling techniques leads to a slight advantage with respect to the DQ in small-sized samples, but the primary advantage still comes from the introduction of the centrality bias in the algorithms. In contrast to trends present for small sample sizes, the nbRW seems to perform best for larges sample sizes, but the differences are far less conclusive at this scale.

Figure 4 shows a similar pattern, with the gap being less pronounced than in the case of the BA network. The advantage of the centrality-based methods could be due to these having a lower hitting time to the network hubs in comparison to the non centrality-based random walks [10]. If the hub is identified quickly, the discovery quotient is expected to be high even for very small sample sizes, since the hub itself will provide routes to the a large portion of the network. Hence, MERW, nbMERW, NBCRW and nbNBCRW are able to *discover* a larger proportion of the network in a small number of steps, while the discovery is more gradual for SRW and nbRW.

Due to the size of the Facebook network, the maximum sample size is taken as 50% of the full network, and the number of sampling runs for averaging is lower. While centrality-based methods are seen to dominate for smaller sample sizes (up to approximately 5% of the network) as in the previous cases, the SRW and nbRW become dominant as the sample size grows past 10% of the full network size.

While the discovery quotient of SRW and nbRW increases gradually with the sample size, the quotient for the rest does not change substatntially as the sample size grows (the variations are likely to smooth out with more robust averaging). Hence, the centrality-based sampling algorithms never *discover* the full network, which is likely to be a side-effect of the community structure in the Facebook network high centrality nodes are likely to be found embedded in communities, and hence centrality-based methods have a higer probability of getting 'stuck' in a community.

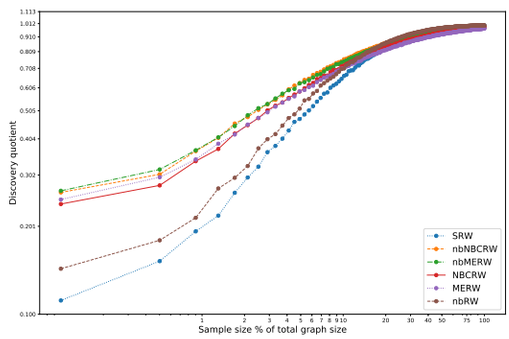

Fig. 3. Discovery quotient in a BA network with 1000 nodes and $m = 4$ for 250 different sample sizes, averaged over a 100 sampling runs for each sample size. Loglog view shows the ranking of the methods for small sample sizes.

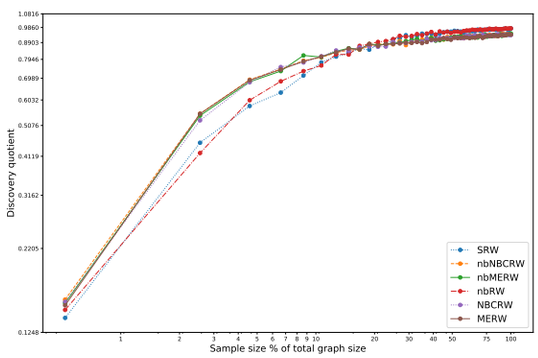

Fig. 4. Discovery quotient in the Jazz network for 50 different sample sizes, averaged over a 50 sampling runs for each sample size.

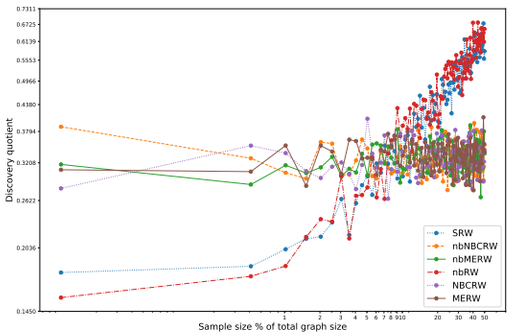

Fig. 5. Discovery quotient for the Facebook network for 100 sample sizes up to 50% of the network size, averaged over 20 sampling runs.

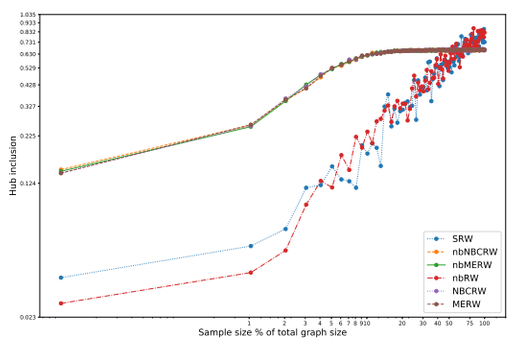

Fig. 6. Hub inclusion for the Facebook network without dangling nodes calculated for 100 sample sizes and averaged over 20 sampling runs for each size.

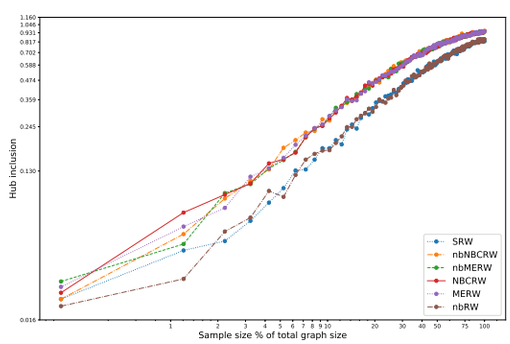

Fig. 7. Hub inclusion for top 5 hubs in a BA netowrk with 1000 nodes and $m = 4$ for 100 different sample sizes, averaged over a 20 sampling runs for each sample size.

Hub inclusion. Hub inclusion is higher for smaller-sized samples for the NBCRW and MERW based algorithms for both the synthetic BA model (figure 7) and the empirical Jazz and Facebook networks, which is consistent with the observation of the NBCRW and MERW having lower mean hitting times to the hub nodes in the network as compared to the SRW [14]. The NBRW exhibits similar behaviour to the SRW, which can be understood on the grounds that neither NBRW nor SRW chose the next node based on some centrality notion, and hence the probability for these strategies to include a high proportion of hubs in small samples is lower.

Although centrality-based algorithms have higher hub inclusion value at lower sample sizes for all three networks, they approach different values in the limit of the sample size becoming of the order of the full network size. All algorithms for the BA and Jazz network eventually reach the hub inclusion value close to one, while the centrality-based algorithms for the Facebook network are seen to plateau from sample sizes below 20 percent of the network size onward. Such trend could be due to community structure which is necessarily present in the Facebook network. While all random walks are susceptible to oversampling a community due to the higher density of edges present within it, this susceptibility can exaggerated/heightened in methods based on centrality if the nodes with high centrality are embedded into communities, and do not have a high density of links to other communities. Hence, the biased random walker can end up being 'stuck' in one part of the network. The $max(hub\_inclusion) < 1$ value for the Facebook network implies that the sampling algorithms never collects all hubs in the sample, which is consistent with the findings that network sampling is particularly challenging in the presence of communities [3]. The absence of such a plateau in the Jazz network could be due to its relatively small size and higher density compared to that of Facebook (see table 1), which is high enough for the walker to have a non-negligible chance of escaping a community to include hubs found in other parts of the network.

Throughout the discussion we have assumed that the leading eigenvector, and hence the non-backtracking centrality was well-defined for all networks under consideration, and we have - effectively ignored the question of the irreducibility and non-uniqueness of the largest non-backtracking eigenvector. This is addressed explicitly in [15] via the introduction of a new centrality measure α, which interpolates between the eigenvector and non-backtracking centrality, and could serve as a useful centrality measure for the analysis of sampling on general networks.

Non-backtracking based random walks we focused on do not account for without replacement sampling, which may be a feature of realistic sampling scenarios. Although forbidding repetition of nodes would account for this requirement, it would not be possible to re-define such sampling as a Markov process, as the walker would require the full knowledge of which nodes were visited since the beginning of the walk to make every step.

Conclusion and Future Work

Combining non-backtracking and centrality bias is seen to perform at least as well as the centrality-based sampling methods on the measures of discovery quotient, hub inclusion and distribution similarity which were considered in this project.

The higher discovery quotient for nbNCRW and nbMERW suggests introducing non-backtracking in centrality-based random walks could offer a slight advantage in cases where only a fraction of the network is accessible, or in case of the network needing to be scaled down to only a fraction of its size.

The results show no identifiable differences between the methods based on non-backtracking versus eigenvector centrality as a result of the localisation that occurs for eigenvec-

tor centrality for the networks and measures considered here. To see if and what kind of effect this may have on sampling processes, we could choose real-world networks based on their power-law exponent, and compare the results of performing sampling methods for networks with exponents which should lead to localisation, and networks in which we do not expect this.

To observe the differences between the performance of the biased random walk strategies in greater detail, these algorithms should be optimised for performance on much larger networks (e.g. of the order of samples obtained from real web crawls). Finally, a more comprehensive study into the suitability of non-backtracking and non-backtracking centrality-based sampling techniques should incorporate measures which account for community structure present in real-world networks, such as measures of community reach.

1. D. D.Heckathorn, and C. J.Cameron, Network Sampling: From Snowball and Multiplicity to Respondent-Driven Sampling, SIAM REVIEW, 43 (2017), pp. 101–119.
2. Mcauley, Jim and Leskovec, Jure, Sampling from Large Graphs, Scientific Reports, 6 (2006).
3. A. S. Maiya and T. Y. Berger-Wolf, Benefits of Bias: Towards Better Characterization of Network Sampling, Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (2011), pp. 105–113.
4. Pan, Jingzhi and Jiang, Fei and Xu, Jin, Influence Maximization in Social Networks Based on Non-backtracking Random Walk, 19 (2016) pp. 3677-3686.
5. M. E. J. Newman, The structure and function of complex networks, SIAM REVIEW, 45 (2003), pp. 167–256.
6. M. Gjoka, M. Kurant, and T. Butts, Carter and Markopoulou, Athina, Walking in Facebook: A Case Study of Unbiased Sampling of OSNs, INFOCOM, 2010 Proceedings IEEE, (2010), pp. 1–9.
7. Alon, Noga and Benjamini, Itai and Lubetzky, Eyal and Sodin, Sasha, Non-backtracking random walks mix faster, Commun Contemp Math, 9 (2006).
8. Lovsz, Lszl, Random Walks on Graphs: A Survey, Combinatorics, Paul Erdos is Eighty, Bolyai Soc. Math. Stud., 2 (1993).
9. M. Travis, X. Zhang and M. E. J. Newman, Localization and centrality in networks, CoRR, abs/1401.5093 (2014).
10. Y. Lin and Z. Zhang, Non-Backtracking Centrality Based Random Walk on Networks, CoRR, abs/1803.03087 (2018).
11. Burda, Zdzislaw and Duda, Jarek and Luck, Jean-Marc and Waclaw, Bartek, Localization of the Maximal Entropy Random Walk, Physical review letters, 102 (2009).
12. R. Pastor-Satorras and C.-Castellano, Distinct types of eigenvector localization in networks, Scientific Reports, 6 (2015).
13. M. Kempton, Non-backtracking random walks and a weighted Iharas theorem, Open J. Discrete Math., 6 (2016), pp. 207–226.
14. Lin, Yuan and Zhang, Zhongzhi, Non-Backtracking Centrality Based Random Walk on Networks, CoRR, 1803.03087 (2018).
15. R. Criado, J. Flores, E. Garca, A. J. Garca del Amo, . Prez and M. Romance, On the -nonbacktracking centrality for complex networks: Existence and limit cases, J. Comp. and Appl. Math., 350 (2019), pp. 35–45.
16. Leskovec, Jure and Faloutsos, Christos, Learning to discover social circles in ego networks, NIPS, 1 (2012), pp. 539–547.
17. Katzir, Liran and Hardiman, Stephen J., Estimating Clustering Coefficients and Size of Social Networks via Random Walk, ACM Trans. Web, 9 (2015), pp. 19:1–19:20.
18. Lee, Chul-Ho and Xu, Xin and Eun, Do Young, Beyond Random Walk and Metropolis-hastings Samplers: Why You Should Not Backtrack for Unbiased Graph Sampling, SIGMETRICS Perform. Eval. Rev., 40 (2012), pp. 319–330.
19. C.D. Sinclair and B.D. Spurr and M.I. Ahmad, Modified anderson darling test, Communications in Statistics - Theory and Methods, 19 (1990) pp. 3677-3686.