Empirical comparison of network sampling: How to choose the most appropriate method?

Neli Blagus, Lovro Šubelj, Marko Bajec

University of Ljubljana, Faculty of Computer and Information Science, Ljubljana, Slovenia

Abstract

In the past few years, the storage and the analysis of large-scale and fast evolving networks presents a great challenge. Therefore, a number of different techniques have been proposed for sampling large networks. Studies on network sampling primarily analyze the changes of network properties under the sampling. In general, network exploration techniques approximate the original networks more accurate than random node and link selection. Yet, link selection with additional subgraph induction step outperforms most other techniques. In this paper, we apply subgraph induction also to random walk and forest-fire sampling and evaluate the effects of subgraph induction on the sampling accuracy. We analyze different real-world networks and the changes of their properties introduced by sampling. The results reveal that the techniques with subgraph induction improve the performance of techniques without induction and create denser sample networks with larger average degree. Furthermore, the accuracy of sampling decrease consistently across various sampling techniques, when the sampled networks are smaller. Based on the results of the comparison, we introduce the scheme for selecting the most appropriate technique for network sampling. Overall, the breadth-first exploration sampling proves as the best performing technique.

Keywords: complex networks, network sampling, comparison of sampling techniques, subgraph induction, sampling accuracy, sampling selection scheme

1. Introduction

Real-world networks are often large and fast evolving. Therefore, not only their storage poses a problem, but their analysis and understanding present a great challenge. In the past few years, a number of different techniques have been proposed for sampling large networks [1, 2]. With sampling, a network is reduced to a smaller sample, suitable for efficient analysis [3, 4] and visualization [5, 6]. Moreover, the knowledge of sampling process in networks can help to understand the network evolution [7] or to train prediction models in link prediction [8, 9]. However, the data about analyzing networks can be incomplete or networks can change quickly over time. Hence, it is of key importance

Email addresses: neli.blagus@fri.uni-lj.si (Neli Blagus), lovro.subelj@fri.uni-lj.si (Lovro Šubelj), marko.bajec@fri.uni-lj.si (Marko Bajec)

to understand the difference between complete original networks and their incomplete variants.

Therefore, a number of studies on network sampling analyze the changes of network properties under the sampling. For example, the preservation of degree distribution [10], clustering distribution [10] or network connectivity [11]. Other studies analyze sampling of specific sorts of networks, like sampling social [12] and online-social networks [13], scale-free [14] and temporal networks [15] or weighted [16] and directed networks [17]. However, despite the efforts, the changes in network structure introduced by sampling are still far from understood. Only a few studies focus on comparing the performance of different sampling techniques. Leskovec et al. [1] observed network properties of the original and sampled networks and compare them based on Kolmogorov-Smirnov D-statistic. Lee et al. [2] analyzed how different sampling techniques under- or overestimate network properties. Doer and Blenn [18] observed the convergence of network properties across different sampling techniques. Hübler et al. [19] compared the properties of the original and sampled networks based on different metrics, like Kolmogorov-Smirnov D-statistic and L_1 norm. On the other hand, Toivonene et al. [20] proposed a family of compression methods for directed networks and compare them based on the time and space complexity. In our previous study [21] we compare different sampling techniques based on the preservation of network community structure. The results reveal that the sampled networks exhibit stronger characterization by community-like groups than the original networks; the changes in the node group structure occur consistently across different sampling techniques.

In general, network exploration techniques like random walk and forest-fire sampling approximate the original networks more accurately than random node and link selection [1]. Yet, Ahmed et al. [22] proposed the link selection with additional subgraph induction step, where the sampled network consists of randomly selected links (i.e., random link selection) and any additional links between their endpoints (i.e., subgraph induction). Not only the performance of the random link selection is improved, but the proposed technique also outperforms several other sampling techniques [22, 23].

In this paper, we apply subgraph induction also to random walk and forest-fire sampling. We evaluate the effects of subgraph induction on the sampling accuracy. In addition, we provide a comparison of various sampling techniques and their suitability for preserving different network properties, including degree and clustering distribution, average degree and density. The results reveal that in most cases the techniques with subgraph induction step improve the performance of techniques without induction. Based on the results of the comparison, we introduce the sampling selection scheme, which supports the selection of the most appropriate technique for sampling particular network.

The rest of the paper is structured as follows. In Section 2, we present the background on network sampling and expose the sampling techniques used in the study. The results of the empirical analysis are reported and formally discussed in Section 3, while Section 4 concludes the paper and suggests directions for further research.

2. Network sampling

Let the network be represented by a simple undirected graph $G = (V, E)$, where V denotes the set of nodes ($n = |V|$) and E is the set of links ($m = |E|$). The goal of network sampling is to create a sampled network $G' = (V', E')$, where $V' \subset V$, $E' \subset E$

Table 1: Abbreviations for the sampling techniques.

RNS	Random node selection
RND	Random node selection based on degree
RLS	Random link selection
RLI	Random node selection with subgraph induction
BFS	Breadth-first exploration sampling
FFS	Forest-fire sampling
FFI	Forest-fire sampling with subgraph induction
RWS	Random walk sampling
RWI	Random walk sampling with subgraph induction
MHRW	Metropolis-Hastings random walk rampling
MHRWI	Metropolis-Hastings random walk sampling with subgraph induction

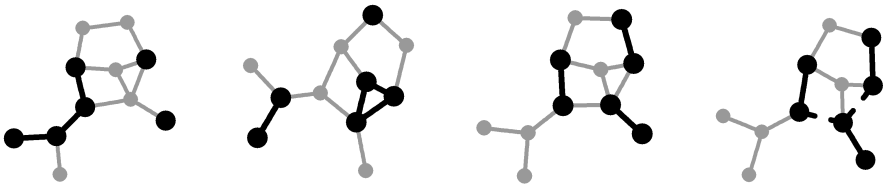
and $n' = |V'| << n$, $m' = |E'| << m$. The sample G' is obtained in two steps. In the first step, nodes or links are sampled using a particular strategy like random selection and network exploration sampling. In the second step, the sampled nodes and links are retrieved from the original network. The sampled network is called a subgraph of the original network, if it consists of sampled nodes or sampled links only. Otherwise, if sampled nodes and all their mutual links are included in the sample or the sample consists of sampled links and any additional links among their endpoints, the sampled network is called an induced subgraph of the original network (i.e., subgraph induction).

A large number of sampling techniques have been proposed in past years, suitable for various purposes and for matching different network properties. The sampling techniques can be roughly divided into two categories: random selection and network exploration techniques. In the first category, nodes or links are included in the sample uniformly at random or proportional to some particular characteristic like degree or PageRank [1]. In the second category, the sample is constructed by retrieving a neighborhood of a randomly selected seed node using different strategies like breadth-first search [2] or random walk [1]. The sampling techniques used in this paper are listed in Table 1 with their abbreviations.

2.1. Random selection

For the purpose of this study, we consider four techniques from the random selection category. We first adopt random node selection [1] (RNS), where the sample consists of nodes selected uniformly at random and all their mutual links (Fig. 1(a)). RNS accurately approximates the degree mixing [2] and preserves the relationship of transitivity and density between the original and sampled networks [24]. Moreover, it shows better performance on larger samples than on smaller [24]. Yet, RNS overestimates the degree and betweenness centrality exponent and fails to match the clustering coefficient [2], degree distribution [14] and the average path length [25] of the original network.

Furthermore, we adopt random node selection by degree [1] (RND), which improves the performance of RNS. Here, the nodes are selected randomly with probability proportional to their degrees and all their mutual links are included in the sample (Fig. 1(b)). RND matches in-degree and out-degree distributions and also spectral properties of the original network better than RNS [1]. Besides, it constructs samples with larger weakly connected component [24]. Nevertheless, despite a fully connected original network, both RNS and RND can construct a disconnected sampled network.



(a) RNS (b) RND (c) RLS (d) RLI

Figure 1: Random selection techniques applied to a small toy network. Highlighted nodes and links represent the samples obtained by different techniques. (a) In random node selection, the sample consists of nodes selected uniformly at random and all their mutual links. (b) In random node selection by degree, the nodes are selected to the sample with probability proportional to their degree, while all their mutual links are included in the sample. (c) In random link selection, links are selected to the sample uniformly at random. (d) In random link selection with subgraph induction, the sample consists of randomly selected links (solid lines) and also any additional links between their endpoints (dashed lines).

Next, we adopt random link selection [1] (RLS), where the sample consists of links selected uniformly at random (Fig. 1(c)). RLS matches well degree mixing [2] and the distribution of sizes of weakly connected components [1]. It constructs sparse samples and accurately approximates the average path length of the original network [23]. Yet, RLS fails to match most of other network properties [1]. RLS overestimates the degree and betweenness centrality exponent and underestimates the clustering coefficient [2].

We last adopt random link selection with subgraph induction [22] (RLI), which improves the performance of RLS. Here, the sample consists of links selected uniformly at random and any additional links between the endpoints of the sampled links (Fig. 1(d)). RLI outperforms several other techniques in matching the degree, path length and clustering coefficient distribution of the original networks [22]. It selects nodes with higher degree more likely than other random selection techniques, which increases the connectivity of the sample. Moreover, RLI is suitable for sampling large networks that can not fit into the main memory and can also be implemented as a technique for sampling streaming networks [23].

2.2. Network exploration

We consider seven sampling techniques from the network exploration category (note that in the literature, this category of sampling techniques is also called topology based sampling [22], traversal based sampling [26] or link-trace sampling [27]). First, we adopt breadth-first exploration sampling [28] (BFS), where a seed node is selected uniformly at random, while its broad neighborhood retrieved from the basic breadth-first search is included in the sample. BFS is biased towards selecting nodes with higher degree [28], yet it underestimates the degree and betweenness centrality exponent [2]. Second, we adopt random walk sampling [1] (RWS), where the random walk is simulated on the network, starting at a randomly selected seed node (Fig. 2(b)). The sample consists of links, which are visited by a random walker and represents a connected subgraph of the original network. RWS outperforms random selection techniques in matching the transitivity [27], clustering coefficient distribution and spectral properties and also shows good performance on smaller samples [1]. Yet, RWS is biased towards selecting nodes with high degree [29] and fails to match the degree distribution [28].

Next, we adopt forest-fire sampling [1] (FFS). Here, a broad neighborhood of a randomly selected seed node is retrieved from partial breadth-first search (Fig. 2(d)). The number of links sampled on each step is selected from a geometric distribution with mean $p/(1-p)$, where p is set to 0.7 [1]. Thus, on average 2.33 links are included in the sample on each step. FFS matches well spectral properties [1] and together with RWS shows the best overall performance among several techniques [1]. However, FFS fails to match the path length and clustering coefficient of the original networks [23].

Moreover, we apply subgraph induction step to random walk and forest-fire sampling, which we term random walk sampling with subgraph induction (RWI) and forest-fire sampling with subgraph induction (FFI). Here, the samples consist of links, sampled with random walk (Fig. 2(c)) or forest-fire (Fig. 2(e)), while also any additional links among the endpoints of the sampled links are included in the sample. To the best of our knowledge, RWI has not been analyzed in any of the previous studies. On the other hand, FFI shows worse performance than RLI in matching the path length, degree and clustering distributions [23]. Still, the performance of FFI has not yet been compared to a larger set of sampling techniques.

Additionaly, we adopt Metropolis-Hastings random walk [30] (MHRW), where the random walk is simulated on the network, starting at a randomly selected seed node (Fig. 2(b)). On each step, the next-hop node is selected uniformly at random among neighbours of current node or random walker performs a self-loop instead of moving to other node. MHRW correct the bias of RWS, where the nodes with higher degree are more likely to be selected to the sample [31]. MHRW performs better on well connected networks and performs poorly on the networks with high community structure, since it frequently stuck in a local community [32].

Lastly, we apply subgraph induction step to MHRW, which we term Metropolis-Hastings random walk with induction (MHRWI). Here, the sample consists of links, sampled with MHRW (Fig. 2(c)), while any additional links among the endpoints of sampled links are also included in the sample. To the best of our knowledge, MHRWI has not been analyzed in any of the previous studies.

3. Analysis and discussion

In the following sections, we describe the adopted social, biological and information networks (Section 3.1), report the results of the empirical analysis and discuss the findings (Section 3.2). Last, we combine the results in the sampling selection scheme (Section 3.3).

3.1. Network data

The analysis is performed on 12 social and information networks. Their main characteristics are presented in Table 2. In collaboration networks, the nodes represent the authors, while undirected links denote that two authors co-authored at least one paper together. The *ca-hep* [33] and *ca-astro* [33] are collaboration networks among researchers, who submitted their papers to arXiv High Energy Physics and Astro Physics category respectively. The *ca-dblp* [34] is a collaboration network among the authors of papers in computer science. Biological networks *yeast* [35] and *human* [35] are the interaction-interaction networks among proteins in *S. cerevisiae* and *H. sapiens*, respectively. The nodes in both networks represent proteins, while the links denote interactions among

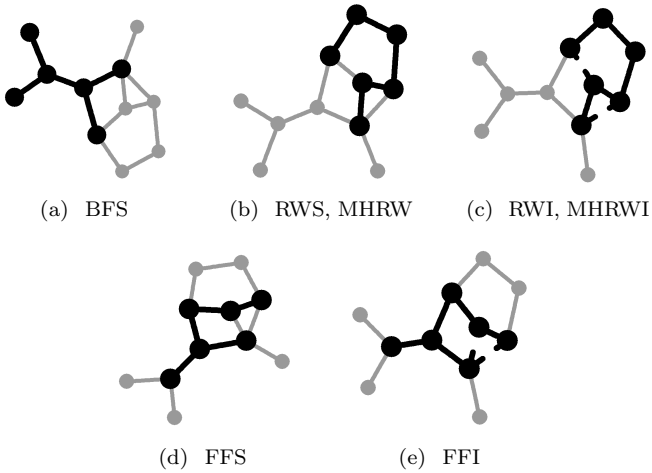


Figure 2: Network exploration techniques applied to a small toy network. Highlighted nodes and links represent the samples obtained by different techniques. (a) In breadth-first exploration sampling the broad neighborhood of a randomly selected seed node is retrieved using breadth-first search. (b) In random walk sampling, the sample consists of links, retrieved from a simulation of a random walker on the network starting at a randomly selected seed node. In Metropolis-Hastings random walk, the random walk is performed similarly to random walk sampling with the possibility of performing a self-loop instead of moving to the other node on each step.(c) In random walk and Metropolis-Hastings random walk sampling with subgraph induction, the sample consists of links selected with random walk sampling or Metropolis-Hastings random walk (solid lines) and also any additional links between their endpoints (dashed lines). (d) In forest-fire sampling, the broad neighborhood of a randomly selected seed node is retrieved using partial breadth-first search, where only a fraction of links is included in the sample on each step. (e) In forest-fire sampling with subgraph induction, the sample consists of links selected with forest-fire sampling (solid lines) and also any additional links between their endpoints (dashed lines).

them. The *cit-hep* [36] is a citation network from the arXiv category High Energy Physics, where the nodes represent papers and the links denote that papers cite each other. The networks *brightkite* [37], *slashdot* [38] and *youtube* [39] are social networks from providers Brightkite, Slashdot and Youtube, respectively. The nodes represent users, while the links denote friendships between them. The *email* [33] network is created using email data from the European Research Institution, where email addresses are linked, if at least one message was sent among them. The *nd.edu* [40] network is the web graph of *nd.edu* domain, where the nodes represent web pages and the links mean hyperlinks between pages. The *flickr* [41] network contains images from image hosting website Flickr. The nodes represent images, while the links denote that the images share the same metadata, for example location, the author or the album of the image.

All networks are considered to be undirected, although some of them are directed. We consider sample sizes from 0.2% to 20% of the original networks (0.2-1% by step of 0.2% and 2-20% by step of 2%). The exception are MHRW and MHRWI sampling techniques, where we limit the number of steps in the algorithm to 0.2% to 20% of the original networks sizes and thus the samples obtained with MHRW and MHRWI are smaller. For each network we perform 100 realizations of each sampling technique and each sample size. For each run of the exploration techniques, the sample was constructed

Table 2: Real-world networks considered in the study.

Network	Nodes	Links	Average degree	Clustering coefficient	Density
yeast	5,717	48,259	16.9	0.068	2.9×10^{-3}
ca-hep	12,008	237,010	39.5	0.660	3.3×10^{-3}
human	15,921	220,019	27.6	0.021	1.7×10^{-3}
ca-astro	18,772	396,160	42.2	0.318	2.2×10^{-3}
cit-hep	27,240	342,437	25.1	0.120	9.2×10^{-4}
brightkite	58,228	214,078	7.4	0.111	1.3×10^{-4}
slashdot	82,168	948,464	23.1	0.024	2.8×10^{-4}
flickr	105,938	2,316,948	43.7	0.402	4.1×10^{-4}
email	265,214	420,045	3.2	0.004	1.2×10^{-5}
ca-dblp	317,080	1,049,866	6.6	0.306	2.1×10^{-5}
nd.edu	325,729	1,497,134	9.1	0.097	2.8×10^{-5}
youtube	1,134,890	2,987,624	5.2	0.006	4.5×10^{-6}

from a new randomly selected seed node.

3.2. Empirical analysis

We observe the different properties of the original and sampled networks, including listed four:

- Degree distribution refers to the probability distribution of degrees of all nodes in the network. The degree of a node is the number of node neighbours, the quantity p_k is the fraction of nodes having degree k, $k > 0$. The quantities p_k represents degree distribution of the network.
- Distribution of clustering coefficient refers to the probability distribution of the proportions of connected neighbors of each node [42]. The clustering coefficient of a node is the ratio of the triangles connected to the node and the maximum number of triangles that could pass through the node, while for distribution a frequency count of the occurrence of each clustering coefficient is provided.
- Average degree refers to the average number of neighbours of nodes over the whole network.
- Density refers to the ratio of existing links to all possible links among all the nodes in the network.

We first analyze the performance of sampling techniques based on the match of the degree and clustering coefficient distributions between the original and sampled networks. To compare the distributions, we use Kolmogorov-Smirnov D-statistic, which is commonly used in similar studies [1, 23, 24]. Kolmogorov-Smirnov test checks the null hypothesis that the distributions of property of the original network and its sampled variant are the same, while the D-statistic measures the distance between the observed distributions. Next, we analyze the performance of sampling techniques in matching the average degree and density between the original and sampled networks. We compare both properties according to the actual values in original and sampled networks. Finally, we compare sampling techniques for each property with the assessment approach proposed in [24], where the sampling techniques are ranked based on the similarities

between the original and sampled networks. In particular, the proposed asses obtains the best technique for preserving particular property with ranking all techniques based on D-statistic (for the degree and clustering distribution) or actual values of property (for the average degree and density): the technique with the smallest D-statistic or value of the property gets rank 0, the second one gets rank 1, and so on. In the next step, the ranks are summed and divide by the greatest possible rank. The technique with the lowest result is assumed as the best for preserving particular property. The results of comparison are presented in Table 3.

The comparison of sampling techniques based on the degree distribution is shown in Fig. 3. We observe a clear difference between the techniques with subgraph induction step (i.e., RNS, RND, RLI, BFS, RWI, FFI) and those without induction (i.e., RLS, RWS, FFS). The first group of techniques approximates the degree distribution of the original networks more accurately. In addition, the techniques with induction improve the performance of the corresponding techniques without it. The induction increases the degrees of the nodes in the sample [22], which contribute to a better match of the degree distribution between the original and sampled networks. In general, the best performing techniques are BFS and RWI (see also Table 3). Among the techniques without induction, RWS shows the best performance, which could be explained by its bias towards selecting high degree nodes and exploring densely connected parts of the network [1]. In contrast, the samples constructed by FFS and RLS are sparse, a large fraction of the nodes in the samples has low degree, while the number of nodes with higher degree is underestimated [22]. Accordingly, both techniques are the least accurate.

The results also reveal that the accuracy of preserving the degree distribution fails for the sampled networks with less than 1% of the nodes from the original networks. Irrespective of used sampling technique and consistently across all networks, under this particular sample size, the sampled networks evolve into small, unconnected networks, which results in a lower similarity between the original and sampled networks. The latter is clear also for other analyzed properties.

The comparison of sampling techniques based on the clustering distribution is shown in Fig. 4. In general, all sampling techniques show weaker performance in preserving the clustering distribution than in the case of the degree distribution. However, the most accurate are again techniques with induction, which improve the performance of techniques without induction. Still, for the *slashdot* network FFS preserves the clustering distribution best, while for *youtube* and *email* FFI and RLI perform the worst. The transitivity of these networks is lower than for other networks (see Table 2). For this reason, the techniques that create samples with larger transitivity perform worse. Among all techniques, BFS and RWI perform the best (see also Table 3), while RNS shows the worst performance. The latter could be explained by its tendency to construct samples with a large number of nodes with low clustering [22].

Fig. 5 shows the comparison of sampling techniques based on the average degree. The results prove that the techniques with induction overestimate the average degree of the original networks, which is the result of including additional links in the sample. In general, BFS and RWI perform the best (see also Table 3). On the other hand, the techniques without induction and RNS tend to underestimate the average degree. They create sampled networks with average degree lower than 5, which follows also from their definition. In detail, RLS creates unconnected and sparse samples with low average degree [23]. The samples constructed with RNS consist of a large fraction of low-degree

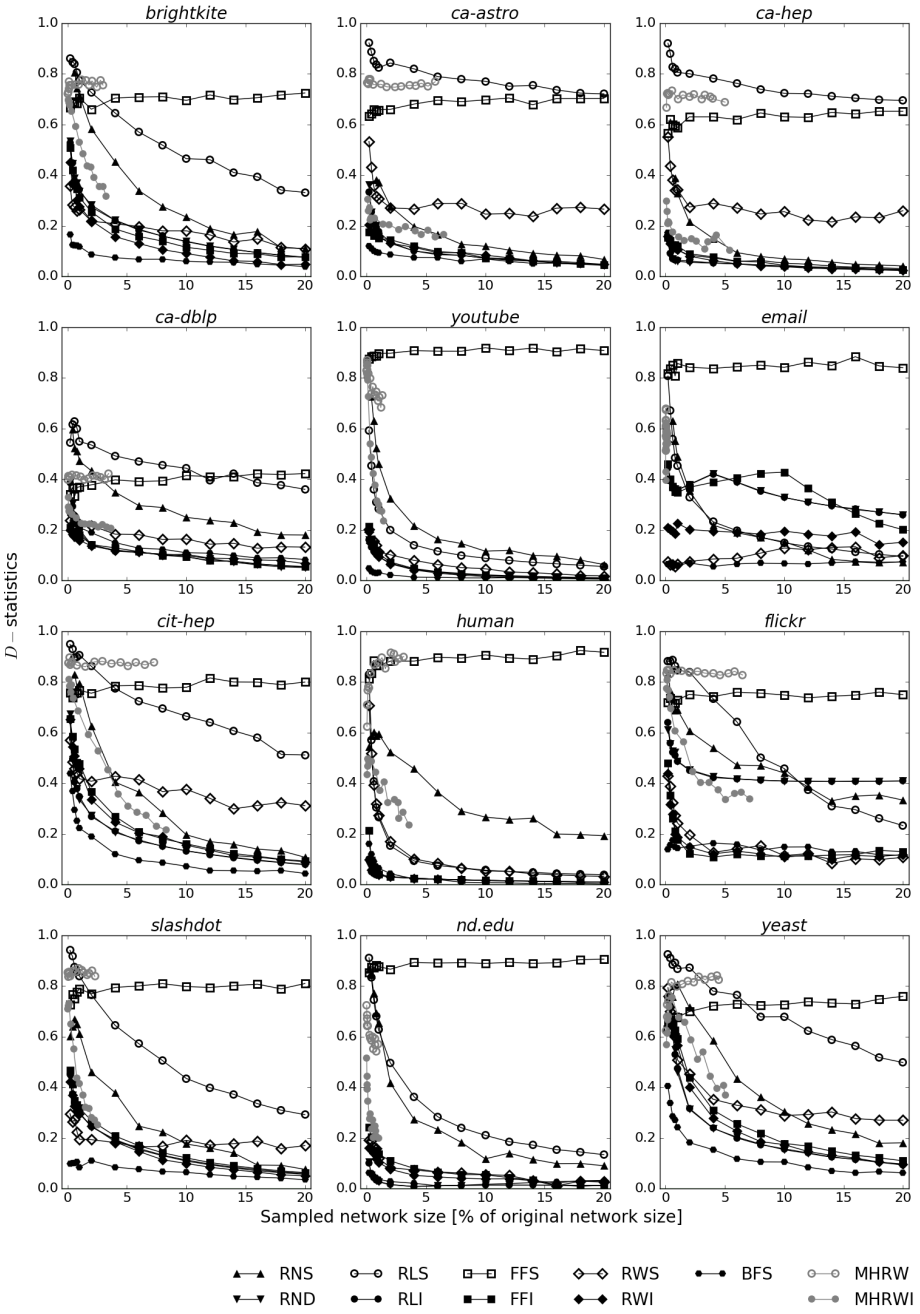


Figure 3: Comparison of the degree distribution of the original networks and their sampled variants obtained by different sampling techniques. Notice that techniques with induction (full markers) approximate the degree distribution more accurately than the techniques without induction (empty markers) in most cases.

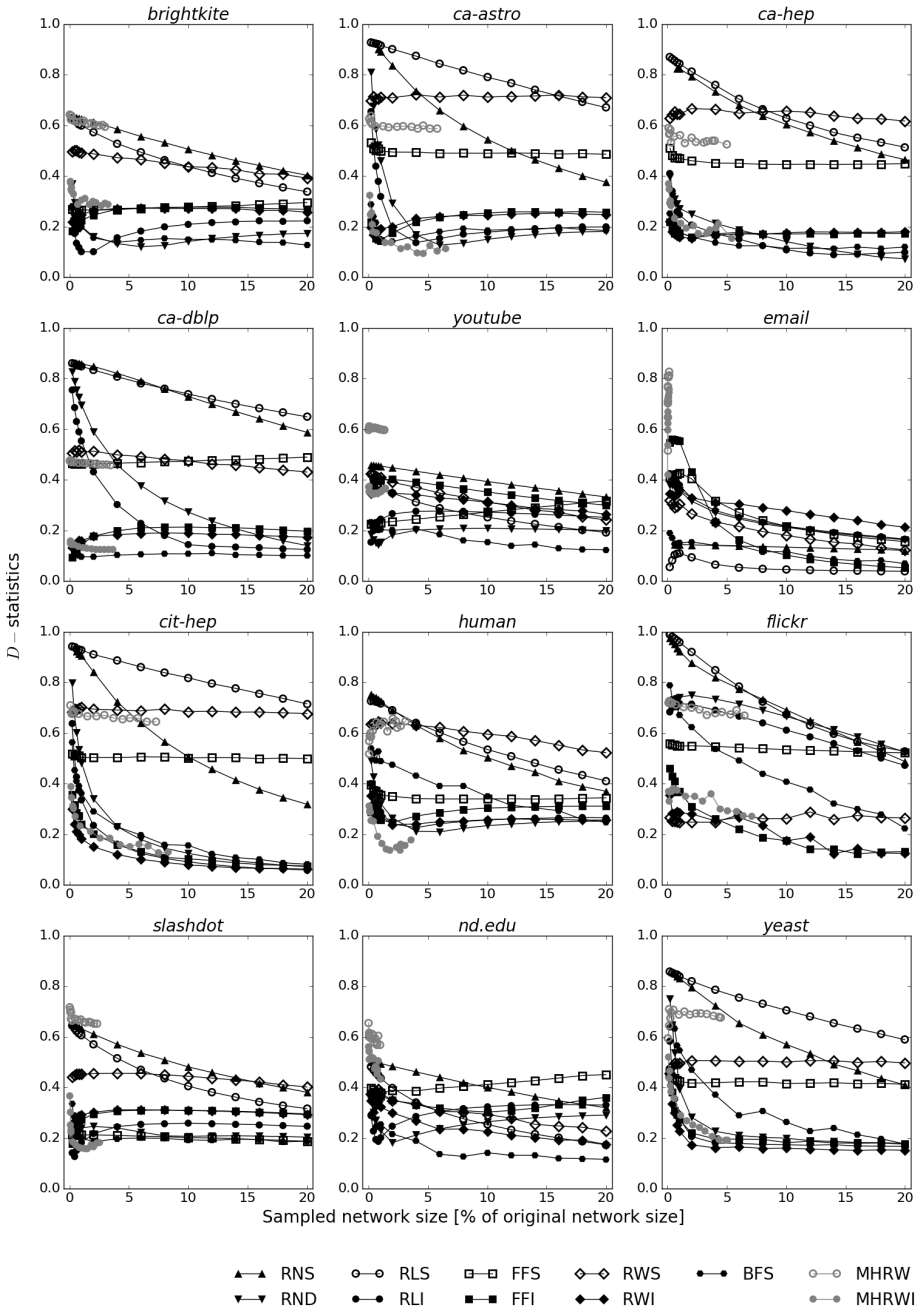


Figure 4: Comparison of the clustering coefficient distribution of the original networks and their sampled variants obtained by different sampling techniques. Notice that techniques with induction (full markers) improve the performance of the corresponding techniques without induction (empty markers).

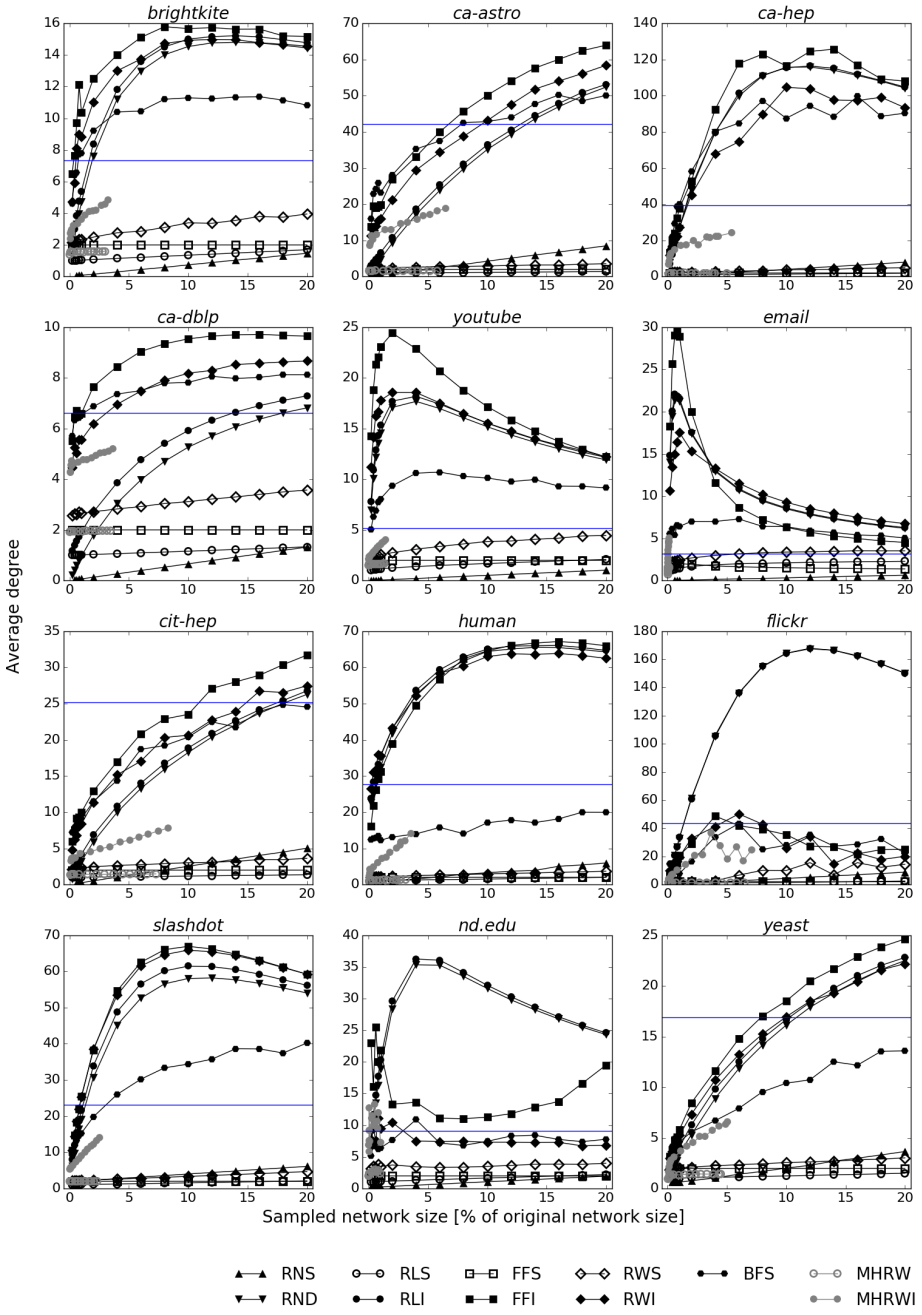


Figure 5: Comparison of the average degree of the original networks and their sampled variants obtained by different sampling techniques. The horizontal (blue) lines mark average degree of the original networks. Notice that techniques with induction (full markers) tend to overestimate the average degree, while the techniques without induction (empty markers) underestimate it.

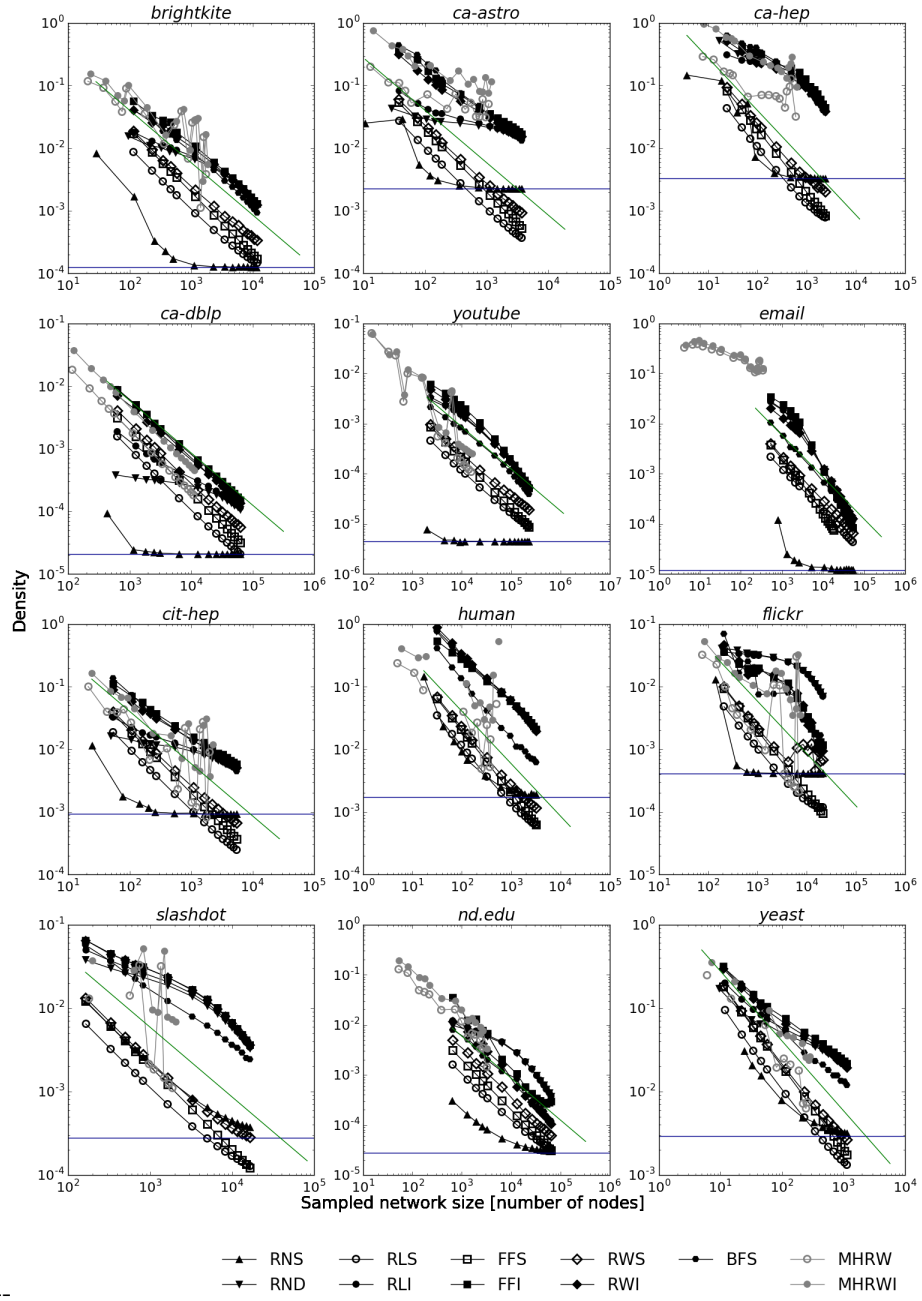


Figure 6: Comparison of the density of the original networks and their sampled variants obtained by different sampling techniques. The horizontal (blue) lines mark density of the original networks, while the diagonal (green) lines mark the power-law relationship between the size and density of real-world networks and their sampled variants [43]. Notice that the sampled networks are denser than the original networks in most cases. The techniques with induction (full markers) create denser samples that the techniques without it (empty markers).

nodes [22] and have a low average degree particularly in the smaller samples [2]. In RWS it is unlikely to occur that the random walker usea same node twice, while at FFS we set the parameter to 0.7 as suggested in [1], thus on each step 2.33 links are selected in the sample on average.

Fig. 6 shows the comparison of the sampling techniques based on the density. In the previous study [43], we proved the power-law relationship between the size and density of real-world networks and their sampled variants. The power-law relationship indicates that network density decreases with its size. In general, all techniques overestimate the density of the original networks. Yet, the techniques with induction create denser samples, while the techniques without induction construct sparser samples than expected by the power-law relationship. Therefore, the accuracy of the sampling techniques based on the density is relative and depends on whether the samples should accurately match the density of the original networks (like RNS) or the sampled networks should be denser than original and their density should follow the power-law relationship between the size and density (like BFS).

Additionally, we adopt MHRW sampling with and without induction, marked with gray markers on Fig. 3, 4, 5 and 6. Since the computationally complex implementation of the algorithm, we limited the number of steps instead of the number visited nodes in the random walk. Thus, the sampled networks with MHRW and MHRWI are smaller than the other samples. The results reveal the technique with induction improve the performance of MHRW without induction for degree and clustering distribution and the average degree. MHRWI shows good performance at preserving the clustering distribution of small samples (for the sample sizes under 5% of the original networks sizes). However, the performance of MHRW is comparable to the performance of FFS and RWS in most cases, which is expected due to the random walk basis of all three techniques. Observing the density preservation, MHRW and MHRWI show the most unstable performance; MHRW creates sparser samples than MHRWI.

In general, BFS and RWI show the best overall performance in preserving degree and clustering distribution and also average degree. The performance of both techniques is stable and on the analyzed set of networks it does not depend on the network type. On the other hand, the preservation of density is not straightforward, since it depends on further use of the sample. However, the results reveal the techniques without induction create sparser samples with a lower average degree. The techniques with induction show the opposite behavior, since they construct denser sampled networks with a higher average degree. In other words, the real average degree and the density of the original networks lay between both regimes. It seems that we could obtain the closest value for the average degree and the density of original network if we create two samples, one with the specific technique without induction and the other with the same technique with induction, and average the values for both samples.

Next, we also investigate two other ideas. First, we analyze the performance of the techniques with partial subgraph induction, where we include a different portion of mutual links between the sampled nodes in the sample (i.e., we randomly select 10%-90% links from all possible links). The results are not described in detail, since the techniques with partial induction did not improve the performance of the techniques with induction and they did not perform worse than techniques without induction. Second, for the case of exploration techniques, we study the influence of seed node selection on sampling accuracy. Particularly, we explore the changes in sampling performance if the nodes with

Table 3: The best and second-best techniques for the preservation of network properties.

Property	Study from [24]		This study	
	Best	Second-best	Best	Second-best
Degree distribution	BFS	RND	BFS (0.152)	RWI (0.281)
In-degree distribution	RND/BFS	RLS	–	–
Out-degree distribution	RND	BFS	–	–
Clustering distribution	RND	BFS	BFS (0.281)	RWI (0.282)
Betweenness	BFS	RND	–	–
Density	RNS	BFS	RNS (0.021)	RLS (0.125)
Degree mixing	BFS	RNS	–	–
Transitivity	RNS	RND	–	–
Average degree	–	–	BFS (0.229)	RWI (0.302)

larger or lower degree are more likely to be selected for the seed. The results reveal that this modification does not affect the performance of sampling and we thus only present results for random seed node selection.

Last, we review the time complexity of the adopted sampling techniques. Due to the several programming languages used for implementation and different computers we worked on, the running times of the algorithms are not comparable. However, assuming an efficient representation of the networks (e.g. adjacency list or edge list representation) and excluding the time needed for construction of the samples (which is proportional to the sample size), the time complexities of the techniques are following:

- RNS, RND: $O(v)$ (assuming edge list representation, so node can be sampled in constant time)
- RLS, RLI: $O(e)$
- BFS, FFS, FFI: $O(v + e)$
- RW, RWI: $O(v + e)$ (assuming node sampling without replacement)
- MHRW, MHRWI: $O(v^3)$ [44] (in worst case)

where v denotes the number of nodes in the sample and e denotes the number of links in the sample.

The running time of the naive algorithm for computing the clustering coefficient is $O(n^3)$ [10], where n denotes number of nodes in the original network. Computing different centrality measures, like betwenness and eigenvector centrality, has similar computational complexity [45]. For example, in the sampled network with 10% of the original network size, the computational time for clustering coefficient is of order 1000 times lower than in the original network.

3.3. Sampling selection scheme

In the following, we compare sampling techniques using a measure, that ranks techniques based on their suitability for preserving different network properties [24]. The comparison is performed for the samples of 10% of the original network sizes. This size proved to be appropriate for suitable preservation of network properties and at the same time the samples are sufficiently small for fast and efficient analysis. The results of the study [24] are presented in Table 3, abbreviations for sampling techniques are explained

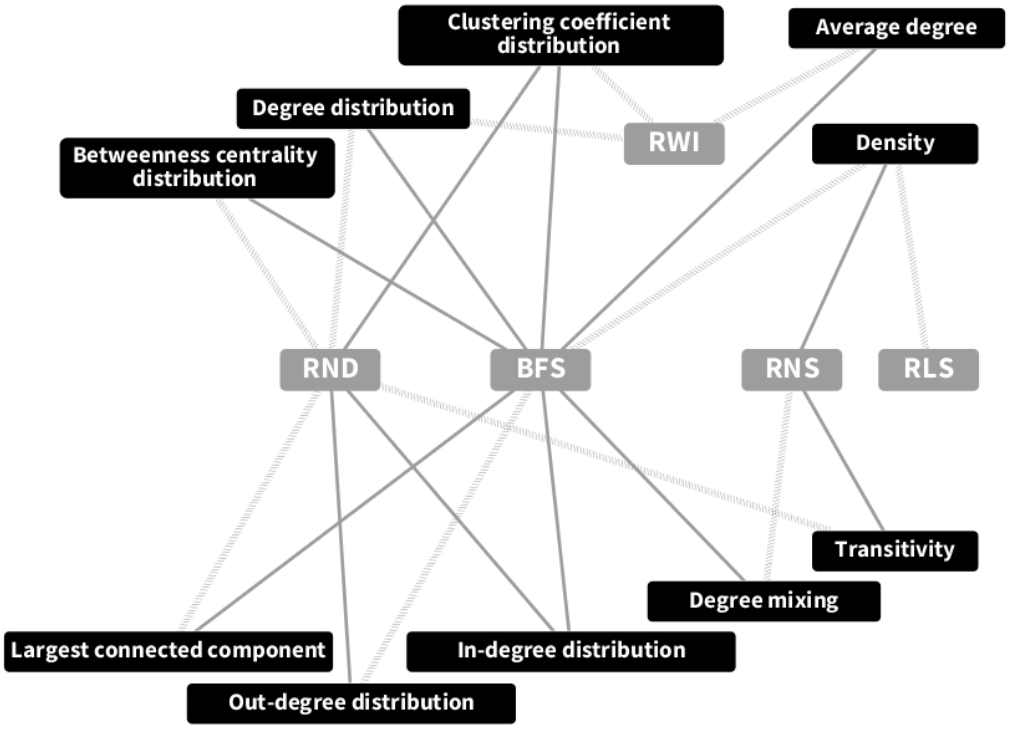


Figure 7: The sampling selection scheme for choosing the most appropriate sampling technique for sampling specific network based on which properties should be preserved under sampling. Darker rectangles denote network properties, while lighter rectangles represent sampling techniques (see Table 1 for abbreviations). Solid links between nodes corresponds to the best technique for preserving specific property, while dashed links denote second-best choice.

in Table 1. Among all, BFS and RND proved the best, since they preserve the most the majority of analyzing properties.

Using described measure we compare the techniques with induction and without it and analyze their suitability for preserving each of the properties (i.e., the degree and clustering distribution, average degree and density). The analysis is performed on the samples of 10% of the original network size as suggested in [24]. The results are presented in Table 3. BFS and RWI prove as the best performing techniques; BFS preserves the most degree and clustering distribution and average degree, while RNS is the most appropriate choice for preserving the density of the original networks. In addition, we observe the rankings of techniques for other sample sizes. The results reveal the same techniques are the best also for other sizes (detailed results are omitted). The noticeable differences occur for smaller samples under 1% of the original network size, at which all sampling techniques perform worse.

Finally, we combine listed results from Table 3 into the sampling selection scheme, presented in Fig. 7. The scheme supports choosing the most appropriate technique for sampling specific network regarding the properties that should be preserved under sampling. In the scheme, rectangles represent sampling techniques and network properties,

while the links between them corresponds to the best or second-best choice for preserving specific property. For example, average degree is best preserved by BFS. On the other hand, clustering distribution is best preserved by BFS or RND. Thus, the choice between both techniques depends also on other properties, which should be preserved under sampling. In general, BFS proves to be the most appropriate technique for network sampling, since it best preserves the majority of observed properties.

4. Conclusion

In this paper, we analyze different real networks and study the changes of their properties introduced by network sampling. We consider six basic sampling techniques, including random node and link selection and exploration techniques based on random walk and breadth-first sampling. We also apply subgraph induction to random link selection, random walk sampling and forest fire and compare the techniques based on the match of properties between the original networks and their sampled variants.

The results reveal that the sampling techniques with induction step approximate the degree and clustering coefficient more accurately than techniques without induction. Moreover, the techniques with induction step improve the performance of the corresponding techniques without induction. The techniques with induction also create denser samples with larger average degree. Particularly, they tend to overestimate the average degree and the density of the original networks. On the other hand, the techniques without induction underestimate the average degree and the density. According to these results, it appears that the performance of the techniques from random selection category compared to network exploration sampling does not differ significantly, while clear differences exist between the techniques with subgraph induction step and without it. Furthermore, the accuracy of sampling decrease consistently across various sampling techniques, when the sampled networks contain less than 1% of nodes from the original networks. Finally, based on the results of this analysis and study in [24], we introduce a sampling selection scheme. The scheme supports the selection of technique for sampling particular network regarding the properties that should be preserved under the sampling. The breadth-first exploration sampling proves as the best performing technique for preserving the majority of the observed properties.

However, the accuracy of sampling techniques does not depend only on the characteristics of the adopted technique, but also on the characteristics of the original networks. For example, the techniques with induction match the degree distribution of the networks with higher average degree more accurately than when the average degree is lower. To confirm the hypothesis, the relation should be observed in a larger set of real-world networks. Besides, a prominent direction for further study is broader analysis of the time and space efficiency of sampling techniques since fitting even a sampled network in the memory becomes challenging with significant growth of real-world networks in the past few years.

Acknowledgment

This work has been supported by the Slovenian Research Agency *ARRS* within the Research Program No. P2-0359.

References

[1] J. Leskovec, C. Faloutsos, Sampling from large graphs, in: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2006, pp. 631–636.

[2] S. H. Lee, P. J. Kim, H. Jeong, Statistical properties of sampled networks, Phys. Rev. E 73 (1) (2006) 016102.

[3] V. Krishnamurthy, M. Faloutsos, M. Chrobak, L. Lao, J.-H. Cui, A. G. Percus, Reducing large internet topologies for faster simulations, in: Proceedings of the 4th International IFIP-TC6 Networking Conference, Springer, 2005, pp. 328–341.

[4] C. Bennett, More efficient classification of web content using graph sampling, in: IEEE Symposium on Computational Intelligence and Data Mining, IEEE, 2007, pp. 485–490.

[5] D. Rafiei, Effectively visualizing large networks through sampling, in: Visualization, IEEE, 2005, pp. 375–382.

[6] D. Hennessey, D. Brooks, A. Fridman, D. Breen, A simplification algorithm for visualizing the structure of complex graphs, in: Proceedings of the 12th International Conference on Information Visualisation, IEEE, 2008, pp. 616–625.

[7] S. Tabassum, J. Gama, Sampling evolving ego-networks with forgetting factor, in: Proceedings of the 17th Conference on Mobile Data Management, Vol. 2, IEEE, 2016, pp. 55–59.

[8] L. Lü, T. Zhou, Link prediction in complex networks: A survey, Physica A 390 (6) (2011) 1150–1170.

[9] L. Pan, T. Zhou, L. Lü, C.-K. Hu, Predicting missing links and identifying spurious links via likelihood analysis, Sci. Rep. 6 (2016) 22955–22955.

[10] M. P. H. Stumpf, C. Wiuf, Sampling properties of random graphs: the degree distribution, Phys. Rev. E 72 (3) (2005) 036118.

[11] F. Zhou, S. Malher, H. Toivonen, Network simplification with minimal loss of connectivity, in: Proceedings of the 10th International Conference on Data Mining, IEEE, 2010, pp. 659–668.

[12] Z. S. Jalali, A. Rezvanian, M. R. Meybodi, Social network sampling using spanning trees, Int. J. Mod. Phys. C 27 (05) (2016) 1650052.

[13] M. B. Zafar, P. Bhattacharya, N. Ganguly, K. P. Gummadi, S. Ghosh, Sampling content from online social networks: Comparing random vs. expert sampling of the twitter stream, ACM T.Web 9 (3) (2015) 12.

[14] M. P. H. Stumpf, C. Wiuf, R. M. May, Subnets of scale-free networks are not scale-free: sampling properties of networks, P. Natl. Acad. Sci. USA 102 (12) (2005) 4221–4224.

[15] M. Génois, C. L. Vestergaard, C. Cattuto, A. Barrat, Compensating for population sampling in simulations of epidemic spread on temporal contact networks, e-print arXiv:1503.04066.

[16] A. Rezvanian, M. R. Meybodi, Sampling algorithms for weighted networks, Soc. Netw. Anal. Min. 6 (1) (2016) 60.

[17] S.-W. Son, C. Christensen, G. Bizhani, D. V. Foster, P. Grassberger, M. Paczuski, Sampling properties of directed networks, Phys. Rev. E 86 (4) (2012) 046104.

[18] C. Doerr, N. Blenn, Metric convergence in social network sampling, in: Proceedings of the 5th ACM workshop on HotPlanet, ACM, 2013, pp. 45–50.

[19] C. Hübler, H. P. Kriegel, K. Borgwardt, Z. Ghahramani, Metropolis algorithms for representative subgraph sampling, in: Proceedings of the 8th International Conference on Data Mining, IEEE, 2008, pp. 283–292.

[20] H. Toivonen, F. Zhou, A. Hartikainen, A. Hinkka, Compression of weighted graphs, in: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2011, pp. 965–973.

[21] N. Blagus, L. Šubelj, G. Weiss, M. Bajec, Sampling promotes community structure in social and information networks, Physica A 432 (2015) 206–215.

[22] N. Ahmed, J. Neville, R. R. Kompella, Network sampling via edge-based node selection with graph induction, Tech. rep., Purdue University (2011).

[23] N. K. Ahmed, J. Neville, R. Kompella, Network sampling: From static to streaming graphs, ACM T. Knowl. Discov. D. 8 (2) (2014) 7.

[24] N. Blagus, L. Šubelj, M. Bajec, Assessing the effectiveness of real-world network simplification, Physica A 413 (2014) 134–146.

[25] N. K. Ahmed, J. Neville, R. Kompella, Reconsidering the foundations of network sampling, in: Proceedings of the 2nd Workshop on Information in Networks, 2010.

[26] P. Hu, W. C. Lau, A survey and taxonomy of graph sampling, e-print arXiv:1308.5865.

[27] A. S. Maiya, T. Y. Berger-Wolf, Benefits of bias: towards better characterization of network sam-

pling, in: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2011, pp. 105–113.
[28] M. Kurant, A. Markopoulou, P. Thiran, On the bias of BFS, in: Proceedings of the 22nd International Teletraffic Congress, IEEE, 2010, pp. 1–8.
[29] L. Lovász, Random walks on graphs: A survey, Combinatorics: Paul Erdös is eighty 2 (1) (1993) 1–46.
[30] D. Stutzbach, R. Rejaie, N. Duffield, S. Sen, W. Willinger, On unbiased sampling for unstructured peer-to-peer networks, IEEE/ACM Trans. Netw. 17 (2) (2009) 377–390.
[31] M. Gjoka, M. Kurant, C. T. Butts, A. Markopoulou, Walking in facebook: A case study of unbiased sampling of osns, in: Proceedings of the 29th Conference on Information Communications, IEEE, 2010, pp. 1–9.
[32] S. Kumar, H. Sundaram, Task-driven sampling of attributed networks, e-print arXiv:1611.00910.
[33] J. Leskovec, J. Kleinberg, C. Faloutsos, Graph evolution: Densification and shrinking diameters, ACM Trans. Knowl. Discov. Data 1 (1) (2007) 1–40.
[34] J. Yang, J. Leskovec, Community-affiliation graph model for overlapping network community detection, in: Proceedings of the 12th International Conference on Data Mining, IEEE, 2012, pp. 1170–1175.
[35] J. Reimand, L. Tooming, H. Peterson, P. Adler, J. Vilo, Graphweb: mining heterogeneous biological networks for gene modules with functional significance, Nucleic acids research 36 (suppl 2) (2008) W452–W459.
[36] KDD Cup '03, `http://www.cs.cornell.edu/projects/kddcup/` (2013).
[37] E. Cho, S. A. Myers, J. Leskovec, Friendship and mobility: user movement in location-based social networks, in: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2011, pp. 1082–1090.
[38] J. Leskovec, K. J. Lang, A. Dasgupta, M. W. Mahoney, Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters, Internet Math. 6 (1) (2009) 29–123.
[39] J. Yang, J. Leskovec, Defining and evaluating network communities based on ground-truth, Knowledge and Information Systems 42 (1) (2015) 181–213.
[40] R. Albert, H. Jeong, A.-L. Barabási, Internet: Diameter of the world-wide web, Nature 401 (6749) (1999) 130–131.
[41] J. McAuley, J. Leskovec, Learning to discover social circles in ego networks, in: Advances in Neural Information Processing Systems 25, 2012, pp. 548–556.
[42] D. J. Watts, S. H. Strogatz, Collective dynamics of "small-world" networks, Nature 393 (6684) (1998) 440–442.
[43] N. Blagus, L. Šubelj, M. Bajec, Self-similar scaling of density in complex real-world networks, Physica A 391 (8) (2012) 2794–2802.
[44] A.-M. Kermarrec, E. Le Merrer, B. Sericola, G. Trédan, Second order centrality: Distributed assessment of nodes criticity in complex networks, Comput. Commun. 34 (5) (2011) 619–628.
[45] M. Al Hasan, Methods and applications of network sampling, in: Optimization Challenges in Complex, Networked and Risky Systems, INFORMS, 2016, pp. 115–139.