Sampling promotes community structure in social and information networks

Neli Blagus*, Lovro Šubelj, Gregor Weiss, Marko Bajec

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

Abstract

Any network studied in the literature is inevitably just a sampled representative of its real-world analogue. Additionally, network sampling is lately often applied to large networks to allow for their faster and more efficient analysis. Nevertheless, the changes in network structure introduced by sampling are still far from understood. In this paper, we study the presence of characteristic groups of nodes in sampled social and information networks. We consider different network sampling techniques including random node and link selection, network exploration and expansion. We first observe that the structure of social networks reveals densely linked groups like communities, while the structure of information networks is better described by modules of structurally equivalent nodes. However, despite these notable differences, the structure of sampled networks exhibits stronger characterization by community-like groups than the original networks, irrespective of their type and consistently across various sampling techniques. Hence, rich community structure commonly observed in social and information networks is to some extent merely an artifact of sampling.

Keywords: complex networks, network sampling, node group structure, communities, modules
PACS: 64.60.aq, 89.75.Fb, 89.90.+n

1. Introduction

Any network found in the literature is inevitably just a sampled representative of its real-world analogue under study. For instance, many networks change quickly over time and in most cases merely incomplete data is available on the underlying system. Additionally, network sampling techniques are lately often applied to large networks to allow for their faster and more efficient analysis. Since the findings of the analyses and simulations on such sampled networks are implied for the original ones, it is of key importance to understand the structural differences between the original networks and their sampled variants.

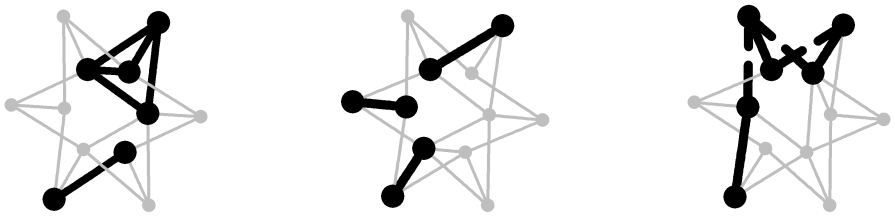
*Corresponding author. Tel.: +386 1 476 81 86.
Email addresses: neli.blagus@fri.uni-lj.si (Neli Blagus), lovro.subelj@fri.uni-lj.si (Lovro Šubelj), gregor.weiss@fri.uni-lj.si (Gregor Weiss), marko.bajec@fri.uni-lj.si (Marko Bajec)

A large number of studies on network sampling focused on the changes in network properties introduced by sampling. Lee et al. [1] showed that random node and link selection overestimate the scale-free exponent [2] of the degree and betweenness centrality [3] distributions, while they preserve the degree mixing [4]. On the other hand, random node selection preserves the degree distribution of different random graphs [5] and performs better for larger sampled networks [6]. Furthermore, Leskovec et al. [7] showed that the exploration sampling using random walks or forest-fire strategy [8] outperforms the random selection techniques in preserving the clustering coefficient [9], different spectral properties [7], and the in-degree and out-degree distributions. More recently, Ahmed et al. [10] proposed random link selection with additional induction step, which notably improves on the current state-of-the-art. Their results confirm that the proposed technique well captures the degree distributions, shortest paths [9] and also the clustering coefficient of the original networks. Lately, different studies also focus on finding and correcting biases in sampling process, for example observing the changes of user attributes under the sampling of social networks [11], analyzing the bias of traceroute sampling [12] and understanding the changes of degree distribution and hubs inclusion under various sampling techniques [13]. However, despite all those efforts, the changes in network structure introduced by sampling and the effects of network structure on the performance of sampling are still far from understood.

Real-world networks commonly reveal communities (also link-density community [14]), described as densely connected clusters of nodes that are loosely connected between [15]. Communities possibly play important roles in different real-world systems, for example in social networks communities present friendship circles or people with similar interest [16], while in citation networks communities can help us to reveal relationships between scientific disciplines [17]. Furthermore, community structure has a strong impact on dynamic processes taking place on networks [18] and thus provides an important insight into structural organization and functional behavior of real-world systems. Consequently, a number of community detection algorithms have been proposed over the last years [19, 20, 21, 22] (for a review see [23]). Most of these studies focus on classical communities characterized by higher density of edges [24]. However, some recent works demonstrate that real-world networks reveal also other characteristic groups of nodes [25, 26] like groups of structurally equivalent nodes denoted modules [25, 27] (also link-pattern community [14] and other [28]), or different mixtures of communities and modules [29].

Despite community structure appears to be an intrinsic property of many real-world networks, only a few studies considered the effects between the community structure and network sampling. Salehi et al. [30] proposed Page-Rank sampling, which improves the performance of sampling of networks with strong community structure. Furthermore, expansion sampling [31] directly constructs a sample representative of the community structure, while it can also be used to infer communities of the unsampled nodes. Other studies, for example analyzed the evolution of community structure in collaboration networks and showed that the number of communities and their size increase over time [32], while the network sampling has a potential application in testing for signs of preferential attachment in the growth of networks [33]. However, to the best of our knowledge, the question whether sampling destroys the structure of communities and other groups of nodes or are sampled nodes organized in a similar way than nodes in original network remains unanswered.

In this paper, we study the presence of characteristic groups of nodes in different



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

Figure 1: Random selection techniques applied to a small toy network, where the samples are shown with highlighted nodes and links. (a) In random node selection by degree, the nodes are selected with probability proportional to their degrees, while their mutual links are included in the sample. (b) In random link selection, the sample consists of links selected uniformly at random. (c) In random link selection with induction, the sample consists of randomly selected links (solid lines) and the links between their endpoints (dashed lines).

social and information networks and analyze the changes in network group structure introduced by sampling. We consider six sampling techniques including random node and link selection, network exploration and expansion sampling. The results first reveal that nodes in social networks form densely linked community-like groups, while the structure of information networks is better described by modules. However, regardless of the type of the network and consistently across different sampling techniques, the structure of sampled networks exhibits much stronger characterization by community-like groups than the original networks. We therefore conclude that the rich community structure is not necessary a result of common belief like for example homophily in social networks.

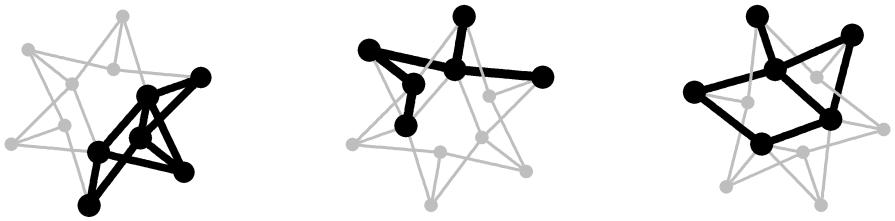
The rest of the paper is structured as follows. In Section 2, we introduce different sampling techniques considered in the study, while the adopted node group extraction framework is presented in Section 3. The results of the empirical analysis are reported and formally discussed in Section 4, while Section 5 summarizes the paper and gives some prominent directions for future research.

2. Network sampling

Network 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 the degree of a node or its PageRank score [34]. In the second category, the sample is constructed by retrieving a neighborhood of a randomly selected seed node using random walks, breadth-first search or another strategy. For the purpose of this study, we consider three techniques from each of the categories.

2.1. Random selection

From the random selection category, we first adopt random node selection by degree [7] (RND). Here, the nodes are selected randomly with probability proportional to their degrees, while all their mutual links are included in the sample (Fig. 1(a)). Note



(a) BFS (b) FFS (c) EXS

Figure 2: Network exploration techniques applied to a small toy network, where the samples are shown with highlighted nodes and links. (a) In breadth-first sampling, a seed node is first selected uniformly at random, while its broad neighborhood retrieved from breadth-first search is included in the sample. (b) In forest-fire sampling, the broad neighborhood of a randomly selected seed node is retrieved from partial breadth-first search, where only a fraction of neighbors is included in the sample. (c) In expansion sampling, the seed node is selected uniformly at random, while the remaining nodes are selected from the neighborhood of sampled nodes with probability proportional to their contribution to the expansion factor (see text).

that RND improves the performance of the basic random node selection [7, 35], where the nodes are selected to the sample uniformly at random. RND fits better spectral network properties [7] and produces the sample with larger weakly connected component [35]. Moreover, it shows good performance in preserving the clustering coefficient and betweenness centrality distribution of the original networks [35]. Nevertheless, it can still construct a disconnected sample network, despite a fully connected original network.

Next, we adopt random link selection [7] (RLS), where the sample consists of links selected uniformly at random (Fig. 1(b)). RLS overestimates degree and betweenness centrality exponent, underestimate the clustering coefficient and accurately matches the assortativity of the original network [1]. The samples created with RLS are sparse and the connectivity of the original network is not preserved, still RLS is likely to capture the path length of the original network [36].

Last, we adopt random link selection with induction [10] (RLI), which improves the performance of RLS. In RLI, the sample consists of randomly selected links as before, while also all additional links between their endpoints (Fig. 1(c)). RLI outperforms several other methods in capturing the degree, path length and clustering coefficient distribution. It selects nodes with higher degree than RLS, thus the connectivity of the sample is increased [10].

2.2. Network exploration

From the network exploration category, we first adopt breadth-first sampling [1] (BFS). Here, a seed node is selected uniformly at random, while its broad neighborhood retrieved from the basic breadth-first search is included in the sample (Fig. 2(a)). The sample network is thus a connected subgraph of the original network. BFS is biased towards selecting high-degree nodes in the sample [37]. It captures well the degree distribution of the networks, while it performs worst in inclusion of hubs in the sample quickly in the sampling process [13]. BFS imitates the snowball sampling approach for

collecting social data used especially when the data is difficult to reach [38]. Selected seed participant is asked to report his friends, which are than invited to report their friends. The procedure is repeated until the desired number of people is sampled.

Next, we adopt a modification of BFS denoted forest-fire sampling [7] (FFS). In FFS, the broad neighborhood of a randomly selected seed node is retrieved from partial breadth-first search, where only some neighbors are included in the sample on each step (Fig. 2(b)). The number of neighbors is sampled from a geometric distribution with mean $p/(1-p)$, where p is set to 0.7 [7]. FFS matches well spectral properties [7], while it underestimates the degree distribution and fails to match the path length and clustering coefficient of the original networks [36]. However, FFS corresponds to a model by which one author collects the papers to cite and include them in the bibliography [8]. The author starts with one paper, explores its bibliography and selects the papers to cite. The procedure is recursively repeated in selected papers until desired collection of citations is reached.

Last, we adopt expansion sampling [31] (EXS), where the seed node is again selected uniformly at random, while the neighbors of the sampled nodes are included in the sample with probability proportional to

$$1 - \beta^{|N(\{v\}) - (N(S) \cup S)|}, \tag{1}$$

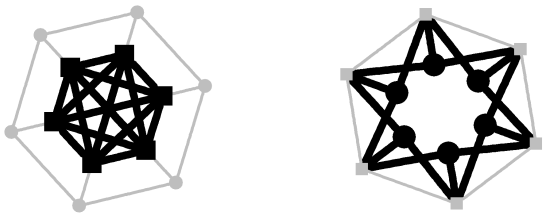
where v is the concerned node, S the current sample and $N(S)$ the neighborhood of nodes in S (Fig. 2(c)). Expression $|N(\{v\}) - (N(S) \cup S)|$ denotes the expansion factor of node v for sample S and means the number of new neighbors contributed by v. The parameter β is set to 0.9 [31]. Note that EXS ensures that the sample consists of nodes from most communities in the original network and that the nodes that are grouped together in the original network, are also grouped together in the sample [15]. EXS imitates the modification of snowball sampling approach mentioned above, where for example we want to gather the data about individuals from different countries. Thus, on each step we include in the sample the individuals, which knows larger number of others from various countries.

3. Group extraction

The node group structure of different networks is explored by a group extraction framework [29, 39, 40] with a brief overview below.

Let the network be represented by an undirected graph $G(V, L)$, where V is the set of nodes and L the set of links. Next, let S be a group of nodes and T a subset of nodes representing its corresponding linking pattern (i.e., the pattern of connections of nodes from S to other nodes [25]), $S, T \subseteq V$. Denote $s = |S|$ and $t = |T|$. The linking pattern T is selected to maximize the number of links between S and T, and minimize the number of links between S and T^C, while disregarding the links with both endpoints in S^C. For details on the group objective function see [29, 41].

The above formalism comprises different types of groups commonly analyzed in the literature (Fig. 3). It consider communities [15] (i.e., link-density community [14]), defined as a (connected) group of nodes with more links toward the nodes in the group than to the rest of the network [24]. Communities are characterized by $S = T$. Furthermore, the formalism consider possibly disconnected groups of structurally equivalent



(a) Community (b) Module

Figure 3: Toy examples of groups of nodes in networks, where groups S and their corresponding linking patterns T are shown with highlighted and squared nodes, respectively (see text). (a) Communities are densely connected groups of nodes with $S = T$. (b) Modules are possibly disconnected groups of structurally equivalent nodes with $S \cap T = \emptyset$. Groups spanning between communities and modules are denoted mixtures.

nodes denoted modules [25, 27] (i.e., link-pattern community [14]), defined as a (possibly) disconnected group of nodes with more links towards common neighbors than to the rest of the network [24]. Modules have $S \cap T = \emptyset$. Communities and modules represent two extreme cases with all other groups being the mixtures of the two [29], $S \cap T \subset S$ and/or $S \cap T \subset T$. The reader may also find it interesting that the core-periphery structure is a mixture with $S \subset T$, while the hub & spokes structure is a module with $t = 1$.

The type of group S can in fact be determined by the Jaccard index [42] of S and its corresponding linking pattern T. The group parameter τ [29], $\tau \in [0, 1]$, is defined as

$$\tau(S, T) = \frac{|S \cap T|}{|S \cup T|}. \tag{2}$$

Communities have $\tau = 1$, while modules are indicated by $\tau = 0$. Mixtures correspond to groups with $0 < \tau < 1$. For the rest of the paper, we refer to groups with $\tau \approx 1$ as community-like and groups with $\tau \approx 0$ as module-like.

Groups in networks are revealed by a sequential extraction procedure proposed in [39, 29, 40]. One first finds the group S and its linking pattern T with random-restart hill climbing [43] that maximizes the objective function. Next, the revealed group S is extracted from the network by removing the links between groups S and T, and any node that becomes isolated. The procedure is then repeated on the remaining network until the objective function is larger than the 99th percentile of the values obtained under the same framework in a corresponding Erdős-Rényi random graph [44]. All groups reported in the paper are thus statistically significant at 1% level. Note that the above procedure allows for overlapping [45], hierarchical [46], nested and other classes of groups.

4. Analysis and discussion

Section 4.1 introduces real-world networks considered in the study. Section 4.2 reports the node group structure of the original networks extracted with the framework described in Section 3. The groups extracted from the sampled networks are analyzed in Section 4.3.

Table 1: Social and information networks considered in the study.

Network	Description	# Nodes	# Links
Collab	High Energy Physics collaborations [8]	9877	25998
PGP	Pretty Good Privacy web-of-trust [47]	10680	24340
P2P	Gnutella peer-to-peer file sharing [8]	8717	31525
Citation	High Energy Physics citations [8]	27770	352807

For a complete analysis, we also observe the node group structure of a large network with more than a million links in Section 4.4.

4.1. Network data

The empirical analysis in the following sections was performed on four real-world social and information networks. Their main characteristics are shown in Table 1.

The *Collab* [8] is a social network of scientific collaborations among researchers, who submitted their papers to High Energy Physics – Theory category on the arXiv in the period from January 1993 to April 2003. The nodes represent the authors, while undirected links denote that two authors co-authored at least one paper together.

The *PGP* [47] is a social network, which corresponds to the interaction network of users of the Pretty Good Privacy algorithm collected in July 2001. The nodes represent users, while undirected links indicate relationships between those, who sign each other's public key.

The *P2P* [8] is an information network, which contains a sequence of snapshots of the Gnutella peer-to-peer file sharing network collected in August 2002. The nodes represent hosts in the Gnutella network, which are linked by undirected links if there exist connections between them.

The *Citation* [8] is an information network, again gathered from the High Energy Physics – Theory category from the arXiv in the period from January 1993 to April 2003 and includes the citations among all papers in the dataset. The network consists of nodes, which represent papers, while links denote that one paper cite another.

4.2. Group structure of original networks

We first analyze the properties of groups extracted from the original networks summarized in Table 2.

The number of groups differs among networks, still the mean group size s (denoted $\langle s \rangle$) is comparable across network types. Groups S in social networks consist of around 64 nodes, while $\langle s \rangle$ in information networks exceeds 150 nodes. The mean linking pattern size t (denoted $\langle t \rangle$) of social networks is comparable to $\langle s \rangle$. The latter relation $\langle t \rangle \approx \langle s \rangle$ is expected due to the pronounced community structure commonly found in social networks [48]. On the other hand, $\langle t \rangle > \langle s \rangle$ is expected for information networks, due to the abundance of module-like groups.

The characteristic group structure of networks is reflected in the group parameter τ. For social networks, its values are around 0.556, which indicates the presence of communities, modules and mixtures of these. In contrast to social networks, the information networks have τ closer to 0 and consist mostly of module-like groups.

Table 2: Groups of nodes extracted from social and information networks. We report the number of groups #, the mean group size s, the mean pattern size t, the mean group parameter τ, the median group parameter denoted m_τ and the distribution over different types of groups (see text). Notice that social networks consist of smaller groups with larger τ than information networks.

Network	Group					Community	Mixture	Module
	#	$\langle s \rangle$	$\langle t \rangle$	$\langle \tau \rangle$	m_τ	Distribution %		
Collab	129	66.9	67.2	0.568	0.554	1.6%	96.8%	1.6%
PGP	87	62.2	61.9	0.568	0.536	4.6%	94.3%	1.1%
P2P	70	154.8	177.0	0.057	0.000	0.0%	44.3%	55.7%
Citation	284	271.7	280.6	0.186	0.120	0.0%	96.8%	3.2%

To summarize, social networks represent people and interactions between them, like a few authors writing a paper together, therefore we can expect a larger number of community-like groups in these networks. On the other hand, in information networks the homophily is less typical and thus the structure of these networks is rather described by module-like groups.

4.3. Group structure of sampled networks

Sampling techniques outlined in Section 2 enable setting the size of the sampled networks in advance. We consider sample sizes of 15% of nodes from the original networks, that provides for an accurate fit of several network properties [7, 35].

Table 3 and 4 present the properties of the node group structure of sampled social and information networks, respectively. Notice that RLS and FFS show different performance than other techniques. The samples obtained with RLS and FFS contain less groups with no more than 36 nodes. Additionally, almost all groups in these samples are modules, which reflects in the mean group parameter τ (denoted $\langle \tau \rangle$) approaching 0 for all networks.

To verify the above findings, we compute externally studentized residuals of the sampled networks that measure the consistency of each sampling technique with the rest. The residuals are calculated for each technique as the difference between the observed value of considered property and its mean divided by the standard deviation. The mean value and standard deviation are computed for all sampling techniques, excluding the observed one (for details see [49]). Statistically significant inconsistencies between techniques are revealed by two-tailed Student $t-$test [50] at $P-$value of 0.1, rejecting the null hypothesis that the values of the considered property are consistent across the sampling techniques.

Statistical comparison of sampling techniques for the number of groups and the mean group parameter τ is shown on Fig. 4. We confirm that the samples obtained with RLS and FFS reveal significantly less groups with significantly smaller $\langle \tau \rangle$ than other sampling techniques. Moreover, if we compare the number of links in the sampled networks, RLS and FFS create samples that contain on average 3% of links from the original networks. In contrast, the samples obtained with RND, RLI, BFS and EXS consist of around 16% of links from the original networks. As mentioned before, the sizes of all samples are 15% of the original networks, thus the sampled networks obtained with RLS and FFS are much sparser than others. In addition, the performance of RLS and FFS can also be explained by their definition. Since in RLS we include only randomly selected links in

Table 3: Groups of nodes extracted from sampled social networks over 100 realizations of different sampling techniques (see text). We report the number of groups # and standard deviation, the mean group size s, the mean pattern size t, the mean group parameter τ and standard deviation, the median group parameter denoted m_τ and the distribution over different types of groups. Notice that sampled networks expectedly consist of smaller groups, but with larger τ than original social networks (see $\langle\tau\rangle$ and m_τ).

Network	Sampling	Group					Community	Mixture	Module
		#	$\langle s\rangle$	$\langle t\rangle$	$\langle\tau\rangle$	m_τ	Distribution %		
Collab	/	129.0	66.9	67.2	0.568	0.554	1.6%	96.8%	1.6%
	RND	65.4 ± 3.7	13.5	13.7	0.851 ± 0.030	0.989	54.7%	41.9%	3.4%
	RLS	1.2 ± 0.5	1.5	4.8	0.047 ± 0.149	0.048	0.0%	8.3%	91.7%
	RLI	74.7 ± 4.6	13.7	13.9	0.846 ± 0.030	0.979	52.7%	43.4%	3.9%
	BFS	104.0 ± 6.5	18.2	18.5	0.787 ± 0.032	0.861	30.3%	66.5%	3.2%
	FFS	4.0 ± 1.6	16.8	29.8	0.000 ± 0.000	0.000	0.0%	0.0%	100.0%
	EXS	87.0 ± 5.8	18.4	18.9	0.741 ± 0.026	0.791	21.4%	76.3%	2.3%
PGP	/	87.0	62.2	61.9	0.568	0.536	4.6%	94.3%	1.1%
	RND	68.2 ± 4.5	15.8	16.0	0.891 ± 0.024	1.000	67.8%	28.7%	3.5%
	RLS	2.8 ± 1.0	5.7	7.6	0.304 ± 0.233	0.263	21.4%	28.6%	50.0%
	RLI	74.3 ± 4.3	15.8	16.1	0.883 ± 0.024	1.000	65.1%	31.1%	3.8%
	BFS	95.4 ± 9.2	17.5	17.7	0.784 ± 0.025	0.909	39.2%	55.6%	5.2%
	FFS	3.6 ± 1.3	13.5	32.6	0.000 ± 0.000	0.000	0.0%	0.0%	100.0%
	EXS	80.9 ± 6.5	15.6	15.8	0.779 ± 0.028	0.873	34.5%	61.2%	4.3%

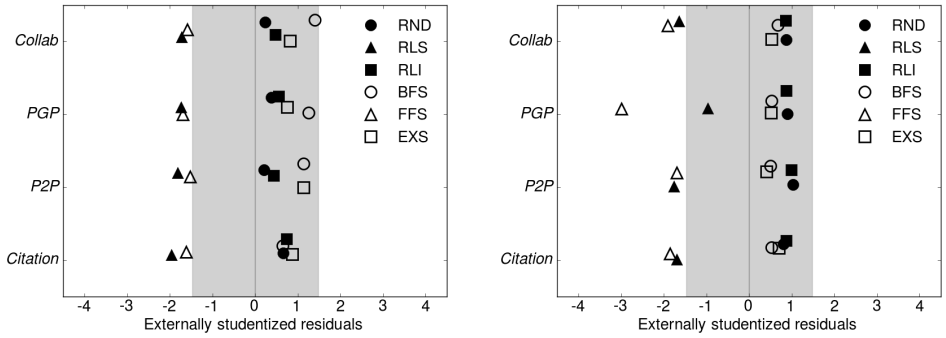
the sample, it commonly contains a large number of sparsely linked components, whose structure is best described as module-like. On the other hand, the samples obtained with FFS consist of one connected component with a low average degree of 2.33. Thus, the sparsely connected nodes also form groups, which are more similar to modules. Due to the above reasons, we exclude RLS and FFS from further analysis. We focus on RND, RLI, BFS, and EXS, whose performance is clearly more comparable.

The selected sampling techniques perform similarly across all networks as shown in Table 3 for social and Table 4 for information networks. The samples consist of various number of groups, still in most cases less than the original networks. The mean sizes s and t are around 40, in contrast to groups with 143 nodes on average in the original networks. Still, $\langle s\rangle \approx \langle t\rangle$ irrespective of network type and the sampling technique, which implies stronger characterization by community-like groups, as already argued in the case of social networks in Section 4.2.

Indeed, the majority of groups found in sampled social networks are community-like, which reflects in the parameter $\tau > 0.7$. In sampled information networks the number of mixtures decreases and communities appear, thus τ is larger than in the original networks. Fig. 5 shows a clear difference in the distribution of τ between the original and sampled networks. Furthermore, to confirm that differences exist between the structure of the original and sampled networks, we compute externally studentized residuals, where we include the value of considered property of the original network in computing the mean over different sampling techniques. We compare the number of groups and the parameter $\langle\tau\rangle$ for the original networks and their samples (Fig. 6). The results prove that the original networks contain a significantly larger number of groups with significantly smaller $\langle\tau\rangle$ than the sampled networks. Yet, larger parameter τ and consequently more community-like groups in sampled social networks and less module-like groups in sampled information networks indicate clear changes in the network structure introduced by sampling. We

Table 4: Groups of nodes extracted from sampled information networks over 100 realizations of different sampling techniques (see text). We report the number of groups # and standard deviation, the mean group size s, the mean pattern size t, the mean group parameter τ and standard deviation, the median group parameter denoted m_τ and the distribution over different types of groups. Notice that sampled networks expectedly consist of smaller groups, but with larger τ than original information networks (see $\langle\tau\rangle$ and m_τ).

Network	Sampling	Group #	Group $\langle s\rangle$	Group $\langle t\rangle$	$\langle\tau\rangle$	m_τ	Community Distribution %	Mixture Distribution %	Module Distribution %
P2P	/	70.0	154.8	177.0	0.057	0.000	0.0%	44.3%	55.7%
	RND	23.3 ± 3.9	24.2	24.4	0.163 ± 0.049	0.034	4.2%	45.8%	50.0%
	RLS	1.6 ± 0.9	1.2	3.6	0.000 ± 0.008	0.000	0.0%	0.0%	100.0%
	RLI	26.2 ± 4.4	27.5	28.1	0.161 ± 0.039	0.035	3.8%	48.8%	47.4%
	BFS	34.1 ± 5.5	31.3	27.9	0.131 ± 0.042	0.034	2.3%	50.7%	47.0%
	FFS	3.6 ± 1.4	17.8	28.3	0.000 ± 0.000	0.000	0.0%	0.0%	100.0%
	EXS	34.0 ± 5.9	36.9	37.3	0.125 ± 0.030	0.035	2.4%	53.8%	43.8%
Citation	/	284.0	271.7	280.6	0.186	0.120	0.0%	96.8%	3.2%
	RND	121.4 ± 4.9	74.9	78.1	0.405 ± 0.016	0.329	0.2%	80.9%	18.9%
	RLS	1.5 ± 1.2	1.4	15.3	0.014 ± 0.073	0.014	0.0%	0.0%	100.0%
	RLI	124.8 ± 5.5	76.3	79.9	0.415 ± 0.014	0.344	0.2%	82.6%	17.2%
	BFS	120.4 ± 7.1	99.2	100.9	0.359 ± 0.047	0.244	0.1%	77.5%	22.4%
	FFS	10.6 ± 4.2	35.5	30.0	0.000 ± 0.000	0.000	0.0%	0.0%	100.0%
	EXS	131.2 ± 6.0	91.4	95.4	0.388 ± 0.019	0.284	0.2%	82.0%	17.8%



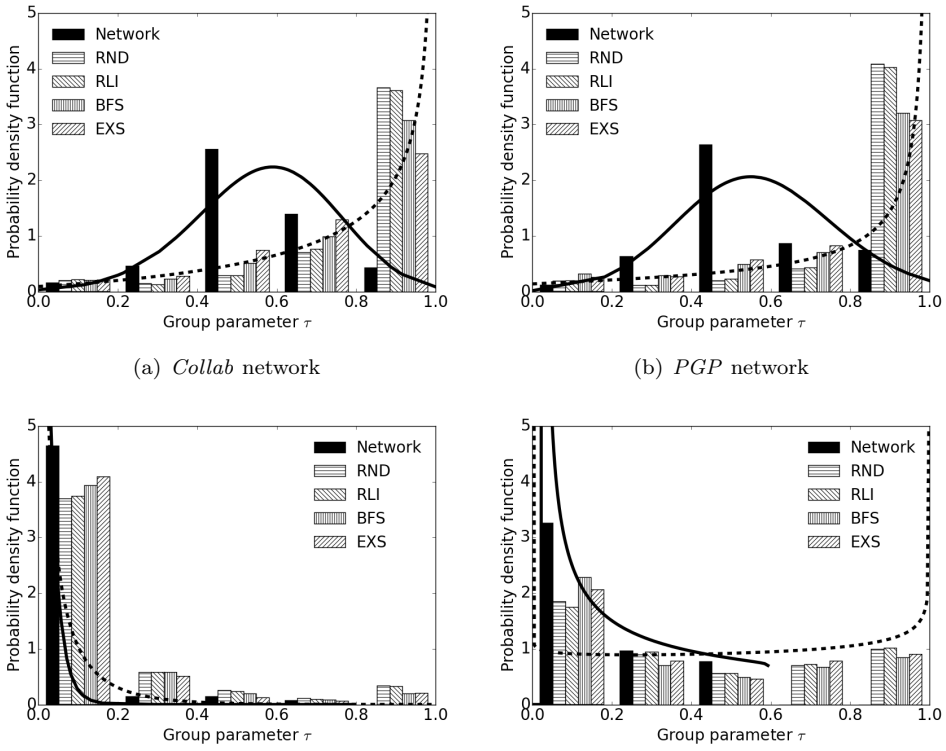
(a) Number of groups #

(b) Group parameter $\langle\tau\rangle$

Figure 4: Statistical comparison of (a) number of groups and (b) mean group parameter τ for the sampled networks obtained with different sampling techniques (see text). We show externally studentized residuals that measure the consistency of each sampling technique with the rest and expose statistically significant inconsistencies between the techniques with two-tailed Student t-test at P-value of 0.1 (shaded regions correspond to 90% confidence intervals). Notice that sampled networks obtained with RLS and FFS reveal less groups (see (a)) with significantly smaller parameter τ (see (b)) than other sampling techniques.

conclude that these changes occur regardless of the network type or the adopted sampling technique.

Notice that the largest τ and thus the strongest characterization by community-like groups is revealed in the sampled networks obtained with both random selection techniques, RND and RLI. In RND nodes with higher degrees are more likely to be selected to the sample by the definition, while RLI is biased in a similar way [10]. Thus,

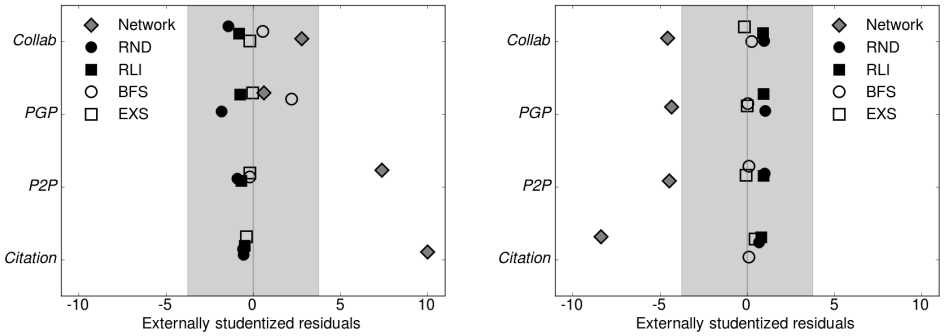


(a) *Collab* network (b) *PGP* network

(c) *P2P* network (d) *Citation* network

Figure 5: Distributions of group parameter τ for the original networks and their sampled representatives obtained with selected sampling techniques (see text). Histograms are derived by standard equidistant binning, while the estimates of a beta distribution for the original (solid lines) and sampled networks (dashed lines) are merely a guide for the eye. Notice that sampled networks are characterized by denser groups with notably larger τ than the original ones. Groups are more community-like in the case of social networks (see (a) and (b)), while less module-like in the case of information networks (see (c) and (d)).

densely connected groups of nodes have a higher chance of being included in the sampled network, while sparse parts of the networks remain unsampled. On the other hand, BFS and EXS sample the broad neighborhood of a randomly selected seed node and thus the sampled network represents a connected component. In the case of BFS, all nodes and links of some particular part of the original network are sampled. The latter is believed to be representative of the entire network [37], yet BFS is biased towards sampling nodes with higher degree [51] and overestimates the clustering coefficient, especially in information networks [1]. On the other hand, EXS ensures the smallest partition distance among several other sampling techniques, which means that nodes grouped together in communities of sampled network are also in the same community in the original network [31]. Therefore, the stronger characterization by community-like groups in sampled networks can also be explained by the definition and behavior of the sampling techniques.



(a) Number of groups #

(b) Group parameter $\langle\tau\rangle$

Figure 6: Statistical comparison of (a) number of groups and (b) mean group parameter τ for the original networks and their sampled representatives obtained with selected sampling techniques (see text). We show externally studentized residuals that measure the consistency of each network with the rest and expose statistically significant inconsistencies between the networks with two-tailed Student t-test at P-value of 0.1 (shaded regions correspond to 90% confidence intervals). Notice that original networks reveal more groups (see (a)) with significantly smaller parameter τ (see (b)) than the sampled networks.

4.4. Group structure of a large network

Due to the relatively high time complexity of the node group extraction framework, we consider only networks with a few thousand nodes. However, our previous study [35] proved that the size of the original network does not affect the accuracy of the sampling. Still, for a complete analysis, we also inspect the changes in node group structure introduced by sampling of a large *NotreDame* network with more than a million links. Due to the simplicity and execution time, we present the analysis for two sampling techniques, RND from random selection and BFS from network exploration category. We also limit the number of groups extracted from the networks to 100 (i.e., we consider top 100 significant groups with respect to the objective function).

The *NotreDame* data are collected from the web pages of the University of Notre Dame – *nd.edu* domain in 1999. The network contains 325,729 nodes representing individual web pages, while 1,497,134 links denote hyperlinks among them.

Table 5 shows the properties of groups, found in the original and sampled networks. The samples consist of smaller groups, still the mean size s remains larger than the mean size t. The majority of groups extracted from the original network are module-like, which reflects in the parameter τ slightly larger than 0. On the other hand, the changes introduced by sampling are clear, since the samples contain less modules, which is revealed by a larger parameter τ. These findings are consistent with the results on smaller networks from previous sections. The *NotreDame* as an information network expectedly consists of densely linked groups similar to modules, while the structure of sampled networks exhibits stronger characterization by community-like groups. That is again irrespective of the adopted sampling technique.

Table 5: Groups of nodes extracted from the original *NotreDame* network and its sampled representatives over 100 realizations of selected sampling techniques (see text). We report the number of groups #, the mean group size s, the mean pattern size t, the mean group parameter τ and standard deviation, the median group parameter denoted m_τ and the distribution over different types of groups. Notice that sampled networks expectedly consist of smaller groups, but with larger τ than original network (see $\langle\tau\rangle$ and m_τ).

Sampling	Group					Community	Mixture	Module
	#	$\langle s\rangle$	$\langle t\rangle$	$\langle\tau\rangle$	m_τ	Distribution %		
/	100	876.8	403.6	0.030	0.028	0.0%	99.0%	1.0%
RND	100	302.5	271.7	0.369 ± 0.010	0.364	0.0%	100.0%	0.0%
BFS	100	411.6	251.7	0.135 ± 0.030	0.119	0.0%	99.5%	0.5%

5. Conclusion

In this paper, we study the presence of characteristic groups of nodes like communities and modules in different social and information networks. We observe the groups of the original networks and analyze the changes in the group structure introduced by the network sampling.

The results first reveal noticeable differences in the group structure of original social and information networks. Nodes in social networks form smaller community-like groups, while information networks are better characterized by larger modules. After applying network sampling techniques, sampled networks expectedly contain fewer and smaller groups. However, the sampled networks exhibit stronger characterization by community-like groups than the original networks. We have shown that the changes in the node group structure introduced by sampling occur regardless of the network type and consistently across different sampling techniques. Since networks commonly considered in the literature are inevitably just a sampled representative of its real-world analogue, some results, such as rich community structure found in these networks, may be influenced by or are merely an artifact of sampling.

Our future work will mainly focus on larger real-world networks, including other types of networks like biological and technological. Moreover, we will further analyze the changes in the node group structure introduced by sampling and explore techniques that could overcome observed deficiencies.

Acknowledgment

This work has been supported in part by the Slovenian Research Agency *ARRS* within the Research Program No. P2-0359, by the Slovenian Ministry of Education, Science and Sport Grant No. 430-168/2013/91, and by the European Union, European Social Fund.

References

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

[2] A. L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (5439) (1999) 509–512.

[3] L. Freeman, A set of measures of centrality based on betweenness, Sociometry 40 (1) (1977) 35–41.

[4] M. E. J. Newman, Assortative mixing in networks, Phys. Rev. Lett. 89 (20) (2002) 208701.

[5] 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.
[6] 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.
[7] 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.
[8] J. Leskovec, J. Kleinberg, C. Faloutsos, Graphs over time: Densification laws, shrinking diameters and possible explanations, in: Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2005, pp. 177–187.
[9] D. J. Watts, S. H. Strogatz, Collective dynamics of small-worldnetworks, Nature 393 (6684) (1998) 440–442.
[10] N. Ahmed, J. Neville, R. R. Kompella, Network sampling via edge-based node selection with graph induction, Tech. rep., Purdue University (2011).
[11] H. Park, S. Moon, Sampling bias in user attribute estimation of osns, in: Proceedings of the 22nd international conference on World Wide Web companion, International World Wide Web Conferences Steering Committee, 2013, pp. 183–184.
[12] A. Lakhina, J. W. Byers, M. Crovella, P. Xie, Sampling biases in ip topology measurements, in: Proceedings of the 22nd Annual Joint Conference of the IEEE Computer and Communications, Vol. 1, IEEE, 2003, pp. 332–341.
[13] A. S. Maiya, T. Y. Berger-Wolf, Benefits of bias: Towards better characterization of network sampling, in: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2011, pp. 105–113.
[14] B. Long, X. Wu, Z. Zhang, P. S. Yu, Community learning by graph approximation, in: Proceedings of 7th IEEE INternational COnference on Data Mining, IEEE, 2007, pp. 232–241.
[15] M. Girvan, M. E. J. Newman, Community structure in social and biological networks, P. Natl. Acad. Sci. USA 99 (12) (2002) 7821–7826.
[16] J. Scott, P. J. Carrington, The SAGE handbook of social network analysis, SAGE publications, 2011.
[17] M. Rosvall, C. T. Bergstrom, Maps of random walks on complex networks reveal community structure, P. Natl. Acad. Sci. USA 105 (4) (2008) 1118–1123.
[18] A. Arenas, A. Díaz-Guilera, C. J. Pérez-Vicente, Synchronization reveals topological scales in complex networks, Phys. Rev. Lett. 96 (11) (2006) 114102.
[19] F. Wu, B. A. Huberman, Finding communities in linear time: a physics approach, Eur. Phys. J. B 38 (2) (2004) 331–338.
[20] M. Rosvall, C. T. Bergstrom, An information-theoretic framework for resolving community structure in complex networks, P. Natl. Acad. Sci. USA 104 (18) (2007) 7327–7331.
[21] U. N. Raghavan, R. Albert, S. Kumara, Near linear time algorithm to detect community structures in large-scale networks, Phys. Rev. E 76 (3) (2007) 036106.
[22] L. Šubelj, M. Bajec, Unfolding communities in large complex networks: Combining defensive and offensive label propagation for core extraction, Phys. Rev. E 83 (3) (2011) 036103.
[23] S. Fortunato, Community detection in graphs, Phys. Rep. 486 (3) (2010) 75–174.
[24] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, D. Parisi, Defining and identifying communities in networks, P. Natl. Acad. Sci. USA 101 (9) (2004) 2658–2663.
[25] M. E. J. Newman, E. A. Leicht, Mixture models and exploratory analysis in networks, P. Natl. Acad. Sci. USA 104 (23) (2007) 9564.
[26] S. Pinkert, J. Schultz, J. Reichardt, Protein interaction networksmore than mere modules, PLoS Computational Biology 6 (1) (2010) e1000659.
[27] L. Šubelj, M. Bajec, Ubiquitousness of link-density and link-pattern communities in real-world networks, Eur. Phys. J. B 85 (1) (2012) 1–11.
[28] J. Reichardt, D. R. White, Role models for complex networks, Eur. Phys. J. B 60 (2) (2007) 217–224.
[29] L. Šubelj, N. Blagus, M. Bajec, Group extraction for real-world networks: The case of communities, modules, and hubs and spokes, in: Proceedings of the International Conference on Network Science (Copenhagen, Denmark, 2013), 2013, pp. 152–153.
[30] M. Salehi, H. R. Rabiee, A. Rajabi, Sampling from complex networks with high community structures, Chaos 22 (2) (2012) 023126.
[31] A. S. Maiya, T. Y. Berger-Wolf, Sampling community structure, in: Proceedings of the 19th international conference on World wide web, ACM, 2010, pp. 701–710.
[32] B. Lužar, Z. Levnajić, J. Povh, M. Perc, Community structure and the evolution of interdisciplinarity in slovenia's scientific collaboration network, PloS One 9 (4) (2014) e94429.

[33] M. Perc, The matthew effect in empirical data, J. Roy. Soc. Interface 11 (98) (2014) 20140378.
[34] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine, Comput. Networks ISDN Syst. 30 (1) (1998) 107–117.
[35] N. Blagus, L. Šubelj, M. Bajec, Assessing the effectiveness of real-world network simplification, Physica A 413 (2014) 134–146.
[36] N. K. Ahmed, J. Neville, R. Kompella, Network sampling: from static to streaming graphs, e-print arXiv:11211.3412.
[37] M. Kurant, A. Markopoulou, P. Thiran, On the bias of BFS, in: Proceedings of the 22nd International Teletraffic Congress, IEEE, 2010, pp. 1–8.
[38] L. A. Goodman, Snowball sampling, Ann. Math. Stat. (1961) 148–170.
[39] Y. Zhao, E. Levina, J. Zhu, Community extraction for social networks, P. Natl. Acad. Sci. 108 (18) (2011) 7321–7326.
[40] G. Weiss, L. Šubelj, nets-nodegroups v1.0, `http://dx.doi.org/10.5281/zenodo.11589` (2014). doi:10.5281/zenodo.11589.
[41] L. Šubelj, S. Žitnik, N. Blagus, M. Bajec, Node mixing and group structure of complex software neworks, Advs. Complex Syst. 17 (2014) 1450022.
[42] P. Jaccard, Étude comparative de la distribution florale dans une portion des alpes et du jura, Bull. Soc. Vaud. Sci. Nat. 37 (1901) 547–579.
[43] S. Russel, P. Norvig, Artificial Intelligence: A Modern Approach (second edition), Upper Saddle River, N. J.: Prentice Hall, 2003.
[44] P. Erdős, A. Rényi, On random graphs i., Publ. Math. Debrecen 6 (1959) 290–297.
[45] G. Palla, I. Derényi, I. Farkas, T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, Nature 435 (7043) (2005) 814–818.
[46] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, A.-L. Barabási, Hierarchical organization of modularity in metabolic networks, Science 297 (5586) (2002) 1551–1555.
[47] M. Boguñá, R. Pastor-Satorras, A. Díaz-Guilera, A. Arenas, Models of social networks based on social distance attachment, Phys. Rev. E 70 (5) (2004) 056122.
[48] M. E. J. Newman, J. Park, Why social networks are different from other types of networks, Phys. Rev. E 68 (3) (2003) 036122.
[49] L. Šubelj, D. Fiala, M. Bajec, Network-based statistical comparison of citation topology of bibliographic databases, Sci. Rep. 4 (2014) 6496.
[50] R. D. Cook, S. Weisberg, Residuals and influence in regression.
[51] M. Najork, J. L. Wiener, Breadth-first crawling yields high-quality pages, in: Proceedings of the 10th international conference on World Wide Web, ACM, 2001, pp. 114–118.