

Assessing the effectiveness of real-world network simplification

Neli Blagus*, Lovro Šubelj, Marko Bajec

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

Abstract

Many real-world networks are large, complex and thus hard to understand, analyze or visualize. Data about networks are not always complete, their structure may be hidden, or they may change quickly over time. Therefore, understanding how an incomplete system differs from a complete one is crucial. In this paper, we study the changes in networks submitted to simplification processes (i.e., reduction in size). We simplify 30 real-world networks using six simplification methods and analyze the similarity between the original and simplified networks based on the preservation of several properties, for example, degree distribution, clustering coefficient, betweenness centrality, density and degree mixing. We propose an approach for assessing the effectiveness of the simplification process to define the most appropriate size of simplified networks and to determine the method that preserves the most properties of original networks. The results reveal that the type and size of original networks do not affect the changes in the networks when submitted to simplification, whereas the size of simplified networks does. Moreover, we investigate the performance of simplification methods when the size of simplified networks is 10% that of the original networks. The findings show that sampling methods outperform merging ones, particularly random node selection based on degree and breadth-first sampling.

Keywords: complex networks, network simplification, sampling, merging, simplification effectiveness

PACS: 64.60.aq, 89.75.Fb, 89.90.+n

1. Introduction

Over the past decade, network analysis [1, 2] has proved to be a suitable tool for describing diverse systems, understanding their structure and analyzing their properties. However, the evolution of the Web and the capability of storing large amounts of data have caused the size of networked systems and thus their complexity to increase. The algorithms for analyzing and visualizing networks appear impractical for addressing very large systems. Therefore, different methods have been proposed for the simplification of complex networks.

*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), marko.bajec@fri.uni-lj.si (Marko Bajec)

Preprint submitted to Physica A

April 29, 2014

Simplification is a process that reduces the size of a network by decreasing the number of nodes and links. The procedure is derived from graph theory (e.g., partitioning [3] and blockmodeling [4]) and was initially developed for compression and efficient graph storage [5, 6]. With the increasing complexity of networks, simplification methods also support clearer visualization [7, 8] and efficient analysis [9, 10]. In addition to these benefits, analyzing the changes undergone by networks under the effects of the simplification process enables us to explore and explain the differences between complete (i.e., original) and incomplete (i.e., simplified) systems (e.g., when only partial insight into the structure of network is available).

Recently, network simplification has been extensively investigated from different perspectives. Some studies have concentrated on the simplification of specific networks, such as simplifying social networks based on stability and retention [11], sampling scale-free [12] or directed networks [13], estimating different properties under social network crawling [14], sampling large dynamic peer-to-peer networks with random walks [15] or simplifying flow networks by removing useless links [16]. Other studies have attempted to provide a sufficient fit to original networks and thus observe the changes in network properties under the effects of simplification, such as preserving the clustering coefficient [17], degree distribution [18], community structure [19], spectral properties [20] or network connectivity [21].

However, only a few studies have focused on comparing simplification methods and measuring their success. Leskovec et al. [9] observed properties of original and simplified networks submitted to several simplification methods and measured their success based on random walk similarity. Lee et al. [10] analyzed basic network properties under the effects of three simplification methods and revealed characteristic patterns of changes in properties. Hübler et al. [22] compared their simplification algorithm to existing ones by measuring the average distance of properties between original and simplified networks. Toivonen et al. [23] studied the compression of weighted networks and measured the method's efficiency according to the running time and cost of the compressed network representation. Doer and Blenn [14] tested the convergence of different properties under three traversal algorithms applied to a single large social network. The findings of the aforementioned analyses indicate that the performances of simplification methods vary; however, the common weakness of these studies is the small set of networks considered.

Despite the above-described efforts, several open questions remain concerning the simplification of complex networks, such as those regarding (Q1) how to evaluate the similarity between original and simplified network, (Q2) how small simplified networks should be and ultimately (Q3) what simplification method should be used. In this paper, we address these questions and propose an approach for assessing the effectiveness of the simplification process. We analyze 30 real-world networks of different size and origin under the effects of six different simplification methods. We compare the original and simplified networks based on several network properties (e.g., degree distribution, clustering coefficient [24], betweenness centrality [25], degree mixing [26] and transitivity [27]) (Q1). The selection of these properties is supported by their common use in similar studies [9, 10]. Moreover, we propose a measure for determining the most appropriate size of simplified networks for preserving the observed properties (Q2) and for determining under which method the simplified networks fit the original ones most closely (Q3). We also study the impact of the original network size and type on the effectiveness of the simplification process.

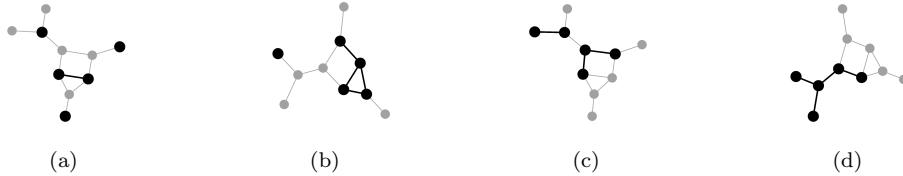


Figure 1: Sampling methods applied to a small sample network for $s = 0.5$. Black nodes represent simplified networks obtained with (a) selecting nodes uniformly at random (RN), (b) selecting nodes with probability proportional to degree (RD), (c) selecting links uniformly at random (RL) and (d) performing breadth-first search starting at a randomly selected node (BF). In the last method, BF ensures a connected network, whereas in other methods, this is not always the case.

The rest of the paper is structured as follows. Section 2 focuses on the simplification methods and real-world networks used in the study and describes the proposed measure. In section 3, we report and formally discuss the results of the analysis. Finally, section 4 concludes the paper and suggests directions for future research.

2. Methods and data

2.1. Simplification methods

Several authors have proposed a broad collection of simplification methods, which can be divided into two general classes. Those in the first class are sampling methods in which a simplified network is represented by a random sample of the original network (e.g., random node selection [28], random link selection [29], snowball sampling [30], random walk sampling [9] and forest fire [9]). Methods in the second class obtain simplified networks by merging nodes and links into supernodes and superlinks based on different characteristics, such as the distance between nodes (e.g., cluster-growing and box-tiling renormalization [31]), node and link attributes (e.g., link weights [32] and node attributes [33]) or community structure (e.g., balanced propagation and modularity optimization [34]).

In this study, we adopt four basic sampling methods (Fig. 1). Random node [28] (RN) and random link selection [29] (RL) create sampled networks with nodes or links selected uniformly at random. Simplified networks under random node selection based on degree [9] (RD) consist of randomly selected nodes, where the probability of selecting a node is proportional to the node’s degree. In breadth-first sampling (BF), a random node with its broad neighborhood is selected into the sample using the breadth-first search strategy. The main advantages of these methods are simplicity, and thus efficient implementation with low time complexity, and adjustability, which enables setting the size of the simplified network in advance.

Sampling methods outperform merging ones in terms of the advantages listed above. Still, we consider two methods from the merging class (Fig. 2). We use merging nodes based on community detection, where supernodes are identified by communities revealed by balanced propagation [35] (BP). We also employ cluster-growing renormalization [31] (CG), which incrementally grows supernodes from randomly selected seed nodes within a distance not larger than c (the nodes within one supernode are at most $2 \cdot c + 1$ steps

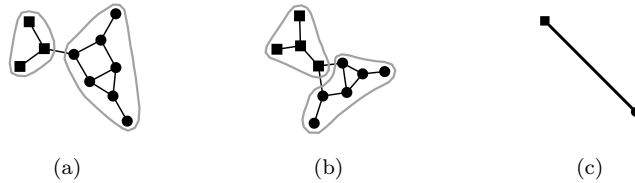


Figure 2: Merging methods applied to a small sample network. The shape of the nodes indicates (a) the nodes' community membership (BP) and (b) whether the nodes are at a distance less than 5 ($c = 2$) within one box (CG). Communities and boxes are marked by a gray contour. The simplified network (shown for both cases in (c)) is obtained by merging nodes inside one community or box into supernodes.

apart). Both methods proved well in analyzing the invariance of network density under different renormalizations [34].

We define s as the number of nodes in the simplified network, measured as the fraction of nodes in the original network. For sampling, we set the sizes of the simplified networks as varying from 1% to 50% of the original networks ($s = 0.01$ and $s = 0.05-0.50$ with a step size of 0.05). For BP, we set the parameters of the algorithm as suggested in [35]. With CG, we cannot control the size of the simplified network; still, we can change the distance between the nodes within one supernode. Therefore, the parameter c ranges from two to six, where smaller values indicate a smaller number of nodes within one supernode and thus a larger simplified network.

2.2. Network data

A diverse set of real-world systems is analyzed. We consider 30 networks of different origin (e.g., information, technological and social) and size (varying from a few thousand to a few hundred thousand nodes), listed in Table 1. Due to the large number of networks considered, a detailed description is omitted here.

For BP, CG and BF, all networks are considered to be undirected, although some of them are directed. To avoid comparing networks of different complexity, we remove self-loops and multiple links from all networks for simplification via merging methods.

2.3. Assessment approach

To perform a fair and sound assessment, we first address the aforementioned questions concerning the comparison approach (Q1) and the size to which a certain network should be simplified (Q2). To address Q1, we select a set of local and global network properties to be observed. To address Q2, we introduce a simple measure that takes into account all of the selected properties and for each network calculates the simplified size that would best preserve the observed properties. The specific size of the simplified networks is then used in a further analysis to compare the selected simplification methods (Q3).

2.3.1. Comparing original and simplified networks

We compare networks based on eight fundamental global and local properties. The global properties are expressed by a single value for each network and include density (the ratio of existing links to all possible links), degree mixing (the tendency of nodes connecting to similar ones [26]) and transitivity (the number of closed triplets over the

Table 1: Real-world networks (n and m correspond to the number of nodes and links, respectively).

Network	Type	n	m
<i>High E. Particle Phys.</i> [36]	Citation	27240	342437
<i>High E. Phys.</i> [37]		34546	421578
<i>NBER US patents</i> [38]		240548	561060
<i>Citeseer publications</i> [39]		384413	1764929
<i>PGP web-of-trust</i> [40]	Collaboration	10680	24340
<i>High E. Phys. archive</i> [41]		12008	237010
<i>Astro Phys. archive</i> [41]		18772	396160
<i>Cond. Matters archive</i> [41]		23133	186936
<i>Computer science</i> [42]		317080	1049866
<i>Digg user reply</i> [43]	Communication	30398	87627
<i>Emails at Enron</i> [44]		36692	367662
<i>Facebook wall post</i> [45]		46952	876993
<i>Emails at EU res. inst.</i> [41]		265214	420045
<i>Amazon products 1</i> [44]	Co-purchase	334863	925872
<i>Amazon products 2</i> [46]		403394	3387388
<i>Flickr images metadata</i> [47]	Co-occurrence	105938	2316948
<i>Oregon aut. systems</i> [48]	Internet	22963	48436
<i>Gnutella file sharing 1</i> [41]		36682	88328
<i>Gnutella file sharing 2</i> [41]		62586	147829
<i>Foldoc dictionary</i> [49]	Information	13356	120238
<i>Wikipedia votes</i> [50]	On-line social	7115	103689
<i>Brightkite friendship</i> [51]		58228	214078
<i>Epinions trust</i> [52]		75879	508837
<i>Slashdot friendship</i> [53]		82168	948464
<i>Wikipedia interactions</i> [54]		186485	740397
<i>Gowalla friendship</i> [51]		196591	1900654
<i>Broad-topic queries</i> [55]		6175	16150
<i>google.com internal</i> [56]		15763	171206
<i>nd.edu domain</i> [57]	Web graph	325729	1497134
<i>Baidu articles</i> [58]		415641	3284387

total number of triplets [27]). The local properties are described by a distribution for all nodes in the network and comprise degree, in-degree and out-degree (the number of neighbors of each node), local clustering coefficient (the proportion of connected neighbors of each node [24]) and betweenness centrality (the number of shortest paths between all nodes going through each node [25]).

For comparison, we define two similarity measures, one based on the selected global properties and one on the selected local properties. The global similarity measure is used to determine how correlated the global properties in the observed original networks and their simplified version are. The correlation is measured with Spearmans correlation coefficient ρ . ρ indicates the extent to which one variable decreases as another increases. In our analysis, we calculate ρ for each selected simplification method and each size of the simplified networks for all networks together.

The comparison based on the selected local properties is expressed using the Kolmogorov-Smirnov D -statistic (Kolmogorov-Smirnov test checks the null-hypothesis, i.e., that the distributions of two properties are the same; the D -statistic measures the distance between the observed distributions). The D -statistic for each network and its simplified version is calculated for each simplification method separately. The values for comparison based on ρ and the D -statistics are averaged over ten simplifications of each network, each simplification method and each size of the simplified networks.

Table 2: An illustrative example of the assessment approach. (left) The results of a comparison between simplified and original networks based on global properties. (right) The results after ranking sizes for each property. P_i denotes properties and S_i sizes of simplified networks.

	P1	P2	P3		P1	P2	P3	Sum	A
S1	0.84	0.69	0.75	S1	5	5	5	15	1.000
S2	0.88	0.89	0.87	S2	4	4	4	12	0.800
S3	0.90	0.92	0.89	S3	3	3	2	8	0.533
S4	0.96	0.95	0.88	S4	0	1	3	4	0.267
S5	0.91	0.94	0.92	S5	2	2	0	4	0.267
S6	0.93	0.96	0.90	S6	1	0	1	2	0.133

The selection of properties and their relevance in assessing the effectiveness of network simplification greatly depends on the purpose of the simplification being performed. The selection of particular properties in this analysis is only supported by their common use in similar studies (e.g., [9, 10]) and serve to demonstrate the effectiveness of the proposed approach. Note that comparing networks based on other sets of properties may lead to different results.

In the literature, we can find studies that have performed similar comparisons to a limited extent. In [12] the authors proved that RN does not preserve the degree distribution of scale-free networks. Moreover, RN and RL sampling are biased toward nodes with high degrees, which affects the degree distribution [9]. However, Lee et al. [10] proved that RN and RL overestimate the degree and betweenness centrality exponent, whereas both methods retain the assortativity of original networks. Merging methods decreases the density [34], but the relationship between density and network size remains invariant after simplification.

2.3.2. Determining simplified network sizes

To determine the size to which a specific network can be decreased while preserving most of the observed properties, the following approach is used. For each simplification method and each global and local property, we rank sizes with respect to ρ and the D -statistic, respectively. The network size that best fits a specific property receives rank 0, the next best one receives rank 1 and so on. Next, we sum the ranks for each size and divide the sum by the greatest possible sum of ranks to normalize the result to the interval $[0, 1]$. Thus, the measure A is defined as

$$A = \frac{1}{(n_s - 1) \cdot n_p} \sum_{i=1}^{n_p} r_i, \quad (1)$$

where n_s denotes the number of different sizes, n_p denotes the number of properties, i indexes the properties (the order is not important) and r_i is the rank of the i -th property. A is thus the normalized total rank assigned to a specific size by the observed properties.

Table 2 shows an example of the measure A calculated by comparing six different sizes for a simplified network, taking into account the measure ρ a specific size receives for each of the three observed global properties. In this example, the most appropriate size for preserving global properties is S6.

2.3.3. Comparing simplification methods

Finally, we compare the different methods for a given size of a simplified network. We rank the methods and measure their effectiveness using a modified version of the measure A described in the previous subsection:

$$A = \frac{1}{(n_m - 1) \cdot n_p} \sum_{i=1}^{n_p} r_i, \quad (2)$$

where n_m is the number of different methods.

With the described measure, we regard all properties as equally important. Still, depending on the purpose of the simplified networks considered and the method by which those networks are analyzed, one property can be more essential than another. With respect to importance, we can assign weights w to the properties; thus, the measure A becomes

$$A_w = \frac{1}{n_m \cdot \sum_{i=1}^{n_p} w_i} \sum_{i=1}^{n_p} r_i \cdot w_i, \quad (3)$$

where w is the vector of weights and thus w_i denotes the weight of property i .

For simplicity, we omit the analysis performed based on the measure A_w and thus assume all properties are equally important.

3. Analysis and discussion

The analysis consists of two stages. First, we determine the size of the simplified networks that ensure adequate preservation of the observed properties. Second, we compare the effectiveness of different methods for a specific size of the simplified networks.

3.1. Effectiveness of the simplification process with respect to the size of the simplified networks

First, we analyze the effect of simplified network size on the effectiveness of the simplification process. As expected, the results reveal that in the majority of cases, the largest simplified networks ($c = 2$ for CG and $s = 0.5$ for sampling methods) are more similar to the original networks and thus better fit the original networks' properties. However, the main goal of the simplification is to sufficiently reduce large networks to allow for easier analysis and understanding, which is achieved when the simplified networks are smaller. Therefore, we define the best size as the local minimum of A achieved at the smallest simplified network size (we assume that $A = 1$ for $s = 0$ and take the global minimum if it is also local).

3.1.1. Analysis of the sampling methods

The analysis of the sampling methods reveals a high level of diversity in their effectiveness (Fig. 3 and Table 3). Fig. 3(a) shows that under simplification methods RN and RL, local properties are best preserved for the largest size of the simplified networks ($s = 0.5$). In contrast, RD and BF perform best for smaller sizes, between $s = 0.01$ and $s = 0.15$, for the majority of the networks (i.e., the local minimum of A is around these values for most of the networks).

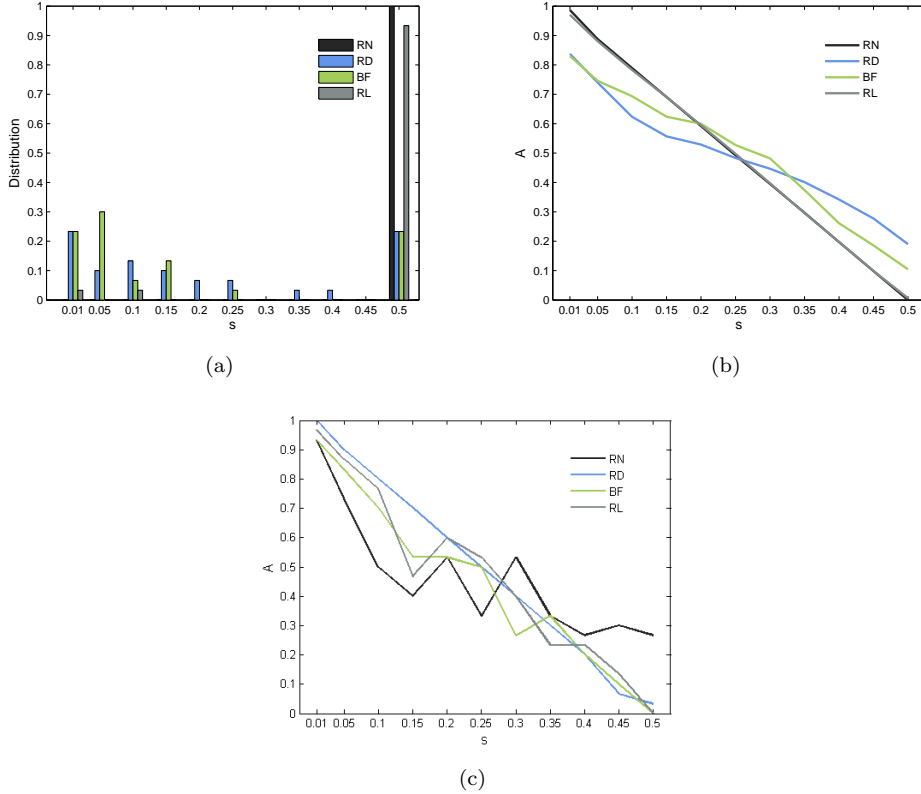


Figure 3: The results for the sampling methods. (a) Portion of networks with the best size equal to s . (b) Distance between the original and simplified networks (average A over all networks) based on the local properties. (c) Distance between original and simplified networks based on the global properties.

Fig. 3(b) and 3(c) shows the average A over all networks for the local and global properties, respectively. For the former, all methods behave in a similar manner. In particular, the best fit of local properties is reached for larger simplified networks; still, RD and BF show some deviation, indicating that for several networks smaller sizes also provide good fits. For the global properties, RN and RL show similar behavior again because the best preservation is achieved on smaller simplified networks ($s = 0.15$). For BF and RD, the local and global minima are reached for larger simplified networks.

Table 4 shows the best sizes of simplified networks for the preservation of each network property. RN and RL perform similarly because both provide better preservation of local properties for the largest simplified networks. On the other hand, for RD, degree is best preserved for smaller networks, whereas for medium-sized networks, out-degree and clustering are best preserved. For BF, distributions of degree, out-degree and in-degree change the least for $s = 0.01, 0.15$. However, the methods behave in a different manner when preserving global properties. Only RN preserves density and degree mixing well on smaller simplified networks, whereas RD, BF and RL work best for $s = 0.5$.

Finally, we analyze how the preservation of local properties depends on the size and

Table 3: The best sizes c or s for the preservation of the local properties with corresponding A .

Network	CG	RD	BF
<i>High E. Particle Phys.</i>	2 (0.00)	0.20 (0.18)	0.05 (0.38)
<i>High E. Phys.</i>	2 (0.08)	0.25 (0.14)	0.10 (0.44)
<i>NBER US patents</i>	2 (0.25)	0.35 (0.20)	0.01 (0.44)
<i>Citeseer publications</i>	2 (0.00)	0.20 (0.08)	0.01 (0.58)
<i>PGP web-of-trust</i>	4 (0.33)	0.25 (0.40)	0.05 (0.77)
<i>High E. Phys. archive</i>	2 (0.00)	0.05 (0.73)	0.05 (0.83)
<i>Astro Phys. archive</i>	2 (0.00)	0.50 (0.17)	0.05 (0.50)
<i>Cond. Matters archive</i>	2 (0.00)	0.50 (0.13)	0.05 (0.70)
<i>Computer science</i>	2 (0.17)	0.50 (0.00)	0.50 (0.00)
<i>Digg user reply</i>	2 (0.25)	0.10 (0.20)	0.05 (0.28)
<i>Emails at Enron</i>	2 (0.00)	0.01 (0.57)	0.50 (0.00)
<i>Facebook wall post</i>	2 (0.00)	0.10 (0.18)	0.01 (0.40)
<i>Emails at EU res. inst.</i>	2 (0.00)	0.50 (0.00)	0.15 (0.60)
<i>Amazon products 1</i>	4 (0.33)	0.50 (0.00)	0.50 (0.00)
<i>Amazon products 2</i>	2 (0.00)	0.50 (0.00)	0.50 (0.02)
<i>Flickr images metadata</i>	3 (0.33)	0.01 (0.80)	0.25 (0.37)
<i>Oregon aut. systems</i>	2 (0.17)	0.40 (0.23)	0.01 (0.93)
<i>Gnutella file sharing 1</i>	5 (0.58)	0.15 (0.34)	0.15 (0.34)
<i>Gnutella file sharing 2</i>	5 (0.42)	0.15 (0.34)	0.10 (0.36)
<i>Foldoc dictionary</i>	2 (0.17)	0.50 (0.00)	0.50 (0.02)
<i>Wikipedia votes</i>	2 (0.00)	0.01 (0.26)	0.01 (0.76)
<i>Brightkite friendship</i>	2 (0.00)	0.05 (0.37)	0.05 (0.83)
<i>Epinions trust</i>	2 (0.00)	0.01 (0.94)	0.01 (0.70)
<i>Slashdot friendship</i>	2 (0.17)	0.01 (0.30)	0.01 (0.58)
<i>Wikipedia interactions</i>	4 (0.47)	0.01 (0.42)	0.05 (0.44)
<i>Gowalla friendship</i>	2 (0.00)	0.05 (0.33)	0.05 (0.03)
<i>Broad-topic queries</i>	4 (0.33)	0.10 (0.26)	0.15 (0.34)
<i>google.com internal</i>	2 (0.00)	0.15 (0.34)	0.15 (0.36)
<i>nd.edu domain</i>	4 (0.08)	0.01 (0.48)	0.50 (0.32)
<i>Baidu articles</i>	2 (0.00)	0.10 (0.22)	0.05 (0.50)

type of the original networks (Table 3). We omit the results for RN and RL because in all cases except two, the best size is $s = 0.5$. In contrast, the effectiveness of RD is partially correlated to the original network size because medium-sized networks ($n = 50000 - 200000$) are best preserved for smaller simplified network sizes ($s = 0.01 - 0.1$), whereas large networks ($n = 200000 - 500000$) are best preserved for larger values of s . However, as indicated by the dependence on network type, the local properties of on-line social networks and Web graphs are best preserved for smaller sizes $s = 0.01 - 0.15$, whereas the local properties of citation and co-purchase networks are best preserved for $s = 0.25 - 0.35$. All differences are statistically significant ($p < 0.05$, one-way ANOVA), which rejects the null hypothesis that there is no dependence between the effectiveness of property preservation and network type. For both RD and BF, only the properties of co-purchase and information networks are best preserved for $s = 0.5$. The results reveal no statistically significant influence of network size or type on the performance of BF.

3.1.2. Analysis of the merging methods

The analysis of CG proves that the local network properties are best preserved when $c = 2$ for 22 out of 30 networks (Fig. 4(a) and Table 3). Fig. 4(b) shows the average A over all networks based on the local and global properties. The local properties are best fitted for larger simplified networks ($c = 2$), whereas for $c = 3, 4$ the simplified networks best fit the global properties of the original networks.

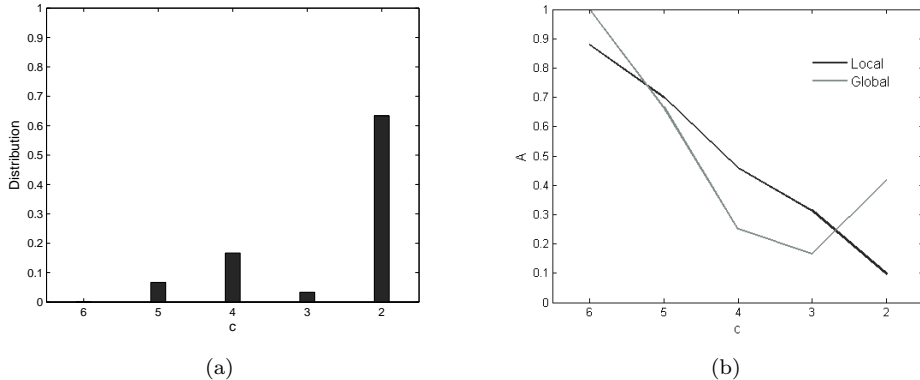


Figure 4: The results for cluster-growing simplification. (a) Portion of networks with the best size equal to c . (b) Distance between the original and simplified networks (A for the global and average A over all networks for the local properties) as a function of c .

Table 4: The best sizes c or s for the preservation of local network properties with corresponding A , and ρ for the global properties.

Property	CG	RN	RD	BF	RL
Degree	2 (0.04)	0.50 (0.00)	0.15 (0.52)	0.15 (0.61)	0.50 (0.00)
In degree	-	0.50 (0.00)	0.50 (0.26)	0.01 (0.75)	0.50 (0.02)
Out degree	-	0.50 (0.00)	0.30 (0.43)	0.01 (0.61)	0.50 (0.01)
Clustering	2 (0.25)	0.50 (0.00)	0.25 (0.44)	0.50 (0.10)	0.50 (0.00)
Betweenness	2 (0.00)	0.50 (0.00)	0.50 (0.08)	0.50 (0.05)	0.50 (0.01)
Density	2 (0.89)	0.10 (0.97)	0.45 (0.95)	0.50 (0.95)	0.50 (0.91)
Degree mixing	3 (0.34)	0.35 (0.66)	0.50 (0.77)	0.50 (0.97)	0.50 (0.63)
Transitivity	4 (0.36)	0.50 (0.99)	0.50 (0.99)	0.50 (0.99)	0.50 (0.83)

Table 4 shows the results obtained for the preservation of each property. Most of the properties are best preserved for larger simplified networks ($c = 2$), with the exception of degree mixing and transitivity, where $c = 5$ and $c = 6$, respectively.

The best size for preserving local network properties (Table 3) does not depend on the original network size or type (i.e., the differences in property preservation, which would depend on the size and type of the original networks, are not statistically significant). Still, if we divide the networks roughly by type, i.e., information, social and technological, the correlation between the type and the effectiveness becomes statistically significant (i.e., the null hypothesis that there are no differences in property preservation, which would depend on network type, is rejected, with $p < 0.05$, one-way ANOVA). Thus, the local properties of social networks are best preserved for $c = 2$, in contrast to the case of technological networks, for which $c > 2$.

3.1.3. Discussion

The findings of the first part of the study confirm the negative correlation between the size of the simplified networks and their similarity to the original networks because larger simplified networks are more similar to the original ones in most cases. The latter has also been proved by other studies, for example, [10]. RD and BF are more effective

Table 5: The best, second-best and worst methods for the preservation of local network properties with corresponding A , and ρ for the global properties.

Property	Best	Second-best	Worst
Degree	BF (0.25)	RD (0.26)	RL (0.84)
In degree	RD/BF (0.26)	RL (0.70)	RN (0.77)
Out degree	RD (0.32)	BF (0.33)	RL (0.70)
Clustering	RD (0.30)	BF (0.35)	RL (0.81)
Betweenness	BF (0.21)	RD (0.27)	BP (0.75)
Density	RN (0.96)	BF (0.91)	BP (0.76)
Degree mixing	BF (0.92)	RN (0.62)	BP (0.21)
Transitivity	RN (0.94)	RD (0.92)	CG (0.22)

for smaller simplified networks, which is consistent with the findings of other authors. Particularly, Doerr and Blenn [14] revealed a solid estimate of an original network for $s = 0.2-0.3$ and $s = 0.1$ in the case of preserving average node degree and the power-law degree exponent, respectively. In addition, Leskovec and Faloutsos [9] obtained a good fit for original networks under several sampling methods for $s = 0.15$. Thus, our results advance those reported in these studies and reveal distinctions in the extent of property preservation among different types and sizes of networks, which are the most obvious for RD.

3.2. Comparison of the effectiveness of the simplification methods

In the second part of our study, we compare the performance of different simplification methods. We focus on size $c = 2$ for CG and $s = 0.1$ for sampling methods for two reasons. First, we select $s = 0.1$ as the middle size among the best sizes determined in the first part of the study. Second, $s = 0.1$ is suitable for the comparison of BP and CG, for which the mean sizes of simplified networks are $s = 0.03$ and $s = 0.12$, respectively.

3.2.1. Analysis

First, we determine the best method for preserving a specific property (Table 5). Global properties are best preserved under RN and BF, whereas merging methods provide the worst preservation. Fig. 5 compares the best, second-best and worst methods with respect to all global properties. For local properties, BF and RD perform the best, particularly BF for the degree and betweenness centrality, whereas RD performs best for the out-degree and clustering. However, RL proves to be the worst method because it preserves the degree, out-degree and clustering to the lowest extent. Examples of local property preservation for the analyzed networks are presented in Fig. 6.

For a complete assessment of the effectiveness of the simplification methods, we compare the performance of the methods for each network based on the preservation of local properties. Results are represented in Table 6. For 23 networks, the best methods are RD and BF. The analysis reveals a dependence between network type and method effectiveness because BP performs the best for on-line social networks and BF performs the best for Internet and co-purchase networks. The differences among the network types are statistically significant ($p < 0.05$, one-way ANOVA). For the second-best methods, the distinctions are less evident. Still, BP proves to be effective for other types of networks (Internet, communication networks, Web graphs). The worst method for preserving local properties is RL (for 22 networks), followed by RN (for 8 networks). On the other

Table 6: The best, second-best and worst methods for preserving local properties of networks with corresponding A .

Network	Best	Second-best	Worst
<i>High E. Particle Phys.</i>	RD (0.10)	BF (0.17)	RL (0.97)
<i>High E. Phys.</i>	BF (0.07)	RD (0.20)	RL (0.96)
<i>NBER US patents</i>	BF (0.07)	BP (0.13)	RN/RL (0.80)
<i>Citeseer publications</i>	RD (0.07)	BF (0.20)	RL (0.93)
<i>PGP web-of-trust</i>	CG (0.13)	BP (0.20)	RN (0.93)
<i>High E. Phys. archive</i>	RD (0.07)	BF (0.20)	RL (0.93)
<i>Astro Phys. archive</i>	RD (0.07)	BF (0.13)	RL (1.00)
<i>Cond. Matters archive</i>	BF (0.00)	RD (0.20)	RL (1.00)
<i>Computer science</i>	BF (0.07)	RD (0.27)	RL (1.00)
<i>Digg user reply</i>	RD (0.17)	CG/BP (0.33)	RL/RN (0.60)
<i>Emails at Enron</i>	BP (0.27)	RD (0.33)	RN (0.73)
<i>Facebook wall post</i>	RD (0.07)	BP (0.17)	RL (1.00)
<i>Emails at EU res. inst.</i>	RL (0.13)	BP (0.20)	RD (0.73)
<i>Amazon products 1</i>	BF (0.00)	BP (0.27)	RL (1.00)
<i>Amazon products 2</i>	BF (0.03)	CG (0.10)	RL (1.00)
<i>Flickr images metadata</i>	RD/BF (0.33)	RN (0.47)	RL (0.73)
<i>Oregon aut. systems</i>	RD (0.07)	BP (0.20)	RN (0.80)
<i>Gnutella file sharing 1</i>	BF (0.13)	BP (0.30)	RL (0.70)
<i>Gnutella file sharing 2</i>	BF (0.13)	BP (0.30)	RL (0.70)
<i>Foldoc dictionary</i>	BF (0.03)	CG/BP (0.13)	RL (1.00)
<i>Wikipedia votes</i>	BP (0.13)	RN (0.27)	BF (0.60)
<i>Brightkite friendship</i>	RD (0.13)	BP (0.20)	RL (0.93)
<i>Epinions trust</i>	BP (0.03)	RL (0.17)	BF (0.87)
<i>Slashdot friendship</i>	BP (0.07)	RD (0.23)	RL (0.83)
<i>Wikipedia interactions</i>	BP (0.07)	BF (0.33)	RN (0.40)
<i>Gowalla friendship</i>	RD (0.07)	BP (0.27)	RL (1.00)
<i>Broad-topic queries</i>	RD (0.07)	BP (0.27)	RN (0.73)
<i>google.com internal</i>	RD (0.10)	BF (0.17)	RL (0.93)
<i>nd.edu domain</i>	CG (0.07)	BP/BF (0.13)	RL/RN (0.80)
<i>Baidu articles</i>	RD (0.00)	BP (0.27)	RL (0.90)

hand, BF is the worst with respect to only two on-line social networks. The results also prove the statistically significant dependence (i.e., reject the null hypothesis that there are no dependencies between the network size and the effectiveness of the simplification methods, with $p < 0.05$, one-way ANOVA) between the worst method and network size. For smaller networks ($n < 50000$), the worst method for preserving local properties is RL, whereas for larger ones, the worst method is RN.

3.2.2. Discussion

The results of the second part of the study reveal several distinctions in the behavior of the simplification methods. RD and BF proved the best for preserving the local properties of networks, whereas for global properties, RN outperforms the other methods. However, RL and merging methods show the worst performance. These findings are consistent with the results of the study reported in [9], where RD had a better performance than RN and RL (other methods are not considered in the aforementioned study).

In addition to comparing the methods for $s = 0.1$, we also compare them for larger simplified networks ($s = 0.5$). The results are not presented because there are no significant changes in the results (i.e., the same methods are the best and the worst for $s = 0.1$).

In addition, we observe how the size of the largest weakly connected component

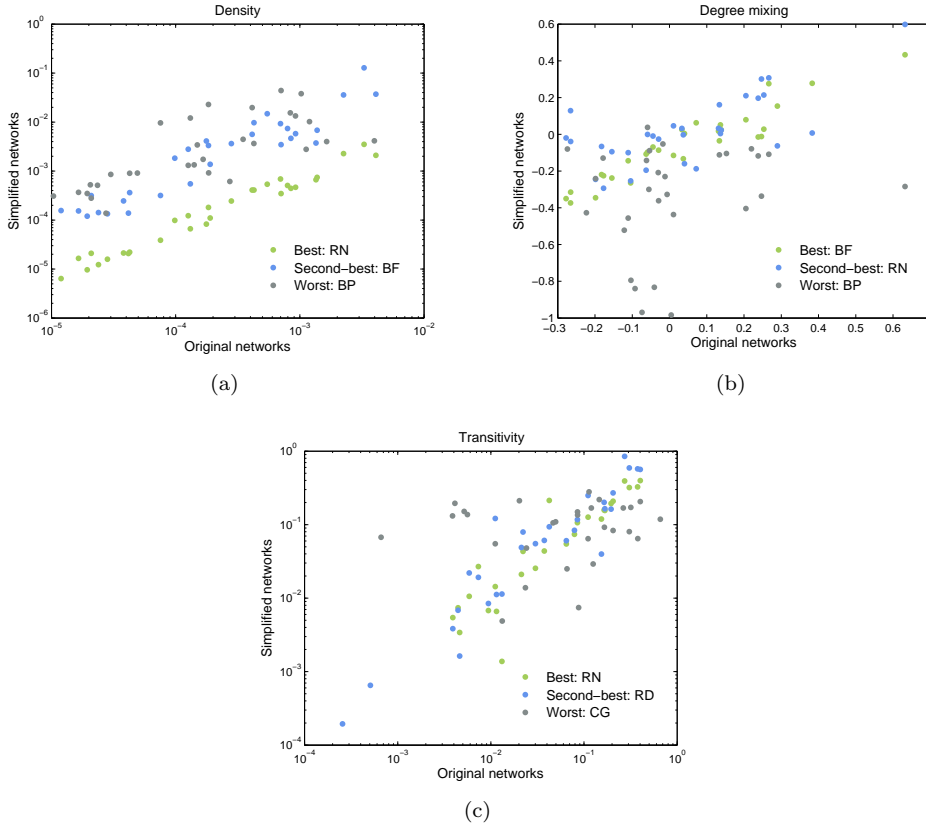


Figure 5: Relationship between the global properties of the original and the simplified networks for the best, second-best and worst method. (a) Density. (b) Degree mixing. (c) Transitivity.

(LWCC) changes under simplification to explain the differences in the methods' performance. The LWCC of the original networks, on average, consists of 59% of all nodes. The size of the LWCC of the simplified networks under all methods depends strongly on the simplified network size (i.e., the size of the LWCC of the smallest simplified networks is the smallest). However, RN and RL show similar performance because the sizes of the LWCC for both methods vary from 1% for $s = 0.1$ to 40% for $s = 0.5$. Still, RL produces the most disconnected components. In contrast, simplification via RD and BF produces networks with a clearly larger LWCC because the sizes vary from 25% for $s = 0.01$ to 60% for $s = 0.5$. Therefore, networks simplified by RD and BF feature a larger LWCC and smaller number of components, which is more similar to the characteristics of the original networks. Based on this finding, the predominance of RD and BF over RN and RL can be confirmed.

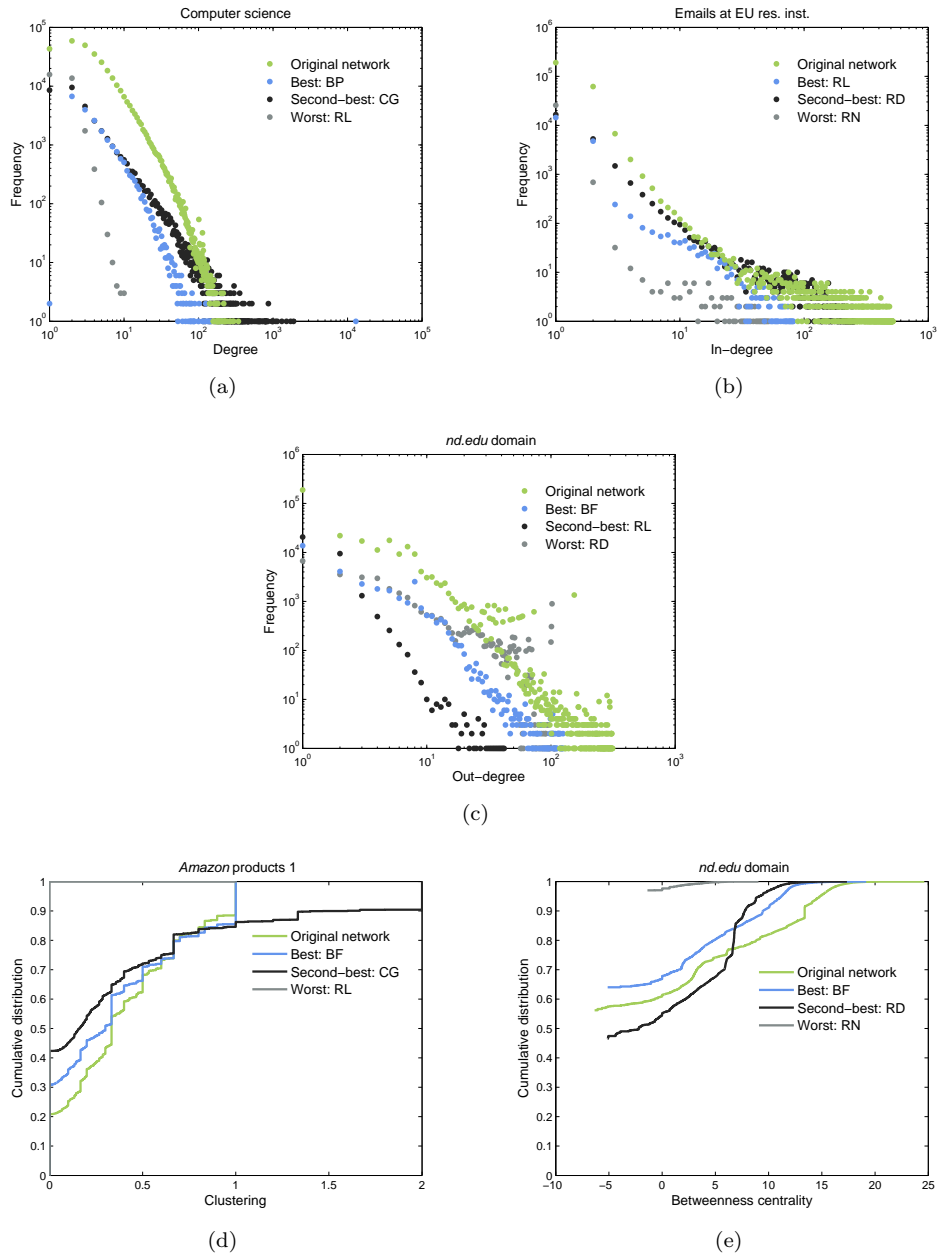


Figure 6: Examples of comparison of the local properties for the original and simplified networks with the best, second-best and worst methods. (a) Degree distribution. (b) In-degree distribution. (c) Out-degree distribution. (d) Cumulative distribution of clustering. (e) Cumulative distribution of of betweenness centrality.

4. Conclusions

Network simplification is an adequate tool for studying large networks for several reasons. In addition to the obvious advantages, including faster analysis and more efficient visualization, the simplification can significantly improve the understanding of large networks. For example, data regarding the systems described by networks can often be missing or incomplete, and thus, networks can be considered a sampled variety of the original systems (e.g., identifying Internet map [59, 60]). For this reason, understanding how similar the original and sampled system are is essential.

This study addressed three aspects of real-world network simplification. First, we focused on a comparison of original and simplified networks. Second, we determined what size of simplified network most adequately fits the properties of the original networks. Finally, we compared the effectiveness of several simplification methods. We analyzed six simplification methods with respect to 30 real-world networks and compared the simplified and original networks based on several properties, including degree, in-degree, out-degree and betweenness centrality distribution, clustering coefficient, density, degree mixing and transitivity.

The results show that the goodness of property preservation depends on the size of the simplified networks. Larger simplified networks fit original networks better; nevertheless, properties are adequately preserved for smaller sizes close to 10% the size of the original networks, especially for random node selection based on degree and breadth-first sampling. Thus, the decision regarding how small a simplified network should be depends on the size of the original network and the purpose of the simplified network. If we can simplify a network by 50%, we can provide for the best fit of the original network properties. However, if the network is large, 50% of the original size is not a sufficient reduction. In that case, 10% of the original network size allows for the adequate preservation of important properties. Furthermore, the findings of this study reveal that random node selection based on degree and breadth-first sampling are the best methods, whereas merging methods performed the worst.

Future work will mainly focus on other characteristics that affect the effectiveness of the simplification process. Moreover, instead of focusing solely on similarities, we will analyze typical distinctions between original and simplified networks. Furthermore, other ways for comparing simplified networks with original for their similarity could also be considered, for example comparing the backbones of networks [61], their community structure [62] or density of edges in subnetworks [63]. Based on this and future studies, a wide range of principles underlying the simplification of real-world networks could be extracted. The application of such principles should allow for the determination of the most suitable simplification method for specific networks, which would allow for more efficient simplification and a better understanding of large real-world networks.

Acknowledgment

The work has been supported by the Slovene Research Agency *ARRS* within the research program P2-0359.

References

- [1] M. E. J. Newman, A.-L. Barabási, D. J. Watts, *The structure and dynamics of networks*, Princeton University Press, 2006.
- [2] M. E. J. Newman, *Networks: An introduction*, Oxford University Press, 2009.
- [3] T. Feder, R. Motwani, Clique partitions, graph compression and speeding-up algorithms, in: *Proceedings of the 23th Annual ACM Symposium on Theory of Computing*, ACM, 1991, pp. 123–133.
- [4] P. Doreian, V. Batagelj, A. Ferligoj, *Generalized blockmodeling*, Cambridge University Press, 2005.
- [5] N. Deo, B. Litow, A structural approach to graph compression, in: *Proceedings of the 23th MFCS Workshop on Communications*, Citeseer, 1998, pp. 91–101.
- [6] M. Adler, M. Mitzenmacher, Towards compressing web graphs, in: *Proceedings of the Data Compression Conference*, IEEE, 2001, pp. 203–212.
- [7] D. Rafiei, Effectively visualizing large networks through sampling, in: *Visualization*, IEEE, 2005, pp. 375–382.
- [8] 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.
- [9] 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.
- [10] S. H. Lee, P. J. Kim, H. Jeong, Statistical properties of sampled networks, *Phys. Rev. E* 73 (1) (2006) 016102.
- [11] M. Kudelka, Z. Horak, V. Snasel, A. Abraham, Social network reduction based on stability, in: *Proceedings of the International Conference on Computational Aspects of Social Networks*, IEEE, 2010, pp. 509–514.
- [12] 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.
- [13] 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.
- [14] C. Doerr, N. Blenn, Metric convergence in social network sampling, in: *Proceedings of the 5th ACM workshop on HotPlanet*, ACM, 2013, pp. 45–50.
- [15] D. Stutzbach, R. Rejaie, N. Duffield, S. Sen, W. Willinger, On unbiased sampling for unstructured peer-to-peer networks, in: *Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement*, 2006, pp. 27–40.
- [16] T. Biedl, B. Brejová, T. Vinař, Simplifying flow networks, *Lect. Notes Comput. Sc.* (2000) 192–201.
- [17] H. Sethu, X. Chu, A new algorithm for extracting a small representative subgraph from a very large graph, e-print arXiv:1207.4825.
- [18] M. P. H. Stumpf, C. Wiuf, Sampling properties of random graphs: the degree distribution, *Phys. Rev. E* 72 (3) (2005) 036118.
- [19] H. Maserrat, J. Pei, Community preserving lossy compression of social networks, in: *Proceedings of the 12th International Conference on Data Mining*, IEEE, 2012, pp. 509–518.
- [20] D. Gfeller, P. De Los Rios, Spectral coarse graining of complex networks, *Phys. Rev. Lett.* 99 (3) (2007) 38701.
- [21] 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.
- [22] 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.
- [23] 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.
- [24] D. J. Watts, S. H. Strogatz, Collective dynamics of small-world networks, *Nature* 393 (6684) (1998) 440–442.
- [25] U. Brandes, A faster algorithm for betweenness centrality, *J. Math. Sociol.* 25 (2) (2001) 163–177.
- [26] M. E. J. Newman, Assortative mixing in networks, *Phys. Rev. Lett.* 89 (20) (2002) 208701.
- [27] S. Wasserman, *Social network analysis: Methods and applications*, Cambridge university press, 1994.
- [28] 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.

- [29] N. Ahmed, J. Neville, R. R. Kompella, Network sampling via edge-based node selection with graph induction, Tech. rep., Purdue University (2011).
- [30] J. Illenberger, G. Flötteröd, Estimating properties from snowball sampled networks, Tech. rep., VSP Working Paper 11-01, TU Berlin, Transport Systems Planning and Transport Telematics (2011).
- [31] L. K. Gallos, C. Song, H. A. Makse, A review of fractality and self-similarity in complex networks, *Physica A* 386 (2) (2007) 686–691.
- [32] H. Toivonen, F. Zhou, A. Hartikainen, A. Hinkka, Network compression by node and edge mergers, in: *Bisociative Knowledge Discovery*, Springer, 2012, pp. 199–217.
- [33] Y. Zhou, H. Cheng, J. Yu, Graph clustering based on structural/attribute similarities, *Proc. VLDB Endowment* 2 (1) (2009) 718–729.
- [34] N. Blagus, L. Šubelj, M. Bajec, Self-similar scaling of density in complex real-world networks, *Physica A* 391 (8) (2012) 2794–2802.
- [35] L. Šubelj, M. Bajec, Robust network community detection using balanced propagation, *Eur. Phys. J. B* 81 (3) (2011) 353–362.
- [36] KDD Cup '03, <http://www.cs.cornell.edu/projects/kddcup/> (2013).
- [37] 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.
- [38] B. H. Hall, A. B. Jaffe, M. Tratjenberg, The NBER patent citation data file (2001).
- [39] K. D. Bollacker, S. Lawrence, C. L. Giles, Citeseer: An autonomous web agent for automatic retrieval and identification of interesting publications, in: *Proceedings of the 2nd International Conference on Autonomous Agents*, ACM, 1998, pp. 116–123.
- [40] 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.
- [41] J. Leskovec, J. Kleinberg, C. Faloutsos, Graph evolution: Densification and shrinking diameters, *ACM Trans. Knowl. Discov. Data* 1 (1) (2007) 1–40.
- [42] 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.
- [43] M. De Choudhury, H. Sundaram, A. John, D. D. Seligmann, Social synchrony: Predicting mimicry of user actions in online social media, in: *International Conference on Computational Science and Engineering*, IEEE, 2009, pp. 151–158.
- [44] J. Yang, J. Leskovec, Defining and evaluating network communities based on ground-truth, in: *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, ACM, 2012, p. 3.
- [45] B. Viswanath, A. Mislove, M. Cha, K. P. Gummadi, On the evolution of user interaction in facebook, in: *Proceedings of the 2nd ACM Workshop on Online Social Networks*, ACM, 2009, pp. 37–42.
- [46] J. Leskovec, L. A. Adamic, B. A. Huberman, The dynamics of viral marketing, *ACM Trans. Web* 1 (1) (2007) 5.
- [47] J. McAuley, J. Leskovec, Learning to discover social circles in ego networks, in: *Advances in Neural Information Processing Systems* 25, 2012, pp. 548–556.
- [48] M. E. J. Newman, Network data, <http://www-personal.umich.edu/~mejn/netdata/> (2013).
- [49] V. Batagelj, A. Mrvar, M. Zaveršnik, Network analysis of texts, Univ. of Ljubljana, Inst. of Mathematics, Physics and Mechanics, Dep. of Theoretical Computer Science, 2002.
- [50] J. Leskovec, D. Huttenlocher, J. Kleinberg, Signed networks in social media, in: *Proceedings of the International Conference on Human Factors in Computing Systems*, ACM, 2010, pp. 1361–1370.
- [51] 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.
- [52] M. Richardson, R. Agrawal, P. Domingos, Trust management for the semantic web, in: *Proceedings of the 2nd International Semantic Web Conference*, Springer, 2003, pp. 351–368.
- [53] 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.
- [54] S. Maniu, T. Abdesslem, B. Cautis, Casting a web of trust over wikipedia: an interaction-based approach, in: *Proceedings of the 20th International Conference Companion on World Wide Web*, ACM, 2011, pp. 87–88.
- [55] J. M. Kleinberg, Authoritative sources in a hyperlinked environment, *J. ACM* 46 (5) (1999) 604–632.
- [56] G. Palla, I. J. Farkas, P. Pollner, I. Derenyi, T. Vicsek, Directed network modules, *New J. Phys.* 9 (6) (2007) 186.

- [57] R. Albert, H. Jeong, A.-L. Barabási, Internet: Diameter of the world-wide web, *Nature* 401 (6749) (1999) 130–131.
- [58] X. Niu, X. Sun, H. Wang, S. Rong, G. Qi, Y. Yu, Zhishi. me-weaving chinese linking open data, in: *Proceedings of the 10th International Semantic Web Conference*, Springer, 2011, pp. 205–220.
- [59] A. Clauset, C. Moore, Accuracy and scaling phenomena in internet mapping, *Phys. Rev. Lett.* 94 (1) (2005) 018701.
- [60] F. Viger, A. Barrat, L. Dall’Asta, C.-H. Zhang, E. D. Kolaczyk, What is the real size of a sampled network? the case of the internet, *Phys. Rev. E* 75 (5) (2007) 056111.
- [61] M. Á. Serrano, M. Boguñá, A. Vespignani, Extracting the multiscale backbone of complex weighted networks, *P. Natl. Acad. Sci. USA* 106 (16) (2009) 6483–6488.
- [62] J.-P. Onnela, D. J. Fenn, S. Reid, M. A. Porter, P. J. Mucha, M. D. Fricker, N. S. Jones, Taxonomies of networks from community structure, *Phys. Rev. E* 86 (3) (2012) 036104.
- [63] L. K. Gallos, N. H. Fefferman, Revealing effective classifiers through network comparison, e-print arXiv:1403.2668.