Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Assessing the effectiveness of real-world network simplification

Q1 Neli Blagus *, Lovro Šubelj, Marko Bajec

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

HIGHLIGHTS

- We explore preservation of network properties under several simplification methods.
- The measure for assessing the effectiveness of simplification process is proposed.
- We compare the original and simplified networks based on global and local properties.
- The simplification on 10% of original networks provide for fair fit of properties.
- Random node selection based on degree and breadth-first sampling proved the best.

ARTICLE INFO

Article history:
Received 13 November 2013
Received in revised form 29 April 2014
Available online xxxx

Keywords:
Complex networks
Network simplification
Sampling
Merging
Simplification effectiveness

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.

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.
E-mail addresses: neli.blagus@fri.uni-lj.si (N. Blagus), lovro.subelj@fri.uni-lj.si (L. Šubelj), marko.bajec@fri.uni-lj.si (M. Bajec).

http://dx.doi.org/10.1016/j.physa.2014.06.065
0378-4371/© 2014 Published by Elsevier B.V.

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.

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