Large networks grow smaller: How to choose the right simplification method?

Neli Blagus, Lovro Šubelj, Gregor Weiss and Marko Bajec
University of Ljubljana, Faculty for computer in information science, Slovenia
{neli.blagus, lovro.subelj, gregor.weiss, marko.bajec}@fri.uni-lj.si

Network simplification proved as effective tool for reducing large real-world networks and at the same time providing for sufficient fit of original network [1, 2]. However, even though a number of analyses have been performed observing the changes of networks under the simplification, broad understanding of the whole process remains partial. The questions such as "How to compare original (i.e., complete) and simplified (i.e., incomplete) network?", "What factors impact the effectiveness of simplification process?", "What size of simplified networks provides for the best fit of original networks?", "What simplification method to use?" are far from solved in the literature.

In our study, we analyze over 30 real-world networks of different size and origin (e.g., social, information, technological) [3]. We reduce networks with several simplification methods (e.g., random node and link selection, breadth-first sampling, merging based on balance-propagation [4, 5]) and observe the changes of several fundamental properties (e.g., degree distribution, clustering coefficient, degree mixing (Fig. 1, above), and density [5]) under simplification. We show that the reduction on about 10% of original network provides for adequate preservation of important properties. The best performing methods prove to be random node selection based on degree and breadth-first sampling. The results also show the size of simplified network influence the effectiveness of simplification method, while the size and type of original network do not.

Besides basic properties, we explore also the changes of network structure under simplification. Particularly, we focus on different groups of nodes [6], commonly observed in real-world networks (e.g., communities, modules and mixtures of the two). In this case, the changes of simplification effectiveness occurs among different types of networks. For example, simplified social networks exhibit even stronger community structure than original networks, while in simplified information networks the number of mixtures increases (Fig. 1, below). However, in general, the proportion of nodes explained by the group structure enlarge in sampled networks, and the goodness of the preservation of node group structure does not depend on the choice of the simplification method.

To summarize, the main advantage of our analysis is large number of networks considered. Therefore we provide for reliable results concerning the effectiveness of simplification process and support a better understanding of the changes of networks under simplification process. In our future work we intend to create a framework for adaptive simplification of real-world networks, which would suggest the best simplification method for a given network based on its properties and the further use of simplified network.

References

[1] Lee, S. H., Kim, P. J., Jeong, H.: Statistical properties of sampled networks. Physical Review E. 73, (2006), 016102.
[2] Leskovec, J., Faloutsos, C.: Sampling from large graphs. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (2006), 631–636.
[3] Blagus, N., Šubelj, L., Bajec M.: Assessing the effectiveness of real-world networks simplification. Physica A, in review.
[4] Šubelj, L., Bajec, M.: Robust network community detection using balanced propagation. The European Physical Journal B. 81, (2011), 353–362.
[5] Blagus, N., Šubelj, L., Bajec, M.: Self-similar scaling of density in complex real-world networks. Physica A. 391, (2012), 2794–2802.
[6] Šubelj, L., Žitnik, S., Blagus, N., Bajec, M.: Node mixing and group structure of complex software networks. Advances in Complex Systems, in review.

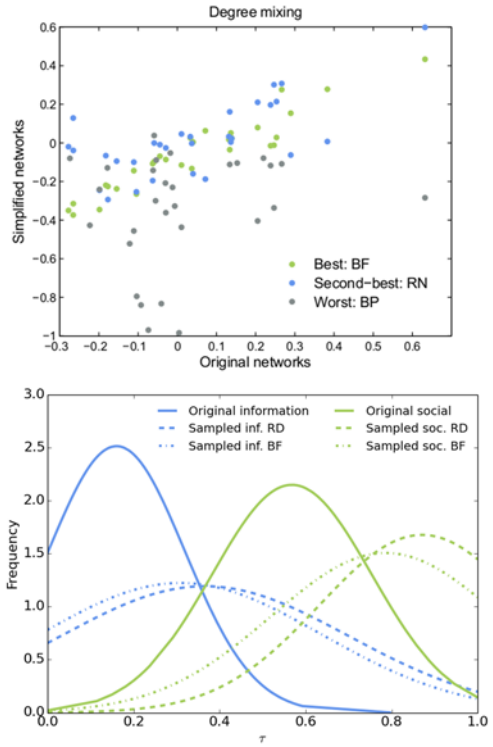

Figure 1: (above) The relationship between the degree mixing of the original and simplified networks. The breadth-first sampling proved as the best method. (below) The distribution of group parameter τ (i.e., τ=1 denotes communities, τ=0 modules and 0< τ<1 mixtures). Sampled networks are characterized by larger number of communities and mixtures than original networks in social and information networks, respectively.