University of Ljubljana
Faculty of Computer and Information Science

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

Neli Blagus*, Lovro Šubelj, Gregor Weiss, Marko Bajec
University of Ljubljana, Faculty of Computer and Information Science
*neli.blagus@fri.uni-lj.si

Systems described by networks can be large and fast evolving. Thus, not only their storage poses a problem but their analysis and understanding presents a great challenge. A natural solution to these problems is provided by the network simplification. Given a large network, the goal of the simplification is to create a smaller sampled network that accurately matches the properties of the original network. In recent years, several simplification methods have been proposed. In general, the methods based on crawling (e.g., random walk, forest fire sampling) outperform the methods based on node or link selection (e.g., random node and link sampling).

What is your goal?

- Estimate particular property
  - The network is hard to explore completely or network data is incomplete and biased
  - How large should the simplified network be?
    - Larger simplified networks provide for better fit of properties of the original networks
    - Large (~50%) → Node sampling
    - Small (10 - 20%) → Sampling via crawling
- Create representative subgraph
  - To evaluate the performance of algorithms or study the processes on networks
  - Do you see the whole network?
    - Original network may be large, decentralized or fast evolving
    - No → Sampling via crawling
    - Yes → Link sampling

Node sampling

In **random node selection** (RN) the sample consists of uniformly at random selected nodes and links among them. RN does not preserve power-law degree distribution [1] and underestimate betweenness centrality [2], while matches the **degree mixing** [2], **transitivity** and **density** (Fig. 1) [3].

In **random node selection based on degree** (RD), we sample nodes proportional to their degree. RD preserves better the **spectral properties** and **out-degree** [3, 4]. RD changes the node group structure (e.g., communities, mixtures, modules [5]). In sampled information networks the number of **mixtures increases**, while the sampled social networks are characterized by **stronger community structure** [6].

Under RD, **hubs are more likely to be selected** and **significant portions of the original network is covered**. RD is sufficient for **detecting landmark nodes** and **outbreaks** [7].

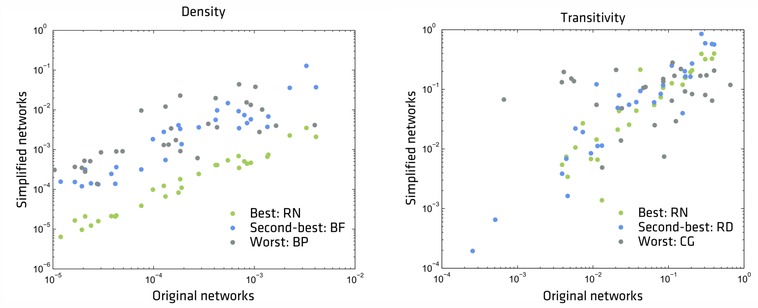

Figure 1: Relationship between (left) the density and (right) the transitivity of the original and the simplified networks [3].

Sampling via crawling

The common idea is to create a simplified network with exploration of neighborhood of a randomly picked start node. In **breadth-first sampling** (BF) nodes are selected using a breadth-first search strategy. BF is commonly used for crawling online networks and preserves the **degree, betweenness centrality distribution** and **degree mixing** (Fig. 3) [3].

In **random walk sampling** (RW) the sampled network is created with simulation of a random walker. RW shows best performance among several sampling methods, at preserving the **clustering coefficient** and for creating a **representative subgraph** [4].

Forest fire sampling (FF) [4] is a combination of BF and RW, where only a fraction of neighbors is selected at each step. FF proves well at preserving the **spectral properties** and in general shows similar performance as RW [4], still the simplified networks contain a high fraction of low clustered nodes [8] and thus fail to preserve transitivity.

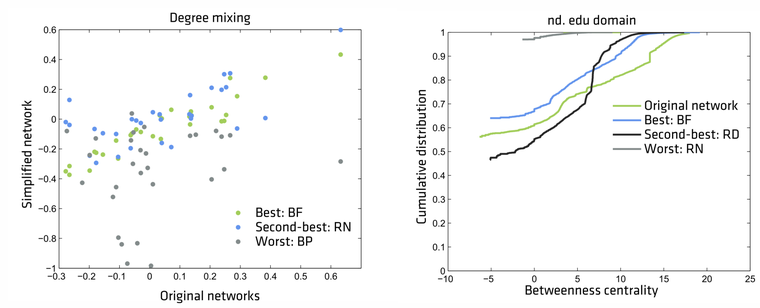

Figure 3: (left) The relationship between the degree mixing of the original and the simplified networks and (right) example of comparison of the betweenness centrality for the original and simplified networks [3].

Link sampling

In **random link selection** (RL) links are sampled uniformly at random. A simplified network can be created in two ways. In first, the sample contains selected links only and thus fits the **distribution of sizes of weakly connected components** [3], **degree mixing** [2] and **path length** of the original network. RL overestimate degree and betweenness centrality exponent [2] and does not preserve connectivity, degree and clustering distribution (Fig. 2) [4].

In second, **totally-induced sampling** (TIES), the simplified network consists of all the links among sampled nodes. Thus, the sample matches the **distribution of degree**, **path length** and **clustering coefficient** more accurately. TIES proves very well on small simplified networks [8].

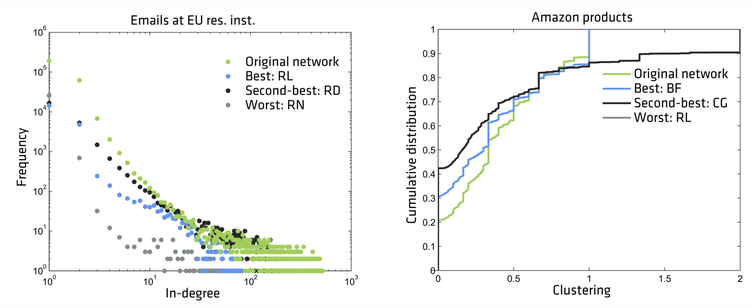

Figure 2: Example of comparison of (above) the in-degree and (below) the clustering coefficient distribution for the original and simplified networks [3].

Summary

The selection of the right simplification method depends on **the further use of the simplified network**. However, sampling based on crawling outperform node and link sampling - **crawling methods create connected samples** that match accurately **larger set of properties**.
Network characteristics are better preserved if the simplified network is larger, still **10-20% of the original network size** is enough for accurately matching the properties, especially crawling methods and link sampling with induction perform well for smaller samples. Node and link sampling work well for **estimating one particular network property** and when the simplified network is larger (~50% of the original network size).

References:
[1] M. P. H. Stumpf, C. Wiuf, R. M. May: "Subnets of scale-free networks are not scale-free: sampling properties of networks". In PNAS, 102 (2005).
[2] S. H. Lee, P. J. Kim, H. Jeong: "Statistical properties of sampled networks". In Phys. Rev. E, 73 (2006).
[3] N. Blagus, L. Šubelj, M.Bajec: "Assessing the effectiveness of real-world network simplification", Submitted to Physica A (2014).
[4] J. Leskovec, C. Faloutsos: "Sampling from large graphs". In ACM SIGKDD (2006).
[5] L. Šubelj, N. Blagus, M. Bajec: "Group extraction for real-world networks: The case of communities, modules, and hubs and spokes". In NetSci (2013).
[6] N. Blagus, G. Weiss, L. Šubelj: "Sampling node group structure of social and information networks". On arXiv:1405.3093 (2014).
[7] A. S. Maiya, T. Y. Berger-Wolf: "Benefits of bias: towards better characterization of network sampling". In ACM SIGKDD (2011).
[8] N. K. Ahmed, J. Neville, R. Kompella: "Network sampling: from static to streaming graphs". On arXiv:1211.3412 (2012).