Can I trust my network? Assessing network estimation uncertainty using local component resolution

Alexandra Jauhiainen[1] **and José Sánchez**[2] **and Rebecka Jörnsten**[2]
[1] Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
[2] Mathematical Statistics, Mathematical Sciences, Chalmers University of Technology and the University of Gothenburg, Göteborg, Sweden

Several network estimation methods have been presented and published in recent years. The methods differ in the type of network estimated (directed, undirected) and the assumptions used. Common to all these methods is that the final solution provided is one estimated network, ignoring estimation uncertainty. This leads to substantial risk of over-interpreting generated networks.

In fact, it is non-trivial to exhaustively assess uncertainty in network estimation. Bootstrap is sometimes used, in which resampling based marginal statistics on whether an edge is present or absent are recorded. Thresholding the individual links, keeping e.g. links which appear in 90% of the bootstrap networks, will give a graph that isn't a member of the set of bootstrap graphs and often is not similar to any of the members. The structure of the thresholded graph is hence not supported by the empirical data.

We propose a paradigm that uses advances in data compression and graph matching to address the uncertainty problem. We exploit resampling bootstrap graphs, not to directly generate the joint high-dimensional edge distribution in graph space (which is complex), but rather to identify a bounded set of bootstrap graphs that capture the large proportion of variation over graphs.

By dividing up the network in tightly interconnected parts, which are referred to as modules, we can produce a set of graphs for each part of the network, depending on the variability of link presence/absence supported by bootstrapping, see illustration in Figure 1. When only one representative graph is generated, the uncertainty is small and the module can be regarded as having good resolution. When several representative graphs are produced for a module, the estimation is more uncertain and the module is of low resolution. Since we use one or several medoid graphs as candidates for a module we can be sure that the structure is supported by the data (unlike for a thresholding graph).

The selection of candidate graphs is based on embedding the bootstrap graphs in a vector space using a graph distance metric. The embedding enables clustering of graphs and assessment of the compression when selecting candidate networks. The final selection of candidate graphs is made using a rate-distortion criterion. Rate-distortion is an information theoretic concept which can be adjusted to function as a model selection tool in clustering and significance analysis (Jörnsten, 2009).

The local number of candidate networks thus captures the different information level that the data provides regarding the local structure, which we refer to as network component resolution (NetCoR). The paradigm offers a way to fairly compare different network estimation methods. That is, the identification of high-resolution network modules can be interpreted as a characteristic of stable and good estimation methods. The paradigm also indicates which nodes in the network to focus on when performing additional experiments in order to achieve better resolution and confidence in the estimated graph. We apply the method to the analysis of cancer genome networks and compare high- and low-resolution modules in terms of functional annotation.

We have developed an R package which implements the NetCoR paradigm from start to end. Several different estimation methods are included in the package, but it is also possible to use with a custom estimation method of choice as the paradigm in itself is estimation method independent.

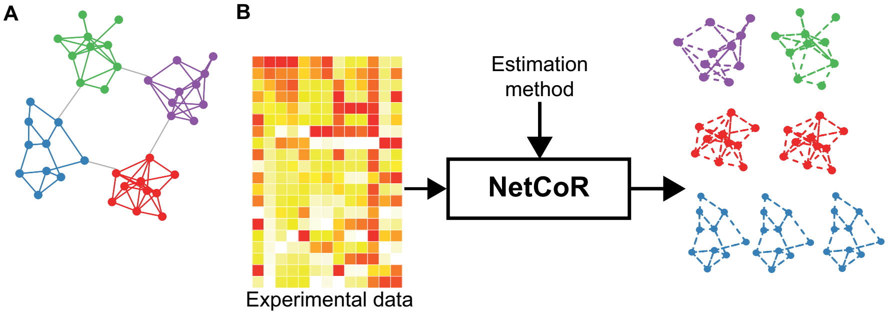

Figure 1: Division of a network into interesting and/or tightly bound modules (A). The NetCoR paradigm produces a number of candidate graphs for each part of the network, indicating whether or not the different parts have high or low resolution (B).

Keywords: Network estimation, Rate-distortion, Local resolution, Compression, Estimation uncertainty

References:

Jörnsten, R. (2009). Simultaneous model selection via rate-distortion theory, with applications to cluster and significance analysis of gene expression data. *Journal of Computational and Graphical Statistics 18*, 613–639.