An Empirical Investigation of Statistical Backbone Filtering Techniques

Ali Yassin[1], Hocine Cherifi[1], Hamida Seba[2], and Olivier Togni[1]
[1]Laboratoire d'Informatique de Bourgogne - Univ. Bourgogne - Franche-Comté, Dijon, France
[2]Univ Lyon, UCBL, CNRS, INSA Lyon, LIRIS, UMR5205, F-69622 Villeurbanne, France

Networks provide an ideal representation for complex systems, yet the density and size of such systems make them challenging for processing and visualization. Backbone extraction techniques address this problem by reducing the network size while preserving the highest amount of "information." Edge Filtering techniques focus on removing edges. They can be classified into "structural" and "statistical" methods. This study [1] analyzes the performance of seven statistical edge filtering techniques in the world air transportation network. Ideally, a good filtering technique should be able to preserve the highest amount of information while filtering as many connections as possible and avoiding the breakup of the system. To evaluate the efficiency of a filtering technique, we consider four indicators: the percentage of edges, weights, isolated nodes, and the size of the largest connected component. First, we compare the seven filtering techniques for a fixed significance level ($\alpha = 0.05$) on a toy network. Second, we experiment for different significance levels to validate the first experiment's results and analyze the evolution of the four indicators in real-world networks. Fig 1 illustrates typical results for the world air transportation network.

In the top left panel, one can see that the Marginal Likelihood Filter retains all the edges even within the strong filtering regime corresponding to $\alpha < 10^{-2}$. The Noise Corrected Filter has similar behavior across different significance levels. Indeed, it preserves 80% of the edges. Both filters model an edge with a binomial distribution. However, they use different definitions of the probability of success. The remaining filters exhibit similar behavior. Indeed, the percentage of edges increases monotonically with the significance levels. However, in a strong regime $\alpha < 10^{-2}$ the Polya Urn Filter, Disparity Filter, and the GLOSS Filter hardly keep any edge. In contrast, the LANS Filter preserves a constant percentage of edges (17%). In comparison, the fraction of edges in the ECM Filter keeps increasing till it reaches 35% for $\alpha = 10^{-2}$. For $\alpha > 10^{-2}$, the percentage of edges increases exponentially except for the ECM Filter, which grows gradually.

Except for the Disparity Filter, the established hierarchy also holds in the top right panel. In the strong regime $\alpha < 10^{-2}$, the Disparity Filter preserves the same percentage of weights as the ECM Filter, although it holds a tiny portion of edges. Instead, the ECM Filter keeps a more significant amount of edges, reaching 40%. Indeed, the Disparity Filter prioritizes high weights, and the ECM Filter preserves small and high weighted edges. For $\alpha > 10^{-2}$, the Disparity Filter retains the same percentage of weights as the LANS Filter, validating the fact that it prioritizes high weights.

The bottom left panel illustrates the fraction of isolated nodes. The Marginal Likelihood Filter and Noise Corrected Filter do not separate any node from the network because they hardly filter any edges. In contrast, except for the ECM Filter, all the other methods isolate a significant portion of nodes in the strong regime. Indeed, the percentage of isolated nodes decays until there are no more isolated nodes as we reach $\alpha = 10^{-2}$. After that, the percentage of isolated nodes decreases with the increase of the significance level, illustrating how they fail to preserve all the nodes while filtering edges.

Unlike the Marginal Likelihood Filter and Noise Corrected Filter, other filters do not retain one giant component, as shown in the last panel. The LANS Filter and GLOSS Filter maintain a giant component with a fixed size in the strong regime. Its size increases gradually to form a unique component only when adding all the edges. In contrast, we notice the emergence of a giant component in the Disparity Filter, Polya Urn Filter, and the ECM Filter as we approach the boundaries of the strong regime. After that, the ECM Filter stops isolating nodes while the others stop only when maintaining all the edges.

To summarize, choosing a good null model is essential. We show how filters based on a binomial distribution generated almost complete networks. In addition, other filters besides the ECM Filter are aggressive in removing edges for reasonable significance levels $10^{-2} \leq \alpha \leq 0.05$, which leads to isolated nodes. At the same time, the ECM Filter lies between these two extreme behaviors. Future work will consider the distribution of the p-values and the topological properties of the backbone's extractors.

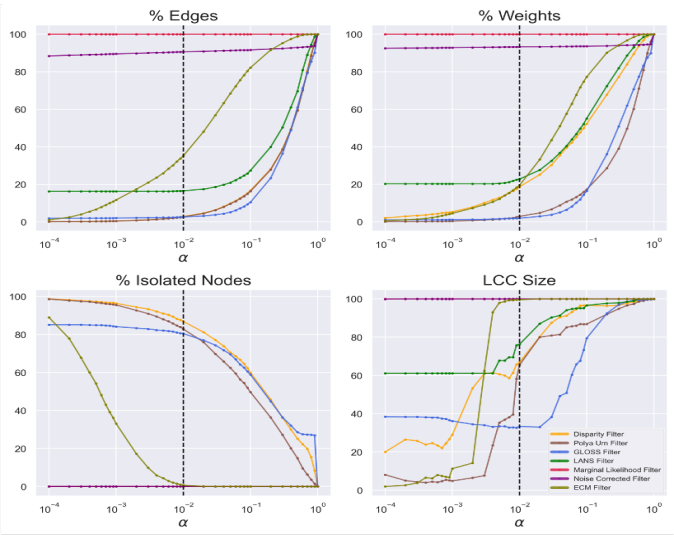

Fig. 1. The percentage of Edges, Weights, Isolated Nodes, and the Largest Connected Component Size in the extracted backbone using different filtering techniques in the world air transportation network as a function of significance level α.

[1]Agence Nationale de Recherche funds this work under grant ANR-20-CE23-0002.