Filtering Real World Networks: A Correlation Analysis of Statistical Backbone Techniques

Ali Yassin, Hocine Cherifi, Hamida Seba, Olivier Togni

Noname manuscript No.
(will be inserted by the editor)

Filtering Real World Networks: A Correlation Analysis of Statistical Backbone Techniques

Ali Yassin · Hocine Cherifi · Hamida Seba · Olivier Togni

Received: date / Accepted: date

Abstract Networks are an invaluable tool for representing and understanding complex systems. They offer a wide range of applications, including identifying crucial nodes, uncovering communities, and exploring network formation. However, when dealing with large networks, the computational challenge can be overwhelming. Fortunately, researchers have developed several techniques to address this issue by reducing network size while preserving its fundamental properties [1–9] . To achieve this goal, two main approaches have emerged: structural and statistical methods. Structural methods aim to keep a set of topological features of the network while reducing its size. In contrast, statistical methods eliminate noise by filtering out nodes or links that could obscure the network's structure, utilizing advanced statistical models.

In a previous work [10] we compared a set of seven statistical backbone filtering techniques in the World Air Transportation network. Results show that the Marginal Likelihood Filter, Disparity Filter, and LANS Filter give more importance to high-weight edges. The other techniques emphasize both small and high-weighted edges.

This study extends the previous research on seven statistical filtering techniques, namely Disparity, Polya Urn, Noise Corrected, Marginal Likelihood, LANS, ECM, and GloSS filters, through the analysis of 39 real-world networks

This material is based upon work supported by the Agence Nationale de Recherche under grant ANR-20-CE23-0002.

A. Yassin
Laboratoire d'Informatique de Bourgogne, University of Burgundy, Dijon, France
E-mail: ali_yassin@etu.u-bourgogne.fr

H. Cherifi
ICB UMR 6303 CNRS - Univ. Bourgogne - Franche-Comté, Dijon, France

H. Seba
Univ Lyon, UCBL, CNRS, INSA Lyon, LIRIS, UMR5205, F-69622 Villeurbanne, France

O. Togni
Laboratoire d'Informatique de Bourgogne, University of Burgundy, Dijon, France

of diverse origins. These networks range in size from 18 to 13,000 nodes and include character, web, biological, economic, infrastructural, and offline/online social networks. In the first experiment, we aim to evaluate and compare the similarities between the seven statistical filtering techniques. Each method assigns a probability value, called a p-value, to each edge. To compare the methods, we use these p-values to conduct correlation analysis. Specifically, we compute the Pearson correlation between each pair of techniques' p-value edges. However, it is important to note that Pearson correlation examines linear relationships, whereas Jaccard similarity compares the similarities of two sets. Therefore, we use Jaccard similarity to compare the fraction of shared edges in each backbone. In a second experiment, we investigate the relationship between edge significance and edge properties. To do this, we compute the Pearson correlation between the p-values and edge properties, including weight, edge degree, and edge betweenness. Fig 1 illustrates these results.

The heatmaps present the mean and standard deviation of Pearson correlation between filtering technique pairs across all networks. The couples (LANS, Disparity filter) and (Noise Corrected, ECM) are well correlated (0.8). Conversely, the Polya Urn filter does not exhibit a noticeable correlation with any other filtering method. The standard deviation heat map shows a low standard deviation validating these findings.

The middle graphs illustrate the typical behaviors of how the mean Jaccard score changes as a function of the top fraction of edges sorted by various backbone filtering techniques. The top left panel shows a low Jaccard score between the Polya Urn filter and the Noise Corrected filter. The other techniques also have a low Jaccard score between the Polya Urn filter. The top right panel shows that the GloSS filter shares at least 20% of its edges with the Marginal Likelihood filter. The other techniques have the same behavior as the Marginal Likelihood filter with the GloSS filter except for the Polya Urn filter. The bottom right panel shows that the set of edges obtained by the Disparity filter shares on average at least 50% of its edges with the LANS filter. The ECM Filter and Marginal Likelihood Filter (ECM-MLF) and ECM Filter and Noise Corrected Filter (ECM-NC) behave similarly. Finally, in the bottom left panel the set of edges obtained by the Marginal Likelihood filter shares on average at least 70% of its edges with the Noise Corrected filter. On the other hand, the couples DF-NC, DF-ECM, DF-MLF, LANS-ECM, LANS-NC, and LANS-MLF behave the same, sharing at least around 30% of the edges. However, they have a high standard deviation.

The boxplots illustrate the Pearson correlation coefficient between edge p-values and edge weights, degrees, and betweenness across all networks. Results indicate a greater demonstration of the distinct behavior of the Polya Urn filter. The edge p-values were found to be uncorrelated with edge weights, degree, and betweenness, with a very low standard deviation. In contrast, the top panel shows that the edge p-values obtained through the Disparity filter and Marginal Likelihood filter were correlated with weights, with average correlation higher than 0.6. This indicates that these techniques prioritize edges with high weights. In the middle panel, the Noise Corrected filter and

ECM filter have an average correlation higher than 0.6. This means that these methods give importance to edges that connect hubs, as these edges have a high edge degree, which is used indirectly by these methods to determine edge significance. Finally, the bottom panel shows that the edge p-values from all techniques have no correlation with edge betweenness, indicating that none of the methods prioritize edges that play a significant role in communication between nodes through the shortest paths.

In conclusion, correlation analysis is crucial in highlighting similarities and differences between backbone edge filtering techniques, identifying areas for improvement, and advancing knowledge in this field. This study can help to identify areas where improvements can be made in the development of new techniques or in the refinement of existing ones.

Keywords Complex Networks · Backbone Filtering Techniques · Network Compression · Graph Summarization · Sparsification

References

1. C.H. Gomes Ferreira, F. Murai, A.P. Silva, M. Trevisan, L. Vassio, I. Drago, M. Mellia, J.M. Almeida, Plos one **17**(9), e0274218 (2022)

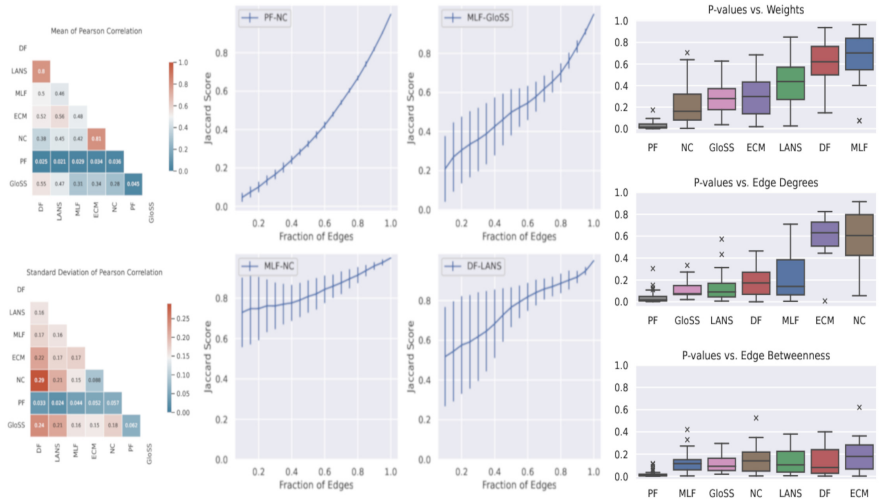

Fig. 1 At the left, the mean and standard deviation heatmap of Pearson Correlation between pairs of filtering techniques p-values across all networks. In the middle the typical behaviors of the mean Jaccard score between the set of edges of pairs of different filtering techniques as a function of fraction of edges preserved across all networks. In the right, the boxplots of Pearson correlation coefficient between edge p-values and edge weights, degrees, and betweenness across all networks. The MLF is the Marginal Likelihood Filter, DF is the Disparity Filter, LANS is the Local Adaptive Network Sparsification, PF is the Polya Urn Filter, NC is the Noise Corrected Filter, and GloSS is Global Statistical Significance Filter. Note that we took the absolute value of the Pearson correlation.

2. Z. Ghalmane, C. Cherifi, H. Cherifi, M. El Hassouni, Scientific Reports **10**(1), 1 (2020)
3. Z. Ghalmane, C. Cherifi, H. Cherifi, M. El Hassouni, in *9th International Conference on Complex Networks and Their Applications* (2020), pp. p–3
4. V. Gemmetto, A. Cardillo, D. Garlaschelli, arXiv preprint arXiv:1706.00230 (2017)
5. Z. Ghalmane, C. Cherifi, H. Cherifi, M. El Hassouni, Information Sciences **576**, 454 (2021)
6. S. Rajeh, M. Savonnet, E. Leclercq, H. Cherifi, in *Network Science: 7th International Winter Conference, NetSci-X 2022, Porto, Portugal, February 8–11, 2022, Proceedings* (Springer International Publishing Cham, 2022), pp. 67–79
7. A. Yassin, H. Cherifi, H. Seba, O. Togni, in *2022 IEEE Workshop on Complexity in Engineering (COMPENG)* (IEEE, 2022), pp. 1–8
8. A. Yassin, H. Cherifi, H. Seba, O. Togni, in *Complex Networks and Their Applications XI: Proceedings of The Eleventh International Conference on Complex Networks and their Applications: COMPLEX NETWORKS 2022—Volume 2* (Springer International Publishing Cham, 2023), pp. 551–564
9. L. Dai, B. Derudder, X. Liu, Journal of Transport Geography **69**, 271 (2018)
10. A. Yassin, H. Cherifi, H. Seba, O. Togni. Air transport network: A comparison of statistical backbone filtering techniques (2023). DOI 10.1007/978-3-031-21131-$7_4$3