Spanning Edge Betweenness in Practice

Andreia Sofia Teixeira, Francisco C Santos, and Alexandre P Francisco

Abstract In this paper we present a study about spanning edge betweenness, an edge-based metric for complex network analysis that is defined as the probability of an edge being part of a minimum spanning tree. This probability reflects how redundant an edge is in what concerns the connectivity of a given network and, hence, its value gives information about the network topology. We apply this metric to distinct empirical networks and random graph models, showing that spanning edge betweenness allows us to identify those edges that are more relevant for connectivity and how removing them leads to disruption in network structure.

Key words: spanning edge betweenness, network analysis, edge centrality measures

1 Introduction

Networks are the simplest representation of interactions and relations between entities. Nevertheless, a network can express very complex processes and behaviours. In this context, understanding structure and dynamics of a network is crucial to extract valuable information. The analysis of complex networks, such as social networks, biological networks, financial networks, electrical networks or even the world wide web, have gathered efforts from mathematicians, physicists, social and computer scientists to build several statistical measures and tools to evaluate the importance of each node and/or each link. Some are very well-known [4, 5, 10]: degree centrality indicates the fraction of connections that a given node has over the entire network; node/edge betweenness states how important a node/edge is through the number of shortest paths between two nodes passing through it; and clustering coefficient is a key measure for social network analysis that for a given node expresses

All authors are with INESC-ID / Instituto Superior Técnico, Universidade de Lisboa, e-mail: {sofia.teixeira,franciscocsantos,aplf}@tecnico.ulisboa.pt

how many of its neighbours are neighbours of each other, evaluating the fraction of possible triangles that the node is a member of. All of these measures can give us information about centrality and connectivity of a network, but they are mostly focused on nodes. On the other hand, although we can evaluate the centrality of an edge by using betweenness centrality, there are many networks whose study can gain new insights if new measures are used for evaluating edge centrality that do not depended on shortest paths, as edge betweenness does. When we address phylogeny, telecommunication/electric networks, among other networks, we are often interested in studying measures that go beyond shortest path properties. If we want to know how resilient a network is, i.e., which links are fundamental to keep the network connected and which are redundant, none of the metrics described before provides that information. In algorithms for inferring phylogenies, we aim to validate the trees that are generated to represent evolution patterns and to identify bridges that connect different groups, in telecommunication/electric networks we are interested to know which links are so important that could cause a breakdown if turned off, or which of them are redundant. Recently, Morone [12] presented a work in which one of the goals is to find the minimal set of nodes that, if removed, would break down the network, but once again, the work is focused on the importance of the nodes and not on the importance of the links.

These problems can be conveniently studied by relying on minimum spanning trees. Recently, a new network measure was proposed for evaluating the importance of edges taking into account information provided by minimum spanning trees – *spanning edge betweenness* [6]. This new metric, which corresponds to the fraction of minimum spanning trees that contains an edge, has the potential to not only help on the evaluation and validation of phylogeny algorithms, for which it was originally proposed, but also to evaluate how redundant an edge is in a given network. Because of its probabilistic property, spanning edge betweenness provides direct information about an edge preventing the relativity inherent to the other measures. Contrary to what is evaluated in edge betweenness, we are not interested in knowing in how many shortest paths the edge is present, but how important the edge is to maintain the network connected. Given an edge, its spanning edge betweenness value can reflect whether the edge removal can cause a disruption in a network or if there are some alternative ways to keep the network connected, reflecting how resilient the network can be and how redundant an edge is in the network. More recently spanning edge betweenness has been object of further studies. An initial study on the importance of the metric in phylogenetic trees was reported in [7]. An improvement in what concerns the efficient computation of spanning edge betweenness was presented in [8]. And Qi *et al.* [9] introduced the concept of spanning tree centrality, that applies the same principles of spanning edge betweenness although applied to the nodes in a weighted network.

In this paper we study the applicability of spanning edge betweenness for evaluating edge redundancy on real and synthetic networks. For this aim, we compare it with previously introduced measures and evaluate how turning on/off the links with highest spanning edge betweenness can affect networks topologies and how can we identify potential bridges that are crucial to ensure networks integrity and connectivity. We use real and artificial networks and for each one we remove all the edges with three criteria: random selection, decreasing order of spanning edge betweenness values, and decreasing order of edge betweenness values. We show that removing edges with high spanning edge betweenness leads to a fast disruption in the networks, rapidly increasing the number of components of the networks.

2 Edge-based Measures on Minimum Spanning Trees

Minimum spanning trees have been used for decades for network design, cluster analysis, among others. Given a network, a minimum spanning tree represents the set of edges with minimum weight that connect all of the nodes. Let $G = (V, E)$ be a connected, undirected and weighted graph, with weight function $w : E \rightarrow \mathbb{R}$, where V is the set of vertices and $E \subset V \times V$ is the set of edges. A minimum spanning tree $T = (V, E')$ is a subgraph of G that is a tree and contains all the vertices of G, i.e., that spans over all vertices in V, with $|E'| = |V| - 1$, and such that $\sum_{e \in E'} w(e)$ is minimum among all spanning trees. For generality, we can assume an unweighed graph as a graph with all edges' weights equal to 1. If the network is a tree, then there is only one minimum spanning tree, otherwise the network can have multiple minimum spanning trees.

When constructing certain networks – such as electrical, computer, transportation, and telecommunication networks – the major concern is to choose the cheaper path for laying the connections. On the other hand, if we already have a network, how can we know which are the links whose presence is imperative to connect all the nodes and which provide a more flexible choice? On other perspective: given a computer network, which connections should we choose to assure its resilience preventing a massive disruption? Which connections/edges are critical? The study of spanning edge betweenness on a network allows us to give some answers for these questions.

The first known edge-based centrality, edge betweenness, was initially proposed by mathematician Anthonisse and later formalized and published by Freeman in 1977 [1]. It was developed in the context of communication networks. For a given edge e it measures how central the edge is, i.e., how many geodesic paths transverse that edge. In 2002, Girvan and Newman [2] applied this metric to the study of finding and evaluating community structures in networks, but little has been done in what concerns exploring new edge importance measures in a network. In 2012, Meo*et al.* [3] developed a k-path centrality, initially developed for nodes, which is based on random walks and is defined as the sum of the frequency with which a message traverses an edge e from a given source to all k-edges-distance possible destinations. These two centrality measures play a central role in reporting knowledge about data flow in a network but few about the structure/topology of the network.

In fact, when analysing minimum spanning trees, shortest-paths or random walks approaches yield insufficient information to infer how much resilient a network is or how redundant are some connections, depending on the subject in study. Recently, Teixeira *t al.* [6] introduced an edge-based centrality measure that relies on minimum spanning trees to evaluate how important is an edge in the structure of a network. Here, we extend the evaluation made on phylogenetic trees [7], providing information about the metric behaviour in real well-known networks, including social, technological and electric networks. The fact that it tells directly the probability of an edge being in a minimum spanning tree, thus reflecting how important it is for the network structure, ensures a high confidence in the analysis of network resilience and edge redundancy.

2.1 Spanning Edge Betweenness

For a given edge e, the *spanning edge betweenness* is defined as:

$$\delta_G(e) = \frac{\tau_G(e)}{\tau_G},$$

where τ_G is the number of different minimum spanning trees for G and $\tau_G(e)$ is the number of different minimum spanning trees for G where e occurs.

There are many applications for this new measure, as exemplified by Teixeira *et al.* [7] in the context of inferring phylogenies. As we said before, a network can have many minimum spanning trees. Spanning edge betweenness comes to help in the confidence evaluation of the tree generated. Because this metric takes values between [0,1] we can infer: 1) if spanning edge betweenness is 1 than the edge has to be on the network to keep it connected; 2) if it is 0, which only can occur in weighted networks, than the edge is completely redundant; 3) being the value between 0 and 1 it means that there are other edges that can keep the network connected, i.e., there is a different minimum spanning tree for the network, thus expressing the redundancy of an edge. As we will see, the proportion of these values can provide information about the network topology.

3 Methods and Results

To evaluate the significance of the spanning edge betweenness we chose eight different networks, with different sizes and from different contexts. Four are real well-known networks (Karate, Power Grid, Political Blogs and NetScience)[1], and four are random networks: two generated from Barabási-Albert model [11] and two networks with community structure[2]. The properties of these networks are in tables 1, 2 and 3. In practice, we compute five measures: node degree centrality, node betweenness, edge betweenness, cluster coefficient and spanning edge betweenness. Than we correlate spanning edge betweenness with the other metrics. Spanning edge be-

[1] http://www-personal.umich.edu/˜mejn/netdata/

[2] https://sites.google.com/site/santofortunato/inthepress2

Table 1 Detail for real networks.

Network	# Nodes	# Edges
Karate	34	78
PowerGrid	4941	6594
Polblogs	1490	2742
NetScience	1589	1252

Table 2 Barabási-Albert model parameters for generating random networks.

# Nodes	# Edges	Average Degree
1000	2975	4
1000	4939	4

Table 3 Model paraemters for generating random networks with community strcuture.

# Nodes	# Edges	Min / Max Degree	Min / Max Community Size
1000	2222	4 / 8	20 / 40
1000	3985	8 / 16	20 / 40

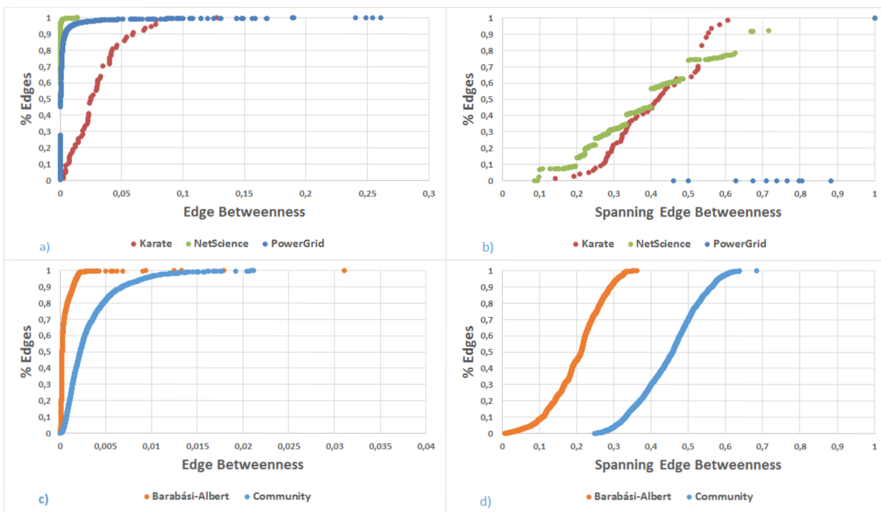

Fig. 1 Edge Betweenness Vs Spanning Edge Betweenness. In panels a) and c) we show the values of edge betweenness for three empirical networks and two random generated networks. In panels b) and d) we show the values of spanning edge betweenness for the same networks. While spanning edge betweenness shows a wide range of values, expressing edge significance in network structure, edge betweenness is limited to a very small set of values not being possible to infer directly information about network structure.

tweenness and edge betweenness were directly correlated; for the other node-based metrics – node betweenness, degree centrality and cluster coefficient – we correlated with the minimum/maximum/average metrics between the source and destination nodes of each edge.

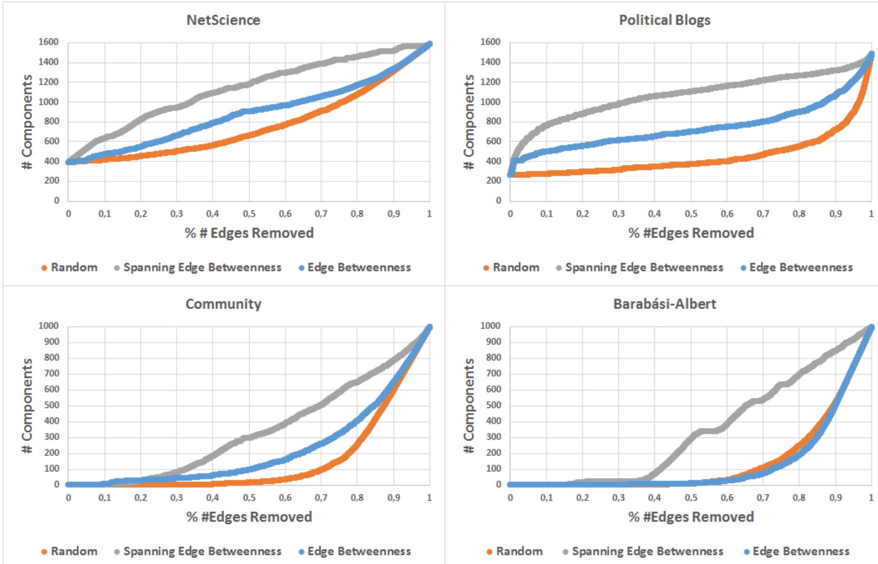

Fig. 2 Analysis of Removing Edges: randomly, in decreasing order of spanning edge betweenness and edge betweenness values in NetScience, PoliticalBlogs, Barabási-Albert and Community networks. It is possible to see that for all networks, empirical and random generated networks, removing edges in decreasing order of spanning edge betweenness leads to a earlier break down into more components of each network when comparing with the other two methods.

The first conclusion is that spanning edge betweenness has no correlation with the other measures. When we correlated it with the other measures mentioned, none of them showed meaningful correlation values. This reinforces the idea that this measure provides novel information that was not given before. In Figure 1, we show that spanning edge betweenness has a different expression than edge betweenness. While spanning edge betweenness took values between 0 and 1, expressing directly the importance of an edge, edge betweenness took all of its values below 0.3. Comparing directly both measures, it is possible to see that the values of edge betweenness do not allow to infer clear information about network structure. Edge betweenness is about how much information flow passes through an edge in shortest paths, while spanning edge betweenness is about the significance of and edge, potentially identifying edges that can break the network and reflecting if the network has a strong or weak redundancy. We can also see that PowerGrid has a very different behaviour from other three chosen networks. This is because the topology of the network is like a tree, or a star, with only ten redundant edges, being one example that if a link is disconnected, most probably the network will break. The other networks illustrate the redundancy that is expected from that kind of networks. As friends are friends from each other, as one cites another, there can be much alternatives to maintain the network connected and reachable between all nodes.

To reinforce the idea that spanning edge betweenness provides information about the redundancy and the connectivity of a network, we also present an evaluation on how removing edges from a network affects network structure. For all networks mentioned before, after we calculated the values of each measure, we sorted them by decreasing order and then, one by one, we removed each edge from the networks, registering the number of connected components after each removal. The result was as expected, removing by decreasing order of spanning edge betweenness speeds up the disruption of the networks when comparing with decreasing order of edge betweenness. In Figure 2, we show four examples – two from real networks and two from generated networks – but for all networks the results were similar on what concerns the number of connected components growth. For the same proportion of edges removed, removing edges with decreasing order of spanning edge betweenness breaks the network structure into more components than with decreasing order of edge betweenness.

4 Final Remarks

Centrality measures are important in a large number of graph applications, from search and ranking to social and biological network analysis. Most of these measures are calculated upon the nodes/vertices, but sometimes our interest is to study the importance of links/edges on a network. Spanning edge betweenness is a useful measure that can be applied both in weighted and unweighed graphs, allowing different types of evaluations – from confidence in phylogenetic trees to the identification of edges that are critical to keep the network connected, passing through the ones that express redundancy and alternative network configurations. In this paper we compared it with another measures, namely with traditional edge betweenness, and on several real and synthetic networks, concluding that spanning edge betweenness performs better at identifying the relevance of edges for maintaining networks connectivity. Since spanning edge betweenness gives direct information about the importance of a link, on further research we plan to investigate other application fields as epidemic spreading, identifying which links are critical in the spreading process, following some of the ideas introduced in [13].

Acknowledgments

This work was partly supported by national funds through FCT – Fundação para a Ciência e Tecnologia, under projects Incentivo/EEI/LA0021/2014, EXCL/EEI-ESS/0257/2012, UID/CEC/50021/2013, PTDC/EEI-SII/5081/2014, and PTDC/MAT/STA/3358/2014.

References

1. L.C. Freeman, A set of measures of centrality based on betweenness. *Sociometry*, 40, 35-41, 1977.
2. M. Girvan and M.E. Newman, Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA*, 99, 7821-7826, 2002.
3. P. De Meo, E. Ferrara, G. Fiumara, A. Ricciardello, A novel measure of edge centrality in social networks. *Know.-Based Syst.* 30, 136-150, 2012.
4. S. P. Borgatti and M. G. Everet, A Graph-theoretic perspective on centrality *Social Networks*, 28.4, 466-484, 2006.
5. L. A. Costa, F. A. Rodrigues, G. Travieso, P. R. Villas Boas, Characterization of complex networks: A survey of measurements. Advances in Physics, 56, 167-242, 2007.
6. A. S. Teixeira, P. T. Monteiro, J. A. Carriço, M. Ramirez, A. P. Francisco, Spanning Edge Betweenness. Eleventh Workshop on Mining and Learning with Graphs, 2013.
7. A. S. Teixeira, P. T. Monteiro, J. A. Carriço, M. Ramirez, A. P. Francisco, Not Seeing the Forest for the Trees: Size of the Minimum Spanning Trees (MSTs) Forest and Branch Significance in MST-Based Phylogenetic Analysis. *PLOS ONE*, 10(3): e0119315, 2015.
8. C. Mavroforakis, R. Garcia-Lebron, I. Koutis, and E. Terzi. Spanning Edge Centrality: Large-scale Computation and Applications. *In Proceedings of the 24th International Conference on World Wide Web (WWW '15)*, 732-742, 2015.
9. X. Qi, E. Fuller, R. Luo, and C. Zhang. 2015. A novel centrality method for weighted networks based on the Kirchhoff polynomial. *Pattern Recogn. Lett.*, 58, C, 2015.
10. R. Albert and A.Barabási, Statistical mechanics of complex networks. *Rev. Mod. Phys*, Vol. 74, No. 47 (2002).
11. A. Barabási and R. Albert, Emergence of Scaling in Random Networks. *Science*, Vol. 286 no. 5439, 509-512, 1999.
12. F. Morone and H. A. Makse, Influence maximization in complex networks through optimal percolation, *Nature*, Vol. 524, 65-68, 2015.
13. D. Grady, C. Thiemann, and D. Brockmann, Robust classification of salient links in complex networks, *Nat. Commun.*, Vol. 3, 864, 2012.