Identifying significant edges via neighborhood information

Na Zhao\textsuperscript{a,b,1}, Jie Li\textsuperscript{a,1}, Jian Wang\textsuperscript{c,1}, Tong Li\textsuperscript{d}, Yong Yu\textsuperscript{a,*}, Tao Zhou\textsuperscript{e,*}

\textsuperscript{a}Key Laboratory in Software Engineering of Yunnan Province, School of Software, Yunnan University, Kunming 650091, P. R. China
\textsuperscript{b}Electric Power Research Institute of Yunnan Power Grid, Kunming 650217, P. R. China
\textsuperscript{c}College of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650504, P. R. China
\textsuperscript{d}School of Big Data, Yunnan Agricultural University, Kunming 650201, P. R. China
\textsuperscript{e}CompleX Lab, University of Electronic Science and Technology of China, Chengdu 61171, P. R. China

Abstract

Heterogeneous nature of real networks implies that different edges play different roles in network structure and functions, and thus to identify significant edges is of high value in both theoretical studies and practical applications. We propose the so-called second-order neighborhood (SN) index to quantify an edge’s significance in a network. We compare SN index with many other benchmark methods based on 15 real networks via edge percolation. Results show that the proposed SN index outperforms other well-known methods.

Keywords: complex networks, significant edges, second-order neighborhood index, edge percolation, robustness

1. Introduction

Many systems in nature and human society, such as communication, social and transportation systems, can be modeled by networks \cite{1}. Given the heterogeneity of real networks \cite{2}, a few nodes and edges play critical roles and largely affect network structure and functions, while the majority of

*Corresponding author

Email addresses: yuy1219@163.com (Yong Yu), zhutou@ustc.edu (Tao Zhou)

\textsuperscript{1}These three authors contributed equally to this work.
them are less important. Numerous methods have been proposed to identify critical nodes (see review articles [3, 4] and the references therein). However, how to measure edge significance receives less attention.

The identification of significant edges is highly valuable in practice. First, it can protect a system from intentional attacks. For example, one can protect a power grid from possible attacks by identifying significant transmission lines, thereby reducing the occurrences of failures. Second, deleting nodes may be too intrusive when we want to prevent failures or propagations. In comparison, edge-cutting strategies may be more applicable in some situations. For example, in financial networks, banks may withdraw certain products or reduce cooperations with some business partners to avoid financial risks, and in air transportation networks, some airlines may be closed to prevent long-range spreading of a certain disease. However, it is highly costly or even infeasible to shut down a bank or an airport. Several methods aiming at uncovering the role of a small set of significant edges in maintaining the network connectivity have been proposed [5, 6, 7, 8, 9], yet we are still in need of more accurate algorithms to locate significant edges.

The majority of current methods in quantifying edge significance based only on structural information. Ball et al. [10] used the effect of removing an edge on the average distance of a network to evaluate the significance of this edge. This method is highly time-consuming and not applicable when the removal causes the network unconnected. Girvan and Newman [11] used edge betweenness (EB) to quantify the significance of an edge, which has been successfully applied in community detection. However, it requires huge computational resource. Holme et al. [5] proposed the measure called degree product (DP), that is, the product of the degrees of two endpoints of an edge. Liu et al. [12] proposed a measure called diffusion intensity (DI) that accounts for the function of an edge in spreading dynamics. Onnela et al. [13] proposed an index named topological overlap (TO), which performs very well in identifying important edges in mobile communication networks. Cheng et al. [14] proposed the bridgeness (BN) index to detect significant edges that boost the communication between two densely connected subnetworks. Yu et al. [15] proposed an algorithm that combines global index of edge betweenness with local index of degree and clique. Other measures include eigenvalues [16], link entropy [17], nearest neighbor connections [18], and so on.

In this study, we propose the so-called second-order neighborhood (SN) index that accounts for the topological overlap of the two endpoints’ second-
order neighborhoods. This index can be considered as an extension of the famous TO index, and a tradeoff between computational complexity and algorithmic accuracy. To validate the significance of the edges selected by SN index, we apply the edge percolation dynamics \[19, 20\] to see whether the removal of a few edges with the highest SN values will lead to the fragmentation of the target network. Experimental results on 15 real networks demonstrate that the SN index remarkably outperforms other benchmark methods.

2. Methods

Consider a connected simple network \( G(V, E) \), where \( V \) and \( E \) are sets of nodes and edges, respectively. The number of nodes and edges are denoted by \( N = |V| \) and \( M = |E| \), and the edge connecting nodes \( i \) and \( j \) is denoted by \( e_{ij} \). The definition of five well-known benchmarks (see more indices in the review article [21]) are introduced below, followed by the description of the SN index.

EB \[11\] is defined as

\[
EB(i, j) = \sum_{s \neq t} \frac{\sigma_{st}(e_{ij})}{\sigma_{st}},
\]

where \( \sigma_{st} \) is the number of shortest paths from node \( s \) to node \( t \), and \( \sigma_{st}(e_{ij}) \) is the number of shortest paths from node \( s \) to node \( t \) that passing through edge \( e_{ij} \). Given that EB requires the calculation of the shortest paths between all node pairs, its time complexity is \( O(N^3) \) and thus it is unsuitable for large-scale networks.

DP \[5\] is defined as

\[
DP(i, j) = k_i k_j,
\]

where \( k_i \) and \( k_j \) represent degrees of nodes \( i \) and \( j \), respectively.

DI \[12\] is defined as

\[
DI(i, j) = \frac{n_{i \setminus j} + n_{j \setminus i}}{2},
\]

where \( n_{i \setminus j} \) is the number of \( i \)'s neighbors that are not connected to \( j \) or being \( j \) itself. The definition of \( n_{j \setminus i} \) is similar. An illustration about how to calculate \( n_{i \setminus j} \) and \( n_{j \setminus i} \) is shown in figure 1.
Figure 1: Illustration of DI index and TO index. Since $n_{i \setminus j}$ is 3 (red nodes) and $n_{j \setminus i}$ is 2 (blue nodes), $DI(i, j) = (3 + 2)/2 = 2.5$. As the number of common neighbors of nodes $i$ and $j$ is $n_{ij} = 2$, $TO(i, j) = 2/((6 - 1) + (5 - 1) - 2) \approx 0.286$.

TO \cite{13} is defined as

$$TO(i, j) = \frac{n_{ij}}{(k_i - 1) + (k_j - 1) - n_{ij}}, \quad (4)$$

where $n_{ij}$ is the number of common neighbors of nodes $i$ and $j$. This index is very similar to the Jaccard index \cite{22}. A simple example about how to calculate $TO(i, j)$ is also shown in figure 1.

BN \cite{14} is defined as:

$$BN(i, j) = \sqrt{S_i S_j/S(e_{ij})}, \quad (5)$$

where $S_i$ is the size of the largest clique containing node $i$ and $S(e_{ij})$ is the size of the largest clique containing edge $e_{ij}$. A clique is a full connected subgraph. A simple example is illustrated in figure 2.

Next, we introduce the SN index for an arbitrary edge $e_{ij}$. Denote $n^{(2)}_{i \setminus j}$ the number of nodes whose distance to node $i$ are 2 in the subgraph $G \setminus \{e_{ij}\}$ obtained by removing edge $e_{ij}$ from $G$. $n^{(2)}_{j \setminus i}$ is defined in a similar way. Then, the $SN$ index is defined as:

$$SN(i, j) = \frac{|n^{(2)}_{i \setminus j} \cap n^{(2)}_{j \setminus i}|}{|n^{(2)}_{i \setminus j} \cup n^{(2)}_{j \setminus i}|}. \quad (6)$$

A simple example is illustrated in figure 3. If the second-order common neighbors of nodes $i$ and $j$ are rare, the edge $e_{ij}$ is more likely to be a potential
Figure 2: Illustration of the BN index. As node \( i \), node \( j \) and edge \( e_{ij} \) are contained in the corresponding largest cliques with sizes being 5, 4 and 3, \( BN(i,j) = \sqrt{5 \times 4}/3 \approx 1.49 \).

Figure 3: Illustration of the SN index, where \( n_{i\setminus j}^{(2)} \) is 3 (red and green nodes), \( n_{j\setminus i}^{(2)} \) is 2 (blue and green nodes), and thus \( SN(i,j) = 1/4 = 0.25 \).

bridge between two different communities, which is crucial in facilitating communications between these two communities \[23, 24\]. Therefore, the smaller the SN index is, the more significant role the edge plays.

3. Results

To validate the role of edges in maintaining the network connectivity, we explore the edge percolation dynamics \[19, 20\], where in each time step an edge is removed from the target network until the remaining network becomes empty. We employ the famous measure named robustness \[25\] to estimate the impact of the edge removal on the network connectivity. The robustness \( R \) is defined as:
\[ R = \frac{1}{N} \sum_{\ell=1}^{N} \gamma_{\ell}, \]  

(7)

where \( \gamma_{\ell} \) is the ratio of the number of nodes in the maximum connected component after the removal of \( \ell \) edges to the number of nodes in the original network. \( 1/M \) is a normalization factor that guarantees the comparison of networks with different sizes. Obviously, a smaller value of \( R \) suggests a faster fragmentation, indicating the corresponding index can better rank the edge significance.

We conduct experiments on 15 real networks from different fields. (i) *Dolphins.* — A social network with frequent associations among 62 dolphins \[26\]. (ii) *Polbooks.* — A network of purchases of political US books \[27\]. (iii) *Adjnoun.* — An adjacent network of common adjectives and nouns in Charles Dickens’ novel *David Copperfield* \[28\]. (iv) *Neural.* — A neural network of *C.elegans* \[29\]. (v) *Jazz.* — A collaborative network of jazz musicians \[30\]. (vi) *Metabolic.* — A metabolic network of *C.elegans* \[31\]. (vii) *Email.* — An e-mail exchange network among the students of the University of Rovira-Virginy \[32\]. (viii) *Yeast.* — A protein-protein interaction network of Yeast \[33\]. (ix) *Friendship.* — A friendship network for users of hamsterster.com \[34\]. (x) *Kohonen.* — A citation network from Pajek \[35\]. (xi) *FissionYeast.* — A yeast fission network \[35\]. (xii) *Dmela.* — A protein-protein interaction network of *D.melanogaster* \[35, 36\]. (xiii) *Openflights.* — The airport transportation network in Openflights.org where two airports are connected by an edge if there is at least one direct flight between them \[35, 37\]. (xiv) *Lederberg.* — A citation network obtained from the Garfield’s collection \[35\]. (xv) *AstroPh.* — A collaborative network of authors having uploaded papers in the arXiv Astrophysics sector \[38\]. Table 1 summarizes the basic topology characteristics of these networks.

Figure 4 shows the collapsing processes of these real networks, resulted from the edge removal by using SN and other benchmark algorithms. Overall speaking, SN leads to much faster collapse than all other algorithms. Table 2 compares the robustness of SN and other benchmarks. Results show that, of all the benchmarks, SN index invariably produces the minimum \( R \) across all the 15 real networks. That is to say, SN index perform best in identifying the most significant edges.
| Networks      | $N$ | $M$  | $\langle k \rangle$ | $C$    | $r$     |
|--------------|-----|------|---------------------|--------|--------|
| Dolphins     | 62  | 159  | 5.1290              | 0.2590 | -0.0436|
| Polbooks     | 105 | 441  | 8.4000              | 0.4875 | -0.1279|
| Adjnoun      | 112 | 425  | 7.5893              | 0.1728 | -0.1293|
| Neural       | 297 | 2148 | 14.4646             | 0.2924 | -0.1632|
| Jazz         | 198 | 2742 | 27.6970             | 0.6175 | 0.0202 |
| Metabolic    | 453 | 2025 | 8.9404              | 0.6465 | -0.2196|
| Email        | 1133| 5451 | 9.6222              | 0.2202 | 0.0782 |
| Yeast        | 2361| 7182 | 6.0839              | 0.1301 | -0.0846|
| Friendships  | 1858| 12534| 13.4919             | 0.0904 | -0.0846|
| Kohonen      | 4470| 12731| 5.6962              | 0.2100 | -0.1204|
| FissionYeast | 2031| 25274| 24.8882             | 0.1874 | -0.1013|
| Dmela        | 7393| 25569| 6.9171              | 0.0118 | -0.0465|
| Openflights  | 2918| 30501| 20.9054             | 0.3967 | 0.0461 |
| Lederberg    | 8843| 41601| 9.4088              | 0.2968 | -0.0996|
| AstroPh      | 16705| 121251| 14.5167             | 0.6387 | 0.2355 |

Table 1: Basic topological features of 15 real networks, where $\langle k \rangle$, $C$ and $r$ represent average degree, clustering coefficient \[29\] and assortativity coefficient \[39\], respectively.
Figure 4: The ratio $\gamma$ versus the fraction of edges being removed for the 15 real networks. The robustness $R$ can be simply interpreted as the area under the $\gamma - f$ curve.
| Networks       | EB  | DP  | DI  | TO  | BN  | SN  |
|---------------|-----|-----|-----|-----|-----|-----|
| Dolphins      | 0.5110 | 0.6746 | 0.6136 | 0.4709 | 0.4953 | **0.3991** |
| Polbooks      | 0.4773 | 0.7584 | 0.7063 | 0.4176 | 0.4921 | **0.4105** |
| Adjnoun       | 0.6841 | 0.7975 | 0.7772 | 0.6118 | 0.6642 | **0.5278** |
| Neural        | 0.6617 | 0.7270 | 0.6962 | 0.6662 | 0.6511 | **0.5216** |
| Jazz          | 0.6660 | 0.9137 | 0.8095 | 0.6398 | 0.7244 | **0.6039** |
| Metabolic     | 0.7706 | 0.8519 | 0.7903 | 0.7522 | 0.7853 | **0.6820** |
| Email         | 0.6252 | 0.8090 | 0.7621 | 0.5532 | 0.6907 | **0.5126** |
| Yeast         | 0.4493 | 0.6354 | 0.5650 | 0.4234 | 0.5100 | **0.3363** |
| Friendships   | 0.5744 | 0.7994 | 0.7598 | 0.5234 | 0.6839 | **0.4062** |
| Kohonen       | 0.5357 | 0.6946 | 0.6470 | 0.4879 | 0.5951 | **0.4616** |
| FissionYeast  | 0.3882 | 0.8079 | 0.6920 | 0.3693 | 0.6641 | **0.2862** |
| Dmela         | 0.5786 | 0.7400 | 0.7300 | 0.6345 | 0.7166 | **0.3761** |
| Openflights   | 0.4114 | 0.7448 | 0.7077 | 0.3727 | 0.5661 | **0.3431** |
| Lederberg     | 0.5111 | 0.7750 | 0.7175 | 0.5244 | 0.6380 | **0.4697** |
| AstroPh       | 0.3777 | 0.7175 | 0.5521 | 0.3485 | 0.3988 | **0.3241** |

Table 2: Comparison of $R$ for different methods on 15 networks. The best result for each network (i.e., the lowest $R$ in the corresponding row) is highlighted in bold.

4. Conclusion

Identification of significant edges is of both theoretical interests and practical importance, yet it receives less attention in comparison with the challenge to dig out critical nodes. This paper provides a novel index, the SN index, to measure the importance of an edge in maintaining the network connectivity. The SN index takes into account the second-order neighborhood of each endpoint of the target edge, which can be considered as a tradeoff between indices using global topological information and indices only accounting for the nearest neighbors. It is not surprising that the SN index performs better than nearest-neighborhood-based indices like TO, while what beyond our expectation is that it performs remarkably better than the global index EB. On the basis of the extensive experiments on real networks from disparate fields, we believe that the SN index is a good candidate in quantifying an edge’s significance. In the future study, one could test the performance of such index in some other dynamical processes. In addition, this work provides a simple yet clear research framework about how to identify significant edges, and thus we hope it could facilitate further studies on this issue.
Acknowledgments

This work has been supported by the National Key Research and Development Program No. 2018YFB2100100, 2018YFC0830100, the National Natural Science Foundation of China under Grant Nos. 11975071 and 61433014. Data Driven Software Engineering innovation team of Yunnan province of China No. 2017HC012, State Key Laboratory of Computer Architecture No. CARCH201813.

References

[1] R. Albert, A.-L. Barabási, Statistical mechanics of complex networks, Reviews of Modern Physics 74 (2002) 47–97.

[2] G. Caldarelli, Scale-free networks: complex webs in nature and technology, Oxford University Press, 2007.

[3] S. Pei, H. A. Makse, Spreading dynamics in complex networks, Journal of Statistical Mechanics: Theory and Experiment (2013) P12002.

[4] L. Lü, D. Chen, X.-L. Ren, Q.-M. Zhang, Y.-C. Zhang, T. Zhou, Vital nodes identification in complex networks, Physics Reports 650 (2016) 1–63.

[5] P. Holme, B. J. Kim, C. N. Yoon, S. K. Han, Attack vulnerability of complex networks, Physical Review E: Statistical Nonlinear and Soft Matter Physics 65 (2002) 056109.

[6] Y. Xia, D. J. Hill, Attack vulnerability of complex communication networks, IEEE Transactions on Circuits and Systems II: Express Briefs 55 (2008) 65–69.

[7] W.-X. Wang, G. Chen, Universal robustness characteristic of weighted networks against cascading failure, Physical Review E: Statistical Nonlinear and Soft Matter Physics 77 (2008) 026101.

[8] J. Platig, E. Ott, M. Girvan, Robustness of network measures to link errors, Physical Review E: Statistical Nonlinear and Soft Matter Physics 88 (2013) 062812.
[9] B. Duan, J. Liu, M. Zhou, L. Ma, A comparative analysis of network robustness against different link attacks, Physica A: Statistical Mechanics and its Applications 448 (2016) 144–153.

[10] M. O. Ball, B. L. Golden, R. V. Vohra, Finding the most vital arcs in a network, Operations Research Letters 8 (1989) 73–76.

[11] M. Girvan, M. E. Newman, Community structure in social and biological networks, Proceedings of the National Academy of Sciences of the United States of America 99 (2002) 7821–7826.

[12] Y. Liu, M. Tang, T. Zhou, Y. Do, Improving the accuracy of the k-shell method by removing redundant links: From a perspective of spreading dynamics, Scientific Reports 5 (2015) 13172.

[13] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, A.-L. Barabási, Structure and tie strengths in mobile communication networks, Proceedings of the National Academy of Sciences of the United States of America 104 (2007) 7332–7336.

[14] X.-Q. Cheng, F.-X. Ren, H.-W. Shen, Z.-K. Zhang, T. Zhou, Bridge-ness: a local index on edge significance in maintaining global connectivity, Journal of Statistical Mechanics: Theory and Experiment (2010) P10011.

[15] E.-Y. Yu, D.-B. Chen, J.-Y. Zhao, Identifying critical edges in complex networks, Scientific Reports 8 (2018) 14469.

[16] J. G. Restrepo, E. Ott, B. R. Hunt, Characterizing the dynamical importance of network nodes and links, Physical Review Letters 97 (2006) 094102.

[17] Y. Qian, Y. Li, M. Zhang, G. Ma, F. Lu, Quantifying edge significance on maintaining global connectivity, Scientific Reports 7 (2017) 45380.

[18] B. Ouyang, Y. Xia, C. Wang, Q. Ye, Z. Yan, Q. Tang, Quantifying importance of edges in networks, IEEE Transactions on Circuits and Systems II Express Briefs 65 (2018) 1244–1248.

[19] D. S. Callaway, M. E. J. Newman, S. H. Strogatz, D. J. Watts, Network robustness and fragility: Percolation on random graphs, Physical Review Letters 85 (2000) 5468–5471.
[20] C. Moore, M. E. J. Newman, Epidemics and percolation in small-world networks, Physical Review E: Statistical Physics Plasmas Fluids and Related Interdisciplinary Topics 61 (2000) 5678–5682.

[21] L. Lü, T. Zhou, Link prediction in complex networks: A survey, Physica A: Statistical Mechanics and Its Applications 390 (2011) 1150–1170.

[22] P. Jaccard, The distribution of the flora in the alpine zone, New Phytologist 11 (1912) 37–50.

[23] M. Granovetter, Getting a job: a study of contacts and careers, University of Chicago Press, 1995.

[24] P. Csermely, Weak links: Stabilizers of complex systems from proteins to social networks, Springer Berlin, 2006.

[25] C. M. Schneider, A. A. Moreira, J. S. Andrade, S. Havlin, H. J. Herrmann, Mitigation of malicious attacks on networks, Proceedings of the National Academy of Sciences of the United States of America 108 (2011) 3838–3841.

[26] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, S. M. Dawson, The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations, Behavioral Ecology and Sociobiology 54 (2003) 396–405.

[27] V. Krebs, Books about US politics network dataset, 2004. URL: http://www.orgnet.com/.

[28] M. E. J. Newman, Finding community structure in networks using the eigenvectors of matrices, Physical Review E: Statistical Nonlinear and Soft Matter Physics 74 (2006) 036104.

[29] D. J. Watts, S. H. Strogatz, Collective dynamics of ‘small-world’ networks, Nature 393 (1998) 440–442.

[30] P. M. Gleiser, L. Danon, Community structure in jazz, Advances in Complex Systems 6 (2003) 565–573.

[31] D. Jordi, A. Alex, Community detection in complex networks using extremal optimization, Physical Review E Statistical Nonlinear and Soft Matter Physics 72 (2005) 027104.
[32] R. Guimera, L. Danon, A. Diaz-Guilera, F. Giralt, A. Arenas, Self-similar community structure in a network of human interactions, Physical Review E: Statistical Nonlinear and Soft Matter Physics 68 (2003) 065103.

[33] D. Bu, Y. Zhao, L. Cai, H. Xue, X. Zhu, H. Lu, J. Zhang, S. Sun, L. Ling, N. Zhang, Topological structure analysis of the protein-protein interaction network in budding yeast, Nucleic Acids Research 31 (2003) 2443–2450.

[34] Hamsterster friendships network dataset – KONECT, 2017. URL: http://konect.uni-koblenz.de/networks/petster-friendships-hamster.

[35] R. A. Rossi, N. K. Ahmed, The network data repository with interactive graph analytics and visualization, in: Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.

[36] R. Singh, J. Xu, B. Berger, Global alignment of multiple protein interaction networks with application to functional orthology detection, Proceedings of the National Academy of Sciences of the United States of America 105 (2008) 12763–12768.

[37] T. Opsahl, Why anchorage is not (that) important: Binary ties and sample selection, 2011. URL: http://wp.me/poFcY-Vw.

[38] M. E. Newman, The structure of scientific collaboration networks, Proceedings of the National Academy of Sciences of the United States of America 98 (2001) 404–409.

[39] M. McPherson, L. Smith-Lovin, J. M. Cook, Birds of a feather: Homophily in social networks, Annual Review of Sociology 27 (2001) 415–444.