Erschienen in: MMB & DFT 2014 : Proceedings of the International Workshops ; Modeling, Analysis and Management of Social Networks and Their Applications (SOCNET 2014) & Demand Modeling and Quantitative Analysis of Future Generation Energy Networks and Energy-Efficient Systems (FGENET 2014) / Fischbach, Kai ... (eds.). - Bamberg : Univ. of Bamberg Press, 2014. - S. 11-18. - (Schriften aus der Fakultät Wirtschaftsinformatik und Angewandte Informatik der Otto-Friedrich-Universität Bamberg ; 16). - ISBN 978-3-86309-208-5

Centrality as a Predictor of Lethal Proteins: Performance and Robustness

David Schoch[1,2] and Ulrik Brandes[1,2]

[1] Department of Computer & Information Science, University of Konstanz
[2] Graduate School of Decision Sciences, University of Konstanz

Abstract. The Centrality-Lethality Hypothesis states that proteins with a higher degree centrality are more likely to be lethal, i.e. proteins involved in more interactions are more likely to cause death when knocked off. This proposition gave rise to several new investigations in which stronger associations were obtained for other centrality measures. Most of this previous work focused on the well known protein-interaction network of *Saccharomyces cerevisiae*. In a recent study, however, it was found that degree and betweenness of lethal proteins is significantly above average across 20 different protein-interaction networks. Closeness centrality, on the other hand, did not perform as well.
We replicate this study and show that the reported results are due largely to a misapplication of closeness to disconnected networks. A more suitable variant actually turns out to be a better predictor than betweenness and degree in most of the networks. Worse, we find that despite the different theoretical explanations they offer, the performance ranking of centrality indices varies across networks and depends on the somewhat arbitrary derivation of binary network data from unreliable measurements. Our results suggest that the celebrated hypothesis is not supported by data.

Key words: Network Centrality, Protein Networks, Centrality-Lethality

1 Introduction

With advances in high-throughput analysis, availability of protein interaction data increased dramatically. This provides opportunities to examine interactions and their properties using network analysis. Substantial interest was sparked by Jeong

et al. [1] who propose that lethal proteins, i.e. proteins causing death if knocked off, tend to have more interactions than non-lethal ones, i.e. they have a higher degree. These findings led to a flurry of follow up studies and a hunt for the centrality best suited to identify lethal proteins [2,3,4,5,6,7,8]. Most of these stuck with the protein-interaction network of *Saccharomyces cerevisiae* used in the original study. Only few studies dealt with different organisms such as *Drosophila melanogaster* or *Caenorhabditis elegans* [9]. In a very recent study, Raman *et al.* [10] reviewed the Centrality-Lethality Hypothesis across protein networks of 20 different organisms. Using a bootstrapping approach, they showed that degree and betweenness centrality of lethal proteins are significantly higher than the network average. In contrast, closeness centrality was found to be less indicative of lethality.

In the following, we reexamine their results, albeit with a variant of closeness centrality correcting for the fact that most of the networks are not connected. Moreover, we use a more detailed evaluation method specifically designed for models with binary outcomes, namely the receiver operating characteristic [15]. Finally, we analyze the robustness of these results when the threshold for high-confidence interactions is varied and discuss theoretical upper bounds for the case when gene attributes are taken into account as well.

2 Methods

2.1 Data

The protein interactions of 20 organisms[3] were obtained from the *STRING Database* (version 9.0). Besides experimentally identified interactions from published literature, the database also contains computationally predicted interactions. Each interaction is given a score which indicates the probability of an actual interaction. We constructed eight networks using $S \in \{600, 650, 700, 750, 800, 850, 900, 950\}$ as lower bounds for the interaction scores for each organism. Lethality data were obtained from the *Database of Essential Genes* (DEG version 5.0).

2.2 Network Analysis

Protein interactions are represented in an undirected graph $G = (V, E)$, where the vertices V represent proteins equipped with a binary attribute indicating lethality

[3] We use the same organisms as in [10], except we choose *D. melanogaster* instead of *S.e.S. typhi*.

and the edges E represent interactions. The cardinalities $|V| =: n$ and $|E| =: m$ denote the number of proteins and interactions respectively. The adjacency matrix $A = (a_{ij})$ encodes the network relation, i.e. $a_{ij} = 1$ if $\{i, j\} \in E$ and $a_{ij} = 0$ otherwise.

For the prediction of lethal proteins we use four standard indices, degree, betweenness, closeness and eigenvector centrality, together with two indices proposed specifically to identify lethal proteins: subgraph centrality [11] and bipartivity [12].

Degree centrality (C_D) is defined as the number of edges incident to a vertex. Betweenness centrality (C_B) quantifies the participation of a node in the shortest paths of the network. It is defined as

$$C_B(v) = \sum_{s \neq t \in V \setminus \{v\}} \frac{\sigma(s, t|v)}{\sigma(s, t)} ,$$

where $\sigma(s, t)$ is the number of shortest paths connecting s and t and $\sigma(s, t|v)$ is the number of shortest paths from s to t passing through v. Closeness centrality (C_C) of a vertex v is defined as the inverse of the sum of its distances to all other vertices in the network,

$$C_C(v) = \frac{1}{\sum_{t \in V \setminus \{v\}} dist(v, t)} .$$

By definition of shortest-path distances, C_C is ill-defined on unconnected networks. Replication confirmed that it was used in [10] nevertheless, which may explain its comparatively poor performance. We therefore use a close variant applicable to both connected and unconnected graphs instead [4],

$$C_C^*(v) = \sum_{t \in V \setminus \{v\}} \frac{1}{dist(v, t)} .$$

Eigenvector centrality (C_E) of a node v is given by the vth entry of the eigenvector corresponding to the largest eigenvalue of A. Again, this formulation is not well-defined for unconnected networks. We therefore calculate C_E for each component separately and scale the values according to the number of nodes in each component. Subgraph centrality (C_S) sums up all closed walks starting and ending at a vertex v. These closed walks are weighted in a way that their contribution decreases as the length increases,

$$C_S(v) = \sum_{k=0}^{\infty} \frac{\left(A^k\right)_{vv}}{k!} = \mathit{trace}(e^A)_v .$$

[4] This variant was proposed, for instance, by Agneessens and Borgatti (presentation at the ASNA 2012 conference)

Bipartivity (β) is defined as the proportion of closed walks of even-length and can be expressed as

$$\beta(v) = \frac{C_{S_{even}}(v)}{C_S(v)} = \frac{\sum_{j=1}^{n} [x_j(v)]^2 \cosh(\lambda_j)}{C_S(v)} ,$$

where $x_j(v)$ is the vth component of the jth eigenvector associated with the eigenvalue λ_j of A. The values of β are confined to the interval $[0.5, 1]$. According to [3], lethal proteins tend to have a low bipartivity score. Therefore we adjust the value by setting it to $1 - \beta(v)$, such that lethal proteins potentially have a higher score.

2.3 Receiver Operating Characteristic

To measure the performance of centrality indices as a predictor for lethal proteins we use the receiver operating characteristic (ROC) [15]. The power of a prediction model can be summarized by the area under the ROC curve (AUC). AUC values are bounded between 0 and 1, where a value of 0.5 is the expected performance of a random classifier and higher (lower) scores indicate a better (worse) prediction than expected by chance.

3 Results

In this section we investigate the predictive power of the six indices to identify lethal proteins. Recall that two of them needed correction to account for disconnectedness. In addition, we examine whether the results are stable with respect to the interaction-confidence threshold S and discuss potential upper bounds for the predictions.

3.1 Prediction Performance

Table 1 shows the AUC values of the six centrality indices for the networks with $S = 700$. In contrast to the results reported in [10], we see that the adjusted closeness performs better than degree and betweenness in most of the networks. However, the efficiency varies strongly across all organisms in general.

Recently it has been argued that the identification of lethal proteins can be improved if centrality indices are combined with further attributes, say from gene expression data [7,13]. In such a scenario, preservation of the neighborhood inclusion preorder remains a necessary condition for a centrality effect to exist. We

thus obtain an upper bound on the performance of any classifier by minimizing the non-lethal/lethal inversions over all linear extension of this preorder. Even though the problem of finding minimum inversion extensions is NP-hard [14], we were able to find rankings with an optimal AUC value of 1 for all organisms. This implies that there is a lot of potential for performance improvement when external attributes are incorporated.

Table 1. AUC values for the protein networks with $S = 700$. Bold values indicate the best performance per organism.

Organism	C_D	C_B	C_C^*	C_E	C_S	β
A. bayli	0.80	0.72	**0.84**	0.77	**0.84**	0.82
A. thaliana	0.48	0.54	0.49	**0.56**	0.51	0.47
B. subtilis	0.80	0.70	**0.84**	0.56	0.72	0.72
C. elegans	0.60	0.58	0.61	0.63	**0.65**	0.53
E. coli	0.63	**0.71**	0.68	0.62	0.65	0.53
F. novicida	0.65	0.60	0.68	**0.71**	0.70	0.64
H. influenzae	0.77	0.68	0.78	0.77	**0.80**	0.79
H. pylori	0.56	0.54	**0.58**	0.55	**0.58**	0.57
D. melanogaster	0.63	0.58	0.64	0.60	**0.65**	**0.65**
M. genitalium	0.66	0.61	0.66	0.64	**0.70**	**0.70**
M. pulmonis	0.78	0.73	**0.82**	0.64	**0.82**	**0.82**
M. tuberculosis	0.67	0.63	0.66	0.68	**0.71**	0.70
P. aeruginosa	0.71	0.65	**0.77**	0.74	**0.77**	**0.77**
S. a.NCTC	0.79	0.73	0.85	0.77	**0.86**	0.85
S. a.s.a.N315	0.81	0.73	0.83	0.78	**0.84**	0.83
S. cerevisiae	**0.71**	0.64	0.70	0.70	0.70	0.50
S. pneumoniae	0.72	0.68	0.76	0.71	0.78	**0.79**
S. sanguinis	0.83	0.77	0.87	0.78	0.89	**0.88**
S. typhimurium	0.65	0.62	0.69	0.69	0.70	**0.71**
V. cholerae	0.65	0.61	0.69	**0.70**	**0.70**	0.69

3.2 Robustness

To test the robustness of the results, we varied the confidence threshold S to construct eight networks for each organism. Figure 1 illustrates that prediction accuracy depends heavily on the chosen threshold. Observe that the index producing the highest AUC value varies with S and that the results exhibit high variability in general.

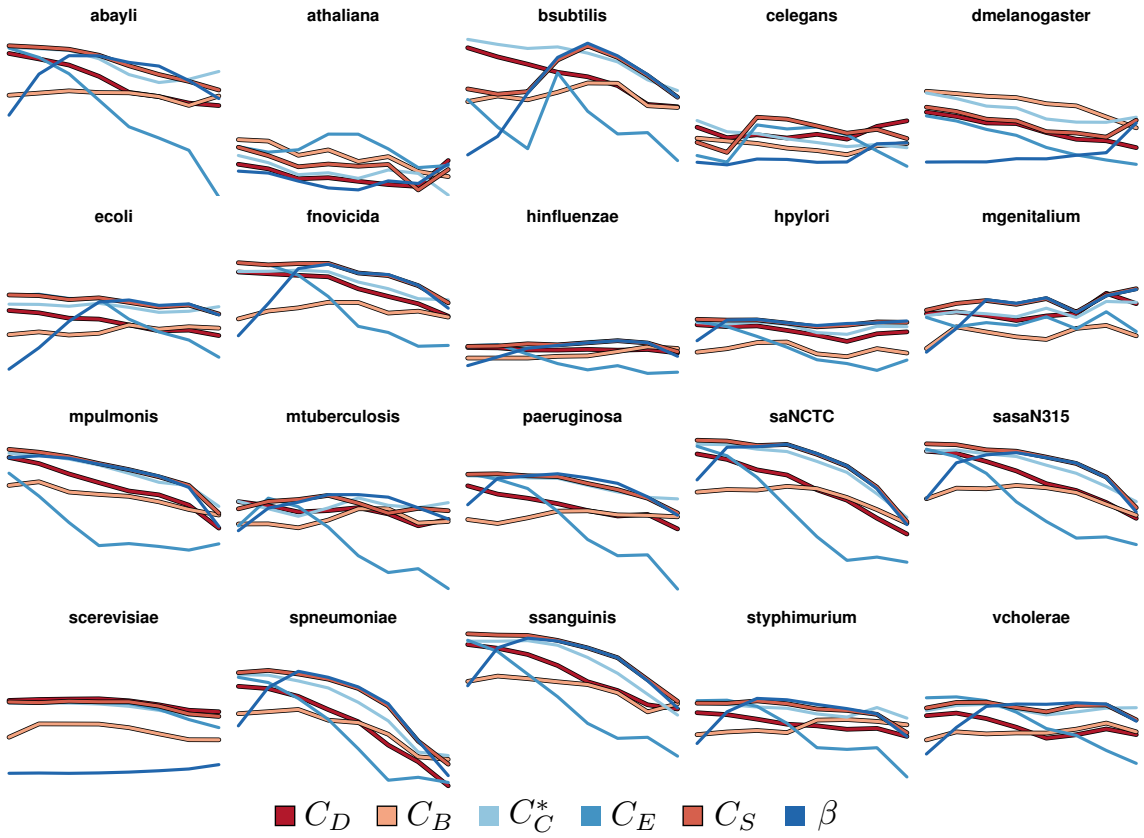

Fig. 1. Performance of the six centrality indices when S is varied from 600 to 950 (shown on the x-axis). The AUC values on the y-axis range from 0.45 to 0.9.

3.3 Discussion

Our reexamination shows that the original results are skewed for two main reasons: inappropriate use of two indices that are ill-defined on disconnected networks, and restriction to a single threshold for interactions. Both are connected to the availability of a finite list of centrality indices, from which instantiations are chosen or new indices are added. Since many of them are defined for connected and unweighted (or otherwise limited classes of) networks, but implementations often output results also for networks outside of this scope, studies need to check carefully whether the aggregate results obtained from such analyses are meaningful.

4 Conclusion

We redesigned and extended a study of Raman *et al.* [10] on the plausibility of the Centrality-Lethality Hypothesis across 20 different organisms. In contrast to [10], we find that (a suitably modified variant of) closeness performs better than

degree and betweenness. We also find, however, that the association of centrality and lethality heavily depends on where the line for high-confidence interactions is drawn.

By minimizing the inversions of lethal/non-lethal proteins over linear extensions of the neighborhood-inclusion preorder, we argued that, at least, there is no principled argument *against* a centrality effect. However, in their consideration of purely structural effects, previous results do not provide sufficient support the Centrality-Lethality Hypothesis either.

Acknowledgments. This research was supported in part by *Deutsche Forschungsgemeinschaft* under grant Br 2158/6-1 and the *Graduate School of Decision Sciences*. We are grateful to Stefan Felsner, Franz J. Brandenburg, and Andreas Gemsa for pointing us to the hardness result for the minimum inversion problem.

References

1. Hawoong Jeong, Sean P. Mason, Albert-László Barabási, and Zoltan N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411:41–42, 2001.
2. Ernesto Estrada. Virtual identification of essential proteins within the protein interaction network of yeast. *Proteomics*, 6(1):35–40, 2006.
3. Ernesto Estrada. Protein bipartivity and essentiality in the yeast protein-protein interaction network. *Journal of Proteome Research*, 5(9):2177–2184, 2006.
4. Gabriel del Rio, Dirk Koschützki, and Gerardo Coello. How to identify essential genes from molecular networks? *BMC Systems Biology*, 3(1):102, 2009.
5. Xue Zhang, Jin Xu, and Wangxin Xiao. A new method for the discovery of essential proteins. *PLoS one*, 8(3):e58763, 2013.
6. Huan Wang, Min Li, Jianxin Wang, and Yi Pan. A new method for identifying essential proteins based on edge clustering coefficient. In *Bioinformatics Research and Applications*, pages 87–98. Springer-Verlag, 2011.
7. Min Li, Jianxin Wang, and Yi Pan. A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data. *BMC Systems Biology*, 6(1):15, 2012.
8. Keunwan Park and Dongsup Kim. Localized network centrality and essentiality in the yeast–protein interaction network. *Proteomics*, 9(22):5143–5154, 2009.
9. Matthew W. Hahn and Andrew D. Kern. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Molecular Biology and Evolution*, 22(4):803–806, 2005.
10. Karthik Raman, Nandita Damaraju, and Govind Krishna Joshi. The organisational structure of protein networks: revisiting the centrality-lethality hypothesis. *Systems and Synthetic Biology*, pages 1–9, 2013.

11. Ernesto Estrada and Juan A. Rodríguez-Velázquez. Subgraph centrality in complex networks. *Physical Review E*, 71(5):056103, 2005.
12. Ernesto Estrada and Juan A. Rodríguez-Velázquez. Spectral measures of bipartivity in complex networks. *Physical Review E*, 72(4):046105, 2005.
13. Xiwei Tang, Jianxin Wang, and Yi Pan. Identifying essential proteins via integration of protein interaction and gene expression data. In *Bioinformatics and Biomedicine (BIBM) 2012 IEEE International Conference on*, pages 1–4, 2012.
14. Eugene L. Lawler. Sequencing jobs to minimize total weighted completion time subject to precedence constraints. *Annals of Discrete Mathematics*, pages 75–90, 1978.
15. Charles E. Metz. Basic principles of ROC analysis. *Seminars in nuclear medicine*, pages 283–298, 1978.