How Do Biological Networks Differ from Social Networks? (An Experimental Study)

Tatiana Gutiérrez-Bunster*, Ulrike Stege†, Alex Thomo‡, John Taylor§
*†‡ Department of Computer Science and §Biology, University of Victoria. Victoria. BC. Canada.
*Departamento de Sistemas de Información, Universidad del Bío-Bío. Concepción. Chile
Email: *tgutierr@uvic.ca, †stege@cs.uvic.ca, ‡ thomo@cs.uvic.ca, §taylorjs@uvic.ca

***Abstract*—In this paper we outline important differences between (1) protein interaction networks and (2) social and other complex networks, in terms of fine-grained network community profiles. While these families of networks present some general similarities, they also have some stark differences in the way the communities are formed. Namely, we find that the sizes of the best communities in such biological networks are an order of magnitude smaller than in social and other complex networks. We furthermore find that the generative model describing biological networks is very different from the model describing social networks. While for latter the Forest-Fire model best approximates their network community profile, for biological networks it is a random rewiring model that generates networks with the observed profiles. Our study suggests that these families of networks should be treated differently when deriving results from network analysis, and a fine-grained analysis is needed to better understand their structure.**

I. Introduction

We focus on the structural differences of protein interaction networks versus social and other complex networks. While there are some similarities between them, there are nevertheless significant differences, mainly in the community structure of these networks. This is in contrast to the widely held belief that biological networks are very similar to social networks and thus tools and insights from the latter can be easily applied to or extended for the former [4]. We show that best communities are smaller by an order of magnitude in biological networks compared to those in social networks.

Community detection is very important not only for social networks, but also for biological networks. This is because communities can provide for a better understanding and insight on the fine grained structure of biological networks and the way their different parts work together.

We compute for our study the *network community profiles* (NCPs) of 11 large protein-interaction networks. NCPs are based on the notion of conductance that captures the ratio of edges connecting nodes within the community with nodes outside the community to edges inside the community. The smaller the conductance of a set of nodes, the more community-like the set is. Conductance is extensively used to measure the cohesiveness of a community and has been shown to have parallels with the theory of random walks on networks.

Along the lines of [13], we investigate the conductance of communities over all the possible size scales. The main question we explore is: What are the best community sizes and community qualities for each network family? The network community profile is one of the best tools to answer this question. Intuitively, NCP extracts the conductance of the best community as a function of the size values considered. While NCP is NP-hard to compute, there are several approximate algorithms that give satisfactory solutions.

We present the following empirical findings. First, the conductance of the best communities for each size scale (k) decreases initially, and the global minimum is typically achieved for $k = 10$. This is in contrast to social networks where the global minimum is reached for $k = 100$ or greater, an order of magnitude bigger than the global minimum for biological networks. Second, at the size of about $k = 10$, the NCP for biological networks exhibits an uptrend, which means that the community structure deteriorates as more nodes are considered in communities. In other words, the communities start blending with each other and gradually disappear. And third, differently from social networks, the generative model explaining this type of behavior is not Forest Fire ([11], [12]), but a random rewiring model [18] conditioned on the same degree distribution as the original graph.

Knowing that the best communities in biological networks are an order of magnitude smaller than communi-

ties in social networks is very important. This is because community structure can help us to decide which are the possible missing links to further investigate. Clearly, there is a higher chance that there is a missing link between nodes within a community than between nodes not in the same community. Exploring missing links in social networks is not particularly expensive. However, it is quite expensive to do so for biological networks. Therefore, the smaller the meaningful communities, the fewer missing links we need to explore in a laboratory setting. A community profile plot helps in better understanding the costs of further investigating missing links in biological networks.

The rest of the paper is organized as follows. In Section II, we describe related work. In Section III, we describe conductance and community profiling. In Section IV, we describe the datasets we use. In Section V, we present our experimental results. In Section VI, we present the modelling results. In Section VII, we discuss other differences between families of networks. Section VIII concludes the paper.

II. Related Works

We focus on protein interaction networks. They have been the theme of numerous works in the research community. Pavlopoulos et al. [20] and Mason et al. [17] studied how to find the most important nodes in large protein networks. They utilize such information for better determining protein functions and identifying drug targets.

Barabassi and Oltvai in [4] study the general properties of the proteins in networks coming from complex interactions. Using network tools allowed them to see a different perspective of proteins and genes. They make the case that with respect to the common measures of network structure, the proteins in these networks and people in social networks behave similarly. In this work, however, we show that this is not always the case.

To understand the biological significance of the systems, many researchers applied different models, approaches, and methods to identify motifs or patterns that indicate common properties. They analyze the networks in detail using measures like network centralities [9], [19], network topologies ([21], [30]), cluster analysis ([15], [22]), or network models [20].

Cluster (community-detection) algorithms are used to understand the organization of networks and their functions through the identification of protein complexes or functional modules ([17], [27]). The clustering algorithms typically join proteins in groups (communities) according to attributes that are shared by the proteins in the group. They show that identifying and predicting communities also helps identifying important nodes (proteins) in the network. There are also comparative analyses of different clustering algorithms to identify those that are better at predicting relevant communities. Wang et al. in [27] present a detailed clustering algorithm comparison for extracting clusters from protein interaction networks. Most of the algorithms focus on protein complexes and functional modules. Some more recent works ([10], [23]) propose improving the prediction of protein function by utilizing protein community information. However, the aforementioned works do not make use of conductance scores as we do.

Centrality measures help to analyze the different communities and evaluate how a gene or protein is relevant for its community, other communities or the complete network [31]. The centrality evaluations give evidence that there is a close relationship between the centrality of a node and its essentiality in the network [17]. One example is presented in Goh et al. [7], where in the biological networks examined the betweenness and degree of nodes are significantly correlated. Also, in Girvan and Newman [6] it is mentioned that when the edges present high betweenness, there is a high probability that the communities are highly interconnected-too.

III. Conductance and Community Profiling

We consider the networks to be undirected graphs. Let G be such a graph with V as set of nodes, and E as set of edges. The conductance of a set $S \subset V$ is

$$\gamma(S) = \frac{|\{e_{ij} \in E : v_i \in S \text{ and } v_j \notin S\}|}{\min\{\sum_{v_i \in S} d(v_i), \sum_{v_j \notin S} d(v_j)\}}$$

where $d(.)$ denotes the degree of a node. The conductance of a set gives a score for the quality of the set as a community. The higher edges that cross the boundaries of a set S, the higher the conductance $\gamma(S)$, and the lower the community structure of S. Hence, for detecting good communities, we look for sets of low conductance. These are sets that are densely connected internally and sparsely connected with the rest of the graph. In Fig. 1, we observe three good communities, 1, 2, and 3.

We note that there are many community measures, however, as noted by several prominent works ([8], [25]), the conductance captures the gestalt of communities [32], and therefore is used frequently to perform

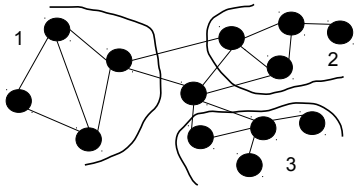


Fig. 1: A networks and three communities. Communities 1, 2, and 3 are densely connected internally and sparsely connected with the rest of the graph.

TABLE I: Biological networks

Biological Networks	Nodes	Edges	Reference
Arabidopsis thaliana	7,050	16,263	BioGrid [26]
Caenorhabditis elegans 1	3,895	7,758	BioGrid [26]
Caenorhabditis elegans 2	2,528	3,706	Harvard [2]
Drosophila melanogaster	8,127	38,839	BioGrid [26]
Echericha coli	2,874	11,538	DIP [29]
H pylo	708	1,357	DIP [29]
Homo sapiens 1	15,337	133,645	BioGrid [26]
Homo sapiens 2	6,711	17,348	Mint [14]
Mus musculus	4,602	9,841	BioGrid [26]
Saccharomyces cerevisie	5,376	24,734	Mint [14]
Schizosaccharomyces pombe	4,008	55,362	BioGrid [26]

community detection ([5], [16], [24]). In community profiling we select the best community for each size and plot their conductance scores. In order to find communities of good conductance, we use the local spectral clustering algorithm of [3] and the bag-of-whiskers clustering algorithm of [13]. Whiskers are sets of nodes connected to the rest of the graph by one edge; bag-of-whiskers are sets of such whiskers. As shown in [13], bags-of-whiskers give communities with very good conductance scores.

IV. Biological Networks Analyzed

We considered many of the reasonable-sized protein interaction networks [1] available. They amount to about 55 in total, coming from 16 species. Due to space constraints, we focus here on 11 of them (Table I). The results for other networks are comparable. The networks we consider have sizes varying from 708 to 15,337 nodes, and from 1,357 to 133,645 edges. The datasets were obtained from various sources (column 'Reference' in Table I).

[1]We will refer to these networks as "biological networks". We note that there are also other types of biological networks that we plan to study as part of our future work.

V. Network Community Plot Analysis

As shown in [13] most social networks exhibit the following community profile structure. Up to a certain size the slope of NCP is downward: as the size increases the conductance values decrease. This in turn means that the best sets become increasing community-like. At size of 100 or more, the NCP reaches a global minimum. This implies that the best communities in social networks are typically of size 100 or more. If larger than that, the NCP of most social networks is upward sloping over several orders of magnitude. This means that after a certain size, typically at least 100, the communities become less meaningful and they blend more and more with the whole network. For other networks, such a power-grid networks, the NCP is almost always slopping downwards. This means the more nodes are added to communities the better they become in terms of conductance.

In Fig. 2 and Fig. 3 we show the community profiles for biological and social networks and a power-grid network. We show the conductance scores of the best communities computed, using the Local Clustering algorithm of [3] and the Bag-of-Whiskers algorithm of [13]. See Table II for statistical data of the networks.

In Fig. 3, we show the community profiles of two social networks, Twitter and Facebook, as well as the community profile of a power-grid network. We observe that the community profiles of the two social networks have a downward slope up to a certain community size, and then they trend upward (similar extensive results for social networks are presented in [13]). Their global minimum is way greater (orders of magnitude) than the global minimums we observe for biological networks. Also, we observe that there are no whiskers for the Twitter and Facebook networks, i.e. there are no communities that are barely connected to the rest of the network.

Regarding the biological networks, we show their NCP in Fig. 2. We observe a similar shape of the NCPs as for social networks. Initially the slope is downward, then upward. However, the global minimum is not reached at size 100 or greater as for social networks, but surprisingly at size about 10, an order of magnitude smaller! Therefore, biological networks present a very different community structure than social networks; they have a much more local structure than social networks. We also observe that whiskers give significantly better communities than Local Clustering. This means that the best communities are only barely connected to the rest of the graph for biological networks.

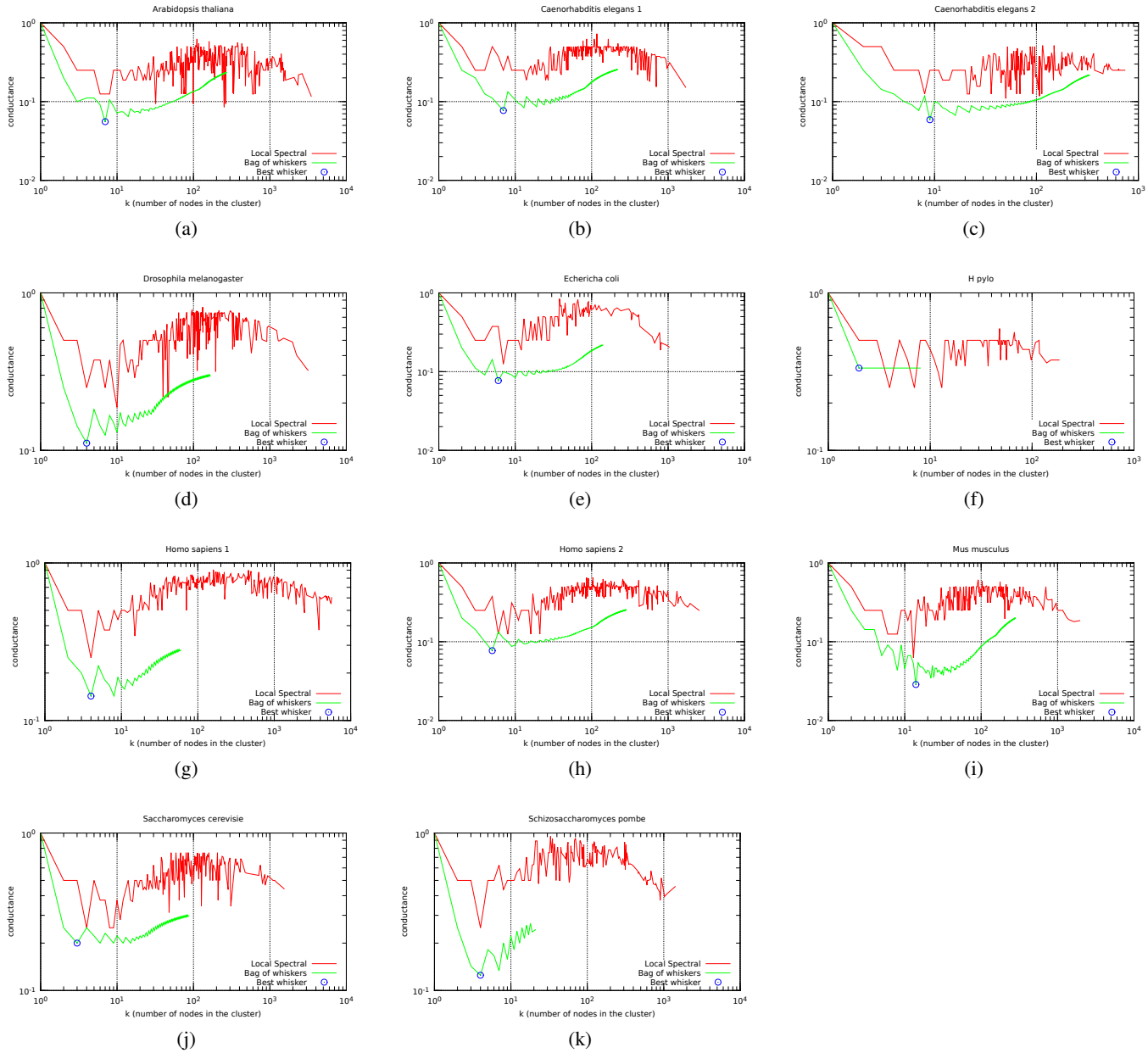
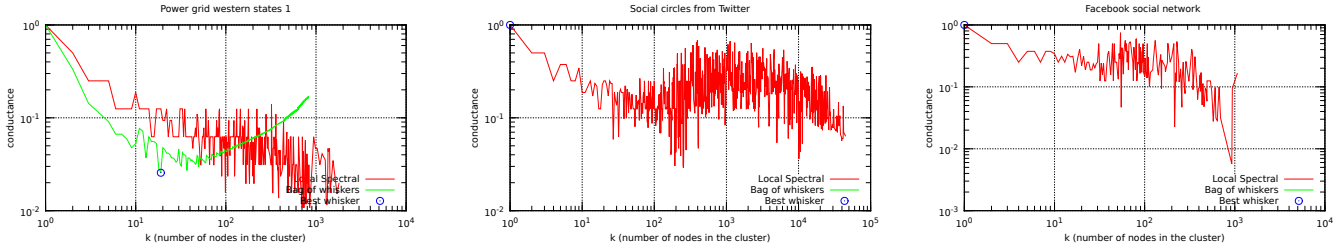


Fig. 2: Network community profiles for biological networks computed using the local spectral clustering (red/dark) and bag-of-whiskers (green/light) algorithms. We have conductance values in axis Y and number of nodes in the cluster in axis X. Both algorithms give a network community profile that is initially downward slopping, then trending upwards. The global minimum for both methods, across most of the biological networks, is at about a community size value of ten. This is in stark contrast to network community profiles for social and other complex networks. Also, observe that whiskers give significantly better communities than local spectral clustering.

VI. Modelling Results

A natural question we would like to answer is: What generative model best fits biological networks? For social networks, [13] shows that a Forest-Fire model, where new edges are added via a recursive burning mechanism in an epidemic-like fashion, generates networks with network profiles that closely resemble profiles of social networks.

In contrast, a Forest-Fire model is not the right choice



(a) (b) (c)

Fig. 3: Network community profiles (red/dark) of two social networks and a power-grid network (green/light). (a) $4,941$ nodes [28], (b) $81,306$ nodes [1], and (c) $4,039$ nodes [1]. The NCPs of Twitter and Facebook are first slopping downward, then after reaching a global minimum, change the slope to be upward. Differently from the biological networks, the mimimum is achieved for much larger values of community size. On the other hand, the NCP of the power-grid is always going downward. Another fact we can observe is that there are no whiskers (green) in the Twitter and Facebook networks. This is very different from what we see in biological networks.

TABLE II: Statistical Data

Networks	Average Degree	Network Diameter	Connected Components	Avg.Clustering Coefficient	Average Path Length	Average #Triangles
A. thaliana	4.61	14	154	0.16	4.46	2.79
C. elegans 1	3.98	13	86	0.11	4.29	1.86
C. elegans 2	2.93	14	147	0.04	5.32	0.37
D. melanogaster	9.56	10	50	0.12	4.09	22.37
E. coli 1	8.03	12	295	0.15	3.97	19.07
H. pylo	3.83	9	17	0.03	4.13	0.33
H. sapiens 1	17.43	8	57	0.32	2.67	85.47
H. sapiens 2	5.17	11	135	0.11	4.41	1.98
M. musculus	4.28	16	124	0.20	4.34	2.35
S. cerevisie	9.20	10	28	0.13	3.87	11.69
S. pombe	27.63	8	6	0.24	2.80	159.21
Twitter	33.00	5	1	0.57	4.59	160.90
Facebook	43.69	8	1	0.62	3.69	1197.33
Power grid western states	2.67	46	1	0.11	18.99	0.40

for biological networks [13]. Surprisingly, we observed that a "rewiring" model, proposed by [18]; can generate networks with a network community profile that closely resembles profiles of biological networks. The rewiring model works as follows. Starting with the original network we randomly select pairs of edges and switch their nodes. By doing this many times, we obtain a random graph with the same degree sequence as the original one.

We show the NCPs with rewiring in Fig. 4 for biological networks, and in Fig. 5 for the two social networks and the power-grid network. We observe that the NCPs for the rewired networks behave similar to those for the original biological networks. On the other hand, the behavior of the NCPs for the rewired social networks and the power-grid network is quite different from their original counterparts. This reinforces once more the fact that the internal structure of communities in biological networks is very different from that observed in social and other complex networks.

VII. Other differences between biological and other networks

We applied Spearman's rank correlation to determine the relation between the centrality measures used in the biological and social networks. We obtained the Spearman's rank correlation between three centrality measures, betweenness, closeness, and degree for all the biological networks, and several social and complex networks. The correlations for each comparison — betweenness (Bet)/closeness (Clo), Bet/degree (Deg), and Deg/Clo — are presented on Table III and IV. We can see that all networks exhibit a strong correlation between degree and

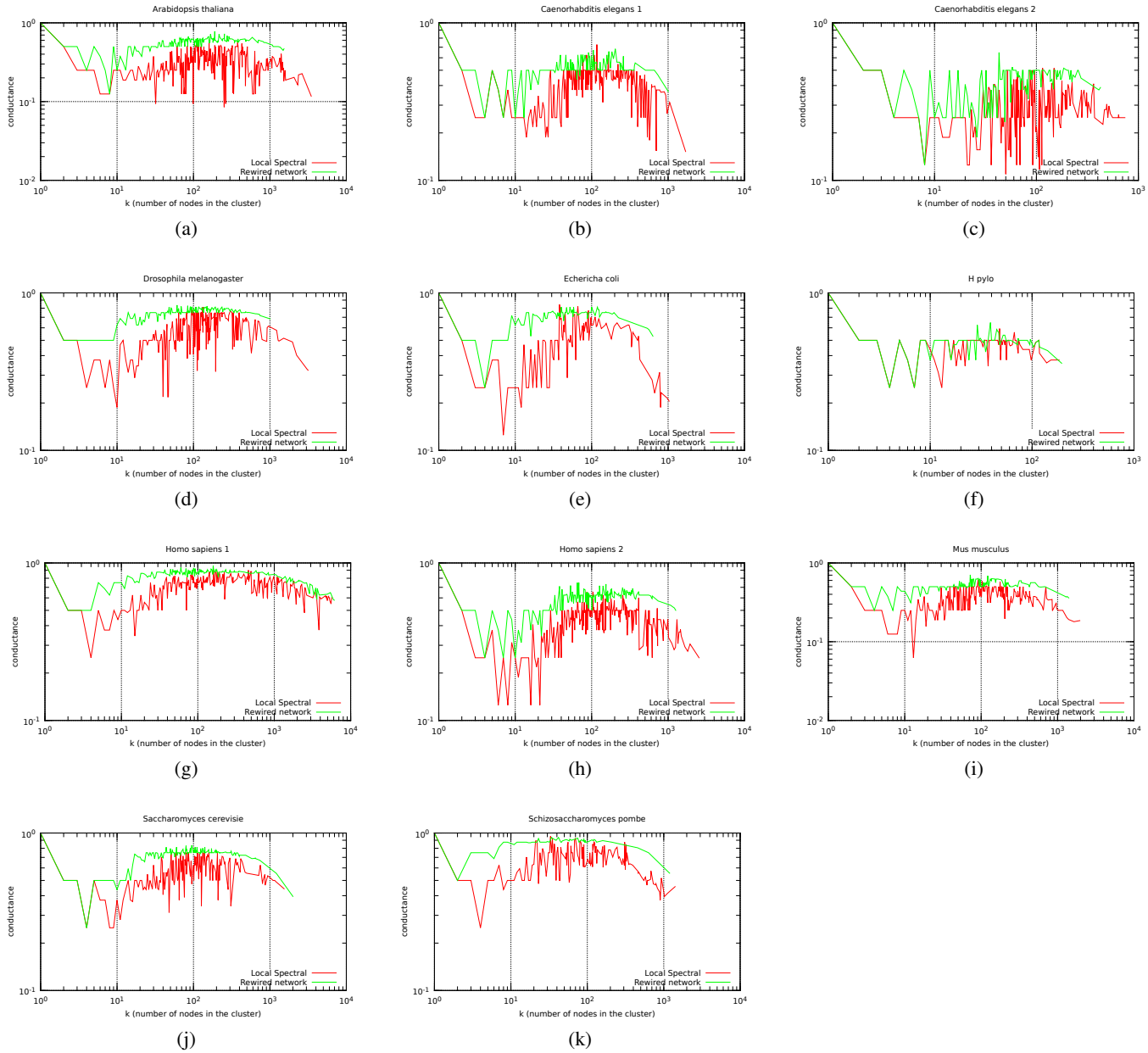
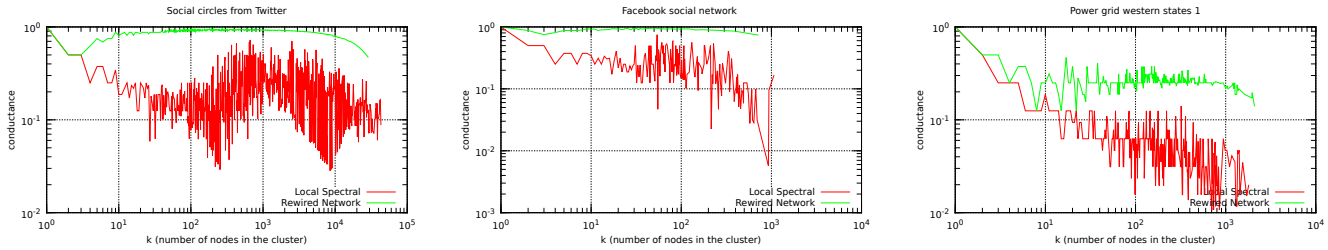


Fig. 4: Network community profiles of biological networks (red/dark) and their rewired (green/light) networks. The profiles of the original networks and their rewired counterparts exhibit a similar nature. This is not the case for social and other complex networks.

betweenness centrality. However, we observe that the biological networks show significantly more correlation between degree and betweenness, and between closeness and betweenness (Fig. 6). This is quite interesting and suggests once more that the structure of these two families of networks is quite different, in contrast to the often held belief that they are pretty much the same in terms of structure.

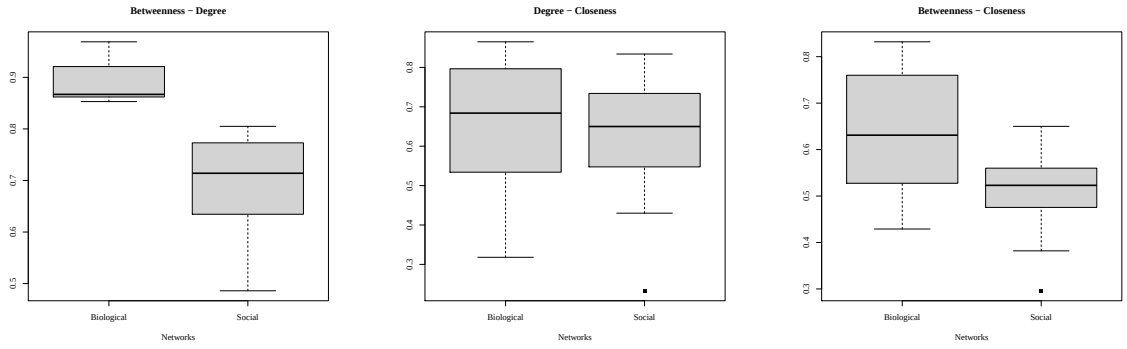
VIII. Conclusions

We presented an empirical study on the fine-grained structural differences between biological networks (namely protein interaction networks) and social and other types of networks (around 100 or more). We revealed surprising differences in terms of the network community profile and correlations of centrality



(a) (b) (c)

Fig. 5: Network community profiles (red/dark) compared to profiles of rewired networks (green/light). The profiles of the rewired networks are different from those of the original networks. Recall, that for biological networks, we observe the opposite.



(a) (b) (c)

Fig. 6: Comparison of Spearman's rank correlations between biological networks and social networks. We observe that the degree and betweenness correlation is significantly more pronounced for biological networks than for social networks (a). Regarding the correlation between degree and closeness, both families of networks exhibit a similar behaviour, with the median for biological networks being slightly higher than the median for social networks (b). Finally, betweenness and closeness correlation is in general higher for biological networks (c).

TABLE III: Spearman correlation between centrality measures for biological networks.

Networks	Clo/Bet	Deg/Bet	Deg/Clo
A. thaliana	0.429	**0.863**	0.318
C. elegans 1	0.484	**0.926**	0.500
C. elegans 2	0.535	**0.963**	0.568
D. melanogaster	0.832	**0.909**	0.865
E. coli 1	0.736	**0.916**	0.848
H. pylo	0.771	**0.969**	0.794
H. sapiens 1	0.594	**0.852**	0.684
H. sapiens 2	0.631	**0.864**	0.603
M. musculus	0.520	**0.861**	0.475
S. cerevisie	0.804	**0.859**	0.799
S. pombe	0.749	**0.867**	0.790

TABLE IV: Spearman correlation between centrality measures for social and complex networks.

Networks	Clo/Bet	Deg/Bet	Deg/Clo
Coauthor ships in science	0.382	0.486	**0.627**
AstroPhysics collaboration 1	0.650	0.714	**0.834**
AstroPhysics collaboration 2	0.562	0.645	**0.748**
Energy Physics, Phenomenology	0.523	0.624	**0.720**
Energy physics, Citation	0.472	0.596	**0.828**
Energy Physics, Theory	0.615	**0.805**	0.650
Condense Matter collaboration	0.558	**0.721**	0.698
R&Quantum Cosmology collab.	0.553	**0.676**	0.589
Enron email	0.516	**0.758**	0.506
Social circles: Facebook	0.479	**0.788**	0.430
Power grid western states 1	0.296	**0.804**	0.233

measures. More specifically, we showed that the best community size in terms of community conductance is at about a size value of ten, and this holds across almost all the available protein networks of a reasonable size.

Such a community size is an order of magnitude lower than that for social and other networks.

On the other hand, the shape of NCPs for both biological and social networks is quite similar; they initially slope downward, then upward. This behaviour is different from that of other networks (neither biological nor social). As future work, we would like to extend our experiments to biological networks of other types, and examine a wider range of network measures at fine levels of granularity.

REFERENCES

[1] Stanford network analysis project. J.Leskovec. http://snap.stanford.edu/index.html.

[2] Dana-farber cancer institute and harvard medical school: *Worm Interactome Database*, 2014.

[3] R. Andersen, F. Chung, and K. Lang. Local graph partitioning using pagerank vectors. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '06, pages 475–486, Washington, DC, USA, 2006. IEEE Computer Society.

[4] A.-L. Barabasi and Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.

[5] U. Brandes and T. Erlebach. *Network Analysis: Methodological Foundations (Lecture Notes in Computer Science)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

[6] M. Girvan and M. E. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[7] K.-I. Goh, E. Oh, B. Kahng, and D. Kim. Betweenness centrality correlation in social networks. *Physical Review E*, 67(1):017101, 2003.

[8] R. Kannan, S. Vempala, and A. Vetta. On clusterings: Good, bad and spectral. *J. ACM*, 51(3):497–515, 2004.

[9] D. Koschützki and F. Schreiber. Comparison of centralities for biological networks. In *German Conference on Bioinformatics*, pages 199–206, 2004.

[10] J. Lee, S. P. Gross, and J. Lee. Improved network community structure improves function prediction. *Scientific reports*, 3, 2013.

[11] J. Leskovec, J. M. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *KDD*, pages 177–187, 2005.

[12] J. Leskovec, J. M. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *TKDD*, 1(1), 2007.

[13] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6(1):29–123, 2009.

[14] L. Licata, L. Briganti, D. Peluso, L. Perfetto, M. Iannuccelli, E. Galeota, F. Sacco, A. Palma, A. Nardozza, E. Santonico, L. Castagnoli, and G. Cesareni. Mint, the molecular interaction database: 2012 update. *Nucleic Acids Research*, 40(Database-Issue):857–861, 2012.

[15] C. Lin, Y.-r. Cho, W.-c. Hwang, P. Pei, and A. Zhang. Clustering methods in protein-protein interaction network. *Knowledge Discovery in Bioinformatics: Techniques, Methods and Application*, pages 1–35, 2007.

[16] U. Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, Dec. 2007.

[17] O. Mason and M. Verwoerd. Graph theory and networks in biology. *Systems Biology, IET*, 1(2):89–119, 2007.

[18] R. Milo, N. Kashtan, S. Itzkovitz, M. E. J. Newman, and U. Alon. On the uniform generation of random graphs with prescribed degree sequences. *eprint arXiv:cond-mat/0312028*, Dec. 2003.

[19] A. Ochoa and L. Arco. Differential betweenness in complex networks clustering. In *Progress in Pattern Recognition, Image Analysis and Applications*, pages 227–234. Springer, 2008.

[20] G. A. Pavlopoulos, M. Secrier, C. N. Moschopoulos, T. G. Soldatos, S. Kossida, J. Aerts, R. Schneider, P. G. Bagos, et al. Using graph theory to analyze biological networks. *BioData mining*, 4(1):10, 2011.

[21] N. Pržulj. Protein-protein interactions: Making sense of networks via graph-theoretic modeling. *Bioessays*, 33(2):115–123, 2011.

[22] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. *PNAS*, 101(9):2658–2663, 2004.

[23] R. Rousseau and L. Zhang. Betweenness centrality and q-measures in directed valued networks. *Scientometrics*, 75(3):575–590, 2008.

[24] S. E. Schaeffer. Survey: Graph clustering. *Comput. Sci. Rev.*, 1(1):27–64, Aug. 2007.

[25] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000.

[26] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. Biogrid: a general repository for interaction datasets. *Nucleic Acids Research*, 34(Database-Issue):535–539, 2006.

[27] J. Wang, M. Li, Y. Deng, and Y. Pan. Recent advances in clustering methods for protein interaction networks. *BMC genomics*, 11(Suppl 3):S10, 2010.

[28] D. J. Watts and S. H. Strogatz. Collective dynamics of small-worldnetworks. *nature*, 393(6684):440–442, 1998.

[29] I. Xenarios, E. Fernandez, L. Salwínski, X. J. Duan, M. J. Thompson, E. M. Marcotte, and D. Eisenberg. Dip: The database of interacting proteins: 2001 update. *Nucleic Acids Research*, 29(1):239–241, 2001.

[30] S.-H. Yook, Z. N. Oltvai, and A.-L. Barabási. Functional and topological characterization of protein interaction networks. *Proteomics*, 4(4):928–942, 2004.

[31] H. Yu, P. M. Kim, E. Sprecher, V. Trifonov, and M. Gerstein. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS computational biology*, 3(4):e59, 2007.

[32] C. T. Zahn. Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Trans. Comput.*, 20(1):68–86, Jan. 1971.