DYNAMICS DAYS SOUTH AMERICA 2010
International Conference on Chaos and Nonlinear Dynamics

INPE – National Institute for Space Research
São José dos Campos – SP – Brazil – July 26-30, 2010

IDENTIFYING ABNORMAL NODES IN PROTEIN-PROTEIN INTERACTION NETWORKS

Bilzã Araújo, Francisco A. Rodrigues, Lilian Berton, Jean Huertas, Thiago C. Silva, Liang Zhao

Institute of Mathematics and Computer Science, University of São Paulo, São Carlos, Brazil,
{bmarques, francisco, lberton, jhuertas, thiagoch, zhao}@icmc.usp.br

keywords: Complex networks, outlier detection, stochastic dynamics, protein-protein interaction networks.

1. INTRODUCTION

In nature and science, data that differs from the mean usually have great relevance. Considering the visual attention of animals, for example, the visual stimulus that call more attention are not the homogeneous part of the visual scene, but the changing of signal intensity, such as the contours of the objects [1]. This kind of data is typically referred to outlier or anomaly [2]. There are several definitions of outlier in the literature. According to [3], an outlier is an object that appears to deviate significantly from other members of the same sample set. Another way to define outlier is that an outlier is an observation that seems to be inconsistent with the rest of the data set [4]. In a data set, outliers occur for several reasons [2]. Some of them are human errors and failures in measuring. Instances from these events usually are considered as noises or impurities in the data set and generally are pruned or eliminated. However, often outliers are relevant data and may contribute significatively to the information in a data set. Intrusion into a monitoring system, novelty finding in image analysis and fraud identification in systems for granting credit are some examples in which outliers represent relevant information. In these cases their identification is an important task.

Over the last decade there has been an increased interest in network research, with the focus shifting away from the analysis of single small graphs and their properties to considers large-scale statistical properties. The first exhaustive and rigorous study on large networks was made by Erdös and Rényi [5], who gave it the name "random graph". In 1998, Watts and Strogatz [6] discovered the small-world property in large scale networks. In 1999, Barabási and Albert [7] discovered that the degree distribution of many complex networks obeys power law, the so called scale-free networks. After these main findings, many researches have been conducted on complex networks and currently it turns out to be a well defined research area. As a consequence, a large amount of research results have been reported [8]. However, little has been studied about the identification of outlier nodes in complex networks.

The identification of outlier nodes in a network can be done considering several criteria. For example, if we consider only the connectivity of the nodes in a scale-free network, the most outlier nodes would be the hubs, because their number of connections is generally much higher than the average connectivity of the network. We analyze the problem of identifying outliers in a network structure and propose an outlier measure by using the random walk distance measure [9] [10] [11] [12] and a dissimilarity index between pairs of vertices. The method determines a "view" to the whole network for each node (the distance measure) and infers that outliers are those nodes whose view differs significantly from majority of the nodes. This perspective, incorporate both local and global information of the network and can give more general outlier detection results.

2. OUTLIER DETECTION METHOD

Considering a network with N nodes where the set of nodes is denoted by $V = \{1, 2, \ldots, N\}$ and the edges between pairs of vertices are represented in the generalized adjacency matrix A, the probability of a particle moving from one node i to a node j in one iteration is given by $P_{ij} = A_{ij} / \sum_{l=1}^{N} A_{il}$. The random walk distance, which is the average number of steps required for a particle moving through the network from i to j, can be calculated by

$$d_{i,j} = P_{ij} + \sum_{m=1}^{\infty} (m+1) \sum_{k_1 \neq j; \ldots; k_m \neq j} P_{ik_1} P_{k_1 k_2} \ldots P_{k_m j}, \tag{1}$$

where m is the number of steps between i and j - the equation considers paths of all sizes. Since the transfer matrix P satisfies the characteristics of a Markov irreducible transition matrix [13], we can apply the convergence theorem to a vector fixed point, where $X = P^n X$ when $n \to \infty$, and obtain the algebraic equation of the distance,

$$[I - B(j)] \begin{pmatrix} d_{1j} \\ d_{2j} \\ \vdots \\ d_{Nj} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \tag{2}$$

where I is the identity matrix $N \times N$ and $B(j)$ is the transfer

matrix P except that $B_{lj} = 0$ for all $l \in V$. Solving this system for all $j \in V$, we obtain the random walk distance matrix. We chose to normalize this matrix into the interval [0, 1]. We obtain the dissimilarity index which measures how different two nodes are viewed by the other all in a network by

$$\Lambda(i,j) = \frac{\sqrt{\sum_{k\neq i,j}^{N}(d_{ki} - d_{kj})^2}}{(N-2)}. \quad (3)$$

Calculating $\Lambda(i,j)$ for every pair $\langle i,j \rangle$ we obtain the symmetric matrix of dissimilarity. We can identify the most unique nodes in the network by establishing an outlier score of each node i. We chose the sum of the dissimilarity index to provide this score, as the following equation,

$$\sigma(i) = \frac{1}{\sqrt{N}} \sum_{l=1}^{N} \Lambda(i,l), \quad (4)$$

where the sum is divided by the square root of N keeping the score close in the interval [0, 1]. From the score $\sigma(i)$, we ranked the set V in descending order, such that the first elements of the ranking have larger score. These elements are the most singular node in the network.

3. OUTLIER DETECTION RESULTS

We have applied the method to several artificial and real networks and interesting results have been obtained. Particularly, we analyzed the results of applying the method to the *Saccharomyces cerevisiae* protein-protein interaction network by using the database by Krogan et al. [14]. Once the choice of the number of outlier nodes is arbitrary, we calculated the number of proteins belonging to each biological function among a given number of outliers (Fig. 1). The proteins related to cell cycle and DNA processing and metabolism represent more than 80% of the $30th$ outliers. For larger number of outliers, the fraction tends to be proportional to their occurrence on the network. In our analysis, we can infer that the proteins related cell cycle and DNA processing and metabolism has the most different "view" through the network, being peripherals nodes in the network.

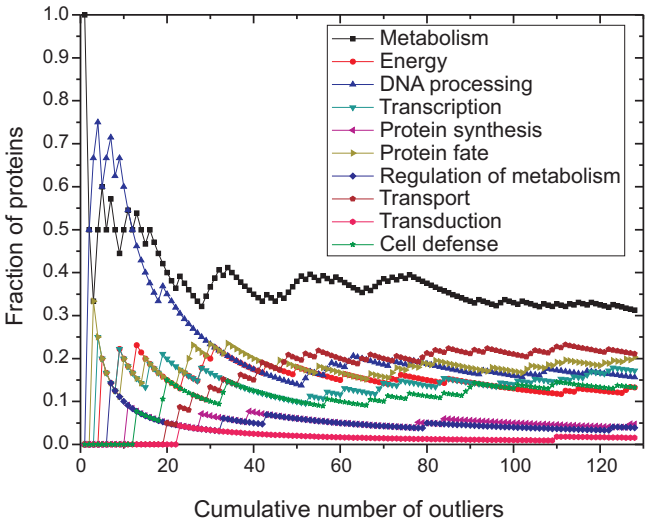

Figure 1 – (color online) The fraction of proteins of each proteins class among the first outliers.

References

[1] D. Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: Freeman, 1982.

[2] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, p. 15, 2009.

[3] F. E. Grubbs, "Procedures for detecting outlying observations in samples," *Technometrics*, vol. 11(1), pp. 1–21, 1969.

[4] V. Barnett and T. Lewis, *Outliers in Statistical Data*, 3rd ed. John Wiley & Sons, 1994.

[5] P. Erdös and A. Rényi, "On the strength of connectedness of a random graph," *Acta Math. Acad. Sci. Hungar.*, vol. 12, pp. 261–267, 1961.

[6] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, pp. 440–442, 1998.

[7] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, pp. 509–512, 1999.

[8] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45(2), pp. 167–256, 2003.

[9] V. V. Anshelevich and A. V. Vologodskii, "Random walk on a one-dimensional inhomogeneous lattice," *J. Phys. A: Math. Gen.*, vol. 15, pp. 185–197, 1982.

[10] W. Woess, *Random walks on infinite graphs and groups*. Cambridge University Press, 2000.

[11] H. Zhou, "Network landscape from a brownian particle's perspective," *Physical Review E*, vol. 67, p. 041908, 2003.

[12] J. Noh and H. Rieger, "Random walks on complex networks," *Physical Review Letters*, vol. 92, p. 118702, 2004.

[13] H. Anton and C. Rorres, *Elementary Linear Algebra with applications*, 5th ed. Wiley, 2005.

[14] N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. Tikuisis *et al.*, "Global landscape of protein complexes in the yeast Saccharomyces cerevisiae," *Nature*, vol. 440, no. 7084, pp. 637–643, 2006.