# Comparing Networks from a Data Analysis Perspective

Wei Li and Jing-Yu Yang

Department of Computer Science, Nanjing University of Science and Technology, Nanjing 210094, P.R. China
liweinust@gmail.com, yangjy@mail.njust.edu.cn

**Abstract.** To probe network characteristics, two predominant ways of network comparison are global property statistics and subgraph enumeration. However, they suffer from limited information and exhaustible computing. Here, we present an approach to compare networks from the perspective of data analysis. Initially, the approach projects each node of original network as a high-dimensional data point, and the network is seen as clouds of data points. Then the dispersion information of the principal component analysis (PCA) projection of the generated data clouds can be used to distinguish networks. We applied this node projection method to the yeast protein-protein interaction networks and the Internet Autonomous System networks, two types of networks with several similar higher properties. The method can efficiently distinguish one from the other. The identical result of different datasets from independent sources also indicated that the method is a robust and universal framework.

**Keywords:** network comparison, complex networks, data analysis, graph theory.

## 1 Introduction

Complex networks are used to depict interaction patterns among units of many real systems [1,2]. With the aim of deeply understanding these systems, researchers have constructed a large number of networks standing for these systems in fields of biology, sociology, physics, etc [3,4,5]. These networks are compared in order to improve universal network models based on their common properties [1,2,6,7], or to infer the unique design principles from particular ones [8,9,10,11,12,13]. In general, we can compare networks by means of: (i) global property statistics, such as degree distribution [6], betweenness centrality [8], assortativity, and clustering coefficient [9,13]; (ii) subgraph enumerating, such as over- and under-represented motifs than in randomized networks [10], graphlet degree distribution [7], joint degree correlations [14] of subgraph, and trained subgraph feature [11].

Each type of comparison method has unique traits, some advantages and some disadvantages. For global property statistics, a network property is usually easy

J. Zhou (Ed.): Complex 2009, Part II, LNICST 5, pp. 1907–1916, 2009.

of compute and effectively depicts a certain attribute of the network. Nevertheless, it could be argued that one single selected property is not sufficient for large network comparison [5,10,11]. Two networks even with widely different substructures sometimes exhibit some similar global properties. Furthermore, it is hard to tell how many and which properties are sufficient to distinguish several arbitrary networks. In order to enhance expressing ability of the global statistic, one creative strategy is synthesizing – first is to perform many measurements, then use pattern recognition tools to extract refined features as comparing criterion [5]. Another strategy is subdividing – to extend the network property on role-based description which can take into consideration the modular structure of the network and depict the network in a more precise manner [12]. Unfortunately, these advanced steps need considerable computational resources. Using subgraph enumerating, the subgraph-based methods can provide abundant local information and reveal more of the essential structure of studied networks. However, for large networks, costly computation makes subgraph enumeration practically infeasible. Finding an informative and efficient method for network comparison still remains a challenge.

Here, we propose a method called "node projection" from the data analysis perspective. The method includes tow steps: first, the node projection method projects each node of the studied network as a data point in a high-dimensional space, thus the original network is linked with a particular data distribution; second, with the help of a sophisticated tool - principal component analysis (PCA) [5,15], the dispersion information of the PCA projections of generated data distribution is used to measure the original network. Experimentally, we apply the current method to the yeast protein-protein interaction (PPI) networks and to the Internet at Autonomous Systems (AS) level, two types of networks with several similar global properties [12,16]. The method can differentiate one type from the other efficiently. In addition, the identical result based on independent datasets further indicated that the present method is robust.

## 2 Methods

As described above, there are two steps, "Node Projection" and "PCA Projection", in our method. First, nodes of the studied network will be projected as data points in a high-dimensional measure space. Second, from the viewpoint of data analysis, PCA is used to analysis the generated data cloud. See Fig. 1 for details. We call the whole method also "Node Projection" for the step one dramatically converted the network analysis problem into a data analysis problem.

### 2.1 Step of "Node Projection"

In order to associate each node of the network with a high-dimensional point, the node projection method assesses nodes using a distance-based quantity called "global influence", which quantifies the influence of each node on the whole network. Formally, if the distance (graph theory distance, i.e. the minimum length

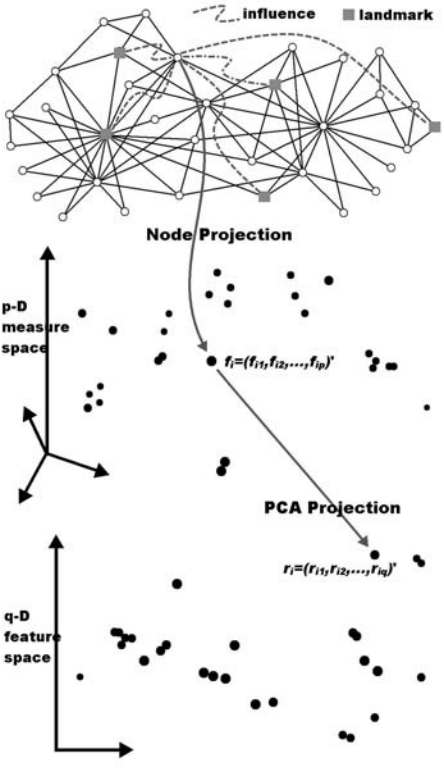


**Fig. 1.** Schematic illustration of the node projection method. First, all nodes of the original network are projected into a $p$-D measure space. The node $i$ is evaluated as $\boldsymbol{f}_i = (f_{i1}, f_{i2}, \ldots, f_{ip})'$ according to its influences on $p$ landmarks. Second, PCA is used to reduce dimensionality and extract dominant structure of the cloud of data points. And, the corresponding data point of node $i$ is $\boldsymbol{r}_i = (r_{i1}, r_{i2}, \ldots, r_{iq})'$ in the $q$-D feature space. On the basis of its distribution patterns, the generated data cloud provides unique information for network analysis.

of the paths connecting nodes) between nodes $i$ and $j$ is $d_{ij}$, the influence coefficient of one node on the other is defined as $f_{ij} = \max\{I - d_{ij}, 0\}$, where $I$ is a pre-defined parameter. (Note, we can control the parameter $I$ to obtain node influence measurement of an appropriate precision and computation for large networks. In this work, $I$ is trivially higher than network diameters.) Then, with $p$ randomly chosen vertices as landmarks, and arrange all influence values of the node $i$ on these landmarks in a certain order, a corresponding influence vector $\boldsymbol{f}_i = (f_{i1}, f_{i2}, \ldots, f_{ip})'$ is obtained. (Reasonably, the influence of the node on a particular landmark can be approximately regarded as the influence of the node on the local part of the network which the landmark belongs to. Therefore, the vector $\boldsymbol{f}_i$ can also be regarded as a coarse grained measure for the "global influence" of the node $i$ on whole network. Now, the node $i$ is projected as a $p$-D data point $\boldsymbol{f}_i$.) After each node obtains a comparable "global influence" vector based

on the same set of landmarks, the original network is linked with a particular data distribution. And node similarity [17] is preserved in the data distribution (Intuitively, similar nodes which share many of the same neighbors will be associated with close vectors.) In what follows, we will analyze the original network based on its corresponding generated data distribution.

### 2.2 Step of "PCA Projection"

Given all vectors of $n$ nodes of the network, the generated data point cloud can be quantified by the node measure matrix $\boldsymbol{F} = (\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_n)$ ($\boldsymbol{F}$ is $p \times n$, corresponding to $p$ landmarks and $n$ nodes, where each column is a global influence vector of a node.) In order to extract dominant structure of the data distribution to characterize the original network, PCA is utilized to process the node measure matrix $\boldsymbol{F}$.

We first normalize $\boldsymbol{F}$ in the measure space $\mathbb{R}^p$ and obtain $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ (mean of each row is 0 and standard deviation is 1.) Then the covariance matrix of $\boldsymbol{X}$ is $\boldsymbol{C} = \boldsymbol{X}\boldsymbol{X}'$. According to the singular value decomposition (SVD) theorem [18], $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}'$, where $\boldsymbol{U}, \boldsymbol{V}$ are orthogonal matrices ($\boldsymbol{U}\boldsymbol{U}' = \boldsymbol{U}'\boldsymbol{U} = \boldsymbol{I}_p, \boldsymbol{V}\boldsymbol{V}' = \boldsymbol{V}'\boldsymbol{V} = \boldsymbol{I}_n$) and $\boldsymbol{\Sigma} = [\boldsymbol{\Sigma}_p, \boldsymbol{0}]$, $\boldsymbol{\Sigma}_p = diag(\sigma_1, \sigma_2, \ldots, \sigma_p)$, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$. Therefore,

$$\boldsymbol{C} = \boldsymbol{X}\boldsymbol{X}' = (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}')(\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}') = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}'$$

where $\boldsymbol{\Lambda} = (\boldsymbol{\Sigma})^2 = diag(\lambda_1, \lambda_2, \ldots, \lambda_p)$, $\lambda_i = {\sigma_i}^2$ and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Choosing first $q$ columns of $\boldsymbol{U}$ (denoting as $\hat{\boldsymbol{U}}$, i.e. $q$ eigenvectors of the covariance matrix corresponding to $q$ largest eigenvalues) as axes of the low-dimensional space and projecting $\boldsymbol{X}$ on them, we get approximate representations of nodes $\boldsymbol{R} = \hat{\boldsymbol{U}}'\boldsymbol{X} = (\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_n)$ in a $q$-D ($q \ll p$) space. And for

$$\boldsymbol{R}\boldsymbol{R}' = (\hat{\boldsymbol{U}}'\boldsymbol{X})(\boldsymbol{X}'\hat{\boldsymbol{U}}) = \hat{\boldsymbol{U}}'(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}')\hat{\boldsymbol{U}} = \boldsymbol{\Lambda}_q$$

where $\boldsymbol{\Lambda}_q = diag(\lambda_1, \lambda_2, \ldots, \lambda_q)$, $\boldsymbol{R}$ is uncorrelated and with most structural features of the original data distribution.

**PCA Measurements.** To characterize the studied network, we use the PCA measurements "variance contribution ratio (VCR)" and "cumulative contribution ratio (CCR)" in this work [15]. Given an arbitrary data distribution, the VCR of the $n$th PCA component is $\rho_n = \lambda_n / \Sigma\lambda$ (where $\lambda_n$ is the $n$th eigenvalue of the covariance matrix $\boldsymbol{C}$ and $\Sigma\lambda$ is the sum of all eigenvalues) and the CCR of first $n$ PCA components is $\eta_n = \rho_1 + \rho_2 + \cdots + \rho_n$. For VCR and CCR are relative measurements, they are reliable in despite of different sizes of compared networks.

### 2.3 Datasets

The datasets used in this work are the yeast PPI networks and the Internet AS networks. They are four yeast PPI networks from four independent data sources

and six Internet AS networks from three different sources. For yeast datasets [19], 'YJ' denote a compositive dataset and 'YC', 'YI', 'YU' denote the CCSB-Y2H, Ito-core, Uetz sets, respectively. For Internet [20], the data sources are the Route Views Project, UCLA Internet Research Lab (IRL) and the DIMES project. We use two Internet datasets from each source and they are in years 2005 and 2007. Say, 'IR5' and 'IR7' stand for two samples from the Route Views Project in 2005 and 2007, respectively. Similarly, 'II5' and 'II7' are IRL data, 'ID5' and 'ID7' are from the DIMES project, in 2005 and 2007. Please note that, for divided network components have no influence on each other, the node projection method only considers the largest connected component of these networks. See the table for basic properties of them.

**Table 1.** Basic statistics of the largest connected components of yeast PPI networks and the Internet AS networks. The properties are: node number $n$; edge number $m$; mean degree $z$; mean node-node distance $l$; exponent of degree distribution $\alpha$; exponent of degree correlation $\beta$; clustering coefficient $cc$; modularity $Q$.

| Network | $n$ | $m$ | $z$ | $l$ | $\alpha$ | $\beta$ | $cc$ | $Q$ |
|---|---|---|---|---|---|---|---|---|
| YJ | 1,458 | 1,948 | 2.67 | 6.81 | 3.03 | 0.58 | 0.07 | 0.82 |
| YC | 9,64 | 1,487 | 3.09 | 5.37 | 2.68 | 0.54 | 0.06 | 0.73 |
| YI | 411 | 497 | 2.42 | 6.18 | 2.74 | 0.53 | 0.04 | 0.79 |
| YU | 473 | 544 | 2.30 | 7.53 | 3.11 | 0.42 | 0.02 | 0.85 |
| IR5 | 19,513 | 41,867 | 4.29 | 3.77 | 2.08 | 0.43 | 0.26 | 0.62 |
| IR7 | 25,050 | 51,856 | 4.14 | 3.89 | 2.08 | 0.46 | 0.22 | 0.65 |
| II5 | 21,838 | 83,686 | 7.66 | 3.52 | 1.94 | 0.34 | 0.43 | 0.56 |
| II7 | 30,214 | 146,862 | 9.72 | 3.44 | 1.98 | 0.32 | 0.45 | 0.56 |
| ID5 | 13,638 | 31,622 | 4.64 | 3.53 | 2.10 | 0.41 | 0.36 | 0.58 |
| ID7 | 19,452 | 48,222 | 4.96 | 3.51 | 2.13 | 0.47 | 0.39 | 0.58 |

It was reported that the yeast PPI network and the Internet AS network were similar of several higher topological properties. First, they both exhibit nontrivial correlation property (both obey $\langle K_1 \rangle_{K_0} \propto K_0^{-0.5}$, where $\langle K_1 \rangle_{K_0}$ is the average degree of neighbors of nodes with the degree $K_0$ [16]. As we only consider the largest network component, most exponent of degree correlation of our results are near 0.5. See $\beta$ column of the table). Second, the yeast networks and Internet networks both have strong modular structure (for all $Q > 0.5$). Third, treated by a module-based analysis, they further exhibit similar role-to-role connectivity pattern [12].

## 3 Results and Discussion

We apply the node projection method on yeast PPI networks and Internet AS networks. From this perspective of data analysis, we first observe the effect of degree distribution and the modules on network structure based on the comparison of real networks and their two random ensembles. Then, from this data

analysis viewpoint, we also compare the distribution profile of generated data clouds and reasonably divide all real networks into two classes. In addition, we discuss the stability and complexity of the node projection method.

### 3.1 Comparison of Real Networks and Random Ensembles

It was previously reported that the degree distribution was an important factor in network features' forming [21,22]. And it was further pointed out that the degree distribution plus the modular structure accounted for most internal structure [12]. To evaluate the effect of the degree distribution and the modules on network structure, we compare each studied network with its two ensembles of random versions.

**Two Random Ensembles of Networks.** One is purely random ensemble and the other is module-kept random ensemble. The first ensemble is with the same degree sequence as the original network by recurrently exchanging one endpoints of each two random edges [16,21]. The second ensemble restricts the swap process in the same module to keep both degree sequence and modular structure unchanged as the initial network [12].

The comparison between real networks and their two ensembles are quantified using two parameters: $\eta_n/\eta r_n$ and $\eta_n/\eta m_n$, where $\eta_n$ is the CCR of first $n$ PCA components of the data distribution generated from real networks while $\eta r_n$ and $\eta m_n$ are the average CCR of the corresponding purely random ensemble and the module-kept random ensemble, respectively. Figures 2(a) and 2(b) show that the two ratios $\eta_n/\eta r_n$ and $\eta_n/\eta m_n$ of all studied networks are approximate 1. Moreover, $\eta_n/\eta m_n$ is more near 1 than $\eta_n/\eta r_n$ (when more data distribution components are counted in, say $n >= 5$, the trend is more clear.) Our results confirm and clarify previous conclusions from a new view: For networks of similar degree distribution (here, the original network and its purely random ensemble) generate data point clouds of similar outlines (Fig. 2(a)), it indicates degree

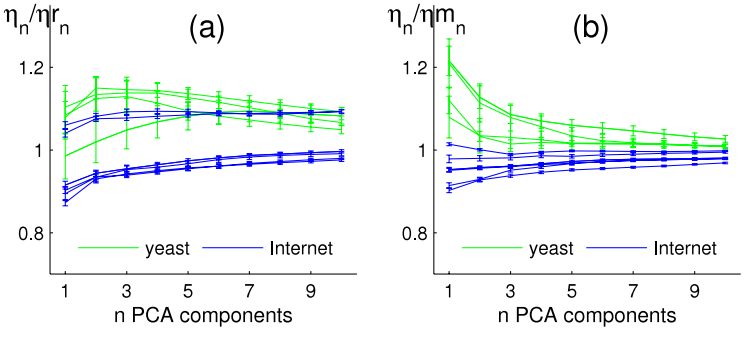


**Fig. 2.** Comparison of real networks and their two random ensembles. (a) $\eta_n/\eta r_n$ gives the ratio of PCA measurements CCR of the generated data distribution to its purely random ensemble. (b) $\eta_n/\eta m_n$ gives the ratio of CCR of the generated data distribution to its module-kept random ensemble. The error bars represent the standard deviation.

distribution is essentially influential on network features [21,22]; The outlines of generated data clouds of the original network and its module-kept ensemble are nearly identical (Fig. 2(b)) indicates that once the modular structure is fixed with degree distribution, real networks are almost determinate without additional internal structure [12].

### 3.2 Comparison between Real Networks

We also compare the yeast PPI networks with the Internet AS networks based on the variance contribution profile (VCP), which reflects over- and under- variance contribution of each PCA component than the module-kept random version. The VCP is described by the Z-score [10,12,16,21] :

$$Z_n = (\rho_n - \langle \rho m_n \rangle)/std(\rho m_n)$$

where $\rho_n$ is the VCR of the $n$th PCA component of the generated data distribution, while $\langle \rho m_n \rangle$ and $std(\rho m_n)$ are the mean and standard deviation of the VCR of the corresponding module-kept random ensemble. And the VCP is the normalized vector of Z-score:

$$VCP_n = Z_n/(\Sigma {Z_n}^2)^{1/2}$$

The VCP of the first ten PCA components (they contain dominant structure information of the generated data cloud, say more than 85% variance contribution) of the yeast PPI networks and the Internet AS networks is present in Fig. 3(a). In spite of they are two types of networks with similar features, their VCP are distinctly dissimilar.

Figure 3(a) shows that subordinate components of the generated data distributions of the yeast PPI network are less significant than the Internet AS network. The result shows that the yeast PPI networks are more regular and more modular than the Internet, and with less local detail. It may be interpreted as follows: To PPI network, the propagation of deleterious perturbation is vital. A strong module structure provides protection against the deleterious

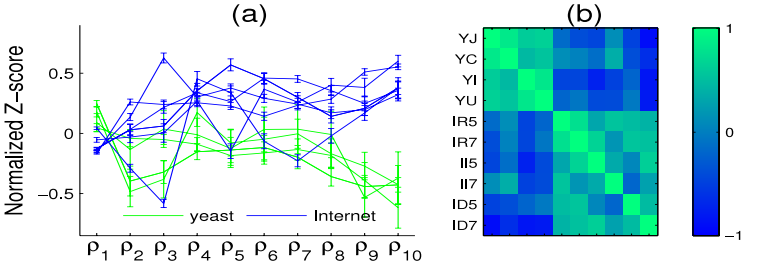


**Fig. 3.** Comparison of the yeast PPI networks and the Internet AS networks. (a) shows the variance contribution profile (VCP) of the yeast PPI networks and the Internet AS networks. The error bars represent the standard deviation. (b) is the correlation coefficient matrix of the VCP in (a).

perturbation at the topology level. Affected by the evolutionary pressure, the PPI networks are composed of biological function modules with little interaction among them, thus less detail. In contrast to the Internet case, complex local interconnections, detail part of the network, provide optional paths for AS-level nodes to access the Internet, which is helpful to avoid congestion and promote communication. That is why subordinate components of the yeast data distribution, which correspond to detail part, show less significance than the Internet's. In Fig. 3(b), the correlations between networks clearly separate the yeast PPI networks and the Internet AS networks. Furthermore, high correlations between independent datasets of the same type network illuminate the node projection method being robust.

### 3.3 Stability and Complexity

Until now, an important question we should discuss is "how many landmarks are needed at least to ensure system stability?" The experiments on datasets of yeast and the Internet AS both indicated that to choose about 200 vertices as landmarks can export acceptably stable result. To be clarity, the experiments on the Internet 1997 dataset was designed to demonstrate the stability trend of increasing vertices (See Fig. 4). Furthermore, the time complexity of the node projection method is $\mathcal{O}(p^3 + p \times |N|)$, where $p$ is the number of landmarks (constant, say 200) and $|N|$ is the number of network nodes. Therefore, the node projection method is with approximately linear complexity of network size.

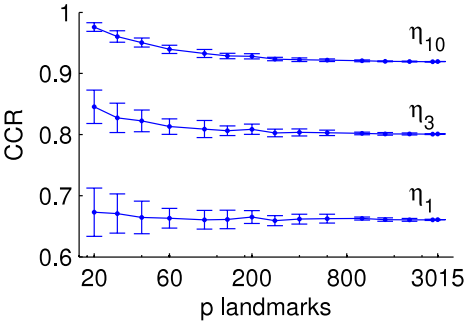


**Fig. 4.** Stability of node measure system. The data set is the Internet AS network in year of 1997. The number of landmarks $p$ increased from 20 to all 3015 vertices. For every $p$, the node projection method ran 100 times to obtain averages and deviations of $\eta_1$, $\eta_3$ and $\eta_{10}$. The X axis is log scale.

## 4 Conclusions

In conclusion, we propose a novel method named "node projection" to compare complex networks from the data analysis perspective. As the studied network is linked with a generated data distribution, we can measure the data distribution

to characterize the original network. This framework provides network comparison with efficiency and convenience by using the sophisticated data analysis tool. The experiments on yeast PPI networks and the Internet AS networks indicate that the node projection can compare networks precisely and robustly. In future work, more data analysis tools can be adopted in this framework, and it can be developed into a universal framework. Interpreting the obtained result in physical sense is another difficult and promising challenge.

**Acknowledgments.** We thank anonymous reviewers for their helpful comments. This work was supported by the Chinese National Natural Science Foundation (No. 60632050).

## References

1. Watts, D.J., Strogatz, S.H.: Collective Dynamics of 'Small-World' Networks. Nature (London) 393, 440–442 (1998)
2. Barabási, A.L., Albert, R.: Emergence of Scaling in Random Networks. Science 286, 509–512 (1999)
3. Newman, M.E.J.: The Structure and Function of Complex Networks. SIAM Rev. 45, 167–256 (2003)
4. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.U.: Complex Networks: Structure and Dynamics. Phys. Rep. 424, 175–308 (2006)
5. Costa, L., Da, F., Rodrigues, F.A., Travieso, G., Villas Boas, P.R.: Characterization of Complex Networks: a Survey of Measurements. Adv. Phys. 56, 167–242 (2007)
6. Amaral, L.A.N., Scala, A., Barthélémy, M., Stanley, H.E.: Classes of Small-World Networks. Proc. Natl. Acad. Sci. U.S.A 97, 11149–11152 (2000)
7. Pržulj, N.: Biological Network Comparison Using Graphlet Degree Distribution. Bioinformatics 23, 177–183 (2006)
8. Goh, K., Oh, E., Jeong, H., Kahng, B., Kim, D.: Classification of Scale-Free Networks. Proc. Natl. Acad. Sci. U.S.A 99, 12583–12588 (2002)
9. Newman, M.E.J., Park, J.: Why Social Networks are Different from Other Types of Networks. Phys. Rev. E 68, 036122 (2003)
10. Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M., Alon, U.: Superfamilies of Evolved and Designed Networks. Science 303, 1538–1542 (2004)
11. Middendorf, M., Ziv, E., Wiggins, C.H.: Inferring Network Mechanisms: The Drosophila Melanogaster Protein Interaction Network. Proc. Natl. Acad. Sci. U.S.A 102, 3192–3197 (2005)
12. Guimerà, R., Sales-Pardo, M., Amaral, L.A.N.: Classes of Complex Networks Defined by Role-to-Role Connectivity Profiles. Nature Phys. 3, 63–69 (2007)
13. McAuley, J.J., Costa, L.D.F., Caetano, T.S.: Rich-Club Phenomenon across Complex Network Hierarchies. Appl. Phys. Lett. 91, 84103 (2007)
14. Mahadevan, P., Krioukov, D., Fall, K., Vahdat, A.: Systematic topology analysis and generation using degree corrlation. In: SIGCOMM (2006)
15. Johnson, R.A., Wichern, D.W.: Applied Multivariate Statistical Analysis. Prentice-Hall, Englewood Cliffs (1998)
16. Maslov, S., Sneppen, K.: Specificity and Stability in Topology of Protein Networks. Science 296, 910–913 (2002)

17. Leicht, E.A., Holme, P., Newman, M.E.J.: Vertex Similarity in Networks. Phys. Rev. E 73, 026120 (2006)
18. Golub, G.H., Van Loan, C.F.: Matrix Computations. Johns Hopkins University Press, Baltimore (1996)
19. The data YJ is from CCNR-Resources, `http://www.nd.edu/~networks/resources.htm` and others, YC, YI and YU are from CCSB Interactome Database, `http://interactome.dfci.harvard.edu/S_cerevisiae/host.php`
20. The Route Views Project, `http://www.routeviews.org/`, UCLA Internet Research Lab, `http://irl.cs.ucla.edu/topology/` and the DIMES project, `http://www.netdimes.org/`
21. Maslov, S., Sneppen, K., Zaliznyak, A.: Detection of Topological Patterns in Complex Networks: Correlation Profile of the Internet. Physica A 333, 529–540 (2004)
22. Park, J., Newman, M.E.J.: Origin of Degree Correlations in the Internet and Other Networks. Phys. Rev. E 68, 026112 (2003)