Some protein interaction data do not exhibit power law statistics

Reiko Tanaka[a,*], Tau-Mu Yi[b], John Doyle[c]

[a] *Bio-Mimetic Control Research Center, RIKEN, Nagoya 223-8522, Japan*
[b] *Developmental and Cell Biology, University of California, Irvine, United States*
[c] *Control and Dynamical Systems, California Institute of Technology, United States*

Received 26 June 2005; revised 27 July 2005; accepted 17 August 2005

Available online 26 August 2005

Edited by Robert B. Russell

Abstract It has been claimed that protein–protein interaction (PPI) networks are scale-free, and that identifying high-degree "hub" proteins reveals important features of PPI networks. In this paper, we evaluate the claims that PPI node degree sequences follow a power law, a necessary condition for networks to be scale-free. We provide two PPI network examples which clearly do not have power laws when analyzed correctly, and thus at least these PPI networks are not scale-free. We also show that these PPI networks do appear to have power laws according to methods that have become standard in the existing literature. We explain the source of this error using numerically generated data from analytic formulas, where there are no sampling or noise ambiguities.

Keywords: Protein–protein interaction networks; Node degree sequence; Power law; Rank–degree plot

1. Introduction

Experimental data on protein–protein interaction (PPI) networks have been extensively gathered with the aim of acquiring a system-level understanding of biological processes [1,2]. Various statistical features of complex graphical structures have received attention, including the size of the largest connected component, the node degree distribution, the graph diameter, the characteristic path length, and the clustering coefficient. However, the feature that has attracted the most attention is the distribution of node degree (the number of links from a node) and whether or not the distribution follows a power law (linear plot on log–log scale). The degree sequences of PPI networks were claimed to follow a power law in [3], and thus PPI networks are claimed to be "scale-free" (SF) [4]. In fact, graph theorists point out that "scale-free" has not been clearly defined in the existing literature, and the results on SF graphs are largely heuristic and experimental studies with "*rather little rigorous mathematical work; what there is sometimes confirms and sometimes contradicts the heuristic results*" [5]. Nevertheless, most treatments assume that a power law node degree distribution is an important, and sometimes defining feature [4].

This paper shows that the node degree sequences of some published PPI networks are better described by an exponential function when properly plotted and analyzed. The problem with previous work is that data were plotted using frequency–degree plots, which are usually ambiguous and misleading but have become standard in much of the scientific literature. This problem can be easily avoided by plotting the same data using rank–degree plots, which are standard within statistics and parts of engineering [6]. We also illustrate the source of the errors using numerically generated data from analytic formulas, where there are no sampling or noise ambiguities. Despite these findings, we expect that power laws will appear ubiquitously in biology [7], just as they do in natural and technological systems [6,8], but that their analysis will require the use of appropriate tools [6].

2. Materials and methods

It is widely accepted that publicly available data for PPI networks represent only an approximation of the real interaction network because of the large number of false positive and false negative interactions. It has been claimed that the self-similarity features of SF networks mean that any appropriately sampled subnetwork is also SF [9]. If this were true, then perhaps the existing data are sufficient to determine whether PPI networks are SF, but this self-similar sampling claim has been disputed elsewhere [10]. Even more fundamental errors, however, may be involved in the basic claim that the publicly available PPI network data possess power law node degree sequences.

A PPI node degree data set consists of a finite sequence of integers $y = (y_1, y_2, \ldots, y_n)$, assumed here without loss of generality always to be ordered such that $y_1 \geqslant y_2 \geqslant \cdots \geqslant y_n$. Note that the rank, the number of nodes P_k with degree equal or greater than y_k, is then simply $P_k = k$. A node degree sequence y follows a power law if

$$k \approx c y_k^{-\alpha}, \tag{1}$$

where k is the rank of y_k, $c > 0$ is a constant, and $\alpha > 0$ is called the scaling index. One important feature of power laws is that for large n the sample means (and mean/median ratio) and variances (and variance/mean ratio) diverge when $\alpha < 1$ and $\alpha < 2$, respectively. Such power laws are approximately straight lines of slope $-\alpha$ on log–log plots of the rank k versus the degree y_k, and these provide a simple test for whether data satisfy (1). In contrast, an exponential

$$k \approx a \exp^{-b y_k}, \tag{2}$$

has finite and convergent sample mean and variance for all constants $a > 0$ and $b > 0$. The k versus y_k plot on a semilog scale approximates a straight line of slope of $-b$ since $\log k = \log a - b y_k$, and thus semilog plots can be used to easily check if (2) holds.

It has been standard practice in studying PPI node degrees to assume that the sequence is drawn from a random ensemble, although no coherent explanation has been offered as to why this is biologically jus-

[*] Corresponding author.
E-mail address: reiko@bmc.riken.jp (R. Tanaka).

Abbreviations: PPI, protein–protein interaction; SF, scale-free; SR, scale-rich; FYI, filtered yeast interactome; HPI, human protein-interaction

0014-5793/$30.00 © 2005 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.febslet.2005.08.024

tified. Indeed, what is known of biology suggests that it is highly nonrandom at every level of organization [11]. Nevertheless, random variable models are central to previous work and their misinterpretations are a classic source of errors. There is a large and sophisticated literature on the theory of stable laws [12], of which power laws are special cases, but we will aim for the simplest possible explanation of the origin of the most common error. A non-negative random variable X with cumulative distribution function (CDF) $F(x) = P[X \leqslant x]$, is said to follow a power law with index $\alpha > 0$ if, as $x \to \infty$,

$$P[X > x] = 1 - F(x) \approx cx^{-\alpha}, \tag{3}$$

for some constant $c > 0$, where $a(x) \approx b(x)$ as $x \to \infty$ if $a(x)/b(x) \to 1$ as $x \to \infty$. If the CDF $F(x)$ satisfying (3) is differentiable, then its derivative, the probability density function (PDF) $f(x) = \frac{d}{dx}F(x)$, satisfies

$$f(x) \approx c'x^{-(1+\alpha)}. \tag{4}$$

A log–log plot of $1-F(x)$ or $f(x)$ versus x thus approximates a straight line of slope $-\alpha$ and $-(1 + \alpha)$, respectively, for large x. Similarly, an exponential F gives an exponential f. While these relationships are true for analytic formulas, differentiation of noisy data amplifies errors, making attempts to create frequency-based data plots of $f(x)$ typically uninformative except in special cases. In the case of the node degree of a graph, the data are not just noisy but inherently discrete.

While a full exposition of these issues is well beyond the scope of a short letter, numerical experiments that are easy to reproduce illustrate the essential points. Even in the most "idealized" cases shown below using numerically generated pseudorandom data, frequency plots can mislead. Fig. 1 shows $n = 1000$ integer values numerically sampled from the distribution $P[X > x] \approx x^{-1}$ for $x \geqslant 1$ and plotted using the following MATLAB program fragment:

```
y=-sort(-floor(1./rand(1,n)));
loglog(y,1:n,'.k');
```

The `rand` function generates (pseudo-)random variables uniformly distributed on (0,1). Suppose Z is such a random variable, then $P[Z < z] = z$. Under mild technical conditions, X with $P[X > x] = h(x)$ for decreasing function $h(x)$ can be generated by $X = h^{-1}(Z)$ since then $P[X > x] = P[h^{-1}(Z) > x] = P[Z < h(x)] = h(x)$ as desired. Thus the first line generates n sorted random integers $y = (y_1, y_2, \ldots, y_n)$ with $y_1 \geqslant y_2 \geqslant \cdots \geqslant y_n$ sampled from $P[X > x] \approx x^{-1}$. The upper (black dots) cumulative or rank–degree plot in Fig. 1 shows that a correct estimate of $\alpha \approx 1$ (solid red line) can be obtained by inspection simply from a plot of the rank k versus the degree y_k. Note also that the false claim that $\alpha \approx 2/3$ (green dotted line, derived from the frequency–degree plot below) is also clearly seen to be incorrect by a large margin.

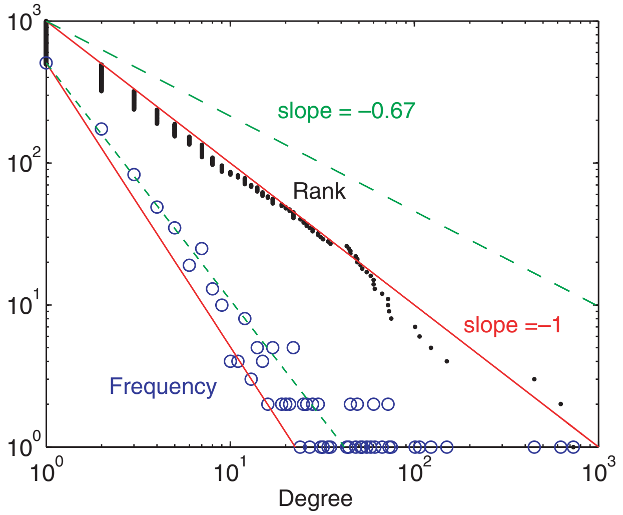

Fig. 1. Rank–degree (upper black dots) and frequency–degree (lower blue circles) plots for integer data numerically sampled from the random variable following a power law $P[X > x] \approx x^{-1}$. Rank–degree plot correctly shows the slope of $\alpha = -1$ (solid red line) whereas the frequency–degree plot incorrectly leads to estimation of $\alpha \approx 0.67$ (dotted green line).

Errors arise however when frequency–degree plots such as the lower (blue circles) plot in Fig. 1 are used. A sample frequency–degree plot of the data can be generated by first creating a vector of unique values $\{u_j\}$, $1 \leqslant j \leqslant m$, from the $\{y_k\}$, and then their sample frequencies $\{f_j\}$ are counted. The smaller values of y_k (large k) have nonunique (repeated) integer values. The following is an example of a MATLAB fragment used to generate the lower part of Fig. 1:

```
u=unique(y); nu=length(u); f=0*u;
for k=1:nu, f(k)=sum(y==u(k)); end;
loglog(u,f,'bo');
```

In this case, we can analytically derive discrete equivalent to Eq. (4) as

$$\begin{aligned} p(x) = P[X = x] &= P[X > x] - P[X > x + 1] \\ &= x^{-1} - (x+1)^{-1} \\ &= x^{-1}(x+1)^{-1}, \quad x \geqslant 1 \\ &\approx x^{-2}, \quad x \gg 1, \end{aligned}$$

and thus it might appear that the true tail index (i.e., $\alpha = 1$) could be inferred from examining the frequency–degree plots. Unfortunately this is not true even in this idealized case. By fitting a straight line with slope $1 + \alpha \approx 5/3$ (dotted green line) to the frequency data, a tail index estimate of $\alpha \approx 2/3$ might appear more plausible than the correct values of $1 + \alpha \approx 2$ (solid red line). At best there is ambiguity. We know in this case however that $\alpha \approx 2/3$ is unambiguously incorrect since we generated the "data" by sampling from an analytically known distribution with $\alpha = 1$, which is confirmed by the rank plot in Fig. 1 (solid red line). Binning and smoothing the data in various ways can occasionally improve the appearance of frequency plots but not their reliability in determining α.

There are a variety of more rigorous, reliable, but unfortunately more complex methods for estimating α [13] than examining rank–degree plots, and since $\{y_k\}$ and the $\{u_k\}$ and $\{f_k\}$ have the same information, appropriate numerical tests on one can in principle be interpreted in terms of the other. The classic errors arise from misinterpretation of frequency plots in ways that do not occur with (cumulative) rank–degree plots. The latter also have additional advantages over more complex tests in that they show the raw data directly, and are also highly robust to a range of measurement errors and noise. Thus experienced readers can tell at a glance from the rank plots such as Fig. 1 if (1) is plausible and over what range it holds.

An equally serious error can occur when the data are exponential, as shown in Fig. 2. This is similar to Fig. 1 except the data are exponentially distributed with $P[X > x] \propto \exp^{-0.1x}$ for $x > 10$, generated by

```
y=-sort(floor(10*(-1+log(rand(1,n)))));
```

and the ranks and frequencies are in separate plots. This is another example of an idealized case since we can analytically derive discrete equivalents to Eq. (4) for $P[X > x] = c\exp^{-\alpha x}$ $(x > x_0)$ as

$$\begin{aligned} p(x) &= c\exp^{-\alpha x} - c\exp^{-\alpha(x+1)} = c(1 - \exp^{-\alpha})\exp^{-\alpha x} \\ &= \tilde{c}\exp^{-\alpha x} \qquad x > x_0. \end{aligned}$$

Thus both $P[X > x]$ and $p(x)$ are exponentials with the same form, yet with very different plots. Fig. 2A is a semilog rank–degree plot confirming that the data are exponentially distributed. Fig. 2B is a log–log frequency–degree plot suggesting incorrectly that the data are power law distributed with $\alpha \approx 1.75$. This kind of misuse of frequency–degree plots is remarkably common in the scientific literature and until recently, frequency plots have been used almost exclusively in the SF literature. That such errors can arise using PPI data are illustrated in Figs. 3 and 4, which is the main focus of this paper.

From among many publicly available studies on PPI networks, we used the filtered yeast interactome (FYI) data set [14] and the predicted human protein-interaction (HPI) map [15]. Much of the original data suffers from numerous false positives and false negatives, but more recent investigations have sought to refine the data. For example, the FYI data set contains high-confidence interactions for yeast, each observed by at least two different methods, thereby enriching for genuine positives. The HPI map was generated using data from seven experimental and four computationally predicted protein-interaction maps from *Saccharomyces cerevisiae* [16–21], *Drosophila melanogaster* [22] and *Caenorhabditis elegans* [23]. The idea is that a human protein

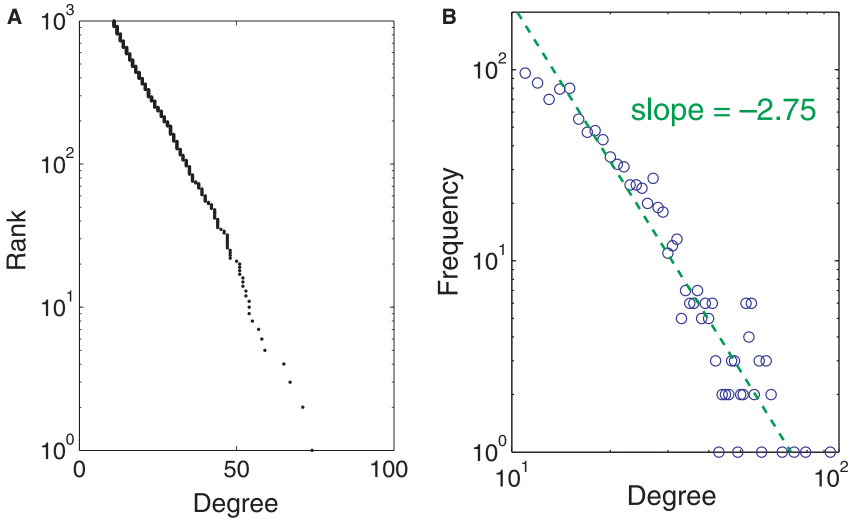

Fig. 2. (A) Semilog rank–degree (black dots), and (B) log–log frequency–degree (blue circles) plots for integer data numerically sampled from the random variable following an exponential $P[X > x] \approx \exp^{(1-0.1x)}$ for $x > 10$. Rank–degree plot (A) correctly shows that the data are exponentially distributed, whereas the frequency–degree plot (B) incorrectly suggests the data follows a power law with the slope $-(1 + \alpha) \approx -2.75$ (dotted green line).

interaction can be predicted if orthologs in a model organism show an interaction. Its accuracy has been assessed in [15]. We consider both FYI and HPI to be refined data sets, and investigate whether their node degree sequences follow a power law, a defining feature of scale-free networks, by rank–degree plots.

3. Results and discussion

The rank–degree plots of the HPI and FYI data are shown in (A) log–log scale and (B) semilog scale in Figs. 3 and 4, respectively. The solid lines and the dotted curve in log–log scale (A) show least-squares fitting of data to a power law with the value of its slope and to an exponential, respectively. The same fittings are depicted as the solid curve and the dotted line in semilog scale (B). From these figures, we can clearly conclude that the node degree sequences of HPI and FYI data are much closer to an exponential (2) for large degrees, and are clearly not power laws (1). More sophisticated statistical analysis can be used to confirm these conclusions. In addition, the rank–degree plots show raw data, without binning or any transformation of the data, and readers can easily judge at a glance the relative suitability of various models.

However, using frequency–degree plots (C) in Figs. 3 and 4 could lead to the erroneous conclusion that the node degree sequence appears to follow a power law. This is essentially the same error as that illustrated in Fig. 2, although there we had the additional benefit of analytic formulas to confirm the analysis. As we have shown in Fig. 1, even if the PPI data were actually a power law, the slope in a frequency–degree plot is not reliably related to the true slope. These results conclusively demonstrate that these two refined PPI data sets are not power laws, and thus these PPI networks are certainly not SF, no matter how this is defined. This is consistent with the claim in [24] that the degree sequence of refined PPI data should not follow a power law.

It is in principle possible that the data studied here is misleading because of the small size of the network and potential experimental errors, and that real PPI networks might have some features attributed to SF networks. At this time we only can draw conclusions about (noisy) subgraphs of the true network since the data sets are incomplete and presumably contain errors. If it is true that appropriately sampled subraphs of a SF graph is SF as was claimed in [9], they possess a power law node degree sequence. That these subgraphs exhibit exponential node degree sequences suggests that the entire network is not SF. Since essentially all the claims that biological networks are SF are based on ambiguous frequency–degree analysis, this analysis must be redone to determine the correct form of the degree sequences. This paper has provided clear examples that ambiguous plots of frequency–degree could lead to erroneous conclusion on the existence and parametrization of power law relationships. We have illustrated how rank–degree plots are more reliable, but of course much more sophisticated analysis is possible and ultimately desirable.

It has also been clearly shown [7,8] that the Internet and cell metabolism, the two most prominent examples of SF networks, plausibly can have power laws for some data sets, but have none of the other features attributed to SF networks. One important feature of the Internet and metabolic networks is the complete absence of centrally located high-degree "hubs" which are claimed to be responsible for global network connectivity and whose removal would fragment the network. This contradicts what has been claimed in the SF literature. Metabolic networks have also been shown to be scale-rich (SR), in the sense that they are far from self-similar [7] despite some power laws in certain node degree sequence. Their power law node degree sequence is a result of the mixture of exponential distributions in each functional module, with carriers playing a crucial role. In principle, PPI networks could have this SR structure as well, since their subnetworks have exponential degree sequence, and perhaps power laws could emerge at higher levels of organization. This will be revealed only when a more complete network is elucidated. Still, the most important point is not whether the node degree sequence follows a power law, but whether the variability of the node degree sequences is high or low [7], and the biological protocols that

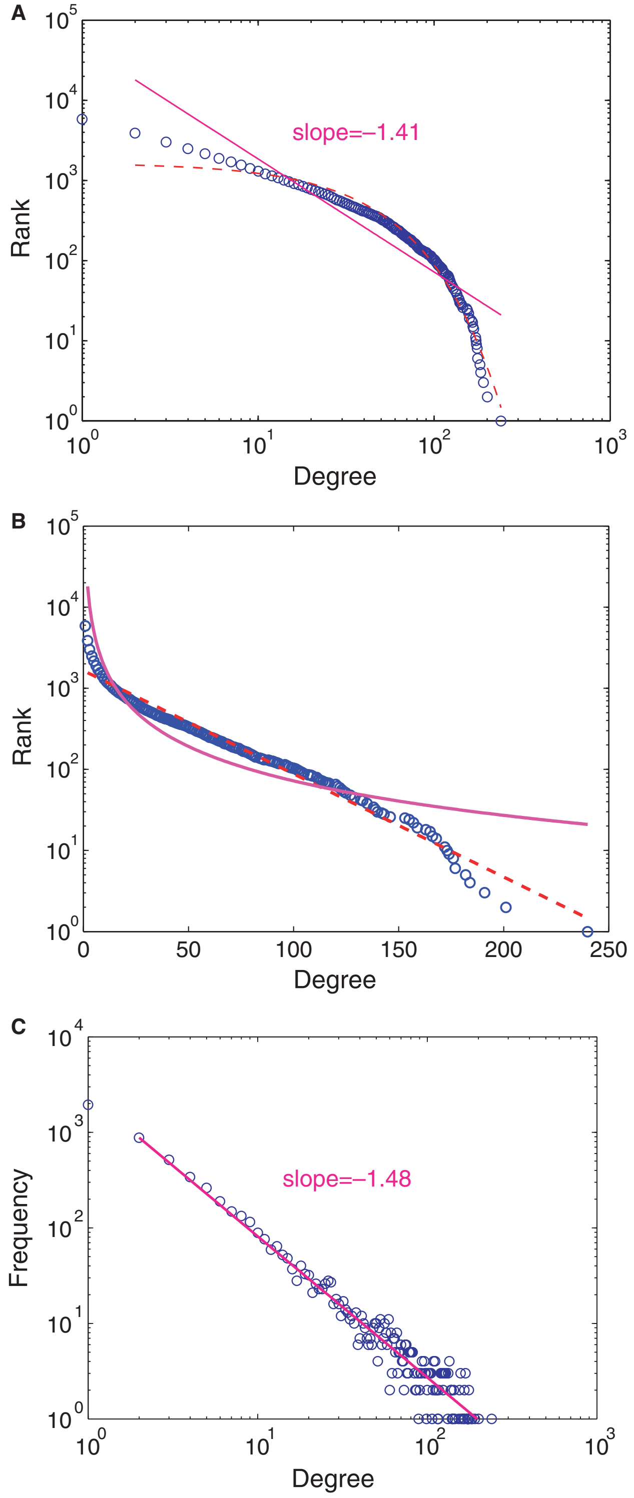

Fig. 3. Node degree distribution of all interactions in human protein-interaction map [15]: (A) rank–degree plot in log–log scale, (B) rank–degree plot in semilog scale, and (C) frequency–degree plot in log–log scale. The rank–degree plots indicate that the degree distribution is exponential. The straight lines (A,C) and the dotted curve (A) in log–log scale are the least-squares fits of the data to the power law (with the value of the slope) and to the exponential distributions, respectively. The straight line and the dotted curve in log–log scale (A) become the curve and the dotted line in semilog scale (B). Still, the frequency–degree plot in (C) might appear visually to follow a power law, and can lead to potential errors of finding power law node degree distribution.

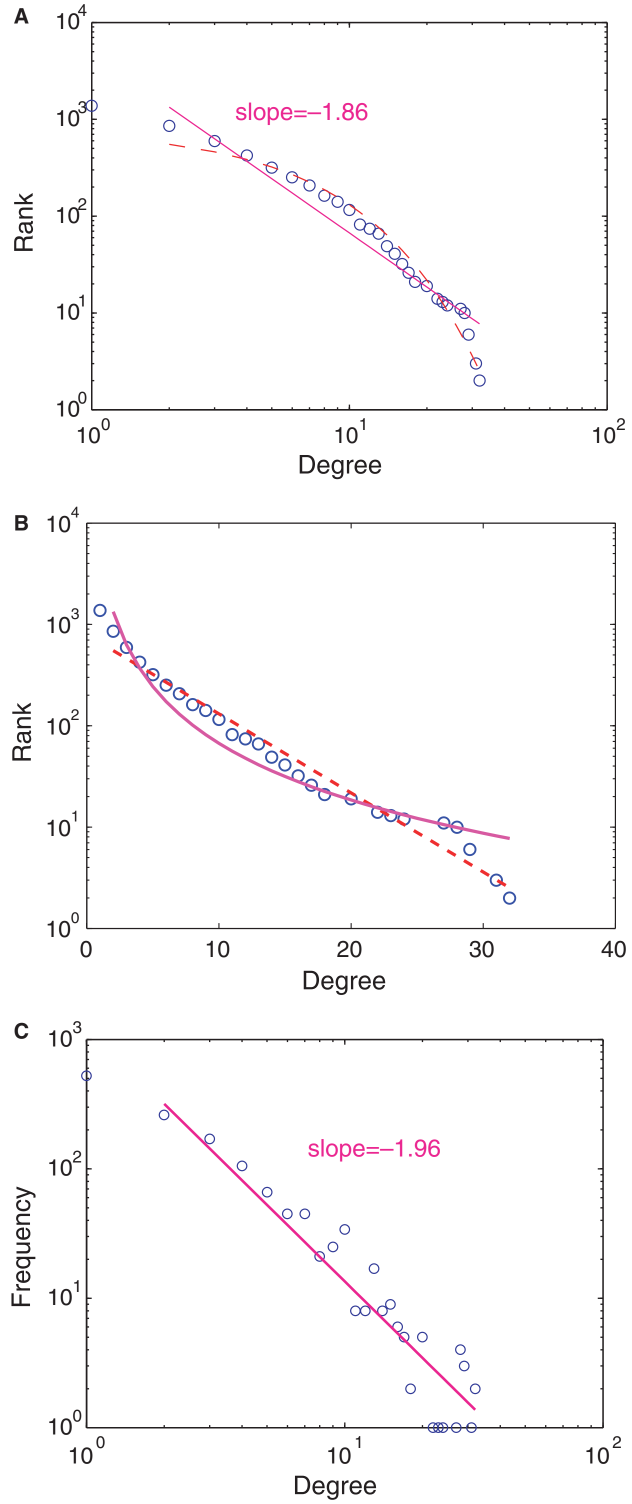

Fig. 4. Node degree distribution of all interactions in 'filtered yeast interactome' (FYI) data set [14]: (A) rank–degree plot in log–log scale, (B) rank–degree plot in semilog scale, and (C) frequency–degree plot in log–log scale. The rank–degree plot (A,B) shows the non-power law distribution, which is not evident in the frequency–degree plot (C). The straight lines (A,C) and the dotted curve (A) in log–log scale are the least-squares fits of the data to the power law (with the value of the slope) and to the exponential distributions, respectively. The straight line and the dotted curve in log–log scale (A) become the curve and the dotted line in semilog scale (B).

necessitate this high or low variability. These issues will be explored in future publications.

Acknowledgements: The authors thank Nicolas Bertin and Marc Vidal for providing FYI data and Nataša Pržulj for some comments. J.C.D. acknowledges support from the Army Institute for Collaborative Biotechnologies, NSF NSF Award: CCF-0326635 "ITR COLLAB: Theory and Software Infrastructure for a Scalable Systems Biology", and AFOSR Award: FA9550-05-1-0032 "Bio Inspired Networks".

References

[1] Uetz, P. and Finley Jr., R.L. (2005) From protein networks to biological systems. FEBS Lett. 579, 1821–1827.

[2] Vidal, M. (2005) Interactome modeling. FEBS Lett. 579, 1834–1838.

[3] Jeong, H., Mason, S.P., Barabási, A.-L. and Oltvai, Z.N. (2001) Lethality and centrality in protein networks. Nature 411, 41–42.

[4] Barabási, A.-L. and Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organization. Nature Rev. Genet. 5, 101–114.

[5] Bollobás, B. and Riordan, O. (2003) Robustness and vulnerability of scale-free random graphs. Internet Math. 1, 1–35.

[6] Li, L., Alderson, D., Doyle, J.C., and Willinger, W. (to appear) Towards a theory of scale-free graphs: definition, properties, and implications. Internet Math.

[7] Tanaka, R. (2005) Scale-rich metabolic networks. Phys. Rev. Lett. 94, 168101.

[8] Li, L., Alderson, D., Doyle, J. and Willinger, W. (2004) A first principles approach to understanding the Internet's router-level topology. Proc. ACM SIGCOMM.

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

[10] Stumpf, M.P.H., Wiuf, C. and May, R.M. (2005) Subnets of scale-free networks are not scale-free: Sampling properties of networks. PNAS 102, 4221–4224.

[11] Berg, J.M., Tymoczko, J.L. and Stryer, L. (2002) Biochemistry, 5th ed, Freeman and Company, New York.

[12] Samorodnitsky, G. and Taqqu, M.S. (1994) Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance, Chapman & Hall, London.

[13] Resnick, S.I. (1997) Heavy tail modeling and teletraffic data. Annal. Stat. 25, 1805–1869.

[14] Han, J.-D., Bertin, N., Hao, T., Goldberg, D.S., Berriz, G.F., Zhang, L.V., Dupuy, D., Walhout, A.J.M., Cusick, M.E., Roth, F.P. and Vidal, M. (2004) Evidence for dynamically organized modularity in the yeast protein–protein interaction network. Nature 430, 88–93.

[15] Lehner, B. and Fraser, A.G. (2004) A first-draft human protein-interaction map. Genome Biol. 5, R63.

[16] Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M. and Pochart, P., et al. (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. Nature 403, 623–627.

[17] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. and Sakaki, Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. USA 98, 4569–4574.

[18] Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M. and Cruciat, C.M., et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415, 141–147.

[19] Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K. and Boutilier, K., et al. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature 415, 180–183.

[20] Tong, A.H., Evangelista, M., Parsons, A.B., Xu, H., Bader, G.D., Page, N., Robinson, M., Raghibizadeh, S., Hogue, C.W. and Bussey, H., et al. (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. Science 294, 2364–2368.

[21] von Mering, C., Krause, R., Snel, B., Cornell, M., Olivier, S.G., Fields, S. and Bork, P. (2002) Comparative assessment of large-scale data sets of protein–protein interactions. Nature 417, 399–403.

[22] Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y.L., Ooi, C.E., Godwin, B. and Vitols, E., et al. (2003) A protein interaction map of *Drosophila melanogaster*. Science 302, 1727–1736.

[23] Li, S., Armstrong, C.M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.O., Han, J.D., Chesneau, A. and Hao, T., et al. (2004) A map of the interactome network of the metazoan *C. elegans*. Science 303, 540–543.

[24] Pržulj, N., Corneil, D.G. and Jurisica, I. (2004) Modeling interactome: scale-free or geometric? Bioinformatics 20, 3508–3515.