Comment on Yu et al., "High Quality Binary Protein Interaction Map of the Yeast Interactome Network." *Science* **322**, 104 (2008).

A. Clauset[1, *]

[1]*Santa Fe Institute, 1399 Hyde Park Road, Santa Fe NM, 87501, USA*

We test the claim by Yu *et al.* —presented in *Science* **322**, 104 (2008)—that the degree distribution of the yeast (*Saccharomyces cerevisiae*) protein-interaction network is best approximated by a power law. Yu *et al.* consider three versions of this network. In all three cases, however, we find the most likely power-law model of the data is distinct from and incompatible with the one given by Yu *et al.* Only one network admits good statistical support for any power law, and in that case, the power law explains only the distribution of the upper 10% of node degrees. These results imply that there is considerably more structure present in the yeast interactome than suggested by Yu *et al.*, and that these networks should probably not be called "scale free."

Protein-interaction networks, where nodes are natural proteins and links represent non-trivial binding affinity between two proteins, hold great promise for pushing forward our understanding of cellular processes. The wide interest in these interactome networks stems mainly from the observation that while modern genetic methods allow us to identify which genes code for proteins, the set of these genes only amounts to a "parts list" of a cell. A more complete understanding of cellular processes requires knowing something about the functional roles and dynamic interactions of these parts [2]. Thus, by considering the patterns of protein interactions, i.e., their network, we can pose and answer meaningful biological questions about complex cellular processes, and their evolution.

Determining the protein-interaction network for a particular species is a highly non-trivial task, and relies upon sophisticated molecular techniques to both build the proteins, and to test their pairwise interactions. The most direct approach would be to individually test each of the n^2 possible interactions for n proteins. But, because n is typically on the order of thousands or tens of thousands, and high-throughput methods are not yet available for these tests, this approach is not used. Instead, researchers use techniques that test multiple interactions at once [3, 4], e.g., the yeast two-hybrid (Y2H) screen [5]. To date, there have been a number of high profile efforts to construct the interactome for yeast (*S. cerevisiae*) [6, 7, 8, 9, 10], and, many would argue, a lot of real progress.

However, these methods also have serious limitations, and have been shown capable of producing high false-positive and high false-negative rates [11, 12]. Some of the techniques also exhibit severe biases, being unable to test for interactions involving entire classes of proteins. The Y2H assay, for instance, is not suitable for transcriptional activators or membrane-bound proteins. As a result of these limitations, some scientists have wondered privately how much real biology our current maps actually capture.

The Yu *et al.* paper attempts to address some of these shortcomings using a new high-throughput Y2H screen. Combining its results with those of previous studies, they produce a new set of interactions (denoted Y2H-union), which they suggest covers about 20% of the entire network. In the interest of completeness, Yu *et al.* include in their subsequent network analyses two alternative versions of the yeast interactome: one is based on co-complex information drawn from raw high-throughput coaffinity purification and mass spectrometry data (denoted Combined-AP/MS), and one is based on a smaller set of literature-curated interactions (denoted LC-multiple) that are assumed to be error free.

This Comment is not concerned with the quality of the laboratory procedures or the accuracy of the inferred interaction data. Rather, the focus is relatively narrow, concerning only the analysis of these networks' degree distributions and the conclusions drawn thereby. In particular, Yu *et al.* claim that the degree distributions of all three networks follow power-law distributions (with parameters given in Table I):

> As found previously for other macromolecular networks, the connectivity or "degree" distribution of all three data sets is best approximated by a power-law. [1]

The implication thus being that the yeast proteome can be considered "scale free", with all that goes along with that label [13]. However, this claim depends on a statistical method—specifically, linear regression—that is known to produce biased and incorrect results in this context. By reanalyzing Yu *et al.*'s data [14] with appropriate statistical tools [15], we show that (i) the most likely power-law models of these networks' degree distribution are distinct from and incompatible with those quoted by Yu *et al.*, (ii) at best, only the 10.3% most connected nodes in the Y2H-union network are plausibly power-law distributed, and (iii) there is considerably more structure present in all of these networks than suggested by Yu *et al.* As a result, these networks should probably not be considered "scale free."
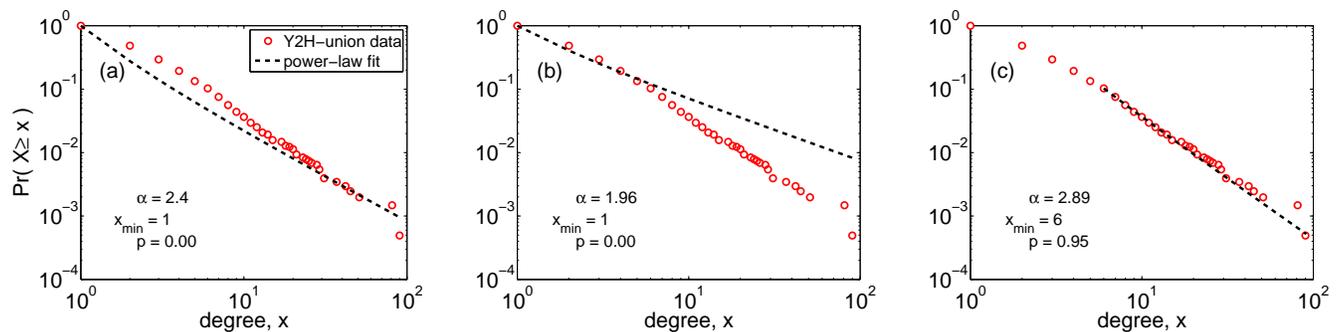
We begin by reanalyzing the Y2H-union data, argued

FIG. 1: Three approaches to fitting a power-law distribution to the Y2H-union data from Yu *et al.* [1]. (a) The parameterization given by Yu *et al.*, derived using standard linear regression on the log-transformed histogram of degree frequencies. (b) A fit derived via maximum likelihood over the entire range of the data, i.e., $x_{\min} = 1$. (c) A fit derived via maximum likelihood for estimating α, and by selecting the $x_{\min}$ that gives the best power-law fit to the upper range of the data. In both (a) and (b), the fitted power-law is not statistically significant ($p = 0.00 \pm 0.01$), indicating that large or systematic deviations from the power-law hypothesis exist. In (c), the upper 10.3% are plausibly power-law distributed ($p = 0.95 \pm 0.03$).

by Yu *et al.* to be the most accurate map among the three. Figures 1a–c show the data along with three different power-law models. The first panel shows the model suggested by Yu *et al.* (with scaling exponent $\alpha = 2.4$) which was derived using a linear regression approach; this model yields a poor fit to the lower-to-middle range of the data. The second panel shows the most likely power-law model over the entire range of data, which fits the lower range and thus the majority of the data relatively well, but yields a poor fit to the upper range. The third panel shows the most likely power-law model for the upper range alone.

Notably, Figures 1a–c plot the data as a complementary cumulative distribution function (CDF). If the data were indeed power-law distributed, this function would be straight on the log-log axes, but the reverse is not true. Being straight on log-log axes is not a sufficient condition for some data to be power-law distributed; many kinds of non-power-law distributed data can look straight on log-log axes. To decide whether some data do or do not follow a power-law distribution, we must use statistical tools that can tell the difference [15]. Linear regression is not one of these.

A straightforward test for whether some data can reasonably be claimed to follow a power-law distribution is a significance test [16, 17]. When we fit the power-law model to the same data that we use to score the power law's plausibility, we induce a correlation between the data and the model; however, this correlation can be controlled using a Monte Carlo procedure [15]. The result of the test is a single value p that represents the plausibility of the fitted model as an explanation of the data: small p-values, conventionally $p < 0.1$, indicate that the data cannot be considered to follow the fitted model, while larger values indicate only that the fitted model is plausible (not that it is correct) [18]. Conducting such a test with the Y2H-union data and the power-law claimed by Yu *et al.* yields $p = 0.00 \pm 0.01$, indicating that this power

law is a terrible model of the data, and that the data deviate in large or systematic ways from the proposed power law.

The poor fit here is partly due to the way the power law was estimated from the data: linear regression for fitting the power law is known to produce biased results [15, 19] mainly by dramatically overweighting the large but rare events. Further, the canonical r^2 value, used to judge the quality of the regression, can easily be high even when the data are not power-law distributed [15]. In short, regression applied in this way makes assumptions about the model that are incompatible with the hypothesis being tested, and thus fails in uncontrolled ways.

A more reliable method for fitting a power-law distribution to data is the method of maximum likelihood [16, 20]. Figure 1b shows a power-law model fitted to the entire range of data, while Figure 1c shows a power law fitted only to the upper range. The second of these requires choosing a point $x_{\min}$ where the power-law behavior starts, but this can be done using appropriate tools [15]. Applying the significance test to both models, we find that even the most likely power law is still a terrible explanation of the entire range of data ($p = 0.00 \pm 0.01$), but that it's an entirely plausible explanation ($p = 0.95 \pm 0.03$) of the 209 (10.3%) most connected nodes, i.e., those with degree $x \geq 6$. A corollary, however, is that the distribution of the degrees $1 \leq x < 6$ is not a power law. (One possibility is that the low-degree data is power-law distributed in a way different from that of the high-degree nodes. Fortunately, the same tools we have already used can be readily adapted to test this hypothesis.)

Figure 2 repeats the analysis of Figure 1c with the data for the alternative interactome maps. For these, we find that the Combined-AP/MS network's degree distribution cannot be considered to follow a power law ($p = 0.01 \pm 0.03$), while the support for the LC-multiple network is marginally significant ($p = 0.15 \pm 0.03$). In both cases,
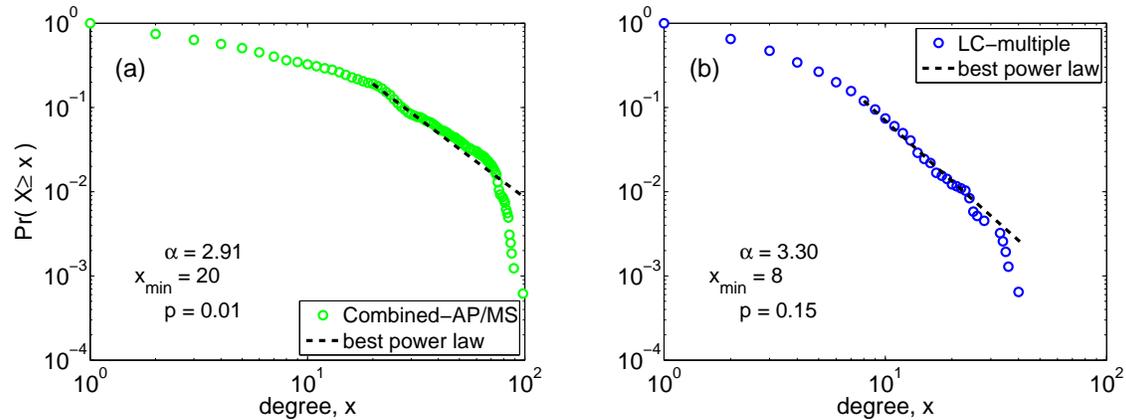
FIG. 2: The Combined-AP/MS and LC-multiple data sets, shown as complementary cumulative distribution functions $\Pr(X \geq x)$ on log-log axes, along with fits of the power-law hypothesis using the same methods as in Fig. 1c.

data set	regression				maximum likelihood				support for
	α	$x_{\min}$	n_{tail}	r^2	α	$x_{\min}$	n_{tail}	p (± 0.03)	power law
Y2H-union	2.4	1	2032	0.96	2.9 ± 0.2	6 ± 1	209	0.95	okay
Combined-AP/MS	1.4	1	1621	0.96	2.9 ± 0.1	20 ± 1	309	0.01	none
LC-multiple	2.1	1	1552	0.92	3.3 ± 0.2	8 ± 2	186	0.15	marginal

TABLE I: Power-law models for the three versions of the yeast interactome. In the first few columns, we quote the models given by Yu *et al.*, which were derived using regression methods. In the second set, we give the best fits derived by maximum likelihood when the range of fit is allowed to vary (as in [15]), along with the corresponding *p*-value. In the final column, we state the support for the conjecture that the corresponding data follows a power-law distribution. In every case, the maximum likelihood power law is distinct and incompatible with that given by Yu *et al.*, and only for the upper 10.3% of the Y2H-union data does the power-law hypothesis have strong statistical support.

the best power-law models cover only the upper range of data, and power laws that cover the entire range have zero support ($p = 0.00 \pm 0.03$). Table I summarizes the results of our reanalysis of the three networks.

As a brief aside, Yu *et al.* also conducted a model-comparison exercise to test whether a power-law distribution with or without an exponential cutoff was a better explanation of the data. For the same reasons given above, the regressions and r^2 values Yu *et al.* use cannot reliably determine which of these models is better. Fortunately, reliable tools for answering such a question do exist, e.g., a likelihood ratio test [15, 21], although we do not apply them here.

With these results in hand, we can now make several novel conclusions about the structure of protein interactions in yeast, and generally clarify the results of Yu *et al.* First, the question of whether the yeast interactome's degree distribution is well-characterized by a power-law distribution is not yet settled. Yu *et al.* argue that the Y2H-union network is the most accurate map to date (more accurate than the Combined-AP/MS or LC-multiple versions), but here the statistics only support the notion that a power law is a plausible model of the degrees of a small fraction of the entire network (10.3%, or the 209 most connected nodes). The LC-multiple network was presented as being a smaller, but generally high-quality,

data set and here there is only marginal statistical support for a power-law degree distribution in the upper range. Notably, the two power laws for these networks are largely incompatible, with $\alpha_{\text{Y2H}} = 2.9 \pm 0.2$ versus $\alpha_{\text{LC}} = 3.3 \pm 0.2$. If the Y2H-union network were merely a more complete version of the LC-multiple network, a greater degree of overlap in these estimates would be expected. The implication is that there are significant differences between these networks that cannot be explained away by simple sampling arguments.

Further, the fact that the best power-law model of the Y2H-union data only explains the distribution of the upper 10.3% of the node degrees implies that there is considerable structure in this network that remains to be explained. This structure may have evolutionary or functional significance, especially considering that its behavior is qualitatively different from that of the large-degree nodes and that it accounts for almost 90% of the network. Additionally, the existence of the cross-over point at $x_{\min} = 6$ from non-power-law to power-law behavior deserves a scientific explanation. In the additional structural analyses performed by Yu *et al.*, the $x_{\min}$ value could serve as a principled threshold by which to quantitatively define "high degree" (see for instance [22]).

In closing, we note that there are many aspects of the Yu *et al.* study that seem entirely reasonable, and much

of the paper concerns the experimental work done to construct the Y2H-union version of the yeast interactome. From this perspective, the paper pushes the field forward in a meaningful way, and the problems discussed here are a small part of a very large project.

On the other hand, the goal of constructing and analyzing the yeast interactome is to ultimately understand the mechanisms that create the observed patterns of interactions, and their implications for higher cellular functions. Scientific progress on these questions certainly depends on high quality experimental work, but it also depends on high quality statistical work: to get the theories right, we must also get the statistics right. Otherwise, we cannot know for sure what the data do and do not say. For testing whether some data do or do not follow a power-law distribution, reliably accurate tools now exist, and their application can shed considerable light on the relevant scientific questions.

[1] H. Yu *et al.*, "High Quality Binary Protein Interaction Map of the Yeast Interactome Network." *Science* **322**, 104 (2008).

[2] Arguably, this level of understanding is still intermediate. A more complete understanding would incorporate knowledge about how genes interact and regulate each other, potentially through intermediate cellular processes.

[3] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics." *Nature* **422**, 198–207 (2003).

[4] E. Phizicky, P. I. H. Bastiaens, H. Zhu, M. Snyder and S. Fields, "Protein analysis on a proteomic scale." *Nature* **422**, 208–215 (2003).

[5] L. McAlister-Henn, N. Gibson and E. Panisko, "Applications of the Yeast Two-Hybrid System." *Methods* **19**, 330–337 (1999).

[6] M. Fromont-Racine, J. C. Rain, P. Legrain, "Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens." *Nature Genetics* **16**, 277–282 (1997).

[7] P. Uetz *et al.*, "A comprehensive analysis of protein-protein interactions in *Saccharomycs cerevisiae*." *Nature* **403**, 623–627 (2000).

[8] T. Ito *et al.*, "A comprehensive two-hybrid analysis to explore the yeast protein interactome." *PNAS* **98**, 4569–4574 (2001).

[9] A.-C. Gavin *et al.*, "Functional organization of the yeast proteome by systematic analysis of protein complexes." *Nature* **415**, 141–147 (2002).

[10] Y. Ho *et al.*, "Systematic identification of protein complexes in *Saccharomycs cerevisiae* by mass spectrometry." *Nature* **415**, 180–183 (2002).

[11] G. D. Bader and C. W. V. Hogue, "Analyzing yeast protein-protein interaction data obtained from different sources." *Nature Biotechnology* **20**, 991–997 (2002)

[12] J. S. Bader, A. Chaudhuri, J. M. Rothberg and J. Chant, "Gaining confidence in high-throughput protein interaction networks." *Nature Biotechnology* **22**, 78–85 (2004)

[13] L. Li, D. Alderson, J. C. Doyle and W. Willinger, "Towards a Theory of Scale-Free Graphs: Definitions, Properties, and Implications." *Internet Mathematics* **2**, 431–523 (2006).

[14] The data we analyze was taken directly from the published histograms in the main text and the appendices of [1], using DataThief (`http://www.datathief.org/`).

[15] A. Clauset, C. R. Shalizi and M. E. J. Newman, "Power-law distributions in empirical data." arXiv:0706.1062 (2007).

[16] L. Wasserman, *All of Statistics*, Springer (2004).

[17] W. H. Press, S. A. Teukolsky, W. T. Vetterling and B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press (1992).

[18] Tests will differ depending on exactly how large a deviation from the power-law model we allow our data to take before we conclude that the data are *not* power-law distributed. The test used here assumes all data are iid; thus, deviations are due to statistical fluctuations. In principle, this assumption could be relaxed to allow for correlations in the data.

[19] M. L. Goldstein, S. A. Morris and G. G. Yen, "Problems with fitting to the power-law distribution." *European Physical Journal B* **41**, 255–258 (2004).

[20] J. Aldrich, "R. A. Fisher and the Making of Maximum Likelihood 1912–1922." *Statistical Science* **12**, 162–176 (1997).

[21] Q. H. Vuong, "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses." *Econometrica* **57**, 307–333 (1989).

[22] A. Clauset, M. Young and K. S. Gledtisch, "On the Frequency of Severe Terrorist Attacks." *Journal of Conflict Resolution* **51**, 58–88 (2007).