BMC Bioinformatics

BioMed Central

Research article

Open Access

# Are scale-free networks robust to measurement errors?

Nan Lin[1] and Hongyu Zhao*[2,3]

Address: [1]Department of Mathematics, Washington University in St. Louis, St. Louis, MO 63143, USA, [2]Department of Epidemiology and Public Health, Yale University, New Haven, CT 06520, USA and [3]Department of Genetics, Yale University, New Haven, CT 06520, USA

Email: Nan Lin - nlin@math.wustl.edu; Hongyu Zhao* - hongyu.zhao@yale.edu

* Corresponding author

Published: 16 May 2005

*BMC Bioinformatics* 2005, **6**:119 doi:10.1186/1471-2105-6-119

Received: 31 October 2004
Accepted: 16 May 2005

This article is available from: http://www.biomedcentral.com/1471-2105/6/119

## Abstract

**Background:** Many complex random networks have been found to be scale-free. Existing literature on scale-free networks has rarely considered potential false positive and false negative links in the observed networks, especially in biological networks inferred from high-throughput experiments. Therefore, it is important to study the impact of these measurement errors on the topology of the observed networks.

**Results:** This article addresses the impact of erroneous links on network topological inference and explores possible error mechanisms for scale-free networks with an emphasis on *Saccharomyces cerevisiae* protein interaction networks. We study this issue by both theoretical derivations and simulations. We show that the ignorance of erroneous links in network analysis may lead to biased estimates of the scale parameter and recommend robust estimators in such scenarios. Possible error mechanisms of yeast protein interaction networks are explored by comparisons between real data and simulated data.

**Conclusion:** Our studies show that, in the presence of erroneous links, the connectivity distribution of scale-free networks is still scale-free for the middle range connectivities, but can be greatly distorted for low and high connecitivities. It is more appropriate to use robust estimators such as the least trimmed mean squares estimator to estimate the scale parameter $\gamma$ under such circumstances. Moreover, we show by simulation studies that the scale-free property is robust to some error mechanisms but untenable to others. The simulation results also suggest that different error mechanisms may be operating in the yeast protein interaction networks produced from different data sources. In the MIPS gold standard protein interaction data, there appears to be a high rate of false negative links, and the false negative and false positive rates are more or less constant across proteins with different connectivities. However, the error mechanism of yeast two-hybrid data may be very different, where the overall false negative rate is low and the false negative rates tend to be higher for links involving proteins with more interacting partners.

## Background

Recent studies have found that many complex networks, ranging from the World-Wide Web [1] and the scientific collaboration network [2] to biological systems such as the yeast protein interaction network [3], are scale-free. The scale-free property states that the distribution of the connectivity $k$ (number of links per node) in a network can be described by the power law, i.e.,

$P(k) = ck^{-\gamma}, c > 0, \gamma > 0.$ (1)

A visual diagnosis of the scale-free behavior can be made through the log-log plot of the connectivity distribution, in which a straight line with slope $-\gamma$ is expected. In scale-free networks, the nodes are not randomly or evenly connected with some highly connected nodes ("hubs"). The ratio of the number of "hubs" to that of nodes in the rest of the network remains constant as the network changes in size. One attractive feature is that scale-free networks are more resistant to random failures compared with random networks due to the existence of a few highly connected "hubs" [4]. Remarkably, it has been observed that the scale parameter $\gamma$ varied only in the narrow range of 2.1 – 4 in the aforementioned real-world networks. All existing studies on scale-free networks assumed that the observed links represented the underlying structure of the network, but paid little attention to the fact that the observed links often involved errors, namely, false positives and false negatives. For example, Jeong *et al.* [3] considered the *Saccharomyces cerevisiae* protein interaction network inferred from yeast two-hybrid (Y2H) experiments. It is well-known that the Y2H system has many false positives as well as false negatives [5]. A natural question to ask is whether a scale-free network is still observed as scale-free in the presence of errors. And if it is, what are the possible underlying error mechanisms and how variable is the observed scale parameter $\gamma$? Answering these questions may lead to further insight to the scale-free property, better understanding and correct usage of the observed network data. For convenience, we will call networks observed with erroneous links as perturbed networks in the rest of this article.

## Results

In this article, we address the above questions by both theoretical derivations and simulation studies using the yeast protein interaction network as a prototype. However, the results apply to general scale-free networks.

### *Connectivity distribution of scale-free networks with erroneous links under a simple model*

We first study how the connectivity distribution of a scale-free network is affected when errors are present. Following previous studies on the reliability of protein interaction networks [6], we assume a simple error mechanism in which the false positive rate ($r_{FP}$) and false negative rate ($r_{FN}$) are the same for all node pairs, and false positives and false negatives are independently generated. The false positive rate and false negative rate of a node pair refer to the probability that the pair of nodes is observed as linked when they are actually not and the probability that the pair of nodes is observed as unlinked when they are actually linked. Under this assumption, every truly linked pair of nodes has a probability $r_{FN}$ to be observed as unlinked nodes, and every truly unlinked pair of nodes has a probability $r_{FP}$ to be observed as linked nodes.

The above assumption is similar to the grand canonical ensembles of random networks in Chapter 4 of Dorogovtsev and Mendes [7], in which networks evolve by removing existing edges and adding new edges with certain probabilities. We can also view the perturbed network as obtained by removing edges (false negative) and adding edges (false positive) from the underlying network. The probability of adding an edge between two non-linked nodes is the false positive rate $r_{FP}$, and the probability of removing the edge between two linked nodes is the false negative rate $r_{FN}$. However, while Dorogovtsev and Mendes mostly discussed the connectivity distribution of equilibrium networks (networks obtained after infinite times edge adding and removing), we focus on the connectivity distribution of the observed network that are obtained by considering removing every existing edge and adding non-existing edges just once.

#### *Connectivity distribution of the perturbed network*

In the following, we will derive the distribution of the observed connectivities for a scale-free network of size $n$ for given values of $r_{FP}$ and $r_{FN}$. Let $N_P$ and $N_T$ denote the observed and true connectivity of a node, respectively. Then the probability to observe a node with $k$ links is

$$P(N_P = k) = \sum_{j=T_{min}}^{T_{max}} P(N_P = k, N_T = j) = \sum_{j=T_{min}}^{T_{max}} P(N_P = k \mid N_T = j) P(N_T = j). \quad (2)$$

The minimum and maximum connectivity of a node, $T_{min}$ and $T_{max}$, are assumed to be the same for all the nodes in the network, and their values depend on the specific network. In general, we set $T_{min} = 0$ and $T_{max} = n - 1$ when expert knowledge is not available, where $n$ denotes the size of the network, i.e., the total number of nodes in the network. The following elucidates how to calculate (2) analytically. Let $N_{FP}$, $N_{TP}$, $N_{FN}$, $N_{TN}$, and $N_N$ be the numbers of false positive links (observed as linked but actually not), true positive links (observed as linked and actually linked), false negative links (observed as unlinked but actually linked), true negative (observed as unlinked and actually unlinked) and negative links (actually unlinked) associated with the node, respectively. Since the observed links of a node consist of both false positive and true positive ones, and the true links consist of true positive and false negative ones, we have $N_P = N_{FP} + N_{TP}$, $N_T = N_{FN} + N_{TP}$, $N_N = N_{FP} + N_{TN}$, and $T_{max} = N_T + N_N$. Furthermore, underour assumed error mechanism, following similar derivations as shown in [7], $N_{FP}$ and $N_{FN}$ follow the binomial distributions $Bin(T_{max} - N_T, r_{FP})$ and $Bin(N_T, r_{FN})$, respectively, for a given value of $N_T$. This implies that $r_{FP} = E(N_{FP})/(T_{max} - N_T) = E(N_{FP})/(N_{FP} + N_{TN})$ and $r_{FP} = E(N_{FN})/N_T = E(N_{FN})/(N_{TP} + N_{FN})$, where $E(X)$ denotes the expectation of random variable $X$. Then the conditional probability $P(N_P = k|N_T = j)$ in (2) can be written as follows.

$$\begin{aligned} P(N_P = k \mid N_T = j) &= \sum_{i=\max(T_{min},j-k)}^{\min(T_{max}-k,j)} P(N_P = k \mid N_{FN} = i, N_T = j) P(N_{FN} = i \mid N_T = j) \\ &= \sum_{i=\max(T_{min},j-k)}^{\min(T_{max}-k,j)} P(N_{FP} = k-(j-i) \mid N_T = j) P(N_{FN} = i \mid N_T = j) \\ &= \sum_{i=\max(T_{min},j-k)}^{\min(T_{max}-k,j)} dBin(k-(j-i); r_{FP}, n-1-j) dBin(i; r_{FN}, j), \end{aligned}$$

where $dBin(k; p, n) = P(X = k)$ with $X \sim Bin(n, p)$. Moreover, the power law of the scale-free network implies that $P(N_T = j) = cj^{-\gamma}$. Hence, the observed connectivity distribution can be calculated by

$$P(N_P = k) = \sum_{j=T_{min}}^{T_{max}} cj^{-\gamma} \sum_{i=\max(T_{min},j-k)}^{\min(T_{max}-k,j)} dBin(k-(j-i); r_{FP}, n-1-j) dBin(i; r_{FN}, j). \qquad (3)$$

### *Simulations*

We next explore the impact of the erroneous links on the topology of the scale-free networks. With an emphasis on the yeast protein interaction network, we compute the distribution of the observed connectivity of scale-free networks with the false positive rate ($r_{FP}$) and false negative rate ($r_{FN}$) similar to the yeast protein interaction network under the assumption of the aforementioned simple error mechanism. We set the scale parameter $\gamma = 3$, the size of the network $n = 1000$ or 7000, and vary $r_{FP}$ from 0.0001 to 0.0003 and $r_{FN}$ from 0.1 to 0.9 on 9 equally spaced values. These ranges of $r_{FP}$ and $r_{FN}$ are based on Deng *et al.* [8], in which the authors estimated the false positive rate and false negative rate to be less than 0.000285 and greater than 0.64, respectively, based on the Y2H data. We consider a larger range of $r_{FP}$ to cover other data sources, such as the MIPS complex data, where false positives are less frequent. In the calculations, we use $T_{min} = 1$ and $T_{max} = n - 1$.

In the log-log plot (Figures 1 and 2) of the observed connectivity distribution of the perturbed networks when ($r_{FP} = 0.0001$, $r_{FN} = 0.3$) and ($r_{FP} = 0.00015$, $r_{FN} = 0.8$), it can be seen that the connectivity distribution after perturbation still maintains the scale-free property in the middle range of the connectivity, but deviates from the original linear pattern at both the small and large connectivity regions. The slope of the linear part is close to the true value -3 (see Tables A.1 and A.2 in Additional file 1). The deviation is more significant in the large connectivity region than that in the small connectivity region. This deviation pattern is consistent across networks of different sizes considered in our calculations (data not shown). Comparisons among the observed connectivity distributions (figures not shown) of perturbed networks with different values of $\mathrm{r}_{FP}$ and $r_{FN}$ suggest that the deviation depends little on $r_{FP}$ but largely on $r_{FN}$. As $r_{FN}$ increases, the deviation of the tail probability becomes more significant. This deviation is also more obvious in a smaller network.

### *Estimation of $\gamma$*

The connectivity distribution of the perturbed network suggests a cautious use of the observed link data, especially on estimating $\gamma$. The scaling parameter $\gamma$, an important characteristic measure of the scale-free network, is commonly estimated using the ordinary least squares (OLS) in the linear model from the log transformation of (1).

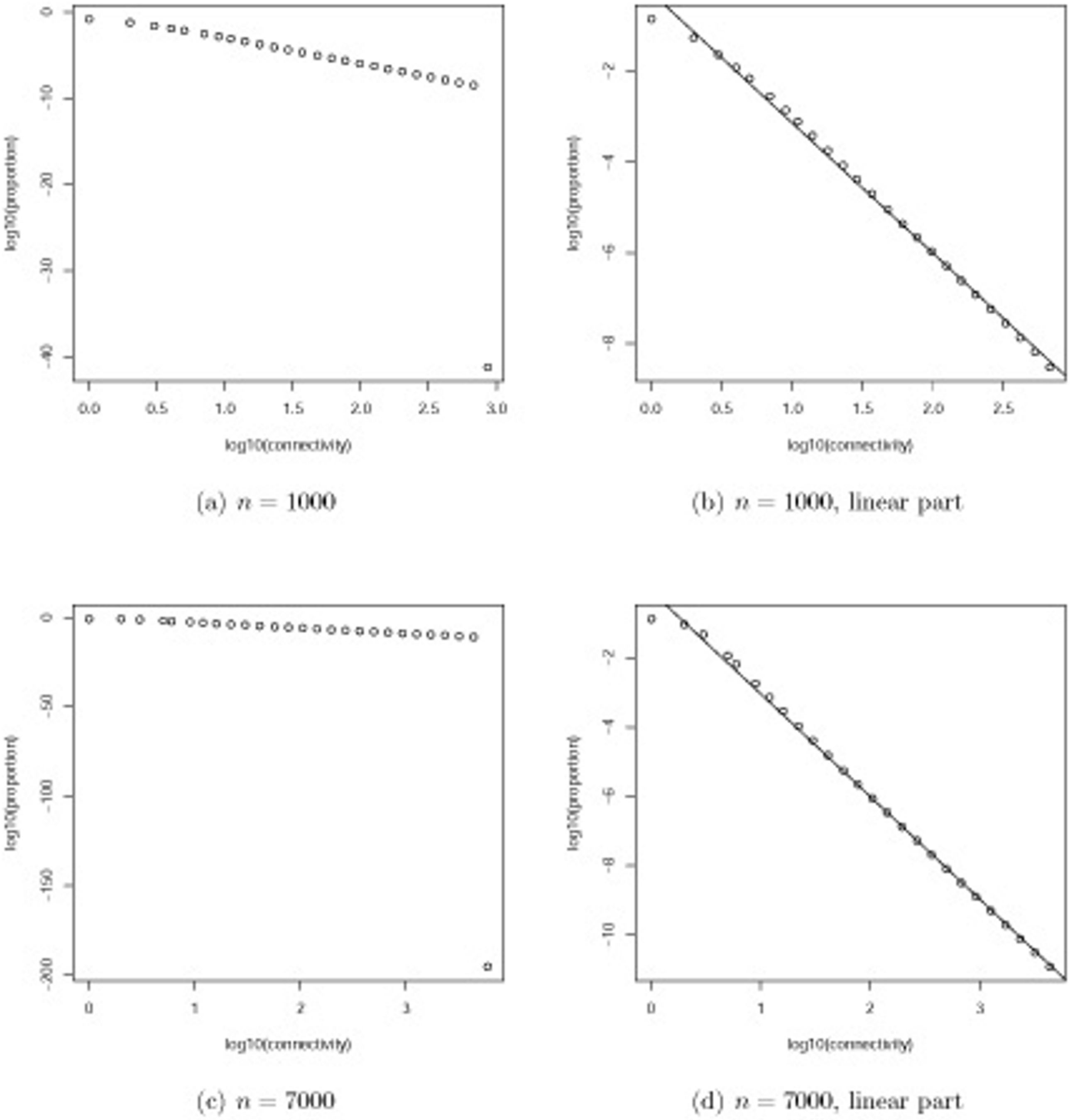
$$\log P(k) = \log c - \gamma \log k. \qquad (4)$$

It is well known that the OLS estimator can be very sensitive to even a small number of outliers. For example, applying the OLS estimator in Figure 1(a) will not be able to capture the linear trend if the point at the last end is included in the estimation. Therefore, robust estimators, such as the M-estimator and the least trimmed squares (LTS) estimator [9] are more proper choices in such situations due to their resistance to outliers. Our simulations suggest that the LTS estimator can correctly capture the linear trend without visual diagnosis of the connectivity distribution, while the OLS and M-estimator often fail to estimate the slope of the linear part correctly. Therefore, we will use the LTS estimator in our following simulation studies.

## ***Exploring error mechanisms of yeast protein interaction networks by simulations***

In the previous section, we found that the scale-free property can be conserved to a large extent under a simple error mechanism. However, the error mechanisms of the real data are often more complicated. For more complicated error mechanisms, theoretical derivations of the connectivity distribution of the perturbed networks are often intractable. But it is also important to know how the empirical connectivity distributions of real networks are affected by the erroneous links. Therefore, we conduct extensive simulation studies to investigate the finite-sample impact of the error mechanisms on the connectivity distribution. Our study focuses on the yeast protein-interaction network data.

For real network data, no matter whether erroneous links are involved or not, the empirical connectivity distribution will not display a linear pattern as clear as the ones in Figure 1 due to sampling variations and its discrete approximation to the tiny probability of nodes with large connectivities. For example, Figure 3 shows the connectivity distribution of a simulated scale-free network $Net_0$ and Figure 4 shows the connectivity distribution of $Net_0$ after perturbation by the simple error mechanism discussed above. In Figure 4, we observe a much larger curvature



(a) $n = 1000$

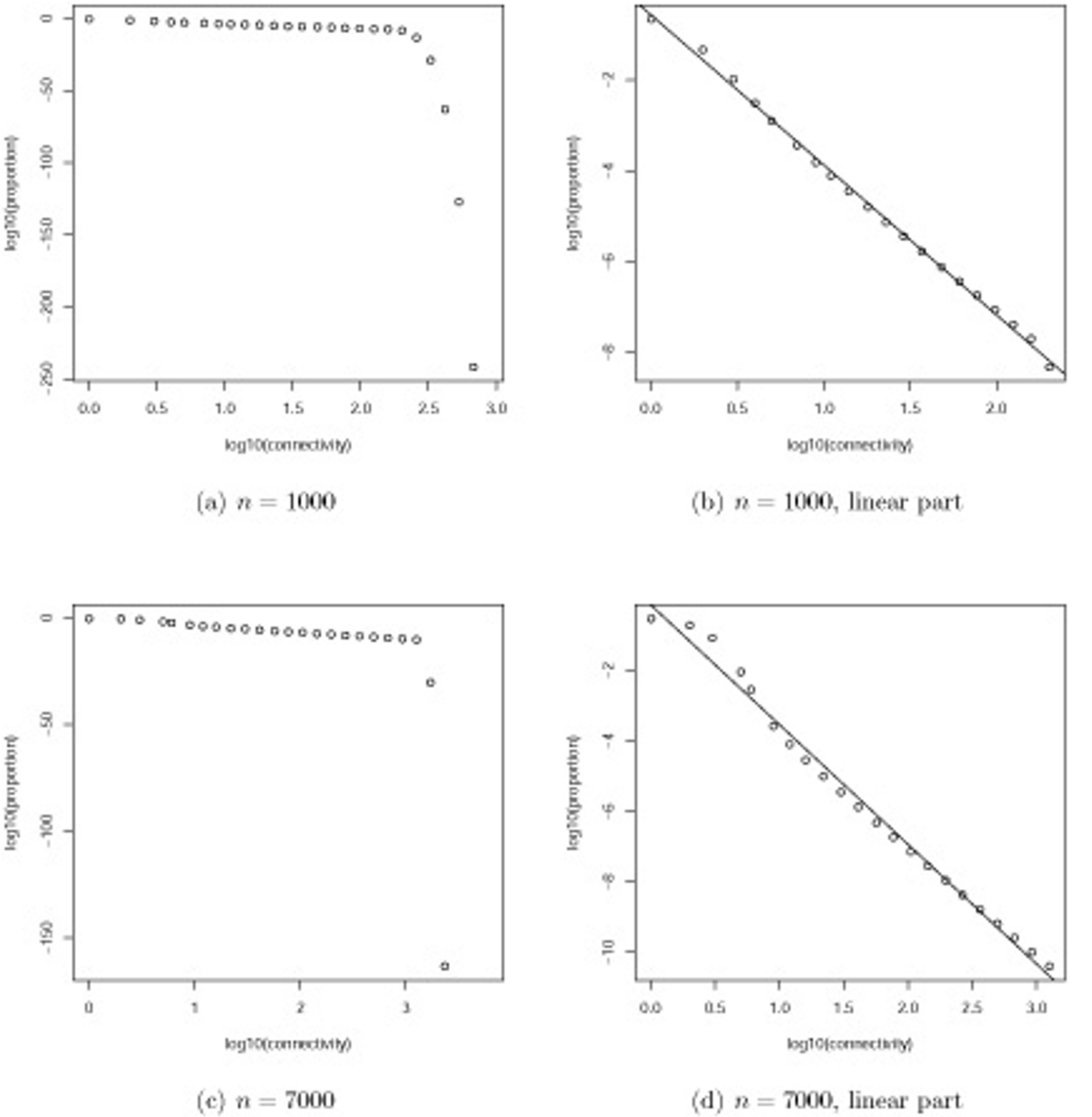
(b) $n = 1000$, linear part

(c) $n = 7000$

(d) $n = 7000$, linear part

**Figure 1**

**Connectivity distribution of the perturbed scale-free networks ($r_{FP} = 0.0001$, $r_{FN} = 0.3$)**. This picture shows the connectivity distribution of the the perturbed networks using (3) provided that $r_{FP} = 0.0001$ and $r_{FN} = 0.3$. Figure 1(b) and 1(d) are the linear parts of Figure 1(a) and 1(c), respectively, imposed with the regression lines fitted by the OLS.



(a) $n = 1000$

(b) $n = 1000$, linear part

(c) $n = 7000$

(d) $n = 7000$, linear part

**Figure 2**

**Connectivity distribution of the perturbed scale-free networks ($r_{FP} = 0.00015$, $r_{FN} = 0.8$)**. This picture shows the connectivity distribution of the the perturbed networks using (3) provided that $r_{FP} = 0.00015$ and $r_{FN} = 0.8$. Figure 2(b) and 2(d) are the linear parts of Figure 2(a) and 2(c), respectively, imposed with the regression lines fitted by the OLS.

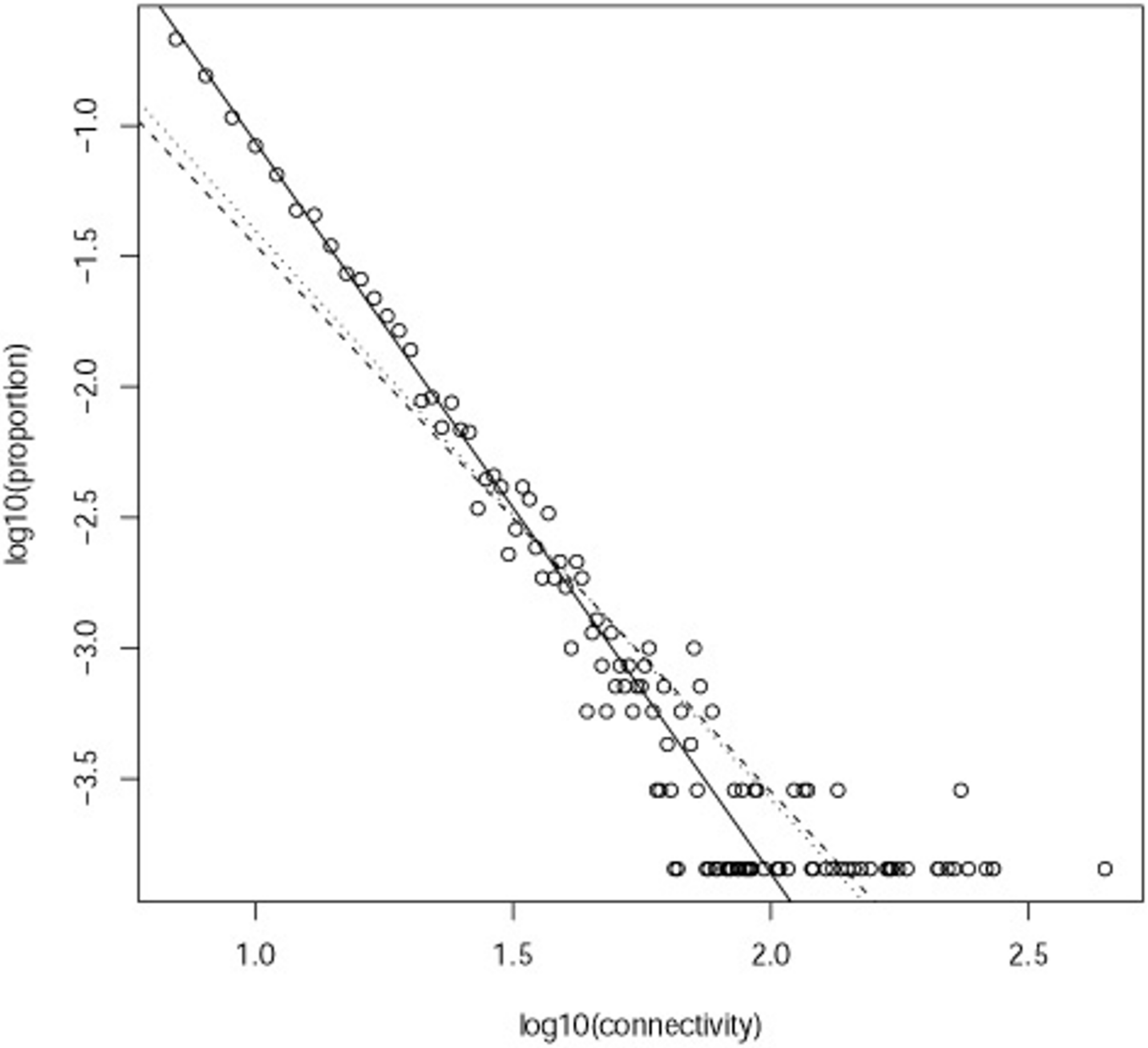


**Figure 3**
**Connectivity distribution of *Net*$_0$.** This picture shows the connectivity distribution of the simulated scale-free random network *Net*$_0$ imposed with regression lines given by different methods (dashed line: OLS; dotted line: M-estimation; solid line: LTS).

deviation from the linear trend at the small connectivity region than that in Figures 1 and 2. It is not clear why the empirical distributions of the simulated networks are so different from the theoretical calculations, but this observation demonstrates that simulation studies are necessary to complement the findings from the theoretical calculations. In addition, simulation studies can also explore possible error mechanisms by comparing the connectivity

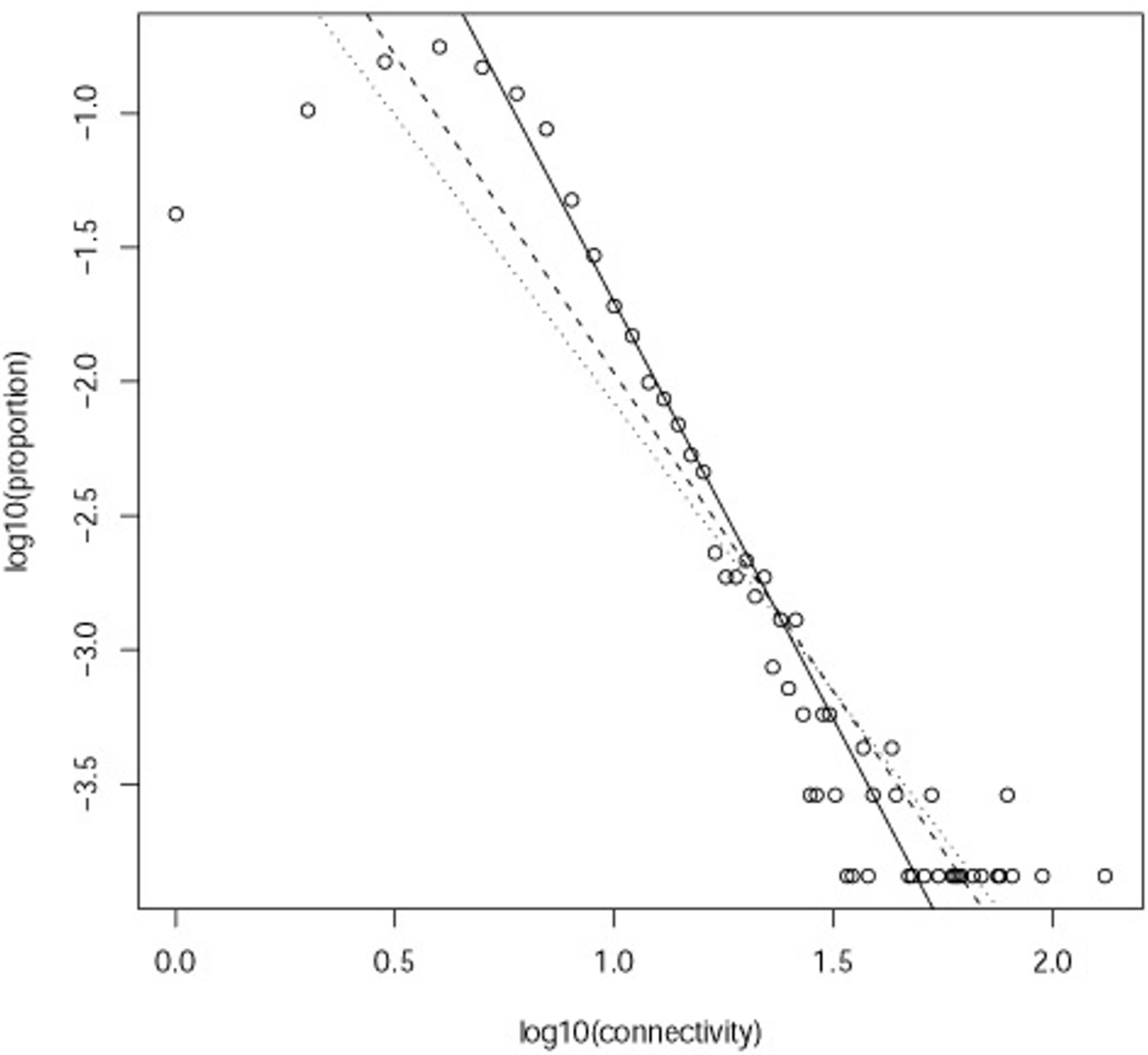


**Figure 4**

**Connectivity distribution of *Net*$_0$ after perturbation ($r_{FP}$ = 0.0002, $r_{FN}$ = 0.7)**. This picture shows the connectivity distribution of the simulated scale-free random network *Net*$_0$ perturbed by the simple error mechanism using $r_{FP}$ = 0.0002 and $r_{FN}$ = 0.7. Regression lines given by different methods are also imposed (dashed line: OLS; dotted line: M-estimation; solid line: LTS).

distributions of simulated perturbed scale-free networks with the observed networks by assuming that their underlying structure are indeed scale-free.

In the following, we investigate the error mechanisms of two real yeast protein interaction network data sets used in Jeong *et al.* [3] and Deng *et al.* [6] by comparing the connectivity distribution of these two networks with that of the simulated network perturbed by different error mechanisms. We assume that the true underlying topology of the yeast protein interaction network is scale-free [3]. Then if we perturb the simulated scale-free network by the error mechanisms similar to the ones of the real yeast protein interaction networks, the resulting connectivity distribution should be similar to the ones of the real networks.

### *MIPS and Y2H yeast protein networks*

Jeong *et al*. derived the yeast protein network from combined, non-overlapping Y2H data [10,11]. This network has 1,870 proteins as nodes, connected by 2,240 identified direct physical interactions [12]. The other network was obtained from the gold standard of yeast protein interactions based on the MIPS complex data [13]. This gold standard data set has 1,376 proteins and 2,876 interacting protein pairs, out of which 2,559 are also recorded in the Yeast Proteome Database (YPD) [14]. The YPD subset has 1,373 proteins. Estimates of $\gamma$ from the Y2H network, the gold standard data and the YPD subset are 2.396, 2.721 and 2.870, respectively. The connectivity distributions of these two networks are shown in Figure 5 and Figure 6, respectively.

### *Error mechanisms*

We consider different error mechanisms in terms of different types of false positive rates ($p_{ij} = P$ ($x_i$ and $x_j$ are observed linked$|x_i$ and $x_j$ are actually unlinked)) and false negative rates ($q_{ij} = P$ ($x_i$ and $x_j$ are observed unlinked$|x_i$ and $x_j$ are actually linked)) for node pair $(x_i, x_j)$, $i = 1,..., n$, $j = 1,..., n$, $i \neq j$. Assume that the overall false positive rate and false negative rate are $r_{FP}$ and $r_{FN}$, in the sense that the expected number of false positive links and false negative links are $E(N_{FP}) = r_{FP} N_N$ and $E(N_{FN}) = r_{FN} N_P$. We consider nine different error mechanisms by letting $p_{ij}$ and $q_{ij}$ be one of the following three different types:

1. **constant**: $p_{ij} = r_{FP}$ and $q_{ij} = r_{FN}$ for all $(x_i, x_j)$;

2. **increasing (with connectivity)**:

$$p_{ij} = \frac{L(x_i)+L(x_j)}{\sum_{(x_i,x_j)\notin S_{TP}}(L(x_i)+L(x_j))} r_{FP} N_N, \; q_{ij} = \frac{L(x_i)+L(x_j)}{\sum_{(x_i,X_j)\in S_{TP}}(L(x_i)+L(x_j))} r_{FN} N_P;$$

3. **decreasing (with connectivity)**:

$$p_{ij} = \frac{[L(x_i)+L(x_j)]^{-1}}{\sum_{(x_i,x_j)\notin S_{TP}}[L(x_i)+L(x_j)]^{-1}} r_{FP} N_N, \; q_{ij} = \frac{[L(x_i)+L(x_j)]^{-1}}{\sum_{(x_i,X_j)\in S_{TP}}[L(x_i)+L(x_j)]^{-1}} r_{FN} N_P,$$

where $L(x)$ denotes the true connectivity of node $x$. For $Net_0$, $N_P = 49, 007$ and $N_N = 24, 503, 521$. The combinations of different structures on false positive rates and false negative rates produce nine error mechanisms in Table 1.

### *Simulation studies*

We simulate a scale-free network $Net_0$ using the preferential attachment growth model [15,16]. In this algorithm, we start from $m_0 = 7$ isolated nodes and add $m = 7$ links to the existing nodes with probability proportional to their connectivity in each of the $T = 7, 000$ evolving steps. $Net_0$ has $L = 49, 007$ links and $n = 7, 008$ nodes. The mean-field theory [15] suggests that the theoretical value of $\gamma$for $Net_0$ is 3, which agrees well with the estimates in Table 2.

We always assume that false positives and false negatives are independently generated. In the simulations, a link is added (false positive) between every two unlinked nodes $(x_i, x_j)$ in $Net_0$ with probability $p_{ij}$, and the link is removed (false negative) between two linked nodes $(x_i, x_j)$ in $Net_0$ with probability $q_{ij}$. We also consider these error mechanisms under high and low overall false positive ($r_{FP}$) and false negative rates ($r_{FN}$). The connectivity distributions of $Net_0$ after perturbation are shown in Figures 7, 8, 9, 10 for different values of $r_{FP}$ and $r_{FN}$: (0.00025, 0.5), (0.00025, 0.8), (0.00015, 0.5), (0.00025, 0.8).

Under the nine different error mechanisms, the connectivity distribution of the perturbed $Net_0$ can be dramatically different. Under error mechanisms $S2$, $S5$, $S6$ and $S9$, the perturbed networks contain a small proportion of nodes with low connectivity, which differs greatly from the observed yeast protein interaction networks (Figures 5 and 6). This finding suggests that these four mechanisms are far different from the true error structure, and we will not discuss them in the following. We also observe that changes in $r_{FP}$ render little impact on the connectivity distribution under all error mechanisms, but a higher value of $r_{FN}$ increases the probability of nodes with small connectivity under $S1$, $S3$ and $S8$. And mechanisms $S4$ and $S7$ are highly stable structures, that is, the connectivity distribution changes little in response to changes in $r_{FP}$ or $r_{FN}$ under these two error mechanisms. This suggests that scale-free networks with constant false negative rates can still provide very credible information about its topological structure. This finding is also confirmed by the fact that the estimates of $\gamma$vary little when $r_{FN}$ changes (see Tables A.5 and A.6 in Additional file 1). The estimated values of $\gamma$ vary only from 2.61 to 3.03 with a standard error of 0.125 under $S4$ and only from 2.56 to 3.31 with a standard error of 0.161 under $S7$, whereas the estimate of $\gamma$ clearly decreases as $r_{FN}$ increases under $S3$ and $S8$ (Tables A. 4 and A. 7 in Additional file 1). Under $S1$, there is no clear pattern on the estimated $\gamma$ as $r_{FN}$ changes (Table A.3 in Additional file 1), but the estimates of $\gamma$vary in a much wider range (1.16 – 4.35) compared with those under $S3$ and $S8$. It is worth noting that our conclusions are restricted to the particular range of $r_{FP}$ and $r_{FN}$ we have studied, however these ranges are believed to be reasonable to describe the Y2H systems.

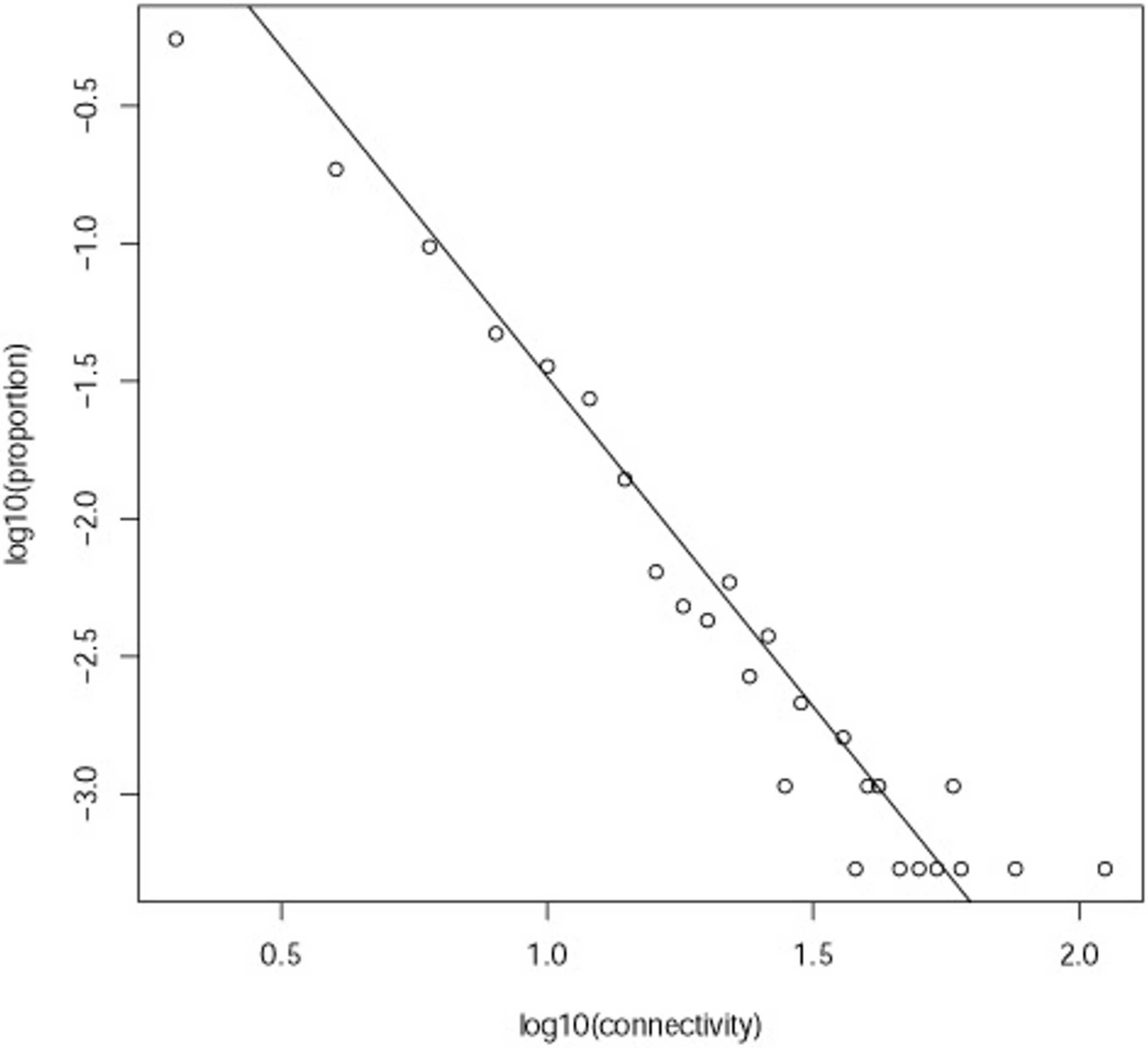


**Figure 5**
**Connectivity distribution of the Y2H yeast protein interaction network**. This picture shows the connectivity distribution of the protein interaction network in Jeong *et al*. [3] inferred from the Y2H data. The imposed regression line is fitted by the LTS method.

The simple error mechanism *S*1 with a high false negative rate produces patterns (Figures 8(a) and 10 (a)) similar to that of the gold standard data (Figure 6). For the Y2H yeast protein interaction network (Figure 5), *S*4 gives the best approximation, but still differs slightly in the probabilities of nodes with small connectivity. This suggests

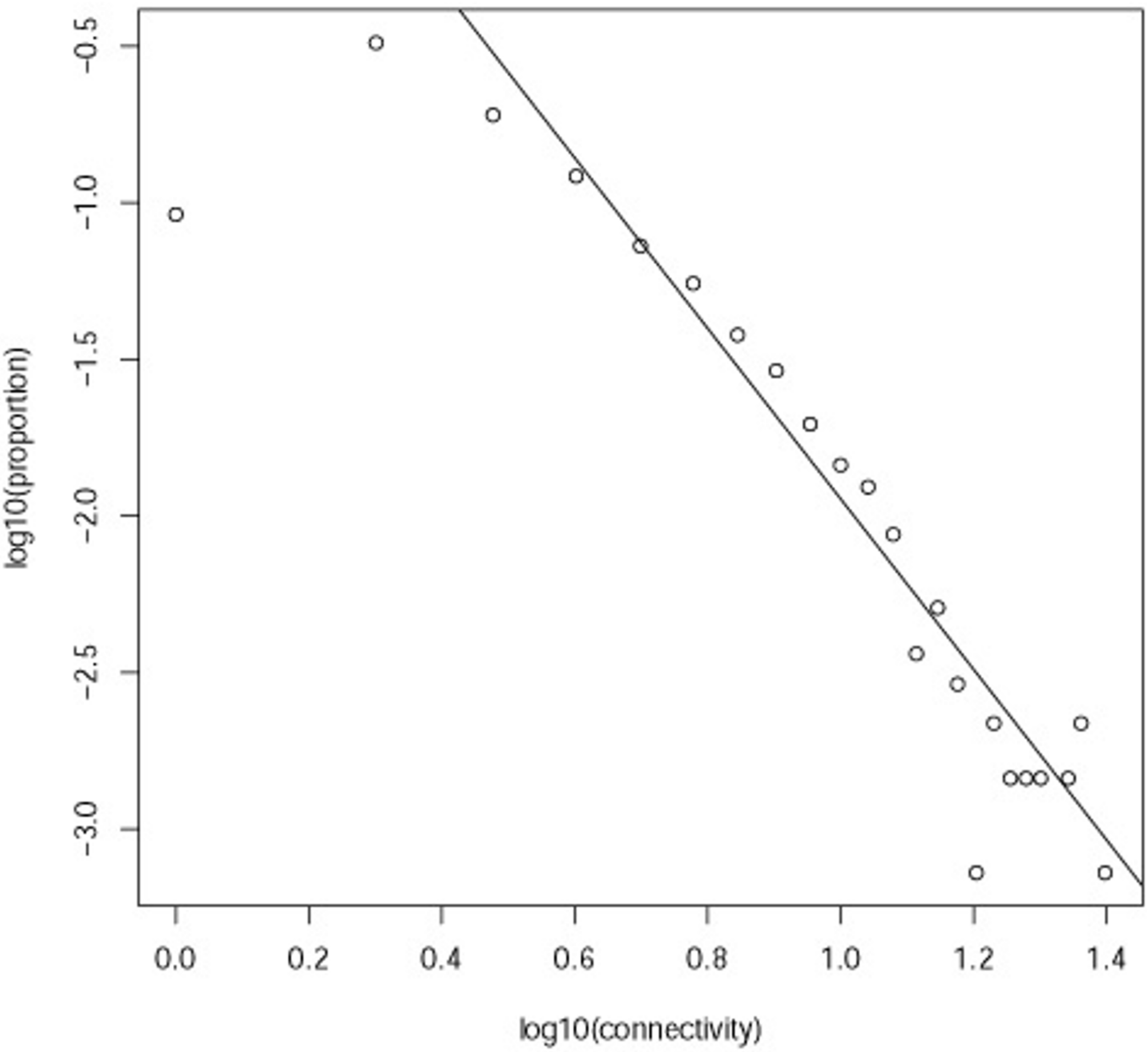


**Figure 6**
**Connectivity distribution of the MIPS yeast protein interaction network**. This picture shows the connectivity distribution of the protein interaction network in Deng *et al*. [6] inferred from the MIPS gold standard data. The imposed regression line is fitted by the LTS method.

that the real error structure of the Y2H analyses may be more complicated than all the simple proposals we have considered.

## Conclusion

This article first investigates the impact of erroneous links on network topological inference. From our theoretical

**Table 1: Nine error mechanisms.**

| Error mechanism | $p_{ij}$ | $q_{ij}$ |
|---|---|---|
| S1 | constant | constant |
| S2 | constant | increasing |
| S3 | constant | decreasing |
| S4 | increasing | constant |
| S5 | increasing | increasing |
| S6 | increasing | decreasing |
| S7 | decreasing | constant |
| S8 | decreasing | increasing |
| S9 | decreasing | decreasing |

**Table 2: Parameter estimates for *Net*$_0$.**

| Parameter | OLS | M-estimation | LTS |
|---|---|---|---|
| log *c* | 1.4600 | 1.7846 | 4.008 |
| $\gamma$ | 2.0918 | 2.1769 | 2.803 |

and simulation results, we find that, under a simple error mechanism, the scale-free property is preserved for moderate connectivities. But the linear pattern is distorted at both the small and large connectivity regions. Accordingly, we recommend to use robust estimators (e.g. LTS) that are more resistant to the outliers at both ends of the distribution to estimate the scale parameter $\gamma$.

Moreover, we have also explored possible error mechanisms of the yeast protein interaction data by simulations considering nine different error mechanisms. The results suggest that changes in the overall false positive rates have little impact on the resulting connectivity distribution, but increasing the overall false negative rates can increase the probability of nodes with small connectivities under some error mechanisms, and hence decrease the scale parameter $\gamma$. The connectivity distribution can be very stable under several error mechanisms when the overall false positive rates and false negative rates change, which suggests that in certain situations the observed data can provide suffcient topological information on the underlying network structure even when the false negative rates are quite high.

The simple error mechanism that assumes that the false positive rate and false negative rate of each protein pair are constants agrees well with the MIPS gold standard data when the false negative rate is high. A different error mechanism is suggested for the Y2H data, where more connected protein pairs tend to have higher false positive rates and lower false negative rates. As this error mechanism provides only a reasonable approximation to the Y2H data, more sophisticated mechanisms might be needed to better capture its error structure.

## Methods

### *Preferential attachment growth model*

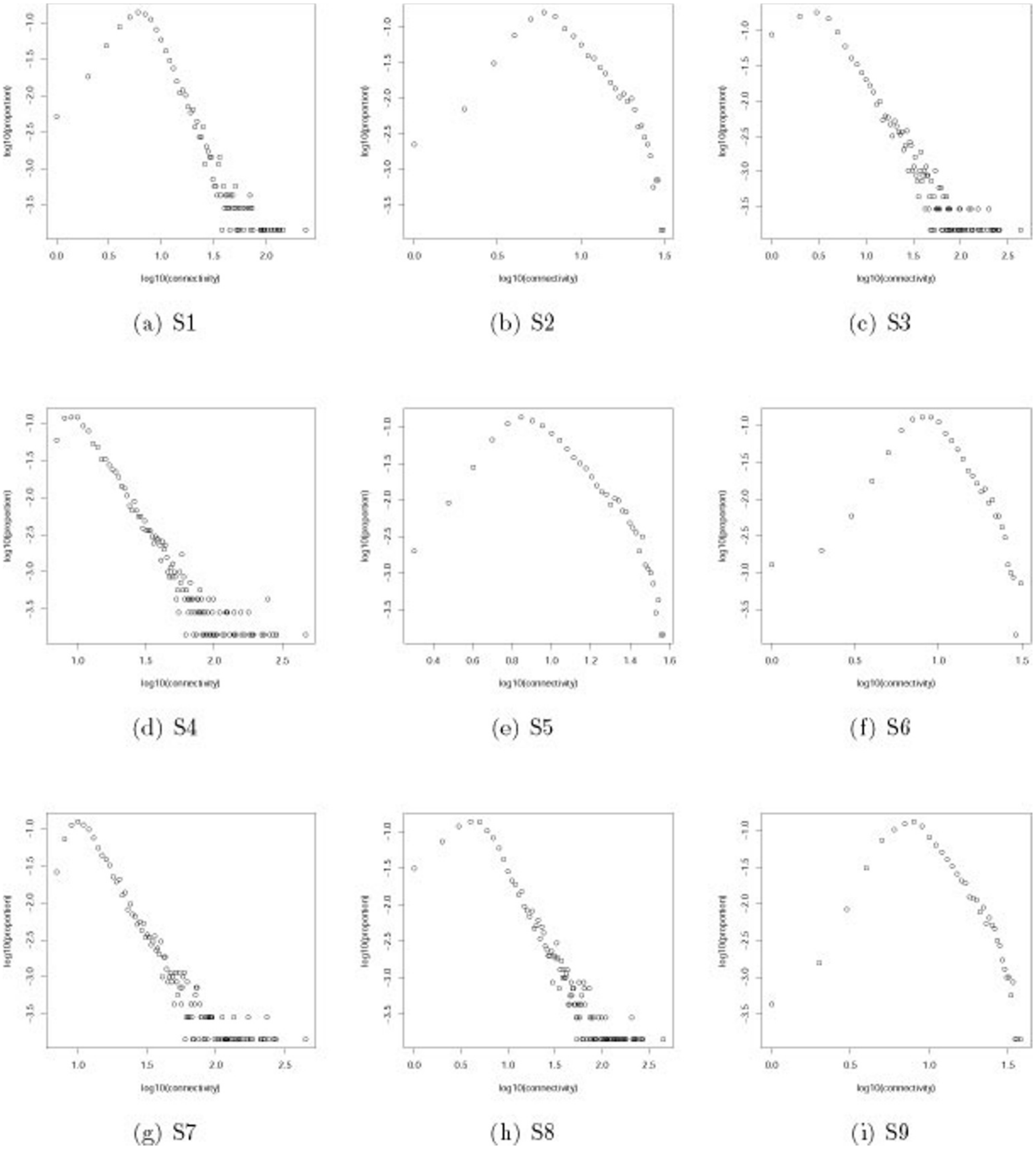
In a series of papers [15,16], Barabási *et al*. demonstrated that a scale-free network could be obtained by growing from a small number of isolated nodes by preferential attachment. The simulation scheme is defined in two steps:

1. *Growth*: starting with a small number ($m_0$) of nodes, add a new node at every time step and connect it to $m$ ($\leq m_0$) nodes already present in the system

2. *Preferential attachment*: The new node is more likely to connect to nodes with larger connectivity. The probability $\Pi_i$ that a new node will be connected to node $i$ depends on its connectivity $k_i$, such that $\Pi(k_i) = \frac{k_i + 1}{\sum_j (k_j + 1)}$.

### *Least Trimmed Squares (LTS)*

The basic idea of LTS estimation is to minimize the sum of $h$ smallest squared residuals instead of all squared residuals in the OLS to achieve robustness and also maintain good effciency. Please refer to [9] for more details of the algorithm, such as practical choices of $h$. In this article, the LTS estimation is performed using the lqs() function implemented in R [17].



(a) S1

(b) S2

(c) S3

(d) S4

(e) S5

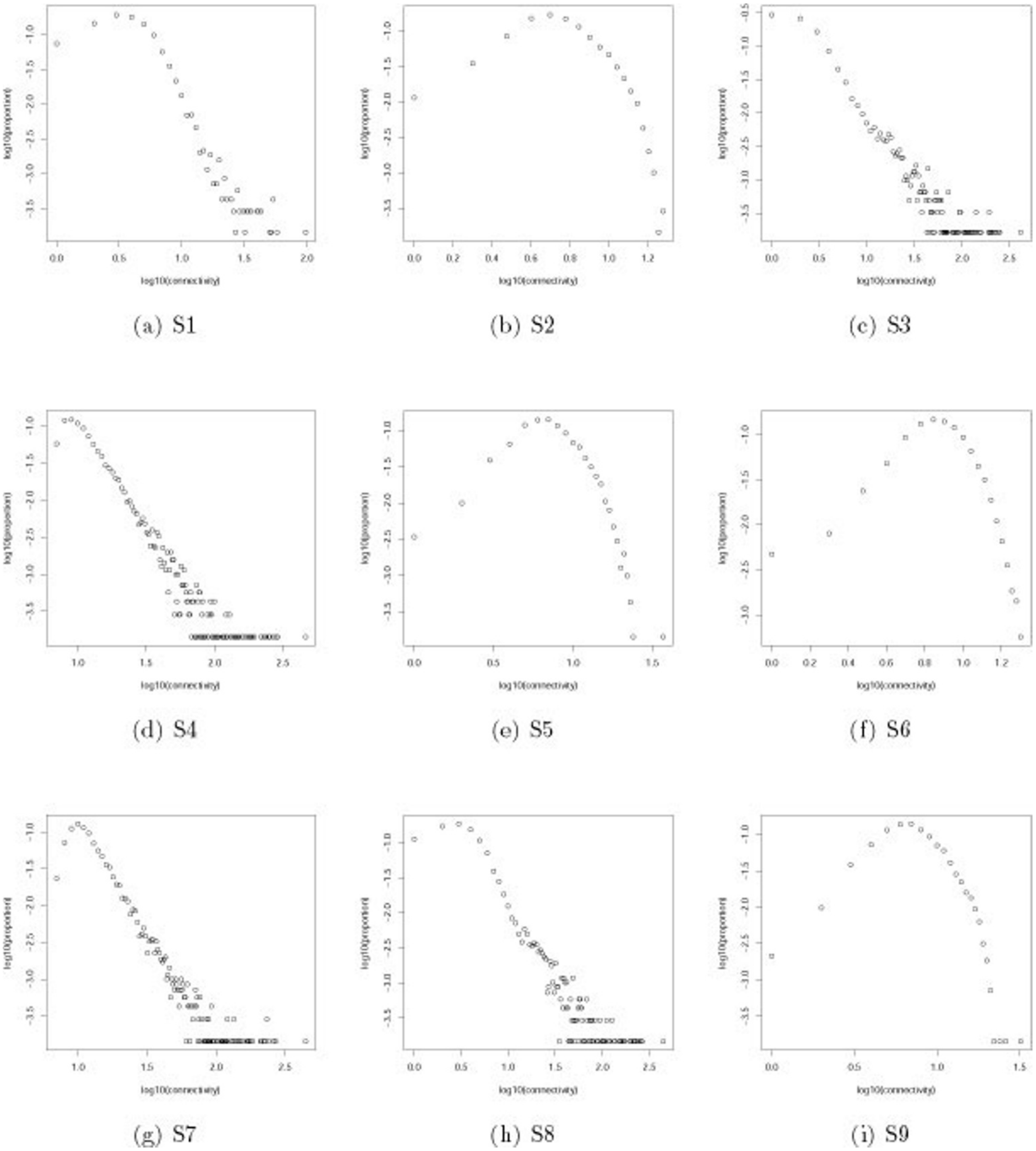
(f) S6

(g) S7

(h) S8

(i) S9

**Figure 7**
Connectivity distribution of *Net*$_0$ perturbed by different error mechanisms ($r_{FP}$ = 0.00025, $r_{FN}$ = 0.5).



(a) S1

(b) S2

(c) S3

(d) S4

(e) S5

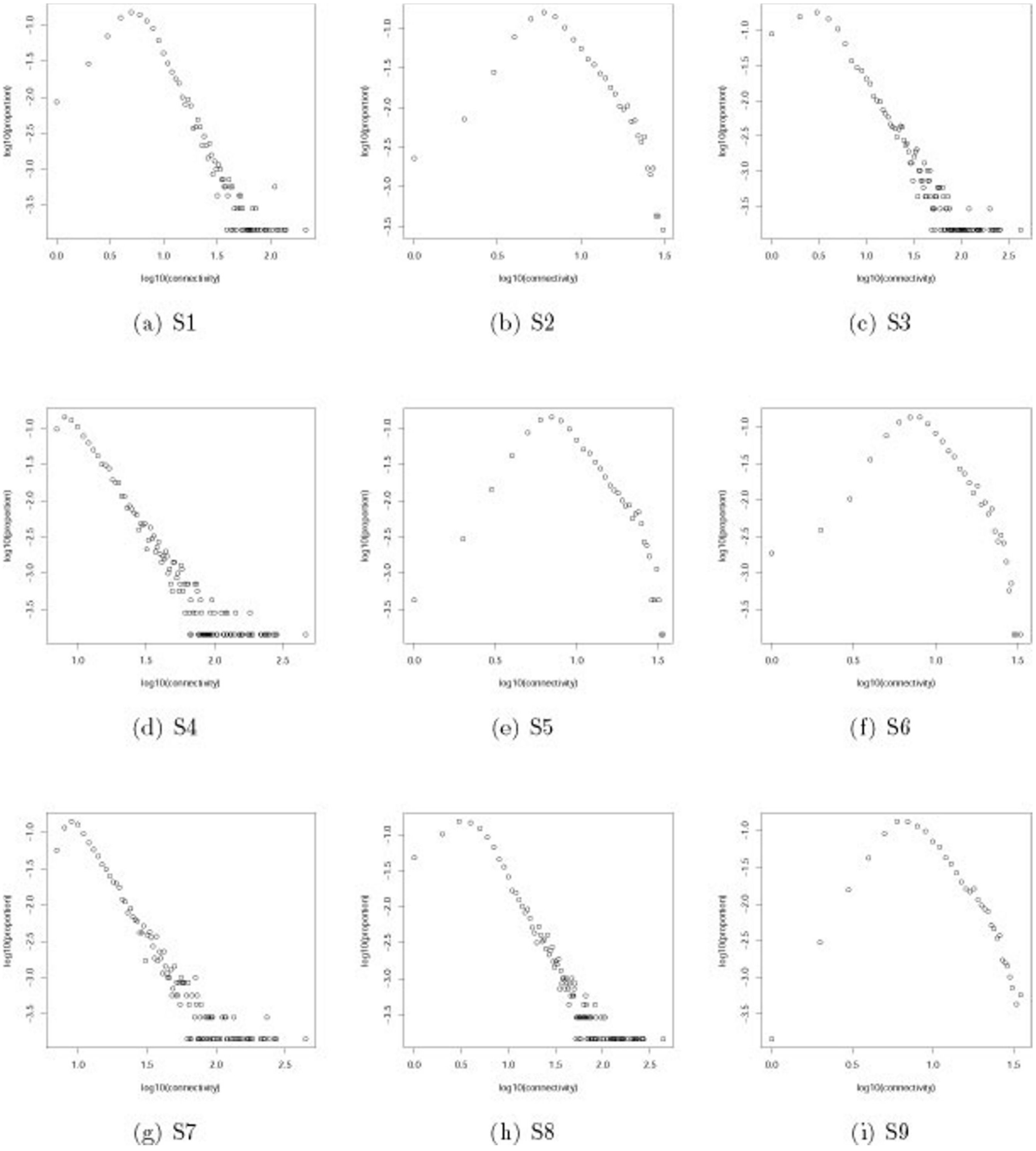
(f) S6

(g) S7

(h) S8

(i) S9

**Figure 8**
Connectivity distribution of *Net*$_0$ perturbed by different error mechanisms ($r_{FP}$ = 0.00025, $r_{FN}$ = 0.8).



(a) S1 (b) S2 (c) S3

(d) S4 (e) S5 (f) S6

(g) S7 (h) S8 (i) S9

**Figure 9**
Connectivity distribution of $Net_0$ perturbed by different error mechanisms ($r_{FP}$ = 0.00015, $r_{FN}$ = 0.5).

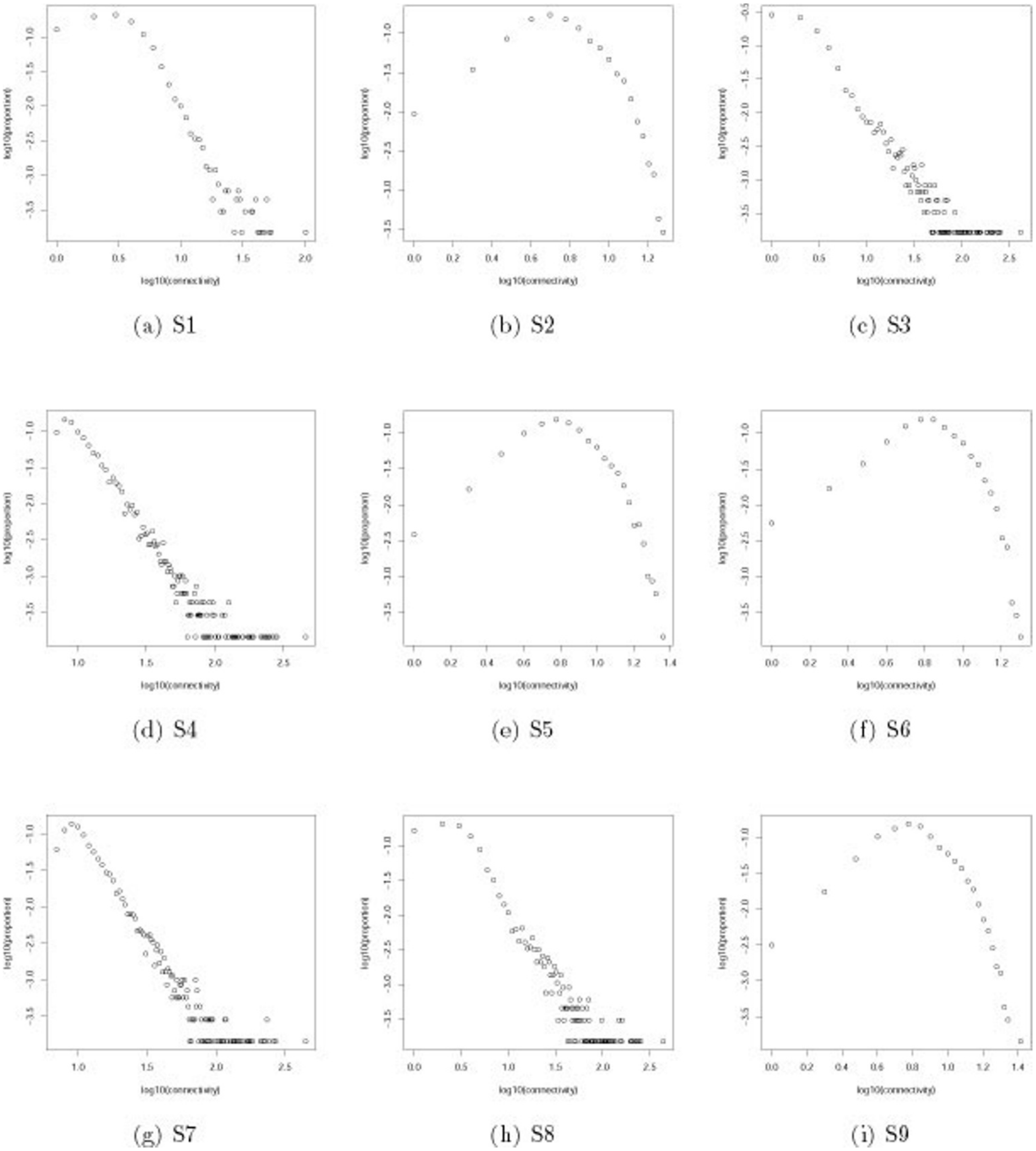


**Figure 10**

Connectivity distribution of $Net_0$ perturbed by different error mechanisms ($r_{FP}$ = 0.00015, $r_{FN}$ = 0.8).

## Authors' contributions

HZ had the initial idea and initiated the study. NL conducted the data analyses, and created all tables and figures, under the supervision of HZ. Both authors read and approved the final manuscript.

## Additional material

### Additional File 1

*Tables of the estimates of the scale parameter γ.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-6-119-S1.pdf]

## Acknowledgements

This work was supported in part by NSF grant DMS 0241160 and NIH grant R01 GM59507.

## References

1. Albert Barabási RAL, Jeong H: **Scale-free characteristics of random networks: The topology of the World Wide Web.** *Physica A* 2000, **281:**69-77.
2. Barabási AL, Jeong H, Néda Z, Revasz E, Schubert A, Vicsek T: **Evolution of the social network of scientific collaborations.** *Physica A* 2002, **311:**590-614.
3. Jeong H, Mason SP, Barabási AL, Oltvai ZN: **Lethality and centrality in protein networks.** *Nature* 2001, **411:**41-42.
4. Albert R, Jeong H, Barabási AL: **Error and attach tolerance of complex networks.** *Nature* 2000, **406:**378-382.
5. Criekinge WV, Beyaert R: **Yeast two-hybrid: State of the art.** *Biol Proced Online* 1999, **2:**1-38.
6. Deng M, Sun F, Chen T: **Assessment of the reliability of protein-protein interactions and protein function prediction.** *Pac Symp Biocomput* 2003:140-151.
7. Dorogovtsev SN, Mendes JFF: *Evolution of Networks* New York: Oxford University Press; 2003.
8. Deng M, Mehta S, Sun F, Chen T: **Inferring domain-domain interactions from protein-protein interactions.** *Genome Research* 2002:1540-1548.
9. Rousseeuw PJ, Leroy AM: *Robust regression and outlier detection* New York: Wiley; 1987.
10. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P: **A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*.** *Nature* 2000, **403:**601-603.
11. Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D: **DIP: the database of interacting proteins.** *Nucleic Acids Res* 2000, **28:**289-291.
12. **Y2H protein interaction network data** [http://www.nd.edu/~networks/database/protein/bo.dat.gz]
13. **MIPS gold standard protein interaction network data** [http://hto-b.usc.edu/~msms/AssessInteraction/MIPSMatchYPD.txt]
14. **Yeast Proteome Database** [http://www.proteome.com/YPDhome.html]
15. Barabási AL, Albert R, Jeong H: **Mean-field theory for scale-free random networks.** *Physica A* 1999, **272:**173-187.
16. Albert R, Barabási AL: **Topology of evolving networks: local events and universality.** *Phys Rev Lett* 2000, **85:**5234-5237.
17. R Development Core Team: **R: A language and environment for statistical computing.** *R Foundation for Statistical Computing, Vienna, Austria* 2004 [http://www.R-project.org].