Supplementary Information

Methods

Interaction retrieval and matrix construction: Proteins can have direct or indirect interactions with one another. Indirect interaction refers to being a member of the same functional module (e.g., transcription initiation complex, ribosome) but without directly binding to one another. In contrast, direct interaction, which is the topic of this paper, refers to two amino acid chains that binds to each other. Obviously, many of these interactions reflect the dynamic state of the cell and are present or absent depending on the particular environment or developmental status of the cell. However, the sum of existing and potential interactions altogether defines the protein network and is ultimately encoded within the genome of a given organism. For our analyses of direct protein-protein interactions in *S. cerevisiae*, three different databases were used: 1. the published Curagen/ Fields Lab database [3](http://curatools.curagen.com/); 2. plus supplemented with newly discovered physical interactions (http://depts.washington.edu/sfields/projects/YPLM/yeast. html), or 3. plus supplemented with the non-overlapping fraction of the Data-base of Interacting Proteins, or DIP (http://dip.doe-mbi.ucla.edu/) [4] for yeast. Note that a large portion of the original DIP was contained in the Curagen database. Thus, the DIP was first screened against the Curagen physical network resulting in a new, small set of interactions, that was then combined with 2. After the interactions were gathered, each dataset (1, 2, and 3) was inserted into an N x N adjacency matrix, where N is the number of interacting proteins in that network. The analyses described in the paper were performed on all three datasets yielding very similar results. In the paper our analyses of dataset 3 is shown.

In all the analyses the protein-protein interactions were considered bi-directional links, with the proteins as the nodes. The most important reason for considering these interactions bi-directional is that they were mostly identified by two-hybrid studies in which many times the function of the particular

yeast protein is not known. Therefore, it is impossible to reliably establish the directionality of the interaction, if any.

At this point there is no database that could offer us reliable data on the directionality of those interactions that indeed should be considered directed. On the other hand, the directionality is not expected to affect the major conclusions of our study. Indeed, our previous studies on the www, citation networks and metabolic networks, which are all directed networks, have shown that the nature of the network can be uncovered whether we look at the directed or the undirected version of the network. If a directed network is scale-free, its non-directed version would also appear scale-free. Thus our result, regarding the nature of the protein interaction network, is not expected to be affected by the network's directness.

Cluster size distribution: The network consists of one big cluster and several isolated clusters. The largest cluster, shown in Fig. 1a in the manuscript, contains 1458 out of 1870 proteins (~78%). Next largest clusters are 4 isolated clusters only with 7 proteins each. After that, there are 168 small isolated clusters with size range from 1 to 6.

Connectivity distribution [$P(k)$]: For each protein in each network, the number of interactions, k, was determined. Histograms were created containing all values of k. Dividing each value of k by the total number of substrates in the given graph we obtained $P(k)$, the probability of a protein having k interactions. The result is shown below, after using logarithmic binning to reduce noise.

There is extensive evidence that in general in networks the scaling is better described by using a law of type $P(k) \sim (k+k_0)^{-\gamma}$. That is, most networks do not follow the power law for small k. Indeed, we have recently demonstrated that if we allow for internal links in the network formation process (that is, if we can sometimes connect two nodes that are already part of the network) then the small k behavior is affected such that there is a small k correction [12]. Similar results were obtained by a number of different

authors as well [13,14]. In the absence of a good understanding of the evolutionary process that produced the protein network, we cannot give k_0 further significance at this point than being a fitting parameter.

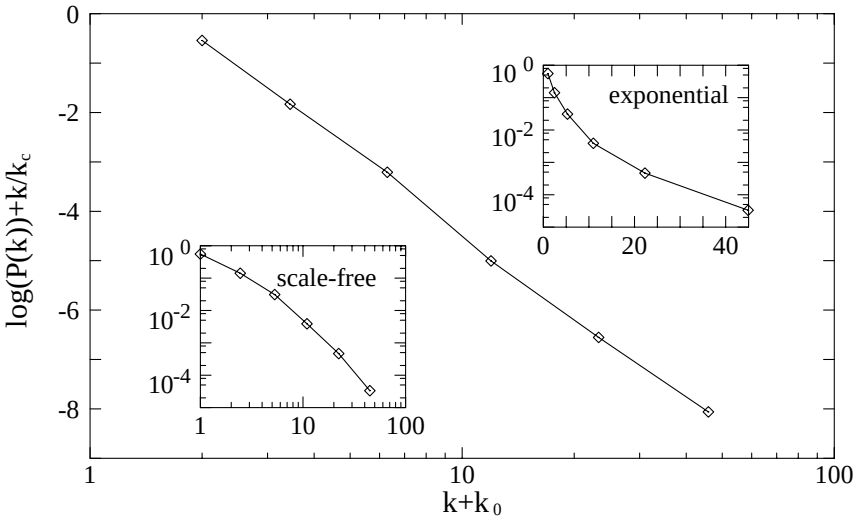

Fig. I. Connectivity distribution *P(k)* of interacting yeast proteins giving the probability that a given protein interacts with *k* other proteins. The main panel shows the connectivity after correction for exponential cutoff indicating that the connectivity follows $P(k) = a \cdot (k + k_0)^{-\gamma} \exp(-(k + k_0)/k_c)$ with $k_0 \cong 1$, $k_c \cong 20$, and $\gamma \cong 2.4$. The linear-log plot (top inset) of the uncorrected *P(k)* indicates that the decay is slower than exponential, while the log-log plot (bottom inset) shows that *P(k)* decays faster than a power-law for large *k*.

Error tolerance and pathway lengths: For all pairs of proteins, the shortest path was determined using a breadth-first search or burning algorithm. A histogram was created, showing the distribution of pathway lengths. From these lengths, the diameter of the network was determined by dividing the total of all path lengths by the total number of paths. To consider error tolerance of the network, we removed the most connected protein and determined the diameter of the remaining network. We then removed the next most connected protein as well, and measured the diameter again. With these measurements, we find that the network diameter increases dramatically as we increase the number of removed proteins, (Fig. II, upper curve) indicating the crucial role of the most connected proteins. On the other hand, under physiological conditions, errors in the networks are expected to occur randomly, i.e., they affect the function of randomly chosen nodes. We have thus performed simulations in which we randomly choose

and eliminate nodes, determining the diameter of the remaining network. As the bottom curve in Fig. II shows, upon the removal of up to 3% randomly selected nodes the diameter remains completely unchanged, indicating that the network structure has a high degree of built-in error tolerance: if nodes randomly malfunction, the overall protein network structure remains essentially unchanged.

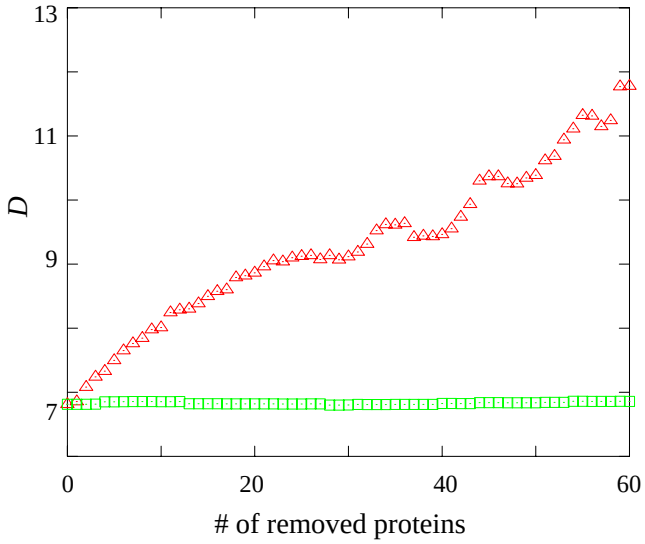

Fig. II. The effect of node removal on the diameter (D), defined as the average of the shortest distance between all pairs of proteins in the largest connected cluster [5]. In the upper curve (Δ) the most connected proteins are removed, while in the bottom curve (□) proteins are removed randomly. M=60 corresponds to ~3% of the total number of interacting proteins in *S. cerevisiae*.

Distribution of lethality: The removal of certain proteins from its proteome can prove fatal to the yeast cell. We hypothesized that those proteins with higher connectivity might have a higher probability of being lethal. Therefore, a list of all proteins and their k values was produced. These were compared with the yeast portion of the Proteome database [11] (www.proteome.com) for lethality and viability. This database contains collated information on yeast mutants generated within systematic gene disruption studies [7, 8], or in studies targeting individual yeast genes. A histogram was created, showing the percentage of lethal proteins for a given k (± SD). In the higher values of k, the results are less reliable, as there are less data points. Thus, $k \geq 18$ were grouped together to produce a more reliable result. Nonetheless, the highest values of k on these histograms are of lower confidence than the lower values.

Statistical analysis of Fig.1c: The linear correlation coefficient is widely used to determine the correlation between two variables. For pairs of quantities (x_i,y_i), [i = 1,2,…N], the linear correlation coefficient r (also called the product-moment correlation coefficient, or Pearson's r) is given by

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{(y_i - \bar{y})^2}}.$$

The value of r lies between –1 and 1. It takes on a value of 1 (-1), termed "complete positive (negative) correlation" when the data point lie on a perfect straight line with positive (negative) slope, with x and y increasing (decreasing) together. The value 1 (-1) holds independent of the magnitude of the slope, while r near zero indicates that the variables x and y are uncorrelated. For our data shown in Fig. 1c the Pearson's r is 0.75, indicating that there is positive correlation (increasing tendency) between the number of links and the percentage of essential proteins. To fit Fig.1c with a function $f(x)$ we use most common employed "chi-square" fitting, $\chi^2 = \sum_{i=1}^{N}\left(\frac{y_i - f(x_i)}{\sigma_i}\right)^2$. Assuming a linear regression, to show this increasing tendency, we change the slope of fitted line from negative to positive, measuring chi-square, as a measure of goodness. Our results show that χ^2 has a clear minimum for positive slope. Finally, we have used several nonlinear fitting functions (power law, logarithmic, exponential), all confirming the positive correlation between lethality and connectivity.

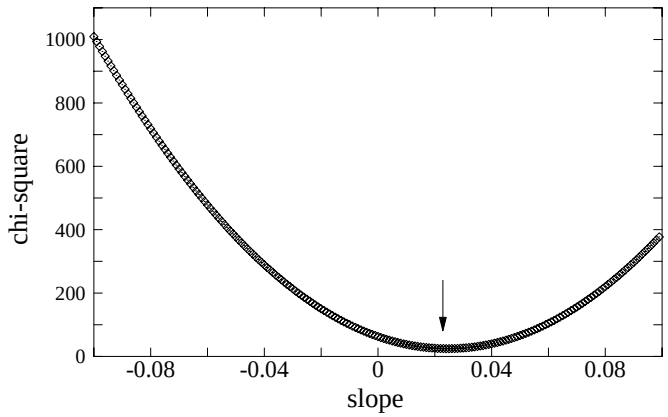

Fig. III. Slope vs goodness of a fit, chi-square.

Full title of references quoted in the manuscript:

1. Hartwell, L.H., Hopfield, J.J., Leibler, S. & Murray, A.W. From molecular to modular cell biology. *Nature* **402**, C47-52 (1999).
2. Eisenberg, D., Marcotte, E.M., Xenarios, I. & Yeates, T.O. Protein function in the post-genomic era. *Nature* **405**, 823-6 (2000).
3. Uetz, P.*, et al.* A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **403**, 623-7 (2000).
4. Xenarios, I.*, et al.* DIP: the database of interacting proteins. *Nucleic Acids Res* **28**, 289-91 (2000).
5. Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. & Barabási, A.-L. The large-scale organization of metabolic networks. *Nature* **407**, 651-654 (2000).
6. Amaral, L.A., Scala, A., Barthelemy, M. & Stanley, H.E. Classes of small-world networks. *Proc. Natl. Acad. Sci.* **97**, 11149-11152 (2000).
7. Winzeler, E.A.*, et al.* Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science* **285**, 901-6 (1999).
8. Ross-Macdonald, P.*, et al.* Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature* **402**, 413-8 (1999).
9. Wagner, A. & Fell, D.A. The small world inside large networks. *preprint # 00-07-041* Santa Fe Institute (2000).
10. Wagner, A. Robustness against mutations in genetic networks of yeast. *Nat Genet* **24**, 355-61 (2000).
11. Costanzo, M.C.*, et al.* The yeast proteome database (YPD) and Caenorhabditis elegans proteome database (WormPD): comprehensive resources for the organization and comparison of model organism protein information. *Nucleic Acids Res* **28**, 73-6 (2000).
12. Albert, R. & Barabási, A.-L. Topology of evolving networks: local events and universality. *Phys. Rev. Lett.* **85**, 5234-7 (2000)
13. Dorogovtsev, S.N., Mendes, J.F.F. & Samukhin, A.N. Structure of Growing Networks with Preferential Linking. *Phys. Rev. Lett.* **85** 4633-6 (2000).
14. Krapivsky, P.L., Redner, S. & Leyvraz. F. Connectivity of Growing Random Networks. *Phys. Rev. Lett.* **85**, 4629-32 (2000).