

# Supplementary

We analyzed the yeast *S. cerevisiae* protein-protein interaction (PPI) data set [1]. It consists of 78390 interactions classified by levels of confidence. The set contains 2455 *high confidence* interactions; these are all interactions seen by at least two methods. These 2455 interactions are on 988 proteins, so they form too small data set to perform statistical analysis on and notice trends. The data set also contains 9400 *medium confidence* interactions, the ones found by one of the following methods: in silico evidence of at least neighborhood 2, or in silico evidence of at least fusion 1, or TAP pulldown data of at least 3, or two-hybrid data of at least 3, or any synthetic lethality interaction. The data set contains 66535 *low confidence* interactions, which include all other yeast protein-protein interactions. Medium and low confidence interactions are ordered so that the ones that were confirmed more times appear first [1].

If in addition to high confidence interactions we were to consider all medium confidence interactions, that would introduce the following bias: the PPI network constructed in this way would contain disproportionately large number of adjacent synthetically lethal pairs for the size of the network. In order to work with a large enough data set with interactions coming from a versatile set of methods, we did most of our analysis on the graph containing all of the high confidence interactions, 6800 of the medium confidence interactions, and 1745 of the low confidence interactions which appear first in the data file [1]. In order to notice trends as the number of interactions grows, we compared several network properties on the following graphs: the graph constructed on the high confidence interactions only (graph A), the graph constructed on the top 11000 interactions that we just described above (graph B), the graph constructed on the top 45000 interactions (graph C), and the graph with all interactions from (graph D) (Table 1) [1].

	number of nodes	number of edges
Graph A	988	2455
Graph B	2401	11000
Graph C	4687	45000
Graph D	5321	78390

Table 1: The four graphs containing top 2455 (only high confidence), top 11000, top 45000, and all interactions respectively.

## 1 Graph Properties of Functional Groups

The analysis described in this section was performed on graph B. We first identified in graph B the following groups of nodes (i.e., proteins) with selected graph properties: 231 articulation points, 141 hubs,

45 hubs which are also articulation points, and 473 siblings (defined in the paper). We linked these graph characteristics of individual proteins in graph B with their function in order to build predictive models.

We analyzed properties of lethal, genetic interaction, and viable proteins, as well as the proteins belonging to the 12 functional groups [2]. There are 558 lethal proteins, 290 genetic interaction, and 1019 viable proteins in graph B. The sizes of the intersections of these three protein groups with the above described five protein groups with selected graph properties are shown in Table 2. The analysis we performed on these three protein groups is summarized in Tables 4, 6, 7, 8, 13, 14, and 15, and described later in this document. There are 98 amino-acid metabolism, 108 cellular fate/organization, 160 cellular organization, 105 energy production, 196 genome maintenance, 258 other metabolism, 176 protein fate, 46 stress and defense, 243 transcription, 98 transcriptional control, 256 translation, 55 transport and sensing, and 562 uncharacterized, proteins in our PPI graph B. The sizes of the intersections of these functional groups with the above described five groups of proteins with selected graph properties are shown in Table 3. The analysis we performed on these functional groups is summarized in Tables 9, 10, 11, 12, 13, 14, and 15, and described later in this document.

	lethal	genetic interaction	viable
art.pts.	55	38	94
hubs	49	21	51
art.pts. $\cap$ hubs	16	8	14
siblings	62	58	202
clusters	82	26	77

Table 2: Number of elements in the intersection of lethal, genetic interaction, and viable protein groups with the five groups of proteins with selected graph properties.

We confirmed scale-free topology previously noted in smaller PPI networks [3] by computing degrees of all nodes in graph B (their mean is 9.16285, minimum is 1, maximum is 114, standard deviation (SD) is 15.52, and the skew is 8.86482) and plotting the histogram of the degree probability distribution (Figure 1). We also computed degree statistics of all of the above five node (protein) groups with selected graph properties, of all of the above functional groups, and of all of the intersections of graph and functional groups. We now describe trends following from this analysis.

Entries in Table 4 show number of lethal, genetic interaction and viable nodes as well as their degree statistics. The cumulative probability distribution functions of degrees of these three protein groups show clear separation between lethal proteins on one side, and genetic interaction and viable proteins on the other (Figure 2). On average, genetic interaction, and viable proteins tend to have degree that is half the degree of the lethal proteins. All three of these degree distributions have large variances and long tails to the right,

	art.pts. $\cap$ hubs	art.pts.	hubs	siblings	clusters
amino-acid metabolism	2	10	7	7	0
cellular fate or organization	2	17	4	11	15
cellular organization	10	26	16	25	28
energy production	2	7	9	17	5
genome maintenance	5	22	10	33	18
other metabolism	3	27	23	53	8
protein fate	3	13	12	44	24
stress and defense	2	6	5	5	3
transcription	2	15	15	26	31
transcriptional control	1	9	3	14	11
translation	2	13	14	68	55
transport and sensing	1	7	3	21	1
uncharacterized	10	55	20	142	23

Table 3: Number of elements in the intersection of the functional groups with the five groups of proteins with selected graph properties.

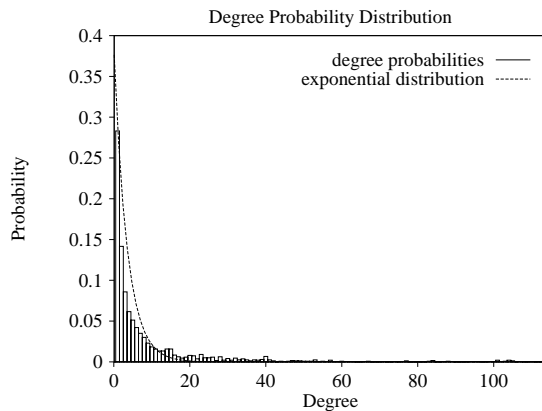


Figure 1: Probability distribution of degrees of graph B.

reflecting degree distribution properties of the whole graph.

Entries in Table 5 show degree statistics of the groups of nodes with specific graph properties. Hubs seem to “boost up” the average degree of articulation points, or equivalently, articulation points lower the average degree of hubs, even though they overlap in only 45 nodes. All of these distributions have large variances and long tails to the right, with hubs having the largest variance and tail, again reflecting degree distribution properties of the whole graph.

Entries in Table 6 are the percentages of “lethality groups” of nodes in the groups of nodes with graph properties. We can see from the table that viable proteins tend to be sibling nodes. This suggests the existence of paths by-passing viable nodes in PPI networks. Table 6 also shows that lethal proteins tend to be articulation points which are hubs, and also they tend to be cluster nodes. More analysis, possibly

	size	mean	min	max	SD	var	skew
lethal	558	14.410	1	114	19.4591	378.655	10.4029
genetic interaction	290	7.3931	1	53	10.7377	115.298	9.09616
viable	1019	6.8449	1	105	12.3247	151.899	9.85924

Table 4: Sizes of lethal, genetic interaction, and viable protein groups and their corresponding degree statistics.

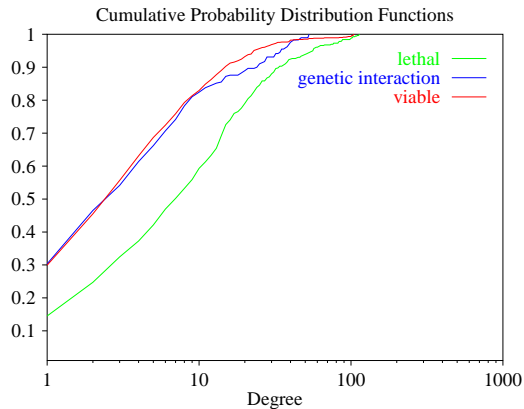


Figure 2: Cumulative probability distribution functions of degrees of lethal, genetic interaction, and viable protein groups in graph B.

on a larger and higher confidence data set, is needed to provide a better separation between lethals and viables with respect to these two properties. However, large percentage of lethal proteins being articulation points that are hubs may suggest that lethal proteins are “bottle necks” of a PPI network whose malfunction disconnects the PPI network, and thus causes death.

Entries in Table 7 show average degrees and standard deviations of lethal, genetic interaction, and viable protein groups as articulation points, hubs, articulation points that are hubs, siblings, and cluster nodes. The table suggests that the members of the three functional groups which happen to be articulation points retain their general average degree properties. However, this does not hold when the functional group nodes belong to the remaining four groups with graph properties. It appears that these four groups of proteins with selected graph properties induce a dramatic degree change in these functional groups, that is, the members of the three functional groups which happen to belong to one of these four “graph groups” exhibit the degree properties similar to the corresponding “graph group”. Cluster nodes seem to exhibit the weakest such behavior.

We ordered nodes of our PPI graph by degree and identified nodes with degrees in the top 3 and 5 percent of this ordering, as well as the nodes of the lowest possible degree (i.e., degree 1; note that around 25% of nodes in graph B are of degree 1). The percentages of lethal, genetic interaction, and viable proteins in

	size	mean	min	max	SD	var	skew
art.pts	231	9.5671	2	99	13.7576	189.273	10.868
hubs	141	22.8369	5	105	21.399	457.952	18.0023
hubs\art.pts	96	24.9375	6	105	21.7206	471.786	20.6003
art.pts\hubs	186	7.44086	2	84	10.7175	114.864	11.5536
art.pts $\cap$ hubs	45	18.3556	5	99	20.2071	408.325	13.9364
siblings	473	6.10994	1	104	14.7779	218.386	7.35202
clusters	225	28.3156	2	114	32.6656	1067.04	11.7306

Table 5: Sizes of the groups of nodes with a specific graph property and their corresponding degree statistics.

	lethal	genetic interaction	viable
art.pts. $\cap$ hubs	35.56	17.78	31.11
siblings	13.12	12.26	42.71
clusters	36.44	11.56	34.22

Table 6: Percentages of lethal, genetic interaction, and viable proteins in the 3 sets of nodes with specific graph properties.

these three degree groups are presented in Table 8. These data confirm our previous observation that lethal proteins tend to have high degree, viable proteins tend to have low degree, and genetic interaction proteins tend to have medium degree.

We repeated the above analysis for all functional groups. Table 9 shows degree statistics of these functional groups in graph B. Translation proteins have the highest and transport and sensing proteins have the lowest average degree. Cumulative probability distribution functions of functional group degrees nicely show these degree separations (Figure 3). This can also be seen from Table 10 which shows the percentages of functional group proteins in the top 3 and 5 percent of nodes ordered by degree, as well as in the set of nodes of degree 1: in the top 3% of all node degrees, 50% are translation proteins, and none of them are amino-acid metabolism, energy production, stress and defense, transcriptional control, or transport and sensing proteins.

Table 11 shows percentages of the functional groups in the five sets with graph properties. It shows that translation proteins tend to be clusters and siblings, cellular organization proteins tend to be articulation points which are hubs, and other metabolism proteins tend to be hubs.

Table 12 shows degree statistics of the intersection of each of the functional groups with each of the five groups of proteins with selected graph properties. We can see a dramatic increase in the average degree of cellular organization proteins which are at the same time hubs and articulation points, exceeding both the functional group and the “graph group” averages. Thus, we could predict that proteins of unknown function which exhibit the property that they are at the same time hubs and articulation points and that their

	lethal	genetic interaction	viable
art.pts.	14.29, $\sigma = 19.11$	6.92, $\sigma = 8.42$	6.69, $\sigma = 5.48$
hubs	29.29, $\sigma = 25.41$	20.00, $\sigma = 15.87$	16.39, $\sigma = 13.82$
art.pts. $\cap$ hubs	22.19, $\sigma = 27.86$	15.50, $\sigma = 13.18$	14.86, $\sigma = 6.59$
siblings	4.03, $\sigma = 13.15$	7.53, $\sigma = 13.44$	6.18, $\sigma = 16.45$
clusters	27.46, $\sigma = 33.30$	15.92, $\sigma = 15.98$	24.75, $\sigma = 33.16$

Table 7: Average degrees and standard deviations ( $\sigma$ ) of the elements in the intersection of the 3 functional group sets and the the 5 sets with graph properties.

	lethal	genetic interaction	viable
top 3	48.61	4.17	20.83
top 5	40.00	9.17	20.00
degree 1 nodes	11.91	12.94	44.85

Table 8: Percentages of our three protein groups in the set of nodes with degrees in the top 3 and 5 percent of all degrees, and in the set of degree 1 nodes in graph B.

average degree is about 3 times the average degree of cellular organization proteins in the PPI graph, and also about 1.25 times the average degree of the set of vertices which are simultaneously articulation points and hubs. For example, there are 15 such articulation point hub proteins, 11 of which have a known functional annotation. Three of these 11 proteins are cellular organization and one is cellular fate/organization. All other functional groups have smaller presence among these 11 proteins. Thus, it is feasible that the 4 proteins of unknown function (BZZ1, YDR214W, RPN13, and ADY3) among these 15 proteins are cellular organization proteins.

Continuing in the same manner, we note that cellular organization proteins which are articulation points have the average degree twice the average degree of all cellular organization points, transcription articulation points have average degree which is about twice the average degree of all articulation points, and translation articulation points have the average degree which is almost a half of the average degree of all translation proteins in the PPI graph. Similarly, translation proteins which are hubs have the average degree twice the average degree of all translation proteins in the PPI graph, and also twice the average degree of all hubs in the PPI graph. Also, siblings of degree one are likely to be cellular fate or organization, genome maintenance, or transcriptional control proteins. On the other hand a subset of siblings which has the average degree similar to the average degree of translation proteins in the PPI graph are likely to be translation proteins, while subsets of sibling nodes which have the average degree similar to the average degree of the whole set of siblings in the PPI graph are likely to be cellular organization, or protein fate proteins. Finally we notice that cluster nodes whose average degree is more than twice the average degree of all cluster nodes in the PPI graph are frequently translation proteins, while cluster nodes whose average degree is less than the

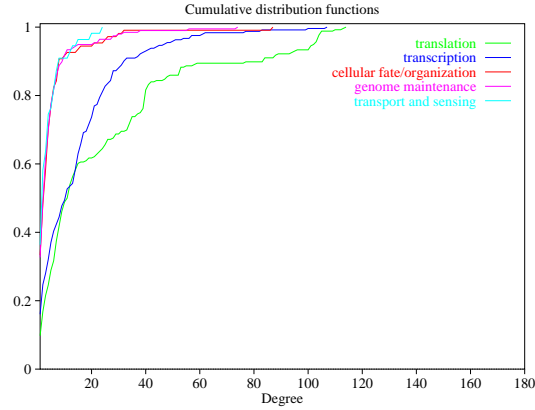


Figure 3: Cumulative probability distribution functions of degrees of 5 functional groups.

	size	mean	min	max	SD	var	skew
amino-acid metabolism	98	7.46939	1	32	6.18146	38.2104	22.7472
cellular fate or organization	108	5.25	1	87	9.81333	96.3014	10.2872
cellular organization	160	7.2875	1	101	14.6873	215.716	8.42152
energy production	105	7.54286	1	46	7.71241	59.4813	15.3843
genome maintenance	196	5	1	74	8.40452	70.6359	10.1437
other metabolism	258	6.10465	1	63	7.64387	58.4287	12.6815
protein fate	176	8.44886	1	101	12.4725	155.563	10.4699
stress and defense	46	4.34783	1	32	5.35503	28.6763	12.2885
transcription	243	15.0165	1	107	16.5642	274.372	13.6707
transcriptional control	98	6.57143	1	38	7.27657	52.9485	12.9562
translation	256	24.8203	1	114	29.5675	874.234	11.3531
transport and sensing	55	4.01818	1	24	4.76286	22.6848	12.1625
uncharacterized	562	6.17794	1	77	11.9402	142.567	7.66294

Table 9: Sizes and corresponding degree statistics of functional groups.

average degree of all genome maintenance nodes in the PPI graph, are more likely to be genome maintenance proteins. These are some of the properties that may be exploited to predict a protein function by analyzing only the graph theoretic properties of proteins in a PPI graph.

In the end we intersected each of the lethal, genetic interaction, and viable protein sets, with each of the functional groups. The numbers of elements in these 39 intersection sets are presented in Table 13. The percentages of each of the functional groups in the sets of lethal, genetic interaction, and viable groups are presented in Table 14. We conclude that amino-acid metabolism, energy production, stress and defense, and transport and sensing proteins are not likely to be lethal. On the other hand, of all functional groups, transcription proteins have the largest presence in the set of lethal nodes on our PPI graph (26.16% of lethals on our PPI graph are transcription proteins), so we conclude that transcription proteins are often lethal. Similarly, amino-acid metabolism, energy production, stress and defense, and transport and sensing pro-

	top 3	top 5	degree 1 nodes
amino-acid metabolism	0	0	1.76
cellular fate or organization	1.39	0.83	5.29
cellular organization	5.56	5.00	6.18
energy production	0	0.83	2.21
genome maintenance	2.78	2.50	9.41
other metabolism	2.78	1.67	8.97
protein fate	2.78	2.50	7.65
stress and defense	0	0	2.21
transcription	15.28	15.83	5.74
transcriptional control	0	0	4.85
translation	50.00	53.33	3.68
transport and sensing	0	0	2.94
uncharacterized	19.44	17.5	36.76

Table 10: Percentages of functional group proteins in the set of nodes with degrees in the top 3 and 5percent of all degrees, and in the set of degree 1 nodes in graph B.

teins are not likely to be genetic interaction, while cellular organization and genome maintenance proteins are largely present among genetic interaction proteins (they each constitute around 16% of genetic interaction proteins). Also, stress and defense, and transport and sensing are not likely to be viable, while other metabolism proteins are largely present among viable proteins (15.41% of viable proteins on our PPI graph are other metabolism proteins).

We show degree statistics of these 39 intersection sets in Table 15. Table 15 indicates that nodes in the intersections of lethals, genetic interactions, and viables with transcription tend to have average degree that is between the average degrees of the intersecting groups. On the other hand, it seems that translation, and transport and sensing groups “override” average degrees of the three groups when intersected with them, and that the three lethality groups “override” the average degree of protein fate group when they get intersected with it. Also, the table indicates that cellular organization, genome maintenance, and other metabolism lower the average degree of lethals, genetic interactions, and viables when intersected with them. In all of the functional groups except for amino-acid metabolism, energy production, stress and defense, transcription, and translation, intersection sets with lethals tend to have the average degree that is twice of the average degree of the intersection sets of the functional groups with genetic interactions, and with viables, reflecting our previous observation that lethal proteins have average degree that is twice the average degree of genetic interaction and viable proteins. Similar to the above, all of these graph characteristics of functional groups may be useful for predicting functionality of proteins by knowing their graph properties.

	art.pts. $\cap$ hubs	art.pts.	hubs	siblings	clusters
amino-acid metabolism	4.44	4.33	4.96	1.48	0
cellular fate or organization	4.44	7.36	2.84	2.33	6.67
cellular organization	22.22	11.26	11.35	5.29	12.44
energy production	4.44	3.03	6.38	3.59	2.22
genome maintenance	11.11	9.52	7.09	6.98	8.00
other metabolism	6.67	11.69	16.31	11.21	3.55
protein fate	6.67	5.63	8.51	9.30	10.67
stress and defense	4.44	2.60	3.55	1.06	1.33
transcription	4.44	6.49	10.64	5.50	13.78
transcriptional control	2.22	3.90	2.13	2.96	4.89
translation	4.44	5.63	9.93	14.38	24.44
transport and sensing	2.22	3.03	2.13	4.44	0.44
uncharacterized	22.22	23.81	14.18	30.02	10.22

Table 11: Percentages of functional group proteins in the sets of nodes with graph properties.

## 2 Predicting genetic interaction pairs

Based on the combined properties of genetic interaction genes, we constructed a model to predict novel genetic interaction pairs. As in any model, we can optimize its performance, by selecting appropriate parameters, in our case filters. We can aim at high recall, that is, predicting all feasible genetic interaction pairs with possibly many false positives; we can also aim at high precision, that is, predicting only highly likely genetic interaction pairs, but possibly missing many potential candidates.

We analyzed properties of known genetic interaction pairs in graph B and constructed a predictive model, as described in the paper. Supplementary Data Table 16 contains the 3225 directly connected pairs whose removal disconnects the graph. Supplementary Data Table 17 contains three sets of 3225 random pairs each and was used to evaluate the quality of the 3225 predicted genetic interaction pairs, as described in the paper. Supplementary Data Table 18 contains the 1234 predicted genetic interaction pairs. Supplementary Data Table 19 contains three sets of 1234 random pairs each and was used to evaluate the quality of the 1234 predicted genetic interaction pairs, as described in the paper.

## 3 Clusters

We ran the HCS [4] [5] algorithm on graphs A, B, and C, as described in section 2.4 of the System and Methods section of the paper. For graph A, we obtained 46 clusters containing the total of 250 nodes and having the largest clusters of sizes 31, 17, 12, 12, 8, 8, 7, and 7. For graph B we obtained 31 clusters containing the total of 225 nodes and having the largest clusters of sizes 65, 22, 22, and 15 and all other

	art.pts. $\cap$ hubs	art.pts.	hubs	siblings	clusters
amino-acid metabolism	16, $\sigma = 1.41$	8.00, $\sigma = 5.10$	17.57, $\sigma = 6.45$	1.43, $\sigma = 0.53$	N/A, N/A
cellular fate or organization	15.50, $\sigma = 13.44$	5.76, $\sigma = 5.33$	34.75, $\sigma = 35.78$	1.00, $\sigma = 0.00$	7.80, $\sigma = 8.33$
cellular organization	23.30, $\sigma = 28.48$	14.58, $\sigma = 24.19$	23.31, $\sigma = 24.37$	6.76, $\sigma = 19.97$	7.46, $\sigma = 18.41$
energy production	11.00, $\sigma = 2.83$	9.14, $\sigma = 6.09$	13.00, $\sigma = 7.18$	2.94, $\sigma = 3.03$	6.40, $\sigma = 2.30$
genome maintenance	12.80, $\sigma = 8.98$	6.09, $\sigma = 6.13$	20.40, $\sigma = 21.10$	1.67, $\sigma = 1.14$	4.17, $\sigma = 4.34$
other metabolism	9.67, $\sigma = 3.06$	6.30, $\sigma = 3.91$	17.17, $\sigma = 14.99$	2.75, $\sigma = 5.57$	12.63, $\sigma = 19.21$
protein fate	12.33, $\sigma = 4.62$	9.31, $\sigma = 5.78$	18.75, $\sigma = 9.55$	6.09, $\sigma = 15.77$	27.00, $\sigma = 22.10$
stress and defense	7.00, $\sigma = 1.41$	4.50, $\sigma = 2.26$	13.60, $\sigma = 10.60$	2.60, $\sigma = 1.14$	4.67, $\sigma = 2.89$
transcription	48.50, $\sigma = 48.79$	20.87, $\sigma = 25.16$	24.80, $\sigma = 20.66$	3.46, $\sigma = 7.26$	25.94, $\sigma = 25.52$
transcriptional control	10.00, N/A	7.78, $\sigma = 4.92$	14.67, $\sigma = 8.08$	1.36, $\sigma = 0.93$	16.36, $\sigma = 8.15$
translation	23.50, $\sigma = 23.33$	15.62, $\sigma = 14.69$	45.14, $\sigma = 31.32$	26.16, $\sigma = 25.86$	70.80, $\sigma = 28.63$
transport and sensing	20, N/A	6.86, $\sigma = 6.94$	11.33, $\sigma = 7.57$	2.62, $\sigma = 2.13$	4.00, N/A
uncharacterized	19, $\sigma = 21.32$	8.8, $\sigma = 13.75$	24.7, $\sigma = 23.59$	1.54, $\sigma = 3.36$	12.43, $\sigma = 16.15$

Table 12: Average degree and standard deviation of the overlaps between functional group proteins and proteins with selected graph properties.

clusters of size 6 or less. For graph C, we got only 7 clusters containing the total of 93 nodes and having the largest cluster of size 69 and all other clusters of size 6 or less. All of the clusters of these three graphs are presented in Supplementary Data Table 20. The number of nodes in the intersection of clusters of these three graphs is illustrated in Figure 4. The 15 vertices common to clusters of all three graphs belong to different clusters of graph A, to the 65-node cluster of graph B, and to the 69-node cluster of graph C.

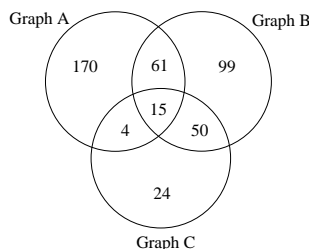


Figure 4: The number of nodes in the intersection of clusters of graphs A, B, and C.

We examined the density of clusters of graph B. By density of a graph we mean the number of edges of the graph compared to the maximum possible number of edges in a graph on the same number of nodes. As mentioned in the paper, all clusters of graph B on 3, 4, 5, and 6 nodes were complete graphs (defined in the paper), except for two 5-node clusters which lack one edge each to be complete graphs. Graph B also has one 15-node and 103-edge cluster which lacks 2 edges to be a complete graph on 15 nodes. Thus, these are already as dense graphs as they can be. The only remaining clusters in graph B are the 65-node cluster on 1646 edges (it lacks 434 edges to be a  $K_{65}$ ), the 22-node cluster on 222-edges (it lacks 9 edges to be a  $K_{22}$ ), and the 22-node cluster on 202-edges (it lacks 29 edges to be a  $K_{22}$ ).

	lethal	genetic interaction	viable
amino-acid metabolism	5	7	59
cellular fate or organization	19	39	43
cellular organization	57	48	44
energy production	4	9	80
genome maintenance	76	49	67
other metabolism	42	12	157
protein fate	58	31	72
stress and defense	1	9	27
transcription	146	33	53
transcriptional control	15	21	55
translation	60	22	96
transport and sensing	7	7	37
uncharacterized	59	4	212

Table 13: Number of elements in the intersection of functional groups with lethal, genetic interaction, and viable groups.

The 65-node cluster of graph B contains a  $K_{19}$ , the denser 22-node cluster contains a  $K_{20}$ , and the sparser 22-node cluster contains a  $K_{16}$ . If we relax the connectivity condition to asking an induced subgraph to have at least  $7n/8$  edges in the smallest set of edges whose removal disconnects the subgraph, where  $n$  is the number of nodes in the subgraph we are considering, the 65-node cluster contains a 36-node and 604-edge subgraph (it lacks 26 edges to be a  $K_{36}$ ), and the sparser 22-node cluster contains a 17-node and 135-edge subgraph (it lacks 1 edge to be a  $K_{17}$ ), while when we apply this connectivity condition to the denser 22-node cluster we get the  $K_{20}$  found in the first test. If we relax the connectivity condition even further to asking an induced subgraph to have at least  $3n/4$  edges in the smallest set of edges whose removal disconnects the subgraph, the 65-node cluster contains a 43-node and 827-edge subgraph (it lacks 76 edges to be a  $K_{43}$ ), the denser 22-node cluster already is of this density, and the sparser 22-node cluster contains a 19-node and 164-edge sub-cluster (7 edges short from being a  $K_{19}$ ). From these observations we conclude that all of our PPI clusters have very dense “cores” surrounded by a less dense neighborhood. The function of these cores of clusters is yet to be examined.

Supplementary Data Table 21 shows overlaps of the 31 HCS clusters of graph B with MIPS database protein complexes. Supplementary Data Table 20 also shows uniformity of function within the HCS clusters of graphs A, B, and C. P-values for clusters of graph B, computed as described in section 2.4 of the paper, are presented in Supplementary Data Table 22. Since graph B had 31 clusters, we constructed three sets of 31 random clusters on the same number of nodes as the HCS clusters of graph B. We used these random cluster sets to evaluate quality of clusters obtained by the HCS algorithm, as described in the paper. These

	lethal	genetic interaction	viable
amino-acid metabolism	0.90	2.41	5.79
cellular fate or organization	3.41	13.45	4.22
cellular organization	10.22	16.55	4.32
energy production	0.72	3.10	7.85
genome maintenance	13.62	16.90	6.58
other metabolism	7.53	4.14	15.41
protein fate	10.39	10.69	7.07
stress and defense	0.18	3.10	2.65
transcription	26.16	11.38	5.20
transcriptional control	2.69	7.24	5.40
translation	10.75	7.59	9.42
transport and sensing	1.25	2.41	3.63
uncharacterized	10.57	1.38	20.80

Table 14: Percentages of functional group elements in the sets of lethal, genetic interaction, and viable nodes.

three sets of random clusters are presented in Supplementary Data Table 23. P-values for these three sets of random clusters are presented in Supplementary Data Table 24. Since random clusters are more functionally heterogeneous than the identified clusters, we computed P-values corresponding to different functional groups which are present in a random cluster (Supplementary Data Table 24).

## 4 Shortest paths

We first considered what happens to the diameter of a PPI graph as we increase the number of interactions. Thus we analyzed shortest paths of graphs A, B, C, and D. We found the shortest paths between each pair of vertices in each of these four graphs. Table 25 shows the summary of our results. As expected, the diameter of the graph (the longest shortest path), as well as the average shortest path length and the variance, decrease as the size of the graph increases. The shortest path distributions for these four graphs are all severely skewed to the right and the skewness grows with the graph size.

We further analyzed graph B as described in section 2.5 of the System and Methods section of the paper. We found that among the top 5% of the most frequent vertices, 16.95% are transcription, 12.71% are cellular organization, 11.86% are other metabolism, and 11.02% are translation proteins. Also, more than half of the top 5% of these most frequent vertices are not viable, indicating the importance of these proteins. The resulting list of the top 5% of the most frequent proteins is presented in Supplementary Data Table 26.

	lethal	genetic interaction	viable
amino-acid metabolism	8.40, $\sigma = 6.66$	11.14, $\sigma = 10.17$	7.36, $\sigma = 5.38$
cellular fate or organization	13.58, $\sigma = 20.25$	3.00, $\sigma = 2.28$	4.14, $\sigma = 5.05$
cellular organization	10.07, $\sigma = 19.68$	5.58, $\sigma = 9.64$	4.32, $\sigma = 5.06$
energy production	6.75, $\sigma = 6.55$	11.11, $\sigma = 13.75$	7.66, $\sigma = 7.25$
genome maintenance	7.55, $\sigma = 12.08$	3.00, $\sigma = 2.57$	3.75, $\sigma = 5.02$
other metabolism	9.95, $\sigma = 11.44$	4.58, $\sigma = 4.42$	5.14, $\sigma = 5.68$
protein fate	12.52, $\sigma = 13.75$	6.48, $\sigma = 8.04$	6.40, $\sigma = 13.05$
stress and defense	3, N/A	6.33, $\sigma = 9.87$	4.00, $\sigma = 3.88$
transcription	18.14, $\sigma = 18.75$	13.28, $\sigma = 12.53$	8.38, $\sigma = 8.62$
transcriptional control	10.00, $\sigma = 6.25$	6.48, $\sigma = 7.23$	5.13, $\sigma = 5.86$
translation	26.00, $\sigma = 32.66$	25.73, $\sigma = 18.16$	21.94, $\sigma = 29.37$
transport and sensing	5.43, $\sigma = 5.00$	3.00, $\sigma = 2.16$	4.24, $\sigma = 5.25$
uncharacterized	16.31, $\sigma = 19.03$	1.75, $\sigma = 0.50$	4.09, $\sigma = 7.58$

Table 15: Average degree and standard deviation of nodes in the intersection of lethal, genetic interaction, and viable proteins with the protein functional groups.

#### PredictedGeneticInteractionPairs1.txt

Table 16: Predicted genetic interaction pairs: 3225 directly connected pairs whose removal disconnects the graph.

## 5 MAPK signaling pathway analysis

MAPK signaling pathway exhibits linearity in structure. Thus, we analyzed it in order to make a predictive model for linear pathways as described in section 2.6 of the System and Methods section of the paper. The predicted pathways are presented in Supplementary Data Table 27. The list of the end node pairs of these pathways with one end node being a transcription factor and the other end node being a transmembrane or sensing protein is presented in Supplementary Data Table 28. The list of the end node pairs of the predicted pathways with one end node being a transcription factor and the other end node being uncharacterized is presented in Supplementary Data Table 29.

## References

- [1] C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887):399–403, 2002.
- [2] H. W. Mewes, D. Frishman, U. Guldener, G. Mannhaupt, K. Mayer, M. Mokrejs, B. Morgenstern, M. Munsterkotter, S. Rudd, and B. Weil. Mips: a database for genomes and protein sequences. *Nucleic Acids Res*, 30(1):31–4, 2002.

### **RandomPairs1.txt**

Table 17: Three sets of 3225 random pairs each.

### **PredictedGeneticInteractionPairs2.txt**

Table 18: Predicted genetic interaction pairs with exactly one protein in each pair being a know genetic interaction protein.

- [3] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabasi. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–4, 2000.
- [4] E. Hartuv and R. Shamir. An algorithm for clustering cdna for gene expression analysis. In *RECOMB'99*, 1999.
- [5] E. Hartuv and R. Shamir. A clustering algorithm based on graph connectivity. *Information Processing Letters*, 76(4-6):175–181, 2000.

### **RandomPairs2.txt**

Table 19: Three sets of 1234 random pairs each.

### **PredictedClusters.txt**

Table 20: The list of all clusters identified on graphs A, B, and C respectively.

### **PredictedClustersMIPSOverlap.txt**

Table 21: Overlaps of the graph B clusters with the MIPS protein complexes.

### **PredictedClustersPvalues.txt**

Table 22: P-values for clusters of the graph B.

### **RandomClusters.txt**

Table 23: Three sets of random clusters.

### **RandomClustersPvalues.txt**

Table 24: P-values for the three sets of random clusters.

	number of paths	mean	min	max	SD	var	skew
Graph A	329,710	5.19392	1	14	1.82624	3.33514	201.325
Graph B	4,565,198	4.9316	1	15	1.5314	2.3452	287.079
Graph C	21,173,896	3.65131	1	10	0.962623	0.926643	459.781
Graph D	27,609,836	3.45237	1	9	0.914903	0.837048	452.885

Table 25: Statistics on the shortest path lengths for the four graphs.

### **MostFrequentProteins.txt**

Table 26: The list of the top 5% of the most frequent proteins.

### **PredictedPathways.txt**

Table 27: The list of predicted pathways.

### **SensingToTranscription.txt**

Table 28: The list of end nodes of predicted pathways between a transcription factor on one end and a transmembrane or sensing protein on the other.

### **UncharacterizedToTranscription.txt**

Table 29: The list of end nodes of predicted pathways between a transcription factor on one end and an uncharacterized protein on the other.