Protein-protein interaction network of *S.Cerevisiae*

Mustapha Bousakla

1 Introduction

The importance of complex networks analysis relies on the fact that it enables a deeper understanding of various real systems, ranging from technological to biological networks. A network is whatever system composed of multiple elements interacting pairwise and a graph is its mathematical representation. More specifically, a network $G = \{V, E\}$ is a set of $\{V\}$ of objects, referred as vertices (or nodes), connected by a set $\{E\}$ of edges. Many systems can be analyzed by means of complex networks, such as gene regulatory circuits, metabolic pathways, large neuronal ensembles, ecological food webs, the electric power grid of a country, etc.

This work is devoted to the study of the protein-protein interaction network in *S.cerevisiae*. And why is this important? Traditionally, protein interactions were studied individually by genetic, biochemical and biophysical techniques focusing on a few proteins at a time, but it is believed now that separating the genetic and biochemical circuitry of a cell prevents us from further understanding of the complicated biological processes involved. Hence, protein-protein interactions are key determinants of protein function and indeed it is believed that all biological processes are essentially carried out through protein-protein interactions. For instance, specific features of this network such as its central nodes may reveal relevant functional groups and predict the function of uncharacterized proteins based on their "ranking" or classification in the network.

2 Methods

In this section it will be briefly reviewed all the network models and parameters presented in the following sections. All calculation were done using the Python NetworkX library and the codes were run in a Jupyter Notebook.

The first computed parameter is the *density*:

$$d = \frac{m}{n(n-1)} \tag{1}$$

being n and m the number of nodes and edges respectively. This expression is specific for undirected networks like ours.

The *average clustering coefficient* (Watts-Strogatz clustering coefficient) estimates the mean clustering in the network. The local clustering of each node is the fraction of triangles that actually exist over all possible triangles in its neighborhood. The average clustering coefficient of a network is the mean of all local clusterings c_ν:

$$C = \frac{1}{n} \sum_{\nu \in G} c_\nu \tag{2}$$

The *transitivity* or clustering coefficient of a network is a measure of the tendency of the nodes to cluster and connect. The transitivity proposed by Newman, Strogatz and Watts is

$$T = 3\frac{N_{\text{triangles}}}{N_{\text{triads}}} \tag{3}$$

being $N_{\text{triangles}}$ the number of triangles and N_{triads} the number of "triads", two edges that share vertex.

The *average path length* is the average number of steps of the shortest paths between all possible pairs of nodes. The mathematical expression used is

$$a = \sum_{s,t \in V} \frac{d(s,t)}{n(n-1)} \tag{4}$$

where $d(s,t)$ is the shortest path or distance between nodes s and t.

On the other hand, the diameter of the network is the maximum eccentricity, that is, the longest distance between two nodes from all possible densities.

Many different *centrality* measures of the network are shown in the results of the following section. Centrality measures the importance or centrality of a node and there are several ways with which a node can be classified as "central":

- The *degree centrality* classifies the nodes with highest degree.
- The *closeness centrality* of a node u measures the average of its inverse distance to all other nodes. Thus, nodes with a high closeness have the shortest distances to all other nodes. The specific equation is

$$C(u) = \frac{n-1}{\sum_{\nu=1}^{n-1} d(v,u)} \tag{5}$$

- The *betweenness centrality* is a way of detecting the nodes that serve as a bridge from one part of a network to another. The algorithm computes the betweenness centrality $c_B(u)$ of a node u by means of the shortest paths $\sigma(s,t)$ between nodes s and t and the shortest path lengths that pass through the node u, denoted as $\sigma(s,t|u)$. More specifically,

$$c_B(u) = \sum_{s,t \in V} \frac{\sigma(s,t|u)}{\sigma(s,t)} \tag{6}$$

- The *eigenvector centrality* of a node measures its importance depending on the importance of its neighbours. Thus, a node connected to neighbours with high degree will have high eigenvector centrality. It is numerically computed by solving an eigenvalue equation ($A\mathbf{x} = \lambda\mathbf{x}$), being A the adjacency matrix and λ the eigenvalue.
- The *Katz index centrality* generalizes the concept of eigenvector centrality as it also computes the centrality of a node based on the centrality of its neighbors. The Katz centrality of a node i is $x_i = \alpha \sum_j A_{ij} x_j + \beta$, where the parameter β controls the initial centrality and the parameter α is taken smaller than the inverse of the maximum eigenvalue, $\alpha < \frac{1}{\lambda_{\max}}$.

- The *PageRank centrality* is an algorithm that was originally created for web pages and it roughly calculates the centrality of a node depending on the structure of the incoming links.

- The *subgraph centrality* of a node n characterizes its participation in all subgraphs in a network and it is the sum of weighted closed walks starting and ending at node n. The NetworkX algorithm computes eigenvectors and eigenvalues of the adjacency matrix and the details can be found in Ref [1].

Another graph property computed in this work is the *degree assortativity*, the tendency of high degree nodes to be connected to other high degree nodes. It is usually quantified by the Pearson correlation coefficient of the degree–degree correlation (Eq 21 of Ref [2]):

$$r = \frac{\sum_{xy} xy(e_{xy} - a_x b_x)}{\sigma_a \sigma_b} \tag{7}$$

More details about the meaning of the parameters can be found in Ref [2]. What is important to us is that r lies in the range $[-1, 1]$ and $r = -1$ and $r = 1$ correspond to perfect dissortativity and perfect assortativity respectively.

The last network property to be calculated is the *bipartivity index* explained in Ref [3]. This index lies in the interval $[0.5, 1]$: 0.5 corresponds to the not bipartite case and 1 is the completely bipartite case.

Regarding the community detection of the network, the *modularity* Q of each community is

$$Q = \sum_{c=1}^{n} \left[\frac{L_c}{m} - \gamma \left(\frac{k_c}{2m} \right)^2 \right] \tag{8}$$

where the sum iterates over all communities c, m is as usual the number of edges, L_c the number of intracommunity links, k_c the sum of degrees of the nodes in community c, and λ the resolution parameter. The modularity ranges from -1 to 1 and positive values close to 1 indicate the presence of communities.

In this work I generated as well 10 random realizations of two important and well-known type of random networks, Erdös-Rény (ER) and Barabási-Albert (BA), and I computed the averages and standard deviations of all the parameters previously explained. In all realizations the number of nodes and edges n and m where the same or approximately the same as the protein interaction network ($n = 2224$, $m = 7049$) so that I could compare both results. In the Erdös-Renyi case I used the $G(n, p)$ model, in which a network is built by connecting nodes randomly and each edge is added with probability p independently from every other edge. The expected number of edges is

$$\overline{m} = \frac{n(n-1)}{2} p \tag{9}$$

This equation enables us to find the probability p to be set in the algorithm in terms of n and m. In my case, I found $p = 0.00285$ and since equation (8) involves an average number of edges $\overline{m}$ (not m itself), the 10 realizations of ER networks did not have exactly the same number of edges as the protein interaction network. The relative difference was, however, negligible, and in fact the 10 random networks generated had in average 7048.4 edges, while the *S. Cerevisiae* network has $m = 7049$.

The Barabási-Albert network is based on a preferential attachment model: a network of n nodes is grown by attaching new nodes each with m edges that are preferentially linked to high degree nodes. More precisely, the probability that a new node is attached to i is:

$$p_i = \frac{k_i}{\sum_j k_j} \tag{10}$$

being k_i the degree of the node i and j the rest of existing nodes. Since we are randomly attaching new nodes, we cannot predict the exact number of edges of the final random network, but after trying several parameters in the NetworkX function that generates BA networks I could find the ones that generated random graphs with approximately the wanted values of n and m.

In the final part of the results section it is simulated a random walk on the protein interaction network and the results are compared to the ones predicted theoretically. In matrix form, the probability that the walkers are in each of the i nodes at time t is

$$\mathbf{p}(t) = \mathbf{A}\mathbf{D}^{-1}\mathbf{p}(t-1) \tag{11}$$

Being A the adjacency matrix and D the diagonal degree matrix. Each walker chooses preferably an edge with probability $1/k_i$ and the final steady state probabilities that we deduced in class are simply:

$$p_i = \frac{k_i}{2m} \tag{12}$$

That is to say, the probability p_i of having a walker at site i is proportional to its degree. Numerically I performed 10 random walks of length $L = 10^4$ per each site (the number of walkers is hence $10n$) and I counted the number of times each node was visited (let's call it N_i). In this way we can numerically compute the probability p_i as the ratio of the number of visits N_i to the total length of all walks:

$$p_i = \frac{N_i}{N_{\text{walkers}} L} \tag{13}$$

The length of the walks and the number of walkers were high enough to reach the steady state and have a perfect agreement with the theoretical result of Eq. 12.

3 Results

3.1 *S. Cerevisiae* network

Table 1: Main properties of the *S. Cerevisiae* protein interaction network.

Density	Diameter	Average clustering coefficient	Newman transitivity	Average path length	Degree assortativity	Bipartivity
0.00285	11	0.138	0.102	4.376	-0.0977	0.500

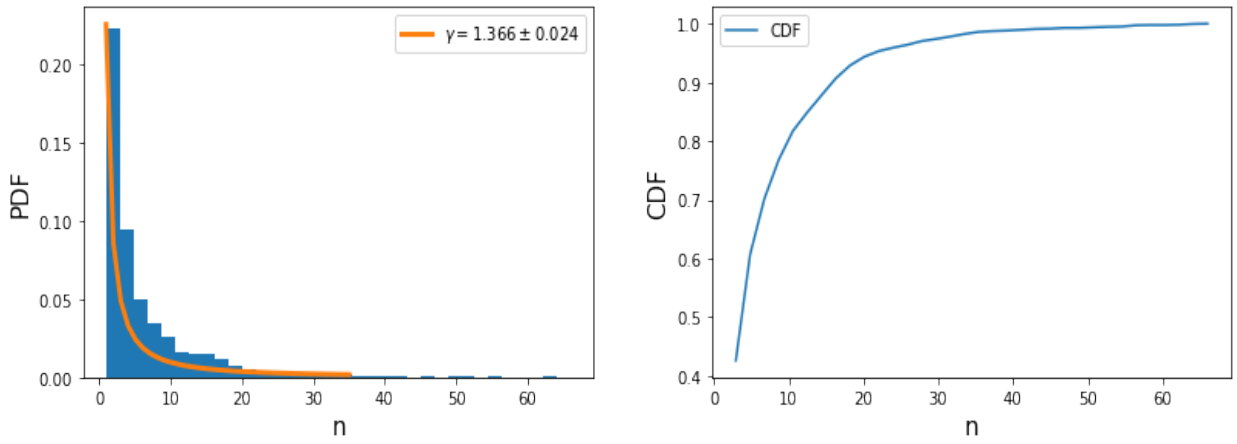

Figure 1: PDF and CDF for the degree distribution. The PDF follows a power law $p(k) \propto k^{-\gamma}$ with exponent $\gamma = 1.34$, which proves that we are dealing with a scale-free network.

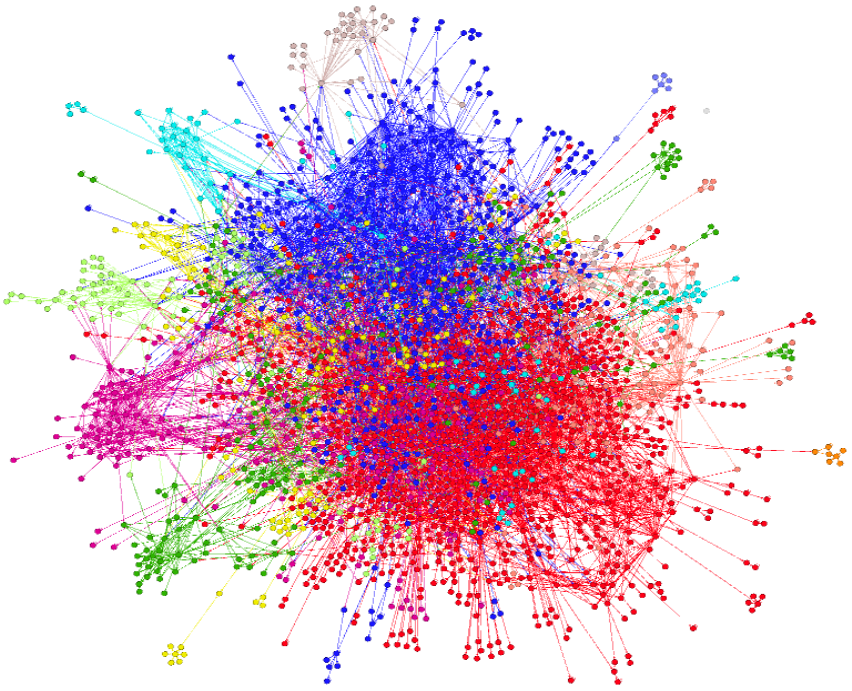

Figure 2: The 10 main communities in the protein-protein network. The representation was done with Gephi because it is more efficient than Python when drawing big networks.

The modularity of the network is high, $Q = 0.589$, and this agrees with the presence of communities shown in Figure 2. With regards to the different centralities, Table 4 in the Appendices contains the listing of the central nodes. Although included in the Appendices due to its size, Table 4 may have the most biologically relevant information of this work and the reason is to be addressed in the next section.

3.2 ER and BA random networks

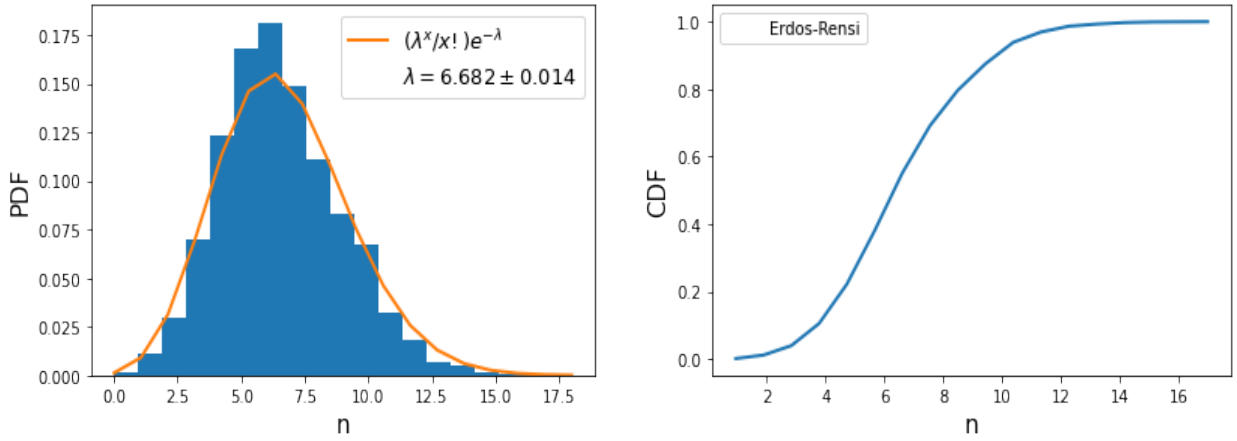

Figure 3: Degree distribution PDF and CDF in a Erdös-Renyi network realization. The PDF follows a Poisson distribution $p(k) = \frac{\lambda^k}{k!}e^{-\lambda}$ with exponent $\gamma = 1.34$. The number of nodes is high enough so that the typical binomial PDF in Erdös-Renyi networks becomes a Poissonian.

Table 2: Mean and standard deviation of the main network properties of 10 ER realizations.

Diameter	Average clustering coefficient	Newman transitivity	Average path length	Degree assortativity	Bipartivity
8.60 ± 0.49	0.00289 ± 0.00051	0.00294 ± 0.00034	4.381 ± 0.024	−0.001 ± 0.010	0.9684 ± 0.0021

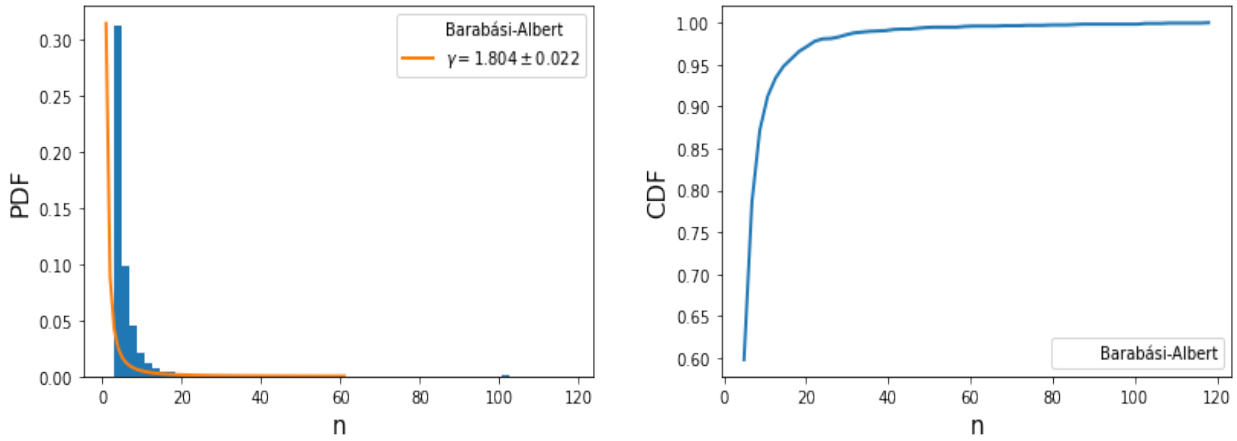

Figure 4: Degree distribution PDF and CDF in a Barabási-Albert network realization. The PDF follows a power law $p(k) \propto k^{-\gamma}$ with exponent $\gamma = 1.8$.

Table 3: Mean and standard deviation of the main network properties of 10 BA realizations.

Diameter	Average clustering coefficient	Newman transitivity	Average path length	Degree assortativity	Bipartivity
6	0.0178	0.00960	3.791	−0.0572	0.5307
±	±	±	±	±	±
0	0.0024	0.00076	0.016	0.0076	0.0068

3.3 Dynamical Analysis: Random Walk

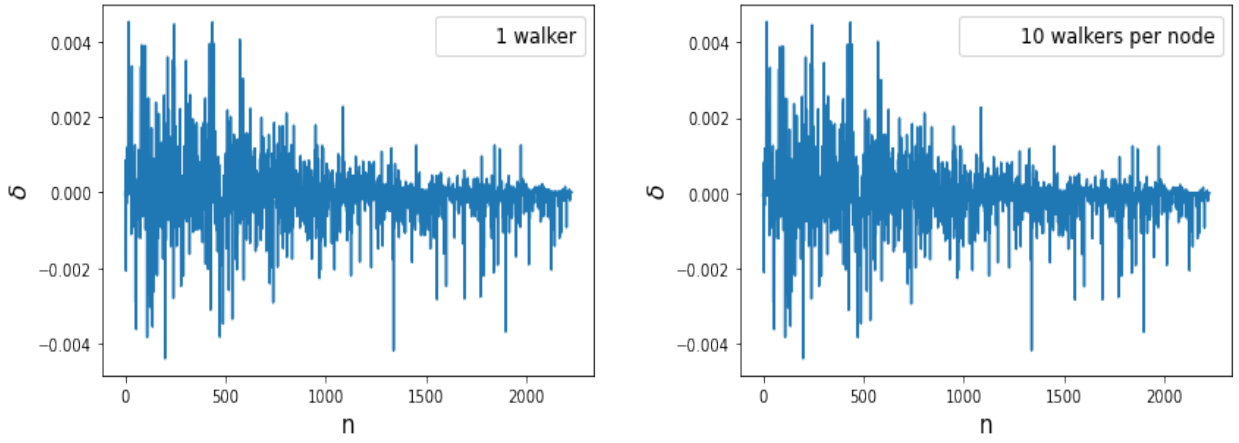

Figure 5: It is represented the difference δ between the theoretical probabilities and the numerical ones, $\delta = p_i^{\text{the}} - p_i^{\text{num}}$. The length of each walk is high, $L = 10^4$ and therefore there makes no statistical difference considering 1 or 10 walkers per node.

4 Discussion

The values of the bipartivity and the assortativity index depicted in Table 1 show that, as expected from a quick visualization of the network, there is no bipartition in the protein interaction network and no relevant assortativity (these properties are indeed more common in social networks). Furthermore, the average path lengths and the diameter are not high despite the size of the network, so the protein interaction network can be considered a small-world network as well. The fact that the degree probability distribution function can be approximated as a power law shows shows, again, that this PDF is the most common one in networks.

When comparing the properties of the ER networks (Table 2) with the ones from the *S.Cerevisiae* case in Table 1, it is worth of mention the high bipartivity and the low average clustering and transitivity found in the ER model. The BA networks seem to be much less bipartite like our protein-protein network, but their tendency to cluster and connect is also 10 orders of magnitude smaller (Table 3). This implies that the structure and topology of our protein network cannot be compared to these random graphs models regardless of their similar number of nodes and edges.

The centrality rankings in Table 4 (Appendices) are more interesting from a biological point of view, specially the degree centrality classification since high-degree nodes in a protein interaction network tend to correspond to proteins that are essential (Ref [6]). However, the main topological determinant of essentiality is not clearly defined: it is not widely accepted any kind of centrality as an infallible and universal determinant of the biological relevance of a protein. Although degree centrality does not unambiguously determine protein essentially for any network, it is more likely that these important proteins will be among the most central ones (Ref [8]), and this is helpful because finding the essential proteins out of thousands of them is time-consuming and expensive.

Betweenness centrality can provide more insight of the proteins essentiality as well and even more accurately than degree centrality [9]. In this work they found that proteins with even low connectivity but high betweenness are more likely to be essential. An accurate analysis of Table 4 reveals that both degree and betweenness centralities share most of their elements and, combining the hypothesis from [9] and [6], we can conclude that these shared proteins are very likely to be among the essentials.

5 Conclusions

The protein interaction network of the yeast *S.Cerevisiae* is a scale-free and a small-world network due to its power law degree distribution and the short average path length. Besides, the network is not bipartite and there can be found several communities. The biological role of these communities may or may not have a biological influence in the protein functions. The most reported and mentioned centralities in the literature for the yeast protein interaction network are the degree and betweenness centralities, and depending on the topology either both or one of them may be helpful for the classification of essential proteins and their functions. Betweenness centrality seems to be, however, more relevant for such purpose as it measures how often a node appears on pathways between two nodes.

Moreover, the random realizations of Erdös-Renyi and Barabási-Albert networks do not have the same topological properties as the protein interaction graph, but the most remarkable common feature is that both BA networks and the *S.Cerevisiae* network are scale-free. Finally, I simulated many random walks on the yeast network and the numerical probabilities of finding a walker on each node agree with the theoretical prediction: these probabilities are proportional to the node degree.

6 References

References

[1] Ernesto Estrada, Juan A. Rodriguez-Velazquez, "Subgraph centrality in complex networks", Physical Review E 71, 056103 (2005).

[2] M. E. J. Newman, Mixing patterns in networks, Physical Review E, 67 026126, 2003.

[3] E. Estrada and J. A. Rodríguez-Velázquez, "Spectral measures of bipartivity in complex networks", PhysRev E 72, 046105 (2005)

[4] von Mering et al. Comparative assessment of large-scale data sets of proteinprotein interactions. Nature 417 (2002) 399-403.

[5] Bu, D. et al. Topological structure analysis of the protein-protein interaction network in budding yeast. Nucleid Acids Res. 31 (2003), 2443-50

[6] Zotenko, E., Mestre, J., O'Leary, D. P., Przytycka, T. M. (2008). Why Do Hubs in the Yeast Protein Interaction Network Tend To Be Essential: Reexamining the Connection between the Network Topology and Essentiality. PLoS Computational Biology, 4(8), e1000140. doi:10.1371/journal.pcbi.1000140

[7] Park, K., Kim, D. (2009). Localized network centrality and essentiality in the yeastâprotein interaction network. PROTEOMICS, 9(22), 5143–5154. doi:10.1002/pmic.200900357

[8] Estrada, E. (2006). Virtual identification of essential proteins within the protein interaction network of yeast. PROTEOMICS, 6(1), 35–40. doi:10.1002/pmic.200500209

[9] Joy, Maliackal Brock, Amy Ingber, Donald Huang, Sui. (2005). High-Betweenness Proteins in the Yeast Protein Interaction Network. Journal of biomedicine biotechnology. 2005. 96-103. 10.1155/JBB.2005.96.

[10] https://networkx.org/documentation/

7 Appendices

Table 4: Top 25 central nodes according to different types of centralities.

Degree	YPR110C,YGL137W,YIL035C,YMR106C,YPL204W, YNL189W,YLR074C,YDL213C,YCR057C,YMR049C, YBR055C,YMR059W, YGR090W,YBR009C,YNL061W, YPR016C,YKR081C,YDR394W,YBR217W,YHR135C, YDL029W,YER133W,YBR017C, YDL047W,YMR125C
Closeness	YNL061W, YMR049C, YKR081C, YPR016C, YPL043W, YIL035C, YDL213C, YGR103W, YNL110C, YHR066W, YLR074C, YHR052W, YDR060W, YMR290C, YGR090W, YOR272W, YPL093W, YER126C, YER006W, YDR496C, YBR142W, YCR057C, YNL132W, YNL230C, YOR206W
Betweenness	YNL189W, YMR106C, YPL204W, YPR110C, YGL137W, YBR009C, YIL035C, YBR055C, YMR059W, YLR074C, YOL139C, YDL213C, YMR049C, YER133W, YMR012W, YDL047W, YPL235W, YML064C, YDL029W, YGR090W, YOL133W, YBR126C, YBR017C, YHR135C, YCR057C
Eigenvector	YNL061W, YMR049C, YKR081C, YPR016C, YPL043W, YIL035C, YDL213C, YGR103W, YNL110C, YHR066W, YLR074C, YHR052W, YDR060W, YMR290C, YGR090W, YOR272W, YPL093W, YER126C, YER006W, YDR496C, YBR142W, YCR057C, YNL132W, YNL230C, YOR206W
Katz	YPR110C, YIL035C, YGL137W, YMR106C, YPL204W, YNL189W, YDL213C, YLR074C, YCR057C, YBR055C, YMR049C, YMR059W, YBR009C, YGR090W, YNL061W, YPR016C, YDR394W, YKR081C, YBR217W, YHR135C, YDL029W, YER133W, YBR017C, YMR125C, YML064C
PageRank	YMR106C,YGL137W,YPR110C,YPL204W,YBR009C, YIL035C,YNL189W,YMR059W,YBR055C,YCR057C, YLR074C,YDL213C,YHR135C,YMR049C,YBR017C, YDL047W,YGR090W,YBR217W,YBR251W,YML064C, YPR016C,YDL029W,YGR040W,YNL061W,YER133W
Subgraph	YNL061W,YMR049C,YKR081C,YPR016C,YPL043W, YIL035C,YDL213C,YGR103W,YNL110C,YHR066W, YLR074C,YHR052W,YDR060W,YMR290C,YGR090W, YOR272W,YPL093W,YER126C,YER006W,YDR496C, YCR057C,YBR142W,YNL132W,YNL230C,YOR206W