Correlation between Transcriptome and Interactome

A major goal of current bioinformatics research is the functional characterization of genes using multiple data sources resulting from high-throughput technology. For example, Ge et al. (2001, 2003) investigate what they call a 'correlation mapping' between the transcriptome, monitored by microarray data, and the interactome, measured by protein-protein interactions (PPIs), for the yeast *Saccharomyces cerevisiae*. By combining information from two large data sources, they conclude that the proteins encoded by genes with similar expression profiles interact more frequently than other proteins, and suggest that the combination of multiple data sources can improve the quality of hypotheses regarding the function of genes.

Ge et al. (2001, 2003) used data from publicly available sources for their study. One of the microarray expression data sets came from a cell-cycle experiment described by Cho et al. (1998) in which 2885 unique genes (2945 probes) were divided into 30 clusters using k-means clustering. A 'literature' data set of 1666 PPIs (315 heterodimers, 1351 homodimers) was constructed from information in YPD and MIPS. Using a binomial distribution to assess statistical significance, Ge et al. (2003) concluded that the protein-protein heterodimers tended to be intracluster interactions rather than intercluster interactions. The same conclusions were made using gene expression data from yeast cells experiencing meiosis (Primig et al., 2000) or various stress conditions (Jelinsky et al., 2000), and PPI data from yeast two-hybrid experiments (Uetz et al., 2000; Ito et al., 2000, 2001).

Rather than use a binomial distribution for statistical inference, the data is better modeled using a hypergeometric distribution. In the 'literature' PPI list, there were 315 heterodimer interactions, 42 of which were between intracluster pairs. Suppose all of the possible pairwise interactions are represented by balls in an urn. Since we have 2885 genes, there are $\binom{2885}{2} = 4160170$ balls in the urn. If all the balls that represent intracluster interactions are red, and the intercluster interaction balls are white, then for the cell cycle data set, there are 156205 red balls and 4003965 white balls. Suppose we select 315 balls at random from this urn. The probability of drawing 42 or more red balls, assuming all balls are drawn independently of each other, can be calculated using the hypergeometric distribution. Specifically,

$$P(\#\text{red balls} \geq 42) = \sum_{i=42}^{315} \frac{\binom{156205}{i}\binom{4003965}{315-i}}{\binom{4160170}{315}} = 1.797187 \times 10^{-12}.$$

We would conclude that it is highly unlikely to observe 42 or more red balls in a random draw of 315 balls. In order to use the binomial distribution, Ge et al. (2003) assumed the balls were drawn with replacement. Despite this discrepancy, the qualitative conclusion is the same.

Graph Representation of the Transcriptome and Interactome

The problem posed in Ge et al. (2001, 2003) can be formulated in terms of graphs. Figure 1 is a graph representation of the 'literature' PPI list. Each gene is represented by a node, and if two proteins are known to interact, an edge is drawn between the two nodes. The 298 nodes in Figure 1 are colored according to membership in the 30 k-means clusters, and the 315 edges correspond to the number of interacting heterodimer protein pairs. There are an additional 2587 nodes of degree zero that are not pictured in Figure 1; these nodes represent genes that were used in the clustering analysis, but were not present in the list of PPIs.

The k-means clustering results can also be represented by a graph in which each node is connected by an edge to any other node in the same cluster. For the cell cycle data, the graph would consist of 30 completely connected subgraphs with the number of nodes in each subgraph corresponding to the number of genes in each cluster. Figure 2 demonstrates the 30 cluster graphs for the cell cycle data, using only the genes that are contained in the list of interacting protein pairs. Note that all connected nodes are the same color, corresponding to cluster membership. Figure 3 gives a closer view of the cluster graphs for clusters 1, 9, and 16.

The number of intracluster edges can be found by intersecting the graphs in Figures 1 and 2, resulting in the graph in Figure 4. The intersection graph keeps all of the same nodes as in the PPI and cluster graphs, but only retains the edges that exist in both graphs. Figure 4 demonstrates the 42 observed edges among 65 nodes that are reported in Ge et al. (2003). (All nodes of degree zero are excluded from the intersection graph).

Statistical Inference using Graphs

The graph representations of the expression cluster and PPI data easily accomodate the type of inference performed in Ge et al. (2001, 2003). Specifically, using the hypergeometric distribution to evaluate the statistical significance of observing 42 intracluster pairs, or 42 edges in the intersection graph, is equivalent to taking the 315 edges from the observed literature PPI graph, randomly reassigning the edges to different node pairs, and then counting the number of edges in the intersection graph of the randomized PPI graph and the cluster graph. Many repetitions of this permute-the-edges (PE) model results in a hypergeometric reference distribution for the number of edges in the intersection graph. Figure 5 shows one example of the random reallocation of the edges in the PPI graph. There were 8 edges in the intersection of this graph with the cluster graph, the arrangement of which are shown in Figure 6.

Notice that the graph in Figure 5 is composed of several components consisting of a very small number of nodes. Given the number of nodes and the number of edges that we are dealing with, this graph structure is consistent with the Erdös-Renyi theory of random graphs in which the degree distribution for the random edge graph is exponential. As suggested in several studies of the structure of PPI networks, however, PPI graphs tend to be scale-free with a degree distribution that follows a power law (Jeong et al., 2000, 2001; Snel et al., 2002; Strogatz, 2001). There is cause for concern in using a random edge model as a

basis for statistical inference since such models are likely not representative of actual PPI networks.

An alternative procedure for generating a reference distribution would be to retain the edge structure of the graph, and randomly permute the node labels (which we will call PN). This guarantees that the structure of the randomized PPI graph is from the sample space of possible PPI graphs, but still leads to a natural reference distribution since the node labels are not preferentially assigned.

Somewhat surprisingly, the PE and PN models result in similar distributions of the number of edges in the intersection graph (see Figure 8). It is possible, however that the number of edges in the intersection graph may not be the most descriptive test statistic. Figure 7 shows a graph with 42 edges and 84 vertices that would give the same p-value result as the observed intersection graph in Figure 4. These two graphs are quite different in terms of the arrangement of the edges. Figure 4 demonstrates a tendency toward connected components with greater numbers of nodes. Test statistics involving node degree and edge structure may give more insight into the structure of the intersection graph, and possibly the relevant biology.

For test statistics on the intersection graph other than the number of edges, Figure 8 shows that the PE and PN methods do give strikingly different distributions. Specifically, node degree ≥ 2, node degree ≥ 3, and the number of 3-cycles all have much different distributions depending on the algorithm. The PN method tends to give nodes with higher degree, and picks up more 3-cycles. Both of these features are evident in the observed intersection graph in Figure 4.

We also evaluated the number of connected components as a test statistic using the PE and PN algorithms. Although the distributions in Figure 8 are not strikingly different, there is some evidence that the PN algorithm tends to give fewer connected components in the intersection graph. This corresponds to finding larger groups of genes, possibly functional modules, that are connected in both the PPI and the cluster graph.

The graph representation of the cluster and PPI data illustrates that some clusters tend to have more intracluster interactions than others, and intercluster communication increases for certain pairs of clusters. Table 1 records the cluster size, the number of intracluster edges for each cluster, and the number of intercluster interactions for a cluster's top two interaction partners for 15 of the clusters. The remaining 15 clusters had no intracluster interactions; the sizes of these clusters and their intercluster activity are described in Table 2.

References

Cho, R., Campbell, M., Winzeler, E., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T., Gabrielian, A., Landsman, D., Lockhart, D. & Davis, R. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell* **2**, 65–73.

Ge, H., Liu, Z., Church, G. & Vidal, M. (2001). Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nature Genetics* **29**, 482–486.

Ge, H., Liu, Z., Church, G. & Vidal, M. (2003). Does mapping reveal correlation between gene expression and protein-protein interaction? -In reply. *Nature Genetics* **33**, 16–17.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., M., H. & Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences USA* **98**, 4569–4574.

Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto, K., Kuhara, S. & Sakaki, Y. (2000). Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between yeast proteins. *Proceedings of the National Academy of Sciences* **97**, 1143–1147.

Jelinsky, S., Estep, P., Church, G. & Samson, L. (2000). Regulatory networks revealed by transcriptional profiling of damaged *Saccharomyces cerevisiae* cells: Rpn4 links base excision repair with proteasomes. *Molecular and Cellular Biology* **20**, 8157–8167.

Jeong, H., Mason, S., Barabási, A.-L. & Oltvai, Z. (2001). Lethality and centrality in protein networks. *Nature* **411**, 41.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. & Barabási, A.-L. (2000). The large-scale organization of metabolic networks. *Nature* **407**, 651–654.

Primig, M., Williams, R., Winzeler, E., Tevzadze, G., Conway, A., Hwang, S., Davis, R. & Esposito, R. (2000). The core meiotic transcriptome in budding yeasts. *Nature Genetics* **26**, 415–423.

Snel, B., Bork, P. & Huynen, M. (2002). The identification of functional modules from the genomic association of genes. *Proceedings of the National Academy of Sciences* **99**(9), 5890–5895.

Strogatz, S. (2001). Exploring complex networks. *Nature* **410**, 268–276.

Uetz, P., Giot, L., Cagney, G., Mansfield, T. et al. (2000). A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627.

Table 1: Intracluster and Intercluster Activity

Cluster #	# of Nodes	# of Intracluster Edges	Highest Interactivity Cluster(# of Edges)	Second Interactivity Cluster(# of Edges)
1	157	9	7,30(4)	6,18,24(3)
2	185	11	23(8)	14,16(6)
5	151	2	4,8,9,13,22(3)	1,2,7,12(2)
7	101	3	2,4(5)	12,16(4)
8	148	2	5(3)	25(2)
10	78	1	11,13(2)	6,12,15,17,28,29(1)
11	94	1	13(4)	2,6,7,9,10,12,14,17,29(2)
13	96	1	11,17(4)	3,5,15,22(3)
14	73	5	2(6)	11,22,30(2)
18	96	1	1(3)	3,7,12,13,14,15,24(1)
19	73	1	9,25(1)	-
22	83	1	2,5,13(3)	14,16(2)
23	63	1	2(8)	12,28(2)
26	49	1	13,16(1)	-
27	63	2	2,11,12,21,22(1)	-

Table 2: Intercluster Activity for Clusters without Intracluster Interactions

Cluster #	# of Nodes	Highest Interactivity Cluster(# of Edges)	Second Interactivity Cluster(# of Edges)
3	103	13(3)	6,7,28(2)
4	169	7(5)	2,5(3)
6	100	1,24(3)	3,11,17(2)
9	146	5(3)	2,4,11,25(2)
12	79	2,7(4)	4,5,11,13,16,23(2)
15	113	13(3)	7,10,11,12,16,17,18,22(1)
16	98	2(6)	7(4)
17	81	13(4)	16(3)
20	84	6,17(1)	-
21	68	1(2)	3,11,12,13,16,27,28,29(1)
24	83	1,6(3)	29(2)
25	74	8,9(2)	4,7,19,24(1)
28	67	3,23(2)	1,10,13,21(1)
29	51	11,13,24(2)	1,6,8,10,12,17,21,22(1)
30	59	1,2(4)	14(2)

Figure 1: Observed Literature Protein-Protein Interaction Graph

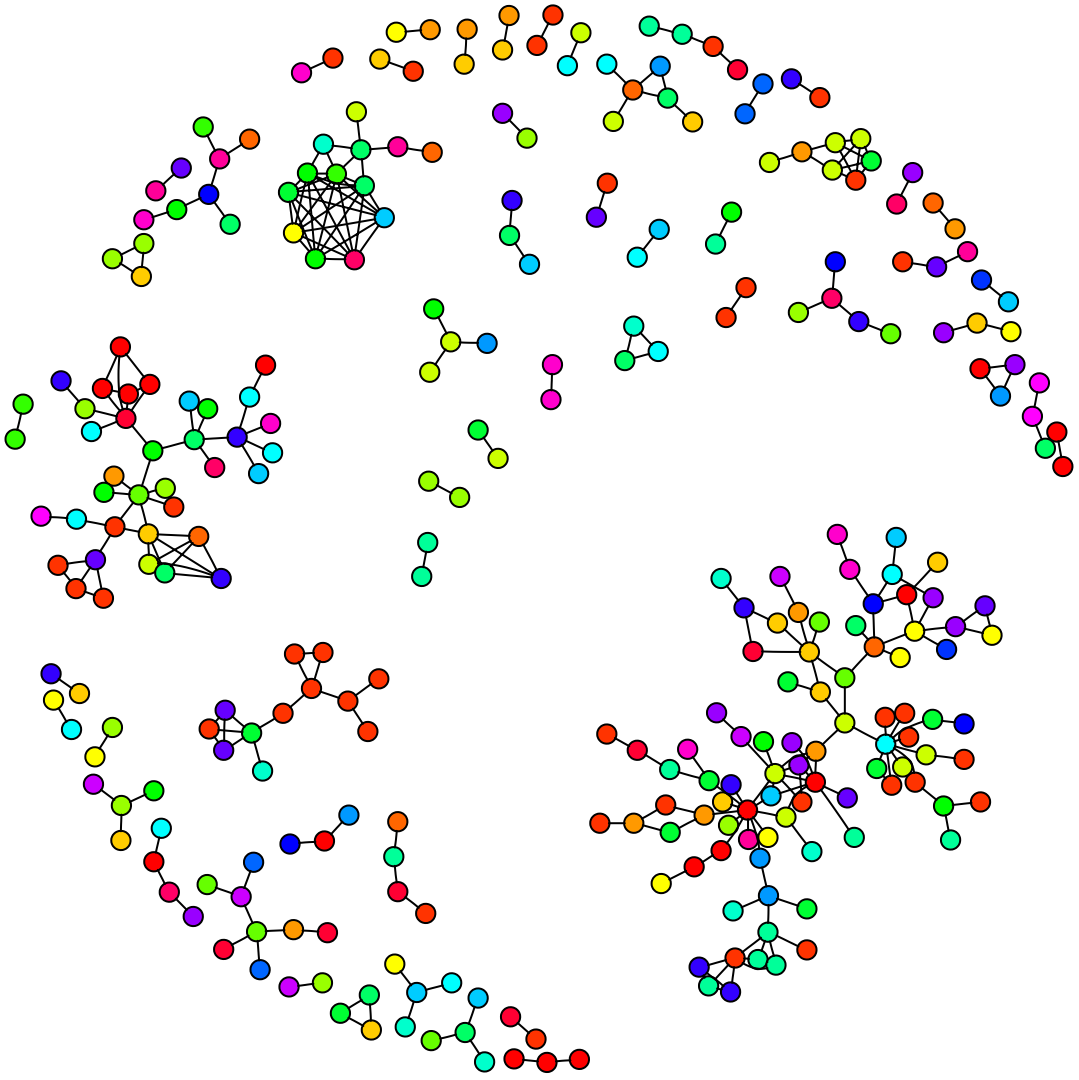

Figure 2: 30 Completely Connected Cell-Cycle Cluster Graphs

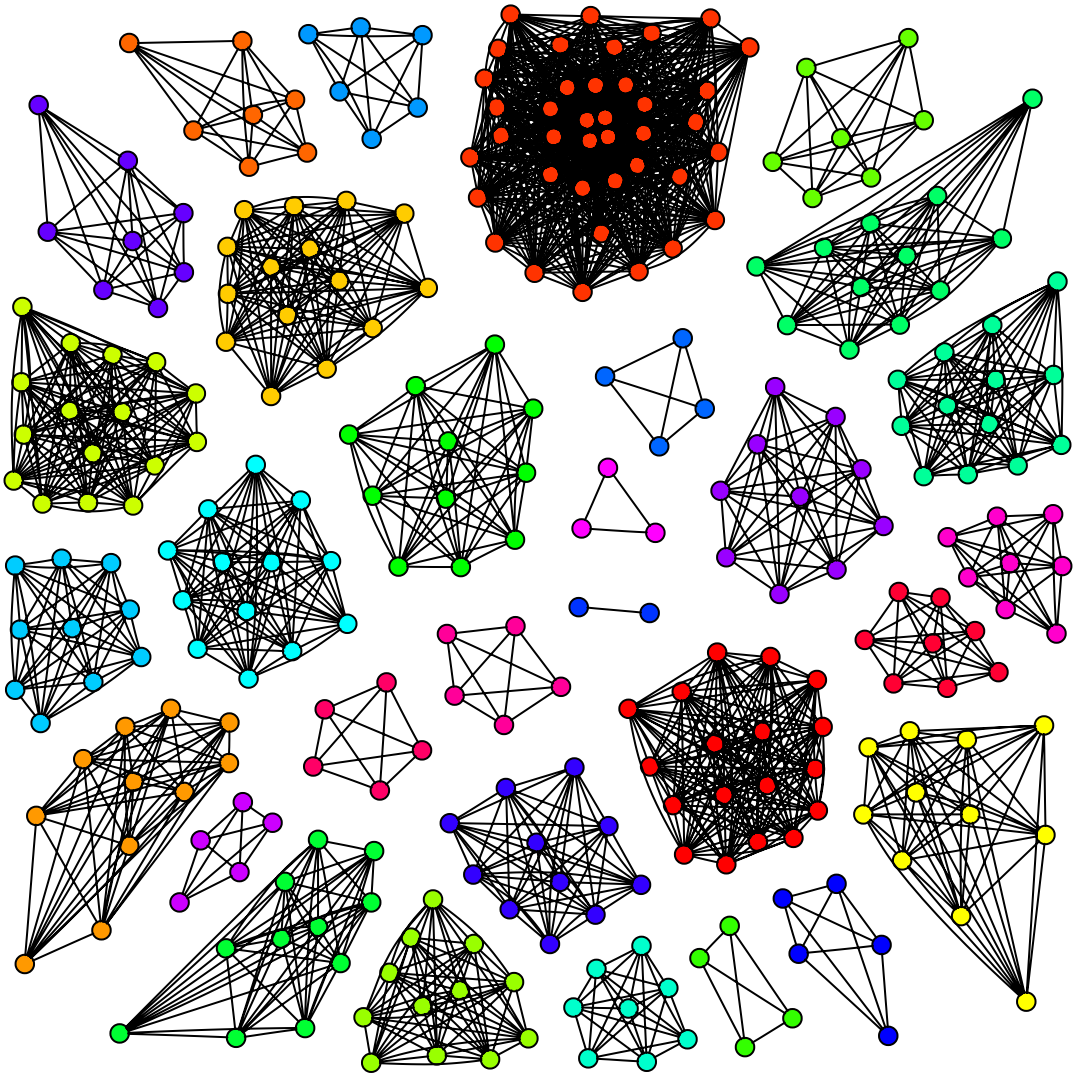

Figure 3: Graphs for Cell-Cycle Clusters 1, 9, and 16

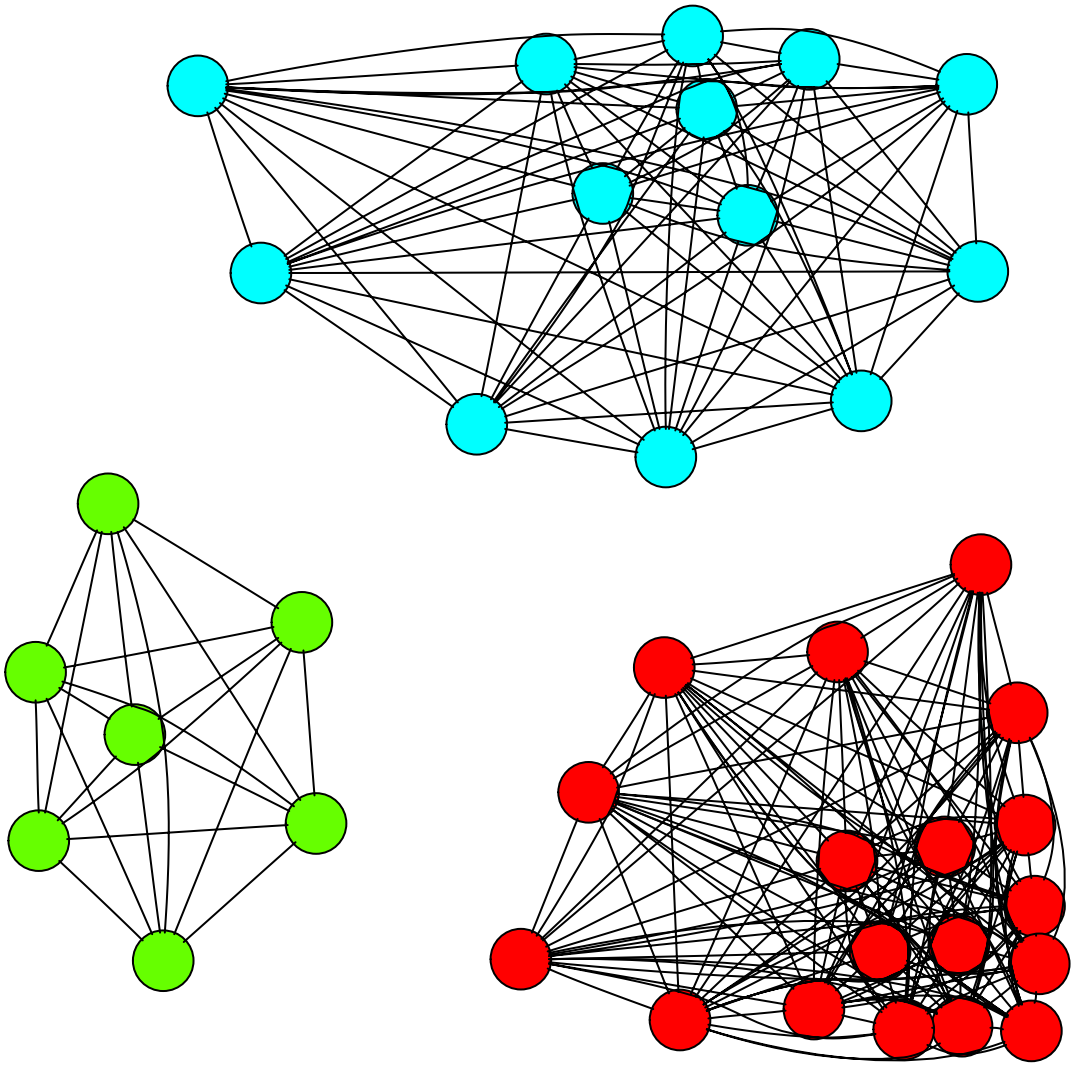

Figure 4: Observed Intracluster Edges in Ge, et al. (2003)

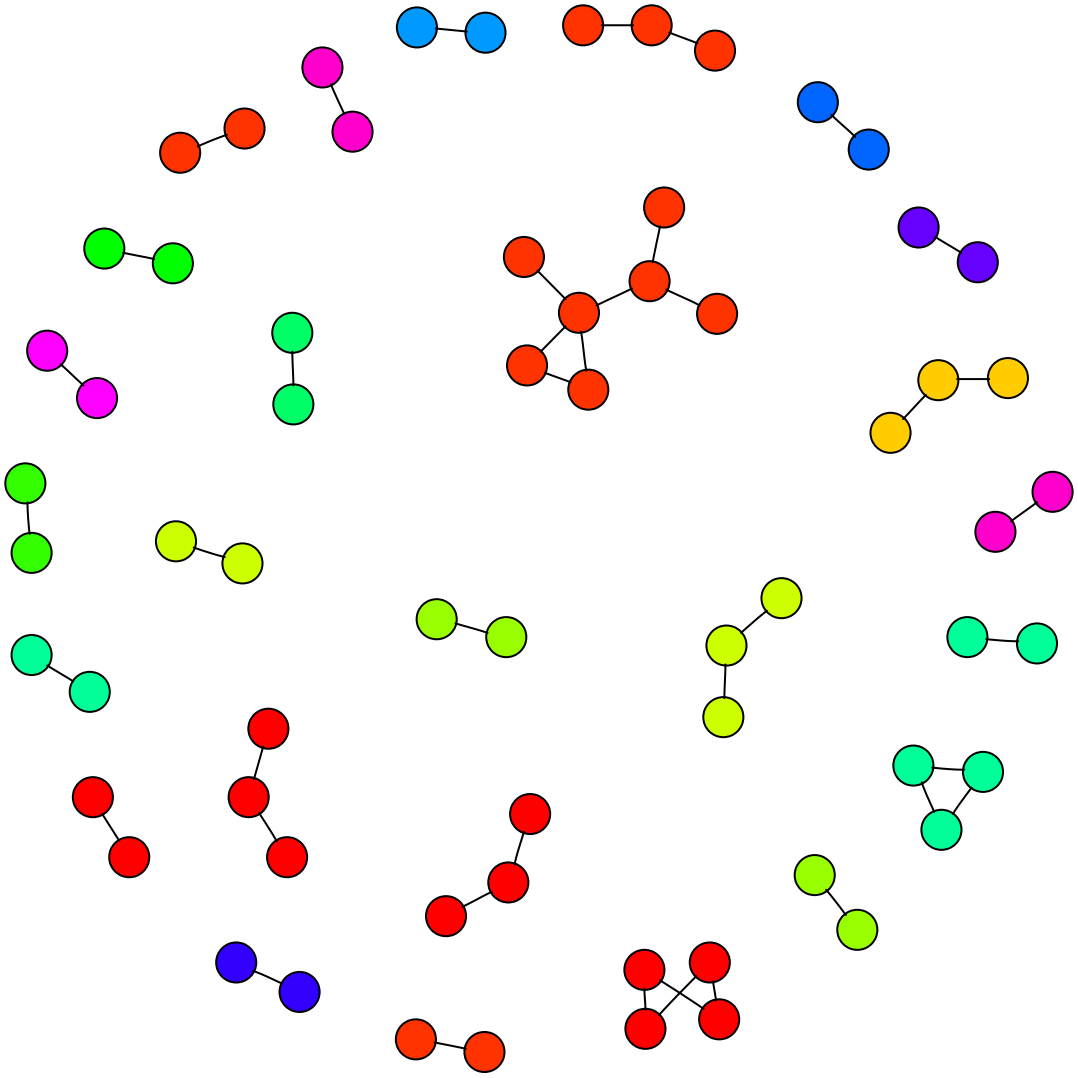

Figure 5: Randomly Reassigned Edges in PPI Graph

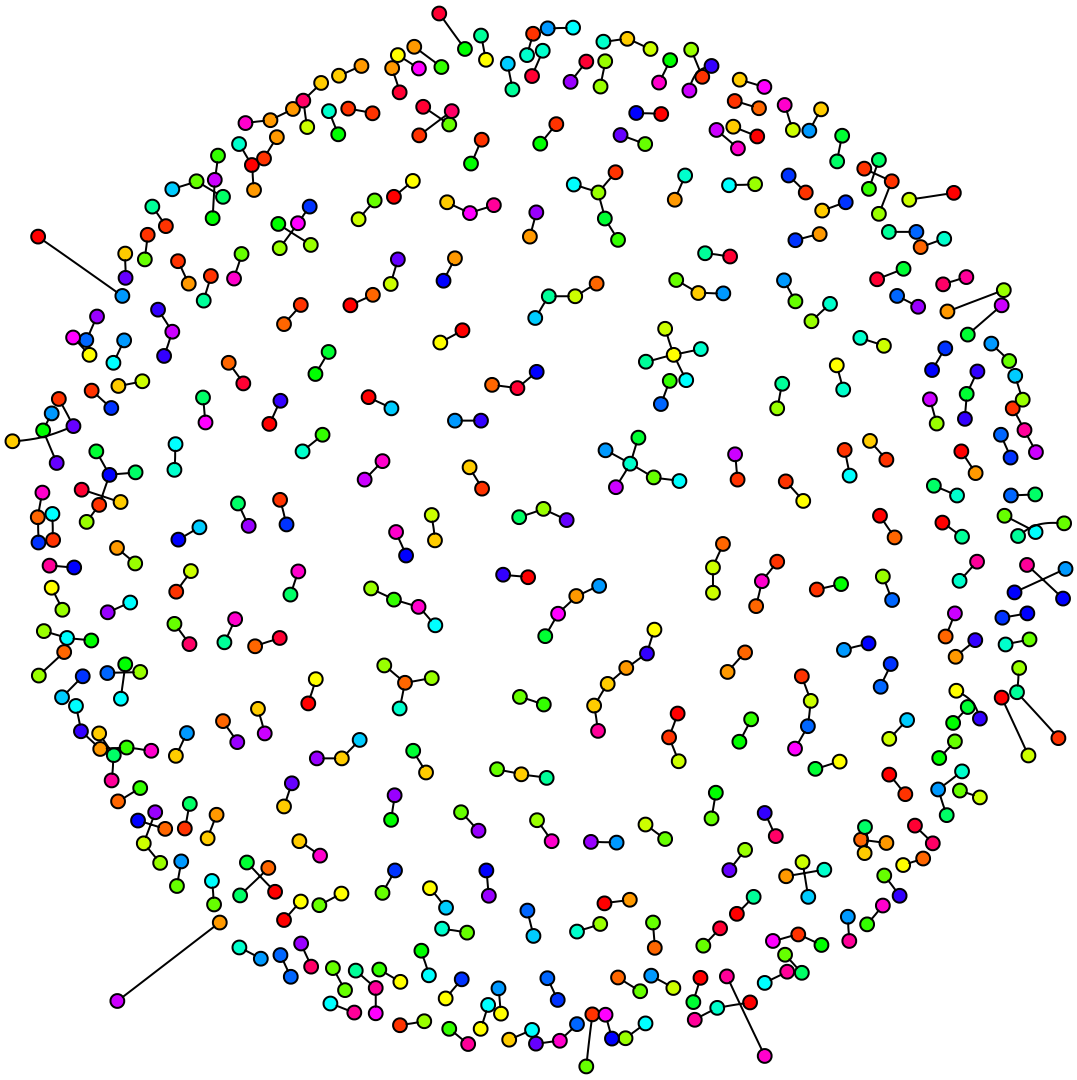

Figure 6: Intersection Graph after Random Edge Reassignment

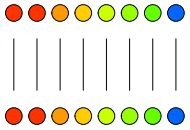

Figure 7: Graph with 42 Edges and 84 Nodes

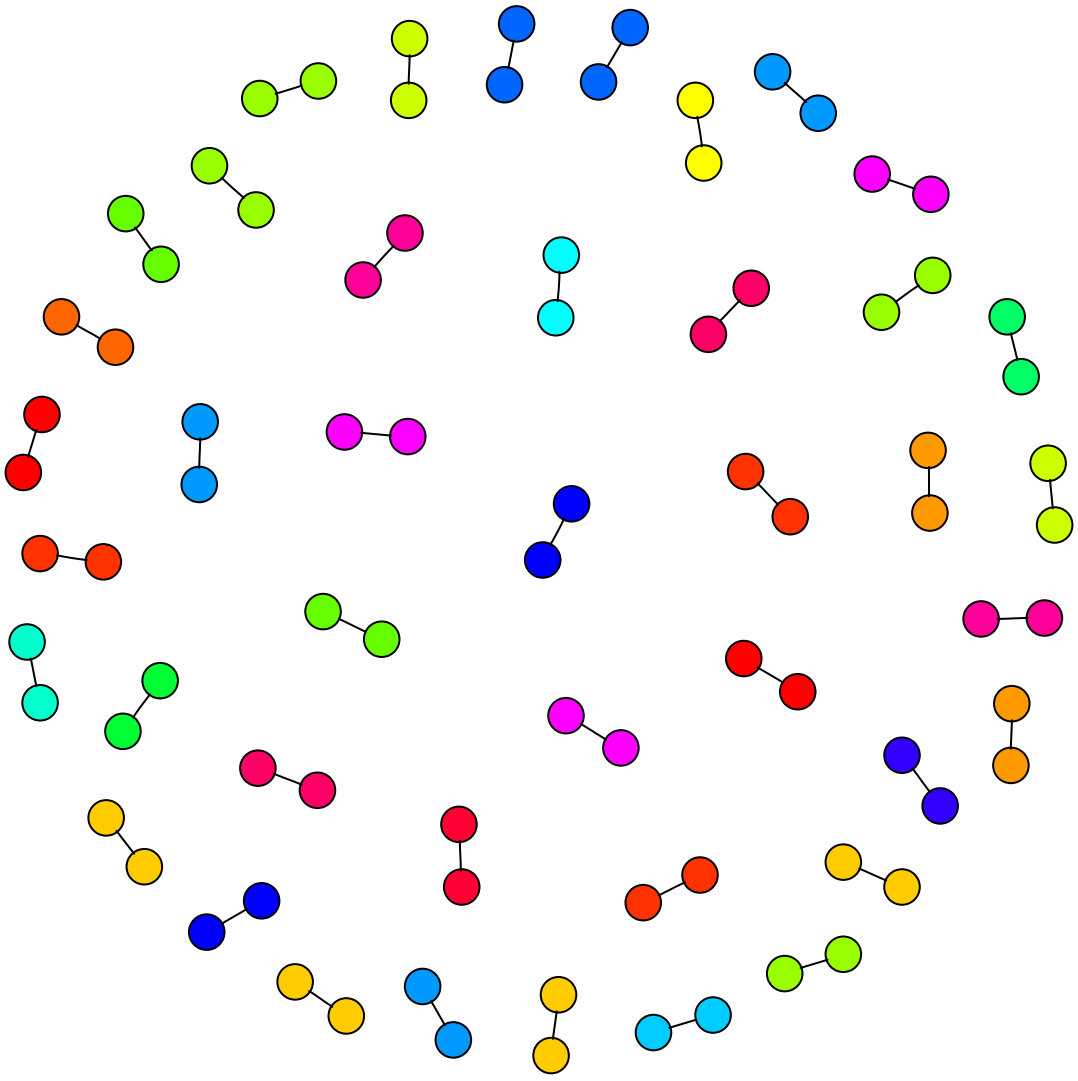

Figure 8: Test Statistics