Supplementary materials for "Identifying network communities with a high resolution"

Jianhua Ruan and Weixiong Zhang

October 4, 2007

Proof of Equation (7) in the main manuscript

Consider moving a vertex v in community c_i to a new community c_j. Rewrite the definition of Q as

$$Q = \sum_{i=1}^{k} Q_i,$$

where $Q_i = \frac{e_{ii}}{M} - (\frac{a_i}{M})^2$ is the contribution made by community c_i. It can be seen easily that after moving vertex v from c_i to c_j, the only terms that will be changed are Q_i and Q_j. Therefore, the difference between the new modualarity Q' and the original modularity can be computed by:

$$Q' - Q = Q'_i + Q'_j - Q_i - Q_j.$$

After the move, the total degrees within community i and j can be given by:

$$\begin{aligned} e'_{ii} &= e_{ii} - 2d_i^v, \text{ and} \\ e'_{jj} &= e_{jj} + 2d_j^v, \end{aligned}$$

where d_i^v is the number of connections that v has in community i. Similarly, the total degrees for the vertices in community i and j can be given by:

$$\begin{aligned} a'_i &= a_i - d^v, \text{ and} \\ a'_j &= a_j + d^v, \end{aligned}$$

where d^v is the degree of v.

Therefore,

$$\begin{aligned} Q'_i + Q'_j &= \frac{e'_{ii} + e'_{jj}}{M} - \frac{(a'_i)^2 + (a'_j)^2}{M^2} \\ &= \frac{e_{ii} - 2d_i^v + e_{jj} + 2d_j^v}{M} - \frac{(a_i - d^v)^2 + (a_j + d^v)^2}{M^2} \\ &= \frac{e_{ii} + e_{jj}}{M} + \frac{2d_j^v - 2d_i^v}{M} - \frac{a_i^2 + a_j^2}{M^2} - \frac{2(d^v)^2 - 2a_j d^v - 2a_i d^v}{M^2} \\ &= Q_i + Q_j + \frac{2}{M}(d_j^v - d_i^v) + \frac{2d^v}{M^2}(a_i - a_j - d^v). \end{aligned}$$

Hence,

$$\Delta Q^{migr}(v,i,j) = Q' - Q = \frac{2}{M}(d_j^v - d_i^v) + \frac{2d^v}{M^2}(a_i + a_j - d^v). \quad (1)$$

Proof of a theorem

Theorem: during the execution of *Qcut*, the algorithm will never move a vertex from a community where it has some connections to a community where it has no connection at all.

Proof: We prove this by contradiction. Suppose that the network has already been partitioned into K communities. Now assume *Qcut* chooses to move a node v, which is currently in community i, to a new community j, where v has no connection, i.e., $d_j^v = 0$. According to Equation (1) above, the change to Q after the move can be calculated as

$$\Delta Q^{migr}(v,i,j) = \frac{-2d_i^v}{M} + \frac{2d^v(a_i - a_j - d^v)}{M^2}, \quad (2)$$

where d_i^v is the number of connections that v has in communities i, d^v is the degree of vertex v, M is the total number of edges in the network, and a_i is the total degree for the vertices in community i.

Suppose that we had moved v to some other community, $k \neq i$, instead of j. Since *Qcut* actually chose j, we must have

$$\Delta Q^{migr}(v,i,k) \leq \Delta Q^{migr}(v,i,j) \text{ for all possible } k, \quad (3)$$

and

$$\Delta Q^{migr}(v,i,j) > 0. \quad (4)$$

Combining Equations (2) and (4), we have

$$\frac{2d^v(a_i - a_j - d^v)}{M^2} > \frac{2d_i^v}{M}.$$

Therefore,

$$d^v(a_i - a_j - d^v) > Md_i^v. \quad (5)$$

From Equations (2) and (3), we have

$$\frac{d_k^v - d_i^v}{M} + \frac{d^v(a_i - a_k - d^v)}{M^2} \leq \frac{-d_i^v}{M} + \frac{d^v(a_i - a_j - d^v)}{M^2}.$$

Therefore,

$$d^v(a_k - a_j) \geq Md_k^v. \quad (6)$$

Combining Equations (5) and (6) and summing over all $k \neq i$, we have

$$\begin{aligned} d^v(a_i - a_j - d^v) + \sum_{k\neq i} d^v(a_k - a_j) &> Md_i^v + \sum_{k\neq i} Md_k^v, \\ d^v(a_i - a_j - d^v) + \sum_{k\neq i} d^v(a_k - a_j) &> Md^v. \end{aligned}$$

Hence,

$$\begin{aligned} a_i - a_j - d^v + \sum_{k \neq i} a_k - \sum_{k \neq i} a_j &> M, \\ (a_i + \sum_{k \neq i} a_k) - (a_j + \sum_{k \neq i} a_j + d^v) &> M, \\ M - (K a_j + d^v) &> M, \\ K a_j + d^v &< 0, \end{aligned}$$

which is impossible, since by definition a network has no negative edges. Therefore, inequalities (3) and (4) cannot be both true. Hence moving v to a community j where $d_j^v = 0$ will not be the best move. It will either reduce Q, or there is a k such that $\Delta Q^{migr}(v, i, k) > \Delta Q^{migr}(v, i, j)$.

Robustness of *HQcut*

The *HQcut* algorithm uses two parameters, *minq* and *minz*. To test the sensitivity of the algorithm's results with respect to the two parameters, we varied the two parameters in a wide range of values, and computed the algorithm's accuracy (Jaccard Index) on a large number of synthetic networks with known community structures. The more than 1000 networks used in this test are the same as the networks used for plotting Fig 2(d) in the main manuscript. As shown in Fig. S1, the accuracy of the algorithm is largely invariant for $10 \geq minz \geq 2$ and is the best for $0.35 \geq minq \geq 0.3$. Therefore, for all experiments, we use $minq = 0.3$ and $minz = 2$ as the default parameters.

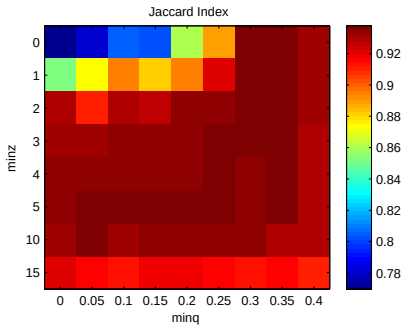

Fig. S1: Accuracy of *HQcut* as a function of *minq* and *minz*.

Other types of accuracy measurement

Besides the Jaccard Index [10], we also measured the accuracy of the algorithms using three other criterion, the Fowlkes-Mallows Index [3], Variation of Information [6], and the Wallace Index [11].

The Fowlkes-Mallows index is a variant of the Wallace Index. Let C_1 and C_2 be the true and predicted community structures, respectively. Let S_1 be the set of vertex pairs in the same community under C_1, and S_2 the set of vertex pairs in the same community under C_2. The Wallace Index is defined as

$$W(C_1, C_2) = \frac{|S_1 \cap S_2|}{|S_1|}, \tag{7}$$

which represents the probability that a pair of vertices which are in the same community under C_1 are also in the same community under C_2. It can be seen that the Wallace Index is asymmetric, i.e., $W(C_1, C_2) \neq W(C_2, C_1)$. The Fowlkes-Mallows Index is defined as the geometric mean of the two:

$$F(C_1, C_2) = \sqrt{W(C_1, C_2) W(C_2, C_1)}. \tag{8}$$

The second measurement, Variation of Information, is defined based on information theory, and is symmetric. It basically measures the amount of information that is lost or gained in changing from C_1 to C_2. For details see [6].

The value of the Fowlkes-Mallows Index is between 0 and 1, and a high value means better accuracy. The value of Variation of Information is always non-negative, and a zero means the best accuracy.

As shown in Figures S2 and S3, the relative performance of the four algorithms (*Newman*, *SA*, *Qcut*, and *HQcut*) based on the Fowlkes-Mallows Index and Variation of Information is similar to that in the main manuscript measured by the Jaccard Index.

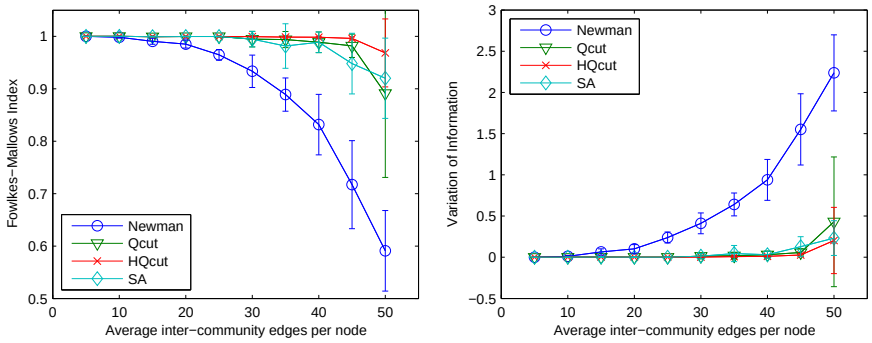

Fig. S2: Accuracy of the algorithms on synthetic networks with equal community sizes. Left: accuracy measured by the Fowlkes-Mallows Index. Right: accuracy measured by Variation of Information.

Figure S4 shows the performance of the algorithms using the Wallace Index. $W(C_1, C_2)$ measures the percentage of vertex pairs in the same community in C_1 that are also in the same community in C_2. On the other hand, $W(C_2, C_1)$ measures the percentage of vertex pairs in the same community in C_2 that are also in the same community in C_1. These two quantities can be considered as

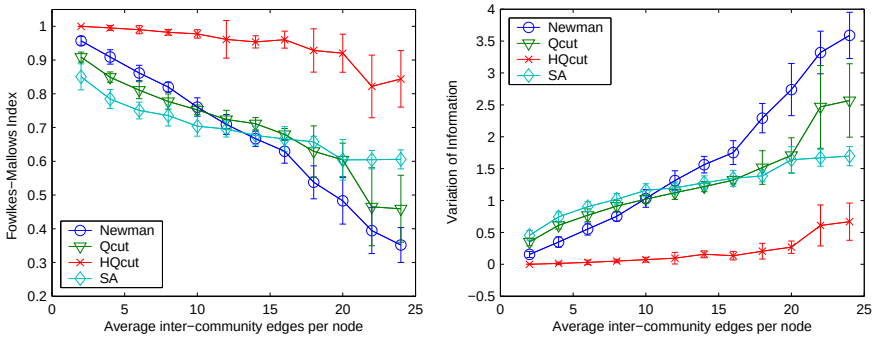

Fig. S3: Accuracy of the algorithms on networks with heterogeneous community sizes. Left: accuracy measured by Fowlkes-Mallows Index. Right: accuracy measured by Variation of Information.

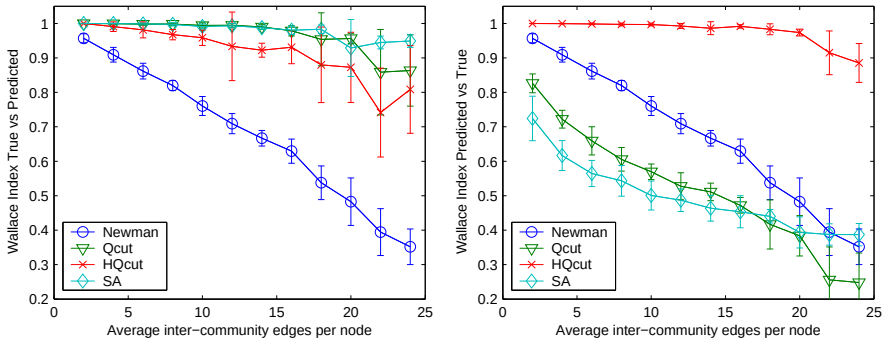

Fig. S4: Wallace Index of the algorithms on the networks with heterogeneous community sizes. Left and right are $W(C_1, C_2)$ and $W(C_2, C_1)$, respectively, where C_1 and C_2 are true and predicted community structures, respectively.

the measurements of recall and precision measurements, if we treat each intra-community vertex pair as an instance.

As shown in the main manuscript, *Qcut* and *SA* often achieve better modularities than *Newman*. Figure S4(a) shows that the former two algorithms also have much higher recall values, i.e., they recovered more intra-community vertex pairs than *Newman*. This is achieved with the price of a lower precision due to more total intra-community vertex pairs being predicted. Note that the low recall of *Newman* is not because it has over-partitioned the networks: the number of communities predicted by *Newman* is still much smaller than that of the true structures (Fig. 2(f) in the main manuscript). On the other hand, *HQcut* has slightly lower recalls than *Qcut* and *SA*, but much higher precisions than all other three algorithms, which means that it is able to successfully separate the merged communities without over-partitioning the true communities.

Information about the real-world networks

- Social: This is a social network of 67 prison inmates, based on their answers to questionnaires. The network was first studied in [5]. Data were downloaded from UCINET IV Datasets (`http://vlado.fmf.uni-lj.si/pub/networks/data/Ucinet/UciData.htm`).
- Neuron: This network represents the neural network of *C. Elegans*. The network was originally described in [13], and was studied by [12]. Data were downloaded from `http://www.weizmann.ac.il/mcb/UriAlon/`.
- Ecoli Reg: Transcriptional regulatory network of *E. coli*. Nodes are operons. An edge is set between A and B if A activates B or B activates A. The network was studied in [8]. Data were obtained from `http://www.weizmann.ac.il/mcb/UriAlon/`.
- Circuit: Electronic circuits. Nodes are electronic components (capacitors, diodes, etc.) and connections are wires. This network was downloaded from `http://www.weizmann.ac.il/mcb/UriAlon/`.
- Yeast Reg: Transcriptional regulatory network of yeast *Saccharomyces cerevisiae*. Similar to *E. coli* transcriptional regulatory network. This network was downloaded from `http://www.weizmann.ac.il/mcb/UriAlon/`.
- Ecoli Met: The largest connected component of the metabolic network of *E. coli*. In this network, nodes represent metabolites and two nodes i and j are connected by a link if there is a chemical reaction in which i is a substrate and j a product, or vice versa. The data was obtained from the KEGG database [4].
- Ecoli PPI: The largest connected component of a protein-protein interaction network of *E. coli*. Network data were obtained from the DIP database [9].
- Internet: The autonomous Systems topology of the Internet [2]. Network data were downloaded from `http://www.cosin.org`.
- Football: Network of United States NCAA division I-A college football teams. Each vertex is a team. An edge between two teams represents a regular-season game played by them in year 2006. The game schedule was obtained from `http://sports.espn.go.com/ncf/schedules`.

Example associated communities

Figures 4–8 show some associated communities, and the corresponding protein complexes in the MIPS database [7]. Nodes in different communities are drawn with different colors. The shape of a node represents the protein complex or

complexes of which it is a member. Annotations of genes and complexes are obtained from SGD (`http://www.yeastgenome.org/`) [1]. As shown, the communities that are statistically associated are not only highly connected physically, but functionally strongly related. They often correspond to different subunits of a large protein complex, or represent protein complexes that share a significant portion of their members and are involved in similar biological processes.

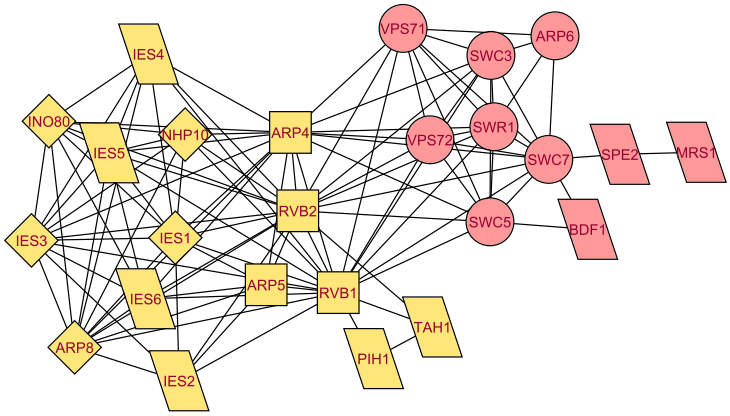

Fig. S5: A pair of associated communities. Diamond: components of the INO80 protein complex. Circle: components of the SWR1 protein complex. Rectangle: shared components of INO80 and SWR1. Parallelogram: proteins that are not components of the INO80 or SWR1 by current knowledge. IES2, IES4, IES5, and IES6 are known to be associated with INO80 under low-salt conditions. Both INO80 and SWR1 have functions in chromatin remodeling.

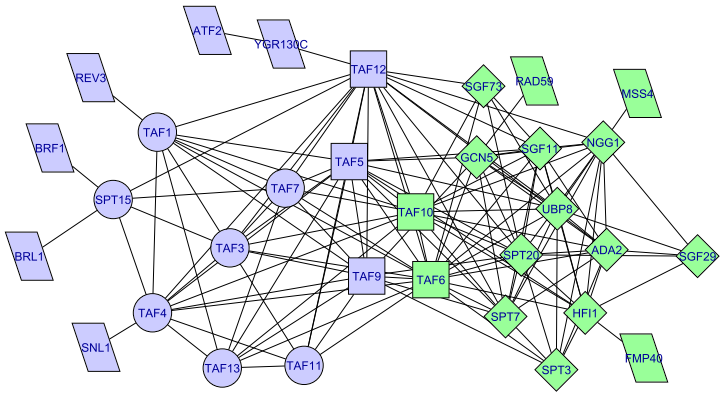

Fig. S6: A pair of associated communities. Diamond: components of the SAGA protein complex. Circle: components of the TFIID protein complex. Rectangle: shared components of SAGA and TFIID. Parallelogram: proteins that are not components of SAGA or TFIID by current knowledge. TFIID is involved in promoter binding and RNA polymerase II transcription initiation. SAGA is a transcription regulatory complex.

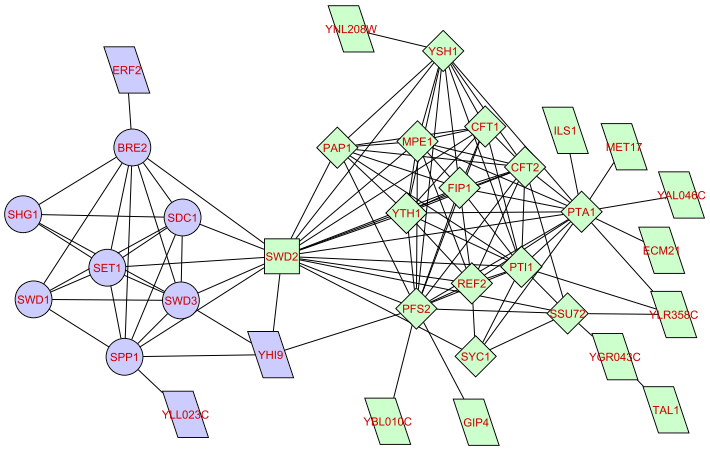

Fig. S7: A pair of associated communities. Diamond: components of the CPF protein complex. Circle: components of the Set1C protein complex. Rectangle: shared components of CPF and Set1C. Parallelogram: proteins that are not components of CPF or Set1C by current knowledge. CPF is involved in mRNA cleavage and polyadenylation. Set1C catalyzes methylation of histone H3, and mediates chromatin remodeling.

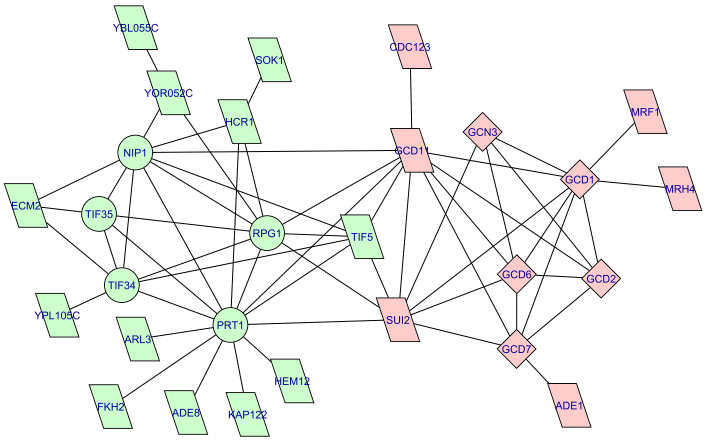

Fig. S8: A pair of associated communities. Diamond: components of the eIF2B protein complex. Circle: components of the eIF3 protein complex. Parallelogram: proteins that are not components of eIF2B or eIF3 by current knowledge. The eIF2B and eIF3 are the two largest complexes involved in cytoplasmic translation initiation. Several highly connected proteins in the network are also translation initiation factors. For example, GCD11 and SIU2 are components of eIF2 complexes; TIF5 encodes translation initiation factor eIF-5; HCR1 is a substoichiometric component of eIF3.

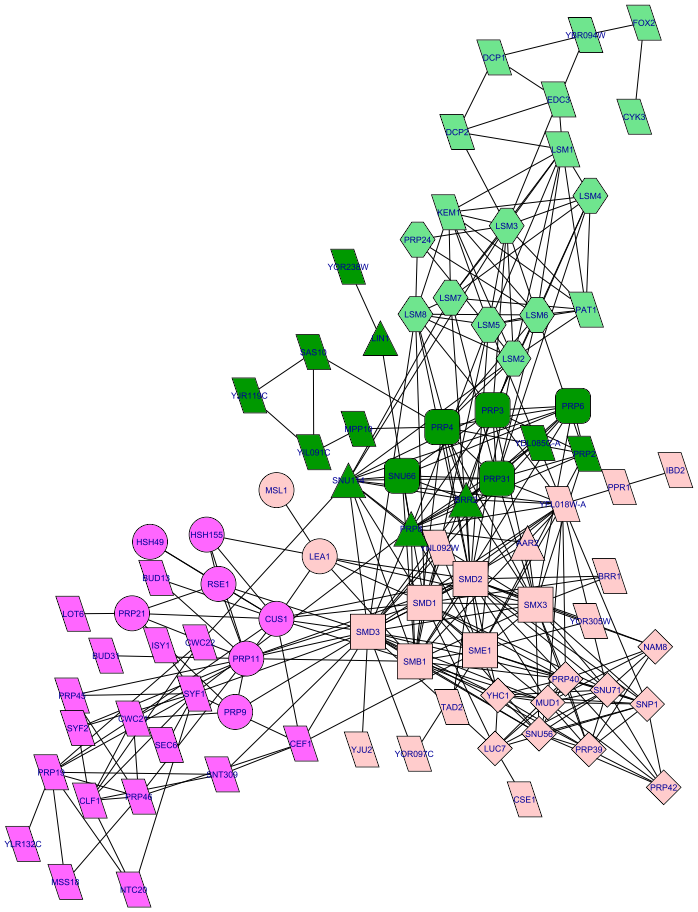

Fig. S9: A set of associated communities corresponding to mRNA spliceosome complexes. Diamond: components of the snRNP U1 protein complex. Circle: components of snRNP U2. Triangle: components of snRNP U5. Hexagon: components of snRNP U6. Rectangle: shared components of snRNP U1 and U5. Rounded rectangle: Components of U4/U6 x U5 tri-snRNP complex. Parallelogram: proteins that are not components of the above protein complexes by current knowledge. However, many of them are known to have functions in mRNA processing.

References

[1] S.S. Dwight, R. Balakrishnan, K.R. Christie, M.C. Costanzo, K. Dolinski, S.R. Engel, B. Feierbach, D.G. Fisk, J. Hirschman, E.L. Hong, L. Issel-Tarver, R.S. Nash, A. Sethuraman, B. Starr, C.L. Theesfeld, R. Andrada, G. Binkley, Q. Dong, C. Lane, M. Schroeder, S. Weng, D. Botstein, and J.M. Cherry. Saccharomyces genome database: underlying principles and organisation. *Brief Bioinform*, 5:9–22, 2004.

[2] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *SIGCOMM*, pages 251–262, 1999.

[3] E.B. Fowlkes and C.L. Mallows. A method for comparing two hierarchical clusterings. *J. Amer. Statist. Assoc.*, 78:553–569, 1983.

[4] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori. The KEGG resource for deciphering the genome. *Nucleic Acids Res*, 32:D277–280, 2004.

[5] J. MacRae. Direct factor analysis of sociometric data. *Sociometry*, 23:360–371, 1960.

[6] M. Meila. Comparing clusterings—an information based distance. *J. Multivar. Anal.*, 98:873–895, 2007.

[7] H.W. Mewes, D. Frishman, K.F. Mayer, M. Munsterkotter, O. Noubibou, P. Pagel, T. Rattei, M. Oesterheld, A. Ruepp, and V. Stumpflen. MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Res*, 34:D169–172, 2006.

[8] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298:824–827, 2002.

[9] L. Salwinski, C.S. Miller, A.J. Smith, F.K. Pettit, J.U. Bowie, and D. Eisenberg. The database of interacting proteins: 2004 update. *Nucleic Acids Res*, 32:D449–451, 2004.

[10] P.N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining.* Addison Wesley, 2005.

[11] D.L. Wallace. Comment. *J. Amer. Statist. Assoc.*, 78:569–576, 1983.

[12] D.J. Watts and S.H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

[13] J. G. White, E. Southgate, J. N. Thompson, and S. Brenner. The structure of the nervous system of the nematode caenorhabditis elegans. *Phil. Trans. R. Soc. London*, 314:1–340, 1986.