A Graph-Theoretic Analysis of the Human Protein-Interaction Network Using Multicore Parallel Algorithms *

David A. Bader and Kamesh Madduri

College of Computing
Georgia Institute of Technology, Atlanta, GA 30332
{bader, kamesh}@cc.gatech.edu

Abstract

Protein-interaction network (PIN) analysis provides valuable insight into an organism's functional organization and evolutionary behavior. In this paper, we study a PIN formed by high-confidence human protein interactions *obtained from various public interaction databases. This is the* largest human PIN *studied to date, comprising nearly 18,000 proteins and 44,000 interactions. A novel contribution of this paper is the computation of betweenness centrality, a graph-theoretic metric that is found to be positively correlated with the essentiality and evolutionary age of a protein. We observe that proteins with high betweenness centrality, but low connectivity are abundant in the human PIN. We have designed an efficient and portable parallel implementation for the calculation of this compute-intensive centrality metric. On the Sun Fire T2000 server with the UltraSparc T1 (Niagara) processor, we achieve a relative speedup of about 16 using 32 threads for a typical instance of betweenness centrality, reducing the running time from several minutes to 13 seconds.*

1 Introduction

Protein interactions play an important role in understanding the functional and organizational principles of biological processes. One of the key goals of functional genomics is to identify the complete protein interaction network of an organism, termed the *interactome*. In recent years, high-throughput experiments have been performed to determine the interactomes of model eukaryotes such as yeast [31, 30], worm [18] and fly [10]. These protein-interaction datasets, mainly derived from the yeast two-hybrid (Y2H) assay, provide evidence that global topological structure and networks features relate to known biological properties [14]. This has motivated several research groups to work on a global map of the human interaction network, in the hope that the interactome would provide insight into development and disease mechanisms at a systems level. There have been several recent efforts on mapping the global human genome [25, 29] using the Y2H assay. However, this system is prone to a high rate of false-positives and the interactions need to be validated with sophisticated techniques. Also, the identity of essential interactions in PINs differ significantly, depending on the experimental methodology [3]. In addition to these global maps, there are a large number of published interactions on individual disease proteins in the last decade. The high-confidence interactions are readily available from online public domain databases (for example, BIND [1], DIP [26] and HPRD [21]). Most of these databases are literature-based and hand-curated with a sizable percentage of overlapping interactions.

The interaction networks of model eukaryotes such as yeast are analyzed extensively [32, 16] using graph-theoretic and complex network analysis concepts. The yeast PIN topology exhibits several interesting features that distinguish it from a random graph. For instance, the distribution of the number of interactions of a protein can be approximated by a power law, and so the PIN may be a scale-free network. The PIN also contains a larger number of highly connected proteins than one would expect in a random Erdős-Rényi network. It is also observed that in the yeast network, the connectivity of a protein appears to be positively correlated with its essentiality [14], i.e., highly connected proteins tend to be more essential to the viability of the organism.

Large-scale network analysis is currently an active area of research in the social sciences [23, 27], and several concepts from this field are being applied to computational biol-

*This work was supported in part by NSF Grants CNS-0614915, CAREER CCF-0611589, DBI-0420513 and ITR EF/BIO 03-31654.

ogy. Important contributions from this field include analytical tools for visualizing networks [4, 24], empirical quantitative indices to determine the key nodes in a network, and clustering algorithms [11]. Betweenness Centrality [8] is a popular quantitative index that has been extensively used in recent years for the analysis of large-scale complex networks. Some applications include biological networks [14, 22], study of sexual networks and AIDS [19], identifying key actors in terrorist networks [7], organizational behavior and transportation networks [12]. Joy et al. [16] report that in the yeast network, proteins with high betweenness are more likely to be essential, and that the evolutionary age of proteins is positively correlated with betweenness. Also, they observe that there are several proteins with low degree but high centrality scores in the yeast PIN.

Gandhi et al. present the first analysis of the human interactome [9]. They study a dataset of about 26,000 human protein interactions obtained from various public databases, compare the human interactome with the yeast, worm and fly datasets, and observe that only 42 interactions were common to all species. Also, they observe that the available data does not support the presumption on the positive correlation between connectivity and essentiality.

We extend the work of Gandhi et al. [9] and Joy et al. [16] in this paper. Our main contributions are the following:

- *Topological study of the largest human PIN constructed to date, comprising nearly 18,000 proteins and 44,000 interactions.* We analyze the global connectivity and clustering properties of a human PIN composed of high-confidence protein interactions.
- *Computation of centrality metrics for the human PIN.* We analyze betweenness centrality scores and find that proteins with high betweenness centrality but low connectivity are abundant in the human PIN. We also observe that this finding cannot be explained by the widely-accepted models for scale-free networks.
- *Applying high performance computing techniques for large-scale PIN analysis.* Our efficient multicore implementation reduces the computation time of betweenness centrality to 13 seconds on 32 processors of the Sun Fire T2000 system, with a relative speedup of 16.

2 Preliminaries

We represent the PIN as an undirected graph $G(V, E)$ in the analysis that follows. The set V represents the proteins, and E the set of interactions. The number of vertices and edges are denoted by n and m, respectively. Since the interaction networks are unweighted, we assume that each edge $e \in E$ has unit weight. A *path* from protein (vertex) s to t is a sequence of interactions (edges) $\langle u_i, u_{i+1} \rangle$, $0 \leq i \leq l$, where $u_0 = s$ and $u_l = t$. The *length* of a path is the sum of the weights of edges. We use $d(s, t)$ to denote the distance between vertices s and t (the minimum length of any path connecting s and t in G). Let us denote the total number of shortest paths between vertices s and t by σ_{st}, and the number passing through vertex v by $\sigma_{st}(v)$.

Betweenness Centrality is a global shortest paths enumeration-based metric, introduced by Freeman in [8]. Let $\delta_{st}(v)$ denote the *pairwise dependency*, or the fraction of shortest paths between s and t that pass through v: $\delta_{st}(v) = \frac{\sigma_{st}(v)}{\sigma_{st}}$. Betweenness Centrality of a vertex v is defined as

$$BC(v) = \sum_{s \neq v \neq t \in V} \delta_{st}(v)$$

This metric measures the control a vertex has over communication in the network, and can be used to identify key vertices in the network. High centrality indices indicate that a vertex can reach other vertices on relatively short paths, or that a vertex lies on a considerable fraction of shortest paths connecting pairs of other vertices.

A straight-forward way of computing Betweenness Centrality is to augment a single-source shortest path algorithm such as Dijkstra's algorithm to compute the pairwise dependencies. Define a set of *predecessors* of a vertex v on shortest paths from s as $pred(s, v)$. Now each time an edge $\langle u, v \rangle$ is scanned for which $d(s, v) = d(s, u) + d(u, v)$, that vertex is added to the predecessor set $pred(s, v)$. Setting the initial condition of $pred(s, v) = s$ for all neighbors v of s, we can proceed to compute the number of shortest paths between s and all other vertices. The computation of $pred(s, v)$ can be easily integrated into breadth-first search (BFS) for unweighted graphs.

To exploit the sparse nature of typical real-world graphs, Brandes [5] gives an algorithm that computes the betweenness centrality score for all vertices in the graph in $O(mn)$ time for unweighted graphs. The main idea is as follows. We define the *dependency* of a source vertex $s \in V$ on a vertex $v \in V$ as $\delta_s(v) = \sum_{t \in V} \delta_{st}(v)$. The betweenness centrality of a vertex v can be then expressed as $BC(v) = \sum_{s \neq v \in V} \delta_s(v)$. It can be shown that the dependency $\delta_s(v)$ satisfies the following recursive relation: $\delta_s(v) = \sum_{w : v \in pred(s,w)} \frac{\sigma_{sv}}{\sigma_{sw}} (1 + \delta_s(w))$.

The algorithm is now stated as follows. First, n BFS computations are done, one for each $s \in V$. The predecessor sets $pred(s, v)$ are maintained during these computations. Next, for every $s \in V$, using the information from the shortest paths tree and predecessor sets along the paths, compute the dependencies $\delta_s(v)$ for all other $v \in V$. To compute the centrality value of a vertex v, we finally compute the sum of all dependency values. The computational complexity of the algorithm is $O(mn)$ and the space requirements are $O(m + n)$.

We present the first parallel centrality algorithm in [2]. Observe that parallelism can be exploited at two levels in the Betweenness Centrality algorithm: The BFS computations from each vertex can be done concurrently, provided the centrality running sums are updated atomically. Also, the actual BFS can be also be parallelized. When visiting the neighbors of a vertex, edge relaxation can be done concurrently. On multicore systems, we perform a coarse-grained partitioning of work and assign each processor a fraction of the vertices. The loop iterations are scheduled dynamically so that work is distributed as evenly as possible. There is synchronization cost involved, as a processor can compute its own partial sum of the centrality value for each vertex, and all the sums can be merged in the end using an efficient reduction operation. The BFS stack S, list of predecessors P and the BFS queue Q, are replicated on each processor, and so the space requirement is $O(mp)$.

3 Analysis

There are several online databases devoted to the human interactome (see Table 1). We derive our human protein interaction map (referred to as HPIN throughout the paper) by merging interactions from Gandhi et al.'s human proteome analysis dataset [9] (updated February 2006), the latest interaction dataset from the Human Protein Reference Database [21] (updated May 2006) and IntAct (updated October 2006). The latest version of the HPRD dataset includes interactions from MIPS, BIND, DIP and MINT. There is a complication using protein complex data (for example, from the MIPS database) to obtain protein interactions, since it is not always known which proteins in a complex interact with each other. We model complex interactions also as pair-wise interactions. This results in a high-confidence protein interaction network of 18869 proteins and 43568 interactions.

3.1 Structural Properties

We first study the topological characteristics of our human interactome HPIN. We primarily focus on the three aspects of network structure that receive most attention: degree distribution, diameter and characteristic path lengths and clustering and modularity.

On analyzing the connectivity properties of the graph, we find that largest connected component has 8510 proteins, and there are 9890 disconnected proteins in the graph (i.e., annotated proteins with no interactions). The average component size is 1.82. There are 85 connected components of size 2, and 14 connected components of size 3.

Previous studies on the yeast and fly datasets have shown that the degree distributions in protein-interaction networks obey a power law of the form $P(k) \approx k^{-\gamma}$. We observe a power-law distribution of the number of interactions in the human PIN, with $\gamma = 1.57$ (Figure 1(a)). We also study an early human PIN described by Lehner and Fraser [17], which includes predicted interactions generated using lower eukaryotic protein interaction data. We observe that this distribution has a heavier tail (a higher number of hub proteins) compared to HPIN. The models proposed to explain scale-free behavior, such as R-MAT [6], can mimic the power-law degree distribution behavior with the right parameters (see Figure 1(d)). Thus, the inferences drawn by studying the degree distributions of relatively small PINs such as yeast (Figure 1(c)) still hold for the human PIN.

Figure 2(a) plots the distribution of the shortest number of links between any two proteins in the network, or the shortest path. The longest short path, or the graph diameter, is 14 links. The average shortest path length is 3.72 in the network. This result is again in agreement with previously studied PINs and social networks. The average path length is an indicator of how readily information can be transmitted through a network. The small average path length, or the *small-world* property, suggests that such networks are efficient in the transfer of biological information. Only a small number of intermediate reactions are necessary for any protein to influence the characteristic behavior of another.

We also calculate the average clustering coefficient, a measure of the tendency of proteins in a network to form clusters or groups. For a vertex v of degree d, the clustering coefficient CC is defined as the $CC(v) = \frac{2k}{d(d-1)}$, where k is the number of links connecting the d neighbors of v, considered pairwise. The average clustering coefficient $CC_a(d)$ for a particular degree d is simply the average of the clustering coefficients of all vertices of degree d. We find that, on an average, CC_a decreases as the number of interactions per protein increases (Figure 2(b)). This indicates that the network has potential for hierarchical organization. Similar results were observed by Stelzl et al. [29] on smaller human protein interaction networks. The sparsely connected proteins are part of highly linked regions, which are connected via a few hubs. A distinguishing feature is that, in our case, we observe a considerable variation (two orders of magnitude) in the clustering coefficient for proteins with interactions between 50 and 200. This may be attributed to the presence of complex interactions in the network (i.e., more than two proteins participate in an interaction). In HPIN, there are several complex interactions, leading to the frequent occurrence of proteins of sizable degree as well as high clustering coefficients in the graph.

3.2 Centrality and Essentiality

The problem of identifying central nodes in large complex networks is of fundamental importance with applications in several areas varying from computational biology to

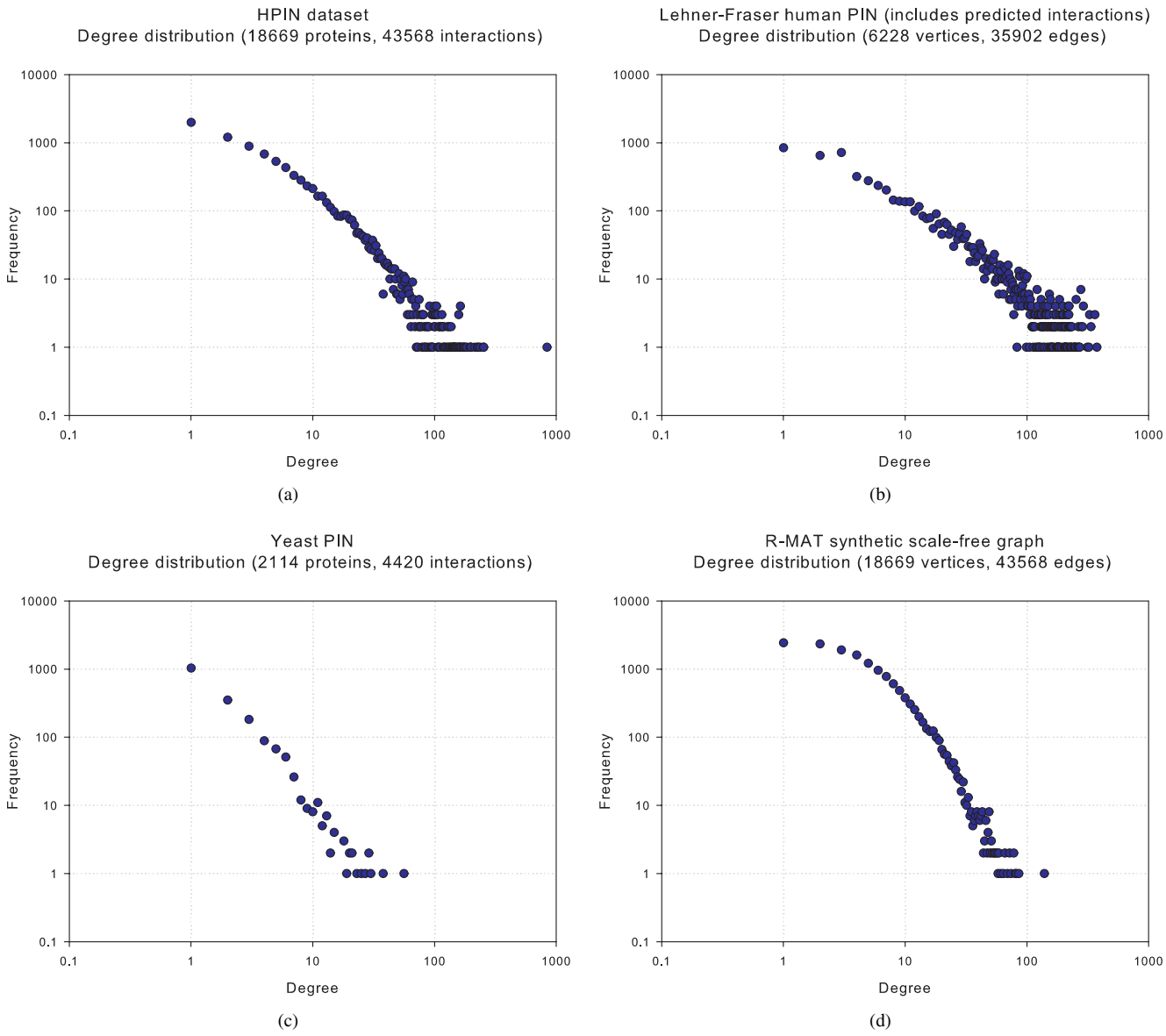
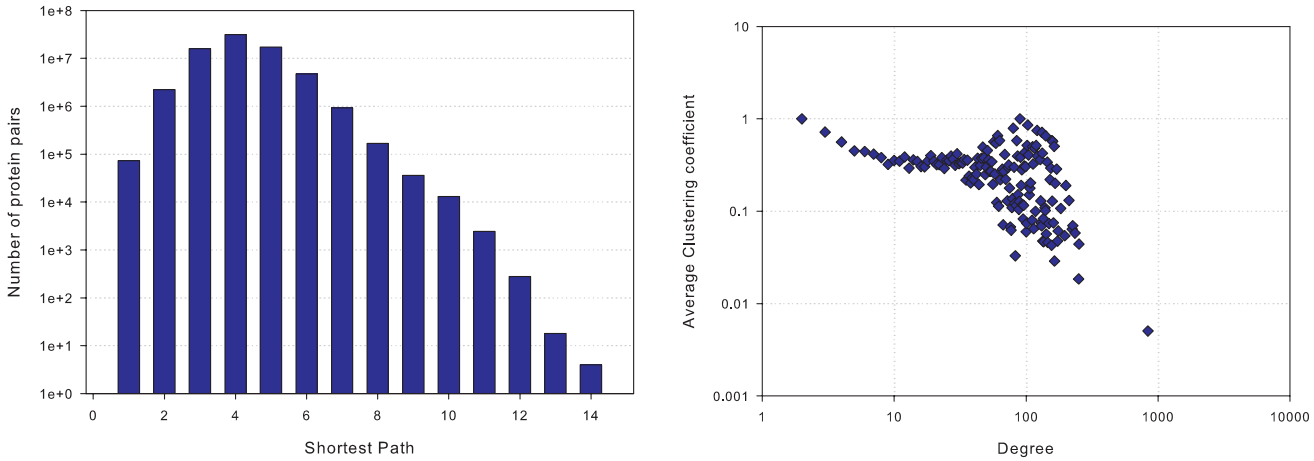


Figure 1. A comparison of degree distributions of various protein-interaction networks

Database	Details
HPRD [21]	Human Protein Reference Database. Experimentally verified protein-protein interactions obtained from manual curation of literature. 25209 proteins and 35262 interactions.
BIND	Biomolecular Interaction Network Database. Collection of molecular interactions including high-throughput data submissions and hand-curated information from the scientific literature. 4644 human protein interactions.
MIPS	Munich Information Center for Protein Sequences. 334 interactions.
MINT	Molecular Interactions Database. 3544 interactions.
IntAct [13]	Freely available, open source database system and analysis tools for protein interaction data. European Bioinformatics Institute. 2420 interactions.
OPHID	Online Predicted Human Interaction Database. Repository of already known experimentally derived human protein interactions, as well as 23,889 additional predicted interactions. This dataset is not included in our human PIN.

Table 1. Popular Online Human Protein Interaction databases



(a) Distribution of shortest paths between pairs of proteins

(b) Average Clustering Coefficient

Figure 2. Structural properties of HPIN

social and business networks. Many quantitative metrics for this purpose have originated from the social network analysis community, commonly referred to as *centrality measures*. We will use the Betweenness centrality metric (see Section 2) to analyze the human interactome. Researchers have paid particular attention to the relation between centrality and *essentiality* or *lethality* of a protein (for instance, [14]). A protein is said to be essential if the organism cannot survive without it. Essential proteins can only be determined experimentally, so alternate approaches to *predicting essentiality* are of great interest and have potentially significant applications such as drug target identification [15]. Previous studies on yeast have shown that proteins acting as hubs (or high-degree vertices) are three times more likely to be essential. So we wish to analyze the interplay between degree and centrality scores for proteins in the human PIN in this section.

Figure 3 plots the betweenness centrality scores of the top 1% (about 100) proteins in two lists, one ordered by degree, and the other by the betweenness centrality score. We observe that there is a strong correlation between the degree and betweenness centrality score: about 65% of the proteins are common to both lists. The protein with the highest degree in the graph also has the highest centrality score. This protein (*Solute carrier family 2 member 4*, Gene Symbol SLC2A4, HPRD ID 00688) belongs to the transport/cargo protein molecular class, and its primary biological function is transport. From Figure 3, it should also be noted that the top 1% proteins by degree show a significant variation in betweenness centrality scores. The scores vary by over four

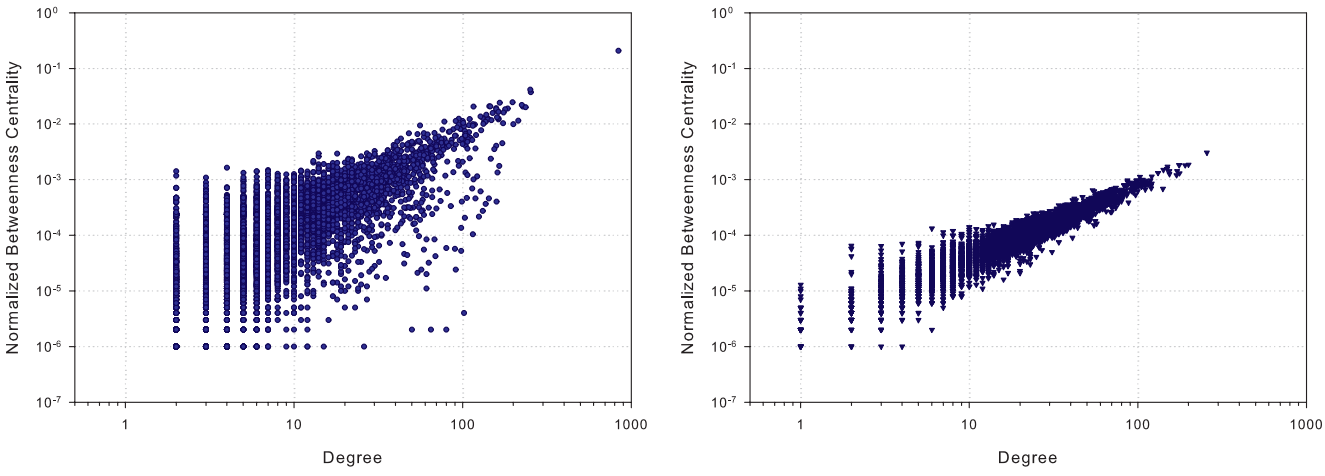


Figure 4. Normalized Betweenness Centrality scores as a function of the degree for HPIN (left) and a synthetic scale-free graph instance (right)

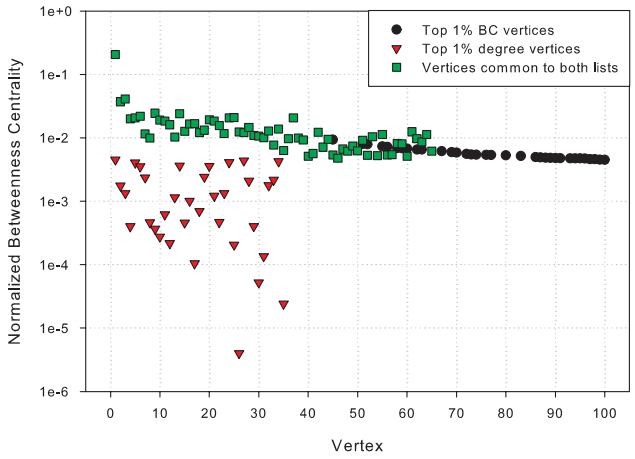


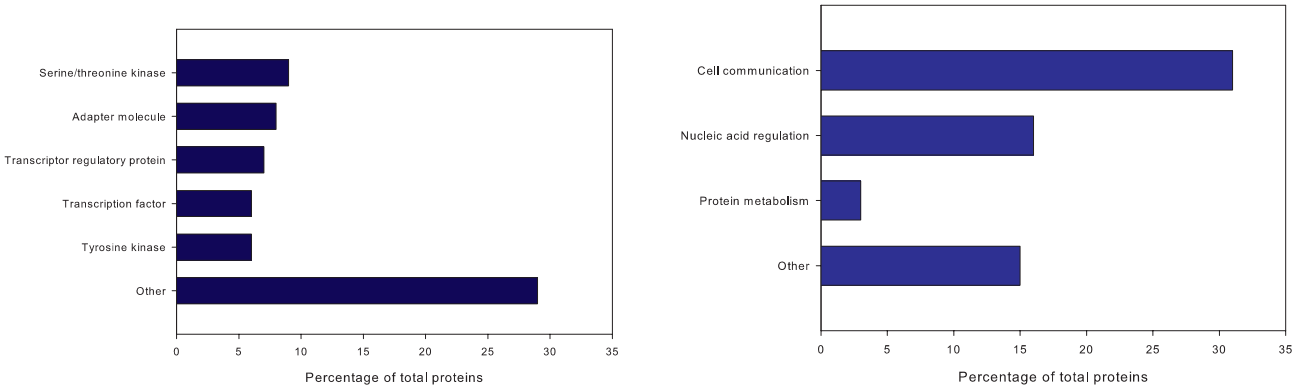
Figure 3. The top 1% proteins in HPIN, sorted by Betweenness Centrality (BC) scores and the number of interactions

orders of magnitude, from 10^{-1} to 10^{-4}.

We next study the correlation of degree with betweenness centrality. Unlike connectivity, which ranges from 1 to 822, the values of betweenness centrality range over several orders of magnitude. The few highly connected vertices have high betweenness values as there are many vertices directly and exclusively connected to these hubs. Thus most of the shortest paths between these nodes go through these hubs. However, the low-connectivity vertices show a significant variation in betweenness values, as evidenced in Figure 4(a). They exhibit a variation of betweenness values up to four orders of magnitude. The high betweenness scores may suggest that these proteins are globally important. Interestingly, these nodes are completely absent in synthetically generated graphs designed to explain scale-free behavior (observe the variation of betweenness centrality scores among low degree vertices in Figure 4(b)).

Our observations are further corroborated by two recent results. As the yeast PIN has been comprehensively mapped, lethal proteins in the network have been identified. Gandhi et al. [9] demonstrate from an independent analysis that the relative frequency of a gene to occur as an essential one is higher in the yeast network than the human PIN. They also observe that the lethality of a gene could not be confidently predicted on the basis of the number of interaction partners. Joy et al. [16] confirm that proteins with high betweenness scores are more likely to be essential, and that there are a significant number of high-betweenness, low-interaction proteins in the yeast PIN.

Figure 5 is a graphical representation of the dominant molecular class and biological function among high betweenness, high connectivity proteins (the common proteins in the top 1% lists). These proteins belong to a variety of molecular classes (Figure 5(a)), with cell communication and signal transduction being the most common biological function (Figure 5(b)).



(a) Molecular Class (b) Biological Function

Figure 5. The dominant molecular classes (left) and biological functions (right) among proteins that are common to both the top 1% betweenness centrality and degree lists

3.3 Parallel Multicore Performance

The sequential complexity for computing betweenness centrality and the graph diameter is $O(mn)$. The parallel algorithms for centrality described in Section 2 are well-suited for implementation on multicore and multiprocessor systems that have high memory bandwidth and a modest number of processors.

We report performance results on the Sun Fire T2000 multicore server, with the Sun UltraSPARC T1 (Niagara) processor. This system has eight cores running at 1.0 GHz, each of which is four-way multithreaded. There are eight integer units with a six-stage pipeline on chip, and four threads running on a core share the pipeline. The cores also share a 3 MB L2 cache, and the system has a main memory of 16 GB. There is only one floating point unit (FPU) for all cores. We compile our codes with the Sun C compiler v5.8 and the flags `-xtarget=ultraT1 -xarch=v9b -xopenmp`. The code is portable to other multicore and multiprocessor systems, and we make it freely available online [20].

Figure 6 plots the execution time and relative speedup achieved on the Sun Fire T2000 for computing the betweenness centrality on HPIN. The performance scales nearly linearly up to 16 threads, but plateaus between 16 and 32 threads. This can be attributed to insufficient memory bandwidth on 32 threads, as well as the presence of only one floating point unit on the entire chip. We use the floating point unit for accumulating pair dependencies and centrality values.

The execution times for betweenness centrality and graph diameter computation differ by a constant multiplicative factor. Betweenness centrality computation is much more involved, as it requires maintaining a BFS stack, a queue and a predecessor list. Also, the BFS tree is traversed twice in the algorithm.

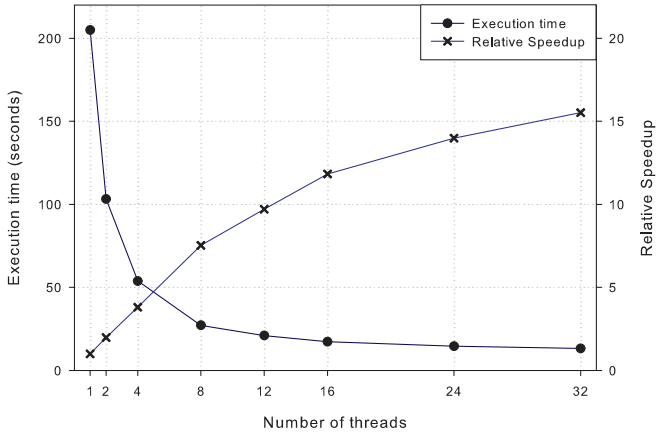


Figure 6. Betweenness Centrality Execution time and Speedup on the Sun Fire T2000 system

4 Conclusions and Future Work

We demonstrate the use of multicore algorithmic techniques for large-scale protein-interaction network analysis. The source code of the various graph analysis programs is freely available online from our web site. We also intend to provide the sequential version of our centrality analysis

codes as plug-ins to the visualization tool Cytoscape [28].

Protein-interaction networks (PINs) provides valuable insight into an organism's functional organization and evolutionary behavior. In recent years, the PINs of model eukaryotes like yeast and fly have been extensively analyzed. Our main contribution is the study of the topological properties of a PIN formed by high-confidence human protein interactions obtained from various public interaction databases. This is the largest human PIN constructed to date, comprising nearly 18,000 proteins and 44,000 interactions.

Predicting essentiality of a protein is of significant interest, and previous studies show that essentiality is highly correlated with betweenness. We compute betweenness and other centrality metrics on the human PIN. On analyzing these scores, we find that proteins with high betweenness and low connectivity are abundant in the human genome. We note that this result cannot be explained by the widely-accepted models for scale-free networks.

References

[1] C. Alfarano and *et al.* The biomolecular interaction network database and related tools: 2005 update. *Nucleic Acids Res.*, 33:D418–D424, 2005.

[2] D. Bader and K. Madduri. Parallel algorithms for evaluating centrality indices in real-world networks. In *Proc. 35th Int'l Conf. on Parallel Processing (ICPP)*, Columbus, OH, Aug. 2006. IEEE Computer Society.

[3] N. Batada and *et al.* Stratus Not Altocumulus: A New View of the Yeast Protein Interaction Network. *PLoS Biol.*, 4(10):e317, October 2006.

[4] V. Batagelj and A. Mrvar. Pajek – program for large network analysis. *Connections*, 21(2):47–57, 1998.

[5] U. Brandes. A faster algorithm for betweenness centrality. *J. Mathematical Sociology*, 25(2):163–177, 2001.

[6] X. Chen, J. Zheng, Z. Fu, P. Nan, Y. Zhong, S. Lonardi, and T. Jiang. Assignment of orthologous genes via genome rearrangement. *IEEE-ACM Trans. Computational Biology and Bioinformatics*, 2(4):302–315, 2005.

[7] T. Coffman, S. Greenblatt, and S. Marcus. Graph-based technologies for intelligence analysis. *Communications of the ACM*, 47(3):45–47, 2004.

[8] L. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, 1977.

[9] T. Gandhi and *et al.* Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nature Genetics*, 38:285–293, 2006.

[10] L. Giot and *et al.* A protein interaction map of drosophila melanogaster. *Science*, 302:1727–1736, 2003.

[11] M. Girvan and M. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences USA*, 99(12):7821–7826, 2002.

[12] R. Guimerà, S. Mossa, A. Turtschi, and L. Amaral. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proceedings of the National Academy of Sciences USA*, 102(22):7794–7799, 2005.

[13] H. Hermjakob and *et al.* IntAct: an open source molecular interaction database. *Nucleic Acids Res.*, 32:D452–D455, 2004.

[14] H. Jeong, S. Mason, A.-L. Barabási, and Z. Oltvai. Lethality and centrality in protein networks. *Nature*, 411:41–42, 2001.

[15] H. Jeong, Z. Oltvai, and A.-L. Barabási. Prediction of protein essentiality based on genomic data. *ComPlexUs*, 1:19–28, 2003.

[16] M. Joy, A. Brock, D. Ingber, and S. Huang. High-betweenness proteins in the yeast protein interaction network. *Journal of Biomedicine and Biotechnology*, 2:96–103, 2005.

[17] B. Lehner and A. Fraser. A first-draft human protein-interaction map. *Genome Biology*, 5(9):R63, 2004.

[18] S. Li and *et al.* A map of the interactome network of the metazoan *C. elegans*. *Science*, 303(5657):540–543, 2004.

[19] F. Liljeros, C. Edling, L. Amaral, H. Stanley, and Y. Åberg. The web of human sexual contacts. *Nature*, 411:907–908, 2001.

[20] K. Madduri. Graph analysis of the human interactome: supplemental results and source code. `http://www.cc.gatech.edu/~kamesh/HumanPPI`, 2006.

[21] S. Peri and *et al.* Development of fuman protein reference database as an initial platform for approaching systems biology in humans. *Genome Research*, 13:2363–2371, 2003.

[22] J. Pinney, G. McConkey, and D. Westhead. Decomposition of biological networks using betweenness centrality. In *Proc. 9th Ann. Int'l Conf. on Research in Computational Molecular Biology (RECOMB 2005)*, Cambridge, MA, May 2005. Poster session.

[23] W. Richards. International network for social network analysis. `http://www.insna.org`, 2005.

[24] W. Richards. Social network analysis software links. `http://www.insna.org/INSNA/soft_inf.html`, 2005.

[25] J.-F. Rual and *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437:1173–1178, 2005.

[26] L. Salwinski and *et al.* DIP: the database of interacting proteins: 2004 update. *Nucleic Acids Res.*, 32:D449–451, 2004.

[27] J. Scott. *Social Network Analysis: A Handbook*. SAGE Publications, Newbury Park, CA, 2000.

[28] P. Shannon and *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504, 2003.

[29] U. Stelzl and *et al.* A human protein-protein interaction network: A resource for annotating the proteome. *Cell*, 122(6):957–968, 2005.

[30] A. Tong and *et al.* Global mapping of the yeast genetic interaction network. *Science*, 303(5659):808–813, 2004.

[31] P. Uetz and *et al.* A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature*, 403:623–627, 2000.

[32] S. Wuchty and E. Almaas. Peeling the yeast protein network. *Proteomics*, 5(2):444–449, 2005.