A Decomposition Approach for Discovering Network Building Blocks*

Qiaofeng Yang
Physical Biosciences Division
Lawrence Berkeley National Laboratory
Berkeley, CA 94720, USA
qyang@lbl.gov

Stefano Lonardi
Department of Computer Science & Engineering
University of California
Riverside, CA 92521, USA
stelo@cs.ucr.edu

ABSTRACT

The increasing availability of biological networks (protein-protein interaction graphs, metabolic and transcriptional networks, etc.) is offering new opportunities to analyze their topological properties and possibly gain new insights in their design principles. Here we concentrate on the problem of *de novo* identification of the building modules of networks, which we refer to as *network modules*.

We propose a novel graph decomposition algorithm based on the notion of edge betweenness that discovers network modules without assuming any *a priori* knowledge. We claim that the knowledge of the distribution of network modules carries more information than the distribution of subgraphs which is commonly-used in the literature. To demonstrate the effectiveness of the statistics based on network modules, we show that our method is capable of clustering more accurately networks known to have distinct topologies, and that the number of informative components in our feature vector is significantly higher. We also show that our approach is very robust to structural perturbations (i.e., edge rewiring) to the network. When we apply our algorithm to protein-protein interaction (PPI) networks, our decomposition method identifies highly connected network modules that occur significantly more frequently than those found in the corresponding random networks. Detailed inspection of the functions of the over-represented network modules in *S. cerevisiae* PPI network shows that the proteins involved in the modules either belong to the same cellular complex or share biological functions with high similarity. A comparative analysis of PPI networks against AS-level Internet graphs shows that in AS-level networks highly connected network modules are less frequent but more tightly connected with each other.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Biology and Genetics; D.2.8 [**Software Engineering**]: Design—*Methodologies*

General Terms

Graph theory

*This project was supported in part by NSF CAREER IIS-0447773, and NSF DBI-0321756.

BIOKDD '07 San Jose, California USA

1. INTRODUCTION

Many real world systems can be modeled as network graphs, and their formal analysis can help us understand the underlying design principles behind each corresponding system. For example, identifying highly connected subgraphs in protein-protein interaction graphs can potentially enable life scientists to discover new protein complexes or speculate about the functions of unknown proteins [3, 23, 6]. In addition, the topological analysis can offer new insights in the roles of structural elements on the network performance, such as, traffic flow or diffusion of computer viruses over the Internet, epidemic diseases or ideas spreading in social networks, error and attack tolerance of various communication networks, etc.

In the past few years, a significant research activity has been focused on studying global and local properties of the network graphs (see, e.g., [7, 4, 27]) and significant breakthroughs have been achieved. For instance, the concept of scale-free networks, and the small world phenomenon have changed the way we model and analyze graphs across many different disciplines, from biological networks, to social networks all the way to communication networks.

In an attempt to understand the design principles of networks, the concept of *network motif* [18] has been recently proposed to represent the subgraphs in the network that occur significantly more often than the number of times they occur in the corresponding random networks. By using the concept of network motif, the authors of [18] were able to show that similar motifs were found in several information processing networks irrespective of their origin. They argued that these motifs may define universal classes of networks. The concept of network motif has been widely adopted to study local properties of various biological networks. For example, the network motifs in the transcriptional regulation network of *E. coli* were studied by Shen-Orr *et al.* [24]. The authors found that three highly significant motifs, namely, the *feed-forward loop*, the *single input module* and the *dense overlapping regulons*, are the main building blocks of the network. They also discovered that each motif is associated with a specific function in determining gene expression. A large collection of metabolic pathway networks were analyzed by Koyuturk *et al.* in [13]. The authors designed

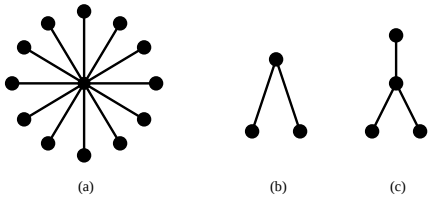

Figure 1: Illustrating the bias introduced by the occurrences of *hubs* (a) on the counts of subgraphs (b) and (c)

an efficient algorithm based on the *frequent itemsets* algorithm [1, 10] to find frequent subgraphs in the metabolic networks of over 150 organisms. Wuchty *et al.* [28] studied the conservation of 678 yeast proteins with the corresponding ortholog proteins in five higher eukaryotic organisms. The authors discovered that the orthologs are not randomly distributed in the yeast protein interaction network but are the building blocks of larger cohesive motifs, which tend to be evolutionarily conserved. They also observed that larger motifs tend to be conserved as a whole, with each of their components having an ortholog. Yeger-Lotem *et al.* [31] proposed the concept of *composite network motifs*, which consist of patterns from both transcription-regulation and protein-protein interaction networks that appear significantly more often than in random networks. They detected two-protein, three-protein, and four-protein motifs that occur in both networks.

Recently, the concept of network motif has been used to classify graphs. Milo *et al.* [17] introduced the concept of *significance profile* which is computed over the small subgraphs of the network and is used to cluster different networks. The profile is a normalized z-score for each subgraph obtained by comparing the number of occurrences of the subgraph to the number of occurrences in corresponding random networks. The authors were able to show that all networks having similar functionality share similar profiles. Surprisingly a few super-families of unrelated networks also share very similar significance profiles. Along the same line, Middendorf *et al.* [16] proposed a discriminative approach to understand the design of complex networks. The authors built a classifier based on alternating decision tree and trained the classifier using raw subgraph counts of 148 subgraphs obtained from seven random graph models. The protein-protein interaction graph (PPI) of *D. melanogaster* was classified as duplication-mutation-complementation network [26].

While this paper was under review, a work by Luo *et al.* [14] appeared in the scientific literature. The authors present an agglomerative algorithm to identify biological modules in PPI based on the concept of betweenness and modularity [9, 19, 21].

We observe that the majority of the approaches mentioned above share two common features, namely (1) they are designed to operate on directed graphs and (2) they are based on the *exhaustive* enumeration of all the subgraphs (up to a given size) in the network. From here on, we refer to exhaustive subgraph enumeration approaches as *Subgraph Counting Network Motif (SCNM)* approaches. We observed that using the raw subgraph counts as an indicator of over-representation has an inherent shortcoming. This arise from the fact

Input: Graph G, integer k and a list L of all subgraphs g_i of size smaller or equal to k
Output: Number of occurrences of each subgraph g_i in L

```
C ← Connected_Components(G)
for each connected component G_d ∈ C do
    Enqueue(Q, G_d)
while Q ≠ ∅ do
    n, G_c ← 1, Dequeue(Q)
    if Num_Vertices(G_c) ≤ k do
        Update_Counts(L, G_c)
    else
        while n = 1 do
            e ← Edge_Betweenness(G_c)
            Remove_Edge(G_c, e)
            C ← Connected_Components(G_c)
            n ← Size(C)
        for each connected component G_d ∈ C do
            Enqueue(Q, G_d)
return L
```

Figure 2: Sketch of the edge betweenness decomposition algorithm

that some subgraphs substantially overlap with each other, which in turn creates strong biases in the absolute counts. For example, *hubs* (nodes with high degree) are quite common in PPI networks [11]. As illustrated in the example of Figure 1, if one hub of degree twelve (a) is present in the network, then we will observe 66 subgraphs of type (b) and 220 subgraphs of type (c). If the network under study has several hubs, then type (b) and type (c) subgraphs will be highly over-represented when compared to random networks and they will dominate the analysis. However, such subgraphs may well be totally irrelevant from a statistical or biological viewpoint.

Here we address this limitation of SCNM approaches by introducing a novel graph decomposition method based on the concept of edge betweenness [9, 19, 21]. Our method decomposes the network into a collection of small subgraphs (called *network modules*), and thereby creates a disjoint partitioning of the nodes. The fact that a node can belong to only one network module solves the problem of counting overlapping subgraphs, and potentially allows us to assign putative biological functions to the nodes involved in the same network module. In order to evaluate objectively the effectiveness of our method to extract important features from the graph, we compare it to SCNM approaches on the problem of graph classification (along the lines of [17]). Results show that our approach is more accurate in distinguishing networks known to have distinct topologies. Our method is also tested for robustness against random perturbations to the network (i.e., edge rewiring), and our findings suggest low sensitivity to small changes in the graph. Finally, we report on preliminary results on the analysis of several protein-protein interaction networks (PPI). We show that highly connected network modules are more over-represented in PPI networks than those found in their random counterparts, and that the proteins involved either belong to the same cellular complex or share highly similar functions.

2. AN EDGE BETWEENNESS DECOMPOSITION ALGORITHM

It is well-known that proteins that are involved in the

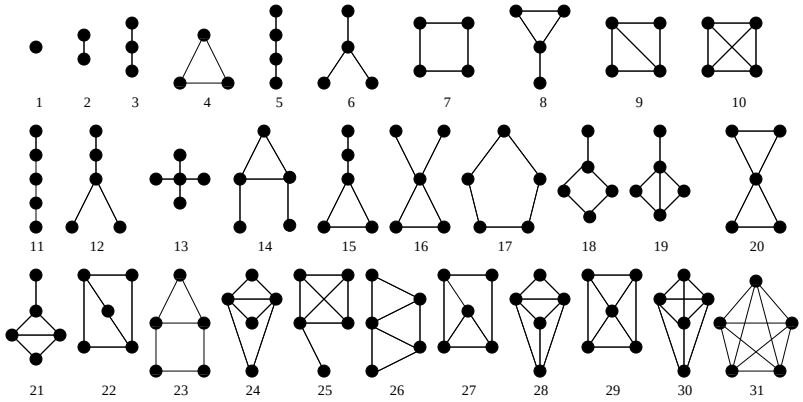

Figure 3: Non-isomorphic subgraphs of size ranging from one to five nodes

same cellular process or reside in the same protein complex are expected to have strong interactions with their partners. At the same time, interactions between distinct functional modules are expected to be suppressed in order to increase the overall robustness of the network by localizing effects of deleterious perturbations [15]. Biological networks are believed to consist of different modules with distinct functions [11, 22]. Here we are interested in identifying the building blocks of these functional modules without any *a priori* biological knowledge.

In this study, the detection of the building modules is based solely on the concept of edge betweenness. Consider the shortest paths between all pairs of vertices in a graph. The *betweenness* of an edge [9] is defined as the number of these shortest paths running through it[1]. When two different functional modules are loosely connected with each other, all shortest paths between vertices in those two modules have to traverse the few links between them. By removing those edges, the functional modules are separated from one another. The effectiveness of the betweenness approach on PPI graph in decomposing the network to find functional modules has been recently reported in [6]. In order to find the basic building modules of the network, we proceed as follows. First, we compute the edge betweenness of all the edges. Then, we start removing the edges with the highest betweenness until the largest connected component of the graph becomes smaller than or equal to some predefined threshold (k). Each time we remove an edge, the betweenness is recomputed from scratch. All the "small" connected components are then classified and counted. We refer to all the classified small subgraphs as *network modules*.

The outline of the algorithm is sketched in Figure 2. The function EDGE_BETWEENNESS computes and returns the edge with the largest edge betweenness. Evaluating the betweenness value for all edges of graph $G = (V, E)$ requires $O(|V||E|)$ time, by running a BFS from each node of the graph. The iterative removal of all $|E|$ edges leads an overall worst-case time complexity of $O(|V||E|^2)$ for our approach. Because of its computational cost, a distributed implementation of EDGE_BETWEENNESS was used [30].

When comparing our approach to Newman and Girvan method [19, 21], several major differences emerge. Although both algorithms employ betweenness to determine the order in which edges have to be removed, Newman and Girvan's relies on a metric that evaluate the quality of the decomposition, called *modularity*. In their method, the final decomposition is obtained by "cutting" the dendrogram of the decomposition at the point in which the value of the modularity peaks. In our method, we keep removing edges until the graph disconnects; only if the component is small enough, we stop the process and classify the module in one of 31 non-isomorphic subgraphs (shown in Figure 3).

Note that in our approach each vertex can only belong to one network module, in contrast to the *network motifs* widely used in the literature [18, 24, 28, 31, 16], which are based on exhaustive subgraph counting (SCNM) approach. To make a distinction between our approach and SCNM approach, we refer to our method as *Graph Decomposition Network Module* (GDNM) approach.

[1]If multiple shortest paths between a pair of nodes exists, each shortest path contributes an equal fraction to the edge-betweenness of their edges [5].

Table 1: The set of graphs used in the experiments

ID	*name*	$\|V\|$	$\|E\|$
1	*H. pylori* PPI	702	1359
2	*H. sapiens* PPI	1059	1318
3	*C. elegans* PPI	2629	3970
4	*S. cerevisiae* PPI	4770	15181
5	*D. melanogaster* PPI	7057	20815
6	*E.coli.* Transcription	418	519
7	*S. cerevisiae* Transcription	688	1078
8	*C. elegans* Neuron Connectivity	202	1952
9	AS1	3522	6324
10	AS2	4885	9276
11	AS3	7246	14629
12	AS4	10515	21455
13	AS5	4686	8772
14	AS6	9200	28957
15	Circuits1	122	189
16	Circuits2	252	399
17	Circuits3	512	819
18	Protein Structure1	95	213
19	Protein Structure2	53	123
20	Protein Structure3	97	212
21	Social1	67	142
22	Social2	32	80
23	Japanese	2704	7998
24	English	7381	44207
25	French	8325	23841
26	Spanish	11586	43065

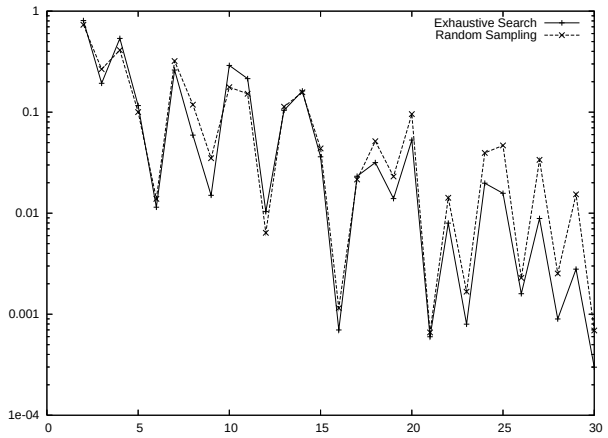

Figure 4: Comparing the exhaustive subgraph enumeration and random sampling on the graph Protein Structure3. The x-axis represents the subgraph index (according to Figure 3), whereas the y-axis represents the subgraph concentration. Subgraph of size 3, 4 and 5 were sampled 100,000 times

3. REPRESENTATION OF GRAPH FEATURES

Since the number of possible subgraphs grows exponentially with the number of nodes, in this study we only consider the number of occurrences of network modules of size up to five nodes (as in papers [20, 28]). As illustrated in Figure 3, there are 31 non-isomorphic subgraphs of size up to $k = 5$. Each subgraph g_i is indexed by an integer $i = 1, \ldots, 31$.

When a graph G is processed by the algorithm in Figure 2 where $k = 5$ and $L = \{g_1, \ldots, g_{31}\}$, a feature vector of 31 components is returned. Note that the number of occurrences of subgraphs of size one and two in the SCNM approaches it is somewhat meaningless, since they correspond respectively to the number of nodes and the number of edges in the graph. As a consequence, the feature vector for the exhaustive subgraph counting is 29-dimensional for $k = 5$. In our approach it is meaningful to keep track of all those 31 counts because when the network is broken down into connected components, some of those components may just have one or two nodes.

Before we can use these feature vectors to classify graphs, we need to normalize the components to remove the dependency on the absolute size of the graph. This will allow us to compare graphs of different sizes. We consider two normalizations, as explained below.

3.1 Subgraph Proportion Normalization

The first normalization tries to capture what proportion of nodes belongs to each subgraph class g_i. Given a graph $G = (V, E)$ and the vector $[n_i]$ of network module counts, the i-th component of the *subgraph proportion* vector is defined as $n_i|g_i|/|V|$ where n_i is the number of occurrences for subgraph class g_i. In the following we will use this normalization for the feature vectors associated with network building modules computed by our GDNM decomposition. Note, that since $\sum_{i=1}^{31} n_i|g_i| = |V|$, the sum of all the components of the subgraph proportion vector is always 1.

3.2 Subgraph Concentration Normalization

The second (alternative) normalization denotes how frequent is one subgraph class with respect to all the other classes with the same number of nodes. Given the vector $[n_i]$ of subgraph counts, the i-th component of the *subgraph concentration* [12] vector is defined as $n_i / \sum_{j:|g_j|=|g_i|} n_j$, where n_i is the number of occurrences of subgraph g_i. It is easy to realize that the sum of all the components of the subgraph concentration vector is always k. In the following we will use this normalization for the vector associated with the exhaustive SCNM approaches, since the subgraph proportion vector is not feasible for it. If we used the subgraph concentration normalization for the GDNM approach, we would loose the information carried by the network modules of size one and two (both components will be one).

4. RESULTS AND DISCUSSION

To test the effectiveness of our GDNM approach, we conducted several experiments and compared the results with the SCNM method. The first set of experiments is about graph classification, both on simulated data and on real networks (see Table 1 for a summary of the dataset). Five PPI networks were obtained from DIP database [29] and the rest of the networks are from [2]. We also performed a robustness test of our technique and computed the over-represented modules in PPI networks. Then, we studied the biological functions associated with the over-represented network modules found by our algorithm on the yeast PPI network.

4.1 Graph classification

4.1.1 Estimating the subgraph counts

Due to the large size of some of the networks in our dataset, the exhaustive subgraph enumeration is not always possible. In order to obtain the network motifs based on subgraph counting, we adopted the sampling algorithm by Kashtan *et al.* [12] to compute the number of occurrences of each subgraph in the network. For completeness of presentation, we briefly review the sampling procedure for a subgraph of size k. (1) Pick an edge $e = (u, v) \in E$ uniformly at random; (2) Set $U = \{u, v\}$ (3) Compute the set F of vertices that are adjacent to the vertices in U; (4) Pick one vertex from F at random and add it to U; (5) Repeat steps (3) and (4), until the target number k of vertices is reached.

Figure 4 shows a comparison between the exhaustive subgraph enumeration and the sampling approach for the "Protein Structure 3" network. The figure shows that the sampling algorithm gives good approximations of the subgraph concentration. We compared the sampling approach to the exhaustive count on many other relatively small graphs and in all cases it was capable of producing good estimates.

4.1.2 Classification of Real Networks

The real-world networks summarized in Table 1 were processed along the same lines as the previous experiment. It is worth noting that we treated all networks as undirected graphs although some of them (i.e., transcription regulation networks, social networks and language networks) are directed. Figure 5 shows the two Pearson correlation coefficient matrices for the 26 networks for our decomposition algorithm (left) and the subgraph counting approach (right).

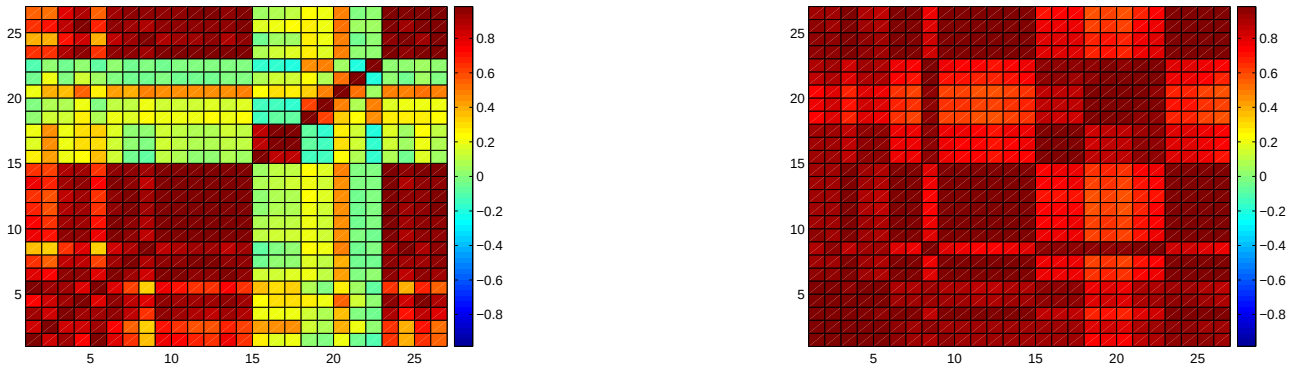

Figure 5: Pearson correlation coefficient matrix on the 26 real networks in Table 1 using decomposition network module approach (LEFT) and subgraph counting network motif approach (RIGHT)

Both pictures use the same scale. An inspection of the right matrix (corresponding to SCNM) shows that almost all networks are significantly correlated with one another. On the other hand, the feature vectors computed with our approach (left) show clearly that there are several distinct families of networks. The first is a big cluster composed of biological networks (PPI, transcriptional and neural), Internet AS-level networks, and languages networks, although the neural network does not share significant similarity with some members of this family. The second consists of circuit networks and the third consists of protein structure networks. The two social networks are not strongly correlated probably due to their small size. Note that circuit, protein structure and social networks are clustered together in the SCNM correlation matrix (right).

4.1.3 *Principal component analysis*

In order to establish an objective measure of the quality of the features extracted by the two approaches, we performed a principal component analysis (PCA) of the covariance matrices for both methods and both datasets (random and real data). The goal of this PCA analysis is to establish the effective dimensionality of the feature vectors obtained by the two methods. Figure 6 shows the distribution of the eigenvalues of the covariance matrix for random (left) and real networks (right). The value of the eigenvalues clearly illustrates that our decomposition method extracts more information from the graph. The analysis shows that our approach has a larger number of significant independent components in the feature vectors. For example on the random dataset, 11 principal components have significant eigenvalues whereas only three are obtained using the subgraph counting approach. On the real network dataset, our method extracts 21 significant components against 14 of the other approach. The fact that we have more "useful" components in our feature vectors can explain why our approach creates sharper and more accurate boundaries between different types of graphs.

4.2 Robustness

To test the sensitivity of the GDNM approach to random perturbation to the graph, we conducted a few experiments in which we swapped some of the edges of the network at random. This process is called *rewiring* [4], and works as follow.

Given a graph $G(V, E)$, randomly pick two edges $(u, v) \in E$ and $(x, y) \in E$. If $(u, x) \notin E$ and $(v, y) \notin E$, add (u, x) and (v, y) to E and delete (u, v) and (x, y) from E. Otherwise, if $(u, y) \notin E$ and $(v, x) \notin E$, add (u, y) and (v, x) to E and delete (u, v) an $\mathrm{d}(x, y)$ from E. If both choices are feasible, then whether we should connect (u, x) and (v, y) or (u, y) and (v, x) is arbitrarily chosen at random.

Figure 7 shows the profile of the vectors computed by our decomposition method before and after random perturbations up to 10% edge-rewiring on the PPI networks of yeast and fly. The figures indicate that our approach is quite robust to random perturbations.

4.3 Enrichment of Network Modules in PPI

We applied our GDNM algorithm to two large biological networks, namely, the protein-protein interaction (PPI) network for *S. cerevisiae* (yeast) and the PPI for *D. melanogaster* (fly). According to [8] the PPI of drosophila was obtained by high-throughput yeast two hybrid assays, whereas the source of the PPI data for yeast is a mix of mass spectrometry and yeast two hybrid assays. Our objective on PPIs is to identify network modules which are over-represented when they are compared to corresponding random networks, and possibly determine whether these over-represented modules are associated with important biological functions. We studied over-represented network modules both analytically and empirically. We performed an analytical analysis based on ER random graph model and an empirical analysis based on scale-free network model. We also report a preliminary comparative analysis of PPI and AS-level networks.

Consider an Erdos-Renyi (ER) random graph $G(V, E)$, which has $|V| = n$ labeled vertices and each pair of vertices is connected with probability p. Given G we want to calculate the expected number of occurrences of subgraphs $H_{r,l}$ with r vertices and l edges. Let $Z_{r,l}$ be the random variable associated with the number of subgraphs $H_{r,l}$ in G. The expected number of occurrences of $H_{r,l}$ can be obtained as follows

$$E(Z_{r,l}) = \binom{n}{r}\binom{r(r-1)/2}{l} p^l (1-p)^{(r(r-1)/2)-l}.$$

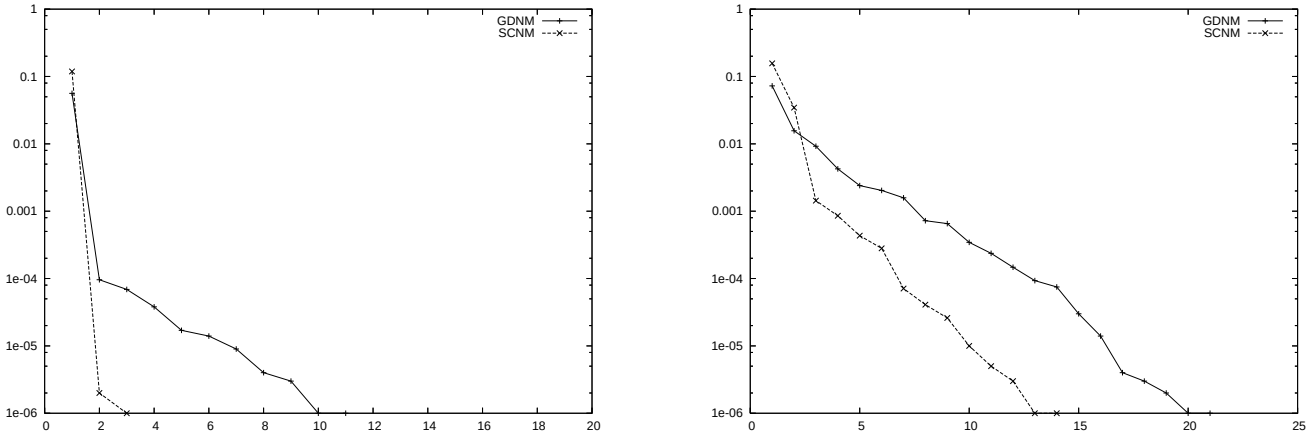

Figure 6: The eigenvalue distribution of the covariance matrix for 20 random networks (LEFT) and 26 real networks (RIGHT). The x-axis represents the ranks of the eigenvalues, the y-axis represent the absolute value of the eigenvalues

Indeed, there are $\binom{n}{r}$ ways of selecting r vertices from n vertices, and the maximum number of edges over r vertices is $\binom{r}{2} = r(r-1)/2$. The probability of observing l edges given r vertices is therefore $\binom{r(r-1)/2}{l} p^l (1-p)^{(r(r-1)/2)-l}$. The value of $E(Z_{r,l})$ is not a tight reference point when used to evaluate the significance of the subgraph counts obtained using our GDNM approach. The reason is that the count captured by $Z_{r,l}$ include overlapping and disconnected subgraphs, whereas our approach only considers non-overlapping and connected subgraphs.

Table 2 lists the observed and expected number of subgraphs $H_{r,l}$ in the yeast PPI network. It is obvious from Table 2 that densely connected subgraphs, such as $g_{28} - g_{31}$, are significantly over-represented when compared with the ER random graph model.

When comparing network module counts with the expected number of subgraphs in the ER random model, another fact need to be taken into account. Since our method removes edges with high betweenness first, it tends to favor highly connected subgraphs to sparser subgraphs. This observation has to be taken into account in the assessment of the statistical significance of these findings. In order to eliminate this bias, we also conducted an empirical analysis of the statistical significance, as described next.

To better understand the distribution of the number of subgraphs when the underlying random graph model has the same degree distribution as the original network, we performed an empirical study based on scale-free network model. The random networks were generated using the same method used to generate the scale-free networks above, but this time the degree distributions are that of the yeast and fly PPI networks. We made sure that the degree distributions are well preserved between real and random networks (statistics not shown). Our GDNM approach was subsequently applied on the scale-free random networks.

Figure 8 shows the profile of the subgraph proportion vectors for yeast (left) and fly (right) networks compared to the subgraph proportion vectors obtained from the random networks with the same degree distribution (averaged over 10 random networks). The comparison shows that large highly-connected subgraphs (i.e., those with high subgraph

Table 2: The observed and expected number of subgraphs with r vertices and l edges.

Network module	r	l	Observed	Expected
g_3	3	2	214	97629
g_4	3	3	8	45
$g_5 - g_6$	4	3	118	1.05 $e6$
$g_7 - g_8$	4	4	20	1085
g_9	4	5	6	0.60
g_{10}	4	6	3	1.38 $e{-4}$
$g_{11} - g_{13}$	5	4	137	1.41 $e7$
$g_{14} - g_{18}$	5	5	18	23430
$g_{19} - g_{23}$	5	6	26	27
$g_{24} - g_{27}$	5	7	18	0.02
$g_{28} - g_{29}$	5	8	7	1.10 $e{-5}$
g_{30}	5	9	18	3.38 $e{-9}$
g_{31}	5	10	29	4.66 $e{-13}$

indices) occur significantly more often in PPI networks than in random networks. This indicates that the occurrences of densely connected modules in PPI networks cannot be explained by chance and may imply important biological roles in the cell. When interpreting these results, we should not forget how the PPI data is collected. For example, since co-immunoprecipitation detects multi-protein complexes, this in turn can possibly bias the number of occurrences of cliques or other highly connected modules. An open question is how to correct for this bias, since the technology used in the collection of protein interaction data is likely to stay with us, at least in the short term.

It is clear from both analytical and empirical approaches that densely connected modules are significantly over-represented. In order to gain some insights in the functions of these modules in PPI networks we concentrated on module g_{31} (5-clique), which is one of the statistically significant modules identified in the yeast network. The functional analysis of the 29 occurrences of module g_{31} obtained by our algorithm reveals two classes of modules. In the first we found cellular protein complexes, such as 26S protease, RNA polymerase II, spliceosome, origin recognition complex, nuclear pore complex, etc. In the second, we found proteins

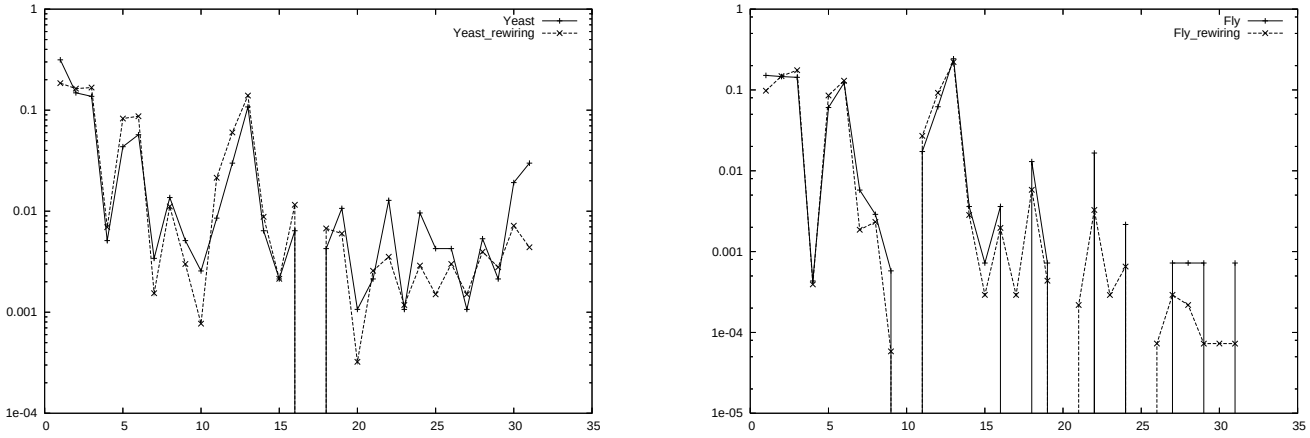

Figure 7: Testing the robustness of our decomposition approach before and after 10% edge rewiring in *S. cerevisiae* (LEFT) and *D. melanogaster* (RIGHT)

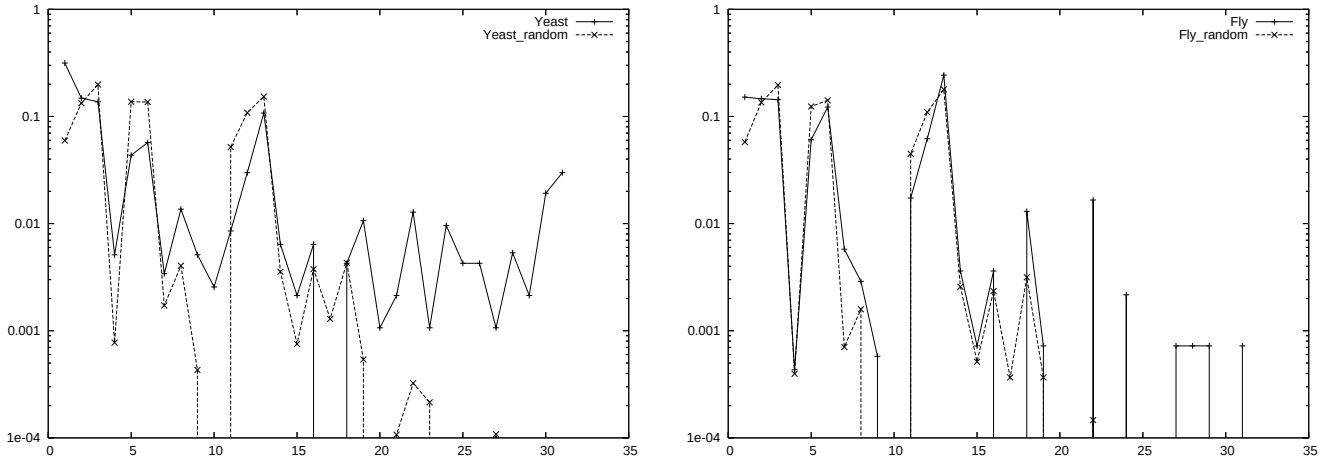

Figure 8: Comparing the occurrences of network modules in *S. cerevisiae* (LEFT) and *D. melanogaster* (RIGHT) against the corresponding random graphs (averaged over 10 random graphs)

that share highly similar functions, of which are involved in transcription regulation, translation initiation, cell cycle control, cellular transportation, mRNA processing, signal transduction cascades, etc. The functional categories of the 29 occurrences of module g_{31} are summarized in Table 3. Examples of the proteins involved in some of the modules g_{31} are given in Table 4. Due to lack of space, we refer the reader to `http://www.cs.ucr.edu/~qyang/` for the complete set of annotations.

We also performed a comparative analysis of the network modules in PPI networks against Internet AS-level networks. The goal of the analysis was to determine whether the over-represented modules in PPI are more or less interconnected than in the AS-level graphs AS4 and AS5. Both PPI and AS-level graphs have a skewed degree distribution. The "rich club connectivity" [32] analysis on the AS4 and AS5 reported one 10-clique among the vertices with the highest degree (data not shown), which is referred as the *core* of the Internet. Figure 9 shows that the yeast PPI has significantly more occurrences of large network modules (e.g., $g_{25}, g_{26}, \ldots, g_{31}$) than AS4 and AS5. Internet AS-level networks are known

Table 3: Distribution of the 5-cliques based on function annotation in *S. cerevisiae* PPI network

Function Category	Number of 5-cliques
Transcription	7
mRNA processing	5
Cell cycle	5
Cellular transportation	4
Metabolism	3
Translation	2
Cytoskeleton	1

to have highly connected core structure, where the links inside the core carry higher amount of communication flow than rest of the links in the network. Therefore, links inside the core will have higher betweenness and will be removed first in the decomposition process. The consequence is that in AS-level networks the resulting decomposition will lack these large network modules. In contrast, the highly con-

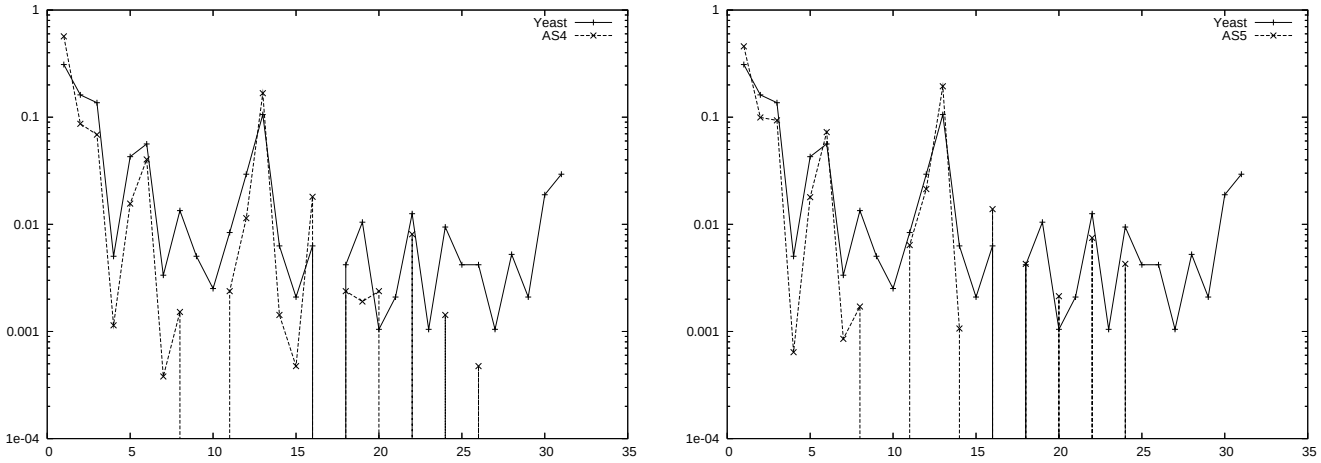

Figure 9: Comparing the occurrences of network modules between *S. cerevisiae* PPI network and the Internet AS-level network AS4 (LEFT) and AS5 (RIGHT)

nected large modules in PPI networks tend to be more frequent and more loosely connected with each other. This may indicate that PPI networks are organized in a decentralized manner across multiple functional domains, inside which strong connections among proteins may constitute the core facility for carrying out specific functions.

5. CONCLUSIONS

In this paper we proposed a new graph decomposition approach that is based on the concept of edge betweenness. The decomposition breaks the network into a set of small network modules, whose frequency of occurrence is then mapped to feature vectors and then normalized. The experiments show that our decomposition method produces normalized feature vectors that more clearly define classes of graphs than the ones produced by the subgraph counting (network motif) approach. More specifically, the analysis of the eigenvalues of the principal components of the covariance matrices shows that our approach extracts a larger number of independent informative features.

Our method turns out to be quite robust to edge rewiring and therefore not over-sensitive to small perturbations to the graph. The analysis of the PPI networks of yeast and fly has identified several over-represented modules when compared to random networks with the same degree distribution, and AS-level Internet graphs. A preliminary investigation on the proteins associated with the cliques found by our decomposition algorithm on the yeast PPI network shows that the proteins involved either belong to the same complex or share similar biological function.

We conclude by addressing some of the limitations of our method that could point to future research direction. The main advantage of a decomposition approach is that one node belongs to only one module, thereby solving the problem of over-counting overlapping subgraphs. However, on PPI graphs this is also a disadvantage because one protein can belong to only one network module, but it is well-known that proteins can be involved in multiple pathways or complexes. In order to capture the notion of "soft-partitioning" on graphs, a radically novel approach might be needed. For example, recent approaches [33] use the notion of information bottleneck [25] to obtain soft partitions of graphs. Also, although our method is not as expensive as the process of counting exhaustively all the subgraphs in a large network, it is still quite computationally intensive. The high computational cost of our method and other graph clustering methods remains an hindrance to their application on large networks.

6. REFERENCES

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. *Proc. 20th Int. Conf. Very Large Data Bases*, pages 487–499, 1994.

[2] U. Alon. `http://www.weizmann.ac.il/mcb/UriAlon/`.

[3] G. D. Bader and C. W. V. Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4, 2003.

[4] A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 2002.

[5] U. Brandes. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25:163–177, 2001.

[6] R. Dunn, F. Dudbridge, and C. M. Sanderson. The use of edge-betweenness clustering to investigate biological function in protein interaction networks. *BMC Bioinformatics*, 6(39), 2005.

[7] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. *ACM SIGCOMM'99 Comput. Commun. Rev.*, 29:251–263, 1999.

[8] L. Giot, J. S. Bader, C. Brouwer, and *et al.* A protein interaction map of *Drosophila melanogaster*. *Science*, 302:1727–1736, 2003.

[9] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *PNAS*, 99(12):7821–7826, 2002.

[10] K. Gouda and M. J. Zaki. Efficiently mining maximal frequent itemsets. *Proceedings of the 2001 IEEE International Conference on Data Mining*, pages 163–170, 2001.

Table 4: Annotations of some 5-cliques in *S. cerevisiae* PPI network (all the annotation can be found at `http://www.cs.ucr.edu/~qyang/`)

	DIP ID	Description of the proteins	Molecular function
1	DIP:1112N	Pre-mRNA splicing factor PRP19	Involved in pre-mRNA splicing and cell cycle control
	DIP:1682N	Pre-mRNA splicing factor ISY1	
	DIP:1681N	Pre-mRNA splicing factor SYF1	
	DIP:1684N	Pre-mRNA splicing factor SYF2	
	DIP:1685N	Pre-mRNA splicing factor CLF1	
2	DIP:2285N	Origin recognition complex subunit 2	Components of origin recognition complex (ORC)
	DIP:2286N	Origin recognition complex subunit 3	
	DIP:2287N	Origin recognition complex subunit 4	
	DIP:2288N	Origin recognition complex subunit 5	
	DIP:2289N	Origin recognition complex subunit 6	
3	DIP:1704N	Eukaryotic translation initiation factor 3 RNA-binding subunit	Eukaryotic translation initiation factors which bind to the 40S ribosome and promote the binding of methionyl-tRNAi and mRNA
	DIP:2519N	Eukaryotic translation initiation factor 3 90 kDa subunit	
	DIP:5870N	Eukaryotic translation initiation factor 3 110 kDa subunit	
	DIP:4532N	Possible eukaryotic translation initiation factor 3 30 kDa subunit	
	DIP:2303N	Eukaryotic translation initiation factor 5	
4	DIP:1587N	26S protease regulatory subunit 6B homolog	Components of 26S protease complex
	DIP:2883N	26S protease regulatory subunit 7 homolog	
	DIP:2100N	26S proteasome regulatory subunit RPN10	
	DIP:5261N	26S proteasome regulatory subunit RPN9	
	DIP:2808N	Proteasome component C11	
5	DIP:866N	Nucleoporin NUP57	Components of the nuclear pore complex (NPC)
	DIP:709N	Nucleoporin NUP49/NSP49	
	DIP:2074N	Nucleoporin NUP145 precursor	
	DIP:2430N	Nucleoporin NUP84	
	DIP:2721N	Nucleoporin NUP120	

[11] J.-D. J. Han, N. Bertin, T. Hao, D. S. Goldberg, G. F. Berriz, L. V. Zhang, D. Dupuy, A. J. M. Walhout, M. E. Cusick, F. P. Roth, and M. Vidal. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430:88–93, 2004.

[12] N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics*, 20:1746–1758, 2004.

[13] M. Koyutürk, A. Grama, and W. Szpankowski. An efficient algorithm for detecting frequent subgraphs in biological networks. *Bioinformatics*, 20:i200–i207, 2004.

[14] F. Luo, Y. Yang, C.-F. Chen, R. Chang, J. Zhou, , and R. H. Scheuermann. Modular organization of protein interaction networks. *Bioinformatics*, 23:207–214, 2007.

[15] S. Maslov and K. Sneppen. Specificity and stability in topology of protein networks. *Science*, 296:910–913, 2002.

[16] M. Middendorf, E. Ziv, and C. H. Wiggins. Inferring network mechanisms: The drosophila melanogaster protein interaction network. *PNAS*, 102:3192–3197, 2005.

[17] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of evolved and designed networks. *Science*, 303:1538–1542, 2004.

[18] R. Milo, S. 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.

[19] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in newtorks. *Physical Review E*, 69, 026113, 2004.

[20] N. Przulj, D. G. Corneil, and I. Jurisica. Modeling interactome: Scale-free or geometric. *Bioinformatics*, 20:3508–3515, 2004.

[21] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. *PNAS*, 101:2658–2663, 2004.

[22] A. W. Rives and T. Galitski. Modular organization of cellular networks. *PNAS*, 100(3):1128–1133, 2003.

[23] R. Sharan, T. Ideker, B. P.Kelley, R. Shamir, and R. M.Karp. Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. *RECOMB*, pages 282–289, 2004.

[24] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of escherichia coli. *Nature Genetics*, 31:64–68, 2002.

[25] N. Tishby, F. Pereira, and W. Bialek. The information

bottleneck method. In *Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999.

[26] A. Vazquez, A. Flammini, A. Maritan, and A. Vespignani. Modelling of protein interaction networks. *ComPlexUs*, 1:38–44, 2003.

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

[28] S. Wuchty, Z. N. Oltvai, and A.-L. Barabási. Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nature Genetics*, 35:176–179, 2003.

[29] L. Xenarios, L. Salwinski, X. J. Duan, P. Higney, S.-M. Kim, and D. Eisenberg. DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Research*, 30:303–305, 2002.

[30] Q. Yang and S. Lonardi. A parallel algorithm for clustering protein-protein interaction networks. *International Journal of Data Mining and Bioinformatics*, 1(3):241–247, 2007.

[31] E. Yeger-Lotem, S. Sattath, N. Kashtan, S. Itzkovitz, R. Milo, R. Y. Pinter, U. Alon, and H. Margalit. Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. *PNAS*, 101:5934–5939, 2004.

[32] S. Zhou and R. J. Mondragon. Accurately modeling the internet topology. *Physical Review E Phys. Rev. E*, 70, 2004.

[33] E. Ziv, M. Middendorf, and C. H. Wiggins. Information-theoretic approach to network modularity. *Phys. Rev. E*, 71, 046117, 2005.