# Bridging Centrality: Graph Mining from Element Level to Group Level

Woochang Hwang *
Department of Computer Science and Engineering,
State University of New York at Buffalo,USA
whwang2@cse.buffalo.edu

Taehyong Kim
Department of Computer Science and Engineering,
State University of New York at Buffalo,USA
thkim7@cse.buffalo.edu

Murali Ramanathan
Department of Pharmaceutical Sciences,
State University of New York at Buffalo,USA
murali@buffalo.edu

Aidong Zhang
Department of Computer Science and Engineering,
State University of New York at Buffalo,USA
azhang@cse.buffalo.edu

## ABSTRACT

Despite the pervasiveness of networks as models for real world systems ranging from the Internet, the World Wide Web to gene regulation and scientific collaborations, only a limited number of metrics capable of characterizing these systems are available. The existing metrics for characterizing networks have broad specificity and lack the selectivity for many applications. The purpose of this paper is to identify and critically evaluate a metric, termed bridging centrality, which is highly selective for identifying bridges in networks. The properties of bridges are unique compared to the other network metrics. For a diverse range of data sets, we found that networks are highly susceptible to disruption but robust to loss structural integrity upon targeted deletion of bridging nodes. A novel graph clustering approach, termed 'bridge cut', utilizing bridging edges as module boundary is also proposed. The modules identified by the bridge cut algorithm are more effective than the other graph clustering methods. Thus, bridging centrality is a network metric with unique properties that may aid in network analysis from element to group level in various areas including systems biology and national security applications.

## Categories and Subject Descriptors

[**Data**]: Data Structure, Graphs and networks

## General Terms

Measurement, Algorithms

*corresponding author


*KDD'08,* August 24–27, 2008, Las Vegas, Nevada, USA.


## Keywords

Bridging centrality, Graph clustering

## 1. INTRODUCTION

Networks emerge in many real world complex systems including technological, biological, and social systems. The Internet, the World Wide Web, gene regulation, business organizations, and scientific collaborations networks are common examples.

Network analyses have been introduced to identify important components or properties of network systems. Network structural properties of real world networks have been investigated and proved that these real world systems have interesting properties, e.g., small world phenomena and hierarchical modularity [4, 30]. The degree distribution of real world networks is inhomogeneous, decays monotonically and contains a limited number of highly connected nodes or hubs [4]. Real world systems are vulnerable to targeted attacks on important components, i.e., the ability to retain network's structural integrity when important nodes or edges are removed is deficient in real world systems. Real world networks are more robust than random networks when nodes are randomly removed [2]. Unfortunately, disruption of the network occurs concomitantly with loss of network integrity.

In element level analysis, several different centrality metrics such as degree, betweeness, clustering coefficient, and pagerank metrics have been proposed as graph theoretic metrics capable of identifying critical nodes [13, 30, 8]. Degree centrality is simply the number of edges that are incident to a node and this metric directly identifies highly connected nodes or hubs. Betweeness is a graph-theoretic metric that assesses the number of shortest paths passing through a given node [13]. Some variations of betweenness were proposed to measure the importance of elements in a network [13, 6, 12]. The clustering coefficient of a node assesses the local connectivity among direct neighbors of a given node [30]. The pagerank is a metric based on eigenvalues of the network connectivity structure that forms the basis for the Google search engine [8].

For many networks, the degree and betweenness metrics produce highly correlated results because the nodes with highest degree frequently also happen to have a large number of shortest paths passing through them. Furthermore, nodes with high values for these metrics have a high tendency to be located and gathered in the core part of a network because high degree nodes are closely connected each other and shortest paths among node pairs would likely pass through the core part of the network. The usefulness of clustering coefficient in identifying critical nodes is limited because high degree nodes typically have low clustering coefficient values. The motivation for this research is to develop a sophisticated metric capable of mediating network disruption without causing substantial loss of network structure and integrity. Such improved approaches for assessing the criticality of individual nodes and edges can enhance the use of network methods for decision-making in various important applications from national security threat assessment to drug discovery.

In group level analysis, several graph clustering approaches have been proposed to find effective functional modules in networks. For example, the reciprocal of the shortest path length and the hitting time for a random walk between two components has been investigated as a distance/similarity measure for distance based clustering [24]. In the maximal clique approach, clustering is reduced to identifying fully connected subgraphs in the graph [27]. To overcome the relatively high stringency imposed by the maximal clique method, the Quasi Clique [9], Molecular Complex Detection (MCODE) [3], Spirin and Mirny [27] algorithms identify densely connected subgraphs rather than fully connected ones either by optimizing an objective density function or by using a density threshold. The Restricted Neighborhood Search Clustering Algorithm (RNSC) [18] and Highly Connected Subgraphs (HCS) algorithms [15] harness minimum cost edge cuts for cluster identification. The Markov Cluster Algorithm (MCL) algorithm finds clusters using iterative rounds of expansion and inflation that promote the strongly connected regions and weaken the sparsely connected regions, respectively [29]. In another direction, statistical approaches to clustering have also been proposed. For example, Samanta and Liang [25] employed a statistical approach to clustering of nodes based on the premise that a pair of nodes sharing a significantly larger number of common neighbors will have high similarity. However, limited available information, e.g., binary nature of interactions among nodes, of real world networks limits the performance in clustering using conventional connectivity, similarity, and other topological and statistical features.

In this paper, we introduce a novel network feature termed bridging and propose a "bridging centrality" metric for identifying and assessing "bridges" that play critical bridging roles between submodules for networks in the element level. The bridging paradigm proposed is intuitive because of its consistency with the everyday notion of bridges in transportation and our results demonstrate that bridges are critical for modulating information flows and interactions between modules of networks. The nodes with high values of the bridging centrality differed in distinctive ways from nodes identified on the basis of betweeness centrality and other metrics. It is shown that bridging nodes are locating on the modulating positions among modules in various types of networks. The vulnerability of bridging nodes is unlike any of the other centrality metrics: they cause network disruption without dismemberment. In group level analysis, a novel graph clustering method called "bridge cut" is also proposed to identify effective modules in networks. Bridging edges that are identified by our bridging centrality are recognized as the boundary among modules in a network. Therefore, a network can be partitioned into effective modules utilizing bridging edges. Our bridge cut algorithm detected more effective modules in various types of networks than the other graph clustering approaches.

# 2. THE METHOD

## 2.1 Terminology and Representation

Let an undirected graph be $G = \{(V, E) \mid V$ is a set of nodes and $E$ is a set of edges, $E \subseteq V \times V$, an edge $e = (i, j)$ connects two nodes $i$ and $j$, $i, j \in V$ , $e \in E$ $\}$. The nodes in real world network's graph representation are the various entities such as persons, computers, or biomolecules in social networks, technical networks or biological networks, respectively. The edges between the nodes represent an interaction or relationship between the underlying entities.

**Network Properties:** The neighbors $N(v)$ of node $v$ is defined to be the set of directly connected nodes to node $v$. The degree $d(v)$ of a node $v$ is the number of the nodes directly connected to node $v$, i.e., cardinality of $N(v)$. A path is defined as a sequence of nodes $(n_1, \ldots, n_k)$ such that from each of its nodes there is an edge to the successor node. The path length is the number of edges in its node sequence. A shortest path between two nodes, $i$ and $j$, is a minimal length path between them. The distance between two nodes, $i$ and $j$, is the length of its shortest path. The clustering coefficient, $C_v$, of node $v$ measures the extent of the interconnectivity between the neighbors of node $v$ and is the ratio of the number of edges between the nodes in the direct neighborhood to the number of edges that could possibly exist among them:

$$C_v = \frac{2 \mid \bigcup_{i,j \in N(v)} e(i,j) \mid}{d(v)(d(v)-1)} : e(i,j) \in E \tag{1}$$

The clustering coefficient of a graph is the average of the clustering coefficients of all nodes in the graph. The density of a graph is defined as follows:

$$density(G) = \frac{2e}{n(n-1)}, \tag{2}$$

where $e$ is the number of edges and $n$ is the number of nodes in graph $G$. The direct neighbor subgraph of node $v$ is a subgraph $S(v)$ such that $S(v) = \{(V, E) \mid V$ is the set of nodes in the direct neighbors of node $v, E$ is the set of edges, $E \subseteq V \times V$ , an edge $e = (i, j)$ connects two nodes $i$ and $j$, $i, j \in V$ ,$e \in E\}$ All network figures were initially obtained using Pajek [5].

## 2.2 Bridging Centrality

We define a bridge to be a node or an edge connecting modular regions in a graph. We introduce a formula, termed bridging centrality, to quantitatively measure the extent of bridging capability of all nodes or edges in the network. The bridges in a graph can then be identified on the basis of their high value of bridging centrality relative to other components on the same graph.

**Definition 1:** A *bridge* is a node or an edge that is located between and connects modules in a graph. In other words, a *bridge* is a node $v$ or an edge $e$ that has high bridging centrality value.

To calculate the bridging centrality of a node $v$ or an edge $e$, we first compute global importance using betweenness centrality in a graph conceptually defined as follows:

$$\Phi(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \quad (3)$$

where $\sigma_{st}$ is the number of shortest paths between node $s$ and $t$ and $\sigma_{st}(v)$ is the number of shortest paths passing through a node $v$ out of $\sigma_{st}$.

Betweenness for an edge $e$ can be defined in the same way as the node case in Equation 3:

$$\Phi(e) = \sum_{s \neq t \in V} \frac{\sigma_{st}(e)}{\sigma_{st}} \quad (4)$$

where $\sigma_{st}$ is the number of shortest paths between node $s$ and $t$ and $\sigma_{st}(e)$ is the number of shortest paths passing through an edge $e$ out of $\sigma_{st}$.

To obtain a metric capable of identifying bridges, we utilize the idea that the number of edges entering or leaving the direct neighbor subgraph of a node $v$ relative to the number of edges remaining within the direct neighbor subgraph of a node $v$ is high at bridges. We use this property to formulate the concept of bridging coefficient for both nodes and edges.

**Definition 2:** The *bridging coefficient* of a node $v$ is defined as the average probability of leaving the direct neighbor subgraph of a node $v$. The *bridging coefficient* of a node $v$ is defined by:

$$\Psi(v) = \frac{1}{d(v)} \sum_{i \in N(v)} \frac{\delta(i)}{d(i) - 1} \quad (5)$$

where $d(v)$ is the degree of a node $v$ and $\delta(v)$ is the number of edges leaving the direct neighbor subgraph of node $v$ among the edges incident to each direct neighbor node $i$ of node $v$.

**Definition 3:** The *bridging coefficient* of an edge $e$ is defined as the product of weighted average of bridging coefficient of two incident nodes $i$ and $j$ for an edge $e$ and the reciprocal of the number of common direct neighbor nodes of nodes $i$ and $j$. The *bridging coefficient* of an edge $e$ is defined by:

$$\Psi(e) = \frac{d(i)\Psi(i) + d(j)\Psi(j)}{(d(i) + d(j))(|C(i,j)| + 1)}, \quad e(i,j) \in E \quad (6)$$

where nodes $i$ and $j$ are the two incident nodes to edge $e$, $d(i)$ is the degree of a node $i$, $\Psi(i)$ is the bridging coefficient of node $i$, $C(i,j)$ is the set of common direct neighbor nodes of nodes $i$ and $j$.

We used the rank product [7] which is defined as the product of rank of the betweeness and the rank of the bridging coefficient for computing the bridging centrality. This normalization procedure corrects for the differences in scale between the betweeness and the bridging coefficient.

**Definition 4:** The *bridging centrality* of a node $v$ is defined by:

$$C_{Br}(v) = R_{\Phi(v)} \cdot R_{\Psi(v)} \quad (7)$$

where $R_{\Phi(v)}$ is the rank of a node $v$ in betweenness and $R_{\Psi(v)}$ is the rank of a node $v$ in bridging coefficient. In rank product normalization, nodes in a graph are ordered by scores measured for each metric, e.g., bridging coefficient and betweenness. Then, we take the rankings of a node $v$ in the sorted order for each metric and bridging centrality uses the product of the rankings in each metric for a node $v$.

**Definition 5:** The *bridging centrality* of an edge $e$ is defined by:

$$C_{Br}(e) = R_{\Phi(e)} \cdot R_{\Psi(e)} \quad (8)$$

where $R_{\Phi(e)}$ is the rank of an edge $e$ in betweenness and is the rank of an edge $e$ in bridging coefficient.

In formulas 7 and 8, the first term, which measures global importance of a node or an edge, represents the fraction of shortest paths passing through a node or an edge. The second term measures the local topological property around a node or an edge, i.e., the probability of leaving the direct neighbor subgraph of a node or an edge. A bridge is a node $v$ or an edge $e$ that has high bridging centrality value relative to other nodes or edges on the same graph.

Based on these definitions, we hypothesize that the bridging centrality is capable of identifying nodes or edges that are located and connect subregions of the network and are therefore potential bottlenecks to information flow between modules.

## 2.3 Bridge Cut Algorithm

**Algorithm 1** BridgeCut(G)

```
1: G‘: A clone of graph G
2: ClusterList: the list of final clusters
3: topEdge: the edge with the highest bridging centrality
4: densityThreshold: subgraph density threshold
5: while G != empty do
6:    Calculate bridging centrality for all edges in graph G
7:    topEdge = The edge with the highest bridging centrality
8:    Remove topEdge
9:    if there is a new isolated module s then
10:      if Density(s,G‘) > densityThreshold then
11:         ClusterList.add(s)
12:         G.remove(s)
13:      end if
14:   end if
15: end while
16: Return ClusterList
```

Bridges are located between modules in a network. Therefore, a graph can be partitioned into submodules using identified bridges if we can utilize bridges as module boundary. In this section, a new novel graph partitioning algorithm utilizing bridging centrality will be introduced.

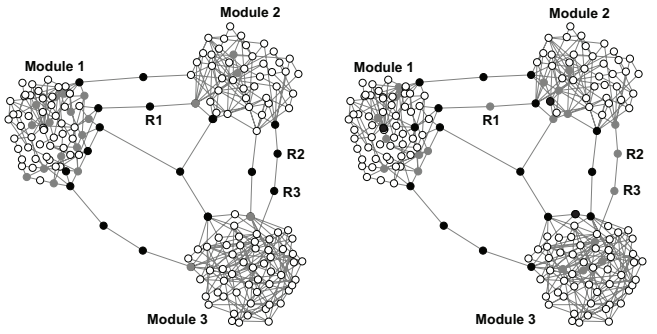
The iterative graph clustering algorithm involves three sequential processes:

**Process 1:** Compute bridging centrality of all edges in graph $G$ and pick edge $e$ with the highest bridging value.

**Process 2:** Remove edge $e$.

**Process 3:** Identify a subgraph $s$ as a final cluster and remove from $G$ if $s$ is isolated from $G$ and the density of $s$ in the intact graph $G‘$ is greater than threshold.

First, bridging centrality is calculated for all edges in a graph $G$ and the highest scored edge $e$ is picked up (Process 1). The highest bridging scored edge $e$ is removed from graph $G$ (Process 2). Finally, subgraph $s$ that is isolated from $G$ after edge $e$ cut, is recognized as a final cluster if the density of $s$ is more than the threshold (Process 3). These three sequential steps are repeated until $G$ is empty. Bridge cut algorithm is described in detail in Algorithm 1.




(a) (b)

**Figure 1: (a) and (b) shows the results of bridging centrality and betweenness centrality in the synthetic network, respectively. The network contains 192 nodes and 466 edges and was created by adding bridging nodes to three independently generated subnetworks. The nodes with the highest 0-10th percentile of values for each metric are highlighted in black circles; the nodes between 10th-20th percentile are colored in gray.**

**Cluster validation**

**$F$-measure:** $F$-measure uses external information, i.e., reference modules that can be compared against the identified modules, for cluster quality assessment. $F$-measure is calculated based on the precision and recall between the detected modules and the reference modules. Precision is the fraction of the elements in a module $C$ that are relevant for a reference module $G$ to the size of the cluster $C$. Recall is the fraction of the elements in a module $C$ that are relevant for a reference module $G$ to the size of the reference module $G$.

$$Precision = \frac{|C \cap G|}{|C|}. \tag{9}$$

$$Recall = \frac{|C \cap G|}{|G|}, \tag{10}$$

Precision shows the extent how well a identified module matches to the reference module and recall measures the extent how well a reference module matches the identified module. $F$-measure is defined as the harmonic mean of recall and precision.

$$F - measure = \frac{2(Precision \cdot Recall)}{Precision + Recall}. \tag{11}$$

**Davies-Bouldin index:** Davies-Bouldin (DB) index [10] measures the quality of a clustering using internal information only, i.e., diameters of each cluster and distance between all cluster pairs. DB index can be used as a cluster quality measure if we don't have any reference information to be compared. It measures the topological quality of the identified clusters in the intact graph.

$$DB = \frac{1}{k}\sum_{i=1}^{k} max_{i \neq j}[\frac{diam(C_i) + diam(C_j)}{d(C_i, C_j)}], \tag{12}$$

where $diam(C_i)$ is the diameter of cluster $C_i$ and $d(C_i, C_j)$ is the distance between cluster $C_i$ and $C_j$. So, $d_{C_i,C_j}$ is small if cluster $i$ and $j$ are compact and theirs centers are far away from each other. Therefore, DB will have a small values for a good clustering.

# 3. EXPERIMENTAL RESULTS

## 3.1 Data Sets

The AT&T Web network is one of the test graphs from the graph drawing competition at the Symposium on Graph Drawing, Berkeley, California, September 18-20, 1996 [1]. The physics collaboration network [22] was kindly provided by Dr. Newman. The school friendship network dataset was kindly provided by Dr. Lind [19]. The yeast metabolic network [14] was kindly provided by Dr. Guimera. The DIP core yeast (*S. cerevisiae*) protein-protein interaction (PPI) dataset is obtained from the DIP database [11]. Functional category data of yeast was obtained from MIPS database [20]. Spellman cell cycle gene expression data for yeast was obtained from [26].

## 3.2 Element Level: Bridging Centrality

### 3.2.1 Performance on Synthetic Data

To obtain a preliminary assessment of the underlying network characteristics identified by bridging centrality, we applied bridging centrality to a synthetic network consisting of 192 nodes and 466 edges shown in Figure 1. The network was created by joining 3 separate synthetic networks and contains key elements such as hub nodes, peripheral nodes, cycles with known bridges. The overall size was kept small so that any patterns present could be easily detected by visual inspection. In Figure 1(a), we have highlighted the nodes with the highest $0-10$th percentile of values for each metric are in black circles; the nodes between 10th$-20$th percentiles are highlighted in gray circles. Visual inspection of the synthetic network reveals that the highest values of bridging centrality occur in the nodes that connect the modules and highly connected parts of the network. Five bridging nodes emerged within Module 1 and one bridging node in Module 2: these are located between highly modular subregions of Module 1 or located on the extremity of effective bridges between modules. Figure 1(b) shows the performance of betweenness centrality on the same network. The same labeling scheme is used except that nodes are colored using the values of betweenness centrality. Betweenness centrality identified some of bridging nodes but failed to identify the major bridges labeled $R1$, $R2$, and $R3$.

### 3.2.2 Performance on Social Networks

**Physics Collaboration Network:** Networks are commonly used to represent social systems and the analysis of these social networks is important in national security applications. Social networks are distinctively different from computer and biological networks in their clustering properties and show positive correlations between degrees of adjacent nodes [22]. We analyzed a social network example to demonstrate that bridging centrality can be used to identify key bridging individuals in a social community.

The physics collaboration network was constructed [23] from the bibliography section of a review by Newman [22]. The bridging nodes (Figure 2) are strategically positioned

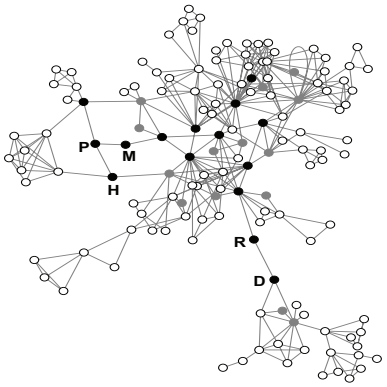



**Figure 2: The results for the physics collaboration network. The nodes with the highest 0-10th percentile of values for the bridging centrality are highlighted in black circles; the nodes in 10th-20th percentile are highlighted in gray circles. The nodes corresponding to Rothman (R), Hermann (H), Penna (P), Dodds (D), and Moukarzel (M) are labeled.**

on the paths between modular subcommunities [21]. The nodes corresponding to the physicists, Rothman and Dodds, have the highest and 4th highest bridging centrality values because the nodes are on the path providing the only connection between the two large communities in the network. The nodes corresponding to the physicists in University of California at Davis, Hermann, Penna and Moukarzel, which have the 2nd, 3rd and 5th highest bridging centrality values, are located between University of California at Davis and european groups.

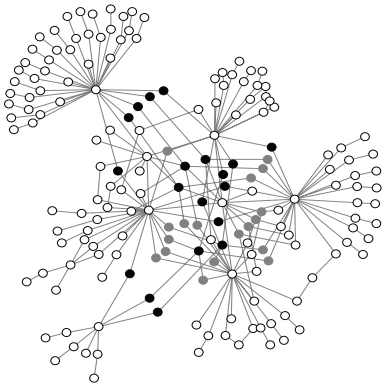


**Figure 3: The results for the AT&T Web network. The nodes with the highest 0-10th percentile of values for the bridging centrality are highlighted in black circles; the nodes in 10th-20th percentile are highlighted in gray circles.**

### 3.2.3 Performance on Technical Networks

**AT & T Web Network**: This network [1] has low modularity, which makes it difficult to differentiate modular regions and their connecting nodes. By visual inspection of Figure 3, it is apparent that bridging centrality successfully identified the bridging nodes in the AT&T Web Network despite the low network modularity.

### 3.2.4 Performance on Biological Networks

High throughput assay methodologies such as microarrays and mass spectrometry have resulted in rapid growth of biological network data sets, the analysis of which can potentially yield insights into the mechanisms of human disease and the discovery of new therapeutic interventions [16]. Biological networks can be diverse in structure but in many cases, involve ordered sequences of interactions rather than inter-connections. The majority of proteins in a given functional category do not have direct physical interaction with other proteins involved in the same function category [16].

**Yeast Metabolic Network:** We extended the results to the much larger well-studied yeast metabolic network [14], which contains 359 nodes and 435 edges in Figure 4. Again, despite the additional complexity and increased size of the network, nodes involved in bridging larger modules to each other are selectively identified.

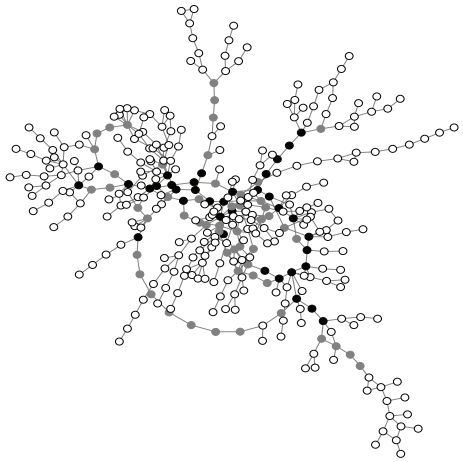
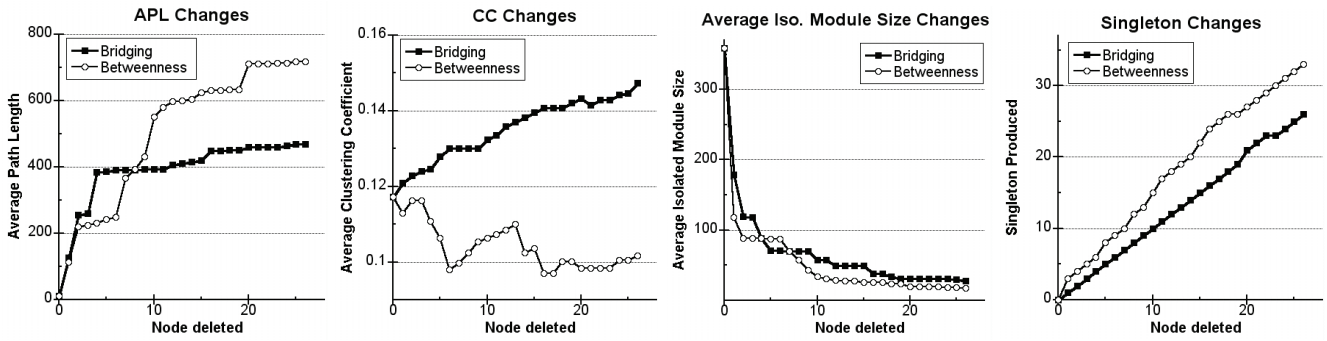


**Figure 4: The results for the yeast metabolic network. The nodes with the highest 0-10th percentile of values for the bridging centrality are highlighted in black circles; the nodes in 10th-20th percentile are highlighted in gray circles.**

### 3.2.5 Assessing Network Disruption, Structural Integrity and Modularity

The main objective of this study is to analyze the potential of bridging centrality score to select the nodes that position on true bridging locations. We use the yeast metabolic network for further analysis since it has better network properties, e.g., power law distribution, small world phenomenon, high modularity, than other examples in the above and also has moderate size that enables us to observe the performances precisely. In order to investigate the topological locality of the bridging nodes picked up by bridging centrality in networks, several network properties, e.g., the average path length, the average clustering coefficient, the average size of isolated module size, and the number of singletons occurrence, were analyzed and compared on sequential node removals for bridging and betweenness centrality in Figure 5. Betweenness is chosen as the competitor because it is the only comparable graph metric that has a similar semantics. In Figure 5, nodes were ordered by each centrality metric and sequentially removed to observe the changes of the network's properties.




(a) (b) (c) (d)

**Figure 5: Analysis on the yeast metabolic Network. (a)Average Path Length Changes, (b)Average custering Coefficient changes, (c) Average isolated module size changes, (d) Singleton Changes. Changes of the average path length, clustering coefficient, average isolated module size, number of singletons followed by the consecutive top 10 percentile high score node removals for two centrality measures (bridging, betweenness).**

As the first evidence of the bridging centrality's superiority on targeting the bridging nodes, we observed the topological properties of the bridging nodes discriminated by bridging centrality from the alternative paths availability and average path length point of view. Figure 5(a) describes the changes of the average path length followed by the removals of top 10 percentile high centrality score nodes. Increment of the average path length by a node removal means that there are some node isolations from the other part of the network or there are some alternative paths but longer than the removed path. The changes of the average path lengths should be also scrutinized along with the changes of the number of singletons, Figure 5(d), to comprehend more precisely. The changes of the average path length for betweenness are increased more than the case of bridging centrality in most of the interval. But it is clear that its increment behaviors are caused by the mass-production of singletons in the same interval as can be seen in Figure 5(d) since the nodes distinguished by betweenness mostly are located on the center of modules that have many peripheral nodes with one degree. Therefore, interrupting the nodes caught by betweenness caused many single node isolations and turned out to be the larger increment of the average path length. On the other hand, the average path length of the interruptions on the bridging nodes discriminated by bridging centrality are also increased significantly with generating only one singleton in the same intervals. Furthermore, the average path length for bridging centrality is increased more than betweenness in top 9 node removals which is most noteworthy. This behavior indicates that interruptions on the bridging nodes resulted in much longer alternative paths or isolations of larger modules.

Figure 5(b) and (c) compares the behaviors of the clustering coefficient of the network and the average isolated module size in the consequence of consecutive removals of top 10 percentile high centrality score nodes for betweenness and bridging centrality. The clustering coefficient and the isolated module size behaviors for these two centralities explain some interesting and important features of the nodes identified by these two different centrality measures. For better understanding of the these behaviors, one needs to observe the behaviors together with the changes of number of singletons. Figure 5(d) shows the changes of the number of singletons produced by the same node removals. The removals by betweenness, did not show monotonic behaviors of the clustering coefficients, and they rather considerably decreased the clustering coefficient about 20%. The average isolated module size also dropped rapidly in the same interval. Furthermore, betweenness produced many more singletons than bridging centrality did in the same intervals. The nodes identified by betweenness are located in the center of modules and the removal of those nodes damaged the modularity of the network and mass-produced singletons, i.e., smaller isolated modules, lower clustering coefficient and more singletons. However, as we removed the highest bridging centrality score nodes one by one, the clustering coefficient of the network is increased about 10% constantly for almost all intervals and fewer singletons are produced. In other words, cutting the high bridging centrality nodes enhanced the modularity of the network without producing many singletons, i.e., the nodes picked up by bridging centrality are located between modules neither on the center of modules nor on the periphery of the network.

### 3.2.6 *Assessing Ability To Occupy Modulating Position*

To demonstrate the unique positioning of bridging nodes at network regions modulating different functional modules, we investigated the functional and topological correlation on the direct neighbors of the nodes identified by bridging and betweenness centrality in Figure 6. The gene expression correlation between proteins was measured for the yeast PPI network [11] using Pearson correlation on Spellman cell cycle data [26]. The results show that the gene expression correlation among the direct neighbors of high bridging centrality nodes is lower than betweenness in top 10 percentile (Figure 6(a)). In addition to the lower biological correlation on the direct neighbors than betweenness, high bridging nodes also have lower topological correlation on the direct neighbors with lower clustering coefficient in top 10 percentile, i.e., the direct neighbors of high bridging nodes are not closely connected each other (Figure 6(b)). These findings support the premise that bridging nodes are positioned between different functional modules and the direct

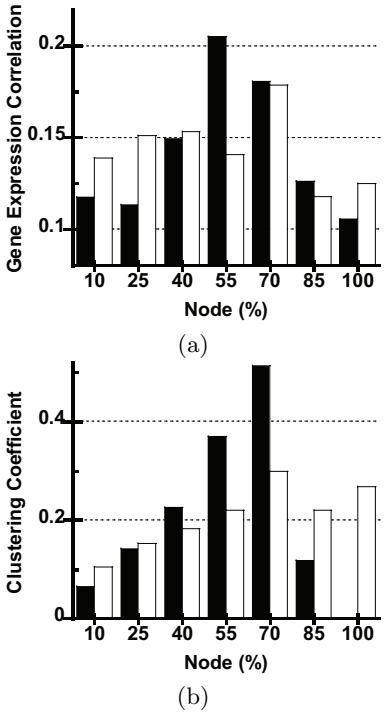



**Figure 6: Figure 6 shows the biological and the topological characteristics of the direct neighbors of the node ordered by two metrics, the bridging centrality (black bar), betweenness centrality (white bar). Figure 6(a) shows the gene expression correlation on the direct neighbors of each percentile. Figure 6(b) shows the average clustering coefficient of the nodes in each percentile.**

neighbors are poorly connected each other whereas the nodes identified by betweenness are located within functional modules that have correlated gene expression patterns and their direct neighbors are well connected each other.

## 3.3 Group Level: Bridge Cut Algorithm

To demonstrate the strengths of bridge cut algorithm, clustering was performed on two different kinds of networks, one biological network and one social network, in this section. We compared our approach to the following six competing clustering approaches: Maximal clique [27], Quasi clique [9], Minimum cut [17], the statistical approach of Samanta and Liang [25], MCL [29], and Rives [24].

### 3.3.1 Performance on a Biological Network

The experimental results on the DIP PPI dataset [11] are presented in Table 1. DIP PPI dataset has 2339 nodes and 5595 edges between them. MIPS complex category data is used as reference modules to measure the clustering quality because a group of proteins has a high probability that forms a protein complex if they have physical interactions among them. By the way, detecting effective modules in a sparse network like DIP PPI network, which has a very low density (0.002045), is challenging because most of the graph clustering methods try to find densely connected regions. Despite its sparse connectivity, bridge cut algorithm detected more effective modules showing higher $F$-measure, 0.53, for MIPS complex category and lower DB index, 4.78, than the other existing approaches. Maximal clique, MCL, and quasi clique methods scores comparable $F$-measures, 0.49, 0.47, and 0.46, respectively. However, maximal clique and quasi clique methods produced many small size clusters highly overlapping each other. They used only 2.7% and 19.2% of the nodes in the dataset discarding huge portion of the dataset. Therefore, maximal clique and quasi clique methods identified many small sized overlapping clusters that are enriched by a small number of same MIPS complex modules. In other words, maximal clique and quasi clique methods have a bad discrimination ability among detected clusters. DB index values for maximal clique and quasi clique are not available due to their discrimination ability deficiency in Tables 1 and 3. MCL method shows a comparable $F$-measure but it shows worse DB index than bridge cut method. The identified clusters by MCL have worse cluster quality from biological and topological viewpoint, less compact and worse separability, than bridge cut. Bridge cut method detected more effective clusters that have better biological enrichment and better compactness and separability topologically.

| Methods | Clusters | Size | MIPS complex (F-measure) | DB |
|---|---|---|---|---|
| **Bridge Cut** | **114** | **7.6** | **0.53** | **4.78** |
| Max Cliq | 120 | 4.7 | 0.49 | N/A |
| Quasi Cliq | 103 | 9.2 | 0.46 | N/A |
| Rives | 74 | 31 | 0.33 | 13.5 |
| Mincut | 227 | 8.7 | 0.35 | 7.23 |
| MCL | 210 | 8.4 | 0.47 | 6.82 |
| Samanta | 138 | 7.2 | 0.43 | 6.8 |

**Table 1: Comparative analysis. Performance of bridge cut method on DIP PPI dataset is compared with six graph clustering approaches (Maximal clique, quasi clique, Rives, minimum cut, Markov clustering, Samanta). The second column indicates the number of clusters detected. The third column shows the average size of each cluster. The fourth column represents the average $F$-measure of the clusters for MIPS complex modules. The average $F$-measure value of detected modules was calculated by mapping each module to the MIPS complex module with the highest $F$-measure value. The fifth column indicates the Davies-Bouldin cluster quality index. Comparisons are performed on the clusters with 4 or more components.**

Figure 7 plots the $F$-measure values and the percentile of matching proteins with the best mapping MIPS complex module for the top 30 highest $F$-measure valued clusters identified by bridge cut. The average $F$-measure value is 0.794 and the average hitting percentile with the best matching MIPS complex module is 75.8%. Table 2 lists the top 10 best $F$-measure valued clusters and their corresponding size, $F$-measure values, hitting percentile onto the best matching MIPS complex module, and the name of the best matching MIPS complex module. Bridge cut algorithm identified effective modules with high enrichment and high accordant percentile onto diverse MIPS complex modules.

### 3.3.2 Performance on a Social Network

To demonstrate the effectiveness of bridge cut algorithm on the other types of network, clustering analysis was performed in a social network dataset, a school friendship net-

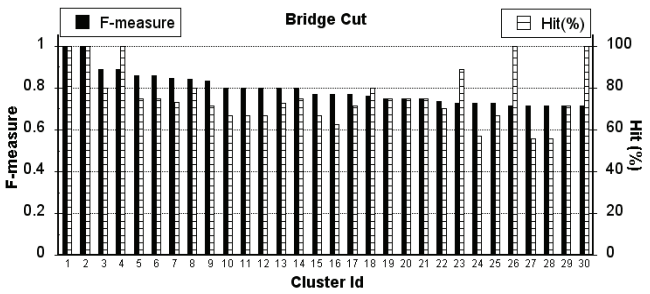



**Figure 7: Top 30 best clusters identified by bridge cut. The $F$-measure values and the percentile of matching proteins with the best mapping MIPS complex module for the top 30 highest $F$-measure valued clusters are illustrated.**

| ID | Size | $F$ | Hit(%) | MIPS complex |
|---|---|---|---|---|
| 1 | 4 | 1.0 | 100 | AP-3 complex |
| 2 | 4 | 1.0 | 100 | CCAAT-binding factor complex |
| 3 | 5 | 0.89 | 80 | AP-1 complex |
| 4 | 4 | 0.89 | 100 | Gim complexes |
| 5 | 8 | 0.86 | 75 | Replication complexes |
| 6 | 4 | 0.86 | 75 | Complex Number 482 |
| 7 | 15 | 0.85 | 73 | Anaphase promoting complex |
| 8 | 20 | 0.84 | 80 | 20S proteasome |
| 9 | 7 | 0.83 | 71 | Tim22p-complex |
| 10 | 6 | 0.8 | 80 | Class C Vps protein complex |

**Table 2: Top 10 best $F$-measure valued clusters identified by bridge cut. The first and the second column indicate the cluster id and the size, respectively. The third and the fourth column represent the $F$-measure and hitting percentile onto the best matching MIPS complex module, respectively. The fifth column shows the best matching MIPS complex module.**

work [19]. This school friendship network has 551 nodes and 2066 edges. The experimental results are shown in Table 3. Bridge cut algorithm produced more effective clusters than the other six approaches from topological viewpoint. There is no ground truth that can be compared to measure the cluster quality in this school friendship network and we conducted a cluster validation using internal information only, e.g., DB. Again, maximal clique and quasi clique show the same behavior as they did in the DIP PPI dataset case. They produced large number of small sized clusters that were highly overlapping with each other and discarded a substantial portion of the dataset. Maximal clique and quasi clique used only 8.8% and 37.1% of the nodes in the dataset, respectively. Overall performance of the other approaches was improved and MCL showed a comparable performance to bridge cut algorithm. The density of the school friendship dataset, 0.01360, is almost 7 times denser than DIP PPI dataset, 0.002045. This higher network density improved the performance of the other approaches because most focus on the densely connected areas. MCL did not show a good performance in a sparse dataset, i.e., DIP PPI dataset. However, bridge cut algorithm identified the clusters with better compactness and separability showing lower DB values topologically than the other approaches in a sparse dataset and in a dense dataset.

| Methods | Clusters | Size | DB |
|---|---|---|---|
| **Bridge Cut** | **40** | **8.6** | **5.46** |
| Max Cliq | 133 | 4.4 | N/A |
| Quasi Cliq | 109 | 9.5 | N/A |
| Rives | 46 | 10.9 | 10.4 |
| Mincut | 53 | 9.3 | 6.29 |
| MCL | 50 | 8.0 | 5.47 |
| Samanta | 40 | 13.5 | 7.1 |

**Table 3: Comparative analysis. Performance of bridge cut method on the school friendship dataset is compared with six graph clustering approaches (Maximal clique, quasi clique, Rives, minimum cut, Markov clustering, Samanta). Column descriptions are the same as Table 1**

# 4. COMPUTATIONAL COMPLEXITY

The computational complexity of the bridging centrality depends on the computational complexity of its two components, computing betweenness centrality and bridging coefficient. Betweenness centrality computation time for a network is bounded by $O(nm)$ and $O(nm + n^2 \log n)$ on unweighted and weighted networks, respectively, where $n$ is the number of nodes and $m$ is the number of edges. $\Theta(n^2)$ in space is required for betweenness centrality calculation for both types of networks [28]. The average computation time for bridging coefficient requires $O(n(\log n)^2)$ because the average degree of nodes is $\simeq \log n$ in real world networks, e.g., scale-free networks. Thus, the total time and space complexity of bridging centrality is bounded by the complexity of betweenness centrality computation, $O(nm)$ and $O(nm+n^2 \log n)$ time on unweighted and weighted networks, respectively, and $\Theta(n^2)$ space.

# 5. CONCLUSIONS

In this paper, we have identified a unique network feature, 'bridging', which has not been characterized in earlier research on networks. The bridging centrality is unique among the many available network metrics because it derives its effectiveness and high selectivity by combining both local and global network properties. Nodes with high bridging centrality occupy critical sites with the networks and connect larger or more densely connected modules to each other. The removal of nodes with high values of bridging centrality selectively causes high levels of network disruption maintaining high modularity. Targeted deletion of bridging nodes caused large changes to the path length distribution that are comparable to betweenness, which indicates that bridging nodes are potent at causing disruption. However, the deletion of bridging nodes generated fewer singletons and the average size of the isolated modules is larger than betweenness. We also proved that bridging centrality has a good ability to take modulating positions between functional modules by showing low correlations in the direct neighbors from the biological and topological viewpoints, low gene expression correlation and clustering coefficient, in yeast PPI network.

A novel graph clustering algorithm called 'bridge cut' is also proposed and evaluated in two different kinds of networks, one biological and one social network examples. Bridge cut algorithm recognizes bridging edges as module boundary in a network. Clustering performance of the bridge cut method is analyzed and compared with the other six graph clustering methods. Bridge cut method produced

more effective modules from a domain knowledge viewpoint and a topological viewpoint, i.e., better $F$-measure values and Davies-Bouldin idex. Furthermore, bridge cut approach showed a good performance in a sparse dataset as well as in a dense dataset.

In conclusion, bridging centrality has many potential applications in a diverse range of research areas as it enables a completely new way of analyzing network structures. In computer networks, it could be used to maximize network performance in the face of computer viruses and hackers. In biological networks, it could be used as an effective method for clustering protein-protein interactions, for identifying cognate functional modules of uncharacterized proteins and for identifying potential drug targets that maximize effectiveness while minimizing side effects. The ability of bridging centrality to identify key individuals and subcommunity structures in social networks could be useful for disease spreading prevention analysis in epidemic networks and law enforcement in national security applications.

## 6. ACKNOWLEDGEMENTS

This work was partly supported by NSF grant DBI-0234895 and NIH grant 1 P20 GM067650-01A1.

## 7. REFERENCES

[1] North, S. in Symposium on Graph Drawing GD'96 . *Springer, Berkely, CA*, 409:, 1996.

[2] Albert, R., Jeong, H. and Barabasi, A. L. Error and attack tolerance of complex networks. *Nature*, 406:378–82, 2000.

[3] Bader, G.D. and Hogue. C.W. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4:2, 2003.

[4] Barabasi, A.L. and Oltvai, Z.N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, 5:101–113, 2004.

[5] Batagelj, V., and Mrvar, A. Pajek . Program for Large Network Analysis. *Connections*, :, 1998.

[6] Brandes, U. and Fleischer, D. Centrality measures based on current flow . In *Proceedings of the 22nd International Symposium on Theoretical Aspects of Computer Science (STACSŠ05)*, volume , page , 2005.

[7] Breitling R, Armengaud P, Amtmann A, Herzyk P. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Letter*, 573:83–92, 2004.

[8] Brin, S. and Page, L. The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of the seventh international conference on World Wide Web*, volume , pages 107–117, 1998.

[9] Bu, D. et al. Topological structure analysis of the protein-protein interaction network in budding yeast. *Nucleic Acids Res.*, 31:2443–2450, 2003.

[10] Davies, D., Bouldin, D. . A cluster separation measure. *IEEE Trans. Pattern Reconit. Machine Intell.*, 1(2):224–227, 1979.

[11] Deane, C.M., Salwinski, L., Xenarios, I. and Eisenberg, D. . Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell Proteomics*, 1:349–356, 2002.

[12] Estrada, E. and Rodríguez-Velázquez, J. Subgraph centrality in complex networks. . *Phys. Rev. E*, 71:056103–1–9, 2005.

[13] Freeman, L.C. A set of measures of centrality based upon betweeness. . *Sociometry*, 40:35Ű41, 1977.

[14] Guimera, R. and Nunes Amaral, L. A. An automated method for finding molecular complexes in large protein interaction networks. *Nature*, 433:895–900, 2005.

[15] Hartuv, E. and Shamir, R. A clustering algorithm based on graph connectivity. *Info. Processing Letters*, 76:175–181, 2000.

[16] Hwang, W. et al. A novel functional module detection algorithm for protein-protein interaction networks. *Algorithms Mol Biol.*, 1:24, 2006.

[17] Johnson, D.B. Efficient algorithms for shortest paths in sparse networks. *J. of the ACM*, 24:1–13, 1977.

[18] King, A.D., Przulj, N. and Jurisica, I. Protein complex prediction via cost-based clustering. *Bioinformatics*, 20:3013–3020, 2004.

[19] Lind, P., Silva, L., Andrade, J., Hermann, H. . Spreading gossip in social networks. *Phys. Rev. E.*, 76:036117, 2007.

[20] Mewes, H. W. . MIPS: analysis and annotation of proteins from whole genome in 2005. *Nucleic Acid Research*, 34:D169–D172, 2006.

[21] Newman, M. E. and Girvan, M. Finding and evaluating community structure in networks. . *Phys Rev E Stat Nonlin Soft Matter Phys*, 69:026113, 2004.

[22] Newman, M.E.J. A measure of betweenness centrality based on random walks. . *arxiv preprint*, 1(3):215Ű239, 2003.

[23] Park, J. and Newman, M. E. Origin of degree correlations in the Internet and other networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 68:026112, 2003.

[24] Rives, A.W. and Galitski, T. Modular organization of cellular networks. *Proc. Natl. Acad. Sci.*, 100(3):1128–33, 2003.

[25] Samanta, M.P. and Liang,S. Redundancies in large-scale protein interaction networks. *Proc. Natl Acad. Sci.*, 100:12579–12583, 2003.

[26] Spellman, P. et al. Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. *Molecular Biology of the Cell*, 9:3273–3279, 1998.

[27] Spirin, V. and Mirny, L.A. . Protein complexes and functional modules in molecular networks. *PNAS*, 100:12123–12128, 2003.

[28] Ulrik Brandes. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2):163–177, 2001.

[29] van Dongen, S. . Technical Report INS-R0010: A cluster algorithm for graphs. *National Research Institute for Mathematics and Computer Science*, 2000.

[30] Watts, D. J. and Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* , 393:440–2, 1998.