

Novel Approaches for Analyzing Biological Networks

Balabhaskar Balasundaram*[‡]; Sergiy Butenko[†]
and Svyatoslav Trukhanov[‡]

Department of Industrial Engineering, Texas A&M University, College Station, Texas 77843, USA

Abstract

This paper proposes clique relaxations to identify clusters in biological networks. In particular, the maximum n -clique and maximum n -club problems on an arbitrary graph are introduced and their recognition versions are shown to be NP -complete. In addition, integer programming formulations are proposed and the results of sample numerical experiments performed on biological networks are reported.

keywords: n -cliques, n -clubs, clique relaxations, social networks, biological networks

1 Introduction

In this post-genomic era, several biological interactions are well captured by networks such as “protein-protein interaction networks” and “gene co-expression networks”. Study of such biological networks and other complex networks such as the internet and the world wide web have received special attention from scientists because of their interesting properties and the information they hold. In this respect, the concept of *scale-free networks* [3, 7] is a recent development. It has been observed that the degree distribution of a large number of such complex networks follow a power law. As a consequence, average degree is no longer representative and a majority of the nodes have few neighbors while a smaller number of nodes have very high degrees. The principle of preferential attachment which suggests that the new nodes have a higher probability to link to nodes that already have a high degree, is used to explain the power-law degree distribution of such scale-free graphs. In addition these networks are also

*E-mail: baski@tamu.edu

†Corresponding author. E-mail: butenko@tamu.edu

‡E-mail: slavik@tamu.edu

hierarchical in the sense that they can be partitioned into a collection of functional modules. Analysis of several biological networks provides strong evidence that biological networks are both scale-free and modular. Identifying large clusters or functional modules in biological networks can aid different objectives depending on the nature of these networks. Clique models have been most popular in this area as they represent “tight clusters” in a network. Cliques have been used to cluster *gene co-expression networks*, where vertices are genes and an edge exists between two vertices if the corresponding genes are co-expressed with correlation higher than a specified threshold [27, 22]. Cliques and high density subgraphs have also been used to cluster *protein interaction networks* in [29, 17]. A protein interaction network is represented by a graph with the proteins as vertices and an edge exists between two vertices if the proteins are known to interact. However, clique models could be overly restrictive in describing clusters in such networks. Graph theoretic clique relaxations that are used in social network analysis for identifying *cohesive subgroups* can provide interesting insights into these networks and provide more information than what is revealed by cliques. Relaxing the restrictions imposed by clique models could reveal new protein interactions. In particular, structures where interactions of proteins occur through a central protein, which are likely to be found in similar biological processes, can be identified [6] by the models suggested in this paper.

Besides biological networks, cohesive subgroups can be used to cluster airline networks where reachability is a critical issue. An important classical application of cohesive subgroups is the study of terrorist and other criminal networks [12, 5, 15]. More recently these models have been used to study web graphs in internet research [30] to facilitate organization and faster retrieval of information from the web. These approaches have also been used in clustering wireless networks [23] and for other graph based data mining applications [13, 31, 16].

In this paper, these models are applied to biological networks to identify large clusters that are distance based relaxations of cliques. Before presenting these clique relaxation models, the notation used and other required definitions will be presented.

Consider a simple undirected graph $G = (V, E)$ with vertex set V and edge set E . Let $d_G(u, v)$ denote the length of a shortest path (in number of edges) between vertices u and v in G and $diam(G) = \max_{u, v \in V} d_G(u, v)$ be the diameter of G . For a subset of vertices $S \subseteq V$, $G(S)$ denotes the subgraph induced by S on G , $G(S) = (S, S \times S \cap E)$.

A *clique* C is a subset of V such that the subgraph $G(C)$ induced by C on G is complete. A clique is *maximal* if it is not a subset of a larger clique, and *maximum* if there is no larger clique in the graph. The maximum clique problem is to find a clique of maximum cardinality. The *clique number* $\omega(G)$ is the cardinality of a maximum clique in G . The maximum clique problem has been extensively studied in terms of its complexity, mathematical programming formulations, exact, approximate and heuristic algorithms, and numerous applications. See [8] for a survey of the maximum clique problem and related references. Among other applications, cliques are often used to represent *clus-*

ters of similar elements. For example, in social networks, a clique represents a group of people such that any two of them have a certain kind of relationship (friendship, acquaintance, etc.) with each other [26]. In fact, some of the earliest works addressing the concept of cliques and methods of their detection were motivated by applications in sociometry [25, 24, 20].

The clustering problems studied in this paper deal with “relaxations” of the idea of a clique, in which, for any two vertices, the requirement of their connectedness is replaced with a less tight condition on the distance between them. We first state the corresponding definitions of n -clique, n -clan and n -club as they originally appeared in the literature. Following which, we will point out some drawbacks in these definitions and modify them according to standard definitions of similar concepts in graph theory. It is not surprising that the clustering concepts of interest first appeared in studying cohesive subgroups in social networks, where the vertices correspond to “actors” in a social network and an edge indicates a relationship between two actors [32].

Luce [24] defines an n -clique of G as a subset of vertices $C \subseteq V$ such that for all $u, v \in C : d_G(u, v) \leq n$ and this subset is maximal by inclusion. In other words, an n -clique C is a set of vertices in which any two vertices are a distance of at most n from each other in G , and no other vertex in the graph is of distance n or less from every other vertex in C . Thus, if two vertices $u, v \in V$ belong to an n -clique C , then $d_G(u, v) \leq n$, however this does not imply that $d_{G(C)}(u, v) \leq n$. For example, Figure 1¹ shows a graph in which the subset of vertices $C_1 = \{1, 2, 3, 4, 5\}$ forms a 2-clique, however the distance between vertices 1 and 5 in the subgraph induced by C_1 is 3. Hence, the concept of n -clique lacks the requirement of “tightness” in the group corresponding to vertices of an n -clique, while such a requirement is essential to applications in social networks. This observation motivated Alba [2] to introduce the concept of a “sociometric clique”, which was later renamed to “ n -clan” by Mokken [26]. An n -clique C is called an n -clan if the diameter of the induced subgraph $G(C)$ is no more than n . Finally, Mokken [26] defines an n -club to be a maximal (by inclusion) subset of vertices, $D \subseteq V$ such that the diameter of the induced subgraph $G(D)$ is at most n . To highlight the differences between the three structures, we turn to the graph in Figure 1. In this graph, the 2-cliques are given by $C_1 = \{1, 2, 3, 4, 5\}$ and $C_2 = \{1, 2, 4, 5, 6\}$. It is easy to see that C_1 is not a 2-clan or 2-club, since the diameter of induced subgraph $G(C_1)$ is 3. Since any n -clan is an n -clique, the only 2-clan in this graph is given by C_2 . Lastly, the 2-clubs of this graph are $D_1 = \{1, 2, 3, 4\}$, $D_2 = \{2, 3, 4, 5\}$ and $D_3 = C_2$. A study of relations between cliques, clans and clubs in a graph can be found in [26].

Even though the concepts just defined are used quite extensively in social networks analysis and are even covered in standard textbooks (see, *e.g.*, [32]), their definitions have some deficiencies from the mathematical viewpoint. One considerable drawback of the n -clan definition is that for some graphs an n -clan

¹A similar example first appeared in Alba [2] and was subsequently adopted by other authors.

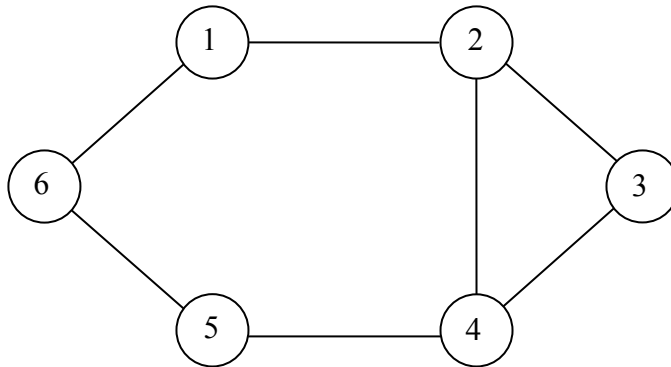


Figure 1: An example graph.

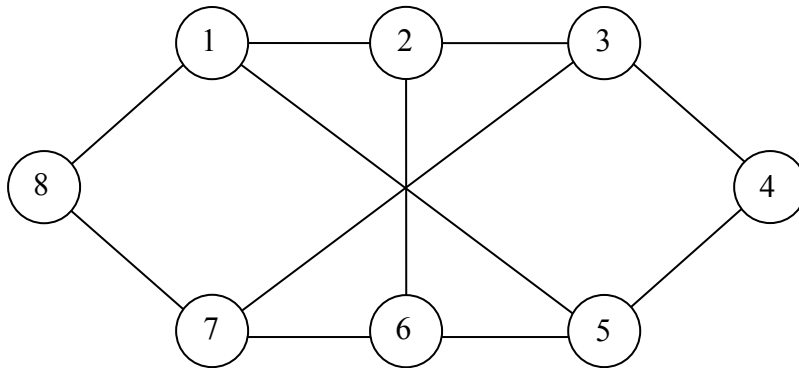


Figure 2: A graph with no 2-clans.

may not exist. This point is illustrated in Figure 2, which shows a graph with two 2-cliques $\{1, 2, 3, 4, 5, 6, 7\}$ and $\{1, 2, 3, 5, 6, 7, 8\}$, neither of which is a 2-clan.

Some other difficulties arise from the requirement of maximality (by inclusion) in all three definitions. In particular, this requirement makes checking whether a given subset of vertices is an n -club a nontrivial matter. Indeed, to check that C is an n -club, it suffices to show that there is no vertex outside C that could be added to C without violating the requirement that all pairwise distances between vertices do not exceed n . A similar criterion would not work for n -clan, however, since in this case the maximality by inclusion is not equivalent to nonexistence of one vertex that could increase the size of the n -club [26]. As an example, consider the graph in Figure 1. For the subset of vertices $C = \{1, 5, 6\}$ of this graph, the subgraph induced by C has diameter 2. Appending one of the vertices 2 or 4 to C would increase the diameter of the induced subgraph, however if both vertices are added, the diameter of the resulting induced subgraph is still 2.

Taking into account that the above definitions of 1-clique, 1-clan and 1-club

all correspond to the standard definition of a *maximal* clique, we propose to modify the definitions of n -clique and n -club accordingly. From now on, by an n -clique of graph $G = (V, E)$ we will mean a subset of vertices C , such that for any $u, v \in C$: $d_G(u, v) \leq n$. Similarly, by an n -club we will understand a subset of vertices D such that $\text{diam}(G(D)) \leq n$. A similar definition of n -clan becomes redundant. The example in Figure 2 suggests the impracticality of such a concept, so we do not consider n -clans in the further discussion. By a maximal n -clique (n -club) we will mean an n -clique (n -club) that is not a subset of a larger n -clique (n -club). Finally, a maximum n -clique (n -club) is an n -clique (n -club) of the largest size in the graph. Given a graph $G = (V, E)$ and a positive integer n , the maximum n -clique problem is naturally defined as the problem of finding a largest n -clique in G . We denote the n -clique number of graph G , which is the cardinality of a maximum n -clique of G , by $\omega_n(G)$. The maximum n -club problem is defined likewise, with $\bar{\omega}_n(G)$ denoting the n -club number of G .

Even though the concept of cohesive subgroups is borrowed from social network analysis, these ideas are applicable to any network, and finding these cohesive subgroups can reveal several important structural aspects of the networks. Despite a number of important practical applications, the combinatorial optimization problems concerned with finding large n -cliques and n -clubs have not been well studied analytically or computationally. In fact, little has been known about the complexity of such problems and mathematical programming approaches have not been developed. This paper addresses some of these issues.

In Section 2 we show that the recognition versions of the maximum n -clique and maximum n -club problems are NP -complete for any fixed n . Section 3 provides mathematical programming formulations of these problems. Section 4 studies the 2-club problem, including its polyhedral properties and the results of sample numerical experiments are presented in Section 5. Finally, Section 6 concludes the paper.

2 Computational Complexity

Before stating the complexity results, we introduce the recognition version of each problem. The n -CLIQUE (n -CLUB) problem is defined as follows: Given a graph $G = (V, E)$ and positive integers n and k , does there exist an n -clique (n -club) of size $\geq k$ in G ?

Theorem 1 *The n -CLIQUE and n -CLUB problems are NP -complete for any positive integer n .*

Proof: To prove NP -completeness of a problem \mathcal{P} , it suffices to show that [18]

1. $\mathcal{P} \in NP$;
2. Some known NP -complete problem is polynomially reducible to \mathcal{P} .

Note that for $n = 1$ both problems coincide with CLIQUE problem, which is a well-known NP -complete problem. So, we consider $n > 1$. Given a “yes” instance of n -CLIQUE (n -CLUB), any n -clique (n -club) of size $\geq k$ can be used as a certificate to verify that this is indeed a “yes” instance in polynomial time. Thus, n -CLIQUE and n -CLUB are in NP . To complete the proof, we reduce CLIQUE, which is a well known NP -complete problem, to n -CLIQUE (n -CLUB). Let $G = (V, E)$ be an instance of CLIQUE which we assume does not contain isolated vertices, as no isolated vertex can be included in any clique of size two or more. We construct a corresponding instance of n -CLIQUE (n -CLUB) which is a $(\lfloor n/2 \rfloor + 2)$ -partite graph $G' = (V', E')$. We define the vertex set as a union of $\lfloor n/2 \rfloor$ copies of V , a copy of E and one more auxiliary vertex 0:

$$V' = \cup_{i=1}^{\lfloor n/2 \rfloor} V^{(i)} \cup E \cup \{0\},$$

where $V^{(i)} = \{1^{(i)}, 2^{(i)}, \dots, N^{(i)}\}$ is the i -th copy of V , $i = 1, \dots, \lfloor n/2 \rfloor$. For any $v \in V$, by $v^{(i)} \in V^{(i)}$ we denote the i -th copy of v . The edge set connects copies of the same vertex in $V^{(i)}$ and $V^{(i+1)}$, $i = 1, \dots, \lfloor n/2 \rfloor - 1$. A vertex $v^{\lfloor n/2 \rfloor}$ in $V^{\lfloor n/2 \rfloor}$ is connected to a vertex $e \in E$ if v is an endpoint of e in G . Finally, all vertices from E in G' are connected to 0. To summarize,

$$\begin{aligned} E' = & \cup_{i=1}^{\lfloor n/2 \rfloor - 1} \{(v^{(i)}, v^{(i+1)}) : v \in V\} \\ & \cup \{(v^{\lfloor n/2 \rfloor}, e) : v \in V, v \text{ is an endpoint of } e \text{ in } G\} \\ & \cup \{(e, 0) : e \in E\}. \end{aligned}$$

Figure 3 shows graph G' corresponding to graph G from Figure 1 for $n = 5$. Graph G' contains $\lfloor n/2 \rfloor |V| + |E| + 1$ vertices and can obviously be constructed in time polynomial with respect to the size of G .

Our reduction is based on the observation that G has a clique of size k if and only if G' has an n -clique (n -club) of size $k + (\lfloor n/2 \rfloor - 1)|V| + |E| + 1$. Note that G' is connected and $\text{diam}(G) \leq 2(\lfloor n/2 \rfloor) + 2$. Indeed, in G' , all vertices of $V' \setminus V^{(1)}$ can be included in any n -clique (n -club). Two vertices $u^{(1)}, v^{(1)} \in V^{(1)}$ belong to the same n -clique (n -club) in G' if and only if $(u, v) \in E$ in G . Thus, n -CLIQUE and n -CLUB are NP -complete problems for any positive integer n . \square

It is known that many massive networks arising in various applications have a relatively small diameter. This observation is commonly referred to as the *small world phenomenon* [4, 33, 34]. Therefore, the clustering problems on graphs of small diameter are of particular interest. This motivates us to consider the n -CLIQUE and n -CLUB problems on graphs of fixed diameter. Note that if $\text{diam}(G) \leq n$ then both the maximum n -clique problem and the maximum n -club problem are trivial as G is the maximum n -clique (n -club), therefore we are only interested in the case where $\text{diam}(G) > n$. For any $d > n$, we define the n -CLIQUE(d) (n -CLUB(d)) problem as follows: Given a graph G of diameter d and positive integers n and k , does there exist an n -clique (n -club) of size $\geq k$ in G ?

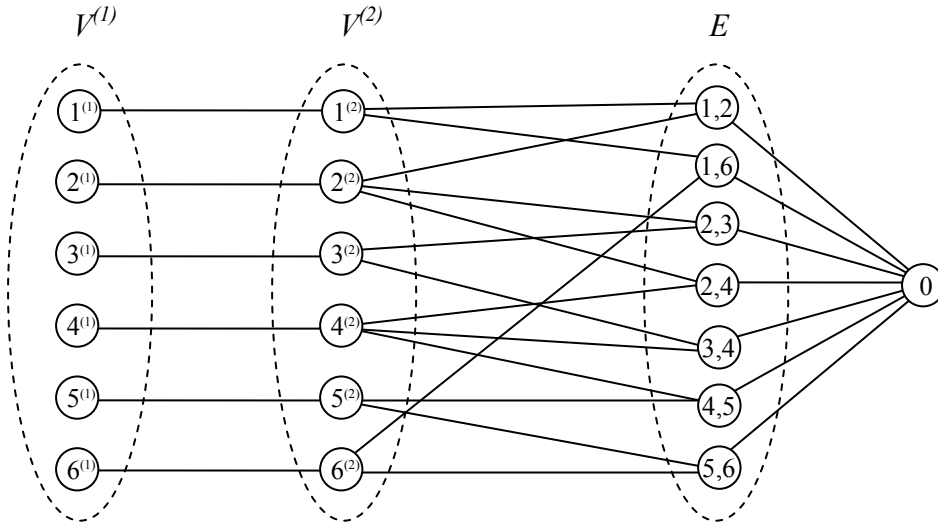


Figure 3: An illustration to the proof of NP -completeness for $n = 5$.

Theorem 2 For any fixed positive integer n and $d > n$, the n -CLIQUE(d) and n -CLUB(d) problems are NP -complete.

Proof: Obviously both considered problems are in NP . To complete the proof we reduce CLIQUE to n -CLIQUE(d) and n -CLUB(d). We first prove the statement for $n = 1$. Given $G = (V, E)$ with no isolated vertices and $d > 1$, we construct a graph $\hat{G} = (\hat{V}, \hat{E})$ of diameter d as follows.

$$\begin{aligned}\hat{V} &= V \cup \{u_i : i = 1, \dots, d\}; \\ \hat{E} &= E \cup \{(v, u_1) : v \in V\} \cup \{(u_i, u_{i+1}) : i = 1, \dots, d-1\}.\end{aligned}$$

Then G has a clique of size k if and only if \hat{G} has a clique of size $k + 1$ and the proof is complete for $n = 1$.

If $n > 1$, we consider two cases, for odd and even n . If n is odd, then we use the same construction of graph G' as in the proof of Theorem 1 to reduce CLIQUE to n -CLIQUE(d) and n -CLUB(d). This is true since $diam(G') \leq n + 1 \leq d$ when n is odd and G has a clique of size k if and only if G' has an n -clique (n -club) of size $k + (\frac{n-1}{2} - 1)|V| + |E| + 1$. If n is even, a similar construction can be used (see Figure 4) to prove the reduction. As before, we use $n/2$ copies of V and a copy of E for the vertex set of the constructed graph $G'' = (V'', E'')$.

$$V'' = \bigcup_{i=1}^{n/2} V^{(i)} \cup E.$$

The edge set E'' is also similar to E' in the previous construction, but instead of connecting vertices from the copy of E to an auxiliary vertex, we make the

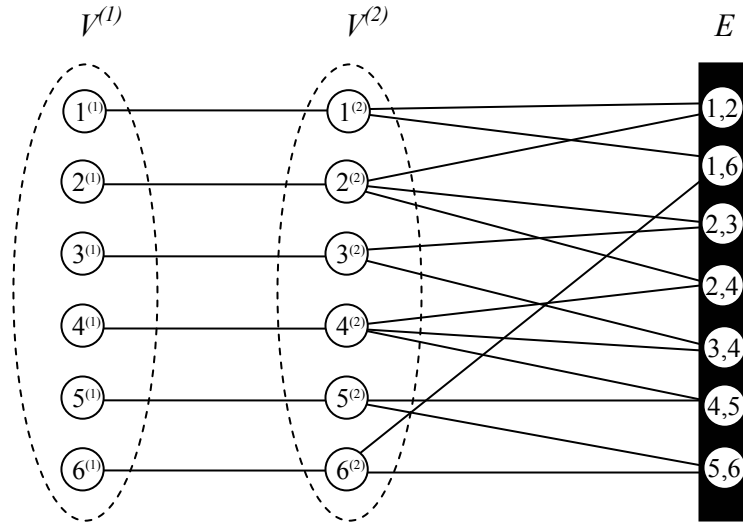


Figure 4: An illustration to the proof of Theorem 2 for $n = 4$.

subset of vertices corresponding to E a clique.

$$\begin{aligned}
 E'' = & \cup_{i=1}^{n/2-1} \{(v^{(i)}, v^{(i+1)}) : v \in V\} \\
 & \cup \{(v^{(n/2)}, e) : v \in V, v \text{ is an endpoint of } e \text{ in } G\} \\
 & \cup \{(e_1, e_2) : e_1, e_2 \in E, e_1 \neq e_2\}.
 \end{aligned}$$

Once again, $\text{diam}(G'') \leq n + 1 \leq d$ and G has a clique of size k if and only if G'' has an n -clique (n -club) of size $k + (n/2 - 1)|V| + |E|$. This completes the proof of NP -completeness on fixed diameter graphs. \square

These complexity results illustrate two important facts. Firstly, these generalizations are hard to solve not only because they generalize cliques, but because they are hard in their own respect (NP -complete for any n). Secondly, the transition in complexity is also sudden, while the problems are easily solved under trivial circumstances when diameter is bounded above by n , but immediately become NP -complete whenever diameter of the graph is strictly larger than n .

3 Integer Programming Formulations

This section presents integer programming (IP) formulations for the maximum n -clique and maximum n -club problems.

3.1 Maximum n -clique problem

There are a number of known mathematical programming formulations of the maximum clique problem [8], including integer programming formulations and

nonlinear programming approaches [1, 10]. Similar formulations can be applied to the maximum n -clique problem on a graph $G = (V, E)$ by reducing it to the maximum clique problem on the n^{th} power of G , $G^n = (V, E^n)$, where $E^n = \{(i, j) : i, j \in V, i < j, d_G(i, j) \leq n\}$. G^n is constructed from the original graph by adding edges corresponding to all pairs of vertices with distance no more than n between them in G . Consider the following formulation for n -clique.

$$\omega_n(G) = \max \sum_{i \in V} x_i \quad (1)$$

subject to:

$$\begin{aligned} x_i + x_j &\leq 1 + \frac{n}{d_G(i, j)} \quad \forall i, j \in V : i < j \\ x_i &\in \{0, 1\} \quad \forall i \in V \end{aligned}$$

The constraint ensures that two vertices with $d_G(i, j) > n$ are not simultaneously included in a n -clique, but becomes redundant for pairs of vertices with $d_G(i, j) \leq n$. Since for all pairs of vertices, the shortest path distance is known, the constraint in the system can be replaced by

$$x_i + x_j \leq 1 \quad \forall (i, j) \in \{(i, j) : i, j \in V, i < j, d_G(i, j) > n\} = \overline{E^n}.$$

The formulation then becomes a maximum clique formulation on G^n . Note that even though the existing heuristics and algorithms for maximum clique problem can be applied to the power of the graph to solve the maximum n -clique problem, their performance may be poorer as the edge density is higher in G^n .

3.2 Maximum n -club problem

The following integer program describes the n -club number.

$$\bar{\omega}_n(G) = \max \sum_{i \in V} x_i \quad (2)$$

subject to:

$$\begin{aligned} x_i + x_j &\leq 1 + \sum_{l: P_{ij}^l \in \mathbb{P}_{ij}} y_{ij}^l \quad \forall (i, j) \notin E \\ x_p &\geq y_{ij}^l \quad \forall p \in V(P_{ij}^l), P_{ij}^l \in \mathbb{P}_{ij}, (i, j) \notin E \\ x_i &\in \{0, 1\} \quad \forall i \in V \\ y_{ij}^l &\in \{0, 1\} \quad \forall P_{ij}^l \in \mathbb{P}_{ij}, (i, j) \notin E \end{aligned}$$

where \mathbb{P}_{ij} is an indexed collection of all paths of length at most equal to n between vertices i, j in G and P_{ij}^l is the path with index l between vertices i, j . The formulation essentially ensures that if two vertices are in a n -club, then all the vertices in at least one path between them with length less than or equal to n are also included in the n -club.

4 Maximum 2-Club Problem

An IP Formulation for the maximum 2-club problem can be induced from the formulation for the maximum n -club problem in the previous section. Thus, the 2-club number $\bar{\omega}_2(G)$ of a graph $G = (V, E)$ admits the following integer programming formulation:

$$\begin{aligned} \bar{\omega}_2(G) &= \max \sum_{i \in V} x_i & (3) \\ \text{s.t.} \quad & x_i + x_j - \sum_{k \in N^\cap(i, j)} x_k \leq 1, \text{ for all } (i, j) \notin E; \\ & x_i \in \{0, 1\}, \text{ for all } i \in V, \end{aligned}$$

where $N^\cap(i, j)$ denotes the common neighborhood of vertices i, j in G , i.e. $N^\cap(i, j) = N(i) \cap N(j)$. The formulation ensures that if two vertices are in a 2-club and they do not have an edge between them, then they have at least one common neighbor inside the 2-club. In the next subsection we study the 2-club polytope and establish some basic results.

4.1 The 2-club polytope

In this section, unless noted otherwise, we denote by $G = (V, E)$ a simple undirected connected graph with vertex set V such that $|V| \geq 2$ and edge set E and by $\bar{G} = (V, \bar{E})$, its complement. Let M_1 be the edge-vertex incidence matrix of \bar{G} . The rows of M_1 correspond to edges $e_{ij} \in \bar{E}$ and the columns correspond to vertices $i \in V$. The entries are all 0 except for (e_{ij}, i) and (e_{ij}, j) for every $e_{ij} \in \bar{E}$ which are 1. Let M_2 be the matrix denoting the common neighborhood of i, j for every $e_{ij} \in \bar{E}$. The rows of M_2 correspond to edges $e_{ij} \in \bar{E}$ and the columns correspond to vertices $i \in V$. All entries are 0 except for (e_{ij}, k) for all $k \in N^\cap(i, j)$ which are 1. Now, let $A = M_1 - M_2$. Then formulation (3) can be rewritten as:

$$\bar{\omega}_2(G) = \max\{\mathbf{1}^T x : Ax \leq \mathbf{1}, x \in \{0, 1\}^{|V|}\} \quad (4)$$

where $\mathbf{1}$ is vector of appropriate dimension with each component equal to 1. Define the set of feasible binary vectors Q as

$$Q = \{x \in \{0, 1\}^{|V|} : Ax \leq \mathbf{1}\} \quad (5)$$

Then the *2-Club Polytope* is given by

$$P_{2C} = \text{conv}(Q) \quad (6)$$

where $\text{conv}(Q)$ denotes the convex hull of Q .

Theorem 3 Let P_{2C} denote the 2-Club polytope of a given graph $G = (V, E)$.

1. $\dim(P_{2C}) = |V|$.
2. $x_i \geq 0$ induces a facet of P_{2C} for every $i \in V$.
3. For any arbitrary $i \in V$, $x_i \leq 1$ induces a facet of P_{2C} if and only if $d_G(i, j) \leq 2 \quad \forall j \in V$.

Proof: We will use following notations in the proof. Let e_i be the unit vector with i^{th} component 1 and the rest 0; $e_{ij} = e_i + e_j$ and $e_{ijk} = e_i + e_j + e_k$.

1. This is trivial and can shown by demonstrating $|V| + 1$ feasible affinely independent points in P_{2C} . The points $\mathbf{0}, e_1, e_2, \dots, e_{|V|}$ are clearly $|V| + 1$ affinely independent points in $P_{2C} \subset \mathbb{R}^{|V|}$. Hence $\dim(P_{2C}) = |V|$.

2. Let $F_i^0 = \{x \in P_{2C} : x_i = 0\}$. Then $\mathbf{0}, e_k$ for all $k \in V \setminus \{i\}$ form $|V|$ affinely independent points in F_i^0 indicating that $\dim(F_i^0) = |V| - 1$ and it is a facet.

3. For a fixed $i \in V$, suppose that $d_G(i, j) \leq 2 \quad \forall j \in V$. We wish to show that $F_i^1 = \{x \in P_{2C} : x_i = 1\}$ is a facet. Then let $S^p = \{j \in V : d_G(i, j) = p\}$. Note that S^0, S^1, S^2 , partition V and $S^0 = \{i\}, S^1 = N(i)$. We now establish the maximality of F_i^1 thereby making it a facet. Suppose there exists a valid inequality $\alpha^T x \leq \beta$ such that, $F = \{x \in P_{2C} : \alpha^T x = \beta\} \supseteq F_i^1$. Note that $e_i, e_{ij} \quad \forall j \in S^1$ are also contained in F_i^1 . Also for every $k \in S^2$, there exists a $j \in S^1$ such that $j \in N^\cap(i, k)$. Hence $e_{ijk} \quad \forall k \in S^2$ for some $j \in S^1$ are also in F_i^1 . These $|V|$ points are also contained in F . $e_i \in F \Rightarrow \alpha_i = \beta$ (by substituting for x in $\alpha^T x = \beta$). Similarly we get $\alpha_i + \alpha_j = \beta \quad \forall j \in S^1 \Rightarrow \alpha_j = 0 \quad \forall j \in S^1$. From the remaining points we obtain $\alpha_i + \alpha_j + \alpha_k = \beta \quad \forall k \in S^2$ and for some $j \in S^1$. Since $\alpha_i = \beta, \alpha_j = 0 \quad \forall j \in S^1$, we get $\alpha_k = 0 \quad \forall k \in S^2$. This shows that $F_i^1 = F$ is a maximal face, i.e. a facet. Alternately, we could argue that the $|V|$ points used are affinely independent.

To establish the other direction, we need to show that, if $F_i^1 = \{x \in P_{2C} : x_i = 1\}$ is a facet then $d_G(i, j) \leq 2 \quad \forall j \in V$. We establish the contrapositive by showing that if there exists a $j \in V$ such that $d_G(i, j) > 2$ then $F_i^1 = \{x \in P_{2C} : x_i = 1\}$ is not a facet. When such a j exists, we know that $(i, j) \notin E$ and $N^\cap(i, j) = \emptyset$. Then for this j , the constraint in the system $x_i + x_j - \sum_{k \in N^\cap(i, j)} x_k \leq 1$ reduces to $x_i + x_j \leq 1$ which dominates $x_i \leq 1$. Hence F_i^1 cannot be a facet. \square

Theorem 4 Let P_{2C} denote the 2-Club polytope of a given graph $G = (V, E)$ and let I denote a maximal 2-independent set in G . Then

$$\sum_{i \in I} x_i \leq 1 \tag{7}$$

induces a facet of P_{2C} .

Proof: Since I is a maximal 2-independent set in G , no two vertices from that set can be simultaneously present in a 2-club, as by definition $d_G(i, j) > 2 \forall i, j \in I$. Hence (7) is a valid inequality of Q and in turn P_{2C} . We now establish the maximality of the face $F_I = \{x \in P_{2C} : \sum_{i \in I} x_i = 1\}$, thereby showing it is a facet. Suppose there exists a valid inequality $\alpha^T x \leq \beta$ such that, $F = \{x \in P_{2C} : \alpha^T x = \beta\} \supseteq F_I$. Since $e_i \in F_I \subseteq F \forall i \in I$, we have $\alpha_i = \beta \forall i \in I$. Now for every $j \in V \setminus I$, there exists a vertex $i \in I$ such that at least one of the following two conditions are satisfied:

1. $(i, j) \in E$ and so $e_{ij} \in F_I \subseteq F$;
2. $N^\cap(i, j) \neq \emptyset$, i.e. they have a common neighbor $k \in V \setminus I$ in which case $e_{ikj}, e_{ik} \in F_I \subseteq F$.

Now for every $j \in V \setminus I$, in the first case we obtain, $\alpha_i + \alpha_j = \beta \Rightarrow \alpha_j = 0$ and in the second case we obtain $\alpha_i + \alpha_k + \alpha_j = \beta$ and $\alpha_i + \alpha_k = \beta \Rightarrow \alpha_j = \alpha_k = 0$. Thus, $F_I = F$ is a facet. \square

4.2 Lower bounds

A lower bound can be obtained by observing that complete bipartite graphs have diameter 2 and form ‘edge essential’ 2-clubs. That is, if we remove any edge from a complete bipartite graph, its diameter increases to 3. Hence we know that, if the size of the largest complete bipartite subgraph (not necessarily induced) of G is b^* , then $\bar{\omega}_2(G) \geq b^* \geq b \geq \Delta + 1$, where b is the size of a largest known complete bipartite subgraph of G and Δ is the maximum degree. A vertex of maximum degree with its neighbors (*star* subgraphs) is an easy-to-find complete bipartite subgraph of G and hence the bound. The need for b , computed from heuristic or other techniques arises because of the fact that finding b^* is *NP*-hard [18, p. 195]

Every clique is also a 2-club (actually with one more connected vertex when it exists) and every 2-club is also a 2-clique. Maximum clique is again a popular *NP*-hard problem and as before, we can say that $\omega_2(G) \geq \bar{\omega}_2(G) \geq \omega(G) \geq c$ where c is the size of a known clique.

5 Preliminary Numerical Results

For the numerical experiments, two popular protein interaction networks were chosen. The first network is the protein-protein interaction map of the yeast *Saccharomyces cerevisiae* [21] and the second is the protein-protein interaction map of a gastric pathogen *Helicobacter Pylori* [28, 9].

As mentioned before, both these networks exhibit power-law degree distribution as shown in Figure 5 and Figure 6. Table 1 and Table 2 contain information on the order and the number of connected components of that order in *S. cerevisiae* protein network and *H. Pylori* protein network respectively. Figure 7 graphically illustrates the protein-protein network of *H. Pylori*.

A maximum clique, 2-clique and 2-club were found on both these networks and a maximum 3-clique was found in *S. cerevisiae* using exact approaches.

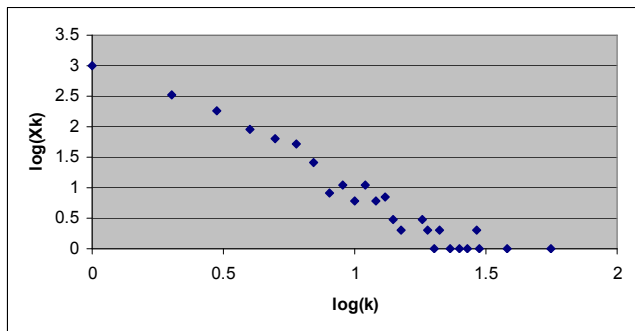


Figure 5: Degree distribution, in logarithmic scale, for the protein network of *S. Cerevisiae*. X_k is the number of vertices of degree k .

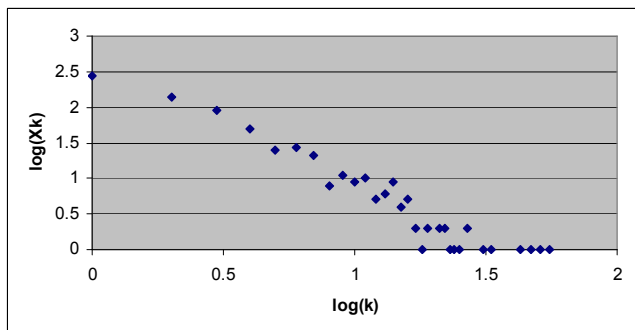


Figure 6: Degree distribution, in logarithmic scale, for the protein network of *H. Pylori*. X_k is the number of vertices of degree k .

Table 3 contains this information. Clique, 2-clique and 3-clique numbers were found by applying Carraghan-Pardalos algorithm [11] for maximum clique on G , G^2 and G^3 respectively. 2-Club number was found by solving the IP formulation (3) using CPLEX[®] [14]. However, in order to solve the maximum 2-club problem on these graphs, some preprocessing techniques were used to reduce the size of the instances.

Denote the closed neighborhood of a set of vertices by

$$N[X] = \left(\bigcup_{i \in X} N(i) \right) \cup X.$$

Since $\bar{\omega}_2(G)$ is bounded below by $\Delta + 1$, a vertex v such that $|N[N[v]]| < \Delta + 1$ cannot be in any optimal solution and can be removed from the graph. This approach can be used to reduce the size of the instance. The reduced instance was decomposed into subproblems externally by setting $x_v = 1$ for a non-leaf vertex v in the reduced graph and deleting all vertices that are at distance 3 or more from the graph. We consider only non-leaf vertices because if a leaf

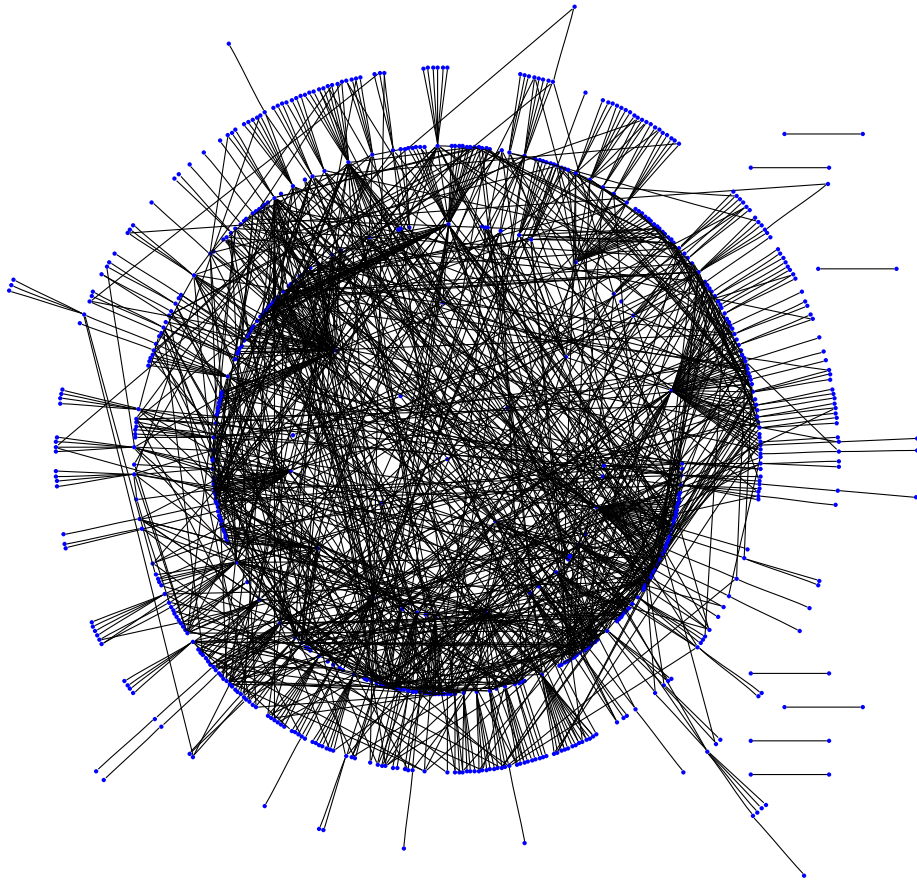


Figure 7: Protein-protein interaction map of *H. Pylori*.

vertex is in a maximum 2-club, then so is its neighbor. This yields the largest 2-club containing v . Now we can delete v and repeat this process. The following pseudo-code summarizes this procedure.

Pre-processing: Input $G = (V, E)$

1. Find maximum degree Δ in G .
2. Find $X = \{v \in V : |N[N[v]]| < \Delta + 1\}$.
3. If $X \neq \emptyset$ then let $G = G(V \setminus X)$ and go to step 1.
4. Return G .

Decomposition Algorithm: Input $G = (V, E)$

1. Let $G = \text{Pre-processing}(G)$.
2. Pick a vertex $v \in V(G)$ with degree at least 2.
3. Solve the IP formulation (3) for $G(N[N[v]])$ with the additional constraint

$x_v = 1$ using CPLEX.

4. Let $G = \text{Pre-processing}(G - v)$.

5. If G has a vertex of degree at least 2 go to step 2; else return the largest 2-club encountered as optimum.

A PENTIUM[®] 4 1.4GHz laptop computer was used in the experiments and the run-times were under a minute in cases where the optimum was obtained. In both biological networks, it turned out that the maximum 2-clique and maximum 2-club correspond to the same solution (subset of vertices). Figure 8 and Figure 9 are the maximum 2-clubs (and maximum 2-cliques) of *S. Cerevisiae* and *H. Pylori* respectively. Figure 10 shows the maximum 3-clique that was found in *S. Cerevisiae*. Observe that it is also a (maximum) 3-club as they are basically three star graphs with their central vertices forming a triangle. Figures 7, 8, 9, 10 were obtained by using the graph visualization software GRAPHVIZ [19].

Table 1: *S. Cerevisiae*. Vertices: 2114; Edges: 2203; Connected components: 417.

Order	#Components	Order	#Components
1	268	5	5
2	101	6	3
3	25	7	4
4	10	1458	1

Table 2: *H. Pylori*. Vertices: 1570; Edges: 1403; Connected components: 858.

Order	Number of Components
1	850
2	7
706	1

Table 3: Clique, 2-Clique, 2-Club, 3-Clique, 3-Club numbers of *S. Cerevisiae* and *H. Pylori* protein maps.

Network	$\omega(G)$	$\omega_2(G)$	$\bar{\omega}_2(G)$	$\omega_3(G)$	$\bar{\omega}_3(G)$
<i>S. Cerevisiae</i>	6	57	57	68	68
<i>H. Pylori</i>	3	56	56	N/A	N/A

6 Conclusion

In this paper we have introduced two cohesive subgroup models, n -cliques and n -clubs from social network analysis as alternatives to cliques for clustering

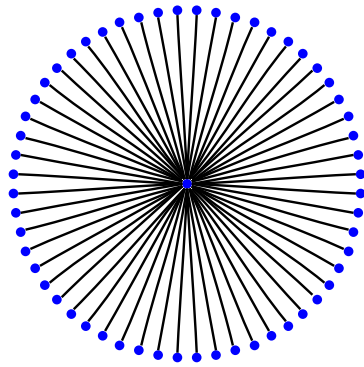


Figure 8: A maximum 2-club and 2-clique of *S. Cerevisiae*.

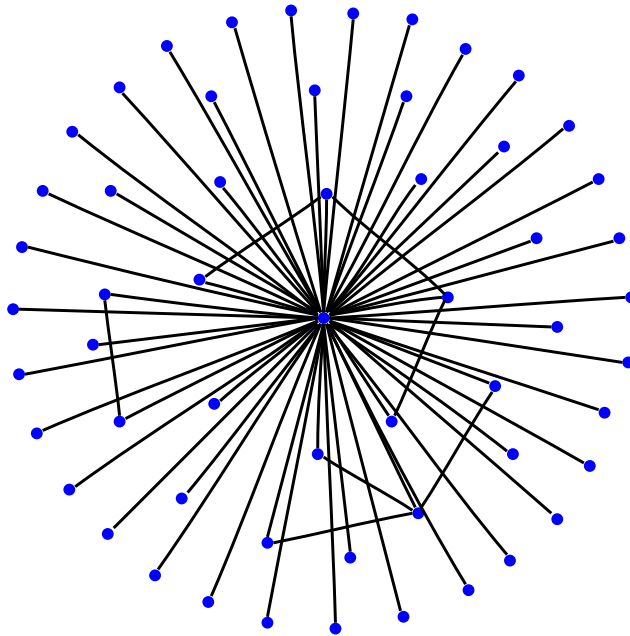


Figure 9: A maximum 2-club and 2-clique of *H. Pylori*.

biological networks. These models can provide useful insights into the structure of these complex networks. *NP*-completeness results for these problems on arbitrary and restricted graph classes are presented and integer programming formulations are proposed for these problems. Basic polyhedral aspects of the maximum 2-club problem have been investigated and trivial and non-trivial facets established. Sample numerical experiments on protein-protein interaction networks of *S. Cerevisiae* and *H. Pylori* are reported. As it can be seen,

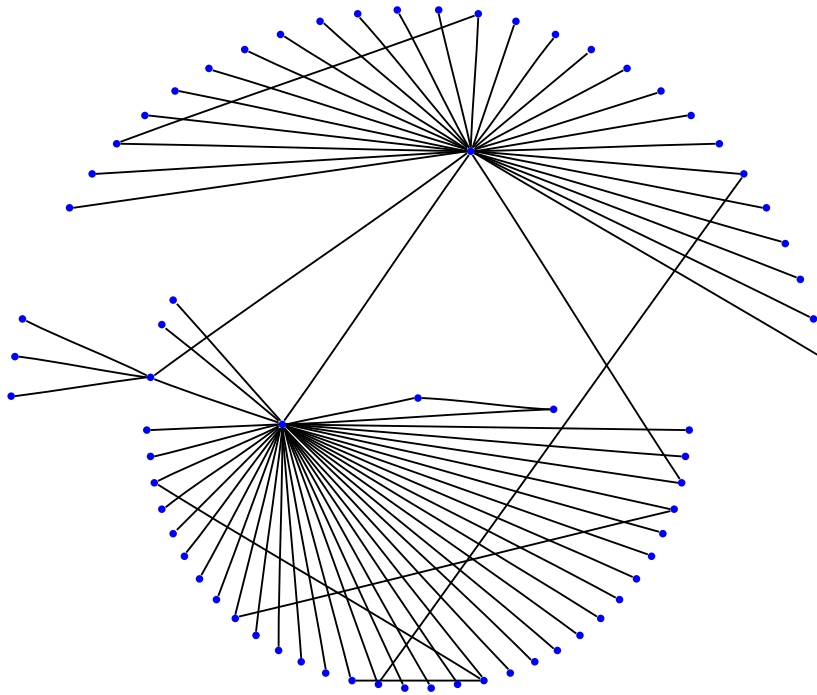


Figure 10: A maximum 3-clique and 3-club of *S. Cerevisiae*.

although the maximum clique sizes in these graphs are very small, relaxing their definitions yielded much larger subsets of vertices that have low diameter. This systematic graph theoretic relaxation of cliques can hence provide useful insights into important substructures in a network.

Acknowledgements The authors would like to thank the anonymous referees for their useful suggestions that improved the content and presentation of this paper.

References

- [1] J. Abello, S. Butenko, P. Pardalos, and M. Resende. Finding independent sets in a graph using continuous multivariable polynomial formulations. *Journal of Global Optimization*, 21:111–137, 2001.
- [2] R. D. Alba. A graph-theoretic definition of a sociometric clique. *Journal of Mathematical Sociology*, 3:113–126, 1973.

- [3] E. Almaas and A.-L. Barabási. Power laws in biological networks. In E. Koonin, editor, *Power laws, scalefree networks and genome biology*. Landes Bioscience, 2005. To appear.
- [4] L. Amaral, A. Scala, M. Barthélemy, and H. Stanley. Classes of small-world networks. *Proc. of National Academy of Sciences USA*, 97:11149–11152, 2000.
- [5] J. Arquilla and D. Ronfeldt. What next for networks and netwars? In J. Arquilla and D. Ronfeldt, editors, *Networks and Netwars: The Future of Terror, Crime, and Militancy*, pages 311–361. RAND Corporation, 2001.
- [6] J. S. Bader, A. Chaudhuri, J. M. Rothberg, and J. Chant. Gaining confidence in high-throughput protein interaction networks. *Nature Biotechnology*, 22(1):78–85, 2004.
- [7] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.
- [8] I. M. Bomze, M. Budinich, P. M. Pardalos, and M. Pelillo. The maximum clique problem. In D.-Z. Du and P. M. Pardalos, editors, *Handbook of Combinatorial Optimization*, pages 1–74. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999.
- [9] Biomolecular Relations in Information Transmission and Expression. Generalized protein interactions. http://www.genome.jp/brite/generalized_interactions.html, 2005. Accessed March 2005.
- [10] S. Busygin, S. Butenko, and P. M. Pardalos. A heuristic for the maximum independent set problem based on optimization of a quadratic over a sphere. *Journal of Combinatorial Optimization*, 6:287–297, 2002.
- [11] R. Carraghan and P. Pardalos. An exact algorithm for the maximum clique problem. *Operations Research Letters*, 9:375–382, 1990.
- [12] H. Chen, W. Chung, J. J. Xu, G. Wang, Y. Qin, and M. Chau. Crime data mining: A general framework and some examples. *Computer*, 37(4):50–56, 2004.
- [13] D. J. Cook and L. B. Holder. Graph-based data mining. *IEEE Intelligent Systems*, 15(2):32–41, 2000.
- [14] Ilog cplex. <http://www.ilog.com/products/cplex/>, 2005. Accessed March 2005.
- [15] R. H. Davis. Social network analysis: An aid in conspiracy investigations. *FBI Law Enforcement Bulletin*, pages 11–19, 1981.

- [16] I. Fischer and T. Meinl. Graph based molecular data mining - an overview. In Wil Thissen, Peter Wieringa, Maja Pantic, and Marcel Ludema, editors, *IEEE SMC 2004 Conference Proceedings*, pages 4578–4582, 2004.
- [17] J. Gagneur, R. Krause, T. Bouwmeester, and G. Casari. Modular decomposition of protein-protein interaction networks. *Genome Biology*, 5(8):R57.1–R57.12, 2004.
- [18] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-completeness*. W.H. Freeman and Company, New York, 1979.
- [19] Graph visualization software. <http://www.graphviz.org/About.php>, 2005. Accessed March 2005.
- [20] F. Harary and I. C. Ross. A procedure for clique detection using the group matrix. *Sociometry*, 20:205–215, 1957.
- [21] H. Jeong, S. P. Mason, A. L. Barabási, and Z. N. Oltvai. Centrality and lethality of protein networks. *Nature*, 411:41–42, 2001. <http://www.nd.edu/~networks/database/index.html>.
- [22] D. Jiang, C. Tang, and A. Zhang. Cluster analysis for gene expression data: A survey. 16(11):1370–1386, 2004.
- [23] P. Krishna, N. Vaidya, M. Chatterjee, and D. Pradhan. A cluster-based approach for routing in dynamic networks. In *ACM SIGCOMM Computer Communication Review*, pages 49–65, 1997.
- [24] R. D. Luce. Connectivity and generalized cliques in sociometric group structure. *Psychometrika*, 15:169–190, 1950.
- [25] R. D. Luce and A. D. Perry. A method of matrix analysis of group structure. *Psychometrika*, 14:95–116, 1949.
- [26] R. J. Mokken. Cliques, clubs and clans. *Quality and Quantity*, 13:161–173, 1979.
- [27] X. Peng, M. A. Langston, A. M. Saxton, N. E. Baldwin, and J. R. Snoddy. Detecting network motifs in gene co-expression networks. 2004.
- [28] J. C. Rain, L. Selig, H. De Reuse, V. Battaglia, C. Reverdy, S. Simon, G. Lenzen, F. Petel, J. Wojcik, V. Schachter, Y. Chemama, A. Labigne, and P. Legrain. The protein-protein interaction map of helicobacter pylori. *Nature*, 409(6817):211–215, 2004. Erratum in: *Nature* 409(6820):553 and 409(6821):743, 2001.
- [29] V. Spirin and L. A. Mirny. Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences*, 100(21):12123–12128, 2003.

- [30] L. Terveen, W. Hill, and B. Amento. Constructing, organizing, and visualizing collections of topically related web resources. *ACM Transactions on Computer-Human Interaction*, 6:67–94, 1999.
- [31] T. Washio and H. Motoda. State of the art of graph-based data mining. *SIGKDD Explor. Newsl.*, 5(1):59–68, 2003.
- [32] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.
- [33] D. Watts. *Small Worlds: The Dynamics of Networks Between Order and Randomness*. Princeton University Press, Princeton, NJ, 1999.
- [34] D. Watts and S. Strogatz. Collective dynamics of “small-world” networks. *Nature*, 393:440–442, 1998.