Social Network Analysis and Community Detection by Decomposing a Graph into Relaxed Cliques

Timo Gschwind*,[a], Stefan Irnich[a], Fabio Furini[b], Roberto Wolfler Calvo[c]

[a] *Chair of Logistics Management, Gutenberg School of Management and Economics, Johannes Gutenberg University Mainz, Jakob-Welder-Weg 9, D-55128 Mainz, Germany.*
[b] *LAMSADE, Université Paris Dauphine, Place du Maréchal de Lattre de Tassigny, F-75016 Paris, France.*
[c] *LIPN, Université Paris 13, F-93430 Villetaneuse, France.*

Abstract

In social network analysis (SNA), relationships between members of a network are encoded in an undirected graph where vertices represent the members of the network and edges indicate the existence of a relationship. One important task in SNA is community detection, that is, clustering the members into communities such that relatively few edges are in the cutsets, but relatively many are internal edges. The clustering is intended to reveal hidden or reproduce known features of the network, while the structure of communities is arbitrary. We propose decomposing a graph into the minimum number of relaxed cliques as a new method for community detection especially conceived for cases in which the internal structure of the community is important. Cliques, that is, subsets of vertices inducing complete subgraphs, can model perfectly cohesive communities, but often they are overly restrictive because many real communities form dense, but not complete subgraphs. Therefore, different variants of relaxed cliques have been defined in terms of vertex degree and distance, edge density, and connectivity. They allow to impose application-specific constraints a community has to fulfill such as familiarity and reachability among members and robustness of the communities. By discussing the results obtained for some very prominent social networks widely studied in the SNA literature we demonstrate the applicability of our approach.

Key words: Community detection, graph decomposition, clique relaxations, social network analysis

1. Introduction

In *social network analysis* (SNA, Wasserman and Faust, 1994; Scott, 2012) there is a growing interest in studying social networks aiming at extracting knowledge from the herein identified structures and characteristic numbers. The analysis of *cohesive groups* also known as *communities* (or *clusters*, *modules*, *blocks*) has received a lot of attention from researchers of different areas like social and computer science, biology, economics, physics, and discrete mathematics. Classically, cliques have been used to model cohesive groups. They can be seen as extremal cohesive groups in the sense that all members are fully connected with each other. This constraint has been found too restrictive in many applications and, therefore, various relaxations of the clique concept, such as s-clique, s-plex, s-club, s-defective clique, and γ-quasi-clique, have been introduced (see Pattillo *et al.*, 2013a, and references given there). We refer to these structures as *relaxed cliques* in the following.

To the best of our knowledge, the graph theory and Operations Research literature on clique relaxations has almost exclusively studied questions related to identifying their structural properties or to solve optimization problems in which a maximal or maximum relaxed clique has to be determined. Surprisingly,

*Corresponding author.
Email address: gschwind@uni-mainz.de (Timo Gschwind)

no research has addressed the related problems of partitioning or covering a graph into/with the smallest number of relaxed cliques, although this type of question was coined by Balasundaram *et al.* (2011, p. 141).

Decomposing a graph into the minimal number of cliques is well-known as the clique partitioning problem. It is equivalent to finding a minimum vertex coloring for the complement graph, and has been discussed intensively, e.g., by Mehrotra and Trick (1998); Nemhauser and Park (1991), and Held *et al.* (2012). We extend this stream of research and propose decomposing a graph into the minimum number of relaxed cliques as a new method for community detection. Note that minimizing the number of relaxed cliques is equivalent to maximizing the average size of the communities.

We study the cases of partitioning (disjoint clusters) as well as covering (overlapping clusters). Any subset of a clique is again a clique, a property known as *hereditary* (Yannakakis, 1978). However, for some classes of relaxed cliques, this is not generally true. As a consequence, partitioning and covering a graph with a minimum number of relaxed cliques can be a different problem, i.e., covering can be a proper relaxation of partitioning. We provide a taxonomy of decomposition problems for the first-order clique relaxations also taking into account that connectivity is often a desirable property of communities that is not automatically fulfilled in some cases.

Our proposed approach is applicable in those cases in which one has a good understanding of the structural properties that a community must have. Aspects such as *familiarity* among members (few strangers), *reachability* among members (quick communication), and *robustness* of the subgroup (not easily destroyable) are often desirable properties of a community (Balasundaram *et al.*, 2011). In a graph-theoretic description, familiarity concerns vertex degrees, reachability concerns distances, and robustness concerns connectivity. Clique relaxations (we give precise definitions in Section 2) like k-core/s-plex, s-club/clique, and k-block/s-bundle, respectively, can model such desired characteristics of subgroups in SNA. Moreover, input data describing a social network may stem from sources that contain errors. In that case using s-defective cliques or γ-quasi-cliques offers a way of absorbing such inaccuracies.

1.1. Literature Review

The idea of decomposing a graph $G = (V, E)$ into clusters is not new. We briefly review related research fields such as graph partitioning, community detection, and graph clustering in order to point out the similarities and differences to the new type of problems introduced within the paper at hand.

In *graph partitioning*, the task is to find a partitioning of the vertex set V into p blocks $V_1, V_2, \ldots, V_p$. Typically, a weight is associated to each vertex of the graph and a maximum capacity of each partition must be respected. We refer to (Garey and Johnson, 1979) where the *graph partitioning problem* has been formally defined and to (Buluç *et al.*, 2013) for a recent comprehensive overview. In addition, balancing constraints requiring that partitions are of (almost) equal size or weight can be considered. If weights are associated to the edges of the graph, the objective can be to minimize the weight of the edge cutsets $E(V_i, V_j)$ comprising all edges connecting different partitions V_i and V_j with $i < j$. Applications of graph partitioning are widespread and include, e.g., the distribution of work in parallel processing, image processing, sparse matrix factorization, very large-scale integration (VLSI) design, and the pre-computation of information to accelerate shortest-path queries in routing. The focus in graph partitioning is on the relationship between different clusters, while the structure of a cluster is almost irrelevant except for its size or weight. In contrast, the primary focus of our work is on the identification of clusters that have very specific structural properties, i.e., that they are relaxed cliques of a specific type.

Community detection is strongly related to our work. The articles by Porter *et al.* (2009) and Schaeffer (2007) and the survey by Fortunato (2010) provide comprehensive overviews showing that diverse methods from various fields have found their way into community detection. Informally speaking, community detection is concerned with clustering the vertex set V of a graph $G = (V, E)$ into communities $V_1, V_2, \ldots, V_p$ such that relatively few edges are found in the cutsets $E(V_i, V_j)$ but relatively many are internal edges $E(V_i)$. In particular, the number p of communities within a given network is typically unknown. In contrast to the balancing constraints in graph partitioning, the communities are generally of unequal size, moreover, their density and structure (if any) can vary. Without explicitly formalizing the relaxed clique partitioning and covering problem, Fortunato (2010) and others offer the concept of relaxed cliques as a viable and reasonable

concept for identifying communities. However, methods for clustering a network into relaxed cliques in a "best possible way" are not described there. This is probably because the standard objectives in community detection assess a clustering by considering both internal edges and edges in the cutsets of the clusters. A very common quality measure in this context is *modularity*, originally introduced by Newman and Girvan (2004): Given p clusters $V_1, V_2, \ldots, V_p$, the *modularity* is $Q(V_1, V_2, \ldots, V_p) = \sum_{i=1}^{p} \left(\frac{|E(V_i)|}{|E|} - exp(V_i) \right)$, where $exp(V_i)$ is the expected fraction of inner-cluster edges of a graph with the same degree distribution as G. A common assumption on the underlying distribution (the so-called *null model*) is that an edge between two vertices i and j appears with probability $|N(i)||N(j)|/|E|$, where $N(i)$ and $N(j)$ is the set of neighboring vertices of i and j, respectively. The modularity value $Q(V_1, V_2, \ldots, V_p)$ varies between -0.5 and 1 depending on the clustering. A value close to 1 indicates that a strong community structure has been identified. Indeed, modularity maximization seems to be appropriate as a general tool for identifying communities with a non-specified structure. Dozens of publications have used modularity to justify newly proposed heuristic clustering methods. Nevertheless, the concept of modularity maximization has also been criticized as it can lead to questionable results (Fortunato and Barthélemy, 2007). Aloise *et al.* (2010) were the first to propose exact algorithms for the modularity maximization problem in networks allowing an absolute evaluation of clustering heuristics. They compare a row-generation algorithm and a direct solution approach using CPLEX to solve a 0-1 MIQP with three column-generation algorithms differing in the underlying formulation and the algorithm used for solving the respective pricing problem. One result is that the column-generation approach with a sparse quadratic subproblem formulation outperforms the other algorithms, while the question of finally producing integer solutions via branching is not addressed in the article.

Clustering methods require a measure of distance (or similarity). In the context of graphs, it means that vertices are considered as (data) points in a metric space. The goal is to find a partitioning of the points V into p clusters (p is typically given) and to minimize or maximize an objective based on distances between points and/or from points to cluster centroids (Fortunato, 2010, p. 93f). Examples are p-means clustering where the average squared distance between points and centroids is minimized, p-centroid where the maximum distance between points and centroids is minimized, and p-clustering sum where the sum of all intra-cluster distances is minimized. While traditional clustering methods need p and additional attributes of vertices as inputs (data points), we solely rely on adjacency information as an input. It is also possible to define a distance by pure adjacency-based measures (e.g. Schaeffer, 2007, p. 36) such as the overlap $|N(i) \cap N(j)|/|N(i) \cup N(j)|$. However, there are no well-defined structural requirements for sharply distinguishing between feasible and infeasible clusters.

Finally, the work of Guo *et al.* (2010) has some relation to the problems studied here, but with a completely different objective and context. The authors consider the so-called *s-plex cluster editing problem* in which, for an undirected graph $G = (V, E)$ and an integer $p \geq 0$, the question is whether G can be modified by up to p edge deletions and insertions into a graph whose connected components are s-plexes. Chang *et al.* (2014) study the complexity of graph partitioning using subgraphs of bounded diameter for restricted classes of graphs, e.g., chordal graphs. They show that the problem is $\mathcal{NP}$-hard in general but also identify some polynomially solvable cases.

1.2. Paper Contribution

The contributions of the paper at hand are the following:

- the formal introduction of a family of partitioning and covering problems with subsets of vertices that are relaxed cliques as a new approach for community detection;
- the presentation of a generic compact formulation of these decomposition problems;
- the introduction of connectivity conditions that each relaxed clique has to respect;
- the first mathematical programming formulations for finding k-blocks and s-bundles, i.e., relaxed cliques defined on the basis of vertex-connectivity;

- the presentation of a computational study on the application of the new models for detecting communities in some real-world social networks that are intensively studied in the SNA and community-detection literature.

The remainder of this paper is organized as follows: In Section 2, we provide an overview over the first-order clique relaxations. Moreover, we discuss a generic and relaxation-specific formulation for finding maximum relaxed cliques in a graph. In Section 3, we formally introduce the decomposition problems and derive a mixed integer programming (MIP) formulation for the partitioning and covering problem, which is compact whenever the formulation for finding a relaxed clique is compact. Results on three widely used social networks are presented and discussed in Section 4, and final conclusions and an outlook close the paper with Section 5.

2. Clique Relaxations

In this section, we introduce the basic notation and different types of relaxed cliques following the taxonomy offered by Pattillo *et al.* (2013a). The precise definition of relaxed cliques, given in Section 2.1, is essential because they form the feasible subsets admissible for the later presented partitioning and covering problems. Problems related to the identification of large relaxed cliques are surveyed in Section 2.2. Section 2.3 then presents known and also new MIP formulations enabling us to specify the generic compact formulation for the graph decomposition problems.

From now on, we assume that a simple graph $G = (V, E)$ with finite vertex set V and edge set E is given. For any subset $S \subseteq V$, the vertex-induced subgraph of S is $G[S] = (S, E \cap (S \times S))$. A graph property Π is *hereditary on vertex induced subgraphs* if for any $S \subseteq V$ with $G[S]$ has property Π it follows that also $G[S']$ has property Π for any $S' \subset S, S' \neq \varnothing$.

In the following, $i \in V$ is any vertex and $S \subseteq V$ is any vertex set. Vertices adjacent to i are denoted by $N(i)$. A set S is a *clique* if $G[S]$ is complete, i.e., all vertices are adjacent. Cliques S form extreme subsets, since all vertices have maximum degree $|S| - 1$, the distance between any two vertices is 1, $G[S]$ has maximum density of 1, and is $(|S| - 1)$-connected.

2.1. Definitions of Relaxed Cliques

The following relaxed cliques are obtained by relaxing a single aspect of the clique definition. In the literature, they are referred to as first-order clique relaxations, while the clique itself is named zero-order clique relaxation (Pattillo *et al.*, 2013a, p. 12).

Relaxing Degree. The *vertex degree* of i is $|N(i)|$ and is denoted by $\deg_G(i)$. The *minimum vertex degree* of G is $\delta(G) = \min_{i \in V} \deg_G(i)$. For $k \geq 0$, S is a *k-core* if $\delta(G[S]) \geq k$. For $s \geq 1$, S is an *s-plex* if $\delta(G[S]) \geq |S| - s$. Every s-plex is an $(|S| - s)$-core, and vice versa. The s-plex clique relaxation has been studied, e.g., in the context of transmission network analysis in Tuberculosis contact investigations by Cook *et al.* (2007).

Relaxing Distance. For two vertices $i, j \in V$, $\text{dist}_G(i, j)$ is the minimum *distance* between i and j, i.e., the minimum length of an i-j-path in G. Note that the length of a path is given by the number of its edges and that $\text{dist}_G(i, j) = \infty$ if i and j are disconnected in G. For $s \geq 1$, S is an *s-clique* if $\text{dist}_G(i, j) \leq s$ for all $i, j \in S$. An s-clique is an ordinary clique (1-clique) in the sth power graph $G^s = (V, \{\{i, j\} : i, j \in V, i < j, \text{dist}_G(i, j) \leq s\})$, and vice versa. Note that any subset of an s-clique is again an s-clique, but since the distance is measured in the given graph G (and not in $G[S]$) the property of being an s-clique is called *weakly hereditary*. The maximum distance is the diameter of G given by $\text{diam}(G) = \max_{i \neq j} \text{dist}_G(i, j)$. For $s \geq 1$, S is an *s-club* if $\text{dist}_{G[S]}(i, j) \leq s$ for all $i, j \in S$ or equivalently $\text{diam}(G[S]) \leq s$. Any s-club is an s-clique, but the reverse it not necessarily true. The s-club and s-clique relaxations have been intensively studied and their relevance for network optimization applications in biology were pointed out by Almeida and Carvalho (2012).

Relaxing Density. For any $S \subseteq V$, the edge set $E(S)$ is the set of edges in G with both endpoints in S. Moreover, the *edge density* of a subgraph $G[S]$ is defined as $\rho(G[S]) = |E(S)|/\binom{|S|}{2}$. For $0 \leq \gamma \leq 1$, S is a *γ-quasi-clique* if $\rho(G[S]) \geq \gamma$. While the density is a relative measure for existing/missing edges, one can also count their number. For $s \geq 0$, S is an *s-defective clique* if $|E(S)| \geq \binom{|S|}{2} - s$. Hence, any s-defective clique has a density of at least $\gamma = 1 - s/\binom{|S|}{2}$, i.e., is a γ-quasi-clique, and vice versa. The s-defective clique has been used, e.g., to identify large protein interaction networks using noisy data collected from large-scale (high-throughput) experiments (Yu *et al.*, 2006).

Relaxing Connectivity. A set $C \subset V$ is a *vertex cut* of a connected graph $G = (V, E)$ if $G[V \setminus C]$ is a disconnected graph. Note that any vertex cut C has at most $|V| - 2$ elements. The vertex connectivity $\kappa(G)$ is the size of a minimum vertex cut. For cliques S, $G[S]$ does not have any vertex cuts, and therefore one defines $\kappa(G[S]) = |S| - 1$. A graph is called k-vertex-connected if its vertex connectivity is k or greater. Let $i, j \in V$ be two different, non-adjacent vertices. The local connectivity $\kappa_G(i, j)$ is the minimum size of a vertex cut C disconnecting i and j in $G[V \setminus C]$. For adjacent vertices i and j, one defines $\kappa_G(i, j) = \infty$. Then, if G is not a clique, $\kappa(G)$ equals the minimum of $\kappa_G(i, j)$ over all pairs of different vertices $i, j \in V$.

Two i-j-paths are called *vertex-disjoint* if they have no vertices in common except i and j. According to Menger's theorem (Menger, 1927), the minimum size of a vertex cut disconnecting i and j is the maximum number of vertex-disjoint paths connecting i and j. Therefore, for non-adjacent vertices i and j, $\kappa_G(i, j)$ is the maximum number of vertex-disjoint i-j-paths. For $k \geq 1$, S is a *k-block* if $\kappa(G[S]) \geq k$. For $s \geq 1$, S is an *s-bundle* if $\kappa(G[S]) \geq |S| - s$. By definition, singleton sets $S = \{i\}$ are no k-blocks but always s-bundles. Connectivity and k-blocks have been comprehensively surveyed by Kammer and Täubig (2005). To the best of our knowledge, the s-bundle relaxation coined in (Pattillo *et al.*, 2013a) has only been studied in (Gschwind *et al.*, 2018).

Table 1 summarizes the definitions of the eight first-order relaxed cliques. In *higher-order clique relaxations*, more than one aspect of the clique definition is relaxed. For example, the (λ, γ)-quasi-clique is a second-order relaxation relaxing degree and density so that each vertex must be connected to at least $\lambda(|S| - 1)$ vertices and the induced subgraph must have a density not smaller than γ. Note that in some cases one property may already result from another property. For an overview of dependencies between first-order relaxations see (Pattillo *et al.*, 2013a, Table 2).

Type of relaxation	Definition	Based on	Clique	Hereditary	Connected
k-core	$\delta(G[S]) \geq k$	Degree	$k = \lvert S\rvert - 1$	no	$\lvert S\rvert \leq 2k + 1$
s-plex	$\delta(G[S]) \geq \lvert S\rvert - s$	Degree	$s = 1$	yes	$\lvert S\rvert \geq 2s - 1$
s-clique	$\mathrm{dist}_G(i, j) \leq s$ for all $i, j \in S$	Distance	$s = 1$	yes, weakly	$s = 1$
s-club	$\mathrm{diam}(G[S]) \leq s$	Distance	$s = 1$	no	always
γ-quasi-clique	$\rho(G[S]) \geq \gamma$	Density	$\gamma = 1$	no	$\left\lceil \gamma\binom{\lvert S\rvert}{2} - \binom{\lvert S\rvert - 1}{2} \right\rceil \geq 1$
s-defective clique	$\lvert E(G[S])\rvert \geq \binom{\lvert S\rvert}{2} - s$	Density	$s = 0$	yes	$\lvert S\rvert \geq s + 2$
k-block	$\kappa(G[S]) \geq k$	Connectivity	$k = \lvert S\rvert - 1$	no	always
s-bundle	$\kappa(G[S]) \geq \lvert S\rvert - s$	Connectivity	$s = 1$	yes	$\lvert S\rvert \geq s + 1$

Table 1: Definition of different clique relaxations, similar to Table 1 in (Gschwind *et al.*, 2017)
Note: The last column gives sufficient conditions for connectivity (Pattillo *et al.*, 2013a, p. 17).

The literature distinguishes between "hereditary on induced subgraphs" (in the proper sense) where the property Π can be directly tested on $G[S]$ without knowing G, and "weakly hereditary" where the property refers to the given graph G. We will use "hereditary" in the comprehensive sense because there are no implications for the algorithmic components that we use.

Requiring Connectivity. In many practical applications, clusters need to be connected. For community detection, e.g., Fortunato (2010, p. 84) stresses that connectedness is a required property. If a community were disconnected, it could be considered as two or more smaller groups. A weakness of general relaxed cliques is that they are not necessarily connected, see last column of Table 1, where sufficient conditions for

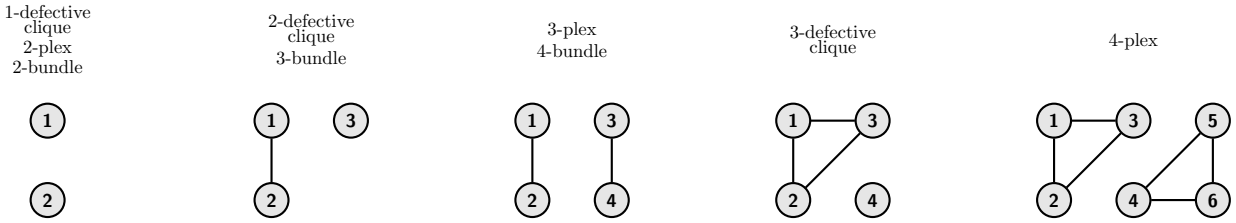


Figure 1: Largest disconnected s-plex, s-defective clique, and s-bundle

connectivity are given. Indeed, arbitrarily large s-cliques can be disconnected because the removal of the central vertex from a star graph induces an edgeless graph, which is however a 2-clique and therefore also an s-clique for all $s \geq 2$. Also, arbitrarily large disconnected γ-quasi-cliques exist resulting from the addition of an isolated vertex to a clique. In contrast, this phenomenon occurs only for small-sized $S \subset V$ in case of s-plex, s-defective clique, and s-bundle, see Figure 1.

As a consequence, we suggest to consider *connected relaxed cliques* S as feasible structures, which result from requiring connectivity of the induced subgraph $G[S]$ in addition to the definition of the respective relaxed clique. Note that for hereditary relaxed cliques (s-plex, s-clique, s-defective clique, and s-bundle) the connectivity requirement makes the resulting structures non-hereditary. For example, a path with three vertices forms a connected 1-defective clique, which becomes disconnected when the middle vertex is removed.

2.2. Large Relaxed Cliques

The scientific literature has focused mainly on finding a largest relaxed clique in a given graph. The attribute large may refer to relaxed cliques S that either are of maximum cardinality, are maximal with respect to inclusion, or have maximum weight. In this section, we briefly review the most successful exact algorithms for this purpose, classify them into MIP-based and others, because only the MIP-based algorithms can serve as a basis for the generic decomposition model that we present later.

Maximum-Cardinality Relaxed Cliques. Before we define the related optimization problems, it is helpful to describe some properties in order to classify types of relaxed cliques. Let Π be the *graph property*, e.g., describing a specific clique relaxation. According to Yannakakis (1978), a property Π is *nontrivial* if it is true for all graphs $G[S]$ induced by singleton sets $S = \{i\}$, but not fulfilled for every graph. Property Π is *interesting* if there exist arbitrarily large graphs satisfying it.

The problem of finding a relaxed clique $S \subseteq V$ with largest cardinality $|S|$ is known as the *maximum (-cardinality) relaxed clique problem* (MC-RC). For nontrivial, interesting, and hereditary (on induced subgraphs, see Section 2) properties Π, Yannakakis (1978) has shown that MC-RC is $\mathcal{NP}$-hard. It is straightforward to see that Π is hereditary for clique, s-plex, s-defective clique, and s-bundle. Consequently, MC-RC is an $\mathcal{NP}$-hard problem for these structures. The same holds for s-clique with $s > 1$ (equivalent to clique in power graph).

The properties Π of being a k-core, s-club (for $s > 1$), γ-quasi-clique (for $\gamma < 1$), or k-block are not hereditary. The theorem by Yannakakis (1978) is therefore not applicable. Indeed, the maximum-cardinality k-core problem is polynomially solvable (see Kosub, 2004). The computation of the k-connected components of a graph (solving the maximum-cardinality k-block problem and the k-block partitioning problem) can be done in polynomial time for fixed k (see Kammer and Täubig, 2005). For s-club, the $\mathcal{NP}$-hardness of MC-RC was proven by (Bourjolly *et al.*, 2002). Recently, Pattillo *et al.* (2013b) showed that MC-RC for γ-quasi-cliques is $\mathcal{NP}$-complete.

Table 2 summarizes the exact solution approaches for MC-RC for the first-order clique relaxations. Exact algorithms for clique are too numerous to be listed here and we refer to (Carraghan and Pardalos, 1990; Abello *et al.*, 1999; Östergård, 2002). Note that these algorithms can solve MC-RC for s-cliques by considering the sth power graph.

Clique relaxation	MIP-based	other
k-core	n.a.	polynom. solvable, see (Kosub, 2004)
s-plex	B&C: (Balasundaram *et al.*, 2011)	CB&B: (Trukhanov *et al.*, 2013; Gschwind *et al.*, 2018)
s-clique	B&C: (Nemhauser and Trotter Jr., 1974, 1975)	(any non-MIP-based for clique)
s-club	MIP: (Bourjolly *et al.*, 2000; Veremyev and Boginski, 2012), B&C: (Almeida and Carvalho, 2012, 2013)	CB&B: (Bourjolly *et al.*, 2002; Mahdavi Pajouh and Balasundaram, 2012; Shahinpour and Butenko, 2013; Moradi and Balasundaram, 2015), SAT:(Wotzlaw, 2014)
γ-quasi-clique	MIP: (Pattillo *et al.*, 2013b; Veremyev *et al.*, 2015)	
s-defective clique	B&C: (Sherali and Smith, 2006)	CB&B: (Trukhanov *et al.*, 2013; Gschwind *et al.*, 2018)
k-block	n.a.	polynom. solvable, see (Kammer and Täubig, 2005)
s-bundle	n.a.	CB&B:(Gschwind *et al.*, 2018)

Table 2: Exact algorithms for MC-RC for different clique relaxations
Note: B&C=branch-and-cut, CB&B=combinatorial branch-and-bound, MIP=(mixed) integer model (no cutting planes), SAT=formulation as a partial max-sat problem

Inclusion Maximal Relaxed Cliques. If a subset $S \subseteq V$ is a largest relaxed clique with respect to inclusion then S is a *maximal relaxed clique*. Obviously, any maximum relaxed clique is also maximal, but the reverse is not necessarily true. For all variants, the question whether or not S induces a relaxed clique is efficiently decidable. Therefore, for hereditary Π, finding inclusion maximal relaxed cliques can be done efficiently by adding vertices in a one-by-one fashion. In contrast, the maximality test regarding subset inclusion is $\mathcal{NP}$-hard for s-club as shown by Mahdavi Pajouh and Balasundaram (2012).

Maximum-Weight Relaxed Cliques. If weights $w_i \in \mathbb{R}$ are given for all vertices $i \in V$, the *maximum-weight relaxed clique* problem (MW-RC) consists of finding a subset $S \subseteq V$ such that $w(S) = \sum_{i \in S} w_i$ is maximum and S is a relaxed clique. Clearly, with unit weights MW-RC reduces to MC-RC. For relaxed cliques with hereditary Π, it is no restriction to assume that weights w_i are non-negative because otherwise a vertex with negative weight can be eliminated from the consideration.

2.3. Mathematical Formulations for Relaxed Cliques

Different formulations for the MC-RC and MW-RC variants have been suggested in the literature (see Pattillo *et al.*, 2012, 2013a, for an overview). All formulations use either variables $x_i \in \{0,1\}$ to indicate that vertex $i \in V$ is in the relaxed clique S, or variables $y_e \in \{0,1\}$ to indicate that $G[S]$ contains edge $e \in E$, or both. The properties defining Π can be formulated using MIP. The relaxed cliques can be described with the help of a polytope describing a set $\mathscr{F}(G)$ of integer points such that $(\mathbf{x}, \mathbf{y}) \in \mathscr{F}(G)$ holds if and only if $G[S]$ with $S = \{i \in V : x_i = 1\}$ fulfills Π. We can write the following generic model for MW-RC:

$$\max \sum_{i \in V} w_i x_i, \quad \text{s.t.} \quad (\mathbf{x}, \mathbf{y}) \in \mathscr{F}(G) \tag{1}$$

A possible way to ensure the compatibility of vertex and edge variables is setting $x_i x_j = y_{ij}$ for all $\{i,j\} \in E$ and to apply the McCormick (1976) linearization for binary variables. We assume that this or any alternative coupling mechanism is already part of the definition of $\mathscr{F}(G)$. Note that additional variables, other than x and y, may be used to define the set $\mathscr{F}(G)$ or that in some formulations the y variables are useless.

Several mathematical formulations describing the first-order relaxed cliques introduced in Section 2 can be found in the literature. An ordinary clique is described by $\mathscr{F}^1(G) = \{x_i \in \{0,1\} : x_i + x_j \leq 1 \text{ for all } i,j \in V, i < j \text{ with } \{i,j\} \notin E\}$. Polyhedral results can be found in (Nemhauser and Trotter Jr.,

1974, 1975). These results transfer directly to s-cliques with $s \geq 2$, since an s-clique is an ordinary clique in the power graph G^s.

s-Plex. Balasundaram *et al.* (2011) provide the following compact formulation for s-plex. Here, $\mathscr{F}^2(G) = \{x_i \in \{0,1\} : \sum_{j \in V \setminus N(i), j \neq i} x_j \leq (s-1)x_i + \bar{d}_i(1-x_i) \text{ for all } i \in V\}$, where the constant $\bar{d}_i$ is defined as $|V \setminus N(i)| - 1$.

s-Defective Clique. The complement of an s-defective clique is a *generalized vertex packing* (GVP-s, Sherali and Smith, 2006). Therefore, s-defective cliques can be detected as GVP-s in the complement graph $\bar{G} = (V, \bar{E})$, where $\bar{E} = \{\{i,j\} : i,j \in V, i < j, \{i,j\} \notin E\}$. The set $\mathscr{F}^3(G)$ is given by $\{x_i \in \{0,1\}, \bar{y}_{ij} \geq 0 : \bar{y}_{ij} \geq x_i + x_j - 1, \{i,j\} \in \bar{E}; \sum_{\{i,j\} \in \bar{E}} \bar{y}_{ij} \leq s\}$. It has additional $\bar{y}_{ij}$ variables for all $\{i,j\} \in \bar{E}$.

γ-Quasi-Clique. Pattillo *et al.* (2013b) describe γ-quasi-cliques by $\mathscr{F}^4(G) = \{x_i \in \{0,1\}, y_{ij} \geq 0 : \sum_{i<j}(\gamma - a_{ij})y_{ij} \leq 0; y_{ij} \leq x_i, y_{ij} \leq x_j, y_{ij} \geq x_i + x_j - 1 \text{ for } i,j \in V, i < j\}$, where (a_{ij}) is the adjacency matrix of G and the y_{ij} variables are defined for every pair of vertices $i,j \in V, i < j$. The authors also present a more compact formulation with $|V|$ binary and $|V|$ continuous variables and $4|V|+1$ constraints; the associated polytope is $\mathscr{F}^5(G)$.

In the more recent paper (Veremyev *et al.*, 2015), four alternative formulations for the γ-quasi-cliques are given and compared against $\mathscr{F}^4(G)$ and $\mathscr{F}^5(G)$. For the sake of brevity, we present only one of the alternative formulations, i.e., the one that was identified as giving the most consistent results and best bounds for maximum-cardinality (Veremyev *et al.*, 2015, p. 210ff). The associated polytope $\mathscr{F}^6(G)$ uses additional binary variables t_s for $s \in \mathcal{S} := \{1, 2, \ldots, |V|\}$ to indicate the size of the γ-quasi-clique. The formulation of the maximum-cardinality γ-quasi clique problem is:

$$\max \quad \sum_{i \in V} x_i \tag{2a}$$

$$\text{s.t.} \quad y_{ij} \leq x_i, y_{ij} \leq x_j \qquad \{i,j\} \in E \tag{2b}$$

$$\sum_{e \in E} y_e \geq \gamma \sum_{s \in \mathcal{S}} \frac{s(s-1)}{2} t_s \tag{2c}$$

$$\sum_{i \in V} x_i = \sum_{s \in \mathcal{S}} s t_s \tag{2d}$$

$$\sum_{s \in \mathcal{S}} t_s = 1 \tag{2e}$$

$$t_s \geq 0 \qquad s \in \mathcal{S} \tag{2f}$$

$$x_i \in \{0,1\}, y_e \geq 0 \qquad i \in V, e \in E \tag{2g}$$

Here, the coupling between vertex and edge indicator variables is established via (2b), the γ-related constraint on the number of edges in the induced graph is (2c), the coupling between the x- and t-variables is given by (2d), and the unique cardinality of the induced graph is enforced via (2e). Veremyev *et al.* (2015) show that a smaller formulation results from replacing the possible sizes $\mathcal{S}$ by $\{l, l+1, \ldots, u\}$ when a lower bound l and an upper bound u is known.

s-Club. Several formulations for maximum-cardinality s-club are known. Mahdavi Pajouh *et al.* (2016) present a tailored model for 2-clubs. The first model for arbitrary $s \geq 2$ is the path-based formulation by Bourjolly *et al.* (2000) which uses indicator variables x_i for the vertices and additional variables for all paths of length at most s. With the coupling of both types of variables, the number of variables and constraints is bounded by the number of paths, which is of the order of $\mathcal{O}(|V|^{s+1})$ for dense graphs. However, for fixed s the formulation is compact, i.e., polynomial in $|V|$ and $|E|$ and valid inequalities together with a branch-and-cut algorithm were presented by Carvalho and Almeida (2011); Almeida and Carvalho (2012). Veremyev and Boginski (2012) proposed the first compact formulation with a polynomial number of variables and constraints (polynomial in s, $|V|$, and $|E|$). Since this formulation (polytope $\mathscr{F}^7(G)$) is relatively spacious, we describe it in the following.

2.3.1. MIP Formulation for Maximum s-Club

Following (Veremyev and Boginski, 2012), we assume that the simple graph $G = (V, E)$ with vertex weights w_i for $i \in V$ is given together with some integer $s \geq 2$. In addition to vertex variables $x_i \in \{0, 1\}$ defining $S = \{i \in V : x_i = 1\}$, there are variables $v_{ij}^{\ell} \in \{0, 1\}$ indicating that an i-j-path of length $\leq \ell$ exists in $G[S]$, i.e., $\text{dist}_{G[S]}(i, j) \leq \ell$. In (Veremyev and Boginski, 2012), the domain of the indices ℓ is not completely defined. We therefore present a sightly modified version of the model in which a minimum number of the path variables v_{ij}^{ℓ} is needed. Since i-j-paths of length one are the edges $\{i, j\}$, the domain of the index ℓ can be defined as $dom_2(i, j) = \{\max\{2, \text{dist}_G(i, j)\}, \ldots, s\}$. Similarly, we define $dom_3(i, j) = \{\max\{3, \text{dist}_G(i, j)\}, \ldots, s\}$. With the definition $\mathcal{U} = \{\{i, j\} : i, j \in V, i < j\}$ for unordered pairs, the formulation of the maximum-weight s-club problem is:

$$\max \quad \sum_{i \in V} x_i \tag{3a}$$

$$\text{s.t.} \quad v_{ij}^{\ell} \leq x_i, \quad v_{ij}^{\ell} \leq x_j \qquad \{i, j\} \in \mathcal{U}, \ell \in dom_2(i, j) \tag{3b}$$

$$\sum_{\ell \in dom_2(i,j)} v_{ij}^{\ell} \geq x_i + x_j - 1 \qquad \{i, j\} \in \mathcal{U} \setminus E \tag{3c}$$

$$v_{ij}^{2} \leq \sum_{p \in N(i) \cap N(j)} x_p \qquad \{i, j\} \in \mathcal{U}, \text{dist}_G(i, j) = 2 \tag{3d}$$

$$v_{ij}^{\ell} \leq \sum_{p \in N(i), \text{dist}(p,j) \leq \ell - 1} v_{pj}^{\ell - 1} \qquad \{i, j\} \in \mathcal{U}, 2 \leq dist_G(i, j) \leq s, \ell \in dom_3(i, j) \tag{3e}$$

$$x_i \in \{0, 1\} \qquad i \in V \tag{3f}$$

$$v_{ij}^{\ell} \in \{0, 1\} \qquad \{i, j\} \in \mathcal{U}, \ell \in dom_2(i, j) \tag{3g}$$

Due to (3b), the selection of a path associated with v_{ij}^{ℓ} is only possible if both endpoints are present. Conversely, the constraints (3c) allow the selection of both vertices i and j if and only if there exists a path of length $\leq s$ between them in $G[S]$. The constraints (3d) and (3e) model the construction of paths in $G[S]$ connecting i and j. The first constraints guarantee that a vertex adjacent to i and j is selected for the distance two, while the latter work recursively. A path of length ℓ between vertices i and j requires the selection of a vertex p adjacent to i together with the presence of another path of length $\ell - 1$ between p and j. The domains of the vertex and path variables are defined by (3f) and (3g). Note that clique-like constraints $x_i + x_j \leq 1$ for incompatible vertices $i, j \in V$ are present in the above formulation: If $\text{dist}_G(i, j) > s$ for $i, j \in V$, no s-club can contain both vertices, and $dom_2(i, j)$ is the empty set by definition so that the corresponding constraint (3c) reduces to $x_i + x_j \leq 1$. Hence, any valid inequalities for the clique polytope of the corresponding power graph G^s are valid and may be used to strengthen the LP relaxation of the model. Moreover, Veremyev and Boginski (2012) presented additional valid inequalities for $\mathscr{F}^7(G)$.

2.3.2. MIP Formulation for Maximum s-Bundle and k-Block

To the best of our knowledge, no MIP formulations for k-block and s-bundle have been presented in the literature. We suggest the following formulations and denote the associated polyhedron by $\mathscr{F}^8(G)$. Recall that a simple graph $G = (V, E)$ with vertex weights w_i for $i \in V$ is given together with some integer $s \geq 2$.

Let $\mathcal{N} = (N, A)$ be an auxiliary network associated with G defined as follows: For each vertex $i \in V$ there exist two vertices i^- and i^+ in $\mathcal{N}$ so that $N = V^+ \cup V^-$. The network $\mathcal{N}$ comprises two types of arcs. First, for all $i \in V$, arcs (i^-, i^+) are present in A. Second, for each edge $\{i, j\} \in E$, the arcs (i^+, j^-) and (j^+, i^-) are in A. Hence, $A = \{(i^-, i^+) : i \in V\} \cup \{(i^+, j^-), (j^+, i^-) : \{i, j\} \in E\}$. All arcs have unit capacity. Now, any two non-adjacent vertices $i, j \in V$ are k-connected in G if and only if there exists a flow of value k between i^+ and j^- in $\mathcal{N}$. The same holds for $G[S]$ and the induced network $\mathcal{N}[S^+ \cup S^-]$ for any $S \subseteq V$.

Three types of decision variables are in the MIP: The binary variables x_i for $i \in V$ indicate whether or not vertex $i \in V$ is in the selected s-bundle $S = \{i \in V : x_i = 1\}$. The continuous variable u describes the number $|S| - s$ of vertex-disjoint paths that must exist between non-adjacent pairs of vertices of S. With

the definition $\mathcal{U} = \{\{i,j\} : i,j \in V, i < j\}$ for unordered pairs, for each $\{i,j\} \in \mathcal{U} \setminus E$, the binary variables $y_a^{ij}, a \in A$ model flows in $\mathcal{N}$ connecting i^+ and j^-.

$$\max \quad \sum_{i \in V} x_i \tag{4a}$$

$$\text{s.t.} \quad u \geq \sum_{i \in V} x_i - s \tag{4b}$$

$$\sum_{a \in \delta^+(i^+)} y_a^{ij} \geq u - M^{ij}(2 - x_i - x_j) \qquad \{i,j\} \in \mathcal{U} \setminus E \tag{4c}$$

$$\sum_{a \in \delta^+(n)} y_a^{ij} - \sum_{a \in \delta^-(n)} y_a^{ij} = 0 \qquad \{i,j\} \in \mathcal{U} \setminus E, n \in N, n \neq i^+, j^- \tag{4d}$$

$$\sum_{a \in \delta^-(j^-)} y_a^{ij} \geq u - M^{ij}(2 - x_i - x_j) \qquad \{i,j\} \in \mathcal{U} \setminus E \tag{4e}$$

$$y_{p^-p^+}^{ij} \leq x_p \qquad \{i,j\} \in \mathcal{U} \setminus E, p \in V \tag{4f}$$

$$x_i \in \{0,1\} \qquad i \in V \tag{4g}$$

$$y_a^{ij} \geq 0 \qquad a \in A, \{i,j\} \in \mathcal{U} \setminus E \tag{4h}$$

$$u \geq 0 \tag{4i}$$

The objective (4a) maximizes the sum of the vertex weights in the selected s-bundle S. The constraint (4b) guarantees $u \geq |S| - s$. The next three groups of constraints (4c)–(4e) ensure a flow of at least u between i^+ and j^- in case that i and j belong to the bundle. Herein, $M^{ij} > 0$ is a sufficiently large number. The coupling constraints (4f) guarantee that flows are positive only in $\mathcal{N}[S^+ \cup S^-]$. The domains of all variables are stated in (4g)–(4i).

In an s-bundle S, every vertex must have a degree $\deg_{G[S]}(i)$ not smaller than $|S| - s$ (see Pattillo *et al.*, 2013a, p. 17). Based on this observation, we can find a feasible, but small value for M^{ij} in constraints (4c) and (4e) in order to tighten the formulation:

$$M^{ij} := \max\{k \in \mathbb{N} : \exists S \subseteq V, |S| - s = k, \forall v \in S : \max\{\deg_{G \setminus \{i\}}(v), \deg_{G \setminus \{j\}}(v)\} \geq k\}$$

This maximum can be computed by simply sorting all vertices decreasingly by the values $\max\{\deg_{G \setminus \{i\}}(v), \deg_{G \setminus \{j\}}(v)\}$.

Note that a similar formulation can be used to find maximum-weight k-blocks. The variable u can be replaced by the constant k so that (4b) and (4i) are obsolete.

2.3.3. Handling Connectivity

As discussed above, some types of relaxed cliques are not necessarily connected. In the presented MIP formulations, a straightforward way to impose connectivity is to add constraints

$$\sum_{i \in S} x_i + \sum_{i \in V \setminus S} (1 - x_i) \leq |V| - 1 \qquad S \subseteq V : \kappa(G[S]) \geq 2. \tag{5}$$

In general, this is an exponential number of constraints requiring a cutting-plane procedure to solve the MIP.

3. Partitioning and Covering a Graph with a Minimum Number of Relaxed Cliques

Community detection consists in partitioning or covering a graph into/with clusters. We propose a new approach for community detection based on decomposing the graph into a minimum number of relaxed cliques. No reasonable problem results for k-core because some vertices may have a degree smaller than k and

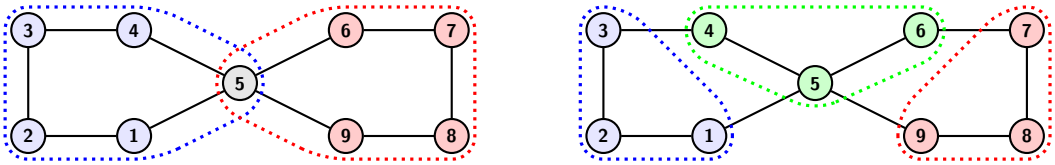


Figure 2: Covering and partitioning into a minimum number of 2-clubs.

	Connected		General		
	Partitioning	Covering	Partitioning	Covering	with
Vertex coloring:	clique				
17 interesting new	s-plex	s-plex	s-plex		$s \geq 2$
decomposition	s-clique	s-clique			$s \geq 2$
problems:			s-club	s-club	$s \geq 2$
	γ-quasi-clique	γ-quasi-clique	γ-quasi-clique	γ-quasi-clique	$0 < \gamma < 1$
	s-defective clique	s-defective clique	s-defective clique		$s \geq 1$
	s-bundle	s-bundle	s-bundle		$s \geq 2$

Table 3: Variants of partitioning and covering with relaxed cliques

cannot belong to any k-core. For k-block, the resulting problem is to determine the k-connected components, for which efficient algorithms exist (Kammer and Täubig, 2005). Therefore, we restrict ourselves to the six remaining first-order clique relaxations.

According to Porter *et al.* (2009) the "detection of network communities that overlap is especially appealing in the social sciences, as people belong simultaneously to several communities (constructed via colleagues, family, hobbies, etc.)". Clearly, covering is always a relaxation of partitioning and this relaxation is proper for non-hereditary structures. The decomposition into 2-clubs shown in Figure 2 is an example.

The non-heredity of a particular structure may either result from the property Π defining the type of relaxed clique or from the connectivity requirement. When connectivity is not already ensured by the definition of the relaxed clique, the variants double and we analyze variants with and without connectivity requirement.

All interesting variants of partitioning and covering with first-order relaxed cliques are summarized in Table 3. Since partitioning and covering are identical for hereditary Π and without connectivity requirement, s-plex, s-defective clique, and s-bundle are listed in the partitioning column only. Moreover, an s-clique is an ordinary clique (1-clique) in the s-th power graph $G^s = (V, E^s)$ with $E^s = \{\{i, j\} : i, j \in V, \mathrm{dist}_G(i, j) \leq s\}$, and vice versa. Hence, we do not consider partitioning and covering with general s-cliques. In contrast, for $s \geq 2$ and with connectivity imposed, partitioning and covering with connected s-cliques differ from clique partitioning in the power graph G^s and differ from the VCP in the complement graph (also the vertex coloring is beyond of the scope of this paper). Figure 3 provides an example.

Finally, it is a non-trivial result that partitioning and covering with connected s-cliques is generally not equivalent. Figure 4 depicts the smallest example that we were able to construct (a graph with 67 vertices and 84 edges). In this example we show that for $s = 4$ partitioning and covering into connected 4-cliques results in a minimum of nine partitions but only eight covering subsets. It is straightforward to generalize the example to $s \geq 5$, but it remains an open question if there exist examples for $s = 2$ and 3 and examples of smaller size.

3.1. Generic Compact Formulation

A generic compact mathematical formulation for all variants needs an upper bound $\bar{rc}(G)$ on the minimum number $rc(G)$ of relaxed cliques in a solution so that they can be numbered by $h \in H = \{1, 2, \ldots, \bar{rc}(G)\}$. Then, binary variables $z^h, h \in H$ indicate whether or not the hth relaxed clique is non-empty in the solution. The sets of variables $(\mathbf{x}^h, \mathbf{y}^h)$ for $h \in H$ model the hth relaxed clique S_h in the sense

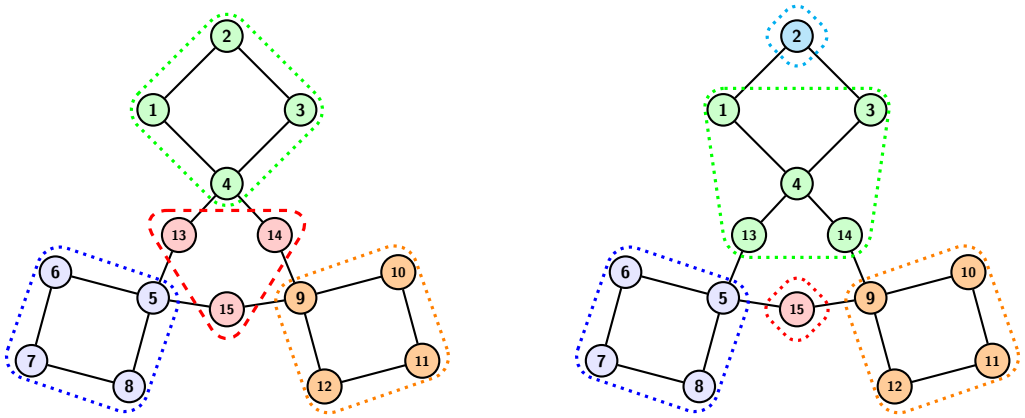


Figure 3: Partitioning into a minimum number of (four) general 2-cliques and (five) connected 2-cliques. Note that $S = \{13, 14, 15\}$ induces the disconnected subgraph $G[S] = (S, \varnothing)$.

that x_i^h and y_e^h expresses that vertex i and edge e belong to $G[S_h]$, respectively. The generic formulation reads as follows:

$$rc(G) = \min \quad \sum_{h \in H} z^h \tag{6a}$$

$$\text{s.t.} \quad \sum_{h \in H} x_i^h = 1 \quad (\text{or} \geq 1) \qquad i \in V \tag{6b}$$

$$z^h \geq x_i^h \qquad i \in V, h \in H \tag{6c}$$

$$(\mathbf{x}^h, \mathbf{y}^h) \in \mathscr{F}(G) \qquad h \in H \tag{6d}$$

$$z^h \in \{0, 1\} \qquad h \in H \tag{6e}$$

The objective (6a) minimizes the number of relaxed cliques in the solution. (6b) are the partitioning/covering constraints. Constraints (6c) ensure that $S_h = \{i \in V : x_i^h = 1\}$ is the empty set whenever $z^h = 0$. The feasibility of S_h is ensured by (6d).

The following theorem shows that formulations eqrefmodel:generic-compact-part-cover have a very weak linear relaxation.

Theorem 1. *For the polyhedra $\mathscr{F}^1(G)$ (clique and s-clique), $\mathscr{F}^2(G)$ (s-plex), $\mathscr{F}^3(G)$ (s-defective clique), $\mathscr{F}^4(G)$, $\mathscr{F}^5(G)$, and $\mathscr{F}^6(G)$ (γ-quasi-clique), $\mathscr{F}^7(G)$ (s-club), and $\mathscr{F}^8(G)$ (s-bundle), let $lp(G)$ be the value of the linear relaxation of the compact formulation (6).*

(a) For every solution $(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}})$ to the linear relaxation, there always exists an equivalent perfectly symmetric solution with $x_i^1 = x_i^2 = \cdots = x_i^{\bar{rc}(G)} = \frac{1}{\bar{rc}(G)} \sum_{h=1}^{\bar{rc}(G)} \hat{x}_i^h$ for each $i \in V$, $y_e^1 = y_e^2 = \cdots = y_e^{\bar{rc}(G)} = \frac{1}{\bar{rc}(G)} \sum_{h=1}^{\bar{rc}(G)} \hat{y}_e^h$ for each $e \in E$, and $z^1 = z^2 = \cdots = z^{\bar{rc}(G)} = \frac{1}{\bar{rc}(G)} \sum_{h=1}^{\bar{rc}(G)} \hat{z}^h$.

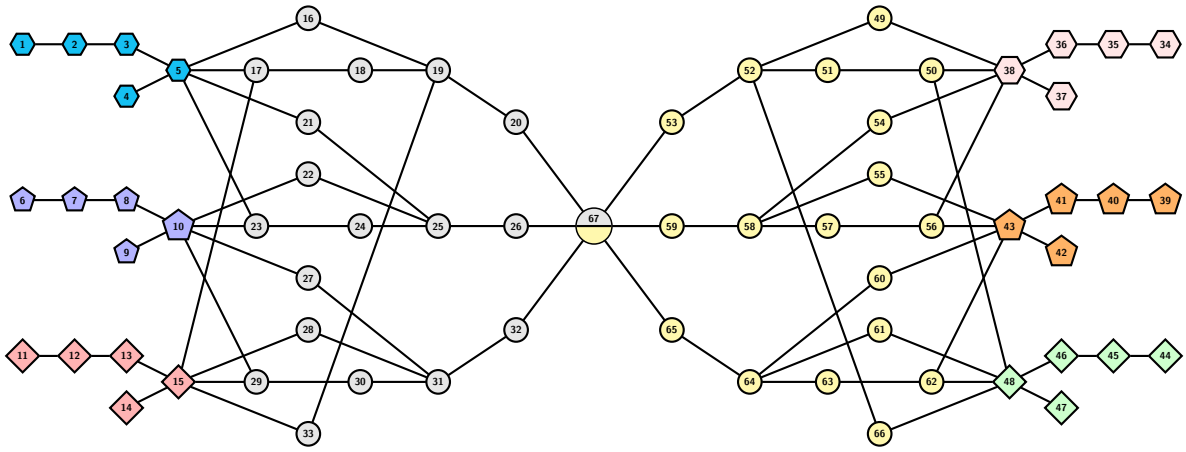
(b) For $\bar{rc}(G) = 1$, the linear relaxation is tight, i.e., $lp(G) = rc(G)$.

(c) For $\bar{rc}(G) \geq 2$, the linear relaxation is not tight and $lp(G) = 1$.

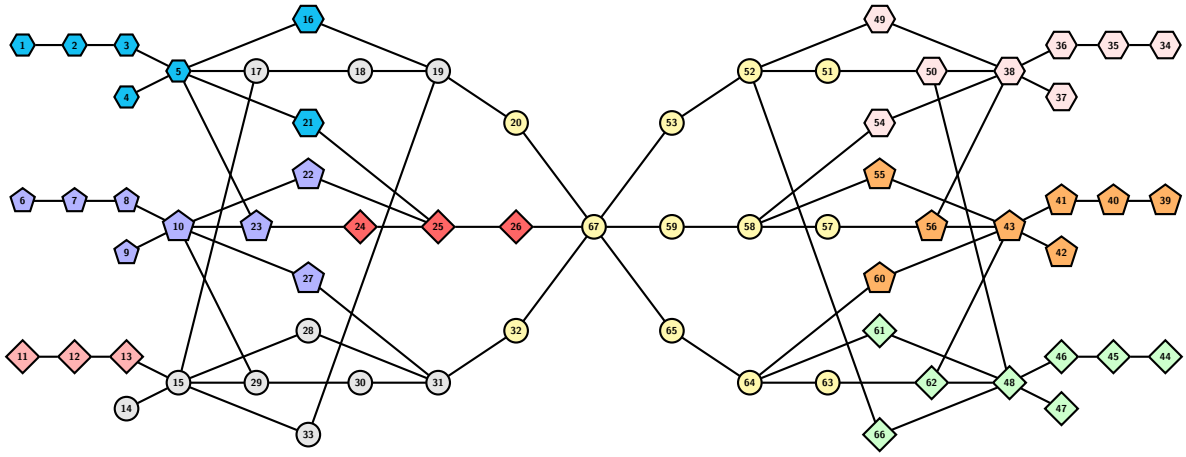
Proof: See Appendix, Section A.

3.2. Set-Partitioning and Set-Covering Formulations

Given the proposed compact formulation (6), a natural Dantzig-Wolfe decomposition can be derived as follows: The partitioning/covering constraints (6b) become the coupling constraints and the constraints



(a)



(b)

Figure 4: Graph with 67 vertices and 84 edges: (a) Covering solution with eight connected 4-cliques, vertex 67 is covered twice, (b) Partitioning solution with nine connected 4-cliques

(6c)–(6e) form the subproblems, identical for each block $h \in H$, and thus blocks can be aggregated (cf. Lübbecke and Desrosiers, 2005). Since each block contains the element $(\mathbf{x}^h, \mathbf{y}^h, z^h) = (\mathbf{0}, \mathbf{0}, 0)$ with cost zero and the number of blocks was chosen sufficiently large, there is no generalized convexity constraint in this Dantzig-Wolfe reformulation. Let Ω be the set of all feasible relaxed cliques. Then, the *integer master program* (IMP) is:

$$\min \sum_{S \in \Omega} \lambda_S \tag{7a}$$

$$\text{s.t.} \sum_{S \in \Omega : i \in S} \lambda_S = 1 \quad (\text{or} \geq 1) \qquad \forall i \in V \tag{7b}$$

$$\lambda_S \geq 0 \quad \text{integer} \qquad \forall S \in \Omega. \tag{7c}$$

The objective (7a) minimizes the number of relaxed cliques, (7b) are the covering/partitiong constraints, and (7c) define the domain of the variables.

Already for relatively small graphs the size of the set Ω becomes huge. Therefore, IMP (7) must be solved using column-generation techniques (Desaulniers *et al.*, 2005). The starting point is a restricted master program (RMP) which is the linear relaxation of (7) defined over a (small) subset $\Omega' \subset \Omega$ of the variables λ_S,

$S \in \Omega'$. The column-generation process alternates between the (re-)optimization of the current RMP and the generation of new variables with negative reduced cost, i.e., $\tilde{c}_S = 1 - \sum_{i \in S} \pi_i < 0$ for $S \in \Omega$ where $\pi = (\pi_i)_{i \in V}$ are optimal dual values to constraints (7b) of the RMP. These variables/columns are added to the RMP as long as at least one negative reduced-cost variable exists. When the column-generation process terminates, a solution to the linear relaxation of (7) is found providing a lower bound to IMP. To produce integer solutions, the integration into branch-and-bound a.k.a. *branch-and-price* is required (for details see Desaulniers *et al.*, 2005). The branching scheme must finally guarantee that any fractional solution to the RMP can be cut off.

The presentation of such a branch-and-price approach is far beyond the scope of this paper at hand. Instead, our companion paper (Gschwind *et al.*, 2017) focuses on the numerous algorithmic difficulties that need to be overcome:

1. The pricing subproblem that has to be solved iteratively is a maximum-weight relaxed clique problem $\max \sum_{i \in V} \pi_i x_i$ subject to $(\mathbf{x}, \mathbf{y}) \in \mathscr{F}(G)$, where the dual solution $\pi = (\pi_i)_{i \in V}$ of the RMP defines the weights, see also Section 2.2.

   Each and every type of first-order relaxed clique requires a relaxation-specific pricing algorithm. Typically only algorithms for the maximum-cardinality problem are described in the literature, and they are often either MIP-based algorithms or combinatorial branch-and-bound (CB&B) algorithms. In the latter case, adaptations to the maximum-weight variant are non-trivial, e.g., for s-club.

   Even worse, for partitioning the graph, some weights π_i can become negative, and for non-hereditary/connected relaxed cliques the known CB&B algorithms are often not able to cope with arbitrary weights. For example, new CB&B need to be developed for connected s-clique, s-plex, s-defective clique, and s-bundle.

   Whenever more that one pricing algorithm is available, the performance of these algorithms has to be compared, using different types of graphs and parameters s or γ.

2. The development of branching schemes is non-trivial. As for pricing, branching rules strongly depend on the clique-relaxation variant, since a desirable branching rule is subproblem-structure preserving. In addition, branching rules also depend on whether a partitioning or covering decomposition has to be found.

   The well-known Ryan-Foster branching for partitioning models is not structure-preserving for relaxed clique, opposed to the situation in vertex coloring and clique partitioning where together and separate decisions can be imposed by some simple graph modifications. The companion paper (Gschwind *et al.*, 2017) compares Ryan-Foster branching for partitioning models with newly invented structure-preserving branching rules for some of the 17 variants. Moreover, a new four-level branching scheme for covering a graph with relaxed cliques is presented.

4. Interpretation of Results in Social Networks

In this section, we test the applicability of our graph decomposition methods for the purpose of detecting community structures (cf. Fortunato, 2010). We have chosen `karate`, `dolphins`, and `football` as three very prominent and intensively studied examples of real social networks for which the true community structure is known and different methods of community detection have been tested. The networks are part of the 10th DIMACS challenge available at `http://dimacs.rutgers.edu/Challenges/`. Some characteristics of these networks are listed in Table 4, such as density $\rho(G)$, minimum degree $\delta(G)$, maximum independent set size $\alpha(G)$, maximum clique size $\omega(G)$, chromatic number $\chi(G)$, chromatic number of the complement graph $\chi(\bar{G})$, and maximal modularity $\mu(G)$ computed with our implementation of the second column-generation algorithm of Aloise *et al.* (2010) (number of clusters in brackets).

The solutions presented in the following were either obtained with the generic compact formulation (6) solved with the CPLEX 12.5.0 MIP solver on a standard PC with an Intel(R) Core(TM) i7-4790 3.6 GHz processor and 8 GB of main memory using a single thread only. Note that we added constraints $z^h \geq z^{h-1}$ for $h < \bar{rc}(G)$ and $x_i^h = 0$ for $i < h$ to the compact formulation in order to break symmetries and facilitate

$G=(V,E)$	$\|V\|$	$\|E\|$	$\rho(G)$	$\delta(G)$	$\alpha(G)$	$\omega(G)$	$\chi(G)$	$\chi(\bar{G})$	$\mu(G)$
`karate`	34	78	0.1390	1	20	5	5	20	0.4198 (4)
`dolphins`	62	159	0.0841	1	28	5	5	28	0.5285 (5)
`football`	115	613	0.0935	7	21	9	9	22	0.6046 (10)

Table 4: Social networks and some of their features

the solution of the model. From the companion paper (Gschwind *et al.*, 2017) we took the results of the branch-and-price algorithm run on the same computer. Table 5 gives a brief overview over lower bounds (LB), upper bounds (UB, bold if optimum), and computation times (*Time*, in seconds). Finally, for the comparison with solutions of the partitioning problem that maximizes modularity, we implemented the branch-and-price of (Aloise *et al.*, 2010).

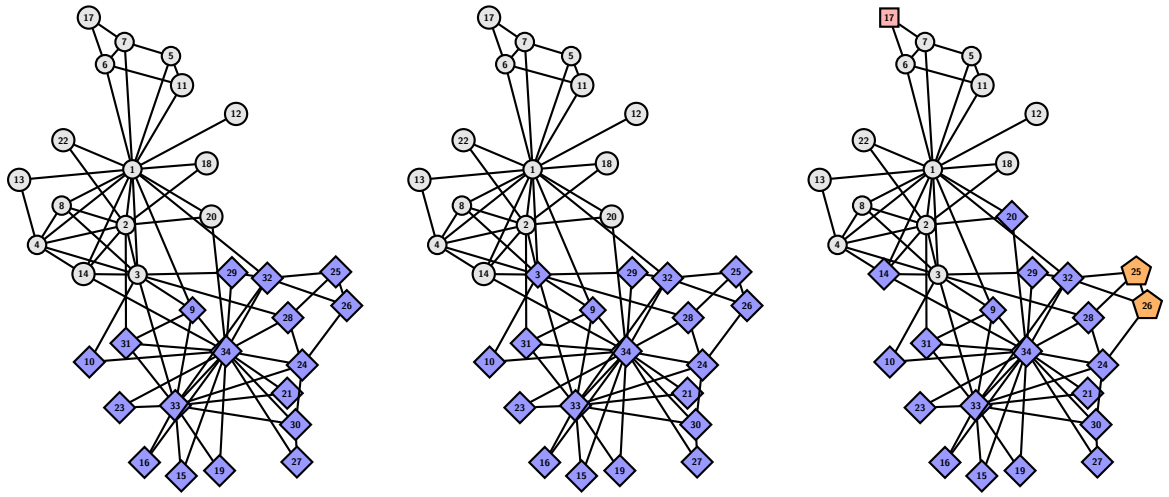
Network	clique relaxation	parameter s	decom-position	connec-tivity	compact formul. (6) with CPLEX			formul. (7) with branch-and-price	
					LB	UB	Time [s]	UB	Time [s]
`karate`	s-club	$s=2$	part.	(yes)		**4**	0.3	**4**	<0.1
		$s=3$	part.	(yes)		**2**	<0.1	**2**	1.1
`dolphins`	s-clique	$s=4$	cover.	yes		**4**	<0.1	**4**	<0.1
		$s=5$	cover.	yes		**2**	<0.1	**2**	<0.1
`football`	s-plex	$s=3$	part.	no	5	–	TL	**16**	6739.3
		$s=4$	part.	no	4	–	MEM	**13**	2802.4
		$s=4$	part.	yes	5	–	TL	**13**	4349.6
`football`	s-plex	$s=3$	part.	yes	5	–	TL	**14**	323.2
(w/o indep. teams)		$s=4$	part.	yes	5	–	TL	**12**	1704.3

Table 5: Results compact formulation with CPLEX vs. branch-and-price from (Gschwind *et al.*, 2017)
Note: MEM=out of memory, TL=time limit of 4 hours reached

4.1. Zachary's Karate Club

Zachary (1977) introduced the formal description of a university-based karate club as an example of a fission of a small anthropological group. The relevant background information is that due to a longer-lasting conflict between the club president and the karate instructor the club finally separated into two new clubs, one supporting the old club's president and the other one following the instructor. Zachary's study, however, focused on the social interaction between members before the fission. He collected the information "if two individuals consistently were observed to interact outside the normal activities of the club". The crisis in the club had the effect of "pulling apart the (sub)networks of friendship ties". The resulting social network has one vertex for each active member of the club, and two vertices are adjacent if and only if the corresponding members consistently interacted. The `Karate` club network is depicted in Figure 5(a). It became a useful benchmark for community detection approaches, since algorithmically computed clusters can be compared with the real memberships in one of the two clubs after the division. For example, Girvan and Newman (2002) applied a hierarchical clustering via tree decomposition. Their first split "corresponds almost perfectly with the actual division of the club members" with only vertex 3 being misclassified (Girvan and Newman, 2002, p. 7823).

The same authors, Newman and Girvan (2004), later introduced modularity in order to measure the quality of a decomposition (see Section 1.1). Their decomposition method (i) calculates the so-called *betweenness* for all edges of the network, (ii) removes an edges with maximum betweenness, (iii) repeats the steps (i) and (ii) for the resulting reduced network until it is edgeless. This creates a hierarchical decomposition of a graph, often displayed using a decomposition tree. With the *shortest-path betweenness*, the method produces a first decomposition into two components with vertex 3 incorrectly classified, while with



(a) (b) (c)

Figure 5: Zachary's karate club (a) Situation as described by Zachary (1977), modularity $\mu = 0.3715$; (b) 3-club partitioning, $\mu = 0.3600$; (c) 2-club partitioning, $\mu = 0.3432$. Note that the graph is a 5-club

random-walk betweenness the two groups are identified correctly. However, with both decomposition methods the clustering into five/four groups achieves a higher modularity. Recall from Table 4 that the maximum modularity is $\mu(G) = 0.4198$ (four clusters) indicating that modularity maximization can lead to many more clusters compared to the real-world community structure.

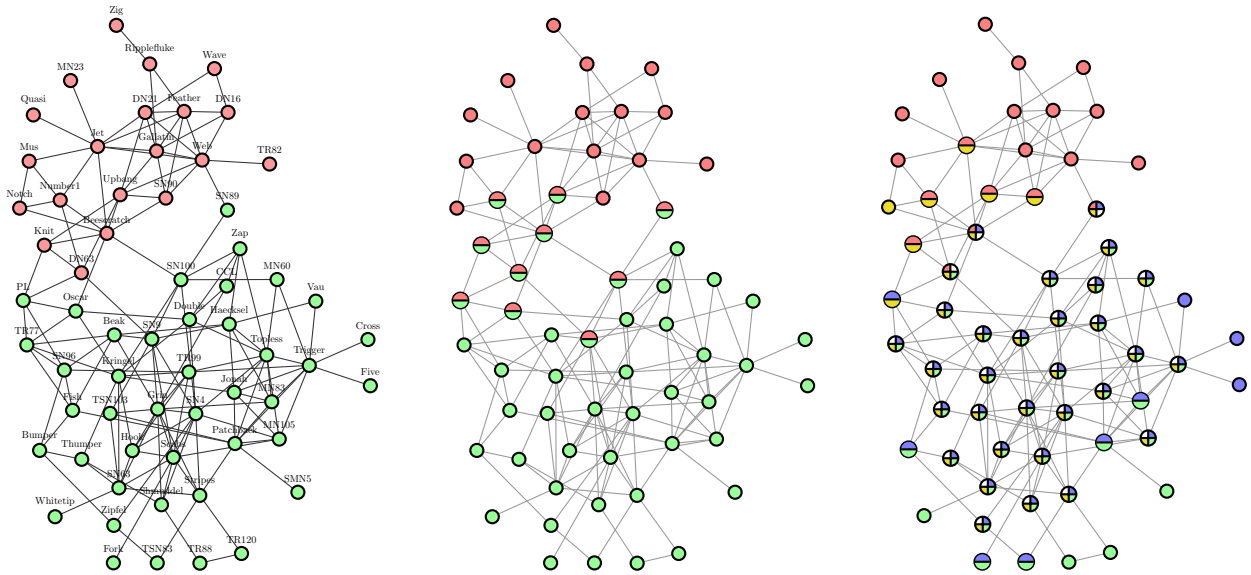
With the knowledge that the new clubs were formed around the polarizing persons that brought the conflict into the club, it seems natural to decompose the graph using a distance-based clique relaxation: The subgroups should have the property that any two members are close to a central person and, therefore, the two members must also be in close distance from each other. Moreover, the resulting subgroups should be connected. Also Almeida and Carvalho (2013) suggest the use of s-clubs in SNA arguing that "social relations are frequently established through intermediaries".

The results of a decomposition into s-clubs ($s = 2$ or 3) are shown in Figure 5(b) and (c). Note that vertex 1 is the club's president and vertex 34 is the instructor. The depicted solutions are at the same time solutions to the covering and partitioning problems. The mismatch of vertex 3 in the 3-club partitioning is actually by chance because the two real groups also form a decomposition into 3-clubs. Moreover, the decomposition into four 2-clubs as depicted in Figure 5(c) is not unique, but fits with the four clusters determined using the method of Newman and Girvan (2004) with random-walk betweenness. Furthermore, the 2-club $\{17\}$ can be enlarged to $\{5, 6, 7, 11, 17\}$, which is one of the clusters identified by Newman and Girvan (2004). The same holds for the 2-club $\{25, 26\}$, which can be extended to $\{25, 26, 29, 32\}$ without making the other 2-clubs infeasible in the resulting partitioning.

4.2. Dolphins

We consider a network of 62 bottlenose dolphins living in Doubtful Sound (New Zealand). Lusseau (2003) defined the edges of the network as indicators of "preferred companionships" meaning that pairs of dolphins were seen together more often than expected by chance. Figure 6(a) depicts the `dolphins` network. After dolphin *SN100* left the place for some time, the dolphins separated into two groups (Lusseau and Newman, 2004) indicated with the two colors.

Similar to the karate club, the dolphin network is an example of a social network in which fission and fusion was observed. As the connection between the dolphins is rather lose, a distance-based clique relaxation seems appropriate for a decomposition. Moreover, the network is larger but less dense compared to the karate club (see Table 4) so that we have chosen larger values of the maximum distance s. As Fortunato (2010,



(a) (b) (c)

Figure 6: Dolphins (a) Real split (b) Covering with connected 5-cliques, two clusters (c) Covering connected 4-cliques, four clusters

Section 11) points out, the identification of overlapping clusters is an important task in community detection. We now show that interesting and interpretable results can be obtained with our graph covering algorithms.

Figure 6(b) shows a decomposition of the dolphins network into connected 5-cliques that are allowed to overlap. Two communities result and are indicated by the red and green colors. The dolphins displayed with bicolored vertices are those that belong to both communities. To be precise, we computed a (non-unique) covering solution and extended each of the two communities to the depicted cardinality-maximal 5-cliques.

Obviously, the result shown in Figure 6(b) perfectly matches the real split into two communities as described in (Lusseau and Newman, 2004). Moreover, our intersection that consists of ten dolphins includes the five dolphins *DN63*, *Knit*, *Oscar*, *PL*, and *SN89* that Lancichinetti *et al.* (2009) identify as members of both groups. Their method is a greedy algorithm, where in an outer loop a single uncovered vertex is randomly chosen and in an inner loop a cluster containing this vertex is determined by maximizing a fitness function.

Finally, we reduced the maximum distance to $s = 4$. The result is four overlapping clusters as depicted in Figure 6(c). Note that no vertex belongs to all four clusters, i.e., vertices are either monochrom, bicolored, or three-colored. Interestingly, Girvan and Newman (2002) also find four communities with their algorithm. Lusseau and Newman (2004) argue that Girvan and Newman (2002) found a natural decomposition of the larger community (green vertices in Figure 6(a)) into three sub-communities, where this subdivision is correlated with the gender and age of the dolphins. In comparison, our depicted decomposition consists of slightly larger clusters, but reflects well that three sub-communities can be identified in the larger community.

4.3. Football

Girvan and Newman (2002) introduced another social network in which the vertices are American Football college teams and edges represent regular-season games between them. The 115 teams are divided into eleven conferences containing between six and 13 teams each. Generally, teams play more intraconference than interconference games so that conferences form clusters. Moreover, there are eight independent teams that do not belong to a specific conference. Their game schedule is less structured than for the conference teams meaning that games among independent teams are as likely as games between independent teams and conference teams. Overall, the interconference games are not uniformly distributed because games between geographically close teams are more frequent. The `football` network is depicted in Figure 7(a).

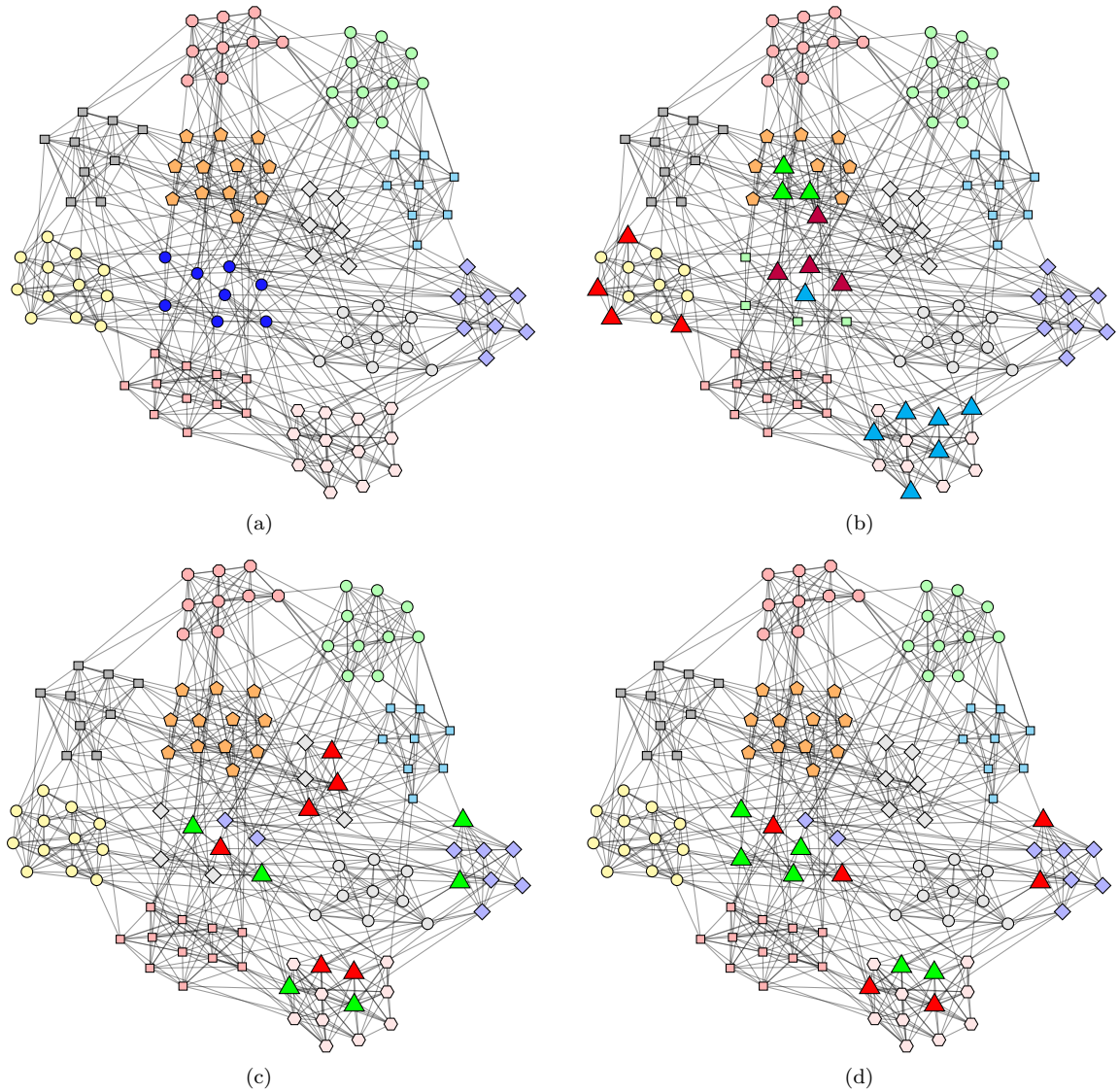


Figure 7: Football (a) Eleven Conferences and Independent Teams (blue •), modularity $\mu = 0.5877$; (b) Partitioning with 3-plexes, 16 clusters, $\mu = 0.4958$; (c) Partitioning with 4-plexes, 13 clusters, $\mu = 0.5345$; (d) Partitioning with connected 4-plexes, 13 clusters, $\mu = 0.5508$

It is important to mention that the instance as provided on Marc Newman's webpage (`http://www-personal.umich.edu/~mejn/netdata/`) incorrectly assigns seven teams to conferences. To be precise, *Boise State* and *Utah State* belong to the conference *Sun Belt (Big West)*, *Texas Christian* belongs to the conference *Western Athletic*, and *Louisiana Tech*, *Louisiana Monroe*, *Middle Tennessee State*, and *Louisiana Lafayette* are independent teams. We used `https://en.wikipedia.org/wiki/2000_NCAA_Division_I-A_football_season` and `http://www.phys.utk.edu/sorensen/cfr/cfr/Output/2000/CF_2000_Main.html` as independent sources. The consequence is that several works base their presentation on an incorrect reference solution (e.g., Girvan and Newman, 2002; Zhou, 2003a,b). However, the 613 games as given by Marc Newman seem to reflect the schedule of the 2000 season. Note that Evans (2010), independently finds similar inconsistencies in this data.

The hierarchical decomposition methods used by Girvan and Newman (2002) and Zhou (2003a,b) provide many possible clusterings but do not answer the question what the "best" number of clusters is. Later, with the definition of modularity, Newman and Girvan (2004) made it possible to assess the quality of different decompositions. With the objective of modularity maximization, Aloise *et al.* (2010) find that a clustering into ten groups is optimal. As can be seen from Figure 11 of Section 4.4, the clustering with maximum modularity cannot discriminate between the smallest conference (*Big West*, 6 teams, gray diamonds) and *Mountain West* (8 teams, light blue boxes), however, team assignments are generally correct. On the contrary, all of our solutions presented next do not combine the two conferences.

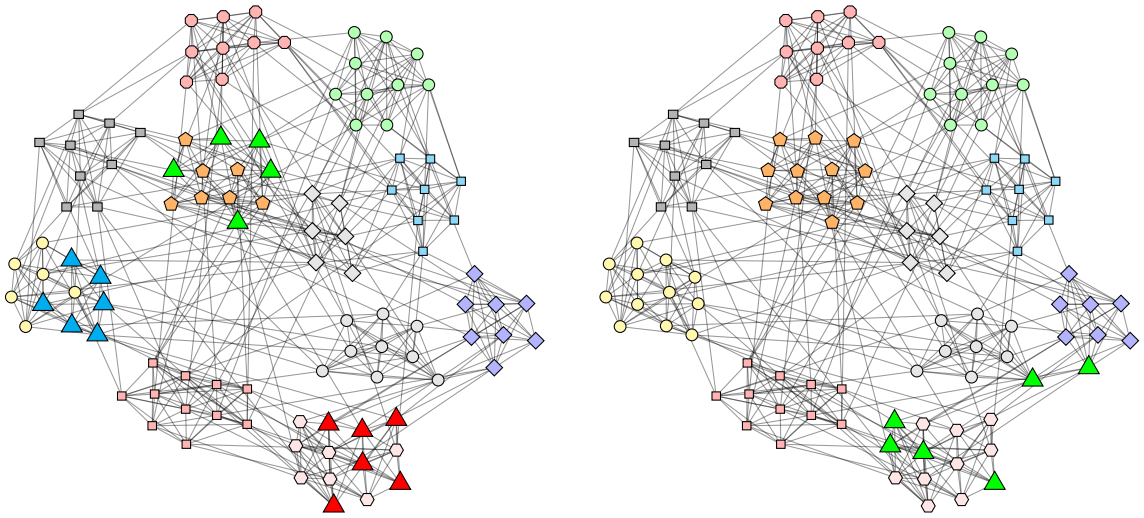
For decomposing the `football` network into relaxed cliques, we expect that conferences are well represented by s-plexes because games within the same conference are predominant. More precisely, with the corrected conference assignment, there are 414 intra-conference games (including 10 games among the independent teams) and 199 inter-conference games. On average, there are ten teams per conference and each team plays seven games within its conference. Thus, the average conference constitutes a 3-plex ($s = 10 - 7$). Due to the above mentioned irregularities, also larger values of s make sense.

We start our analysis of the `football` network with a partitioning into 3-plexes. Figure 7(b) shows that in this case the minimum number of partitions is 16. While eight of the eleven conferences are detected, the remaining three are structured into two or three groups that also contain some of the independent teams. The split conferences are the three largest conferences which are actually subdivided into two divisions of six or seven teams each. We later see that good decompositions can uncover this type of substructure.

In order to better meet the correct number of conferences, we partition the network using 4-plexes, see Figure 7(c). Also here three conferences are mixed with independent teams. However, only nine conference teams are misclassified compared to 14 conference teams in the 3-plex solution. Even with this improvement, the solution has the defect that one cluster is disconnected (teams depicted as red triangles). We therefore impose connectivity. The resulting partitioning into connected 4-plexes is shown in Figure 7(d). The number of clusters does not increase compared to the disconnected solution (13 groups). Now, one more conference is correctly detected.

In the three solutions given in Figure 7(b)–(d), the independent teams are assigned very differently. This may be an indication that the independent teams do not form a community encoded by the graph. In a series of additional experiments, we therefore removed the independent teams from the network. The new network consists of 107 vertices and 551 edges, as shown in Figure 7(a) but with the independent teams removed (dark blue circles). For the sake of brevity, we omit the explicit depiction of the new network.

We present the partitioning with connected 3-plexes and connected 4-plexes in Figure 8(a) and (b). The 3-plex partitioning consists of 14 clusters. Eight of them perfectly reproduce the smaller conferences, while three pairs of the remaining six clusters exactly form the three largest conferences. Recall that these three conferences do have subdivisions in reality. The 4-plex partitioning identifies twelve groups, one more than there are conferences. However, this is the best solution in the sense that only six teams are misclassified. The conference with the most mismatches is *Mid American* (depicted with rose hexagons). It is the only conference with 13 teams and it does not form an s-plex for $s \leq 5$ because four teams have a degree of seven. Thus, the real conferences do not form a feasible partitioning into 5-plexes. If we run our algorithm for partitioning with 5-plexes, the solution perfectly matches the correct number of eleven conferences, but it groups teams of four conferences incorrectly.



(a) (b)

Figure 8: Football without Independent Teams with maximum modularity $\mu(G) = 0.6392$ (ten clusters) and modularity of the real-world solution of $\mu = 0.6381$ with eleven clusters (a) Partitioning into connected 3-plexes, 14 clusters, $\mu = 0.5301$; (b) Partitioning in connected 4-plexes, twelve clusters, $\mu = 0.5850$

4.4. Solutions from Modularity Maximization

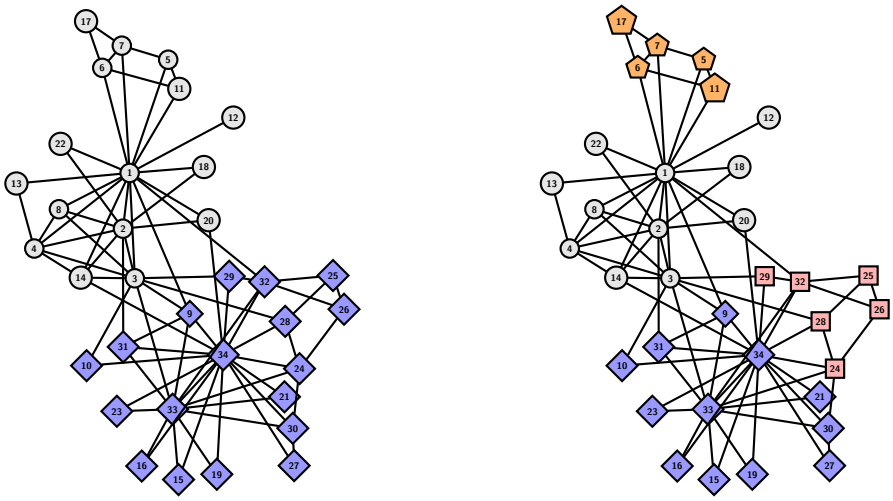
We have implemented the second column-generation algorithm of Aloise *et al.* (2010) to obtain the clusterings resulting from modularity maximization. These results are compared with the real-world community structures. Figure 9 shows the comparison for the network `Karate`, Figure 10 for `Dolphins`, and Figure 11 for `Football`.

5. Conclusions

In this paper, we have introduced the problem of decomposing a graph into a minimum number of relaxed cliques as a new method for community detection. While in prior works the resulting clusters generally do not have any structure, the different clique relaxations allow to impose application-specific constraints a cluster has to fulfill. Using the eight types of first-order clique relaxations as defined by Pattillo *et al.* (2013a), we identified 17 new relevant types of decompositions with first-order relaxed cliques. In particular, for non-hereditary relaxed cliques one must distinguish between partitioning and covering the network. Moreover, since a basic requirement for communities is connectivity, we have introduced the concept of connected relaxed cliques. As a consequence, decomposing into connected or general relaxed cliques gives rise to different problem variants. Our type of approach is useful in cases where one has a good understanding of what defines a community. For three prominent examples from social network analysis, we have demonstrated that decomposition into relaxed cliques reproduces some known features of the networks.

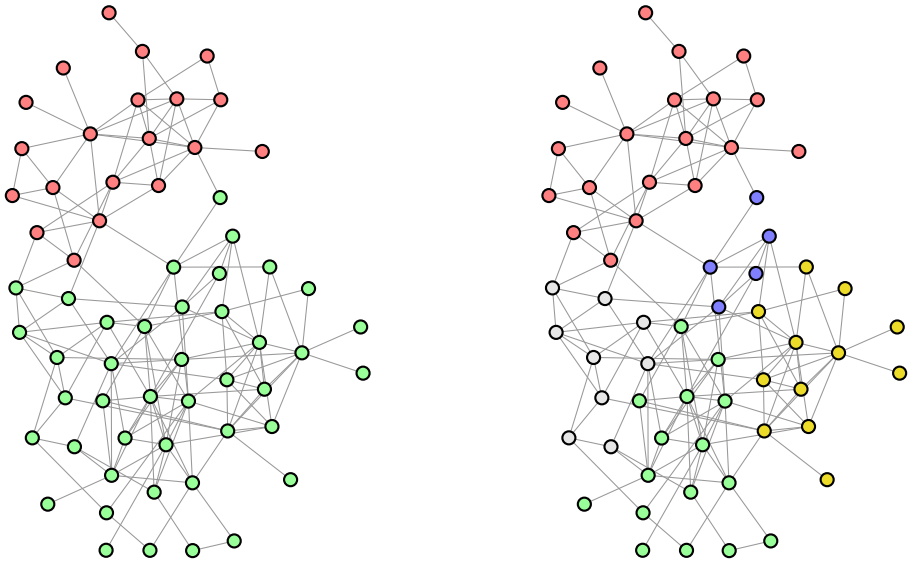
Modularity maximization is the predominant method in community detection to assess the quality of a clustering. Our decomposition approach is independent of modularity and might be a valid alternative to overcome the limitations of modularity maximization as discussed by Fortunato and Barthélemy (2007). They prove that modularity maximization can incorrectly identify clusters in some cases. For example, for a network composed of the union of sufficiently large cliques K_n arranged in a cycle, modularity maximization joins pairs of K_n. Our (relaxed) clique partitioning approach would correctly identify each K_n as a single cluster for reasonable choices of s or γ.

From an optimization point of view, decomposing into relaxed cliques is a hard problem. A brief computational analysis has shown that applying a commercial MIP solver to a standard compact formulation of the problem is not a viable approach. Instead, tailored solution algorithms are needed to tackle at least medium-sized instances. Our companion paper (Gschwind *et al.*, 2017) proposes a branch-and-price framework that



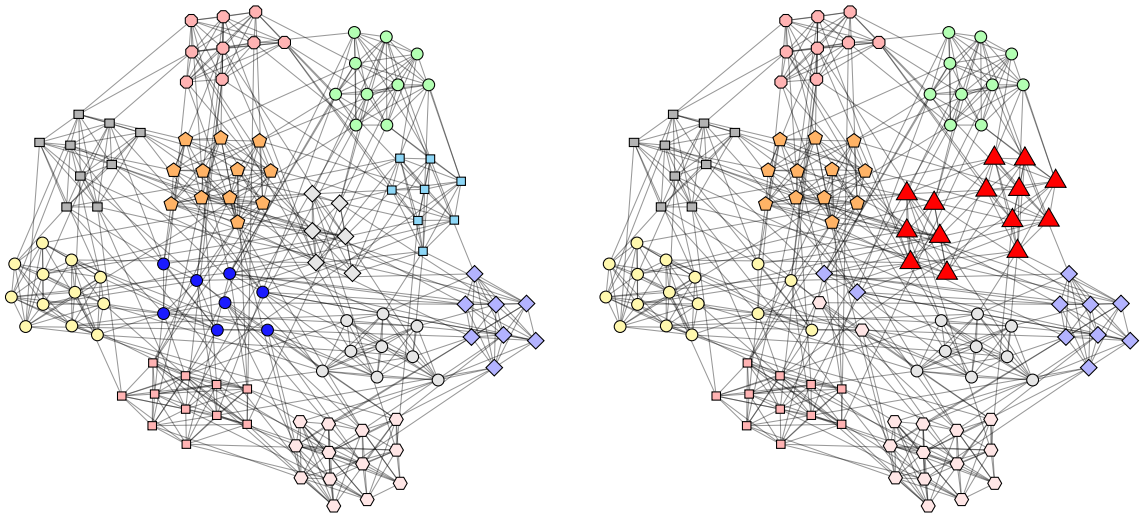
(a) (b)

Figure 9: Zachary's karate club (a) Real-world split, modularity $\mu(G) = 0.3715$; (b) Decomposition with maximum modularity $\mu = 0.4198$



(a) (b)

Figure 10: Dolphins (a) Real-world split, modularity $\mu(G) = 0.3735$; (b) Decomposition with maximum modularity $\mu = 0.5285$



(a) (b)

Figure 11: Football (a) Real-world split, modularity $\mu(G) = 0.5877$; (b) Decomposition with maximum modularity $\mu = 0.6046$

provides promising results for the exact solution of the different decomposition problems on some medium-sized networks. Clearly, large-scale networks require heuristic and metaheuristic solution approaches.

There is also room for alternative structures that define the clusters, e.g., additional types of relaxed cliques such as second-order relaxed cliques and k-connected/k-hereditary relaxed cliques (see Pattillo *et al.*, 2013a). Alternatively, two or more different types of relaxed cliques can be allowed meaning that a cluster can, e.g., either be a 2-plex or a 5-defective clique. Moreover, new relaxed clique definitions result when additional attributes are associated with vertices or edges. An example is a distance-d-clique defined as a set of vertices with a pairwise distance not exceeding d (measured by the sum of edge distances d_{ij}). Also, the overall objective of minimizing the number of clusters can be replaced by one in which the clusters receive a weight, e.g., computed as a function (maximum, sum, average, or product) of its vertex and edge weights.

Appendix

A. Strength of LP Bounds of the Compact Decomposition Model

Proof of Theorem 1:
(a) For every solution $(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}})$ to the linear relaxation of formulation (6), we can arbitrarily permute the indices $h \in H$. The resulting solution is then also a solution to the linear relaxation. Since any convex combination of solutions is again a solution to the linear relaxation, it follows the uniform convex combination over all permutations gives $x_i^1 = x_i^2 = \cdots = x_i^{\bar{r}c(G)}$ for all $i \in V$, $y_{ij}^1 = y_{ij}^2 = \cdots = y_{ij}^{\bar{r}c(G)}$ for all $\{i,j\} \in E$, and $z^1 = z^2 = \cdots = z^{\bar{r}c(G)}$.
(b)+(c) We now prove that $x_i^h = \frac{1}{\bar{r}c(G)}$ for all $i \in V, h \in H$ is feasible. Note first that for $\bar{r}c(G) = 1$ the entire graph G is a relaxed clique of the considered type. Therefore, there exists a feasible solution in $\mathscr{F}(G)$ with $x_i = 1$ for all $i \in V$, which proves the theorem in this case.

From now on, we assume that $\bar{r}c(G) \geq 2$ holds. We consider the polytopes one by one and show first that setting $x_i = \frac{1}{\bar{r}c(G)}$ is feasible for every polytope. For the remainder of the proof, we skip the superscript h for simplicity.
Case $\mathscr{F}^1(G)$ (clique and s-clique): The only constraints are $x_i + x_j \leq 1$ for $i, j \in V, i < j$ with $\{i,j\} \notin E$, which are fulfilled for $x_i = \frac{1}{\bar{r}c(G)} \leq \frac{1}{2}$.
Case $\mathscr{F}^2(G)$ (s-plex): Recall the definition of the number $\bar{d}_i = |V \setminus N(i)| - 1$ for all $i \in V$. The only constraints are $\sum_{j \in V \setminus N(i), j \neq i} x_j \leq (s-1)x_i + \bar{d}_i(1 - x_i)$ for all $i \in V$. Inserting $x_i = \frac{1}{\bar{r}c(G)}$ gives the value

$\frac{\bar{d}_i}{\bar{rc}(G)}$ of the left-hand side. The right-hand side is

$$\frac{s-1}{\bar{rc}(G)} + \frac{\bar{d}_i(\bar{rc}(G)-1)}{\bar{rc}(G)} = \underbrace{\frac{s-1}{\bar{rc}(G)}}_{\geq 0} + \frac{\bar{d}_i}{\bar{rc}(G)} \underbrace{(\bar{rc}(G)-1)}_{\geq 1} \geq \frac{\bar{d}_i}{\bar{rc}(G)},$$

proving the statement.

Case $\mathscr{F}^3(G)$ (s-defective clique): Recall that $\bar{G} = (V, \bar{E})$ is the complement graph of G and that there are non-negative variables $\bar{y}_{ij} \geq 0$ for all $\{i,j\} \in \bar{E}$. The constraints of the polytope are given by $\bar{y}_{ij} \geq x_i + x_j - 1$ for all $\{i,j\} \in \bar{E}$ and $\sum_{\{i,j\}\in\bar{E}} \bar{y}_{ij} \leq s$. Setting $\bar{y}_{ij} = 0$ for all $\{i,j\} \in \bar{E}$ and $x_i = \frac{1}{\bar{rc}(G)}$ is clearly feasible.

Case $\mathscr{F}^4(G)$ (γ-quasi-clique): Recall that there are additional variables $y_{ij} \geq 0$ for $i,j \in V, i < j$ and that (a_{ij}) is the adjacency matrix of G. The polytope is described by $\sum_{i<j}(\gamma - a_{ij})y_{ij} \leq 0$, $y_{ij} \leq x_i$, $y_{ij} \leq x_j$, and $y_{ij} \geq x_i + x_j - 1$ for all $i,j \in V, i < j$. Setting $y_{ij} = 0$ for all $i,j \in V, i < j$ and $x_i = \frac{1}{\bar{rc}(G)}$ is feasible, proving the statement.

Case $\mathscr{F}^5(G)$ (γ-quasi-clique): First, note that there are additional variables y_i for all $i \in V$ and recall that (a_{ij}) is the adjacency matrix of G. The constraints of the polytope are (cf. Pattillo *et al.*, 2013b):

$$\sum_{i\in V} y_i \geq 0 \tag{8a}$$

$$y_i \leq (1-\gamma)\sum_{j\in V} a_{ij}x_i \qquad i \in V \tag{8b}$$

$$y_i \geq -\left(n - 1 - \sum_{j\in V} a_{ij}\right)\gamma x_i \qquad i \in V \tag{8c}$$

$$y_i \geq \gamma x_i + \sum_{j\in V}(a_{ij}-\gamma)x_j - (1-\gamma)\sum_{j\in V} a_{ij}(1-x_i) \qquad i \in V \tag{8d}$$

$$y_i \leq \gamma x_i + \sum_{j\in V}(a_{ij}-\gamma)x_j + \left(n - i - \sum_{j\in V} a_{ij}\right)\gamma(1-x_i) \qquad i \in V \tag{8e}$$

With $x_i = \frac{1}{\bar{rc}(G)}$ and $y_i = 0$ for all $i \in V$ the first three constraints (8a)–(8c) are clearly fulfilled. To show that also the latter two are satisfied define $d_i = \sum_{j\in V} a_{ij}$ for all $i \in V$ for convenience. The right-hand sides of constraints (8d) then write

$$\begin{aligned} & \frac{\gamma}{\bar{rc}(G)} + \frac{d_i}{\bar{rc}(G)} - \frac{n\gamma}{\bar{rc}(G)} - (1-\gamma)d_i\left(1 - \frac{1}{\bar{rc}(G)}\right) \\ = \; & \frac{\gamma}{\bar{rc}(G)} - \frac{n\gamma}{\bar{rc}(G)} + d_i \underbrace{\left(\frac{1}{\bar{rc}(G)} + (\gamma-1)\left(1 - \frac{1}{\bar{rc}(G)}\right)\right)}_{=:\,\alpha}. \end{aligned}$$

If $\alpha < 0$, then the expression is maximized by choosing the smallest possible value for d_i, i.e., $d_i = 0$. Because $\frac{\gamma}{\bar{rc}(G)} - \frac{n\gamma}{\bar{rc}(G)} \leq 0$, the constraint is clearly satisfied in this case. In the case $\alpha \geq 0$, the right-hand side is maximized by setting d_i to its largest value $d_i = n - 1$ resulting in

$$\begin{aligned} & \frac{\gamma}{\bar{rc}(G)} + \frac{n-1}{\bar{rc}(G)} - \frac{n\gamma}{\bar{rc}(G)} - (1-\gamma)(n-1)\left(1 - \frac{1}{\bar{rc}(G)}\right) \\ = \; & (\gamma + n - 1 - n\gamma)\frac{1}{\bar{rc}(G)} - (1-\gamma)(n-1)\left(1 - \frac{1}{\bar{rc}(G)}\right) \\ = \; & (n-1)(1-\gamma)\frac{1}{\bar{rc}(G)} - (1-\gamma)(n-1)\left(1 - \frac{1}{\bar{rc}(G)}\right) \end{aligned}$$

$$
\begin{aligned}
&= \underbrace{(1-\gamma)}_{\geq 0}\underbrace{(n-1)}_{\geq 0}\underbrace{\left(\frac{2}{\bar{rc}(G)}-1\right)}_{\leq 0} \\
&\leq 0
\end{aligned}
$$

showing that $x_i = \frac{1}{\bar{rc}(G)}$ and $y_i = 0$ fulfill constraints (8d).

The right-hand sides of constraints (8e) write

$$
\begin{aligned}
&\frac{\gamma}{\bar{rc}(G)} + \frac{d_i}{\bar{rc}(G)} - \frac{n\gamma}{\bar{rc}(G)} + (n-1-d_i)\gamma\left(1-\frac{1}{\bar{rc}(G)}\right) \\
= &\underbrace{\frac{\gamma}{\bar{rc}(G)} - \frac{n\gamma}{\bar{rc}(G)} + (n-1)\gamma(1-\frac{1}{\bar{rc}(G)})}_{(n-1)\gamma\left(1-\frac{2}{\bar{rc}(G)}\right)} + d_i \underbrace{\left(\frac{1+\gamma}{\bar{rc}(G)}-\gamma\right)}_{=:\,\beta}
\end{aligned}
$$

The case $\beta \geq 0$ means that $d_i = 0$ minimizes the right-hand side and because $(n-1)\gamma(1-\frac{2}{\bar{rc}(G)}) \geq 0$ the constraint is satisfied. In the case $\beta < 0$, the value $d_i = n-1$ has to be verified resulting in a right-hand side of

$$
\begin{aligned}
&\frac{\gamma}{\bar{rc}(G)} + \frac{n-1}{\bar{rc}(G)} - \frac{n\gamma}{\bar{rc}(G)} + (n-1-n-1)\gamma\left(1-\frac{1}{\bar{rc}(G)}\right) \\
= &\frac{(n-1)(1-\gamma)}{\bar{rc}(G)} \\
\geq &\ 0
\end{aligned}
$$

which completes the proof.

Case $\mathscr{F}^6(G)$ (γ-quasi-clique): We set $x_i = \frac{1}{\bar{rc}(G)}$ for all $i \in V$ and set $y_e = \frac{1}{\bar{rc}(G)}$ for all $e \in E$. With this inequalities (2b) are fulfill with equality. Now we consider two cases.

If $\frac{|V|}{\bar{rc}(G)}$ is integer, we define $\bar{s}$ as this integer number and set $t_{\bar{s}} = 1$ and $t_s = 0$ for all other $s \in \mathcal{S} \setminus \{\bar{s}\}$. Obviously, we have now $\sum_{s\in\mathcal{S}} t_s = 1$ and $\sum_{s\in\mathcal{S}} s t_s = \bar{s} = \frac{|V|}{\bar{rc}(G)}$ so that (2e) and (2d) are fulfilled. It remains to show that (2c) is fulfilled. We have

$$
\sum_{e\in E} y_e = \frac{1}{\bar{rc}(G)}\cdot|E| \geq \frac{1}{\bar{rc}(G)}\cdot\sum_{p=1}^{\bar{rc}(G)}|E(V_i)| \geq \frac{1}{\bar{rc}(G)}\sum_{p=1}^{\bar{rc}(G)}\gamma\frac{|V_i|(|V_i|-1)}{2} = \gamma\cdot\frac{1}{\bar{rc}(G)}\sum_{p=1}^{\bar{rc}(G)}\frac{|V_i|(|V_i|-1)}{2}
$$

for any partitioning $V = V_1 \cup V_2 \cup \cdots \cup V_{\bar{rc}(G)}$ into $\bar{rc}(G)$-many γ-quasi-cliques. As $\bar{rc}(G)$ is an upper bound, we know that such a feasible partitioning exists. Now, the RHS in the above inequality can be interpreted as γ times the average lower bound on the size of a γ-quasi-clique in the partitioning. It is straightforward to proof that this latter value is minimal if all terms in the average are equal. Equal terms implies $|V_i| = \frac{|V|}{\bar{rc}(G)}$. Hence, the value $|V_i| = \bar{s} = \sum_{s\in\mathcal{S}} s t_s$, which proves that (2c) is fulfilled.

If $\frac{|V|}{\bar{rc}(G)}$ is fractional, we can no longer define a unique $\bar{s}$. Instead we define $\underline{s} = \lfloor\frac{|V|}{\bar{rc}(G)}\rfloor$ and $\bar{s} = \underline{s}+1 = \lceil\frac{|V|}{\bar{rc}(G)}\rceil$. Moreover, with

$$
t_{\underline{s}} = \bar{s} - \frac{|V|}{\bar{rc}(G)}
$$

and $t_{\bar{s}} = 1 - t_{\underline{s}}$, we again obtain $\sum_{s\in\mathcal{S}} t_s = 1$ and $\sum_{s\in\mathcal{S}} s t_s = \frac{|V|}{\bar{rc}(G)}$ so that (2e) and (2d) are fulfilled. Note that because $s(s-1)/2$ is a convex function in s, the convex combination $t_{\underline{s}}\cdot\frac{\underline{s}(\underline{s}-1)}{2} + (1-t_{\underline{s}})\cdot\frac{\bar{s}(\bar{s}-1)}{2}$ is not smaller than the function at the respective intermediate point between $\underline{s}$ and $\bar{s}$, in particular for $\frac{|V|}{\bar{rc}(G)}$. The remainder of the proof can follow the same line of arguments as in the first case.

Case $\mathscr{F}^7(G)$ (s-club): The set of all unordered pairs is $\mathcal{U} = \{\{i,j\} : i,j \in V, i < j\}$ and the index set for paths lengths is $dom_2(i,j) = \{\max\{2, \mathrm{dist}_G(i,j)\}, \ldots, s\}$. There are additional variables v_{ij}^ℓ for $\{i,j\} \in \mathcal{U}$ and $\ell \in dom_2(i,j)$. Similar to γ-quasi-clique, setting $v_{ij}^\ell = 0$ for all $\{i,j\} \in \mathcal{U}$ and $\ell \in dom_2(i,j)$ and $x_i = \frac{1}{\bar{r}c(G)}$ is feasible for constraints (3b) and (3c). The remaining constraints (3d) and (3e) are obviously also fulfilled.

Case $\mathscr{F}^8(G)$ (s-bundle): Recall that the auxiliary network $\mathcal{N} = (N, A)$ associated with G has, for each vertex $i \in V$, two vertices i^- and i^+ in $\mathcal{N}$. The arc set A is $\{(i^-, i^+) : i \in V\} \cup \{(i^+, j^-), (j^+, i^-) : \{i,j\} \in E\}$. As before, $\mathcal{U} = \{\{i,j\} : i,j \in V, i < j\}$.

There are additional variables $u \geq 0$ and y_a^{ij} for each $a \in A$ and $\{i,j\} \in \mathcal{U} \setminus E$. Inserting $x_i = \frac{1}{\bar{r}c(G)}$ into constraint (4b) gives a lower bound on u. Accordingly, we set $u = (\frac{n}{\bar{r}c(G)} - s)^+$.

The right-hand side of constraints (4c) and (4e) is $u - M^{ij}(2 - x_i - x_j)$, which is at least $\frac{n}{\bar{r}c(G)} - s - M^{ij}$. Herein, $M^{ij} + s$ is an upper bound on the size of an s-bundle in $G \setminus \{i\}$ or $G \setminus \{j\}$. Since $\frac{n}{\bar{r}c(G)}$ is the average size of an s-bundle in a decomposition we have $\frac{n}{\bar{r}c(G)} - s \leq M^{ij}$. Consequently, the right-hand sides do not exceed zero so that constraints (4c) and (4e) are fulfilled by setting $y_a^{ij} = 0$ for all $a \in A$ and $\{i,j\} \in \mathcal{U} \setminus E$. The remaining constraints (4d) and (4f) are trivially satisfied.

In summary, setting $x_i^1 = x_i^2 = \cdots = x_i^{\bar{r}c(G)} = \frac{1}{\bar{r}c(G)}$ for each $i \in V$ and $z^1 = z^2 = \cdots = z^{\bar{r}c(G)} = \frac{1}{\bar{r}c(G)}$ is a feasible solution to the linear relaxation of the decomposition problem (6). Note that the coupling constraints (6c) are fulfilled with equality. Since the solution is feasible for partitioning it is also feasible for covering.

The described solution has objective value 1. Summing (6c) over $h \in H$ we see $lp(G) \geq \sum_{h \in H} z^h \geq \sum_{h \in H} x_i^h \geq 1$, where the last inequality results from (6b). □

References

Abello, J., Pardalos, P., and Resende, M. G. C. (1999). On maximum clique problems in very large graphs. In J. M. Abello and J. S. Vitter, editors, *External Memory Algorithms and Visualization. DIMACS Series on Discrete Mathematics and Theoretical Computer Science*, pages 119–130. American Mathematical Society, Boston, MA.

Almeida, M. T. and Carvalho, F. D. (2012). Integer models and upper bounds for the 3-club problem. *Networks*, **60**(3), 155–166.

Almeida, M. T. and Carvalho, F. D. (2013). An analytical comparison of the LP relaxations of integer models for the k-club problem. *European Journal of Operational Research*, **232**(3), 489–498.

Aloise, D., Cafieri, S., Caporossi, G., Hansen, P., Perron, S., and Liberti, L. (2010). Column generation algorithms for exact modularity maximization in networks. *Physical Review E*, **82**(4), 046112.

Balasundaram, B., Butenko, S., and Hicks, I. V. (2011). Clique relaxations in social network analysis: The maximum k-plex problem. *Operations Research*, **59**(1), 133–142.

Bourjolly, J.-M., Laporte, G., and Pesant, G. (2000). Heuristics for finding k-clubs in an undirected graph. *Computers & Operations Research*, **27**(6), 559–569.

Bourjolly, J.-M., Laporte, G., and Pesant, G. (2002). An exact algorithm for the maximum k-club problem in an undirected graph. *European Journal of Operational Research*, **138**(1), 21–28.

Buluç, A., Meyerhenke, H., Safro, I., Sanders, P., and Schulz, C. (2013). Recent advances in graph partitioning. *CoRR*, **abs/1311.3144**.

Carraghan, R. and Pardalos, P. M. (1990). An exact algorithm for the maximum clique problem. *Operations Research Letters*, **9**(6), 375–382.

Carvalho, F. D. and Almeida, M. T. (2011). Upper bounds and heuristics for the 2-club problem. *European Journal of Operational Research*, **210**(3), 489–494.

Chang, J. M., Yang, J. S., and Peng, S. L. (2014). On the complexity of graph clustering with bounded diameter. In *2014 International Computer Science and Engineering Conference (ICSEC)*, pages 18–22.

Cook, V. J., Sun, S. J., Tapia, J., Muth, S. Q., Argüello, D. F., Lewis, B. L., Rothenberg, R. B., and McElroy, P. D. (2007). Transmission network analysis in tuberculosis contact investigations. *Journal of Infectious Diseases*, **196**(10), 1517–1527.

Desaulniers, G., Desrosiers, J., and Solomon, M., editors (2005). *Column Generation*. Springer, New York, NY.

Evans, T. S. (2010). Clique graphs and overlapping communities. *Journal of Statistical Mechanics: Theory and Experiment*, **2010**(12), P12037.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, **486**(3–5), 75–174.

Fortunato, S. and Barthélemy, M. (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, **104**(1), 36–41.

Garey, M. and Johnson, D. (1979). *Computers and Intractability*. Freeman, New York.

Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, **99**(12), 7821–7826.

Gschwind, T., Irnich, S., Furini, F., and Wolfler Calvo, R. (2017). A branch-and-price framework for decomposing graphs into relaxed cliques. Technical Report LM-2017-07, Chair of Logistics Management, Gutenberg School of Management and Economics, Johannes Gutenberg University Mainz, Mainz, Germany.

Gschwind, T., Irnich, S., and Podlinski, I. (2018). Maximum weight relaxed cliques and Russian doll search revisited. *Discrete Applied Mathematics*, **234**, 131–138.

Guo, J., Komusiewicz, C., Niedermeier, R., and Uhlmann, J. (2010). A more relaxed model for graph-based data clustering: s-plex cluster editing. *SIAM Journal on Discrete Mathematics*, **24**(4), 1662–1683.

Held, S., Cook, W., and Sewell, E. C. (2012). Maximum-weight stable sets and safe lower bounds for graph coloring. *Mathematical Programming Computation*, **4**(4), 363–381.

Kammer, F. and Täubig, H. (2005). Connectivity. In U. Brandes and T. Erlebach, editors, *Network Analysis: Methodological Foundations [outcome of a Dagstuhl seminar, 13-16 April 2004]*, volume 3418 of *Lecture Notes in Computer Science*, pages 143–177. Springer.

Kosub, S. (2004). Local density. In U. Brandes and T. Erlebach, editors, *Network Analysis: Methodological Foundations [outcome of a Dagstuhl seminar, 13-16 April 2004]*, volume 3418 of *Lecture Notes in Computer Science*, pages 112–142. Springer.

Lancichinetti, A., Fortunato, S., and Kertész, J. (2009). Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, **11**(3), 033015.

Lübbecke, M. and Desrosiers, J. (2005). Selected topics in column generation. *Operations Research*, **53**(6), 1007–1023.

Lusseau, D. (2003). The emergent properties of a dolphin social network. *Proceedings of the Royal Society B: Biological Sciences*, **270**(Suppl_2), S186–S188.

Lusseau, D. and Newman, M. E. J. (2004). Identifying the role that animals play in their social networks. *Proceedings of the Royal Society B: Biological Sciences*, **271**(Suppl_6), S477–S481.

Mahdavi Pajouh, F. and Balasundaram, B. (2012). On inclusionwise maximal and maximum cardinality k-clubs in graphs. *Discrete Optimization*, **9**(2), 84–97.

Mahdavi Pajouh, F., Balasundaram, B., and Hicks, I. V. (2016). On the 2-club polytope of graphs. *Operations Research*.

McCormick, G. P. (1976). Computability of global solutions to factorable nonconvex programs: Part I – convex underestimating problems. *Mathematical Programming*, **10**(1), 147–175.

Mehrotra, A. and Trick, M. (1998). Cliques and clustering: A combinatorial approach. *Operations Research Letters*, **22**, 1–12.

Menger, K. (1927). Zur allgemeinen Kurventheorie. *Fund. Math.*, **10**, 96–115.

Moradi, E. and Balasundaram, B. (2015). Finding a maximum k-club using the k-clique formulation and canonical hypercube cuts. *Optimization Letters*. doi:10.1007/s11590-015-0971-7.

Nemhauser, G. and Park, S. (1991). A polyhedral approach to edge coloring. *Operations Research Letters*, **10**, 315–322.

Nemhauser, G. and Trotter Jr., L. (1974). Properties of vertex packing and independence system polyhedra. *Mathematical Programming*, **6**, 48–61.

Nemhauser, G. and Trotter Jr., L. (1975). Vertex packings: Structural properties and algorithms. *Mathematical Programming*, **8**(1), 232–248.

Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, **69**, 026113.

Östergård, P. R. (2002). A fast algorithm for the maximum clique problem. *Discrete Applied Mathematics*, **120**(1-3), 197–207.

Pattillo, J., Youssef, N., and Butenko, S. (2012). Clique relaxation models in social network analysis. In M. T. Thai and P. M. Pardalos, editors, *Handbook of Optimization in Complex Networks*, volume 58 of *Springer Optimization and Its Applications*, pages 143–162. Springer New York.

Pattillo, J., Youssef, N., and Butenko, S. (2013a). On clique relaxation models in network analysis. *European Journal of Operational Research*, **226**(1), 9–18.

Pattillo, J., Veremyev, A., Butenko, S., and Boginski, V. (2013b). On the maximum quasi-clique problem. *Discrete Applied Mathematics*, **161**(1–2), 244–257.

Porter, M. A., Onnela, J.-P., and Mucha, P. J. (2009). Communities in networks. *Notices of the AMS*, **56**(9), 1082–1097.

Schaeffer, S. E. (2007). Graph clustering. *Computer Science Review*, **1**(1), 27–64.

Scott, J. (2012). *Social Network Analysis*. Sage, London, UK, 3rd edition edition.

Shahinpour, S. and Butenko, S. (2013). Algorithms for the maximum k-club problem in graphs. *Journal of Combinatorial Optimization*, **26**(3), 520–554.

Sherali, H. D. and Smith, J. C. (2006). A polyhedral study of the generalized vertex packing problem. *Mathematical Programming*, **107**(3), 367–390.

Trukhanov, S., Balasubramaniam, C., Balasundaram, B., and Butenko, S. (2013). Algorithms for detecting optimal hereditary structures in graphs, with application to clique relaxations. *Computational Optimization and Applications*, **56**(1), 113–130.

Veremyev, A. and Boginski, V. (2012). Identifying large robust network clusters via new compact formulations of maximum k-club problems. *European Journal of Operational Research*, **218**(2), 316–326.

Veremyev, A., Prokopyev, O. A., Butenko, S., and Pasiliao, E. L. (2015). Exact MIP-based approaches for finding maximum quasi-cliques and dense subgraphs. *Computational Optimization and Applications*, **64**(1), 177–214.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis*. Cambridge University Press.

Wotzlaw, A. (2014). On solving the maximum k-club problem. Technical Report arXiv:1403.5111v2, Institut für Informatik, Universität zu Köln, Köln, Germany.

Yannakakis, M. (1978). Node-and edge-deletion NP-complete problems. In *STOC '78: Proceedings of the 10th Annual ACM*

Symposium on Theory of Computing, pages 253–264, New York, NY. ACM Press.

Yu, H., Paccanaro, A., Trifonov, V., and Gerstein, M. (2006). Predicting interactions in protein networks by completing defective cliques. *Bioinformatics*, **22**(7), 823–829.

Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, **33**(4), 452–473.

Zhou, H. (2003a). Distance, dissimilarity index, and network community structure. *Physical Review E*, **67**(6), 061901.

Zhou, H. (2003b). Network landscape from a Brownian particle's perspective. *Physical Review E*, **67**(4), 041908.