HHS Public Access

Author manuscript

IEEE ACM Trans Netw. Author manuscript; available in PMC 2018 November 23.

Published in final edited form as:

IEEE ACM Trans Netw. 2017 October ; 25(5): 3219–3234. doi:10.1109/TNET.2017.2728638.

Latent Network Features and Overlapping Community Discovery via Boolean Intersection Representations

Hoang Dau [Member] and **Olgica Milenkovic [Senior Member, IEEE]**

Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, 1308 W. Main Street, Urbana, IL 61801, USA

Abstract

We propose a new latent Boolean feature model for complex networks that captures different types of node interactions and network communities. The model is based on a new concept in graph theory, termed the Boolean intersection representation of a graph, which generalizes the notion of an intersection representation. We mostly focus on one form of Boolean intersection, termed *cointersection*, and describe how to use this representation to deduce node feature sets and their communities. We derive several general bounds on the minimum number of features used in cointersection representations and discuss graph families for which exact cointersection characterizations are possible. Our results also include algorithms for finding optimal and approximate cointersection representations of a graph.

I. Introduction

A. Background

An important task in network analysis is to understand the mechanism behind the formation of a given complex network. *Latent feature models* for networks seek to explain the observed pairwise connections among the nodes in a network by associating to each node a set of features and by setting rules based on which pairs of nodes are connected according to their features. Inference of latent network features not only allows for the discovery of community structures in networks via association with features but also aids in predicting unobserved connections. As such, feature inference is invaluable in the study of social networks, protein complexes and gene regulatory modules.

Probabilistic latent feature models for networks are usually studied via machine learning techniques; known problems and analytic approaches include the Binary Matrix Factorization model [1], the Mixed-Membership Stochastic Block model [2], the Infinite Latent Feature/Attribute model [3], [4], the Multiplicative Attribute Graph model [5], the Attribute Graph Affiliation model [6], and the Cluster Affiliation model (or BIGCLAM) [7]. In contrast, almost nothing is known about deterministic, combinatorial latent feature models.

Part of the results was presented at ISIT 2016, Barcelona.

In the recent work of Tsourakakis [8], a probabilistic latent feature model for networks was proposed that implicitly uses the notion of intersection representations of graphs [9], [10], [11] and builds upon the overlapping community detection approach of Bonchi *et al.* [12]. More specifically, in this model one fixes the total number of features and tries to assign to each vertex a subset of features in a way that maximizes a certain *score*. Here, the score of a specific feature assignment is the count of unordered pairs of vertices (u, v) that satisfies the so-called *Intersection Condition*, which states that u and v are adjacent if and only if they share at least one common feature. In particular, if one insists on a *perfect* score, i.e., a score equal to $\binom{n}{2}$, then the minimum number of features required reduces to the *intersection number* of the graph [9]. An assignment of sets of features to vertices that achieves the perfect score is known as an *intersection representation* of a graph (see Fig. 1)[1]. If in the Intersection Condition one insisted on u and v sharing at least $p \geq 1$ common features, achieving a perfect score would require a minimum number of features equal to the p-intersection number of the graph [10], [11]. Intersection representations elucidate *overlapping community structures* via a simple generative principle: one feature - one community. As an illustrative example, each feature in Fig. 1 may describe one community; the triangle forms one community defined by feature a_1, and the remaining two edges are defined by features a_2 and a_3, respectively. Note that all communities are cliques, and that they may overlap (intersect).

B. Our Contribution

We propose to extend the combinatorial variant of the model studied by Bonchi *et al.* [12] and by Tsourakakis [8] to a much more general setting by using Boolean functions of features that can express more complicated interactions among nodes (vertices). For instance, suppose that there are three different types of features, namely 'Family member', 'City', and 'Hobby'. The Boolean function $f(x_1, x_2, x_3) = x_1 \vee (x_2 \wedge x_3)$ can be used to express the connection rule that two people are Facebook friends if and only if either they are family members or they have lived in at least one common city and shared at least one common hobby. As such, it asserts that the 'Family' feature is more relevant than either of the 'City' or 'Hobby' features. More generally, we can use any Boolean function $f = f(x_1, \ldots, x_r)$ together with a vector $\boldsymbol{p} = (p_1, \ldots, p_r)$, $p_i \geq 1$, to describe a connectivity rule based on r different types of features in which the requirement 'sharing at least one common feature of type $\mathscr{A}_i$' is replaced by the requirement 'sharing at least p_i common features of type $\mathscr{A}_i$'.

In the scope of this paper, we mostly focus on a basic building block of Boolean functions, namely the AND function of two variables $f(x_1, x_2) = x_1 \wedge x_2$. It is straightforward to see that the Boolean OR function leads to results identical to those obtained for the simple intersection problem, and results obtained for AND functions allow one to easily extend all the proposed approaches to the case of Boolean functions that include both AND and OR operations. For simplicity, we also consider $(p_1, p_2) = (1, 1)$. To illustrate the latent feature model arising in this setup, we consider the example in Fig 2. The network has five nodes,

[1]The intersection representation of graph arises in numerous problems such as the keyword conflict problem, the traffic phasing problem, and the competition graphs from food webs, to name a few, and has been extensively studied in the literature (see, for instance [13], [14]).

which represent five different people. Each person is assigned *two* distinct sets of features, one representing the hobbies that the person has and the other representing the cities that the person has lived in. For instance, let $\mathscr{A} = \{a_1, a_2\}$ be such that a_1 stands for *fishing* and a_2 stands for *playing soccer*, and let $\mathscr{B} = \{b_1, b_2\}$ be such that b_1 stands for *Hanoi* and b_2 stands for *Champaign*. Then Person 4 is assigned two sets of features, namely $\{a_2\}$ and $\{b_1, b_2\}$, which states that this person has soccer as a hobby and has lived in both Hanoi and Champaign (to avoid notational clutter, we use $\{a_2 \mid b_1, b_2\}$ to denote pairs of sets). Suppose that two people are connected if and only if they share at least one common hobby AND they have lived in at least one common city. For instance, Person 3 and Person 4 are connected because they have soccer as a common hobby and they both have lived in Hanoi. However, Person 3 and Person 5 are not connected, even though they both like playing soccer, because they have not lived in the same city.

Given the nodes' corresponding sets of features and the rules as of how to connect two nodes, it is clear how the graph emerges. The problem of interest is the opposite: under the assumption that the graph is given and that each node is assigned two subsets of features from $\mathscr{A}$ and $\mathscr{B}$, where $\mathscr{A}$ and $\mathscr{B}$ are two disjoint sets of features, and that two nodes are connected if and only if they share at least one feature from $\mathscr{A}$ and at least one feature from $\mathscr{B}$, how can we infer the latent features assigned to the nodes? Usually, the latent features are abstracted as elements from a discrete set, and the mapping between the elements and the real features is determined based on available data.

Our first aim is to determine the smallest possible number of features $\min(|\mathscr{A}| + |\mathscr{B}|)$ needed to explain a given graph. We refer to this quantity as the *cointersection number* of a graph. Note that the notions of cointersection number and cointersection representation of graphs have not been studied before in the literature. We then proceed to establish general lower and upper bounds on the cointersection number of a graph via its intersection number. In addition, we derive several explicit bounds for various families of graphs, including stars, paths, cycles, ring lattices, Newman-Watts small-world graphs, multipartite graphs, and graphs with bounded degrees (Section III and Section IV). In particular, we describe an interesting connection between the cointersection representations of certain complete multipartite graphs and affine planes. We provide an exact algorithm to find an optimal cointersection representation of a graph by using SAT solvers (Section V-B). In addition, we develop a randomized algorithm to find an approximate cointersection representation of a graph in Section V-C. Finally, we extend the bounds on the cointersection number for the case when a general Boolean function is used instead of the AND function (Section VI). Open problems are discussed in Section VII.

C. Applications of the Cointersection Model

Apart from its principal application in network community detection, the cointersection model may be used in other applications, such as resource allocation. In such applications, having multilple options for the assignments is desired, as it allows more flexibility in the system design. We outline another pertinent problem in this area below.

Key distribution for sensor networks—Suppose that n wireless sensors are deployed in the field, each of which is preset with a set of secret keys from a key pool and a set of time slots of being ON (each sensor alternatively switches between ON and OFF to save energy). Two sensors can establish secure communication if and only if they share a common ON-time slot *and* a common secret key. If a certain topology of secure communication among the sensors needs to be imposed, i.e. given the target communication graph, one would want to find an assignment of keys and time slots to all sensors that uses a minimum number of keys and timeslots. One can even fix either the number of time slots or keys and minimize the value of the other parameter. Clearly, a feasible assignment is a cointersection representation of the given graph. Note that using one common key for all sensors (as would be the case for a communication graph that is complete) imposes security risks for the whole network: even if only one sensor is compromised, all communications may be exposed.

The application of a random intersection representation in key distribution for distributed sensor network was originally studied in the highly cited work of Eschenauer and Gligor [15].

II. Preliminaries

We start by formally introducing our new latent feature model and describing its relevant properties.

A. The Cointersection Model

Definition 1—Let $\mathscr{A}$ and $\mathscr{B}$ be two disjoint nonempty subsets of features of cardinalities α and β, respectively. An $(\alpha \mid \beta)$-*cointersection representation* (CIR) for a graph $\mathscr{G} = (\mathscr{V}, \mathscr{E})$ is a family $\mathscr{R} = \{(A_v \mid B_v) : v \in \mathscr{V}\}$, where $A_v \subseteq \mathscr{A}$, $B_v \subseteq \mathscr{B}$, that satisfies the so-called *Cointersection Condition*:

$$(u, v) \in \mathscr{E} \Leftrightarrow A_u \cap A_v \neq \varnothing \text{ and } B_u \cap B_v \neq \varnothing .$$

Let $\theta^{\mathscr{C}}(\mathscr{G}) = \min_{\mathscr{R}}(|\mathscr{A}| + |\mathscr{B}|)$, where the minimum is taken over all cointersection representations $\mathscr{R}$ of $\mathscr{G}$. Then $\theta^{\mathscr{C}}(\mathscr{G})$ is called the *cointersection number* of $\mathscr{G}$. A cointersection representation that uses exactly $\theta^{\mathscr{C}}(\mathscr{G})$ features is called *optimal*.

It is clear that the cointersection number of a graph is precisely the smallest number of features used to describe the network in the Boolean AND model (see Section VI).

Fig. 2 depicts a (2 | 2)-CIR. We can verify easily that for this graph, $\theta^{\mathscr{C}} = 4$, and hence, this representation is optimal. If we refer to the set of nodes that have a particular common feature as a *community*, then the community structure induced by this representation is illustrated in Fig. 3. Note that in this setting communities are no longer restricted to be cliques, which is a more realistic modeling assumption. Furthermore, u and v are adjacent if and only if they belong to the *intersection* of one community of type $\mathscr{A}$ and another community of type $\mathscr{B}$. Note that communities may also be defined by pairs of features, in which case they form cliques and represent intersections of individual feature communities.

B. The Intersection Number and the p-Intersection Number

In this subsection, we review the concepts and some well-known results on the intersection number and its generalization, the p-intersection number.

Clearly, an $(\alpha \mid 1)$-CIR of a graph is equivalent to an *intersection representation* of the same graph that uses α features [9]. An intersection representation of a graph is equivalent to an *edge clique cover*, i.e. a set of complete subgraphs (cliques) of a graph that covers every edge at least once. The *intersection number* of a graph $\mathscr{G}$, denoted by $\theta_1(\mathscr{G})$, is the smallest number of features used in an intersection representation of the graph, or the size of a smallest edge clique cover of that graph. The *p-intersection number* of a graph, denoted by $\theta_p(\mathscr{G})$, is the smallest possible number of features to assign to the vertices such that two vertices are adjacent if and only if they share at least p common features (see, e.g. [10], [11], [16]). We list below a couple of well-known results on the intersection number and the p-intersection number of a graph.

Theorem 1—(Erdös, Goodman, and Pósa [9]). If $\mathscr{G}$ is any graph, then $\theta_1(\mathscr{G}) \leq \lfloor n^2/4 \rfloor$.

Theorem 2—(Alon [17]). Let $\mathscr{H}$ be a graph on n vertices with maximal degree at most d and minimal degree at least one, and let $\mathscr{G} = \overline{\mathscr{H}}$ be its complement. Then $\theta_1(\mathscr{G}) \leq 2e^2(d+1)^2 \log_e n$.

Theorem 3—(Eaton, Gould, and Rödl [16]). For $p \geq 2$ and any graph $\mathscr{G}$ on n vertices, $\binom{\theta_p(\mathscr{G})}{p} \geq \theta_1(\mathscr{G})$.

Theorem 4—(Eaton, Gould, and Rödl [16]). Let $\mathscr{G}$ be a graph on n vertices with maximum vertex degree d and $p > 1$ be an integer, then $\theta_p(\mathscr{G}) \leq 3epd^2(d+1)^{1/p}n^{1/p}$.

III. Lower and Upper Bounds on the cointersection Numbers of Graphs

We now turn our attention to deriving upper bounds on the cointersection numbers θ^c of arbitrary graphs, and explicit bounds on θ^c for bipartite graphs, chordal graphs, and graphs with bounded vertex degrees.

Lemma 1

For any graph $\mathscr{G}$, one has $\theta^c(\mathscr{G}) \leq 1 + \theta_1(\mathscr{G})$.

Proof—Given an optimal intersection representation of $\mathscr{G}$, which uses θ_1 features, we may create a $(\theta_1 \mid 1)$-CIR of $\mathscr{G}$ as follows. If in the intersection representation of $\mathscr{G}$ the vertex v is assigned the set of features $\{a_1, \ldots, a_r\}$, then in the corresponding cointersection representation of $\mathscr{G}$, we assign to v the sets of features $\{a_1, \ldots, a_r \mid b\}$, where $b \notin \{a_1, \ldots, a_{\theta_1(\mathscr{G})}\}$. It is easy to verify that this feature assignment is indeed a $(\theta_1 \mid 1)$-CIR of $\mathscr{G}$.

Lemma 1 immediately implies some explicit upper bounds on the cointersection number of graphs. For instance, the following upper bound for *complement of a sparse graph* is an obvious corollary of Lemma 1 and [17, Theorem 1.4]: if $\mathscr{G}$ is a graph on n vertices with

maximum degree at most $n-1$ and minimum degree at least $n-d$ then $\theta^c(\mathscr{G}) \leq 1 + 2e^2(d+1)^2 \ln n$. Another immediate consequence of Lemma 1 and [18, Corollary 3.2] is that if $\mathscr{G}$ is a chordal graph on n vertices with largest clique of size r then $\theta^c(\mathscr{G}) \leq 1+\theta_1(\mathscr{G}) \leq n-r+2$.

We show next that a graph of bounded degree has a cointersection representation that uses $\mathscr{O}(\sqrt{n})$ features. Our probabilistic proof is based on the analysis in [16, Theorem 11].

Theorem 5

Let $\mathscr{G}$ be a graph on n vertices, with edge set $\mathscr{E}$ and maximum vertex degree $\Delta(\mathscr{G}) \leq d$. Then $\theta^c(\mathscr{G}) \leq 16d^{5/2}\sqrt{n}$.

Proof—Let $\mathscr{A}$ and $\mathscr{B}$ be two disjoint sets of features of the same cardinality $\alpha = \beta = 8d^{5/2}n^{1/2}$. Our goal is to show the existence of an $(\alpha \mid \beta)$-CIR of $\mathscr{G}$.

We independently assign to every edge e of $\mathscr{G}$ a randomly chosen pair of features $\{a(e) \mid b(e)\}$, where $a(e) \in \mathscr{A}$ and $b(e) \in \mathscr{B}$. For each vertex $v \in \mathscr{V}$, let

$$A_v = \{a(e) : e = (u, v) \in \mathscr{E}\}, \quad (1)$$

$$B_v = \{b(e) : e = (u, v) \in \mathscr{E}\}. \quad (2)$$

We aim to show that with a positive probability, the feature assignment $\{(A_v \mid B_v) : v \in \mathscr{V}\}$ co-represents $\mathscr{G}$. Clearly, if $e = (u, v) \in \mathscr{E}$ then by (1) and (2), we have $a(e) \in A_u \cap A_v$ and $b(e) \in B_u \cap B_v$. Therefore, $A_u \cap A_v \neq \varnothing$ and $B_u \cap B_v \neq \varnothing$. In order for the Cointersection Condition to be satisfied, we need to show that with a positive probability, for every $(u, v) \notin \mathscr{E}$, either $A_u \cap A_v = \varnothing$ or $B_u \cap B_v = \varnothing$. To this end, we make use of the Lovász Local Lemma [19].

The classical Lovász Local Lemma may be stated as follows. Suppose that there are m *bad* events $E_1, E_2, \ldots, E_m$, each occurring with probability at most P. Moreover, each event is dependent on at most D other events. If $PD \leq 1/4$ then

$$\text{Prob}(\cap_{i=1}^{m} \overline{E_i}) > 0.$$

In other words, with a positive probability, we can avoid all *bad* events simultaneously.

We define our set of *bad* events as follows. For each $(u, v) \notin \mathscr{E}$, we let $E_{u,v}$ denote the event that $A_u \cap A_v \neq \varnothing$ and $B_u \cap B_v \neq \varnothing$. For each event $E_{u,v}$, we need to find an upper bounds on the probability that it happens and the number of other events that it may depend on.

First, we estimate the probability that each $E_{u,v}$ occurs. Since $\Delta(\mathscr{G}) \leq d$, each vertex $v \in \mathscr{V}$ is incident to at most d edges. Therefore, by (1) and (2), $|A_v| \leq d$ and $|B_v| \leq d$, for every $v \in \mathscr{V}$. To obtain an upper bound on the probability that $A_u \cap A_v \neq \varnothing$, we may assume that $|A_u|$ and

$|A_v|$ are as large as possible, i.e. $|A_u| = |A_v| = d$. Moreover, since u and v do not have any incident edges in common, their sets of $\mathscr{A}$-features are independent. Therefore, we can treat A_u and A_v as two arbitrary subsets of $[\alpha]$ of sizes d. Then we have

$$\text{Prob}(A_u \cap A_v \neq \varnothing) \leq \frac{d\binom{\alpha}{d-1}}{\binom{\alpha}{d}} = \frac{d^2}{\alpha - d + 1}.$$

Similarly,

$$\text{Prob}(B_u \cap B_v \neq \varnothing) \leq \frac{d\binom{\beta}{d-1}}{\binom{\beta}{d}} = \frac{d^2}{\beta - d + 1}.$$

Thus, we deduce that for $(u, v) \notin \mathscr{E}$,

$$\text{Prob}(E_{u,v}) = \text{Prob}(A_u \cap A_v \neq \varnothing) \times \text{Prob}(B_u \cap B_v \neq \varnothing) \leq P = \frac{d^4}{(\alpha - d + 1)(\beta - d + 1)}. \quad (3)$$

Second, we evaluate the number of other events that a certain event $E_{u,v}$ is dependent of. If $(u, v) \notin \mathscr{E}$ and $(w, x) \notin \mathscr{E}$ then the two events $E_{u,v}$ and $E_{w,x}$ are dependent if and only if either there exist $z \in \{u, v\}$ and $z' \in \{w, x\}$ such that $(z, z') \in \mathscr{E}$ or $|\{u, v, w, x\}| \leq 3$. For each $(u, v) \notin \mathscr{E}$, there are at most $2dn$ pairs $\{w, x\}$ that meet the first criteria and at most $2n$ pairs that meet the second. Therefore, each event $E_{u,v}$ is dependent of at most $D = 2n(d + 1)$ other events.

By Lovás Local Lemma, it remains to prove that $PD \leq 1/4$. Recall that we assumed that $\alpha = \beta = 8d^{5/2}n^{1/2}$. Hence, we need to show that

$$(8d^{5/2}n^{1/2} - d + 1)^2 \geq 8d^4(d + 1)n. \quad (4)$$

This claim may be established as follows:

$$\begin{aligned}
&(8d^{5/2}n^{1/2} - d + 1)^2 \geq (8d^{5/2}n^{1/2} - 2\sqrt{2}d)^2 = 8d^2(2\sqrt{2}d^{3/2}n^{1/2} - 1)^2 = 8d^2(8d^3n - 4\sqrt{2}d^{3/2}n^{1/2} + 1) \\
&\geq 8d^2((d^3n + d^2n) + (7d^3n - d^2n - 4\sqrt{2}d^{3/2}n^{1/2})) = 8d^2\left(d^2(d + 1)n + ((7d - 1)d^{1/2}n^{1/2} - 4\sqrt{2})d^{3/2}n^{1/2}\right) \\
&> 8d^4(d + 1)n.
\end{aligned}$$

The last inequality is due to the fact that for $n \geq d \geq 1$, we have $(7d - 1)d^{1/2}n^{1/2} \geq 6 > 4\sqrt{2}$. This completes the proof.

For triangle-free d-regular graphs $\mathscr{G}$ on n vertices, by Corollary 1, $\theta^{c}(\mathscr{G}) \geq 2\sqrt{\theta_1(\mathscr{G})} = \sqrt{2d}\sqrt{n}$. Therefore, in this case, the upper bound given by Theorem 5 is optimal up to a constant factor depending on d.

Recall that $\theta_2(\mathscr{G})$ denotes the 2-intersection number of $\mathscr{G}$. As already pointed out, Eaton *et al.* [16] showed that $\theta_2(\mathscr{G}) \leq 1+\theta_1(\mathscr{G})$ for a general graph and $\theta_2(\mathscr{G}) \leq 3epd^2(d+1)^{1/2}\sqrt{n}$ for a graph of bounded degree d. The former bound is the same as the upper bound for $\theta^{c}(\mathscr{G})$ in Lemma 1 and the latter is essentially the same as the upper bound for $\theta^{c}(\mathscr{G})$ in Theorem 5. However, $\theta^{c}(\mathscr{G})$ and $\theta_2(\mathscr{G})$ can be vastly different for certain families of graphs. For instance, we establish in Proposition 3 in Section IV that for a complete balanced bipartite graph with edge set $\mathscr{V}$, while $\theta^{c}(\mathscr{G}) = |\mathscr{V}|$, $\theta_2(\mathscr{G})$ is quadratic in $|\mathscr{V}|$ (see Chung and West [11] for the latter claim).

Next, we show that the cointersection number of a bipartite graph is at most its order. Since the intersection representation of a bipartite graph is equal to its size, the bound stated in Lemma 2 improves the bound stated in Lemma 1 when the graph has more edges than vertices.

Lemma 2

$\theta^{c}(\mathscr{G}) \leq |\mathscr{V}|$ if $\mathscr{G} = (\mathscr{V}, \mathscr{E})$ is a bipartite graph.

Proof—As $\mathscr{G}$ is a bipartite graph, we can partition the set of vertices into two parts, say $U = \{1, 2, \ldots, n_1\}$ and $V = \{n_1 + 1, n_1 + 2, \ldots, n\}$, for some $1 \leq n_1 < n$, so that $\mathscr{E} \subseteq \{(u, v) : u \in U, v \in V\}$. Set $\mathscr{A} = \{a_u : u \in U\}$ and $\mathscr{B} = \{b_v : v \in V\}$. We assign to each $u \in U$ two sets of features, namely $A_u = \{a_u\}$ and $B_u = \{b_v : (u, v) \in \mathscr{E}\}$. Similarly, we assign to each $v \in V$ two sets of features, namely $A_v = \{a_u : (u, v) \in \mathscr{E}\}$ and $B_v = \{b_v\}$. Then it is straightforward to verify that $\mathscr{R} = \{(A_v, B_v) : v \in \mathscr{V}\}$ is an $(n_1, n - n_1)$-CIR of $\mathscr{G}$. As this cointersection representation uses n features in total, the proof follows.

We prove next a lower bound on θ^{c} via θ_1.

Lemma 3

If $\mathscr{R}$ is an $(\alpha \mid \beta)$-CIR of $\mathscr{G}$ then $\alpha\beta \geq \theta_1(\mathscr{G})$. As a consequence, $\theta^{c}(\mathscr{G}) \geq \min_{\alpha\beta \geq \theta_1(\mathscr{G})}(\alpha + \beta)$.

Proof—Suppose we have a cointersection representation $\mathscr{R} = \{(A_v \mid B_v) : v \in \mathscr{V}\}$ of $\mathscr{G}$ with two disjoint sets of features $\mathscr{A}$ and $\mathscr{B}$, where $|\mathscr{A}| = \alpha$ and $|\mathscr{B}| = \beta$. For each pair $(a, b) \in \mathscr{A} \times \mathscr{B}$, the set of vertices $\mathscr{C}_{a,b} = \{v \in V : a \in A_v, b \in B_v\}$ forms a clique of $\mathscr{G}$. Moreover, it is obvious that any edge of $\mathscr{G}$ must be covered by one such clique. Therefore, $\mathscr{C} = \{\mathscr{C}_{a,b} : (a, b) \in \mathscr{A} \times \mathscr{B}\}$ is an edge clique cover of $\mathscr{G}$. As $\theta_1(\mathscr{G})$ is the number of cliques in a minimum edge clique cover of $\mathscr{G}$, we have

$$\alpha\beta = |\mathscr{A}||\mathscr{B}| = |\mathscr{C}| \geq \theta_1(\mathscr{G}).$$

Therefore, $\theta^{c}(\mathscr{G}) \geq \min_{\alpha\beta \geq \theta_1(\mathscr{G})}(\alpha + \beta)$.

The following is immediate from Lemma 1 and Lemma 3.

Corollary 1

For any graph $\mathscr{G}$ we have

$$\lceil 2\sqrt{\theta_1(\mathscr{G})}\rceil \leq \theta^c(\mathscr{G}) \leq 1 + \theta_1(\mathscr{G}). \quad (5)$$

Note again that both θ^c and θ_2 (the 2-intersection number) have quite similar lower bounds in terms of θ_1. Indeed, based on the aforementioned bound $\binom{\theta_2(\mathscr{G})}{2} \geq \theta_1(\mathscr{G})$, one arrives at $\theta_2(\mathscr{G}) \geq \sqrt{2\theta_1(\mathscr{G})}$. Corollary 1 gives us $\theta^c(\mathscr{G}) \geq 2\sqrt{\theta_1(\mathscr{G})}$. The two lower bounds for θ_2 and θ^c differ from each other only by a multiplicative factor of $\sqrt{2}$.

IV. Tightness of the Bounds

We discuss next the tightness of the bounds on $\theta^c(\mathscr{G})$ for several families of graphs. In addition, we link the existence of cointersection representations of certain complete multipartite graphs that achieve the lower bound with the existence of specific affine planes.

A. Graphs with small θ_1

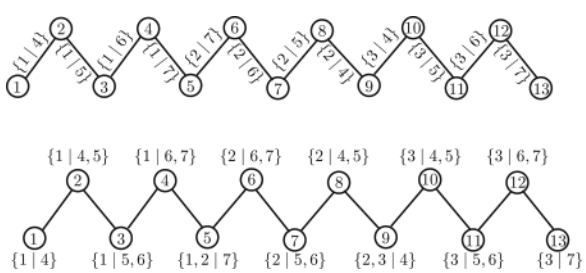
The first result shows that for graphs with very small θ_1, the upper bound $\theta^c(\mathscr{G}) \leq 1 + \theta_1(\mathscr{G})$ is actually tight.

Proposition 1—The upper bound $\theta^c(\mathscr{G}) \leq 1 + \theta_1(\mathscr{G})$ stated in Lemma 1 is tight when $\theta_1(\mathscr{G}) \leq 3$.

Proof: It is obvious that when $\theta_1(\mathscr{G}) \leq 3$, the left-hand side and the right-hand side of (5) are coincide.

B. Stars, Paths, and Cycles

Next, we demonstrate that for some simple graphs, the lower bound $\alpha\beta \geq \theta_1(\mathscr{G})$ established in Lemma 3 is also sufficient for the existence of an $(\alpha \mid \beta)$-CIR. As θ_1 is known for these graphs, θ^c can be determined explicitly.



Proposition 2—If $\alpha\beta \geq \theta_1(\mathscr{G})$ then there exists an $(\alpha \mid \beta)$-CIR of $\mathscr{G}$ when $\mathscr{G}$ is a star $\mathscr{S}_n$, a path $\mathscr{P}_n$, or a cycle $\mathscr{C}_n$.

Proof: Suppose that $\mathscr{G} \equiv \mathscr{S}_n$ is a star graph on n vertices. Let $\mathscr{A}$ and $\mathscr{B}$ be two disjoint subsets of features of sizes α and β, respectively. First, suppose that $\mathscr{S}_n$ has edges (1, 2), (1, 3), …, (1, n). Since $|\mathscr{A}||\mathscr{B}| \geq n - 1 = \theta_1(\mathscr{S}_n)$, we can assign distinct pairs $(a, b) \in \mathscr{A} \times \mathscr{B}$ to the edges of $\mathscr{S}_n$. For each vertex $\upsilon \in \{2, \ldots, n\}$, let $A_\upsilon = \{a_{1,\upsilon}\}$, $B_\upsilon = \{b_{1,\upsilon}\}$, where $\{a_{1,\upsilon} \mid b_{1,\upsilon}\}$ are the features assigned to the edge $(1, \upsilon)$. Also, let $A_1 = \mathscr{A}$ and $B_1 = \mathscr{B}$. It is clear that this is an $(\alpha \mid \beta)$-CIR of $\mathscr{S}_n$.

Next, suppose that $\mathscr{G} \equiv \mathscr{P}_n$ is a path on n vertices and that it has edges $(\upsilon, \upsilon+1)$, $1 \leq \upsilon < n$. Recall that $\theta_1(\mathscr{P}_n) = n-1$. To simplify the notation, we assume that $\alpha\beta = \theta_1(\mathscr{P}_n) = n - 1$. The case when we have strict inequality can be proved in the same manner. Furthermore, let $\mathscr{A} = \{a_1, \ldots, a_\alpha\}$, and $\mathscr{B} = \{b_1, \ldots, b_\beta\}$.

We describe next an $(\alpha \mid \beta)$-CIR of $\mathscr{P}_n$. We first split $n - 1$ edges of $\mathscr{P}_n$ into α equal-sized groups, each consisting of precisely β consecutive edges. We then assign $\{a_1 \mid b_1\}$, $\{a_1 \mid b_2\}$, …, $\{a_1 \mid b_\beta\}$ as features to the first group of β edges in that order. For the next group of β edges, we assign the sequence of features $\{a_2 \mid b_\beta\}$, $\{a_2 \mid b_{\beta-1}\}$, …, $\{a_2 \mid b_1\}$. For the third group of β edges, we use the sequence $\{a_3 \mid b_1\}$, $\{a_3 \mid b_2\}$, …, $\{a_3 \mid b_\beta\}$. Note that we used an *increasing* order for the indices of the sequence b_j in the first group, and a *decreasing* order for the second group, and again an *increasing* order for the third group. We continue to assign features in this way until reaching the last group of edges. We illustrate this feature assignment for the edges of $\mathscr{P}_{13}$ in the figure below. Here, we set $\mathscr{A} = \{1, 2, 3\}$ and $\mathscr{B} = \{4, 5, 6, 7\}$. We use $\{a(e) \mid b(e)\}$ to denote the pair of features assigned to an edge e. Then we assign to each vertex $\upsilon \in \mathscr{P}_n$ two feature sets $A_\upsilon = \{a(e) : e \text{ is incident to } \upsilon\}$ and $B_\upsilon = \{b(e) : e \text{ is incident to } \upsilon\}$. For example, the features of the vertices of $\mathscr{P}_{13}$ are given in the figure below. We can verify that this is an $(\alpha \mid \beta)$-CIR of $\mathscr{P}_n$. Due to the way we assign features to the vertices, each vertex has precisely the feature pairs $\{a, b\}$, where $a \in \mathscr{A}$ and $b \in \mathscr{B}$ assigned to the edges incident to that vertex. Moreover, different edges are assigned different feature pairs. Consequently, two distinct vertices share a common feature pair only if they share a common edge. The proof for cycles proceeds along the same lines as the proof for paths, except for one added modification. Recall that $\theta_1(\mathscr{G}) = n$ if $\mathscr{G} \equiv \mathscr{C}_n$ is a cycle on n vertices. Suppose that $\alpha\beta = n$ (the case $\alpha\beta > n$ can be dealt with in the same manner). We split the n edges of $\mathscr{C}_n$ into α equal-sized groups, each consisting of β consecutive edges. As demonstrated for paths, the key idea is to assign features to edges so that different edges receive different pairs of features and moreover, the set of the feature pairs each vertex has consists precisely of the feature pairs assigned to its two adjacent edges. When α is even, we assign features to α groups of edges of $\mathscr{C}_n$ and then deduce the set of features assigned to each vertex in the same way we do for paths. When α is odd, this feature assignment may no longer work, because now the vertex 1 of the cycle would be assigned two sets of features $\mathscr{A}_1 = \{a_1, a_\alpha\}$ and $\mathscr{B}_1 = \{b_1, b_\beta\}$; as a result, it would have four instead of two feature pairs, namely $\{a_1 \mid b_1\}$, $\{a_1 \mid b_\beta\}$, $\{a_\alpha \mid b_1\}$, $\{a_\alpha \mid b_\beta\}$. As a consequence, this vertex may share a common pair of features with some other vertices that are not adjacent to it. For instance, for $n = 9 = 3\times3$, the currently discussed feature assignment for $\mathscr{C}_9$, demonstrated in Fig. 4, violates the Cointersection Condition.

We correct this issue as follows. Suppose that $\alpha \geq 3$ (the case $\alpha = 1$ and $\beta = n$ is trivial, due to Lemma 1). We assign features to the first $\alpha - 2$ groups of edges of $\mathscr{C}_n$ in the same way as

for paths. For the $(\alpha - 1)$th group, instead of assigning $\{a_{\alpha-1} \mid b_\beta\}, \ldots, \{a_{\alpha-1} \mid b_1\}$, we assign $\{a_{\alpha-1} \mid b_\beta\}, \ldots, \{a_{\alpha-1} \mid b_3\}, \{a_{\alpha-1} \mid b_1\}, \{a_{\alpha-1} \mid b_2\}$ to the edges in this order. For the αth group, instead of assigning $\{a_\alpha \mid b_1\}, \ldots, \{a_\alpha \mid b_\beta\}$, we assign $\{a_\alpha \mid b_2\}, \{a_\alpha \mid b_3\}, \ldots, \{a_\alpha \mid b_\beta\}, \{a_\alpha \mid b_1\}$ to the edges. In this way, we guarantee that the vertex 1 is also assigned two feature pairs as the others, and hence, two vertices share a common feature pair if and only if they are adjacent to the same edge. We illustrate this feature assignment in Fig. 5.

Corollary 2—If $\mathscr{G}$ is a star, a path, or a cycle, then $\lceil 2\sqrt{\theta_1(\mathscr{G})} \rceil \leq \theta^c(\mathscr{G}) \leq 2\lceil \sqrt{\theta_1(\mathscr{G})} \rceil$.

Proof: By Corollary 1, we have $\theta^c(\mathscr{G}) \geq \lceil 2\sqrt{\theta_1(\mathscr{G})} \rceil$. Moreover, by Proposition 2, if $\mathscr{G}$ is a star, a path, or a cycle, then there exists a $(\lceil \sqrt{\theta_1(\mathscr{G})} \rceil \mid \lceil \sqrt{\theta_1(\mathscr{G})} \rceil)$-CIR of $\mathscr{G}$, which uses $2\lceil \sqrt{\theta_1(\mathscr{G})} \rceil$ features in total. Hence, $\lceil 2\sqrt{\theta_1(\mathscr{G})} \rceil \leq \theta^c(\mathscr{G}) \leq 2\lceil \sqrt{\theta_1(\mathscr{G})} \rceil$, which establishes our assertion for stars, paths, and cycles.

C. Ring Lattices and Newman-Watts Random Graphs

A *ring lattice* $\mathscr{L}(n, k)$ is a graph obtained by taking a cycle on n vertices and connecting each vertex to its neighbors at most k edges away, forming a $2k$-regular graph (an $\mathscr{L}(20, 2)$ is depicted in Fig. 6). The ring lattice is an essential component in the construction of the random graph in the *Watts-Strogatz* model [20]. In this model, a random graph is created by taking a ring lattice $\mathscr{L}(n, k)$ and *rewiring* every existing edge (u, v) to a random new edge (u, w) with probability q. Note that when $q = 0$ the resulting graph is the same as the ring lattice, and when $q = 1$, it resembles the classic Erdös-Rényi random graph $G(n, q')$ [21], in which every pair (u, v) is an edge with probability $q' = nk/\binom{n}{2}$ [21]. The random graphs in theWatts-Strogatz model have two important properties of a *small-world* graph: small typical path length and large typical clustering coefficient [20]. A more mathematically tractable variant of the model was proposed by Newman and Watts [22], where instead of rewiring lattice edges (nkq edges, on average), shortcuts are added to the lattice ring without removing any existing edges (nkq shortcuts are added, on average). This is referred to as the *Newman-Watts* model and a random graph in this family is denoted by $\mathscr{L}_q(n, k)$.

We provide next an upper bound on the cointersection number of ring lattices, which in turn leads to a probabilistic upper bound on that of the random graph in the Newman-Watts model.

Theorem 6—If $\alpha\beta' \geq n$, $\alpha \geq 4$, $\beta' \geq 2k + 1$, and $k \geq 2$, then there exists an $(\alpha \mid \beta \triangleq \beta' + (k - 1)(\alpha - 1))$-CIR of $\mathscr{L}(n, k)$.

Proof: -assigning procedure for cycles in the following way. Suppose that $\alpha\beta' = n$ (the case $\alpha\beta' > n$ can be dealt with in the same manner). There are n $(k + 1)$-cliques, each of which consists of $k + 1$ consecutive vertices along the ring. Let C_i, $i \in [n]$, be the $(k+1)$-clique formed by the vertex i and its k right-neighbors. We partition these n cliques into α groups, each of which consists of β' consecutive cliques. Set $\mathscr{A} = \{a_1, \ldots, a_\alpha\}$, and $\mathscr{B} = \{b_1, \ldots, b_{\beta'+(k-1)(\alpha-1)}\}$.

We assign pairs of features to cliques one by one and group by group, following similar rules as those used for cycles, with some additional changes. The cliques in the first group are assigned the features $\{a_1|b_1\}, \ldots, \{a_1|b'_\beta\}$ as in the case of cycles. From the second group to the second to last group, the following rules are applied. *Transition Rule* (the same as for cycles): the first clique in group i ($2 \le i \le \alpha - 1$) is assigned the transitional pair of features $\{a_i \mid b_j\}$, where $\{a_{i-1} \mid b_j\}$ has been assigned to the last clique in the previous group. *New Feature Rule* (differs from the assignment rule for cycles): from the second clique to the kth clique in group $i \ge 2$, we assign $\{a_i \mid b_{\beta'+(i-1)(k-1)+1}\}, \ldots, \{a_1 \mid b_{\beta'+i(k-1)}\}$. Note that these $k - 1$ $\mathcal{B}$-features are new, as they have not been used in the previous steps. *Greedy Rule*: the features for the $(k + 1)$th clique to the βth clique in this group are assigned such that the smallest possible $\mathcal{B}$-feature is chosen for the each considered clique (the $\mathcal{A}$-feature is always a_i), in a way that avoids obvious violation of the Cointersection Condition. If we are in the second to last group ($i = \alpha - 1$), we also have to avoid assigning any of the features $b_1, \ldots, b_{k-1}$ to the cliques that share at least one vertex with the cliques in the last group. *Inverse Rule*: for the last group, we apply the Transition Rule, but ignore the New Feature Rule. In fact, we apply the Greedy Rule from the second clique onwards, and only use the novel features $b_{\beta+(\alpha-2)(k-1)+1}, \ldots, b_{\beta'+(\alpha-1)(k-1)}$ for the last $k-1$ cliques.

It can be shown that the above procedure always produces a valid cointersection representation, which uses $\alpha+\beta'+(k-1)(\alpha - 1)$ features. We omit the details of the proof.

Following the proof of Theorem 6, a cointersection representation of $\mathcal{L}(20, 2)$ is depicted in Fig. 6 ($\alpha = 4$ and $\beta' = 5$).

Corollary 3—Given that $2 \le k \le (\lceil\sqrt{n}\rceil - 1)/2$, for all $\delta > 0$, we have

$$\Pr\left[\theta^c(\mathcal{L}_q(n,k)) \ge 2\lceil\sqrt{n}\rceil + (k-1)(\lceil\sqrt{n}\rceil - 1) + 2(1+\delta)\mu\right] \le \exp\left(-\delta^2\mu/(2+\delta)\right),$$

where $\mu \triangleq nkq$. In particular, when $q \approx 1/\left(2(1+\delta)\sqrt{n}\right)$, $\theta^c\left(\mathcal{L}_q(n,k)\right) \le (2k+1)\lceil\sqrt{n}\rceil$ with probability at least $1 - \exp\left(-\delta^2 k\lceil\sqrt{n}\rceil/(2(1+\delta)(2+\delta))\right)$.

Proof: Set $\alpha = \beta' = \lceil\sqrt{n}\rceil$. According to Theorem 6, there exists a cointersection representation for $\mathcal{L}(n, k)$ using at most $2\lceil\sqrt{n}\rceil + (k-1)(\lceil\sqrt{n}\rceil - 1)$ features.

For each added random edge $e = (u, v)$, we can extend the current representation by introducing a pair of new features $\{a_e \mid b_e\}$ and assign this pair to both u and v. Let X be the random variable representing the number of random edges added to $\mathcal{L}(n, k)$. Then X follows the binomial distribution $B(nk, q)$. Moreover, $\mu = \mathbb{E}(X) = nkq$. Applying the Chernoff bound, we deduce

$$\Pr\left[X \ge (1+\delta)\mu\right] \le \exp\left(-\delta^2\mu/(2+\delta)\right).$$

The proof follows.

Note that as $\theta_1(\mathcal{L}(n, k) = n$, by Corollary 5, $\theta^c(\mathcal{L}(n,k)) \geq \lceil 2\sqrt{n} \rceil$. Theorem 6, on the other hand, establishes that $\theta^c(\mathcal{L}(n,k)) \leq (k+1)\lceil \sqrt{n} \rceil$, by setting $\alpha = \beta' = \lceil \sqrt{n} \rceil$. The random graph $\mathcal{L}_q(n, k)$ in the Newman-Watts model, which contains $\mathcal{L}(n, k)$ as a subgraph, has the same lower bound and almost the same (probabilistic) upper bound, as stated in Corollary 3, for a specific regime of q. For both graphs, the established lower bound and the (probabilistic) upper bound differ by a linear factor in k.

D. Multipartite Graphs

Note that for a complete bipartite graph $\mathcal{K}_{n,n}$, we have $\theta_1(\mathcal{K}_{n,n}) = n^2$, which is precisely the number of edges. We henceforth denote the set $\{1, 2, \ldots, m\}$ by $[m]$.

Proposition 3—If $n = ts$ then a (t, ts^2)-CIR exists for $\mathcal{K}_{n,n}$. As a consequence, $\theta^c(\mathcal{K}_{n,n}) = 2n = 2\sqrt{\theta_1(\mathcal{K}_{n,n})}$.

Proof: The explanation that the second assertion follows from the first assertion is as follows. Let $t = n$ and $s = 1$. Then an (n, n)-CIR of $\mathcal{K}_{n,n}$ exists which uses exactly $2n$ features. Combining this result with Corollary 1, we have

$$2n = 2\sqrt{\theta_1(\mathcal{K}_{n,n})} \leq \theta^c(\mathcal{K}_{n,n}) \leq 2n,$$

which implies that

$$\theta^c(\mathcal{K}_{n,n}) = 2n = 2\sqrt{\theta_1(\mathcal{K}_{n,n})}.$$

Note that this equality may also be deduced by combining Corollary 1 and Lemma 2.

We now prove the first assertion of the proposition. Let $\mathcal{A} = \{a_1, \ldots, a_t\}$ and $\mathcal{B} = \{b_1, \ldots, b_{ts^2}\}$. Let $R_1, \ldots, R_s$ be disjoint subsets of size ts of $\mathcal{B}$ that partition $\mathcal{B}$. Moreover, let $C_1, \ldots, C_{ts}$ be disjoint subsets of size s of $\mathcal{B}$ that partition $\mathcal{B}$. In addition, let $|R_i \cap C_j| = 1$ for every $i \in [s]$ and $j \in [ts]$. For instance, if we arrange the ts^2 elements of $\mathcal{B}$ in a $s \times (ts)$ matrix, then we can simply let R_i be the set of ts elements in the ith row and let C_j be the set of s elements in the jth column.

We assign feature sets to each vertex in $\mathcal{K}_{n,n}$ as follows. Suppose that $\mathcal{V}(\mathcal{K}_{n,n}) = \{1, \ldots, n\} \cup \{n+1, \ldots, 2n\}$, and let $\mathcal{E}(\mathcal{K}_{n,n}) = \{(i, j) : 1 \leq i \leq n, n+1 \leq j \leq 2n\}$. First, for a vertex $i \in \{1, \ldots, n\}$, we write $i = (i_a - 1)s + i_b - 1$, where $1 \leq i_a \leq t$ and $1 \leq i_b \leq s$. Then we assign $A_i = \{a_{i_a}\}$ and $B_i = R_{i_b}$. For a vertex $i \in \{n+1, \ldots, 2n\}$, we assign $A_i = \mathcal{A} = \{a_1, \ldots, a_t\}$ and $B_i = C_i$. Recall that $n = ts$, which is precisely the number of sets C_j's that we have. For example, when $n = 6$, $t = 2$, and $s = 3$, then the sets R_i and C_j consist of elements in the correspondingly indexed rows and columns, respectively, of the matrix given below. The resulting (2, 18)-CIR of $\mathcal{K}_{6,6}$ constructed as described above is illustrated in Fig. 7.

We now proceed to verify that this feature assignment is indeed a cointersection representation of $\mathcal{K}_{n,n}$.

We first verify that the Cointersection Condition holds for non-edges of $\mathscr{K}_{n,n}$. For $1 \le i \ne i' \le n$, either $i_a \ne i'_a$ or $i_b \ne i'_b$. If $i_a \ne i'_a$ then $A_i \cap A_{i'} = \{a_{i_a}\} \cap \{a_{i'_a}\} = \varnothing$. If $i_b \ne i'_b$ then $B_i \cap B_{i'} = R_{i_b} \cap R_{i'_b} = \varnothing$, because the sets R_i form a partition. In either case, we have $A_i \cap A_{i'} = \varnothing$ or $B_i \cap B_{i'} = \varnothing$ For $n + 1 \le i \ne i' \le 2n$, we always have $B_i \cap B_{i'} = C_i \cap C_{i'} = \varnothing$, since all the pairs of sets C_i are disjoint.

Next, we verify that the Cointersection Condition holds for edges of $\mathscr{K}_{n,n}$. Indeed, for $1 \le i \le n$ and $n + 1 \le j \le 2n$, we have $A_i \cap A_j = \{a_{i_a}\} \cap \mathscr{A} = \{a_{i_a}\} \ne \varnothing$, and moreover, $B_i \cap B_j = R_{i_b} \cap C_j \ne \varnothing$, because we assume that $|R_i \cap C_j| = 1$ for every $i \in [s]$ and $j \in [ts]$. Thus, we constructed a (t, ts^2)-CIR of $\mathscr{K}_{n,n}$.

Before proceeding with our discussion, we review a few definitions from the theory of combinatorial designs (see, e.g. [23, VI.40]). Let $n \ge k \ge 2$. A 2-$(n, k, 1)$ *packing* is a pair $(\mathscr{X}, \mathscr{S})$, where $\mathscr{X}$ is a set of n elements (points) and $\mathscr{S}$ is a collection of subsets of size k of $\mathscr{X}$ (blocks), such that every pair of points occurs in *at most* one block in $\mathscr{S}$. A 2-$(n, k, 1)$ packing $(\mathscr{X}, \mathscr{S})$ is *resolvable* if $\mathscr{S}$ can be partitioned into *parallel classes*, each comprising n/k blocks that partition $\mathscr{X}$. We provide an example for a 2-(9, 3, 1) resolvable packing below.

The following simple lemma describes a property of a 2-$(k^2, k, 1)$ resolvable packing that will be of importance in the proof of upcoming Theorem 7.

Lemma 4—Let $(\mathscr{X}, \mathscr{S})$ be a 2-$(k^2, k, 1)$ resolvable packing. If $S \in \mathscr{S}$ and $S' \in \mathscr{S}$ are two blocks from different parallel classes, then $|S \cap S'| = 1$.

Proof: By the definition of a packing, every pair of points is contained in exactly one block. Therefore, any two different blocks have at most one point in common. Hence, $|S \cap S'| \le 1$. Suppose that S and S' belong two different parallel classes $\mathscr{C}$ and $\mathscr{C}'$, respectively. Note that each parallel class consists of precisely $k = k^2/k$ disjoint blocks. These k blocks together partition the set $\mathscr{X}$. Therefore, if $S' \notin \mathscr{C}$ then it must intersect each block in $\mathscr{C}$ at at least one point, for otherwise

$$|S'| = |\cup_{S \in \mathscr{C}} S' \cap S| = \sum_{S \in \mathscr{C}} |S' \cap S| < \sum_{S \in \mathscr{C}} 1 = k,$$

a contradiction. Hence, $|S \cap S'| \ge 1$. Thus, $|S \cap S'| = 1$.

Theorem 7—If there exists a 2-$(k^2, k, 1)$-resolvable packing with at least $r \ge 2$ parallel classes then $\theta^c(\mathscr{K}_n[r]) = 2n$, where $n = k^2$, and $\mathscr{K}_n[r]$ is the complete r-partite graph $\mathscr{K}_{n, \ldots, n}$.

Proof: Note that for $r \ge 2$, $\mathscr{K}_{n,n}$ is an induced subgraph of $\mathscr{K}_n[r]$. Therefore, by Proposition 3, we have

$$\theta^c(\mathscr{K}_{n}[r]) \ge \theta^c(\mathscr{K}_{n,n}) = 2n.$$

Hence, it remains to prove that we can co-represent $\mathscr{K}_{n[r]}$ by using $2n$ features if a certain resolvable packing exists.

Let us assume that a 2-($n = k^2$, k, 1)-resolvable packing ($\mathscr{X}$, $\mathscr{S}$) with at least r parallel classes, say $\mathscr{C}_1, \ldots, \mathscr{C}_r$, exists. Let $\mathscr{A} = \{a_x : x \in \mathscr{X}\}$ and $\mathscr{B} = \{b_x : x \in \mathscr{X}\}$. Then $|\mathscr{A}| = |\mathscr{B}| = n$. We assign to the vertices of $\mathscr{K}_{n[r]}$ features from $\mathscr{A}$ and $\mathscr{B}$ as follows. Consider n vertices in the ℓth part P_ℓ of the graph ($\ell \in [r]$). We partition these $n = k^2$ vertices into k groups, each of which consists of precisely k vertices. Let $G_i^\ell = \{v_{i,j}^\ell : j \in [k]\}$ denote the ith vertex group of P_ℓ, for $i \in [k]$ and $\ell \in [r]$. The vertices in P_ℓ are then assigned features according to the blocks in the ℓth parallel class $\mathscr{C}_\ell = \{S_1^\ell, \ldots, S_k^\ell\}$ in the following way. The vertex $v_{i,j}^\ell$ in the ith group G_i^ℓ has feature sets $A_{v_{i,j}^\ell} = \{a_x : x \in S_i^\ell\}$ and $B_{v_{i,j}^\ell} = \{b_x : x \in S_j^\ell\}$.

We show next that the above feature assignment indeed satisfies the Cointersection Condition.

First, we verify this condition for the non-edges of $\mathscr{K}_{n[r]}$. Consider each part P_ℓ of the graph. If $v_{i,j}^\ell$ and $v_{i,j'}^\ell$, where $j \neq j'$, are two distinct vertices that belong to the same group G_i^ℓ, then

$$|B_{v_{i,j}^\ell} \cap B_{v_{i,j'}^\ell}| = |S_j^\ell \cap S_{j'}^\ell| = 0.$$

The reason is that when $j \neq j'$, S_j^ℓ and $S_{j'}^\ell$ are two distinct blocks in the same parallel class $\mathscr{C}_\ell$ of the packing, and hence must be disjoint. If $v_{i,j}^\ell$ and $v_{i',j'}^\ell$ belong to different groups G_i^ℓ and $G_{i'}^\ell$, respectively, where $i \neq i'$, then

$$|A_{v_{i,j}^\ell} \cap A_{v_{i',j'}^\ell}| = |S_i^\ell \cap S_{i'}^\ell| = 0,$$

because S_i^ℓ and $S_{i'}^\ell$ are two distinct blocks in the same parallel class $\mathscr{C}_\ell$. Thus, every pair of vertices from the same part P_ℓ ($\ell \in [r]$) has either no $\mathscr{A}$-features or no $\mathscr{B}$-features in common.

Second, we verify the Cointersection Condition for the edges of $\mathscr{K}_{n[r]}$ that connect vertices in different parts. Suppose that $v_{i,j}^\ell \in P_\ell$ and $v_{i',j'}^{\ell'} \in P_{\ell'}$, where P_ℓ and $P_{\ell'}$ are different parts of the complete r-partite graph. Then we have

$$|A_{v_{i,j}^\ell} \cap A_{v_{i',j'}^{\ell'}}| = |S_i^\ell \cap S_{i'}^{\ell'}| = 1.$$

The validity of the above claim follows from the observation that for $\ell \neq \ell'$, the two blocks S_i^{ℓ} and $S_{i'}^{\ell'}$, which are from different parallel classes of the packing, must intersect at one point (according to Lemma 4). Similarly, we have

$$|B_{v_{i,j}^{\ell}} \cap B_{v_{i',j'}^{\ell'}}| = |S_j^{\ell} \cap S_{j'}^{\ell'}| = 1 .$$

Therefore, the Cointersection Condition is satisfied for all edges of the graph. Thus, the assigned features form an (n, n)-CIR of $\mathscr{K}_n[r]$, which uses precisely $2n$ features, as desired.

Example 1—To illustrate the idea of Theorem 7, we consider $\mathscr{K}_{9,9,9,9}$ and the 2-(9, 3, 1) resolvable packing with four parallel classes $\mathscr{C}_1, \mathscr{C}_2, \mathscr{C}_3, \mathscr{C}_4$ given in Fig. 8. Note that by Theorem 7,

$$\theta^{c}(\mathscr{K}_{9,9,9,9}) = \theta^{c}(\mathscr{K}_{9,9,9}) = \theta^{c}(\mathscr{K}_{9,9}) = 2\sqrt{\theta_1(\mathscr{K}_{9,9})} = 18 .$$

We omit the edges of the graph and provide a (9, 9)-CIR of $\mathscr{K}_{9,9,9,9}$ in Fig. 9. Note that in this figure, instead of a_i and b_j, we simply use i and j, respectively.

A 2-(n, k, 1) resolvable *design* (see, e.g. [23, II.7]) is equivalent to a 2-(n, k, 1) resolvable packing defined earlier, except that one requires that every pair of points appear in *exactly* one block. An *affine plane* of order k is a 2-(k^2, k, 1) resolvable design. So far, only affine planes of orders that are prime powers are known (see, e.g. [23, VII.2.2]).

Corollary 4—If there exists an affine plane of order k then $\theta^c(\mathscr{K}_n[r]) = 2n$, for every $r \leq k + 1$, where $n = k^2$. As a consequence, this equality holds when k is a prime power.

Proof: It is well known that a 2-(k^2, k, 1) resolvable design has precisely $k + 1$ parallel classes. As an affine plane of order k is a 2-(k^2, k, 1) resolvable design, which is also a packing, by Theorem 7, the first assertion of the corollary follows. The last assertion also holds because an affine plane of a prime power order always exists. The resolvable packing used in Example 1 is in fact an affine plane of order three.

In light of Corollary 4, it is apparently nontrivial to prove (theoretically or computationally) that $\theta^c(\mathscr{K}_n[r]) > 2n$, where $n = k^2$, $r = k + 1$, when k is not a prime power. Indeed, such a proof (if any) would imply that an affine plane of order k does not exist. Note that the question whether an affine plane of an order which is not a prime power exists is still a widely open question in finite geometry. It is not even known whether an affine plane of order 12 or 15 exists (see, e.g. [23, VII.2.2]).

Corollary 5—$\theta^c(\mathscr{K}_{n,n,n}) = 2n$ for every $n = k^2$, where $k \geq 2$ is not necessarily a prime power.

Proof: By Theorem 7, it suffices to construct a 2-(k^2, k, 1) resolvable packing with three parallel classes for every $k \geq 2$. Let $\mathscr{X} = [k^2]$. We can arrange these k^2 points into a $k \times k$

matrix. Then the k blocks containing the points along the rows of this matrix form the first parallel class. The k blocks containing the points along the columns of this matrix form the second parallel class. The k blocks containing the points along the direction of the main diagonal form the third parallel class. It is easy to verify that these blocks and the three parallel classes form a 2-(k^2, k, 1) resolvable packing. For example, when $k = 4$, the three parallel classes of this packing are given in Fig. 10.

Until this point, we have focused on providing several examples of graphs which meet the lower bound on $\theta^{\mathcal{C}}$ established in Lemma 3. However, as we establish in subsequent propositions, the lower bound many not always be achievable. Note that by Corollary 5, $\theta^{\mathcal{C}}(\mathcal{K}_{n,n,n}) = 2n$ for $n = 4, 9, 16, \ldots$ This is, in contrast, not true for $n = 2, 3$.

We first need to prove the following lemma, which states an important property of cointersection representations of triangle-free graphs (e.g. bipartite graphs) that meet the lower bound on $\theta^{\mathcal{C}}$ in Lemma 3. Recall that if $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a triangle-free graph, then $\theta_1(\mathcal{G}) = |\mathcal{E}|$.

Lemma 5—If there exists an $(\alpha|\beta)$-CIR of a triangle-free graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\alpha\beta = |\mathcal{E}|$, then

$$|A_v||B_v| = \deg(v),$$

for every $v \in V$. Moreover, if $(u, v) \in \mathcal{E}$, then $|A_u \cap A_v| = |B_u \cap B_v| = 1$.

Proof: Suppose that $\{(A_v, B_v) : v \in \mathcal{V}\}$ is an $(\alpha \mid \beta)$-CIR of $\mathcal{G}$, where $\alpha\beta = |\mathcal{E}|$. For each edge $(u, v) \in \mathcal{E}$, choose an arbitrary feature $a_{u,v} \in A_u \cap A_v$ and an arbitrary feature $b_{u,v} \in B_u \cap B_v$ and assign the pair $\{a_{u,v} \mid b_{u,v}\}$ to this edge.

We claim that different edges must have different pairs of features. Indeed, if (u, v) and (u', v') are two different edges of $\mathcal{G}$ such that $a_{u,v} = a_{u',v'}$ and $b_{u,v} = b_{u',v'}$, then the four vertices u, v, u', v' have a pair of features in common, namely $\{a_{u,v} \mid b_{u,v}\}$. This implies that any three distinct vertices among these four must form a triangle in $\mathcal{G}$, which contradicts our assumption that $\mathcal{G}$ is triangle-free. Thus, different edges must be assigned different pairs of features, as claimed. A consequence of this claim is that for every vertex $v \in \mathcal{V}$, the number of pairs of features $\{a \mid b\}$, where $a \in A_v$ and $b \in B_v$, must be greater than or equal to the number of edges incident to v. In other words, $|A_v||B_v| \geq \deg(v)$, for every $v \in \mathcal{V}$.

Moreover, by our assumption, the number of possible pairs of features $\{a \mid b\}$, where $a \in \mathcal{A}$ and $b \in \mathcal{B}$, is $\alpha\beta$, which is the same as the number of edges. Therefore, each such pair of features must be used exactly once, as features of some edge. It is now clear that if $(u, v) \in \mathcal{E}$, then $|A_u \cap A_v| = 1$ and $|B_u \cap B_v| = 1$. For otherwise, we could replace the assigned features $\{a_{u,v} \mid b_{u,v}\}$ for (u, v) by a different pair of features $\{a' \mid b'\}$, where $a' \in A_u \cap A_v$ and $b' \in B_u \cap B_v$. But as proved earlier, $\{a' \mid b'\}$ must already have been used as a pair of features of some other edge $(u', v') \neq (u, v)$. That would imply a triangle formed by some three distinct vertices among u, v, u', and v', which, again, contradicts our assumption that $\mathcal{G}$ is triangle-free.

Finally, suppose that $|A_v||B_v| > \deg(v)$ for some $v \in \mathscr{V}$. Then there must be a pair of features $\{a \mid b\}$, where $a \in A_v$ and $b \in B_v$, that is not assigned to any edge incident to v. However, as shown earlier, this pair of features $\{a \mid b\}$ must be used as features of some edge, say (u, w), that is not incident to v. Then u, v, and w share the common features $a \in \mathscr{A}$ and $b \in \mathscr{B}$ and hence must form a triangle in $\mathscr{G}$, which is impossible. Thus, $|A_v||B_v| = \deg(v)$ for every $v \in \mathscr{V}$, as stated.

Proposition 4—$\theta^c(\mathscr{K}_{n,n,n}) > 2n$ for $n = 2, 3$. Hence, for the given graphs, the lower bound $\theta^c \geq \min_{\alpha\beta \geq \theta_1} (\alpha + \beta)$ established in Lemma 3 is not tight.

Proof: Since the graphs under consideration are small, one can determine their cointersection numbers by using the algorithm of Section V-B, resulting in $\theta^c(\mathscr{K}_{2,2,2}) = 5$ and $\theta^c(\mathscr{K}_{3,3,3}) = 8$. This fact may also be proved theoretically, based on the previously derived results for the induced subgraphs $\mathscr{K}_{2,2}$ and $\mathscr{K}_{3,3}$. The details of the proof are omitted due to lack of space.

Proposition 5—Let $\mathscr{K}^M_{n,n}$ be a bipartite matrix obtained from $\mathscr{K}_{n,n}$ by removing a maximum matching. Then

$$2n - 1 \leq \theta^c(\mathscr{K}^M_{n,n}) \leq 2n .$$

The lower bound is attained when $n = 2, 3$. If $n - 1$ is an odd prime, then $\theta^c(\mathscr{K}^M_{n,n}) = 2n$.

Proof: Let $\mathscr{K}^M_{n,n} = (\mathscr{V}, \mathscr{E})$ and let $U = \{u_1, \ldots, u_n\}$ and $V = \{v_1, \ldots, v_n\}$ be two parts of $\mathscr{V}$ such that $\mathscr{E} = \{(u_i, v_j) : 1 \leq i \neq j \leq n\}$. By Lemma 2 we have

$$\theta^c(\mathscr{K}^M_{n,n}) \leq 2n . \quad (6)$$

Note that

$$\theta_1(\mathscr{K}^M_{n,n}) = |\mathscr{E}(\mathscr{K}^M_{n,n})| = n^2 - n = (n-1)n .$$

Therefore, by Lemma 3,

$$\theta^c(\mathscr{K}^M_{n,n}) \geq \min_{\alpha\beta \geq n(n-1)} (\alpha + \beta) = (n-1) + n = 2n - 1 . \quad (7)$$

When $n = 2, 3$, the above lower bound on θ^c is attained. Examples of $(n - 1, n)$-CIRs of $\mathscr{K}^M_{n,n}$ when $n = 2, 3$ are given in Fig. 11.

It remains to show that if $n-1$ is an odd prime then $\theta^c(\mathscr{K}^M_{n,n}) \neq 2n-1$. Suppose, by contradiction, that $\theta^c(\mathscr{K}^M_{n,n}) = 2n-1$. Then there must exist an $(n-1, n)$-CIR of $\mathscr{K}^M_{n,n}$. Let $\mathscr{A} = \{a_1, \ldots, a_{n-1}\}$ and $\mathscr{B} = \{b_1, \ldots, b_n\}$. Note that every vertex of this graph has degree $n-1$. By Lemma 5, for every vertex v,

$$|A_v|\,|B_v| = \deg(v) = n-1\,.$$

As $n-1$ is a prime number, we deduce that either $|A_v| = 1$ and $|B_v| = n-1$ or $|A_v| = n-1$ and $|B_v| = 1$. We consider the following three cases, distinguished by the number of vertices that have only one $\mathscr{A}$-feature, and aim to obtain a contradiction in each case.

Case 1—$|A_v| = n-1$ and $|B_v| = 1$ for all $v \in \mathscr{V} = U \cup V$. Since $A_{u_i} = \mathscr{A}$ for all $i \in [n]$ and there are no edges between these vertices u_i elements, $B_{u_i} \cap B_{u_j} = \varnothing$ whenever $i \neq j$. Similarly, $B_{v_i} \cap B_{v_j}$ whenever $i \neq j$. However, as u_1 is adjacent to $v_2, \ldots, v_n$, these vertices must have the same $\mathscr{B}$-feature as u_1. We arrive at a contradiction.

Case 2—There exists one vertex, say u_i, satisfying $|A_{u_i}| = 1$, while other vertices in the same part have $A_{u_j} = \mathscr{A}$, $j \neq i$. By Lemma 5, $|B_{u_i}| = n-1$ and $|B_{u_j}| = 1$ for $j \neq i$. Moreover, as u_i is not adjacent to u_j for $j \neq i$, $B_{u_i} \cap B_{u_j} = \varnothing$. As $|\mathscr{B}| = n$ and $|B_{u_i}| = n-1$, this implies that $B_{u_j} = \mathscr{B} \setminus B_{u_i}$ for all $j \neq i$. Since $A_{u_j} = \mathscr{A}$ for all $j \neq i$ as well, the corresponding elements u_j must be all adjacent, which is not true. We arrive at a contradiction.

Case 3—There exist two vertices, which we without loss of generality label as u_i and u_j, that are in the same part of the graph, and which satisfy $|A_{u_i}| = |A_{u_j}| = 1$. Then by Lemma 5, $|B_{u_i}| = |B_{u_j}| = n-1$. Since $n > 2$, $B_{u_i} \cap B_{u_j} \neq \varnothing$. Therefore, $A_{u_i} \cap A_{u_j} = \varnothing$. Without loss of generality, let $A_{u_i} = \{a_i\}$ and $A_{u_j} = \{a_j\}$. For any $h \in [n] \setminus \{i, j\}$, since v_h is connected to both u_i and u_j, we deduce that $\{a_i, a_j\}$ is a subset of both A_{v_h}. As $|A_{v_h}| \in \{1, n-1\}$, we deduce that $A_{v_h} = \mathscr{A}$, for all $h \neq i, j$. Then $|B_{v_h}| = 1$ and $B_{v_h} \cap B_{v_k} = \varnothing$ for every $h \neq k$, $h, k \in [n] \setminus \{i, j\}$. We can set $B_{v_h} = \{b_h\}$ and $B_{v_k} = \{b_k\}$. As $n-1$ is an odd prime, $n \geq 4$. Therefore, we can choose h and k such that h, k, i, j are distinct.

Since v_h and v_k are not adjacent to v_i, and moreover, since $A_{v_h} = A_{v_k} = \mathscr{A}$, we deduce that $B_{v_i} \cap \{b_h, b_k\} = \varnothing$. Therefore, $|B_{v_i}| \leq n-2$. Since $|B_{v_h}| \in \{1, n-1\}$, we deduce that $|B_{v_i}| = 1$. Similarly, $|B_{v_j}| = 1$. We can set $B_{v_i} = \{b_i\}$ and $B_{v_j} = \{b_j\}$. For any $r \neq i, j$, since u_r is adjacent to v_i and v_j, the set $\{b_i, b_j\}$ is a subset of B_{u_r}. Therefore, $|B_{u_r}| = n-1$, and hence, $|A_{u_r}| = 1$, for all $r \in [n]$. By the pigeon hole principle, among the n vertices $u_1, \ldots, u_n$, there must be two distinct vertices, say u_r and u_s, that satisfy $A_{u_r} = A_{u_s}$. Moreover, as $|B_{u_r}| = |B_{u_s}| = n-1$, we must have $B_{u_r} \cap B_{u_s} \neq \varnothing$ as well. We obtain a contradiction, since the Cointersection Condition is violated.

Thus, if $n-1$ is an odd prime then $\theta^c(\mathscr{K}^M_{n,n}) \neq 2n-1$. Therefore, due to (6) and (7), we have $\theta^c(\mathscr{K}^M_{n,n}) = 2n$.

An obvious corollary of Proposition 5 is that there exists infinitely many bipartite graphs where the lower bound $\theta^{\mathcal{C}} \geq \min_{\alpha\beta \geq \theta_1} (\alpha + \beta)$ established in Lemma 3 is not attained.

V. Algorithms for the Cointersection Model

In what follows, we develop two algorithms for finding (exact and approximate) cointersection representations of a graph. The first algorithm is based on a transformation to instances of the Satisfiability Problem (SAT) and outputs an *optimal* cointersection representation, which uses exactly $\theta^{\mathcal{C}}$ features. The second algorithm is based on the well known simulated annealing approach, which produces an *approximate* cointersection representation of a graph. More specifically, this algorithm inputs $\mathcal{G}$, α, and β, and outputs feature assignments to all vertices of the graph so as to maximize, as much as possible, the score of the representation, i.e. the number of pairs (u, v) that satisfy the Cointersection Condition.

A. Uniqueness of Optimal Cointersection Representations

Before presenting the two algorithms, we briefly discuss the question of uniqueness of an optimal cointersection representation of a graph. Throughout our analysis, we tacitly assume that $\alpha \leq \beta$ for all (α, β)-CIRs.

Two cointersection representations are considered *equivalent* if one can be obtained from the other by possibly swapping the set of $\mathcal{A}$-features and the set of $\mathcal{B}$-features (only if $|\mathcal{A}| = |\mathcal{B}|$), and by permuting features within each set. A graph is said to be *uniquely cointersectable* if all of its *optimal* cointersection representations are equivalent. The issue of unique cointersection representations is of importance in practical applications, where different feature assignment algorithms may construct diverse solutions and where we would like to understand how many different solutions are possible. The related concept of *uniquely intersectable* graphs was studied in [24], [25]. It was proved in [25, Thm. 3.2] that every *diamond-free* graph is uniquely intersectable (more precisely, *uniquely intersectable with respect to a multifamily*). Note that a diamond is obtained by removing one edge in $\mathcal{K}_4$. The problem of finding a *necessary and sufficient* condition for a graph to be uniquely intersectable is widely open.

Some examples of uniquely cointersectable graphs include:

- Cliques $\mathcal{K}_n$, $n \geq 2$, which have a unique (1, 1)-CIR with all vertices having features $\{a_1 \mid b_1\}$,
- $\mathcal{K}_n - e$, $n \geq 2$, where $e = (u, v)$ is an arbitrary edge. This graph has a unique (1, 2)-CIR in which u is assigned the pair of features $\{a_1 \mid b_1\}$, v is assigned $\{a_1 \mid b_2\}$, while all other vertices (if any) are assigned the set $\{a_1 \mid b_1, b_2\}$.
- The path $\mathcal{P}_5$ has a unique (2, 2)-CIR, where the vertices from 1 to 5 are respectively assigned the following sets of features: $\{a_1 \mid b_1\}$, $\{a_1 \mid b_1, b_2\}$, $\{a_1, a_2 \mid b_2\}$, $\{a_2 \mid b_1, b_2\}$, and $\{a_2 \mid b_1\}$,

- The cycle $\mathscr{C}_4$ has a unique (2, 2)-CIR, where the vertices from 1 to 4 are respectively assigned the following sets of features: $\{a_1, a_2 \mid b_1\}$, $\{a_1 \mid b_1, b_2\}$, $\{a_1, a_2 \mid b_2\}$, $\{a_2 \mid b_1, b_2\}$.

A graph may not have a unique cointersection representation, even if we restrict ourselves to *optimal* (α, β) cointersection representations, where α and β are fixed, and $\alpha + \beta = \theta^{\mathcal{F}}$. An example of two optimal (2, 3)-CIRs of the path $\mathscr{P}_7$ that are not equivalent is presented in Fig. 12. In fact, we prove in Corollary 6 that *every* path $\mathscr{P}_n$, $n \geq 4$, except $\mathscr{P}_5$, is *not* uniquely cointersectable. A similar result also holds for cycles, but we omit the proof due to lack of space. In fact, most paths have at least exponentially many nonequivalent optimal cointersection representations (Theorem 8). Note that a path or a cycle, which is obviously diamond free, is always uniquely intersectable. These results suggest that uniquely cointersectable graphs are even scarcer than uniquely intersectable ones. The problem of finding a necessary and/or sufficient condition for a graph to be uniquely cointersectable is also open.

Theorem 8—Every path $\mathscr{P}_n$ with $n \geq 6$ has at least $(\lceil\sqrt{n-1}\rceil - 1)!$ nonequivalent optimal cointersection representations.

<u>Proof:</u> The main idea behind the proof is to construct a list of at least $(\lceil\sqrt{n-1}\rceil - 1)!$ optimal cointersection representations of $\mathscr{P}_n$, and then show that for every pair of representations, there exist two vertices whose sets of assigned features intersect in a nonequivalent manner.

Two nonequivalent optimal (2, 3)-cointersection representations of $\mathscr{P}_7$ are shown in Fig. 12. If we delete the last vertex and edge in the paths, we obtain two nonequivalent representations for $\mathscr{P}_6$.

Now suppose that $n \geq 8$ and that we have an optimal (α, β)-CIR of $\mathscr{P}_n$. If $\beta \geq \alpha + 2$, then $(\alpha + 1)(\beta - 1) > \alpha\beta$, and hence by Proposition 2, there is another optimal $(\alpha+1, \beta - 1)$-CIR of $\mathscr{P}_n$. We can repeat this argument to obtain an optimal representation with $\alpha \leq \beta \leq \alpha + 1$ (Note that this argument also reveals that for paths, there always exists a balanced optimal cointersection representation). By Lemma 3, $\alpha(\alpha + 1) \geq \alpha\beta \geq \theta_1(\mathscr{P}_n) = n - 1 \geq 7$. Hence, $\beta \geq \alpha \geq 3$. We also have $\beta \geq \lceil\sqrt{n-1}\rceil$.

We describe next a list of $(\beta - 1)!$ (α, β)-cointersection representations of $\mathscr{P}_n$ and proceed to prove that the representations are pairwise nonequivalent. Each of these representations corresponds to a particular permutation σ of the set $\{1, 2, \ldots, \beta - 1\}$, denoted by $\mathcal{R}_\sigma$. Following the proof of Proposition 2 for paths, we partition the set of $n - 1$ edges into α groups of β consecutive edges each, except for possibly the last group, which may contain less than β edges if $\alpha\beta > n - 1$. In all representations, we assign β pairs of features $\{a_1, b_1\}$, $\{a_1, b_2\}$, …, $\{a_1, b_\beta\}$ to the *first* group of β consecutive edges in that order. In the representation $\mathcal{R}_\sigma$, we continue to assign β pairs of features $\{a_2, b_\beta\}$, $\{a_2, b_{\sigma(\beta-1)}\}$, $\{a_2, b_{\sigma(\beta-2)}\}$, …, $\{a_1, b_{\sigma(1)}\}$ to the next group of β consecutive edges in that order. Similarly, the third group of edges is assigned pairs of features $(a_3, b_{\sigma(1)})$, $(a_3, b_{\sigma(2)})$, …, in $\mathcal{R}_\sigma$, and so forth. In general, the rule is to assign different features a_i to different groups of edges, and to assign the features b_j in such a way that the *last edge* of one group is assigned the same b_j as

the *first edge* of the following group. This process is continued until all edges are assigned one pair of features each. Upon completion of this procedure, each vertex is assigned the union of the sets of features assigned to its adjacent edges. According to the argument used in the proof of Proposition 2 for paths, each $\mathcal{R}_\sigma$ represents an (α, β)-cointersection representation of $\mathcal{P}_n$.

It remains to prove that for two different permutations σ and σ' of $\{1, 2, \ldots, \beta - 1\}$, there exist two distinct vertices u and v whose sets of assigned features intersect differently in the two representations. More specifically, u lies within the first group of vertices and v lies within the second group of vertices. Let $j \in [\beta - 1]$ be the largest index satisfying $z \triangleq \sigma(j) \neq t \triangleq \sigma'(j)$. Then $y \triangleq \sigma(j + 1) = \sigma'(j + 1)$. Note that if $j = \beta - 1$, one may set $y = \beta$. Without loss of generality, let us also assume that $t > z$. We select v (see Fig. 13 and Fig. 14) to be the vertex adjacent to the two consecutive edges in the second group which are assigned features $\{a_2, b_y\}$ and $\{a_2, b_z\}$ in $\mathcal{R}_\sigma$. In $\mathcal{R}_{\sigma'}$, v is adjacent to two edges with assigned features $\{a_2, b_y\}$ and $\{a_2, b_t\}$. As $\alpha \geq 3$, both groups have β edges and vertices u and v as described above always exist.

We consider two cases which correspond to different choices of u. It suffices to show that in both cases, u and v have a different number of common features in $\mathcal{R}_\sigma$ and $\mathcal{R}_{\sigma'}$.

Case 1—$t = z + 1$. We select u (see Fig. 13) as the vertex adjacent to the two consecutive edges in the *first* group that are assigned features $\{a_1, b_t\}$ and $\{a_1, b_{t+1}\}$ in both $\mathcal{R}_\sigma$ and $\mathcal{R}_{\sigma'}$. Note that $t \leq \beta - 1$, and hence $t+1 \leq \beta$. Since $y \notin \{z, t\}$, we consider the following two sub-cases. If $y < z$ or $y > t + 1$, then in $\mathcal{R}_\sigma$ the vertices u and v do not share any features, while in $\mathcal{R}_{\sigma'}$, they do share one common feature, namely b_t. If $y = t+1$, then in $\mathcal{R}_\sigma$ the vertices u and v share precisely one feature, namely b_{t+1}, while in $\mathcal{R}_{\sigma'}$, they share two features, b_t and b_{t+1}.

Case 2—$t > z + 1$. We select u (see Fig. 14) as the vertex adjacent to the two consecutive edges in the *first* group that are assigned $\{a_1, b_z\}$ and $\{a_1, b_{z+1}\}$ in both $\mathcal{R}_\sigma$ and $\mathcal{R}_{\sigma'}$. If $y < z$ or $y > z + 1$ then in $\mathcal{R}_\sigma$ the vertices u and v share one feature, namely b_z, while in $\mathcal{R}_{\sigma'}$, they do not share any features. If $y = z + 1$, then in $\mathcal{R}_\sigma$, the vertices u and v share precisely two features, namely b_z and b_{z+1}, while in $\mathcal{R}_{\sigma'}$, they share only one feature, namely b_{z+1}.

This completes the proof.

Corollary 6—None of the paths $\mathcal{P}_n$, $n \geq 4$, except for $\mathcal{P}_5$, is uniquely cointersectable.

<u>Proof:</u> By Proposition 2, $\mathcal{P}_4$ has a (1, 3)-CIR as well as a (2, 2)-CIR, both of which are optimal. Hence, $\mathcal{P}_4$ is not uniquely cointersectable. For $n \geq 6$, according to Theorem 8, $\mathcal{P}_n$ has at least $2 = (\lceil\sqrt{6-1}\rceil - 1)!$ nonequivalent optimal cointersection representations, and is hence not uniquely cointersectable.

B. Feature Assignments via SAT Solvers

For arbitrary α and β, it is an NP-complete problem to determine if an (α, β)-CIR exists; indeed, when $\alpha = 1$, the problem becomes whether there exists an intersection representation

that uses β features, which is known to be NP-complete [26]. We discuss below a means of determining the cointersection number in a constructive manner, which also results in feature assignments for the vertices. The idea is to restate the cointersection problem as a Satisfiability Problem (SAT).

Given α, β, and a graph $\mathscr{G}$ on n vertices, we construct an instance of a SAT problem that is satisfiable if and only if there exists an (α, β)-CIR of $\mathscr{G}$. An optimal pair (α, β), therefore, can be determined via a simple binary search. We use the variables $x_{u,a}$ and $y_{u,b}$, for $u \in [n]$, $a \in [\alpha]$, $b \in [\beta]$, where $x_{u,a} = 1$ and $y_{u,b} = 1$ mean that the vertex u is assigned a feature $a \in \mathscr{A} = [\alpha]$ and a feature $b \in \mathscr{B} = [\beta]$, respectively. For each edge (u, v), we want the formula

$$\left(\vee_{a \in [\alpha]} (x_{u,a} \wedge x_{v,a})\right) \wedge \left(\vee_{b \in [\beta]} (y_{u,b} \wedge y_{v,b})\right) \quad (8)$$

to be satisfiable, which is equivalent to the requirement that u and v have some common features $a \in \mathscr{A}$ and $b \in \mathscr{B}$. To turn this formula into a conjunctive form, we introduce the variable $A_{u,v,a}$ and add one more requirement that $A_{u,v,a} \leftrightarrow (x_{u,a} \wedge x_{v,a})$, which stands for

$$(\overline{A_{u,v,a}} \vee x_{u,a}) \wedge (\overline{A_{u,v,a}} \vee x_{v,a}) \wedge (A_{u,v,a} \vee \overline{x_{u,a}} \vee \overline{x_{v,a}}) . \quad (9)$$

Similarly, we include $B_{u,v,b} \leftrightarrow (y_{u,b} \wedge y_{v,b})$, which stands for

$$(\overline{B_{u,v,b}} \vee y_{u,b}) \wedge (\overline{B_{u,v,b}} \vee y_{v,b}) \wedge (B_{u,v,b} \vee \overline{y_{u,b}} \vee \overline{y_{v,b}}) . \quad (10)$$

One may hence rewrite (8) as

$$(\vee_{a \in [\alpha]} A_{u,v,a}) \wedge (\vee_{b \in [\beta]} B_{u,v,b}) . \quad (11)$$

If (u, v) is not an edge, we introduce the variables $C_{u,v}$ and $D_{u,v}$ and the following clauses

$$\overline{C_{u,v}} \vee \overline{D_{u,v}}, \quad (12)$$

$$C_{u,v} \vee \overline{x_{u,a}} \vee \overline{x_{v,a}}, \text{ for every } a \in [\alpha], \quad (13)$$

$$D_{u,v} \vee \overline{y_{u,b}} \vee \overline{y_{v,b}}, \text{ for every } b \in [\beta] . \quad (14)$$

These clauses impose the condition that u and v either have no common feature in $\mathscr{A} = [\alpha]$ or have no common feature in $\mathscr{B} = [\beta]$. Using (9)–(14), we can now create an instance of

SAT in the conjunctive normal form (CNF), which may be solved by Minisat [27]. The interested reader is referred to [28] for a related discussion on intersection representations.

C. A Simulated Annealing Algorithm for Approximate Cointersection Representation Inference

It is important to have approximate cointersection representations of a graph, especially when the graph is constructed from a real world data set, where data is usually noisy and an exact representation is, therefore, not necessary. Moreover, for large graphs, an approximate representation may still provide insight into the structure of the data, without over-representing the graphs with too many features. In this subsection, we present a randomized algorithm based on simulated annealing that produces an approximate (α, β)-cointersection representation of a graph, for any fixed pair (α, β) given as an input. We also illustrate an applications of the algorithm to a real world network and discuss the structure of overlapping communities induced by the output representation which coincides with the ground truth.

The randomized algorithm (Fig. 15) first assigns to each vertex $v \in \mathscr{V}$ a random set of $\mathscr{A}$-features, namely A_v, and a random set of $\mathscr{B}$-features, namely B_v, both of which should be nonempty. This is referred to as the feature assignment $\mathscr{L}$. Subsequently, it enters a loop of N rounds, where N is set to $b\, n \log_2(n)$ with some constant b. In each round, it chooses a random vertex u and generates two random sets $A'_u \subseteq \mathscr{A}$ and $B'_u \subseteq \mathscr{B}$. Let $\mathscr{L}'$ be the feature assignment obtained from $\mathscr{L}$ by replacing A_v by A'_u and B_u by B'_u. The *score* s of any feature assignment $\mathscr{L}$ is defined as the number of edges/non-edges of the graphs that match $\mathscr{L}$, according to the Cointersection Condition. If $s(\mathscr{L}') > s(\mathscr{L})$ then we set $\mathscr{L} := \mathscr{L}'$. Otherwise, we do it with probability $e^{c(s(\mathscr{L}')-s(\mathscr{L}))}$. We usually set c to be a constant, for example, $c = 10$ in our subsequent examples. For a more detailed discussion of the role of c in the convergence speed of the underlying Markov chain, the reader may refer to the work of Tsourakakis [8] on intersection representation of graphs. At any time, $\mathscr{L}_{\max}$ records the feature assignment with maximum score seen so far.

Example 2—We consider the social network of friendships among 34 members of an university-based Karate club, introduced by Zachary [29]. Each individual is represented by a node in the network and two nodes are joined by an edge if and only if the two corresponding individuals were consistently observed to interact outside the normal activity time of the club (Fig. 16). As a result of a dispute between the instructor (Node 1) and the club president (Node 34), the members of the clubs were split into two groups, one supporting the president and the other supporting the instructor. This fission naturally induced two communities inside the club, corresponding to the aforementioned groups. As some form of “the ground truth” community structure is known, this data set has become a well known benchmark for community detection algorithms.

Applying the randomized algorithm to this network, with $\alpha = \beta = 2$, a community structure is revealed as illustrated in Fig. 16. The set of nodes with feature a_1 corresponds to the supporters of the instructor (Node 1), while the set of nodes with feature a_2 corresponds to the supporters of the club president (Node 34). Each of these two sets is further divided into

overlapping sub-communities, marked by different colors, where the overlapping nodes, marked with a mix of two colors, correspond to the club president and the instructor. Thus, in this case, the algorithm produces an "error-free" result if we look at communities defined via features a_1 and a_2.

As demonstrated in the example, the feature set $\mathscr{A}$ is more relevant in identifying the two-part community structure of the Karate network. However, the feature set $\mathscr{B}$ reveals additional two overlapping sub-communities within each part, and hence, refines the structure of the network. Furthermore, the feature set that clusters the graph vertices into clusters with the smaller number of cross-edges may be seen as the dominant feature set, while the feature set with larger number of cross-edges may represent the feature set of lesser importance. While there is no distinction between the two feature sets in our model, this example suggests that the two feature sets may correspond to different structural aspects and be of different relevance to the community structure of the network.

Remark 1—Note that if we set $\alpha = 1$ then the randomized algorithm coincides with the algorithm developed in Tsourakakis's work [8] for intersection representation. In Example 2, if we set $\alpha = 1$ and $\beta = 2$, then the algorithm also outputs two communities that correspond perfectly to the ground truth.

Example 3—We ran the simulated annealing algorithm on the Newman-Watts small-world graphs of small and medium sizes (see Section IV-C for the definition). The three standard criteria to measure the quality of the output representation include precision, recall, and the F-score, defined as follows.

$$\text{precision} = \frac{\text{number of correct edges induced by the CIR}}{\text{total number of edges induced by the CIR}},$$

$$\text{recall} = \frac{\text{number of correct edges induced by the CIR}}{\text{total number of edges in the graph}},$$

$$\text{F} - \text{score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}.$$

The closer these are to one, the better the representation. We observe that the algorithm performs better for larger k and smaller q, which means more regularity and less randomness. This is not a surprise, as one would expect that the cointersection number θ^c is large for random graphs (this was known to be true for the intersection number θ_1), and hence, it is more likely for the simulated annealing algorithm to yield a low quality approximate representation. Indeed, on the one hand, if we use significantly fewer features than needed, then the approximate representation would contain lots of unfit edges/non-edges. On the other hand, if we use close to θ^c features, the search space becomes so large that the likelihood of reaching a good solution is reduced. It is also clear that one can trade the running time, by increasing the number of rounds N in the algorithm, for better quality

of the output, i.e. higher F-score, precision, and recall. However, note that a low F-score output can still provide useful information about the community structure of the network. For instance, for the Karate-club network in Example 2, the F-score is only 0.57. Nevertheless, the output representation already gives us the ground-truth partition of the network. In Table I, we choose $k \approx 2 \log_2(n) > \log_2(n)$, as assumed in [20]. In all cases, $\alpha = \beta \in \{5, 10, 15\}$. The number of rounds $N = b\, n \log_2(n)$, where $b \in \{500, 1000\}$.

VI. Extension to General Boolean Functions

We extend the bounds developed for the cointersection model in Section III, which is based on the AND Boolean function, to cater to models based on more general Boolean functions.

Let $f = f(x_1, x_2, \ldots, x_r)$ be a Boolean function in the *full* disjunctive normal form. In other words, the corresponding logical formula of the Boolean function is a disjunction ($\vee$) of one or more conjunctions ($\wedge$) of one or more literals, where each variables appears exactly once in every clause. Some examples are $f(x_1, x_2, x_3) = x_1 \vee (x_2 \wedge x_3)$ and $f(x_1, x_2, x_3, x_4) = (x_1 \wedge x_2) \vee (\neg x_1 \wedge x_3 \wedge x_4)$. We first discuss the meanings of the AND ($\wedge$) operator, the OR ($\vee$) operator, and the NEGATION ($\neg$) operator, and then proceed to describe the model corresponding to a general Boolean function in its full disjunctive normal form.

The AND function $f(x_1, x_2) = x_1 \wedge x_2$

Let $\mathscr{A}^1$ and $\mathscr{A}^2$ be two pairwise disjoint nonempty sets of features of cardinalities α_1 and α_2, respectively. In an $(\alpha_1 \mid \alpha_2)$-AND-intersection representation of a graph $\mathscr{G} = (\mathscr{V}, \mathscr{E})$, each vertex $v \in \mathscr{V}$ is assigned two sets $A_v^i \subseteq \mathscr{A}^i$, $i \in [2]$, such that for every $u \neq v$, $u, v \in \mathscr{V}$, it holds that $(u, v) \in \mathscr{E}$ if and only if $A_u^1 \cap A_v^1 \neq \varnothing$ and $A_u^2 \cap A_v^2 \neq \varnothing$. The AND-intersection number of $\mathscr{G}$ is the smallest number of features used, i.e. $\alpha_1 + \alpha_2$, in any $(\alpha_1 \mid \alpha_2)$-AND-intersection representation of the graph. The AND-intersection number of $\mathscr{G}$ is precisely the cointersection number of the graph.

The OR function $f(x_1, x_2) = x_1 \vee x_2$

Let $\mathscr{A}^1$ and $\mathscr{A}^2$ be two pairwise disjoint nonempty sets of features of cardinalities α_1 and α_2, respectively. In an $(\alpha_1 \mid \alpha_2)$-OR-intersection representation of a graph $\mathscr{G} = (\mathscr{V}, \mathscr{E})$, each vertex $v \in \mathscr{V}$ is assigned two sets $A_v^i \subseteq \mathscr{A}^i$, $i \in [r]$, such that for every $u \neq v$, $u, v \in \mathscr{V}$, it holds that $(u, v) \in \mathscr{E}$ if and only if $A_u^1 \cap A_v^1 \neq \varnothing$ or $A_u^2 \cap A_v^2 \neq \varnothing$. The OR-intersection number of $\mathscr{G}$ is the smallest number of features used, i.e. $\alpha_1 + \alpha_2$, in any $(\alpha_1 \mid \alpha_2)$-OR-intersection representation of the graph. Note that as $\mathscr{A}^1$ and $\mathscr{A}^2$ are disjoint, we can simply let $\mathscr{A} = \mathscr{A}^1 \cup \mathscr{A}^2$, $\alpha = \alpha_1 + \alpha_2$, and for each vertex v, let $A_v = A_v^1 \cup A_v^2$. Then an $(\alpha_1 \mid \alpha_2)$-OR-intersection representation of $\mathscr{G}$ simply corresponds to a way to assign to each vertex v a set $A_v \subseteq \mathscr{A}$ of features such that for every $u \neq v$, $u, v \in \mathscr{V}$, it holds that $(u, v) \in \mathscr{E}$ if and only if $A_u \cap A_v \neq \varnothing$. This is precisely the definition of an intersection representation of $\mathscr{G}$. Thus, the OR-intersection number of a graph is the same as its intersection number, as long as the intersection number is at least two.

NEGATION function $f(x) = \neg x$

Let $\mathscr{A}$ be a nonempty set of features of cardinality a. In an (a)-NEGATION-intersection representation, each vertex $v \in \mathscr{V}$ is assigned a set $A_v \subseteq \mathscr{A}$ such that for every $u \neq v$, $u, v \in \mathscr{V}$, it holds that $(u, v) \in \mathscr{E}$ if and only if $A_u \cap A_v = \varnothing$. The NEGATION-intersection number of $\mathscr{G}$ is the smallest number of features a used in any (a)-NEGATION-intersection representation of $\mathscr{G}$. It is immediate that this number is the same as the intersection number of the complement of $\mathscr{G}$.

Suppose we have a general Boolean function $f = f(x_1, x_2, \ldots, x_r)$ written in the full disjunctive normal form, which involves three operators $\vee$, $\wedge$, and $\neg$. Let $\mathscr{A}^1, \mathscr{A}^2, \ldots, \mathscr{A}^r$ be disjoint sets of features of cardinalities $a_1, a_2, \ldots, a_r$, respectively. In an $(a_1 \mid a_2 \mid \cdots \mid a_r)$-$f$-intersection representation of $\mathscr{G}$, each vertex $v \in \mathscr{V}$ is assigned r sets $A_v^i \subseteq \mathscr{A}^i$, $i \in [r]$, such that for every $u \neq v$, $u, v \in \mathscr{V}$, it holds that $(u, v) \in \mathscr{V}$ if and only if the intersections of the sets A_u^i and the sets A_v^i follow the rule set by the propositional formula of f. For example, when $f(x_1, x_2, x_3) = x_1 \vee (x_2 \wedge x_3)$, it is required that $(u, v) \in \mathscr{E}$ if and only if the following statement is satisfied.

$$(A_u^1 \cap A_v^1 \neq \varnothing) \vee \left((A_u^2 \cap A_v^2 \neq \varnothing) \wedge (A_u^3 \cap A_v^3 \neq \varnothing)\right).$$

In words, u and v are adjacent if and only if they share either an $\mathscr{A}^1$-label or both an $\mathscr{A}^2$-label and an $\mathscr{A}^3$-label. For another example, take $f(x_1, x_2, x_3, x_4) = (x_1 \wedge x_2) \vee (\neg x_1 \wedge x_3 \wedge x_4)$. Then in a corresponding representation of $\mathscr{G}$, two vertices are adjacent if and only if either of the following two cases happens: (1) they share both an $\mathscr{A}^1$-label and an $\mathscr{A}^2$-label; or (2) they do not share any $\mathscr{A}^1$-label, but they share both an $\mathscr{A}^3$-label and an $\mathscr{A}^4$-label. The *f-intersection number* of $\mathscr{G}$ is defined to be the smallest number of features used, namely $\sum_{i=1}^{r} a_i$, in any $(a_1 \mid a_2 \mid \cdots \mid a_r)$-$f$-intersection representation of the graph.

It is not immediately clear that the negation function has sufficiently strong relevance as the AND and OR functions in the context of social network analysis. Hence, we focus on Boolean functions that involve $\vee$ and $\wedge$ operations only and provide the following proposition generalizing Lemma 3.

Proposition 6—Let $f = f(x_1, x_2, \ldots, x_r)$ be a Boolean function in the full disjunctive normal consisting only of $\vee$ and $\wedge$. Let $g_f = g_f(a_1, a_2, \ldots, a_r)$ be an integer-valued function on r non-negative integral variables $a_1, a_2, \ldots, a_r$, obtained from f by replacing x_i by a_i ($i \in [r]$), $\vee$ by $+$, and $\wedge$ by $\times$. Then the f-intersection number of a graph $\mathscr{G}$ is bounded from below by the optimal value of the objective function of the integer programming problem given below:

$$\begin{aligned} (IP)\ \ minimize\ \ & \sum_{i=1}^{r} \alpha_i \\ subject\ to\ \ & g_f(\alpha_1, \alpha_2, \ldots, \alpha_r) \geq \theta_1(\mathscr{G}), \\ & \mathbb{Z} \ni \alpha_i \geq 1, \forall i \in [r]. \end{aligned}$$

Proof: Suppose that we have an $(a_1 \mid a_2 \mid \cdots \mid a_r)$-$f$-intersection representation of the graph $\mathscr{G}$ with the corresponding sets of labels $\mathscr{A}^1, \mathscr{A}^2, \ldots, \mathscr{A}^r$. For any clause $x_{i_1} \wedge x_{i_2} \wedge \cdots \wedge x_{i_s}$ of f, a tuple $(a^{i_1}, a^{i_2}, \ldots, a^{i_r})$ where $a^{i_j} \in \mathscr{A}^{i_j}$ corresponds to a clique in $\mathscr{G}$, which consists of all vertices $v \in \mathscr{V}$ that have $a^{i_1}, a^{i_2}, \ldots, a^{i_r}$ in their feature sets. Note that there are in total $g_f(a_1, a_2, \ldots, a_r)$ such cliques. As each edge of $\mathscr{G}$ must belong to one of these cliques, these cliques form an edge clique cover of $\mathscr{G}$. Therefore, $g_f(a_1, a_2, \ldots, a_r) \geq \theta_1(\mathscr{G})$.

If we ignore the condition that $a_i \in \mathbb{Z}$ in the integer programming problem (IP) stated in Proposition 6, we obtain a real-valued programming problem, referred to as (P). An optimal solution to (P) also provides a lower bound on the f-intersection number of the graph. Generally, we can find necessary conditions for a solution of (P) to exist by using either the method of Lagrange multipliers or the Karush-Kuhn-Tucker (KKT) conditions. We illustrate this observation with the following example.

Example 4—Let $f(x_1, x_2, x_3) = x_1 \vee (x_2 \wedge x_3)$. Using the notation in Proposition 6, $g_f(a_1, a_2, a_3) = a_1 + a_2 a_3$. Then the optimal value of the objective function of the following programming problem serves as a lower bound for the f-intersection number of a graph $\mathscr{G}$:

$$\begin{aligned} \text{(P)} \quad & \text{minimize} \quad \alpha_1 + \alpha_2 + \alpha_3 \\ & \text{subject to} \quad \alpha_1 + \alpha_2\alpha_3 \geq \theta_1(\mathscr{G}), \\ & \qquad \mathbb{R} \ni \alpha_i \geq 1, \forall i \in [3]. \end{aligned}$$

In order to use the method of Lagrange multipliers, we first introduce the slack variables β_i, $i \in [4]$, to convert the inequality constraints into equality constraints as follows. The constraint $a_i \geq 1$ is converted into the new constraint $\alpha_i - \beta_i^2 - 1 = 0$, for each $i \in [3]$, and the constraint $a_1 + a_2 a_3 \geq \theta_1$ is converted into the new constraint $\alpha_1 + \alpha_2\alpha_3 - \beta_4^2 - \theta_1 = 0$. Let λ_i, $i \in [4]$, be the Lagrange multipliers. We formulate the Lagrangian

$$\mathscr{L}(\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3, \beta_4, \lambda_1, \lambda_2, \lambda_3, \lambda_4) = \sum_{i=1}^{3} \alpha_i + \sum_{i=1}^{3} \lambda_i(\alpha_i - \beta_i^2 - 1) + \lambda_4(\alpha_1 + \alpha_2\alpha_3 - \beta_4^2 - \theta_1).$$

The method of Lagrange multipliers states that if we examine all stationary points of the Lagrangian, at which $\nabla\mathscr{L} = \mathbf{0}$, where $\nabla\mathscr{L} = \left(\frac{\partial\mathscr{L}}{\partial\alpha_1}, \ldots, \frac{\partial\mathscr{L}}{\partial\alpha_3}, \frac{\partial\mathscr{L}}{\partial\beta_1}, \ldots, \frac{\partial\mathscr{L}}{\partial\beta_4}, \frac{\partial\mathscr{L}}{\partial\lambda_1}, \ldots, \frac{\partial\mathscr{L}}{\partial\lambda_4}\right)$, then the one that leads to the minimum objective value $\sum_{i=1}^{3} \alpha_i$ is an optimal solution to (P). Therefore, using this method, we arrive at the following system $\nabla\mathscr{L} = \mathbf{0}$ of equations:

$$\begin{cases} 1+\lambda_1+\lambda_4=0, \\ 1+\lambda_2+\lambda_4\alpha_3=0, \\ 1+\lambda_3+\lambda_4\alpha_2=0, \\ \lambda_i\beta_i=0, \quad i\in[4], \\ \alpha_i-\beta_i^2-1=0, \quad i\in[3], \\ \alpha_1+\alpha_2\alpha_3-\beta_4^2-\theta_1=0\,. \end{cases} \qquad \text{(15a), (15b), (15c), (15d), (15e), (15f)}$$

A straightforward way to obtain all the solutions of the system (15) is by examining all 16 cases, each of which captures whether $\lambda_i = 0$ or $\beta_i = 0$, $i \in [4]$ (from (15d)). We can ignore certain cases due to symmetry. As a consequence, we find that the objective function $\sum_{i=1}^{3}\alpha_i$ is minimized when $\alpha_1 = 1$ and $\alpha_2 = \alpha_3 = \sqrt{\theta_1 - 1}$, which gives us the lower bound $1 + 2\sqrt{\theta_1 - 1}$ on the f-intersection number of $\mathscr{G}$.

Another example we considered is $f = (x_1 \wedge x_2) \vee (x_1 \wedge x_3) \vee (x_2 \wedge x_3)$. Again, applying the method of Lagrange multipliers and Proposition 6, it may be shown that the f-intersection number of $\mathscr{G}$ is at least $\sqrt{3\theta_1}$.

An upper bound on the f-intersection number of a graph of bounded degree, where f only involves the $\vee$ and $\wedge$ operations, may be obtained in the same way as that for the cointersection number, in Theorem 5. We present this fact below.

Theorem 9—Let $\mathscr{G}$ be a graph on n vertices with $\Delta(\mathscr{G}) \le d$. Let $f = f(x_1, x_2, \ldots, x_r)$ be a Boolean function in the full disjunctive normal consisting of only $\vee$ and $\wedge$. Let s be the largest number of literals that appear in any clause of f. Then the f-intersection number of $\mathscr{G}$ is at most $c(d, r, s)n^{1/s} + r - s$, where $c(d, r, s)$ is a function of d, r, and s.

Proof: We can assume that no clause $\mathscr{C}'$ of f is a sub-clause of another clause $\mathscr{C}$ (i.e., that all of the literals of $\mathscr{C}'$ also appear in $\mathscr{C}$), as otherwise we can always remove $\mathscr{C}'$ and obtain an equivalent formula of f.

Now let $\mathscr{C}$ be a clause of f with s literals, referred to as the *leading* clause. Relabeling the indices if necessary, we can assume that $\mathscr{C} = \wedge_{i=1}^{s} x_i$. Let $A^1, \ldots, A^r$ be r pairwise disjoint sets of features such that $a_i \triangleq |A^i| = c'(d, r, s)n^{1/s}$ for $i \in [s]$, while $a_j \triangleq |A^j| = 1$ for all $j > s$, $j \in [r]$. Here $c'(d, r, s)$ is a function of d, r, and s, which will be determined later. Similar to the proof of Theorem 5, we show that there exists an $(a_1 \mid a_2 \mid \cdots \mid a_r)$-$f$-intersection representation of $\mathscr{G}$ by invoking the Lovász Local Lemma [30]. As a consequence, the f-intersection number of $\mathscr{G}$ is at most $c(d, r, s)n^{1/s} + r - s$, where $c(d, r, s) \triangleq sc'(d, r, s)$.

We independently assign to every edge e of $\mathscr{G}$ a randomly chosen set of features $\{a^1(e), a^2(e), \ldots, a^s(e)\}$. Note that we do not assign to e any label $a^j \in \mathscr{A}^j$, for $j > s$. For every vertex $v \in \mathscr{V}$ and for every $i \in [r]$, let

$$A_v^i = \{a^i(e) : e = (u, v) \in \mathscr{E}\}.$$

Then $A_v^j = \varnothing$ for $j > s$. Hence, $A_u^j \cap A_v^j \neq \varnothing$ for every $u \neq v$ and $j > s$. Moreover, we know that for any clause $\mathscr{C}' \neq \mathscr{C}$, there must exist a $j > s$ such that $\mathscr{C}'$ contains x_j, for otherwise, $\mathscr{C}'$ would be a sub-clause of $\mathscr{C}$. Therefore, this feature assignment is an f-intersection representation of $\mathscr{G}$ if and only if for every $u \neq v$, $u, v \in \mathscr{E}$, it holds that

$$(u, v) \in \mathscr{V} \Leftrightarrow A_u^i \cap A_v^i \neq \varnothing, \text{ for all } i \in [s]. \quad (16)$$

In other words, we can focus only on the leading clause $\mathscr{C} = \vee_{i=1}^{s} x_i$ and ignore all other clauses of f.

It is clear that (16) is satisfied for all pairs $(u, v) \in \mathscr{E}$. We now define for each pair $(u, v) \notin \mathscr{E}$ a bad event $E_{u,v}$ where $A_u^i \cap A_v^i \neq \varnothing$ for all $i \in [s]$. The goal is to show that there exists a function $c'(d, r, s)$ of d, r, and s, so that $PD \leq 1/4$, where $\text{Prob}(E_{u,v}) \leq P$ and each bad event is dependent on at most D other bad events. Then by the Lovász Local Lemma [30], we may conclude that there exists a way to assign features to the edges of $\mathscr{G}$ that leads to an f-intersection representation of $\mathscr{G}$. Just as in the proof of Theorem 5, we have

$$\text{Prob}(E_{u,v}) = \prod_{i=1}^{s} \text{Prob}(A_u^i \cap A_v^i \neq \varnothing) \leq P \triangleq \left(\frac{d^2}{\alpha_i - d + 1}\right)^s = \frac{d^{2s}}{(c'(d, r, s)n^{1/s} - d + 1)^s}.$$

We also have $D = 2n(d + 1)$. It is straightforward to verify that for $c'(d, r, s) \triangleq (8d^{2s+2})^{1/s} + d - 1$, we have $PD \leq 1/4$.

VII. Discussion

We established a new Boolean feature model to study the complex community structure in networks. This model allows for the discovery of overlapping and hierarchical communities, thanks to the use of different feature sets, each of which may play a different role in highlighting the overall network structure. The newly developed concept of Boolean intersection representation, in particular, cointersection representation, generalizes the well-known intersection representation (a.k.a. edge clique cover) in the literature of graph theory. We obtained tight lower and upper bounds on the cointersection numbers of numerous families of graphs, and also showed that a graph as simple as a path can have exponentially many optimal cointersection representations. We also developed an exact algorithm and a heuristic algorithm for finding optimal and approximate cointersection representations of a graph. The latter was evaluated on the standard Karate-club network and on the Newman-Watts small-world graphs. A number of problems remain open for future research.

Problem 1

Find a generic upper bound on the cointersection number $\theta^{\mathcal{F}}$, similar to the one on the intersection number (Theorem 1). The evidence suggests that $\theta^{\mathcal{F}}(\mathcal{G}) \leq n$ for all G.

Problem 2

Find matching lower bounds and upper bounds for the cointersection numbers of random graphs in various models: Erdös-Rényi, Newman-Watts, Watts-Strogatz, and the preferential attachment model. The upper bound obtained in this work for the random Newman-Watts graphs can most likely be tightened.

Problem 3

Develop more scalable algorithms for exact and approximate cointersection/Boolean intersection representations of large-sized graphs. The objective is to test the new model with large-scale real-world networks. The current algorithms only run well on small graphs: at most 20 vertices for the *exact* SAT-based algorithm and a few thousand vertices for the simulated annealing algorithm. It would be interesting to see if the data reduction technique developed for the intersection number by Gramm *et al.* [31] could be extended to the cointersection number.

Problem 4

Extend the current model to include continuous features. The notion of continuous features has been studied before in the literature, e.g. in the context of interval graphs and unit disk graphs. For instance, an interval graph (see, for example, [32]) is a graph where each vertex is assigned an interval on the real line and an edge exists between any two vertices if and only if the two corresponding intervals intersect. It is of interest to study the generalizations of these graphs under the general Boolean intersection model.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank Gregory J. Puleo and Charalampos Tsourakakis for helpful discussions. This work was funded by NIH BD2K Grant 1U01CA198943-01 and NSF Grant IOS 1339388 and NSF 239 SBC Purdue 41010-38050.

References

1. Meeds E, Ghahramani Z, Neal RM, Roweis ST. Modeling dyadic data with binary latent factors. Proc. Adv. Neural Inf. Process. Syst. (NIPS). 2006:977–984.
2. Airoldi EM, Blei DM, Fienberg SE, Xing EP. Mixed membership stochastic blockmodels. J. Mach. Learn. Res. 2008; 9:1981–2014. [PubMed: 21701698]
3. Miller K, Jordan MI, Griffiths TL. Nonparametric latent feature models for link prediction. Proc. Adv. Neural Inf. Process. Syst. (NIPS). 2009:1276–1284.
4. Palla K, Knowles DA, Ghahramani Z. An infinite latent attribute model for network data; Proc. Int. Conf. Mach. Learn. (ICML); 2012.
5. Kim M, Leskovec J. Modeling social networks with node attributes using the multiplicative attribute graph model; Proc. Conf. Uncertain. Artificial Intelligence (UAI); 2011. 400–409.

6. Yang J, Leskovec J. Community-affiliation graph model for overlapping network community detection; Proc. IEEE Int. Conf. Data Min. (IDCM); 2012. 1170–1175.
7. Yang J, Leskovec J. Overlapping community detection at scale: A nonnegative matrix factorization approach; Proc. ACM Int. Conf. Web Search Data Min. (WSDM); 2013. 587–596.
8. Tsourakakis C. Provably fast inference of latent features from networks: With applications to learning social circles and multilabel classification; Proc. Int. Conf. World Wide Web (WWW); 2015. 1111–1121.
9. Erdös P, Goodman AW, Pósa L. The representation of a graph by set intersections. Canad. J. Math. 1966; 18(1):106–112.
10. Hefner KA, Jones KF, Kim S, Lundgren JR, Roberts FS. (i, j) competition graphs. Discrete Applied Mathematics. 1991; 32(3):241–262.
11. Chung MS, West DB. The *p*-intersection number of a complete bipartite graph and orthogonal double coverings of a clique. Combinatorica. 1994; 14(4):453–461.
12. Bonchi F, Gionis A, Ukkonen A. Overlapping correlation clustering; Proc. IEEE Int. Conf. Data Min. (IDCM); 2011. 51–60.
13. Pullman NJ. Combinatorial Mathematics X, ser. Lect. Notes Math. Vol. 1036. Springer; Berlin Heidelberg: 1983. Clique coverings of graphs A survey; 72–85.
14. Roberts FS. Applications of edge coverings by cliques. Discrete Appl. Maths. 1985; 10(1):93–109.
15. Eschenauer L, Gligor VD. A key-management scheme for distributed sensor networks; Proc. ACM Conf. Comput. Comm. Security; 2002. 41–47.
16. Eaton N, Gould RJ, Rödl V. On *p*-intersection representations. J. Graph Theory. 1996; 21(4):377–392.
17. Alon N. Covering graphs by the minimum number of equivalence relations. Combinatorica. 1986; 6(3):201–206.
18. Erdös P, Ordman ET, Zalcstein Y. Clique partitions of chordal graphs. Combin. Probab. Comput. 1993; 2(04):409–415.
19. Erdös P, Lovász L. Problems and results on 3-chromatic hypergraphs and some related questions. In: Hajna A, , et al., editorsInfinite and Finite Sets. Colloquia Mathematica Societatis János Bolyai; 1975. 609–627.
20. Watts DJ, Strogatz SH. Collective dynamics of ’small-world’ networks. Nature. 1998; 393:440–442. [PubMed: 9623998]
21. Erdös P, Rényi A. On random graphs. Publicationes Mathematicae. 1959; 6:290–297.
22. Newmand MEJ, Watts DJ. Renormalization group analysis of the small-world network model. Physics Letters A. 1999; 263:341–346.
23. Colbourn CJ, Dinitz JH. Handbook of Combinatorial Designs, Second Edition (Discrete Mathematics and Its Applications). Chapman & Hall/CRC; 2006.
24. Alter R, Wang CC. Uniquely intersectable graphs. Discrete Math. 1977; 18(3):217–226.
25. Mahadev N, Wang T-M. On uniquely intersectable graphs. Discrete Math. 1999; 207(1–3):149–159.
26. Orlin J. Contentment in graph theory: Covering graphs with cliques. Indagationes Mathematicae (Proceedings). 1977; 80(5):406–424.
27. Eén N, Sörensson N. An extensible SAT-solver; Proc. 6th Int. Conf. Theory Appl. Satisf. Testing; 2003.
28. Berg J, Jarvisalo M. Optimal Correlation Clustering via MaxSat; Proc. IEEE Int. Conf. Data Min Workshops (ICDMW); 2013. 750–757.
29. Zachary WW. An information flow model for conflict and fission in small groups. Journal of Anthropological Research. 1977; 33:452–473.
30. Lovász L. On coverings of graphs. In: Hajna A, , et al., editorsProceedings of the Colloquium held at Tihany, Hungary. 1968. 231–236.
31. Gramm J, Guo J, Hüffner F, Niedermeier R. Data reduction and exact algorithms for clique cover. J. Exp. Algorithmics. 2009; 13:2:2.2–2:2.15.
32. Golumbic MC. Interval graphs and related topics. Discrete Mathematics. 1985; 55(2):113–121.

Biographies

Hoang Dau received the B.S. degree in applied mathematics and informatics from Vietnam National University, Hanoi, Vietnam, in 2006, and the M.S. and Ph.D. degrees in mathematical sciences from Nanyang Technological University, Singapore, in 2009 and 2012, respectively. He is currently a Post-doctoral Research Associate with the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign. His research interests include coding theory, network coding, distributed storage systems, and combinatorics.

Olgica Milenkovic is a professor of Electrical and Computer Engineering at the University of Illinois, Urbana-Champaign (UIUC), and Research Professor at the Coordinated Science Laboratory. She obtained her Masters Degree in Mathematics in 2001 and PhD in Electrical Engineering in 2002, both from the University of Michigan, Ann Arbor. Prof. Milenkovic heads a group focused on addressing unique interdisciplinary research challenges spanning the areas of algorithm design and computing, bioinformatics, coding theory, machine learning and signal processing. In 2013, she was elected a UIUC Center for Advanced Study Associate and Willett Scholar. In 2015, she became Distinguished Lecturer of the Information Theory Society.

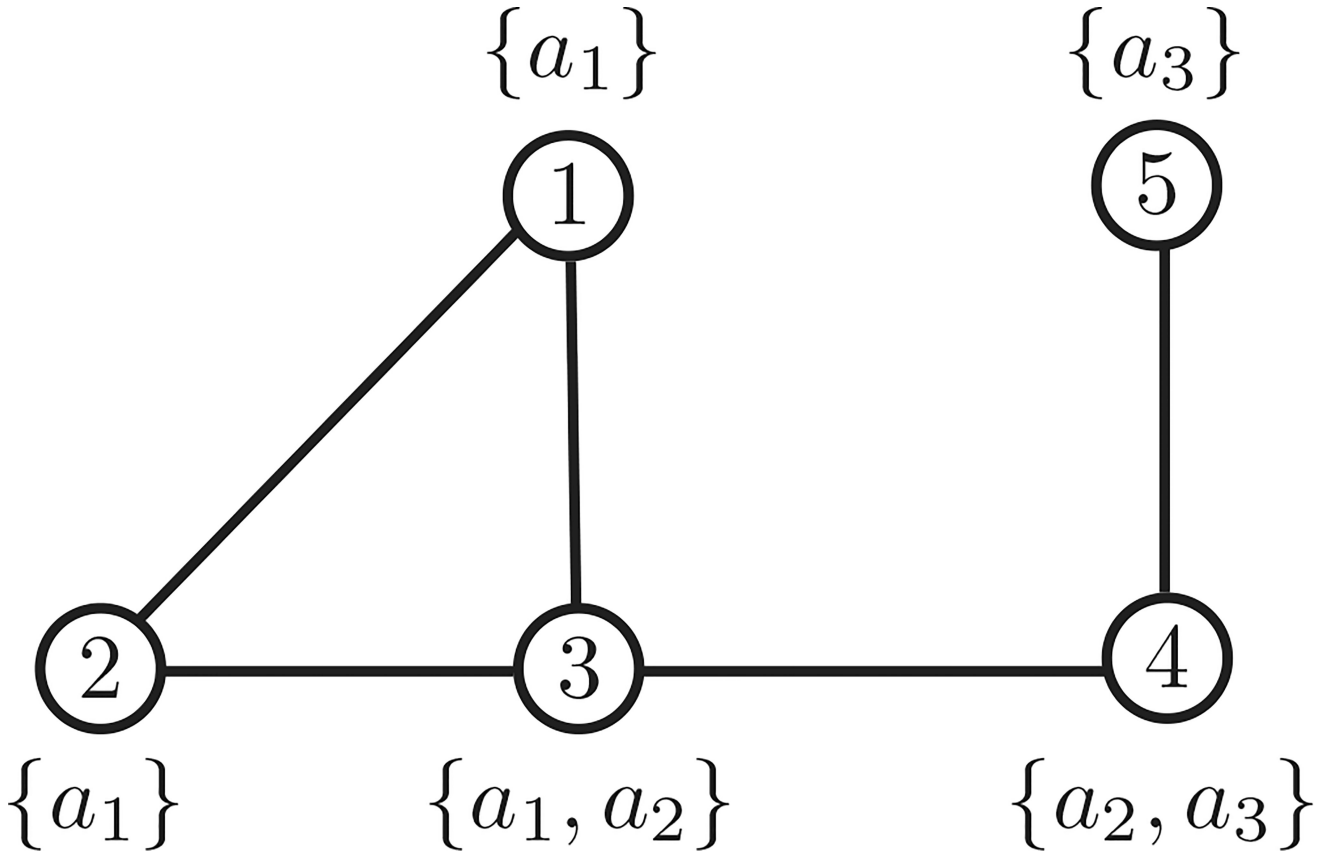


Fig. 1.

Illustration of an intersection representation of a graph from [8]. Vertices are assigned subsets from the feature set $\mathscr{A} = \{a_1, a_2, a_3\}$ so that two vertices are adjacent if and only if they share at least one common feature. In this case, the intersection number is three.

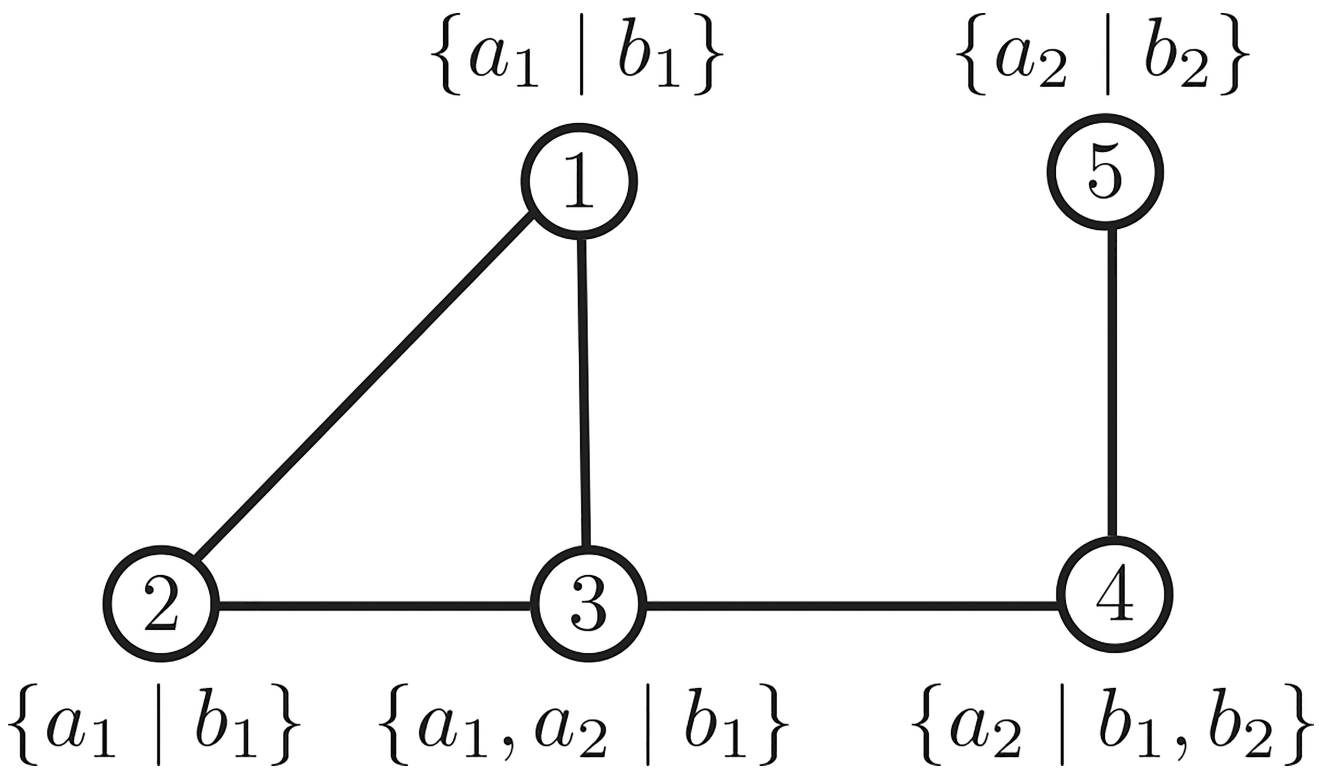


Fig. 2.

Each node is assigned a set of features from $\mathscr{A} = \{a_1, a_2\}$ and a set of features from $\mathscr{B} = \{b_1, b_2\}$. Two nodes are connected by an edge if and only if they share at most one feature from $\mathscr{A}$ and one feature from $\mathscr{B}$.

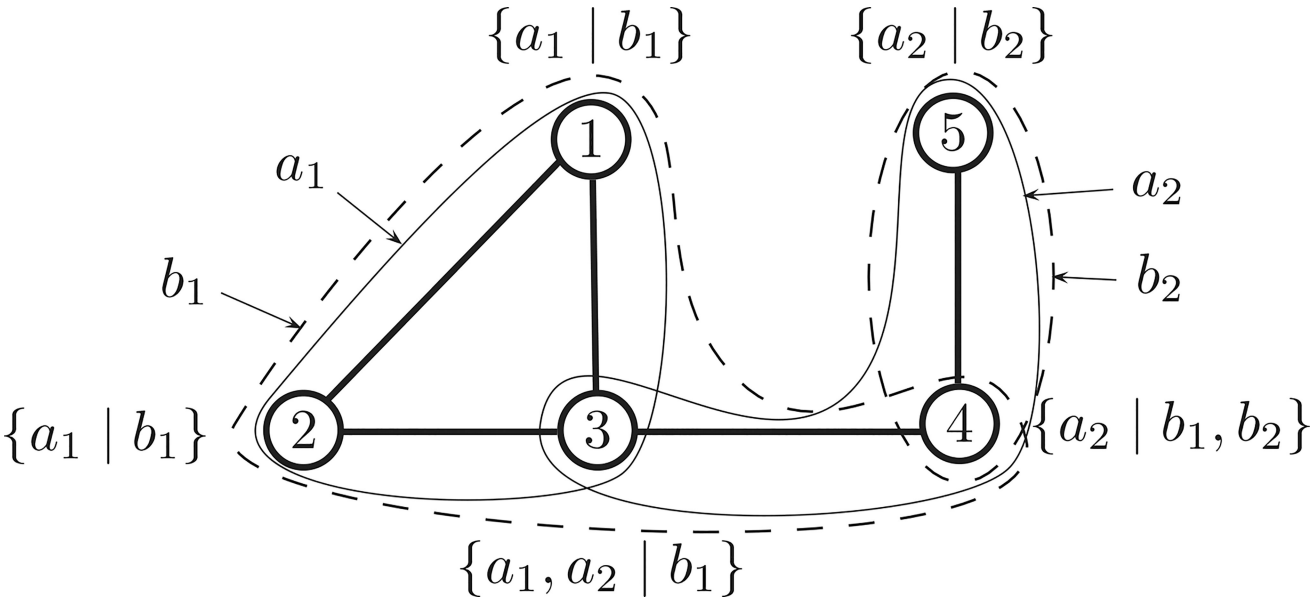


Fig. 3.

The community structure induced by the features in a cointersection representation of the graph. The vertices are grouped into different communities, each of which corresponds to an $\mathscr{A}$-feature (solid closed curve) or a $\mathscr{B}$-feature (dashed closed curve). The pair (u, v) is an edge if and only if both u and v belong to a common $\mathscr{A}$-community and a common $\mathscr{B}$-community. In other words, every edge lies inside both a solid curve and a dashed curve.

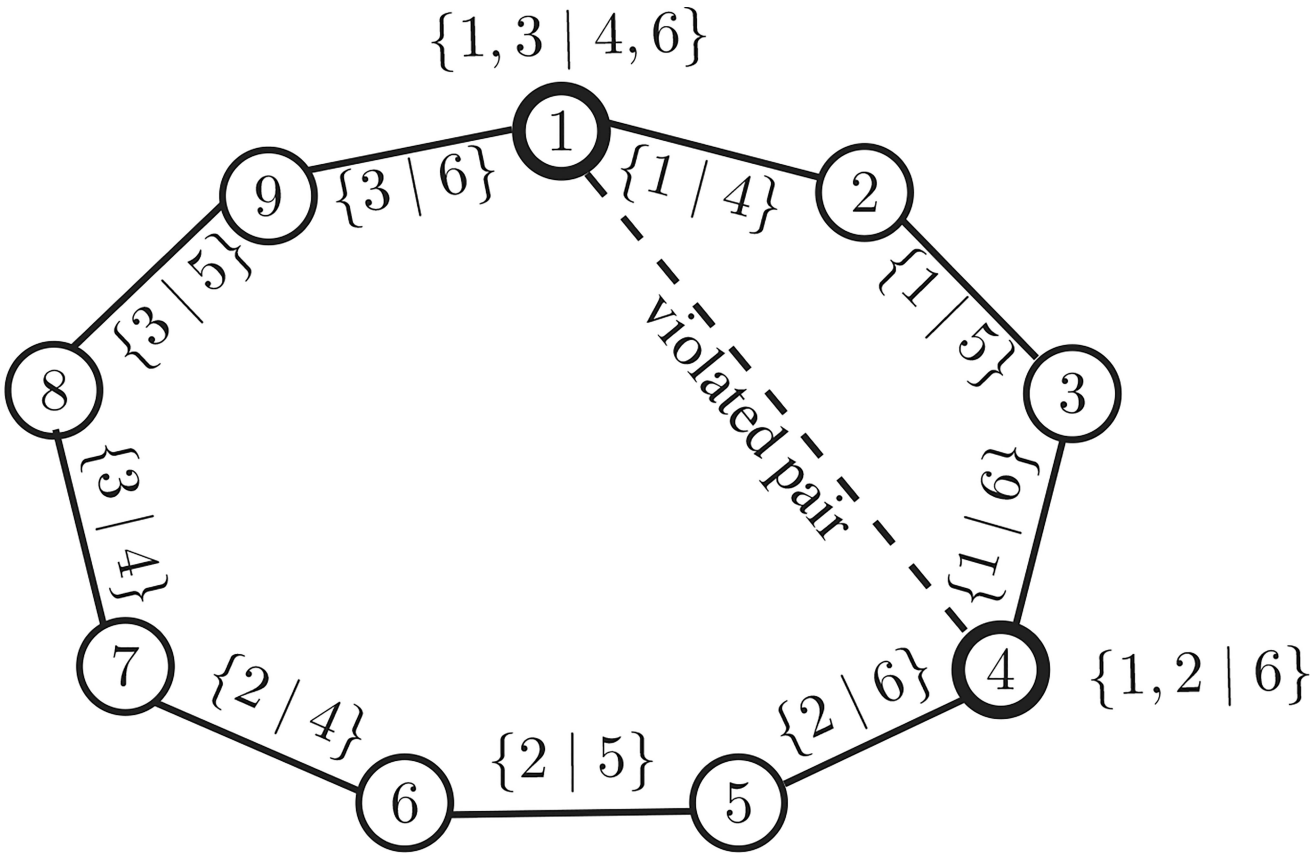


Fig. 4.

An example where the discussed feature assignment for paths does not apply for the case of a cycle, say $\mathscr{C}_9$. Two vertices 1 and 4 share a pair of common features {1 | 6}, even though they are not adjacent. Here we set $\mathscr{A}$ = {1, 2, 3} and $\mathscr{B}$ = {4, 5, 6}.

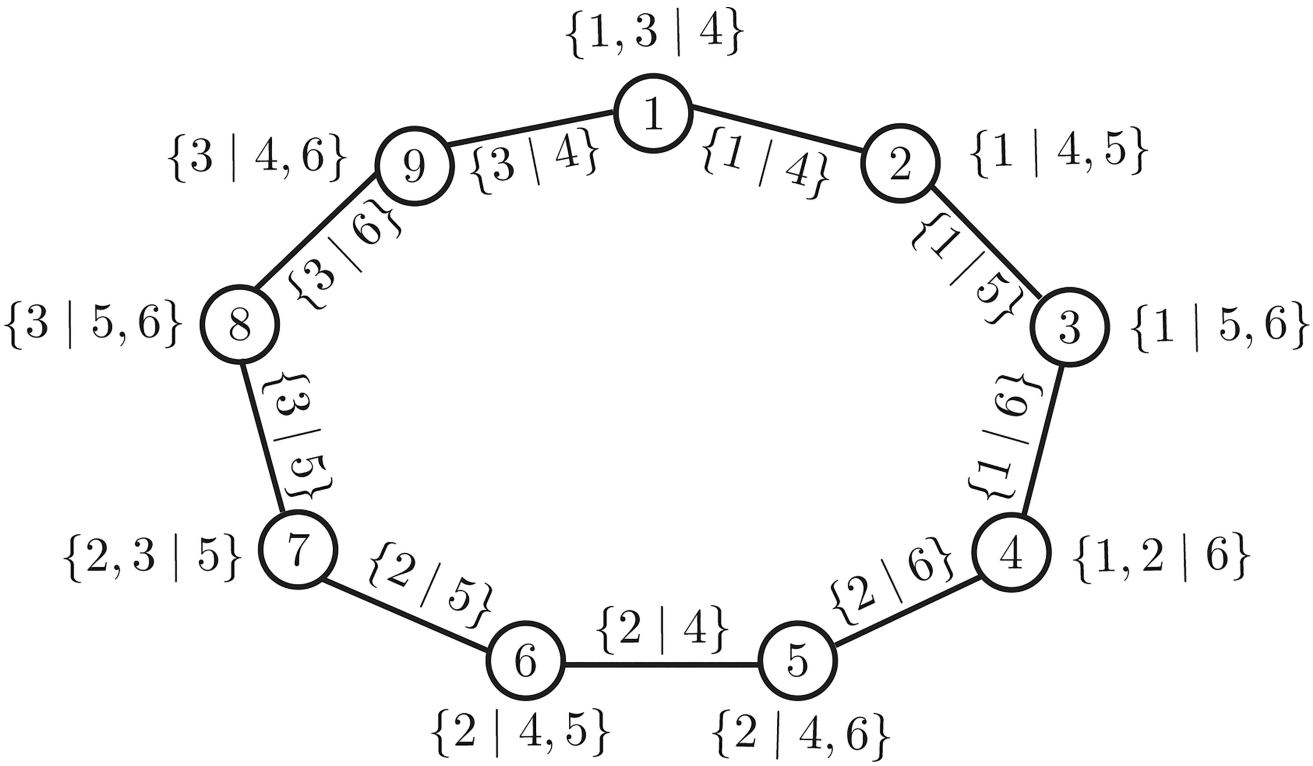


Fig. 5.

An example of a (3 | 3)-cointersection representation of $\mathscr{C}_9$. Here we set $\mathscr{A}$ = {1, 2, 3} and $\mathscr{B}$ = {4, 5, 6}.

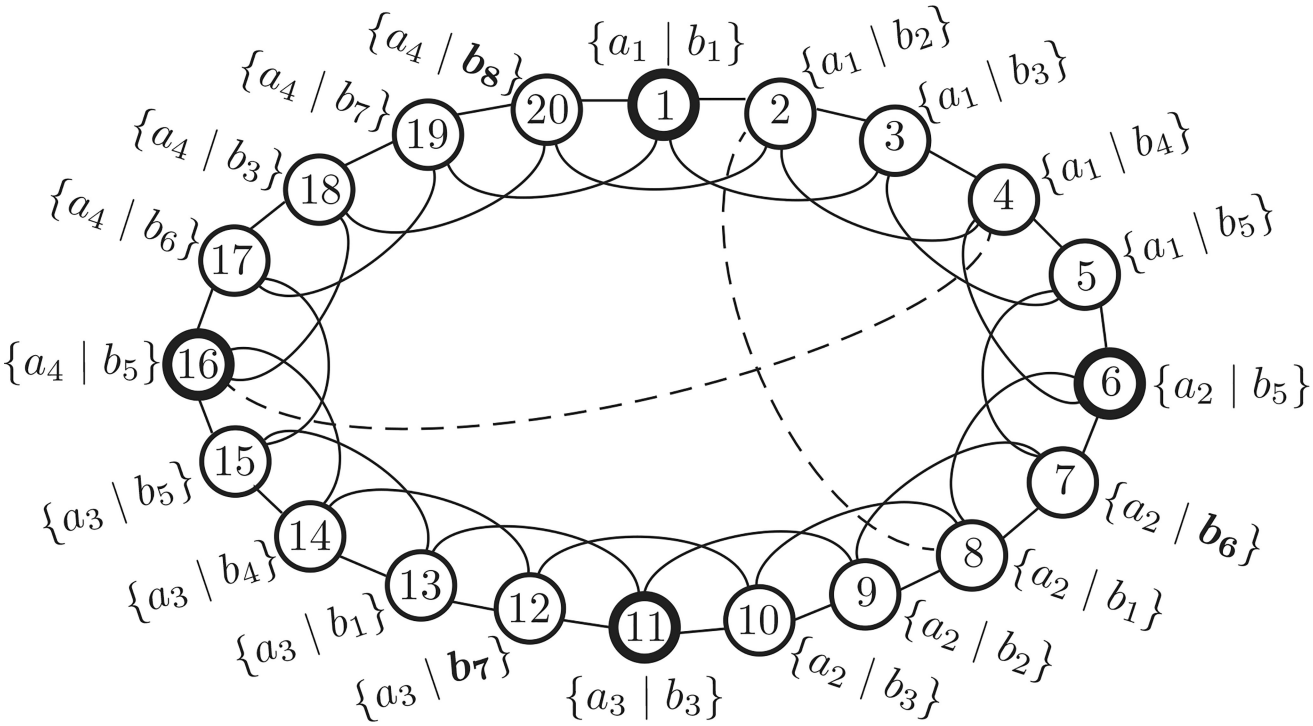


Fig. 6.

A (4 | 8)-cointersection representation of the ring lattice $\mathcal{L}(20, 2)$. Here $\mathcal{A} = \{a_1, \ldots, a_4\}$, $\mathcal{B} = \{b_1, \ldots, b_8\}$. The pair of features assigned to each 3-clique C_i is given at vertex i. The feature sets assigned to each vertex are unions of those assigned to all three cliques containing that vertex. The bold vertices correspond to the first clique in each group, where the Transition Rule applies. The bold $\mathcal{B}$-features signal where the Novel-Feature or the Inverse Rule applies. Adding random shortcuts (dashed) produces a Newman-Watts random graph.

	C_1	C_2	C_3	C_4	C_5	C_6
R_1	b_1	b_2	b_3	b_4	b_5	b_6
R_2	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}
R_3	b_{13}	b_{14}	b_{15}	b_{16}	b_{17}	b_{18}

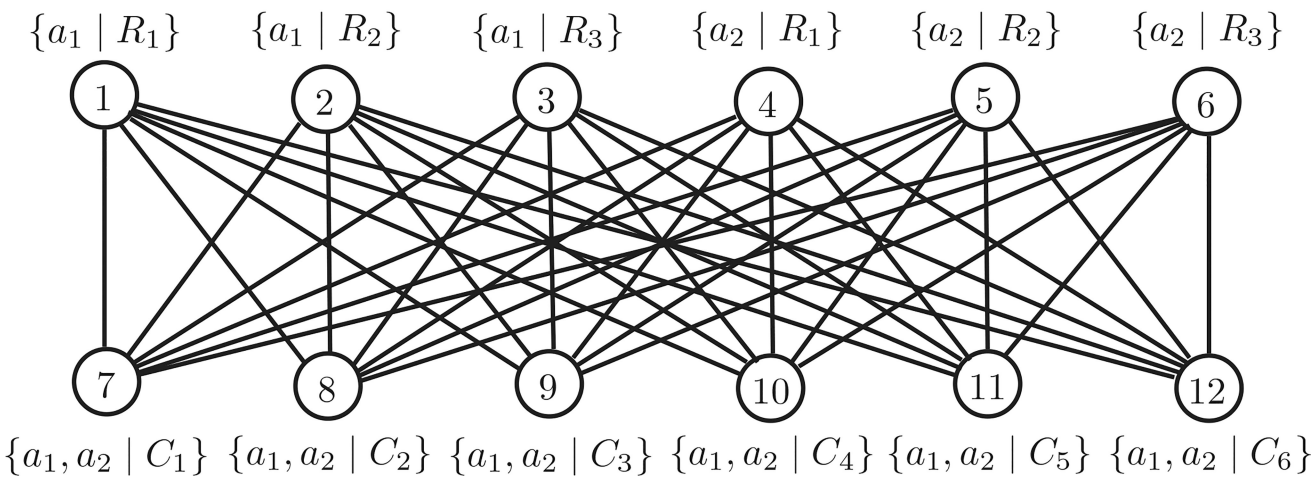


Fig. 7.

A (2, 18)-CIR of $\mathscr{K}_{6,6}$. The sets R_1, R_2, and R_3 are pairwise disjoint. The sets $C_1, \ldots, C_6$ are also pairwise disjoint. Each pair of sets R_i and C_j has an intersection of size one. Both R_i's and C_j's are subsets of $[b_1, \ldots, b_{18}]$.

$\{1,2,3\}$	$\{1,4,7\}$	$\{1,5,9\}$	$\{1,6,8\}$
$\{4,5,6\}$	$\{2,5,8\}$	$\{2,6,7\}$	$\{2,4,9\}$
$\{7,8,9\}$	$\{3,6,9\}$	$\{3,4,8\}$	$\{3,5,7\}$

Fig. 8.
A 2-(9, 3, 1) resolvable packing with four parallel classes.

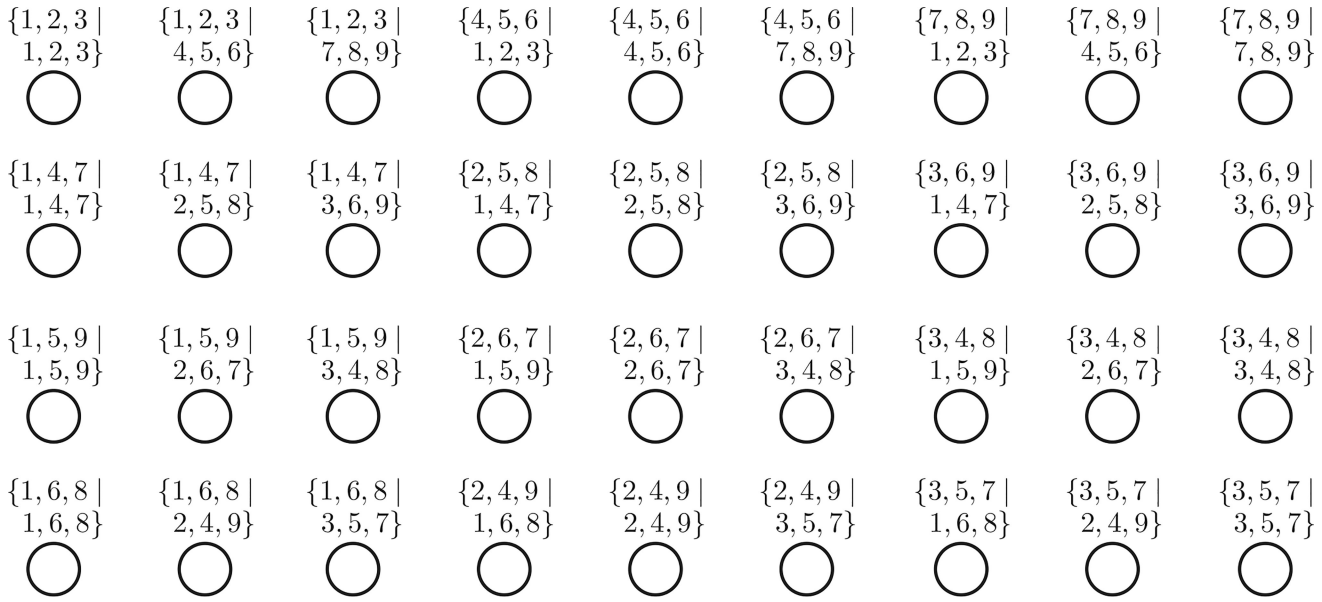


Fig. 9.

An optimal (9, 9)-CIR of $\mathscr{K}_{9,9,9,9}$ via a 2-(9, 3, 1) resolvable packing with four classes. In fact, this is a 2-(9, 3, 1) resolvable design, which is also an affine plane of order 9.

$\{1,2,3,4\}$	$\{1,5,9,13\}$	$\{1,6,11,16\}$
$\{5,6,7,8\}$	$\{2,6,10,14\}$	$\{2,7,12,13\}$
$\{9,10,11,12\}$	$\{3,7,11,15\}$	$\{3,8,9,14\}$
$\{13,14,15,16\}$	$\{4,8,12,16\}$	$\{4,5,10,15\}$

Fig. 10.
A 2-(16, 4, 1) resolvable packing with three parallel classes.

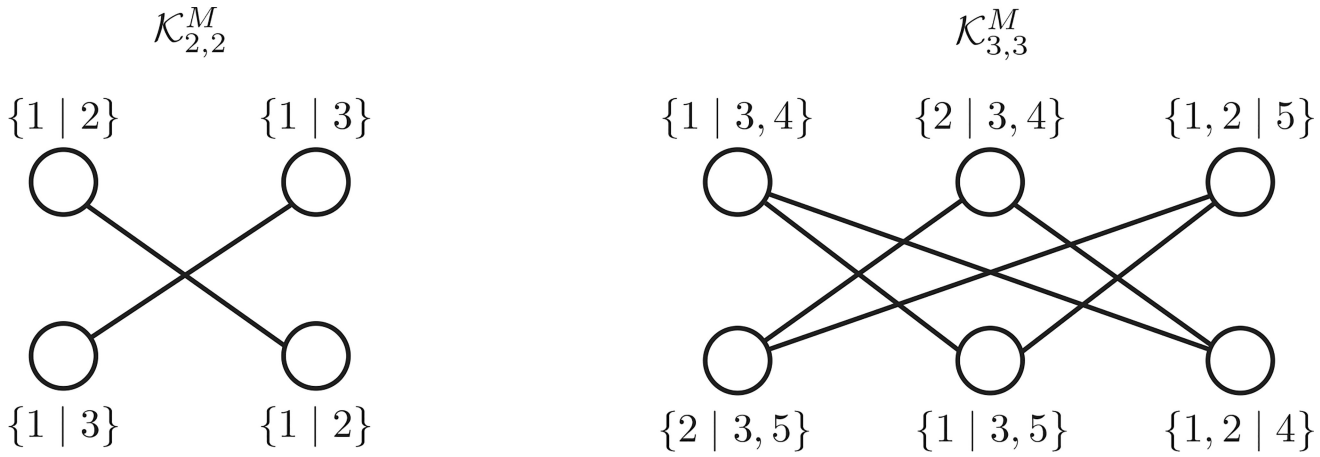


Fig. 11.

A (1, 2)-CIR of $\mathscr{K}^M_{2,2}$ (left) and a (2, 3)-CIR of $\mathscr{K}^M_{3,3}$ (right).

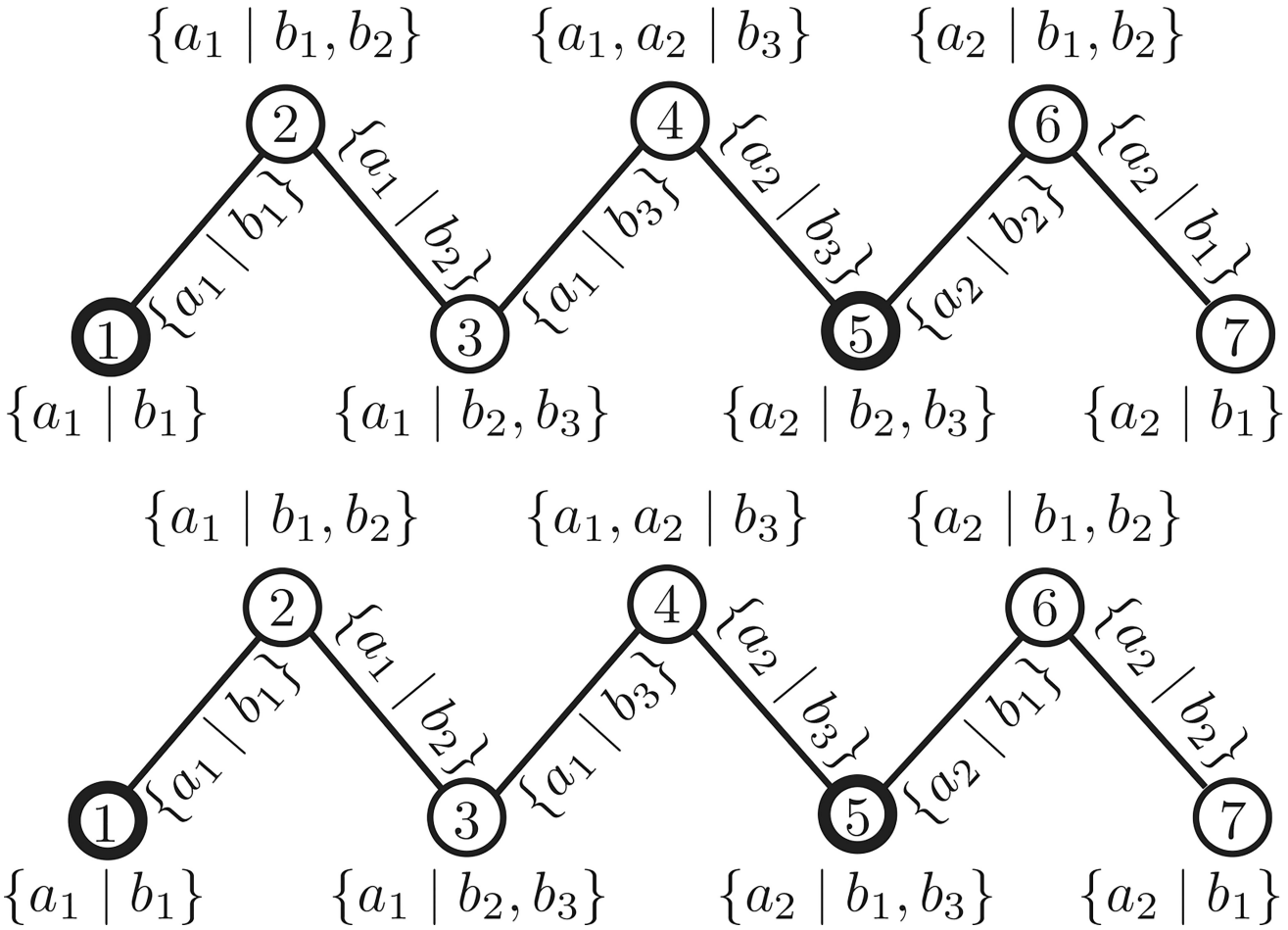


Fig. 12.

An illustration of two nonequivalent, optimal (2, 3)-CIRs of the path $\mathscr{P}_7$. In the first (top) representation, vertex 1 and vertex 5 do not share any features, while in the second (bottom) representation, they do share one feature, b_1.

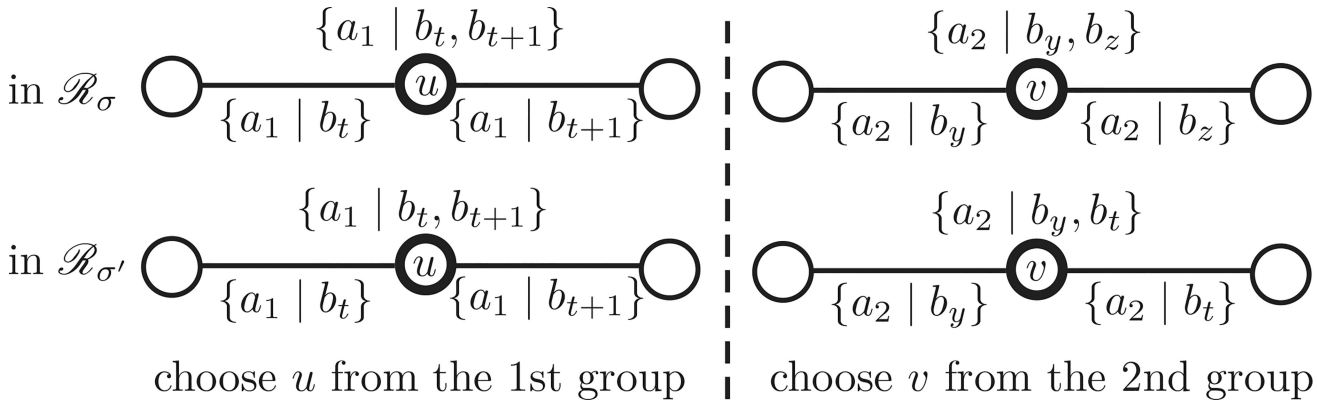


Fig. 13.

(Case 1) The feature sets of u and v with respect to $\mathscr{R}_\sigma$ and $\mathscr{R}_{\sigma'}$.

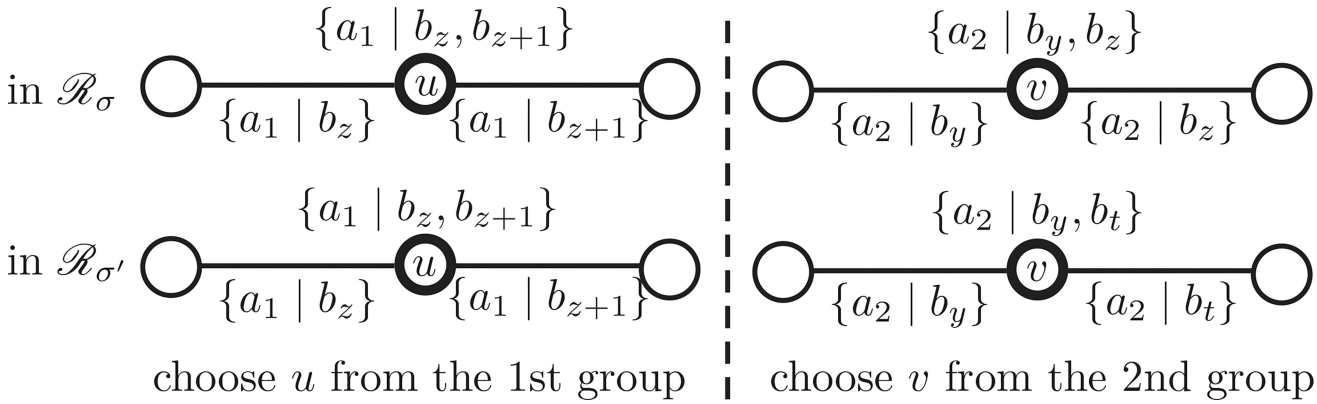


Fig. 14.
(Case 2) The feature sets of u and v with respect to $\mathscr{R}_\sigma$ and $\mathscr{R}_{\sigma'}$.

A Randomized Algorithm

Input: A graph $\mathcal{G}$, integer parameters α, β, mixing parameter c, number of rounds N;

Initialization:

- Assign to all $v \in \mathcal{V}$ nonempty sets $A_v \subseteq \mathcal{A} = \{a_1, \ldots, a_\alpha\}$ and $B_v \subseteq \mathcal{B} = \{b_1, \ldots, b_\beta\}$, chosen uniformly at random;
- Initially, let both $\mathcal{L}$ and $\mathcal{L}_{\max}$ denote the chosen random feature assignments;

repeat

  Choose a vertex $u \in \mathcal{V}$ uniformly at random;

  Select $\varnothing \neq A'_u \subseteq \mathcal{A}$ and $\varnothing \neq B'_u \subseteq \mathcal{B}$ uniformly at random;

  Let $\mathcal{L}' \leftarrow \mathcal{L}$ by assigning A'_u and B'_u to u;

  Set $\mathcal{L} = \mathcal{L}'$ with probability $\min\{1, e^{c\left(s(\mathcal{L}') - s(\mathcal{L})\right)}\}$;

  if $\mathcal{L}$ is replaced by $\mathcal{L}'$ and $s(\mathcal{L}') > s(\mathcal{L}_{\max})$ **then**

    Set $\mathcal{L}_{\max} = \mathcal{L}'$;

  end if

until the loop has run for N rounds;

Output: $\mathcal{L}_{\max}$;

Fig. 15.

A simulated annealing algorithm for determining approximate cointersection representations of graphs. The score $s(\mathcal{L})$ counts the number of edges/non-edges of $\mathcal{G}$ that match the feature assignment $\mathcal{L}$ according to the Cointersection Condition.

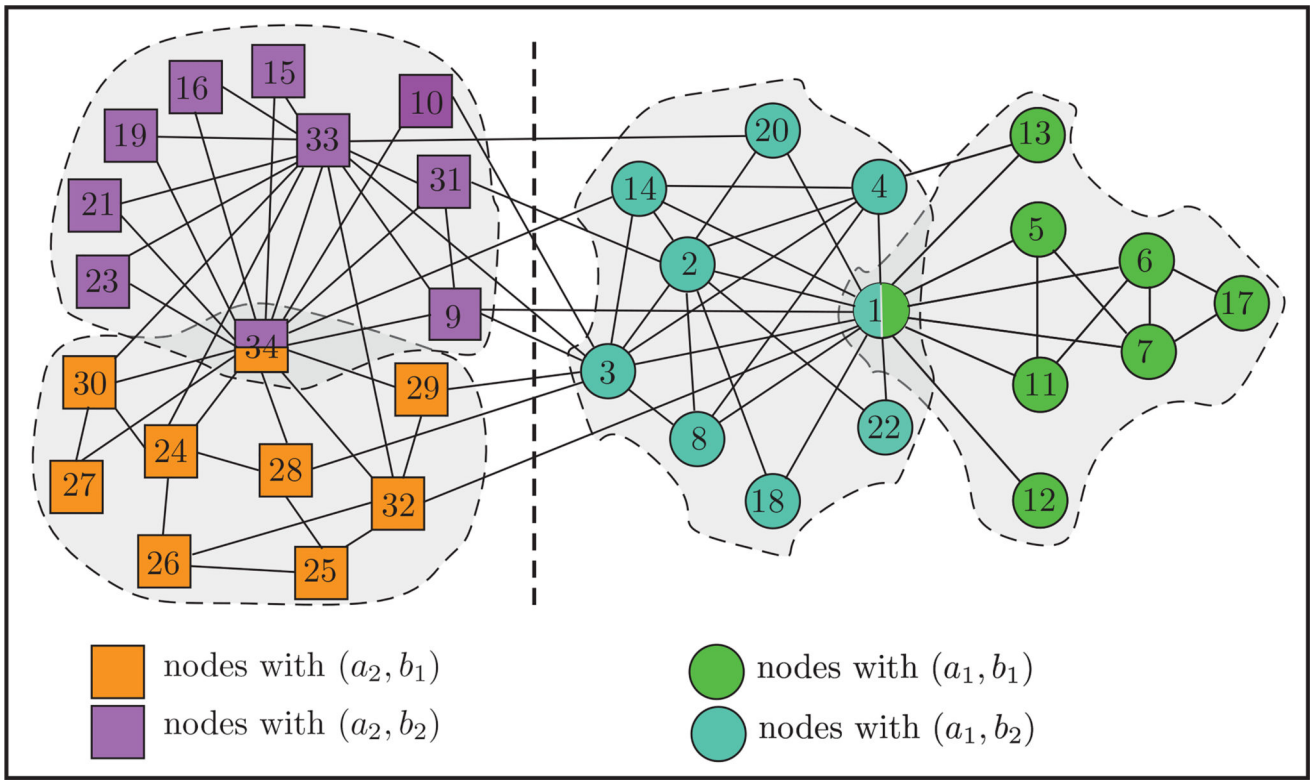


Fig. 16.

The social network of friendships in a Karate club. The members of the club were naturally divided into two groups, the one on the left supporting the president (Node 34), and the one on the right supporting the instructor (Node 1). Given $\alpha = \beta = 2$ as input parameters, the randomized algorithm recovered a community structure, with two disjoint communities which correspond exactly to the two groups of supporters as discussed. But the algorithm provided more information, as within each community two further overlapping sub-communities, marked by different colors, where identified. The only overlapping was in terms of Node 34 and Node 1, marked with a mix of two colors, correspond to the club president and the instructor. This suggests that there were two sub communities within each community held together by the president and the instructor.

TABLE I

Performance of the simulated annealing algorithm on the Newman-Watts random graphs. Each entry from the second column onward contains four measurements, which correspond to four different choices of q from $\{0.1, 0.4, 0.7, 1.0\}$.

(*n*, *k*)	F-score	Precision	Recall	CPU (sec.)
(100, 14)	.9|.8|.77|.79	.94|.92|.84|.77	.86|.71|.71|.8	≤ 7
(200, 16)	.87|.75|.67|.65	.92|.89|.88|.9	.83|.65|.54|.5	≤ 18
(500, 18)	.78|.69|.61|.56	.89|.84|.86|.84	.71|.58|.47|.42	≤ 92
(1000, 20)	.71|.62|.56|.46	.74|.74|.75|.76	.68|.54|.45|.33	≤ 912
(2000, 22)	.59|.5|.45|.41	.71|.7|.7|.7	.5|.4|.33|.29	≤ 6880