



# Asymptotically Smaller Encodings for Graph Problems and Scheduling

Bernardo Subercaseaux   

Carnegie Mellon University, Pittsburgh PA

---

## Abstract

We show how several graph problems (e.g., vertex-cover, independent-set,  $k$ -coloring) can be encoded into CNF using only  $\mathcal{O}(|V|^2/\lg |V|)$  many clauses, as opposed to the  $\Omega(|V|^2)$  constraints used by standard encodings. This somewhat surprising result is a simple consequence of a result of Chung, Erdős, and Spencer (1983) about biclique coverings of graphs, and opens theoretical avenues to understand the success of *Bounded Variable Addition* (Manthey, Heule, and Biere, 2012) as a preprocessing tool. Finally, we show a novel encoding for independent sets in some dense interval graphs using only  $\mathcal{O}(|V|\lg |V|)$  clauses (the direct encoding uses  $\Omega(|V|^2)$ ), which we have successfully applied to a string-compression encoding posed by Bannai et al. (2022). As a direct byproduct, we obtain a reduction in the encoding size of a scheduling problem posed by Mayank and Mondal (2020) from  $\mathcal{O}(NMT^2)$  to  $\mathcal{O}(NMT + MT^2 \lg T)$ , where  $N$  is the number of tasks,  $T$  the total timespan, and  $M$  the number of machines.

**2012 ACM Subject Classification** Mathematics of computing → Combinatorial optimization; Mathematics of computing → Graph algorithms; Theory of computation → Automated reasoning; Theory of computation → Constraint and logic programming

**Keywords and phrases** Independent set, CNF encodings, Biclique covering, Disjoint intervals

**Digital Object Identifier** 10.4230/LIPIcs.ModRef.2025.23

**Related Version** <https://arxiv.org/abs/2506.14042>

**Funding** This research is supported by the NSF under grant DMS-2434625.

## 1 Introduction

Using a more compact CNF encoding can make the entire difference between a combinatorial problem being solvable (even in many CPU years) and it being intractable [16, 17, 31, 33, 34]. However, besides a few very general principles [6, 29], it seems that the “*art of encodings*” is still mostly explored through problem-specific ideas, and it is not clear how to systematically obtain smaller encodings for combinatorial problems. Furthermore, lower bounds on the size of encodings have been elusive, with very few exceptions on relatively simple constraints such as *Parity* [10] or *At-Most-One* [22], making it difficult to predict whether the direct encoding for a given problem is already optimal or not.

In this article, I will show that several standard graph problems can be encoded more efficiently than through their direct formulation, and more importantly, that the tools used can shed light into theoretical questions about encodings.

As a representative example, consider first the *independent set* problem. The input is a graph  $G = (V, E)$ , together with an integer  $k$ , and the goal is to find a subset of the vertices  $S \subseteq V$  such that  $\binom{S}{2} \cap E = \emptyset$  (i.e., no two vertices in  $S$  are neighbors) and  $|S| = k$ . The “*direct encoding*” is thus to create, for each vertex  $v \in V$ , a variable  $x_v$  representing whether  $v \in S$ . Then, the direct encoding consists of enforcing the independent-set property

$$\bigwedge_{\{u,v\} \in E} (\overline{x_u} \vee \overline{x_v}), \quad (1)$$



© Bernardo Subercaseaux;

licensed under Creative Commons License CC-BY 4.0

The 24th workshop on Constraint Modelling and Reformulation.

Editors: María Andreína Francisco Rodríguez and Ian Gent; Article No. 23; pp. 23:1–23:21



Leibniz International Proceedings in Informatics

LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

and then the cardinality constraint  $\sum_{v \in V} x_v = k$ . While cardinality constraints are known to admit compact encodings with  $\mathcal{O}(n)$  clauses [32], and even arc-consistency in  $\mathcal{O}(n \lg^2 n)$  clauses [2], Equation (1) amounts to  $\Theta(|E|)$  clauses, which is  $\Omega(|V|^2)$  for dense graphs.

Our first contribution (in Section 2) is to show that this encoding can be improved to  $\mathcal{O}(|V|^2 / \lg |V|)$  clauses, and consequently, that several structurally similar graph problems can be encoded compactly as well:

► **Theorem 1 (Informal).** *The independent set, vertex cover,  $k$ -coloring, and clique problems can be encoded into CNF using  $\mathcal{O}(|V|^2 / \lg |V|)$  many clauses.*

This result improves upon an idea of Rintanen [30], and then Ignatiev, Morgado, and Marques-Silva [19], who used *clique coverings* to encode the independent-set property by observing that, for any clique  $K_t$  of  $G$ , at most one vertex of the clique can be part of  $S$ , which can be encoded using  $\mathcal{O}(t)$  clauses as opposed to the  $\Omega(t^2)$  clauses used by Equation (1). However, as the authors themselves note, this idea is not enough to obtain an encoding with  $o(|V|^2)$  clauses in all graphs, since for example a complete bipartite graph has  $\Omega(|V|^2)$  edges and yet not cliques of size larger than 2. We overcome this limitation by using *biclique coverings* of graphs, leveraging the fact that any graph with  $\Omega(|V|^2)$  edges must contain a biclique  $K_{t,t}$  with  $t = \Omega(\lg |V|)$  [8].

Then, in Section 3, we compare more in detail the *biclique covering* framework with the *clique covering* framework of Ignatiev, Morgado, and Marques-Silva [19] as well as with *Bounded Variable Addition* (BVA) [23], a successful preprocessing technique for reducing encoding sizes. As a cornerstone of this comparison, we study in Section 4 how to encode that a selection of intervals  $[i, j]$  for  $1 \leq i < j \leq n$  is *pairwise disjoint*, i.e., that no two intervals overlap. This corresponds to encoding the independent-set property of a corresponding interval graph  $\mathcal{I}_n$ . We show that, despite the fact that  $|E(\mathcal{I}_n)| = \Omega(n^4)$ , it is possible to obtain a much more compact encoding.

► **Theorem 2 (Informal).** *The independent-set property of the interval graph  $\mathcal{I}_n$  can be encoded into CNF using  $\mathcal{O}(n^2 \lg n)$  clauses.*

We show that this is more efficient than what is obtained via either the clique covering framework or the biclique covering framework. Moreover, we show that while BVA can obtain a more compact encoding in terms of the number of clauses (experimentally, since we do not have a theoretical guarantee), the structured encoding we present works better in practice. This is reminiscent of the idea of Haberlandt, Green, and Heule for *Structured Bounded Variable Addition* (SBVA), which ended up winning the SAT competition 2023 [15].

## 1.1 Notation and preliminaries

We will use  $\top$  and  $\perp$  to denote *true* and *false*, respectively. The negation of a variable  $x$  is denoted  $\bar{x}$ , and a *literal* is either a variable or the negation of a variable. Literals  $x$  and  $\bar{x}$  are said to be complementary, for any variable  $x$ . A *clause* is a set of non-complementary literals, and a *formula* is a set of clauses (we will thus identify  $\wedge$  with  $\cup$  when operating over clauses, and  $\ell_1 \vee \ell_2$  with  $\{\ell_1\} \cup \{\ell_2\}$  when operating over literals). The size of a formula is simply its number of clauses. We denote the set of variables appearing in a formula  $F$  as  $\text{Var}(F)$ . Given a set  $\mathcal{V}$  of variables, an assignment is a function  $\tau : \mathcal{V} \rightarrow \{\perp, \top\}$ . For a variable  $x \in \mathcal{V}$ , we say  $\tau \models x$  if  $\tau(x) = \top$ , and similarly,  $\tau \models \bar{x}$  if  $\tau(x) = \perp$ . For a clause  $C$ , we say  $\tau \models C$  if  $\bigvee \ell \in C \tau \models \ell$ , and for a formula  $F$ ,  $\tau \models F$  if  $\bigwedge_{C \in F} \tau \models C$ . When writing this, we assume implicitly that  $\text{Var}(F) \subseteq \mathcal{V}$ . For an assignment  $\tau : \mathcal{V} \rightarrow \{\perp, \top\}$  and a formula  $F$ , now with  $\mathcal{V} \subsetneq \text{Var}(F)$ , we denote by  $F|_\tau$  the formula obtained by eliminating

from  $F$  each clause satisfied by  $\tau$ , and then from each remaining clause eliminating every literal  $\ell$  such that  $\tau \models \bar{\ell}$ . Note that  $\text{Var}(F|_\tau) = \text{Var}(F) \setminus \mathcal{V}$ . We will write  $\text{SAT}(F)$  to say that  $\tau \models F$  for some assignment  $\tau$ , and  $\text{UNSAT}(F)$  to mean that no such assignment exists.

## 2 Subquadratic Encodings through Biclique Coverings

### 2.1 Clique Covering Encodings

Arguably, the *At-Most-One* constraint (AMO) for variables  $x_1, \dots, x_n$  is the most elemental example of an encoding whose direct formulation can be asymptotically improved. The naïve formulation, often referred to as the *pairwise encoding*, is simply

$$\text{AMO}(x_1, \dots, x_n) := \bigwedge_{1 \leq i < j \leq n} (\bar{x}_i \vee \bar{x}_j),$$

using  $\binom{n}{2}$  clauses, analogously to the dense case of Equation (1). On the other hand, several formulations using  $\mathcal{O}(n)$  clauses are known [27, 36]. The most compact is Chen’s *product encoding*, which uses  $2n + 4\sqrt{n} + \mathcal{O}(\sqrt[4]{n})$  clauses [7], and is essentially tight [22]. We will use notation  $\text{AMO}_{\text{PE}}(x_1, \dots, x_n)$  to denote the formula resulting from Chen’s product encoding over variables  $x_1, \dots, x_n$ .<sup>1</sup> As noted by Ignatiev, Morgado, and Marques-Silva [19], we can interpret this result as saying that “the independent-set property can be encoded in  $\mathcal{O}(n)$  clauses for a complete graph  $K_n$ ”. Let us now formalize what this means.

► **Definition 3** (Encoding the ISP). *Given a graph  $G = (V, E)$ , a set of variables  $X = \{x_v \mid v \in V\}$ , and a potentially empty set of (auxiliary) variables  $Y = \{y_1, \dots, y_m\}$ , we say that a formula  $F$  with  $\text{var}(F) = X \cup Y$  encodes the “independent-set property” (ISP) if for every assignment  $\tau : X \rightarrow \{\perp, \top\}$ ,*

$$\text{SAT}(F|_\tau) \iff \{v \in V \mid \tau(x_v) = \top\} \text{ is an independent set of } G.$$

The trivial observation now is that for a complete graph  $K_n$ , a subset  $S \subseteq V(K_n)$  of its vertices is independent if and only if  $|S| \leq 1$ , and thus  $\text{AMO}_{\text{PE}}(V(K_n))$  encodes the ISP of  $K_n$  with  $\mathcal{O}(n)$  clauses. Note that we have written  $\text{AMO}_{\text{PE}}(V(K_n))$ , identifying each vertex  $v$  with a corresponding variable  $x_v$ , and will do this repeatedly to avoid cluttering the notation.

To extend this idea to more general graphs, Ignatiev, Morgado, and Marques-Silva used *clique coverings*: a “clique covering” of a graph  $G$  is a set  $\{C_1, \dots, C_m\}$  of subgraphs of  $G$ , each of which must be a clique, such that every edge  $e \in E(G)$  belongs to some subgraph  $C_i$ . That is,  $\bigcup_{i=1}^m E(C_i) = E(G)$ . We thus define a clique-covering-ISP (CC-ISP) encoding as follows:

► **Definition 4** (CC-ISP encoding). *Given a graph  $G$ , and a clique covering  $\mathcal{C}$  of  $G$ , the formula*

$$F_{\mathcal{C}} := \bigwedge_{C \in \mathcal{C}} \text{AMO}_{\text{PE}}(V(C))$$

*is said to be a “CC-ISP encoding” for  $G$ .*

Note that the number of clauses of  $F_{\mathcal{C}}$  is  $|F_{\mathcal{C}}| = \sum_{C \in \mathcal{C}} f(|V(C)|)$ , where  $f(n)$  is the number of clauses used by  $\text{AMO}_{\text{PE}}$  over  $n$  variables, and thus  $f(n) = \mathcal{O}(n)$ .

<sup>1</sup> We will rather arbitrarily use Chen’s product encoding throughout this section, but any encoding using a linear number of clauses would work as well.

► **Lemma 5** (Implicit in [19]). *Any CC-ISP encoding  $F_C$  for a graph  $G$  indeed encodes the independent-set property (see Definition 3) for  $G$ .*

**Proof.** Let  $\tau$  be an assignment of the variables  $\{x_v \mid v \in V(G)\}$ , and  $S_\tau = \{v \in V(G) \mid \tau(x_v) = \top\}$ . Now, assume  $S_\tau$  is an independent set of  $G$ . As the intersection of any independent set and a clique has at most one vertex, we have  $|S_\tau \cap V(C)| \leq 1$  for every  $C \in \mathcal{C}$ , and thus  $F_C|_\tau$  is clearly satisfiable. Conversely, if  $S_\tau$  is not an independent set of  $G$ , then some edge  $e = \{u, v\}$  of  $G$  has  $|S_\tau \cap e| = 2$ , but by definition of clique covering,  $e \in E(C)$  for some  $C \in \mathcal{C}$ , and thus  $e \subseteq V(C)$ . But this implies  $|S_\tau \cap V(C)| \geq 2$ , and thus the subformula  $\text{AMO}_{\text{PE}}(V(C))|_\tau$  is already unsatisfiable, implying  $F_C|_\tau$  is unsatisfiable. ◀

Unfortunately, this re-encoding technique is not useful for worst-case graphs. For example, in a complete bipartite graph (or *biclique*)  $K_{n,n}$ , the maximum clique has size 2, and thus any clique covering  $\mathcal{C}$  of  $K_{n,n}$  covers at most one edge per clique, leading to  $|\mathcal{C}| \geq |E(K_{n,n})| = n^2$ , and thus  $|F_C| \geq n^2$ . Recall that  $|E(K_{n,n})| = n^2$  is the number of clauses used by the pairwise encoding (Equation (1)), and thus no improvement is achieved. It is worth noting that clique coverings do yield an improvement for random graphs  $G(n, p)$ , using a result by Frieze and Reed [12] (errors corrected in its arXiv version [13]) or also a stronger result of Guo, Patton, and Warnke [14, Theorem 7].

► **Proposition 6.** *Let  $p \in [0, 1]$  be a constant, and  $G \sim G(n, p)$  be a random graph on  $n$  vertices obtained by adding each edge independently with probability  $p$ . Then, with high probability, there exists a CC-ISP encoding for  $G$  with  $\mathcal{O}\left(\frac{n^2}{\lg n}\right)$  clauses.*

**Proof Sketch.** The proof is a direct consequence of the existence of a clique cover with at most  $\mathcal{O}\left(\frac{n^2}{\lg^2 n}\right)$  cliques, proved by Frieze and Reed [13], and improved by Guo, Patton, and Warnke [14, Theorem 7]. Since in a  $G(n, p)$  the maximum clique size is  $\Theta(\lg n)$  with high probability, the result follows. ◀

Note that as  $G(n, \frac{1}{2})$  corresponds to the uniform distribution on  $n$ -vertex graphs, Proposition 6 implies that most graphs have CC-ISP encodings with  $\mathcal{O}\left(\frac{n^2}{\lg n}\right)$  clauses.

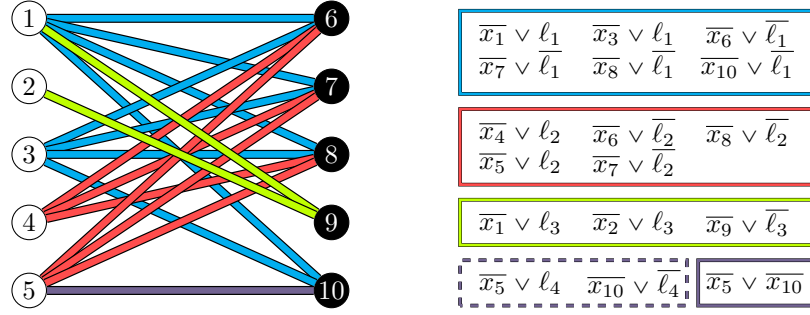
We will next see that by using biclique coverings instead of clique coverings we get a worst-case asymptotic improvement.

## 2.2 Biclique Covering Encodings

For a biclique  $K_{a,b}$  (complete bipartite graph with  $a$  vertices on one part and  $b$  on the other part), the independent-set property can also be encoded efficiently. Let  $A$  and  $B$  be the parts of  $K_{a,b}$ . Then, introduce an auxiliary variable  $x_A$  which intuitively represents that some vertex of  $A$  belongs to the desired independent set  $S$ . The desired formula is then

$$\text{BIS}(K_{a,b}) := \left( \bigwedge_{v \in A} (\overline{x_v} \vee x_A) \right) \wedge \left( \bigwedge_{v \in B} (\overline{x_A} \vee \overline{x_v}) \right). \quad (2)$$

Intuitively, the first part of the formula is saying that if some vertex  $v \in A$  is selected, then  $x_A$  will be true, and the second part enforces that  $x_A$  being true forbids any vertex in  $B$  from being selected. Note immediately that  $|\text{BIS}(K_{a,b})| = a + b = |V(K_{a,b})|$ , as opposed to the direct encoding which uses  $|E(K_{a,b})| = a \cdot b$ . As can be observed in Figure 1 (dashed purple box), this wastes a clause when  $a = b = 1$ , so we shall assume that in this case  $\text{BIS}(K_{1,1})$  will be the direct encoding. We again lift the biclique encoding to arbitrary graphs by coverings. A *biclique covering*  $\mathcal{B}$  for a graph  $G$  is simply a set of bicliques  $B_1, \dots, B_m$  that are subgraphs of  $G$  and such that  $\bigcup_{i=1}^m E(B_i) = E(G)$ .



■ **Figure 1** A biclique covering of a bipartite graph with 10 vertices and 17 edges, resulting in a formula with 15 clauses.

► **Proposition 7.** Let  $\mathcal{B} = \{B_1, \dots, B_m\}$  be a biclique covering of a graph  $G$ , and  $S \subseteq V(G)$  some set of vertices. Then,  $S$  is an independent set of  $G$  if and only if for every  $1 \leq i \leq m$ ,  $S \cap V(B_i)$  is an independent set of  $B_i$ .

**Proof.** Suppose first that  $S$  is an independent set of  $G$ . Then trivially  $S \cap V(B_i)$  is also an independent set of  $G$  for every  $1 \leq i \leq m$ , and as each  $B_i$  is a subgraph of  $G$ , the sets  $S \cap V(B_i)$  are also independent sets of  $B_i$ . For the opposite direction, assume that  $S$  is not an independent set of  $G$ , and thus  $|e \cap S| = 2$  for some  $e \in E(G)$ . Then, as  $e \in E(B_i)$  for some  $i \in \{1, \dots, m\}$  by definition of covering, we have  $e \cap V(B_i) = e$  and thus

$$|e \cap (S \cap V(B_i))| = |(e \cap V(B_i)) \cap S| = |e \cap S| = 2,$$

implying  $S \cap V(B_i)$  is not independent in  $B_i$ . ◀

Using Proposition 7 we directly obtain the following analog to Definition 3 and Lemma 5:

► **Proposition 8.** Given a graph  $G$ , and a biclique covering  $\mathcal{B}$  of  $G$ , the formula  $F_{\mathcal{B}} := \bigwedge_{B \in \mathcal{B}} \text{BIS}(B)$  is said to be a “BC-ISP” encoding for  $G$ . The formula  $F_{\mathcal{B}}$  encodes the independent-set property of  $G$ , and has size  $\sum_{B \in \mathcal{B}} |V(B)|$ .

The key difference now is made by a result stating that every graph admits a biclique covering that is asymptotically smaller than just taking its set of edges.

► **Theorem 9** (Erdős, Chung, and Spencer [8]). Every graph  $G$  on  $n$  vertices has a biclique covering  $\mathcal{B}$  with  $\sum_{B \in \mathcal{B}} |V(B)| = \mathcal{O}(n^2/\lg n)$ .

Combining Theorem 9 and Proposition 8 we get the main result of this section.

► **Theorem 10** (Formal version of Theorem 1). For every graph  $G$  on  $n$  vertices, there is a formula  $F$  that encodes the independent-set property of  $G$  such that  $|F| = \mathcal{O}(n^2/\lg n)$ .

Figure 1 illustrates a biclique covering encoding, showing a reduction in size from the direct encoding. By avoiding the re-encoding of  $K_{1,1}$  cliques (lime green and purple in Figure 1), the size would go down to 13.

We can trivially extend Theorem 10 to other graph problems as we show next.

- **(Vertex Cover)** Since a set of vertices  $S \subseteq V(G)$  is a vertex cover if and only if  $V(G) \setminus S$  is an independent set, it suffices to invert the polarity of each literal  $x_v$  (resp.  $\overline{x_v}$ ) to  $\overline{x_v}$  (resp.  $x_v$ ) in the formula  $F$  from Theorem 10, resulting on a formula of the same size.

- ( **$k$ -Coloring**) A  $k$ -coloring of a graph  $G$  is the same as a partition of  $V(G)$  into  $k$  independent sets. Naturally, if we have variables  $x_{v,c}$  for  $v \in V(G)$  and  $c \in \{1, \dots, k\}$  that indicate assigning color  $c$  to vertex  $v$ , then we simply use the conjunction of  $k$  formulas obtained from Theorem 10, where the  $c$ -th of them is obtained by replacing each variable  $x_v$  by  $x_{v,c}$ .
- (**Clique**) It suffices to use that a set of vertices  $S \subseteq V(G)$  is a clique of  $G$  if and only if  $S$  is an independent set of the complement graph  $\overline{G}$ .

We conclude this section by noting that Theorem 10, and its analog to the other graph problems mentioned above can be made constructive by using a result of Mubayi and Turán [26], which states that a biclique covering with the asymptotic guarantee of Theorem 9 can be computed deterministically in polynomial time.

### 3 Covering Frameworks

Even though the presented biclique-covering encodings are more efficient than clique-covering encodings over worst-case graphs, this is not necessarily the case on every graph. The main example being again  $K_n$ , for which we have the trivial clique covering  $\mathcal{C} = \{K_n\}$ , but for which every biclique covering uses at least  $\lceil \lg n \rceil$  many bicliques [11]. Furthermore, the trivial clique covering results in formula with  $\mathcal{O}(n)$  clauses, whereas any biclique covering results in  $\Omega(n \lg n)$  clauses.

► **Proposition 11.** *There is a BC-ISP encoding for  $K_n$  using  $\mathcal{O}(n \lg n)$  clauses, and any BC-ISP encoding for  $K_n$  uses  $\Omega(n \lg n)$  many clauses.*

**Proof.** Both the lower and upper bound were indirectly proven first by Katona and Szemerédi [21], but for completeness we present a recursive proof of the upper bound. For the upper bound, we construct a biclique covering  $\mathcal{B}$  recursively. First, we separate  $V(K_n)$  into two parts  $L$  and  $R$ , with  $|L| = \lfloor n/2 \rfloor$  and  $|R| = \lfloor n/2 \rfloor$ . Then, we add to  $\mathcal{B}$  the biclique  $K(L, R)$  between the vertices in  $L$  and the vertices in  $R$ , and proceed recursively on both  $L$  and  $R$  obtaining biclique coverings  $\mathcal{B}_L$  and  $\mathcal{B}_R$  respectively. The total biclique covering of  $K_n$  is then given by  $\mathcal{B} = K(L, R) \cup \mathcal{B}_L \cup \mathcal{B}_R$ . Thus, if we let  $g(n)$  denote the number of clauses used in an optimal BC-ISP encoding for  $K_n$  (i.e., one that minimizes  $\sum_{B \in \mathcal{B}} |V(B)|$ ), we have  $g(n) \leq n + g(\lfloor n/2 \rfloor) + g(\lfloor n/2 \rfloor) \leq n + 2g(\lfloor n/2 \rfloor)$ , from where it follows that  $g(n) = \mathcal{O}(n \lg n)$ . ◀

#### 3.1 Bounded Variable Addition

Interestingly, it is possible to do better for  $K_n$  by using a slight generalization of biclique coverings, as the *Bounded Variable Addition* (BVA) [23] preprocessing technique does. Let us explain first how BVA operates, which requires the following definition adapted from [15].

► **Definition 12** (Grid product). *Given a clause  $L$ , and set of clauses  $\Gamma$ , the “grid product” of  $L$  and  $\Gamma$  is the set of clauses  $L \bowtie \Gamma := \bigcup_{\ell \in L} \bigcup_{\gamma \in \Gamma} \gamma \cup \{\ell\}$ .*

BVA works by iteratively identifying subsets of a formula  $F$  that can be expressed as a grid product  $L \bowtie \Gamma$  (note that this does not require  $L \in F$  nor  $\Gamma \subseteq F$ ), and introducing a new auxiliary variable  $y$  which allows replacing  $L \bowtie \Gamma$  by the clauses  $\bigwedge_{\ell \in L} (\bar{y} \vee \ell) \wedge \bigwedge_{\gamma \in \Gamma} (y \vee \gamma)$ . Naturally, resolving on the new variable  $y$  yields the replaced set of clauses, and thus the new formula is equisatisfiable to the original one.



► **Example 13.** Let  $L = (x_1 \vee x_2)$  and  $\Gamma = \{(p \vee q), (q \vee r), (\bar{p} \vee \bar{r} \vee \bar{q})\}$ , then the corresponding grid product is

$$L \bowtie \Gamma = \{(x_1 \vee p \vee q), (x_1 \vee q \vee r), (x_1 \vee \bar{p} \vee \bar{r} \vee \bar{q}), (x_2 \vee p \vee q), (x_2 \vee q \vee r), (x_2 \vee \bar{p} \vee \bar{r} \vee \bar{q})\}.$$

The new variable  $y$  can be used to replace this set of clauses by the following ones:

$$(\bar{y} \vee x_1) \wedge (\bar{y} \vee x_2) \wedge (y \vee p \vee q) \wedge (y \vee q \vee r) \wedge (y \vee \bar{p} \vee \bar{r} \vee \bar{q}).$$

► **Definition 14.** Given a formula  $F$ , and a subset of  $F$  that can be expressed as a grid product  $L \bowtie \Gamma$ , we say that the formula  $F' := F \setminus (L \bowtie \Gamma) \wedge \bigwedge_{\ell \in L} (\bar{y} \vee \ell) \wedge \bigwedge_{\gamma \in \Gamma} (y \vee \gamma)$  is obtainable from  $F$  by BVA, which we denote by  $F \xrightarrow{\text{BVA}} F'$ . If there is a sequence of formulas  $F_1, \dots, F_k$  such that  $F \xrightarrow{\text{BVA}} F_1 \xrightarrow{\text{BVA}} \dots \xrightarrow{\text{BVA}} F_k$ , we say that  $F_k$  is a potential BVA re-encoding of  $F$  and write  $F \rightsquigarrow^{\text{BVA}} F_k$ .

Note that in the case of  $\Gamma$  being a set of unit clauses, this matches Equation (2), from where we immediately have the following result:

► **Proposition 15.** For every graph  $G$  on  $n$  vertices, there is a formula  $F$  such that  $\bigwedge_{\{u,v\} \in E(G)} (\bar{x}_u \vee \bar{x}_v) \rightsquigarrow^{\text{BVA}} F$  and  $|F| = \mathcal{O}(n^2 / \lg n)$ .

An important difference between BVA and covering re-encodings is that the new variables created by BVA can also be part of the identified grid products, whereas the (bi)clique covering encodings only use the original set of variables. We will show that this difference is enough to obtain a linear encoding for the independent-set property of  $K_n$ , and thus showing that BVA reencodes pairwise cardinality constraints into a linear number of clauses. While this was already observed without proof in the original BVA paper of Manthey, Heule, and Biere [23], we remark that the authors have declared that their results were only empirical [5]. We now provide a formal proof, noting that it applies to an idealized version of BVA, as opposed to its actual implementation, which is more subtle.

► **Proposition 16.** There is a formula  $F$  with  $|F| = \mathcal{O}(n)$  such that  $\text{AMO}(x_1, \dots, x_n) \rightsquigarrow^{\text{BVA}} F$ .

**Proof.** We prove that for  $n \geq 3$ , there is such an  $F$  with  $|F| = 3n - 6$ . The proof is by induction on  $n$ . For  $n = 3$  and  $n = 4$ , we simply take  $F := \text{AMO}(x_1, \dots, x_n)$ , which has size  $\binom{3}{2} = 3 = 3 \cdot 3 - 6$  and  $\binom{4}{2} = 6 = 3 \cdot 4 - 6$  respectively. For  $n \geq 5$ , consider first the grid product  $\{x_1, x_2, x_3\} \bowtie \{x_4, x_5, \dots, x_n\}$ , which is clearly in  $\text{AMO}(x_1, \dots, x_n)$ . The replacement of this grid product by BVA can be split into the following sets of clauses:

1. The clauses  $(\bar{x}_1 \vee \bar{x}_2)$ ,  $(\bar{x}_1 \vee \bar{x}_3)$ , and  $(\bar{x}_2 \vee \bar{x}_3)$ , unaltered by the replacement.
2. The clauses  $(\bar{y} \vee \ell)$  for  $\ell \in \{\bar{x}_1, \bar{x}_2, \bar{x}_3\}$ .
3. The clauses  $(y \vee \gamma)$  for  $\gamma \in \{\bar{x}_4, \bar{x}_5, \dots, \bar{x}_n\}$ .
4. The clauses  $(\bar{x}_i \vee \bar{x}_j)$  for  $4 \leq i < j \leq n$ , also unaltered by the replacement.

There are only 6 clauses of types (1) and (2), and for the clauses of types (3) and (4), the key observation is that they correspond exactly to the formula  $\text{AMO}(\bar{y}, x_4, \dots, x_n)$ , which by the induction hypothesis admits a BVA re-encoding  $F'$  of size  $3(n-2) - 6$ . Thus, denoting by  $S$  the set of clauses of types (1) and (2), we have

$$\text{AMO}(x_1, \dots, x_n) \xrightarrow{\text{BVA}} S \cup \text{AMO}(\bar{y}, x_4, \dots, x_n) \rightsquigarrow^{\text{BVA}} S \cup F',$$

and  $|S \cup F'| = 6 + 3(n-2) - 6 = 3n - 6$ , as desired. ◀

### 3.2 A Lower Bound for the Independent-Set Property

We now turn our attention to whether BVA, or some other re-encoding technique, could yield an asymptotic improvement over Theorem 10. While we do not manage to answer this question in full generality, we show that if we restrict ourselves to 2-CNF formulas (as the ones obtained by BVA, or the covering encodings), then we can prove a matching lower bound of  $\Omega(n^2/\lg n)$ , using a similar argument to Jukna [20, Theorem 1.7]. In fact our results holds for  $k$ -CNF (at most  $k$  literals per clause) for any fixed  $k$ .

► **Proposition 17.** *Fix an integer  $k \geq 2$ . Then, for any sufficiently large  $n$ , there is a graph  $G_n$  on  $n$  vertices such that any  $k$ -CNF formula encoding the independent-set property of  $G_n$  has size  $\Omega(n^2/\lg n)$ .*

**Proof Sketch.** There are  $2^{\binom{n}{2}}$  many distinct  $n$ -vertex graphs, and we will prove first that each of them requires a different formula to encode its independent-set property. Assume, expecting a contradiction, that some formula  $F$  encodes the independent-set property for two distinct  $n$ -vertex graphs  $G$  and  $G'$ . Then, take an edge  $\{u, v\}$  that is in  $G$  but not in  $G'$  (without loss of generality), and consider the assignment  $\tau$  defined by  $\tau(x_w) = \begin{cases} \top & \text{if } w \in \{u, v\}, \\ \perp & \text{otherwise.} \end{cases}$  Then,

according to Definition 3, since  $F$  encodes the independent-set property for  $G$  we have that  $F_\tau$  is satisfiable iff  $\{u, v\}$  is an independent set in  $G$ , which is not the case since  $\{u, v\} \in E(G)$ , thus making  $F_\tau$  unsatisfiable. However, since  $F$  also encodes the independent-set property for  $G'$ , we conclude that  $F_\tau$  is satisfiable, which is a contradiction. Now, we fix without loss of generality the set of variables for the  $k$ -CNF formulas encoding the independent-set property to be  $\{x_1, \dots, x_n, y_1, \dots, y_r\}$ , where  $r$  is the maximum number of auxiliary variables used in any of the formulas. Consider now that there are  $2^r \binom{m}{r}$  many possible  $r$ -ary clauses from a set of  $m$  variables (since each of the  $k$  literals can have two polarities), and thus, the number of  $k$ -CNF formulas with  $t$  clauses is at most

$$\left( \sum_{r=1}^k \frac{2^r \binom{m}{r}}{t} \right) \leq \left( k 2^k \binom{m}{k} \right)^t \leq \left( k 2^k \binom{t}{k} \right)^t,$$

where the last inequality assumes without loss of generality that the formulas do not contain pure literals (which can be simply removed), and thus the number of variables  $m$  is at most the number of clauses  $t$ . As we proved that each different graph needs a different formula, we have that

$$\begin{aligned} \left( k 2^k \binom{t}{k} \right)^t &\geq 2^{\binom{n}{2}} \iff t \cdot \lg \left( k 2^k \binom{t}{k} \right) \geq \binom{n}{2} \\ &\iff t \cdot \lg(t) \geq c \cdot n^2 \iff t \geq c \cdot n^2 / \lg(t). \end{aligned} \quad (\text{for some constant } c > 0)$$

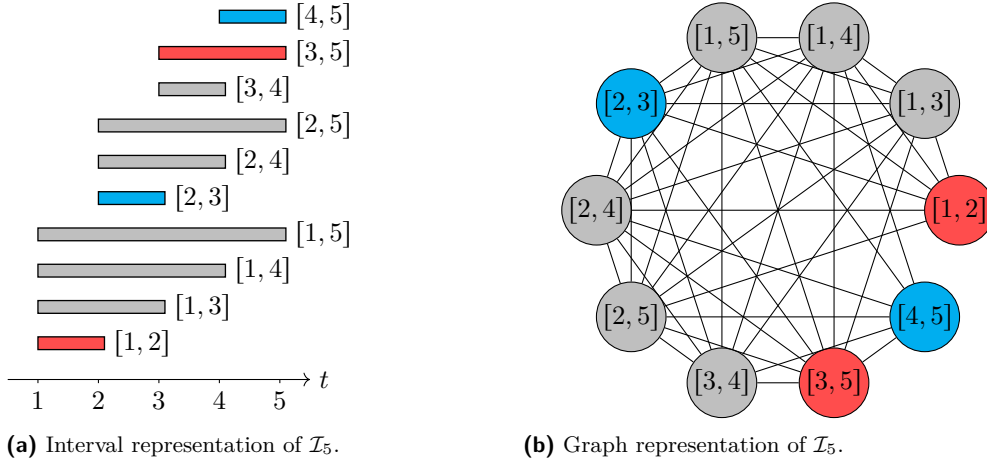
But unless  $t = \Omega(n^2)$ , in which case we can directly conclude the result, we have that  $\lg(t) \leq 2 \lg(n)$ , and thus we conclude that  $t = \Omega(n^2/\lg n)$ . ◀

## 4 Disjoint Intervals

We now study the independent-set property over a class of graphs that is relevant for a variety of scheduling-related problems: *interval graphs*.

► **Definition 18.** *For any  $n \geq 2$ , we define the full and discrete interval graph  $\mathcal{I}_n$  as the graph whose vertices are all the intervals  $[i, j]$  for integers  $1 \leq i < j \leq n$ , and there are edges between any two intervals that intersect.*





**Figure 2** Illustration of the interval graph  $\mathcal{I}_5$  in two different representations, where the only two independent sets of size larger than 1 are depicted, one in red and one in cyan.

An example is illustrated in Figure 2. Naturally, the independent sets of  $\mathcal{I}_n$  correspond to sets of intervals that do not intersect, and thus for which tasks with competing resources can be all scheduled. Note that  $\mathcal{I}_n$  has  $\Omega(n^4)$  edges, as for each subset  $\{i, j, k, \ell\} \subseteq \{1, \dots, n\}$  with  $i < j < k < \ell$  we have an edge between vertices  $[i, k]$  and  $[j, \ell]$ . Therefore, a direct encoding of the independent-set property is very large even for small values of  $n$ . We will prove that this can be drastically improved to  $\mathcal{O}(n^2 \lg n)$ , but first we analyze how well clique coverings do for this family of graphs.

► **Proposition 19.** *There is a CC-ISP encoding for  $\mathcal{I}_n$  using  $\mathcal{O}(n^3)$  clauses, and no CC-ISP encoding can be asymptotically more compact.*

**Proof.** First, we note that for every  $k \in \{2, \dots, n-1\}$ , all the intervals  $[i, j] \in V(\mathcal{I}_n)$  with  $i \leq k \leq j$  intersect, thus forming a clique that we denote  $K_{\cap k}$ . Then, observe that the collection of cliques  $K_{\cap k}$ , for  $2 \leq k \leq n-1$ , is a clique covering of  $\mathcal{I}_n$ . Indeed, any edge  $e = ([i, j], [a, b])$  must either have  $a \leq j \leq b$ , in which case  $e$  is covered by  $K_{\cap j}$ , or  $i \leq a \leq j$ , in which case  $e$  is covered by  $K_{\cap a}$ . Each clique  $K_{\cap k}$  has  $\mathcal{O}(n^2)$  vertices (there are  $\mathcal{O}(n^2)$  vertices in the entire  $\mathcal{I}_n$ ), and we thus this clique covering results in  $\sum_{k=2}^{n-1} |\text{AMO}_{\text{PE}}(V(K_{\cap k}))| = \mathcal{O}(n^3)$  many clauses. For the lower bound, consider an arbitrary clique covering  $\mathcal{C}$  of  $\mathcal{I}_n$ . Then, observe that each interval  $x := [i, j]$  is adjacent to all the intervals  $[i + 2t, i + 2t + 1]$  for  $t \in \{0, \dots, \lfloor (j-i-1)/2 \rfloor\}$ . Moreover, all these intervals  $[i + 2t, i + 2t + 1]$  are pairwise disjoint, which implies that for each  $t$ , the edge  $\{x, [i + 2t, i + 2t + 1]\}$  must be covered by a different clique  $C_t \in \mathcal{C}$ . Thus, we have that  $|\{C \in \mathcal{C} : x \in C\}| \geq \lfloor (j-i-1)/2 \rfloor + 1 \geq (j-i)/3$ . We can now conclude since

$$\sum_{C \in \mathcal{C}} |\text{AMO}_{\text{PE}}(C)| \geq \sum_{C \in \mathcal{C}} |C| = \sum_{x \in V(\mathcal{I}_n)} |\{C \in \mathcal{C} : x \in C\}| \geq \sum_{i=1}^n \sum_{j=i+1}^n (j-i)/3,$$

from where the change of variables  $d := j - i$  yields

$$\sum_{i=1}^n \sum_{j=i+1}^n (j-i)/3 = \sum_{i=1}^n \sum_{d=1}^{n-i} d/3 \geq \frac{1}{3} \sum_{i=\lfloor n/2 \rfloor}^n \sum_{d=1}^{\lfloor n/2 \rfloor} d = \frac{1}{3} \lfloor n/2 \rfloor \cdot \frac{\lfloor n/2 \rfloor (\lfloor n/2 \rfloor + 1)}{2} = \Omega(n^3). \quad \blacktriangleleft$$

## 4.1 The Interval Propagation Trick

The proof of Theorem 2, the main result of this section, is quite technical, and it is worth isolating one of its ingredients which might be of independent interest. Consider the following encoding problem:

*We have variables  $x_{i,j}$ , representing that an interval  $[i, j]$  is “selected”, for  $1 \leq i < j \leq n$ , and also variables  $t_\ell$ , for  $1 \leq \ell \leq n$ , whose intended semantics are that  $t_\ell$  is true if and only if the index  $\ell$  is contained in some selected interval. The problem is how to efficiently encode this relationship between the  $x_{i,j}$  and  $t_\ell$  variables, without enforcing any other conditions on either the  $x$ - or  $t$ -variables.*

For example, if  $x_{2,4}$  and  $x_{7,9}$  are the only  $x$ -variables assigned to  $\top$ , then  $\{t_2, t_3, t_4, t_7, t_8, t_9\}$  should be assigned to  $\top$ , and every other  $t_\ell$  variable to  $\perp$ . The fact that  $t_\ell$  implies that some interval containing  $\ell$  is selected is trivial to encode, by just adding the  $\mathcal{O}(n)$  following clauses:

$$\overline{t_\ell} \vee \bigvee_{[i,j] \ni \ell} x_{i,j}, \quad \forall 1 \leq \ell \leq n.$$

The other direction admits a nice trick. The naïve way of encoding the implication from the  $x$ -variables toward the  $t$ -variables is to simply add clauses of the form  $(\overline{x_{i,j}} \vee t_\ell)$ , for every  $1 \leq i < j \leq n$  and every  $i \leq \ell \leq j$ , which amounts to  $\sum_{i=1}^n \sum_{j=i+1}^n (j-i) = \Omega(n^3)$  many clauses, by the same analysis of the sum used in the proof of Proposition 19. It turns out, however, that we can achieve this with  $\mathcal{O}(n^2)$  many clauses, using what we denote the “interval propagation trick”. First, we create variables  $z_{i,j}$  for each  $1 \leq i < j \leq n$ , and then add the following clauses:

1.  $\overline{x_{i,j}} \vee z_{i,j}$ , for every  $1 \leq i < j \leq n$ .
2.  $(\overline{z_{i,i+1}} \vee t_i)$  and  $(\overline{z_{i,i+1}} \vee t_{i+1})$ , for every  $1 \leq i < n$ .
3.  $(\overline{z_{i,j}} \vee z_{i+1,j})$  and  $(\overline{z_{i,j}} \vee z_{i,j-1})$ , for every  $1 \leq i < j-1 < n$ .
4.  $(\overline{z_{i,j}} \vee x_{i,j} \vee z_{i-1,j} \vee z_{i,j+1})$ , for every  $1 \leq i < j \leq n$ , and removing the non-sensical literal  $z_{i-1,j}$  when  $i = 1$ , and  $z_{i,j+1}$  when  $j = n$ .

To formalize correctness, let us denote by  $\text{NIP}_n$  the formula resulting from the aforementioned clauses in the naïve encoding (i.e., of the forms  $\overline{t_\ell} \vee \bigvee_{[i,j] \ni \ell} x_{i,j}$  and  $(\overline{x_{i,j}} \vee t_\ell)$ ), and  $\text{IPT}_n$  the formula resulting from the clauses of the form  $\overline{t_\ell} \vee \bigvee_{[i,j] \ni \ell} x_{i,j}$  together with the clauses of types (1-4) above. Note that  $|\text{IPT}_n| \leq 6n^2 = \mathcal{O}(n^2)$ , and let us now state the desired form of “equivalence” between these formulations.

► **Proposition 20.** *Let  $\tau : \text{var}(\text{NIP}_n) \rightarrow \{\perp, \top\}$  and assignment. Then, we have that*

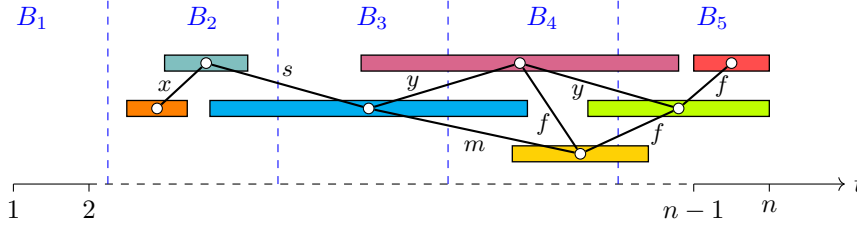
$$\tau \models \text{NIP}_n \iff \text{SAT}(\text{IPT}_n|_\tau),$$

*and moreover, any satisfying assignment  $\theta$  for  $\text{IPT}_n|_\tau$  must assign  $\theta(z_{a,b}) = \top$  if and only if there is some  $[i, j]$  such that  $\tau(x_{i,j}) = \top$  and  $[a, b] \subseteq [i, j]$*

The proof of Proposition 20 is a rather tedious induction, and thus we defer it to Appendix A.

## 4.2 Better Encodings for $\mathcal{I}_n$

Our next result, which uses only  $\mathcal{O}(n^2 \lg n)$  many clauses, requires a more careful encoding, which starts by decomposing  $[1, n]$  into blocks of size at most  $b$ , which we do by assigning to each position  $1 \leq i \leq n$  a block number  $B(i) := \lceil i/b \rceil$ . For now we will keep  $b$  as a parameter, and denote by  $k := \lceil n/b \rceil$  the number of blocks. We can characterize the different edges of  $\mathcal{I}_n$  in terms of the blocks of their vertices, as the following lemma states.



■ **Figure 3** Illustration of Lemma 21. The type of each edge is indicated by its label, and blocks are separated by dashed blue lines.

► **Lemma 21.** *Each edge  $e := \{[i_1, j_1], [i_2, j_2]\} \in E(\mathcal{I}_n)$  with  $i_1 \leq i_2$  must be part of exactly one of the following cases:*

1.  $B(i_1) = B(i_2)$  and  $B(j_1) = B(j_2)$ , and  $i_2 \leq j_1$ , in which case we say  $e$  is an  $x$ -edge.
  2.  $B(i_1) < B(i_2) < B(j_1)$ , in which case we say  $e$  is a  $y$ -edge.
  3.  $B(i_1) = B(i_2)$  and  $i_2 \leq j_1$ , but  $B(j_1) \neq B(j_2)$ , in which case we say  $e$  is an  $s$ -edge.
  4.  $B(i_1) < B(i_2) = B(j_1) = B(j_2)$  and  $i_2 \leq j_1$ , in which case we say  $e$  is an  $f$ -edge.
  5.  $B(i_1) < B(i_2) = B(j_1) \neq B(j_2)$ , and  $i_2 \leq j_1$ , in which case we say  $e$  is an  $m$ -edge.
- Moreover, any tuple  $(i_1, j_1, i_2, j_2)$  with  $i_1 \leq i_2$  that satisfies one of these cases implies  $\{[i_1, j_1], [i_2, j_2]\} \in E(\mathcal{I}_n)$ .

An illustration of Lemma 21 is provided in Figure 3, and the proof is just case analysis and thus deferred to Appendix B.

► **Theorem 22.** *The independent-set property of  $\mathcal{I}_n$  can be encoded using at most  $26n^2 \lg n$  clauses.*

**Proof.** We will prove a slightly stronger statement in order to have a stronger inductive hypothesis. Let  $\mathcal{I}_n^0$  be the graph whose vertices are all the intervals  $[i, j]$  for integers  $1 \leq i < j \leq n$ , but now with edges only between intervals whose intersection has cardinality at least 2. That is,  $\{[1, 3], [3, 5]\}$  is not an edge in  $\mathcal{I}_5^0$ , but it is in  $\mathcal{I}_5$ . The proof is by (strong) induction on  $n \geq 2$ . The base case  $n = 2$  is trivial since we can use the direct encoding then. In fact, it is easy to see that the direct encoding uses at most  $3\binom{n}{4} = \frac{n^4}{8}$  clauses, and one can computationally check that for  $n \leq 32$ ,  $\frac{n^4}{8} \leq 26n^2 \lg n$ , and thus the direct encoding is enough for the result. We thus assume  $n > 32$  from now on. We now focus on the inductive case, where we will generally assume that we are encoding the independent-set property for  $\mathcal{I}_n$ , but indicate whenever a slight change is needed for  $\mathcal{I}_n^0$ , since the two cases are almost identical. The encoding will consider each of the four types of edges from Lemma 21 separately. The base variables are  $x_{i,j}$  (for  $1 \leq i < j \leq n$ ), representing that the interval  $[i, j]$  is part of the independent set.

- ( **$x$ -edges**) Consider a fixed choice of  $B(i_1) = B(i_2) = \ell$  and  $B(j_1) = B(j_2) = r$ , and note that there will be  $k^2$  such choices, since there are  $k$  blocks in total. If  $\ell = r$ , then it is easy to see that the  $x$ -edges whose endpoints are in block  $\ell$  form a graph isomorphic to  $\mathcal{I}_b$  (resp.  $\mathcal{I}_b^0$ ), and thus by inductive hypothesis they can be encoded using at most  $26b^2 \lg b$  clauses. If  $\ell < r$ , then we consider the graph  $G_{\ell,r}$  formed by the  $x$ -edges whose endpoints are in blocks  $\ell$  and  $r$ , even if they are both in  $\ell$  or both in  $r$ . This time  $G_{\ell,r}$  is isomorphic to  $\mathcal{I}_{2b}$  (resp.  $\mathcal{I}_{2b}^0$ ), and thus by the inductive hypothesis all these  $x$ -edges can be encoded using at most  $26(2b)^2 \lg(2b) = 104b^2 \lg b + 104$  clauses. As there are  $k^2$  choices for  $\ell, r$ , we can encode all the  $x$ -edges of  $\mathcal{I}_n$  using at most  $k^2 \cdot (104b^2 \lg b + 104)$  clauses.

- **(y-edges)** We create, for each pair of block-indices  $1 \leq \ell < r \leq k$ , an auxiliary variable  $y_{\ell,r}$  that represents that there is some interval  $[i, j]$  in the independent set such that  $B(i) = \ell$  and  $B(j) = r$ . To enforce these semantics, we add clauses

$$\overline{x_{i,j}} \vee y_{B(i),B(j)}, \quad \forall 1 \leq i < j \leq n, \quad (3)$$

$$\left( \overline{y_{\ell,r}} \vee \bigvee_{i,j: B(i)=\ell, B(j)=r} x_{i,j} \right), \quad \forall 1 \leq \ell < r \leq k. \quad (4)$$

The key observation now is that the graph  $G_y$  whose vertices are the  $y_{\ell,r}$  variables and has edges between variables  $y_{\ell_1,r_1}, y_{\ell_2,r_2}$  whenever  $\ell_1 < \ell_2 < r_1$ , is isomorphic to  $\mathcal{I}_k^0$ . Therefore, by the inductive hypothesis, we can encode all the  $y$ -edges using  $\binom{n}{2} + \binom{k}{2} + 26k^2 \lg k$  many clauses.

- **(s-edges)** We create, for each block-index  $2 \leq r \leq k$ , and position  $i$  such that  $B(i) < r$ , an auxiliary variable  $s_{i,r}$  that represents that there is some interval  $[i, j]$  in the independent set such that  $B(j) = r$ . We encode these semantics using clauses  $(\overline{x_{i,j}} \vee s_{i,B(j)})$  and  $(\overline{s_{i,r}} \vee \bigvee_{j \text{ s.t. } B(j)=r} x_{i,j})$ , which amount to at most  $2n^2$  clauses. Then, we add clauses

$$\overline{s_{i_1,r_1}} \vee \overline{s_{i_2,r_2}}, \quad \forall i_1 \leq i_2, \text{ s.t. } B(i_1) = B(i_2), \forall r_1 > B(i_1), r_2 > B(i_1) \text{ with } r_1 \neq r_2. \quad (5)$$

There are at most  $b^2 \cdot k^3$  clauses from Equation (5), since we need to choose  $B(i_1), r_1, r_2$ , for which there are  $k^3$  possibilities, and then  $b^2$  possibilities for  $i_1, i_2$  such that  $B(i_1) = B(i_2)$ . Unfortunately, this is not enough to encode all  $s$ -edges, since Equation (5) misses the cases where one of the two intervals is entirely contained in one block, so either  $B(i_1) = B(j_1)$  or  $B(i_2) = B(j_2)$ . The naïve solution would be to add the following clauses:

$$\overline{x_{i_1,j_1}} \vee \overline{s_{i_2,r}}, \quad \forall 1 \leq i_1 \leq i_2 \leq n, \forall j_1 > i_1 \text{ s.t. } B(i_1) = B(j_1) = B(i_2), \forall r > B(i_2), \quad (6)$$

$$\overline{x_{i_2,j_2}} \vee \overline{s_{i_1,r}}, \quad \forall 1 \leq i_1 \leq i_2 \leq n, \forall j_2 > i_2 \text{ s.t. } B(i_1) = B(i_2) = B(j_2), \forall r > B(i_1). \quad (7)$$

Equation (6) uses at most  $b^3 \cdot k^2$  clauses, since we need to choose  $B(i_1) = B(j_1) = B(i_2)$  and  $r$ , and for each such choice there are at most  $b^3$  possibilities for  $i_1, i_2, j_1$  within the same block. Analogously, Equation (7) also incurs in at most  $b^3 \cdot k^2$  clauses. However, it will turn out that  $b^3 \cdot k^2$  clauses would be too large, since we will end up setting  $k = \Theta(\lg n)$ , and thus we need a slightly better way to encode Equations (6) and (7). The solution is to use the “interval propagation trick” from Section 4.1 independently in each block of index  $1 \leq d \leq k$ , thanks to which we can assume variables  $t_\ell^d$  that represent whether some  $x_{i,j}$  with  $\ell \in [i, j]$  is true with  $B(i) = B(j) = d$  using at most  $6b^2$  clauses. In total over the  $k$  blocks this incurs in at most  $6b^2 \cdot k$  clauses. Now, we can replace Equations (6) and (7) by

$$\overline{t_\ell^{B(i)}} \vee \overline{s_{i,r}}, \quad \forall 1 \leq i \leq \ell \leq n, \text{ s.t. } B(i) = B(j), \forall r > B(i). \quad (8)$$

Equation (8) only requires  $k \cdot b^2$  many clauses, since there are at most  $k$  choices for  $r$ , and  $b^2$  for  $i, \ell$ . We thus cover all  $s$ -edges using a total of  $2n^2 + b^2 k^3 + 7b^2 k$  clauses.

- **(f-edges)** This case is fully symmetrical to the  $s$ -edges, this time using variables  $f_{\ell,j}$  that represent the presence of some interval  $[i, j]$  in the independent set such that  $B(i) = \ell$ . We add the symmetrical clauses, e.g., the symmetrical of Equation (5) is

$$\overline{f_{\ell_1,j_1}} \vee \overline{f_{\ell_2,j_2}}, \quad \forall 1 \leq j_1, j_2 \leq n, \text{ s.t. } B(j_1) = B(j_2), \forall \ell_1 < B(j_1), \ell_2 < B(j_1) \text{ with } \ell_1 \neq \ell_2.$$

and similarly with the symmetrical of Equation (8). A minor saving is that we do not need to pay any extra clauses for having the variables  $t_i^d$ , and therefore the combination of this case with the  $s$ -edges incurs in a total of  $4n^2 + 2b^2k^3 + 8b^2k$  clauses.

- **( $m$ -edges)** In this case we are dealing with intervals  $[i_1, j_1]$  and  $[i_2, j_2]$  such that  $B(j_1) = B(i_2)$ , which we call  $d := B(j_1)$ , and the other two endpoints not being in the block  $d$ . We can cover these by using both our  $s$  and  $f$  variables. Indeed, it suffices to add clauses

$$\overline{f_{\ell, j_1}} \vee \overline{s_{i_2, r}}, \quad \forall 1 \leq i_2 \leq j_1 \leq n \text{ s.t. } B(j_1) = B(i_2), \forall \ell < B(j_1), \forall r > B(i_2), \quad (9)$$

which amounts to at most  $b^2k^3$  clauses, since we have to choose  $\ell, r, B(j_1)$ , for which there are  $k^3$  options, and conditioned on  $B(j_1)$  there are at most  $b^2$  choices for  $j_1, i_2$ .

Adding the total number of clauses over all types, we get a total of

$$k(104b^2 \lg b + 104) + \binom{n}{2} + \binom{k}{2} + 26k^2 \lg k + 4n^2 + 2b^2k^3 + 8b^2k + b^2k^3$$

clauses. Using  $n = kb$ , this is at most  $104nb \lg b + 104k + 4.5n^2 + 0.5k^2 + 26k^2 \lg k + 3n^2k + 8nb$ , and then taking  $k = \lfloor \lg n \rfloor$  and  $b = n/\lfloor \lg n \rfloor$ , this is at most

$$3n^2 \lg n + n^2 \left( 108.5 + \frac{\lg n + 8.5 \lg^2(n) \lg \lg n}{n^2} + \frac{8}{\lg n} \right) \leq 3n^2 \lg n + 109.5n^2, \quad (\text{since } n \geq 32)$$

but as  $n > 32$ , we have  $\lg n > 5$ , and thus  $109.5n^2 < 22n^2 \lg n$ , from where  $3n^2 \lg n + 109.5n^2 < 25n^2 \lg n$ , and thus we conclude our result. ◀

### 4.3 Applications to Scheduling Problems

We consider the *non-preemptive schedulability* question treated by Mayak and Mondal [25], where there are  $N$  tasks, the  $i$ -th of which has an integer duration  $d_i$  and must be both started and finished in the interval  $[r_i, e_i]$ , with  $1 \leq r_i \leq e_i \leq T$ , and moreover, there are  $M$  machines which can do tasks in parallel. They present several SAT encodings, all of which use  $\Omega(NMT^2)$  clauses [25, Table 2]. Theorem 22 allows us to do better:

► **Theorem 23 (Informal).** *The non-preemptive schedulability problem can be encoded using  $\mathcal{O}(NMT + MT^2 \lg T)$  clauses.*

**Proof.** Create variables  $x_{i,t,m}$  that represents that task  $i$  is assigned to start in machine  $m$  at time  $t$ , and auxiliary variables  $y_{m,t_1,t_2}$  that represent that there is some task assigned to machine  $m$  for exactly the interval  $[t_1, t_2]$ . The semantics of the  $y$ -variables are encoded by clauses

$$\overline{x_{i,t,m}} \vee y_{m,t,t+d_i}, \quad \forall i \in [1, N], t \in [r_i, e_i], m \in [1, M],$$

using  $\mathcal{O}(NMT)$  clauses. We partition the tasks according to their duration, with  $D_t = \{i : d_i = t\}$ , so that we can enforce  $\text{AMO}_{\text{PE}}(\{x_{i,t',m} : i \in D_t\})$ , for each  $t' \in [1, T], t \in [1, T - t']$  and  $m \in [1, M]$ , which uses

$$\sum_{t'=1}^T \sum_{t \in [1, T-t']} \mathcal{O}(|D_t|) \cdot M \leq MT \cdot \mathcal{O} \left( \sum_{t \in [1, T-t']} |D_t| \right) = \mathcal{O}(NMT)$$

many clauses. Then, for each index  $i$ , we enforce that task  $i$  is done at some point, in some machine, with  $\mathcal{O}(N)$  clauses:

$$\bigvee_{\substack{m \in [1, M] \\ t \in [r_i, e_i - d_i]}} x_{i,t,m}, \quad \forall i \in [1, N],$$

We then use, independently for each  $m$ , the encoding of Theorem 22 to encode that the variables  $y_{m,t_1,t_2}$  assigned to true make for disjoint intervals. This results in  $\mathcal{O}(NMT + MT^2 \lg T)$  clauses. Correctness follows from the facts (i) we explicitly enforce that each task is done at some point, in some machine, and respecting its time constraints, (ii) the  $\text{AMO}_{\text{PE}}$  constraints ensure that no two tasks are assigned to the same machine during the same time interval, and (iii) the *disjoint intervals* encoding on the  $y_{m,t_1,t_2}$  variables ensures that no machine is used for two overlapping time intervals. that by the  $\text{AMO}_{\text{PE}}$  constraints we are forbidding different  $\blacktriangleleft$

When  $N$  and  $M$  are  $\Theta(T)$ , Theorem 23 results in  $\mathcal{O}(T^3 \lg T)$  clauses, as opposed to the  $\Omega(T^4)$  clauses of Mayak and Mondal [25]. It is worth noting that, as done for example by Marić [24], an alternative option would be to add for each time  $t$ , and machine  $m$ , a constraint

$$\text{AMO}_{\text{PE}}(\{x_{i,t',m} : i \in [N], t' \in [t - d_i, t]\}),$$

however  $\{x_{i,t',m} : i \in [N], t' \in [t - d_i, t]\}$  is a set of size  $\Omega(NT)$ , and thus this would result in  $\Omega(NMT^2)$  clauses.

## 5 Discussion and Future Work

We have proposed a theoretical framework that will hopefully be helpful in the quest for understanding the limits of CNF encodings. To do so, we considered “covering encodings”, based on covering the graph of incompatible pairs of literals with either cliques or bicliques, leading to different encodings. We showed how both clique coverings and biclique coverings have different advantages, where clique coverings are more efficient in the complete graph but biclique coverings are more efficient in the worst-case (Theorem 10). This difference is essentially surmounted for random Erdős-Renyi graphs (Proposition 6). Moreover, it is worth noting that clique coverings are also very efficient when the graph is “very close to being complete”, meaning that every vertex has degree at least  $n - \Theta(1)$ , as in this case a nice result of Alon [1] gives a covering with  $\mathcal{O}(\lg n)$  cliques and thus an encoding with  $\mathcal{O}(n \lg n)$  clauses. We have shown a modest lower bound in Proposition 17, which only applies to encodings using constant-width clauses. Extending this to general encodings is an interesting direction for future work. Even though our study here has been theoretical in nature, I have implemented clique-covering and biclique-covering algorithms. In particular, I tested the algorithm of Mubayi and Turán [26] for biclique coverings with the guarantee of Theorem 9, however, the algorithm was designed to prove its asymptotic quality and gives poor results for small  $n$ . Therefore, a natural next step is to design a more practically efficient algorithm to find good biclique coverings. For clique coverings, I tested both using the greedy approach of selecting the maximum clique at a time (for which I used the **Cliquer** tool [28]), and the specific clique-covering tool from Conte, Grossi and Marino [9]. The resulting quality of the coverings was not too different, but the latter algorithm was orders of magnitude faster for formulas with  $\approx 1000$  variables. A more thorough experimental evaluation is in order, especially considering the several more modern maximum clique algorithms, with their



different trade-offs [35]. In terms of related work to our covering ideas, besides its nearest neighbor being the works of Rintanen [30] and Ignatiev, Morgado, and Marques-Silva [19], we highlight that Jukna provides an in-depth treatment of the relationship between biclique coverings and the complexity of formulas representing graphs (especially bipartite graphs) [20]. Rintanen seems to be the earliest occurrence of the clique/biclique covering idea, although it is worth noting that his work explicitly states that his biclique representation still yields  $\Omega(n^2)$  clauses for graphs of  $n$  vertices [30, Section 6].

We have also studied how our framework applies to the case of complete interval graphs which are useful for encoding planning and scheduling problems. In fact, the *primitive* of encoding disjoint intervals seems to be useful in diverse contexts: a prior version of our encoding, that uses  $\mathcal{O}(n^{8/3}) = \mathcal{O}(n^{2.666\dots})$  clauses, was successfully used to improve an  $\mathcal{O}(n^4)$  encoding for *Straight Line Programs* (SLPs) from Bannai et al. [3], where a slight variant of the disjoint-intervals property was the bottleneck, and the improved encoding led to a total of  $\mathcal{O}(n^3)$  clauses, making a different constraint the bottleneck. The improved SLP encoding incorporating our ideas is currently under review [4]. To obtain  $\mathcal{O}(n^{8/3})$  clauses, the encoding is essentially the same as the one in Theorem 22, but except of proceeding recursively for the  $x$ -edges and the  $y$ -edges, we encode those directly. This leads to  $\mathcal{O}(k^2b^4)$  for the  $x$ -edges, since for any pair of blocks  $\ell, r$  (of which there are  $k^2$ ), we forbid the  $x$ -edges between intervals  $[i, j]$  and  $[i', j']$  such that  $B(i) = B(i') = \ell$  and  $B(j) = B(j') = r$ , and there are at most  $b^4$  choices for  $i, j, i', j'$  conditioned on their blocks being  $\ell$  and  $r$ . For the  $y$ -edges, this leads to  $\mathcal{O}(k^4)$  clauses, since we need to forbid the  $y$ -edges between pairs of blocks that guarantee an interval overlap. Therefore, the total number of clauses is  $\mathcal{O}(k^2b^4 + k^4 + n^2 + k^3b^2)$ , for which a simple calculus argument shows that the optimal choice is  $k = n^{2/3}$  and  $b = n^{1/3}$ , leading to the number of clauses being

$$\mathcal{O}(n^{4/3}n^{4/3} + n^{8/3} + n^2 + n^{6/3}n^{2/3}) = \mathcal{O}(n^{8/3} + n^{8/3} + n^2 + n^{8/3}) = \mathcal{O}(n^{8/3}).$$

The difference between the  $\mathcal{O}(n^{8/3})$  encoding and the result in Theorem 22 is the use of recursion. Indeed, the  $\mathcal{O}(n^{8/3})$  encoding can be formulated as a carefully crafted biclique covering, making for another example of the difference between BVA-style encodings, which allow for covering auxiliary variables, and encodings that only cover the base variables. Running BVA on top of the  $\mathcal{O}(n^{8/3})$  encoding led to smaller encodings that either of them alone, reinforcing the idea that BVA can operate recursively on top of an initial covering (see Table 1 in Appendix C). Experimental results for the SLP problem of Bannai et al. [3] are presented in Table 2 (Appendix C).

The recursion of blocks in Theorem 22 has a nice interpretation for scheduling: if one were to schedule events on e.g., a calendar year, starting on day  $d_1$  and ending on day  $d_2$ , it would be convenient to first catalogue the events based on their starting and ending month: *anything starting in January and ending in May is incompatible with anything starting in March and ending in September*. Then, for more granularity, the same technique decomposes months into weeks, and so on. The concrete decomposition in Theorem 22 used  $\lg n$  blocks in the first recursive step, which would be  $\lg 365 \approx 8.5$  blocks for a calendar year, on the order of magnitude of months. On the other hand, the decomposition for the  $\mathcal{O}(n^{8/3})$  version, where only one level of recursion is used, takes  $k = n^{2/3}$ , which would be roughly  $365^{2/3} \approx 51$  for a calendar year, so basically the number of weeks. Studying the practical applicability of our encoding for scheduling problems is an interesting direction for future work.

Finally, we note that our ideas can be readily applied to *Integer Linear Programming* formulations, and hopefully to other forms of constraint programming too.

## References

- 1 Noga Alon. Covering graphs by the minimum number of equivalence relations. *Combinatorica*, 6(3):201–206, September 1986. doi:10.1007/BF02579381.
- 2 Roberto Asín, Robert Nieuwenhuis, Albert Oliveras, and Enric Rodríguez-Carbonell. Cardinality Networks: A theoretical and empirical study. *Constraints*, 16(2):195–221, April 2011. doi:10.1007/s10601-010-9105-0.
- 3 Hideo Bannai, Keisuke Goto, Masakazu Ishihata, Shunsuke Kanda, Dominik Köppl, and Takaaki Nishimoto. Computing NP-Hard Repetitiveness Measures via MAX-SAT. In Shiri Chechik, Gonzalo Navarro, Eva Rotenberg, and Grzegorz Herman, editors, *30th Annual European Symposium on Algorithms (ESA 2022)*, volume 244 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 12:1–12:16, Dagstuhl, Germany, 2022. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. URL: <https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ESA.2022.12>, doi:10.4230/LIPIcs.ESA.2022.12.
- 4 Hideo Bannai, Keisuke Goto, Masakazu Ishihata, Shunsuke Kanda, Dominik Köppl, Takaaki Nishimoto, and Bernardo Subercaseaux. Computing NP-Hard Repetitiveness Measures via MAX-SAT (under review), 2025.
- 5 Armin Biere. exactly one clauses · Issue #39 · arminbiere/kissat — github.com. <https://github.com/arminbiere/kissat/issues/39#issuecomment-1686043817>. [Accessed 06-06-2025].
- 6 Magnus Björk. Successful SAT encoding techniques. *J. Satisf. Boolean Model. Comput.*, 7(4):189–201, 2011. URL: <https://doi.org/10.3233/sat190085>, doi:10.3233/SAT190085.
- 7 Jingchao Chen. A new sat encoding of the at-most-one constraint. *Proc. of the Tenth Int. Workshop of Constraint Modelling and Reformulation.*, page 8, 2010.
- 8 F. R. K. Chung, P. Erdős, and J. Spencer. On the decomposition of graphs into complete bipartite subgraphs. In Paul Erdős, László Alpár, Gábor Halász, and András Sárközy, editors, *Studies in Pure Mathematics: To the Memory of Paul Turán*, pages 95–101. Birkhäuser, Basel, 1983. doi:10.1007/978-3-0348-5438-2\_10.
- 9 Alessio Conte, Roberto Grossi, and Andrea Marino. Clique covering of large real-world networks. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing, SAC '16*, page 1134–1139, New York, NY, USA, 2016. Association for Computing Machinery. doi:10.1145/2851613.2851816.
- 10 Gregory Emdin, Alexander S. Kulikov, Ivan Mihajlin, and Nikita Slezkin. CNF Encodings of Parity. In Stefan Szeider, Robert Ganian, and Alexandra Silva, editors, *47th International Symposium on Mathematical Foundations of Computer Science (MFCS 2022)*, volume 241 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 47:1–47:12, Dagstuhl, Germany, 2022. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. doi:10.4230/LIPIcs.MFCS.2022.47.
- 11 Peter C. Fishburn and Peter L. Hammer. Bipartite dimensions and bipartite degrees of graphs. *Discrete Mathematics*, 160(1):127–148, November 1996. doi:10.1016/0012-365X(95)00154-0.
- 12 Alan Frieze and Bruce Reed. Covering the edges of a random graph by cliques. *Combinatorica*, 15(4):489–497, December 1995. doi:10.1007/BF01192522.
- 13 Alan Frieze and Bruce Reed. Covering the edges of a random graph by cliques, 2011. URL: <https://arxiv.org/abs/1103.4870>, arXiv:1103.4870.
- 14 He Guo, Kalen Patton, and Lutz Warnke. Prague Dimension of Random Graphs. *Combinatorica*, 43(5):853–884, October 2023. doi:10.1007/s00493-023-00016-9.
- 15 Andrew Haberlandt, Harrison Green, and Marijn J. H. Heule. Effective Auxiliary Variables via Structured Reencoding. In Meena Mahajan and Friedrich Slivovsky, editors, *26th International Conference on Theory and Applications of Satisfiability Testing (SAT 2023)*, volume 271 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 11:1–11:19, Dagstuhl, Germany, 2023. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. doi:10.4230/LIPIcs.SAT.2023.11.

- 16 Marijn J. H. Heule and Manfred Scheucher. Happy Ending: An Empty Hexagon in Every Set of 30 Points. In Bernd Finkbeiner and Laura Kovács, editors, *Tools and Algorithms for the Construction and Analysis of Systems*, pages 61–80, Cham, 2024. Springer Nature Switzerland. doi:10.1007/978-3-031-57246-3\_5.
- 17 Marijn J. H. Heule and Stefan Szeider. A SAT Approach to Clique-Width. *ACM Trans. Comput. Logic*, 16(3):24:1–24:27, June 2015. doi:10.1145/2736696.
- 18 Daniel Donnelly (<https://cs.stackexchange.com/users/12373/daniel-donnelly>). Is the smallest grammar problem over the singleton alphabet known to be np-complete or ...? Computer Science Stack Exchange. URL:<https://cs.stackexchange.com/q/171713> (version: 2025-04-12). URL: <https://cs.stackexchange.com/q/171713>, arXiv:<https://cs.stackexchange.com/q/171713>.
- 19 Alexey Ignatiev, Antonio Morgado, and Joao Marques-Silva. Cardinality encodings for graph optimization problems. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI'17*, pages 652–658, Melbourne, Australia, August 2017. AAAI Press.
- 20 Stasys Jukna. Computational Complexity of Graphs. In *Advances in Network Complexity*, chapter 5, pages 99–153. John Wiley & Sons, Ltd, 2013. doi:10.1002/9783527670468.ch05.
- 21 Gyula Katona and Endre Szemerédi. On a problem of graph theory. *Studia Scientiarum Mathematicarum Hungarica*, 2:23–28, 1967.
- 22 Petr Kučera, Petr Savický, and Vojtěch Vorel. A lower bound on CNF encodings of the at-most-one constraint. *Theoretical Computer Science*, 762:51–73, 2019. doi:10.1016/j.tcs.2018.09.003.
- 23 Norbert Manthey, Marijn J. H. Heule, and Armin Biere. Automated reencoding of boolean formulas. In *Haifa Verification Conference*, pages 102–117. Springer, 2012.
- 24 Filip Marić. Timetabling based on sat encoding: a case study. <https://poincare.matf.bg.ac.rs/~filip/phd/sat-timetable.pdf>, 2008.
- 25 Jaishree Mayank and Arijit Mondal. Efficient SAT encoding scheme for schedulability analysis of non-preemptive tasks on multiple computational resources. *Journal of Systems Architecture*, 110:101818, November 2020. doi:10.1016/j.sysarc.2020.101818.
- 26 Dhruv Mubayi and György Turán. Finding bipartite subgraphs efficiently. *Information Processing Letters*, 110(5):174–177, February 2010. doi:10.1016/j.ipl.2009.11.015.
- 27 Van-Hau Nguyen, Van-Quyet Nguyen, Kyungbaek Kim, and Pedro Barahona. Empirical study on sat-encodings of the at-most-one constraint. In *The 9th International Conference on Smart Media and Applications, SMA 2020*, page 470–475, New York, NY, USA, 2021. Association for Computing Machinery. doi:10.1145/3426020.3426170.
- 28 Sampo Niskanen and Patric Östergård. Cliquer homepage — users.aalto.fi. <https://users.aalto.fi/~pat/cliquer.html>, 2002. [Accessed 09-06-2025].
- 29 Steven Prestwich. *CNF Encodings*, chapter 2. IOS Press, February 2021. URL: <http://dx.doi.org/10.3233/FAIA200985>, doi:10.3233/faia200985.
- 30 Jussi Rintanen. *Compact representation of sets of binary constraints*, pages 143–147. Frontiers in Artificial Intelligence and Applications. IOS Press BV, Netherlands, 2006.
- 31 Andre Schidler and Stefan Szeider. SAT-based Decision Tree Learning for Large Data Sets. *Journal of Artificial Intelligence Research*, 80:875–918, July 2024. doi:10.1613/jair.1.15956.
- 32 Carsten Sinz. Towards an Optimal CNF Encoding of Boolean Cardinality Constraints. In Peter van Beek, editor, *Principles and Practice of Constraint Programming - CP 2005*, pages 827–831, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.
- 33 Bernardo Subercaseaux and Marijn J. H. Heule. The Packing Chromatic Number of the Infinite Square Grid is 15. In Sriram Sankaranarayanan and Natasha Sharygina, editors, *Tools and Algorithms for the Construction and Analysis of Systems - 29th International Conference, TACAS 2023, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2022, Paris, France, April 22-27, 2023, Proceedings, Part I*, volume 13993 of *Lecture Notes in Computer Science*, pages 389–406. Springer, 2023. doi:10.1007/978-3-031-30823-9\_20.

- 34 William J. Wesley. Lower Bounds for Book Ramsey Numbers, October 2024. [arXiv:2410.03625](#), [doi:10.48550/arXiv.2410.03625](#).
- 35 Qinghua Wu and Jin-Kao Hao. A review on algorithms for maximum clique problems. *European Journal of Operational Research*, 242(3):693–709, May 2015. [doi:10.1016/j.ejor.2014.09.064](#).
- 36 Neng-Fa Zhou. Yet another comparison of sat encodings for the at-most-k constraint, 2020. URL: <https://arxiv.org/abs/2005.06274>, [arXiv:2005.06274](#).

## A

 Proof of Proposition 20

► **Proposition 20.** *Let  $\tau : \text{var}(NIP_n) \rightarrow \{\perp, \top\}$  and assignment. Then, we have that*

$$\tau \models NIP_n \iff SAT(IPT_n|_\tau),$$

*and moreover, any satisfying assignment  $\theta$  for  $IPT_n|_\tau$  must assign  $\theta(z_{a,b}) = \top$  if and only if there is some  $[i, j]$  such that  $\tau(x_{i,j}) = \top$  and  $[a, b] \subseteq [i, j]$*

**Proof.** For the ( $\implies$ ) direction, we assume that  $\tau \models NIP_n$ , and we build the assignment  $\theta : \text{var}(IPT_n|_\tau) \rightarrow \{\perp, \top\}$  as described in the statement of the proposition:  $\theta(z_{a,b}) = \top$  if and only if there is some  $[i, j]$  such that  $\tau(x_{i,j}) = \top$  and  $[a, b] \subseteq [i, j]$ . Then the clauses (1-4) can be easily checked to be satisfied by  $\theta$ . For the ( $\impliedby$ ) direction, we assume a satisfying assignment  $\theta$  for  $IPT_n|_\tau$  and assume expecting a contradiction that  $\tau \not\models NIP_n$ . Since  $IPT_n|_\tau$  is satisfiable, and  $IPT_n$  contained the clauses  $\bar{t}_\ell \vee \bigvee_{[i,j] \supseteq \{\ell\}} x_{i,j}$  whose variables were assigned by  $\tau$ , the only possibility is that  $\tau \not\models (\bar{x}_{i,j} \vee t_\ell)$  for some  $1 \leq i < j \leq n$  and  $\ell \in [i, j]$ . Thus, we have that  $\tau(x_{i,j}) = \top$  and  $\tau(t_\ell) = \perp$ . But by the clauses of type (1),  $\tau(x_{i,j}) = \top$  implies  $\theta(z_{i,j}) = \top$ , so it suffices to prove that  $\theta(z_{i,j}) = \top$  contradicts  $\tau(t_\ell) = \perp$ . We do this by showing that  $\theta(z_{i,j}) = \top$  implies  $\tau(t_\ell) = \top$ , for any  $\ell \in [i, j]$ . The proof is by induction over  $d := j - i$ . If  $d = 1$ , then the clauses of type (2) directly give us  $\tau(t_\ell) = \top$ , and if  $d > 1$ , the clauses of type (3) give us that  $\theta(z_{i+1,j}) = \top$  and  $\theta(z_{i,j-1}) = \top$ , and as  $\ell$  belongs to either  $[i+1, j]$  or  $[i, j-1]$ , we conclude by the inductive hypothesis.

Let us now show that no other  $\theta$  works. Indeed, we first prove that  $\tau(x_{i,j}) = \top$  implies that  $\theta(z_{a,b}) = \top$  for every  $[a, b] \subseteq [i, j]$ , by induction on  $d := (a - i) + (j - b)$ . If  $d = 0$ , then  $a = i$  and  $j = b$ , so by the clause of type (1) we are done. Otherwise, we need to prove that the statement for  $d$  implies the case for  $d + 1$  for any  $d < j - i - 1$  (since  $j - i - 1$  is the maximum possible value of  $d$ ). We assume  $\tau(x_{i,j}) = \top$  and note that by the inductive hypothesis this implies  $\theta(z_{a,b}) = \top$  for any  $1 \leq a < b \leq n$  such that  $(a - i) + (j - b) = d$ . Then, any interval  $[a', b'] \subseteq [i, j]$  with  $(a' - i) + (j - b') = d + 1$  must be of the form  $[a + 1, b]$  or  $[a, b - 1]$  with  $(a - i) + (j - b) = d$ , and as by the clauses of type (3) we have  $\theta(z_{a+1,b}) = \top$  and  $\theta(z_{a,b-1}) = \top$ , we are done. On the other hand, assume that  $\theta(z_{a,b}) = \top$  for some pair  $1 \leq a < b \leq n$ , and let us show that  $\tau(x_{i,j}) = \top$  for some  $[i, j] \supseteq [a, b]$ . This time the induction is over  $d := n - (b - a)$ , with the base case  $d = 1$  implying that  $a = 1, b = n$ , from where (4) directly yields  $z_{1,n} \rightarrow x_{1,n}$  which proves the base case. For the inductive case, note that a clause of type (4) guarantees that either  $\tau(x_{a,b}) = \top$ , in which case we are done, or that either  $\theta(z_{a-1,b}) = \top$  or  $\theta(z_{a,b+1}) = \top$ . But  $n - (b - (a - 1)) = n - (b + 1 - a) = d - 1$ , and thus by inductive hypothesis we have that  $\tau(x_{i,j}) = \top$  for some  $[i, j]$  such that  $[i, j] \supseteq [a - 1, b] \supset [a, b]$  or  $[i, j] \supseteq [a, b + 1] \supset [a, b]$ . ◀

## B

 Proof of Lemma 21

► **Lemma 21.** *Each edge  $e := \{[i_1, j_1], [i_2, j_2]\} \in E(\mathcal{I}_n)$  with  $i_1 \leq i_2$  must be part of exactly one of the following cases:*

1.  $B(i_1) = B(i_2)$  and  $B(j_1) = B(j_2)$ , and  $i_2 \leq j_1$ , in which case we say  $e$  is an  $x$ -edge.
2.  $B(i_1) < B(i_2) < B(j_1)$ , in which case we say  $e$  is a  $y$ -edge.
3.  $B(i_1) = B(i_2)$  and  $i_2 \leq j_1$ , but  $B(j_1) \neq B(j_2)$ , in which case we say  $e$  is an  $s$ -edge.
4.  $B(i_1) < B(i_2) = B(j_1) = B(j_2)$  and  $i_2 \leq j_1$ , in which case we say  $e$  is an  $f$ -edge.
5.  $B(i_1) < B(i_2) = B(j_1) \neq B(j_2)$ , and  $i_2 \leq j_1$ , in which case we say  $e$  is an  $m$ -edge.

Moreover, any tuple  $(i_1, j_1, i_2, j_2)$  with  $i_1 \leq i_2$  that satisfies one of these cases implies  $\{[i_1, j_1], [i_2, j_2]\} \in E(\mathcal{I}_n)$ .

**Proof.** First, observe that in all cases we are assuming  $i_1 \leq i_2$ , and thus as all cases except (2) explicitly require  $i_2 \leq j_1$ , they all imply the intersection of the intervals  $[i_1, j_1]$  and  $[i_2, j_2]$  is non-empty. For case (2), note that  $B(i_2) < B(j_1)$  implies  $i_2 < j_1$ , which together with  $i_1 \leq i_2$  implies again that the intersection of the intervals  $[i_1, j_1]$  and  $[i_2, j_2]$  is non-empty. Thus, all cases correspond to actual edges in the graph  $\mathcal{I}_n$ . To see exhaustiveness, we present in Figure 4 a decision tree that cases on the relationships between the blocks. ◀

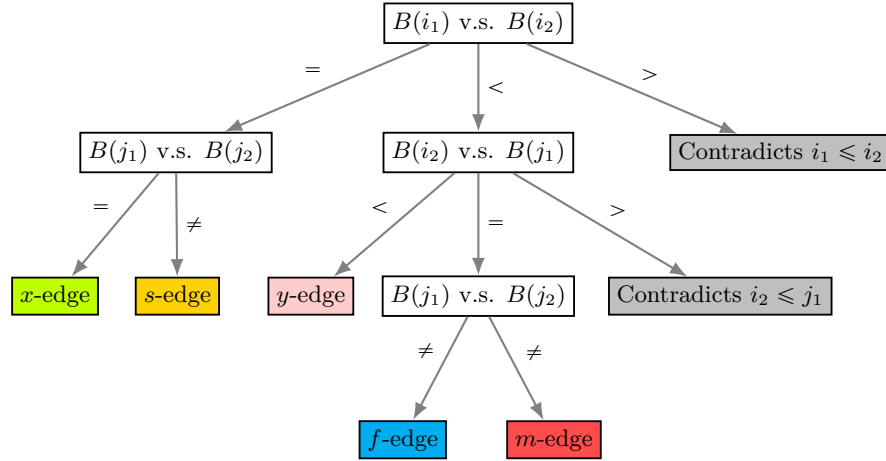
Furthermore, as a sanity check, we present in Code 1 a Python script that validates the correctness for some finite values.

**Code 1** Code to validate Lemma 21.

```

1 def B(i, b):
2     # ceil(i/b) can be written as (i + b - 1)//b
3     return (i + b - 1) // b
4
5 def classify(i1, j1, i2, j2, b):
6     """Return which case (1,2,3,4) the edge {[i1,j1],[i2,j2]} belongs to."""
7     bi1, bi2 = B(i1,b), B(i2,b)
8     bj1, bj2 = B(j1,b), B(j2,b)
9
10    # the five predicates from the lemma
11    is_x = (bi1 == bi2) and (bj1 == bj2)
12    is_y = (bi1 < bi2) and (bi2 < bj1)
13    is_s = (bi1 == bi2) and (bj1 != bj2)
14    is_f = (bi1 < bi2) and (bj1 == bj2) and (bi2 == bj1)
15    is_m = (bi1 < bi2) and (bi2 == bj1) and (bj1 != bj2)
16
17    flags = [is_x, is_y, is_s, is_f, is_m]
18    if sum(flags) < 1:
19        # either none or more than one case matched
20        return flags
21    return flags.index(True) + 1 # return 1,2,3,4, or 5.
22
23 def check_lemma(n, b):
24     bad = []
25     for i1 in range(1, n+1):
26         for i2 in range(i1, n+1):
27             for j1 in range(i1+1, n+1):
28                 for j2 in range(i2+1, n+1):
29                     # only look at intersecting intervals
30                     if i2 <= j1:
31                         cls = classify(i1, j1, i2, j2, b)
32                         if cls is None or type(cls) is not int:
33                             bad.append((i1, j1, i2, j2))
34
35     if bad:
36         print("Found unclassified or multiply-classified edges:")
37         for quad in bad:
38             print(quad, "-> (blocks)", [B(q, b) for q in quad])
39     else:
40         print(f"All edges classified correctly! (n={n},b={b})")
41
42 if __name__ == "__main__":
43     # Example for n = 50, block size of 10.
44     check_lemma(n=50, b=10)
45
46     # Example without exact divisibility
47     check_lemma(n=50, b=8)

```



■ **Figure 4** A decision tree for the cases of Lemma 21.

## C Application to String Compression with SLPs

Note that in the problem of Bannai et al. [3], the intervals can overlap if one is strictly contained in the other. This can be achieved by minor modifications in the indices of our constraints, without any new conceptual ideas.

We display some preliminary experimental results in Table 2, where we show the usage of the  $\mathcal{O}(n^{8/3})$  encoding for disjoint intervals in replacement of constraint 8 of the SLP encoding of Bannai et al. [3]. Our experiments are over families of string that make standard examples for string compression, and their descriptions can be found in the paper of Bannai et al. [3], and the concrete strings are publicly available in their repository: <https://github.com/kg86/satcomp>. Perhaps the most striking example is that of Fibonacci binary strings, defined recursively as  $F_0 = 0$ ,  $F_1 = 1$ , and  $F_n = F_{n-1}F_{n-2}$  (concatenation) for  $n \geq 2$ , where as shown in Table 2, the improved encoding reduces the number of clauses from 118 millions to 12 millions (`fib12.txt`). Generally speaking, the impact of the encoding for the disjoint (or strictly contained) intervals is not always as large as the asymptotics suggest, since the constraint in the encoding from [3, Equation (8)] are only applied based on a condition that depends on the repetitiveness of the input string. The impact of the encoding is thus maximal for strings of the form  $a^n$ , for some symbol  $a$ . These strings are interesting from a theoretical point of view, since the complexity of the smallest SLP (which is NP-hard to compute over arbitrary strings) is not known [18].

■ **Table 1** Comparison of encoding methods for the independent-set property of  $\mathcal{I}_n$ . We use the acronym ‘bd’ for the  $\mathcal{O}(n^{8/3})$  encoding, standing for “block decomposition”.

$n$	naïve		bva		bd		bd+bva	
	vars	clauses	vars	clauses	vars	clauses	vars	clauses
10	55	330	82	143	130	196	132	194
20	210	5,985	402	862	398	984	430	801
30	465	31,465	1,015	2,499	850	2,714	952	1,893
40	820	101,270	1,935	4,996	1,315	5,774	1,687	3,095



■ **Table 2** Comparison of encodings for smallest SLPs

File	T	Base Encoding			Base + $\mathcal{O}(n^{8/3})$ disjoint intervals		
		Time (s)	Clauses	Vars	Time (s)	Clauses	Vars
fib6.txt	21	<b>0.15</b>	3,754	494	0.25	3,808	719
fib7.txt	34	<b>0.93</b>	16,764	1,513	1.39	14,465	1,987
fib8.txt	55	<b>3.67</b>	84,554	4,716	3.88	52,746	5,716
fib9.txt	89	<b>5.14</b>	472,269	14,604	5.17	203,392	16,683
fib10.txt	144	6.97	2,849,938	44,848	<b>6.57</b>	798,982	49,430
fib11.txt	233	33.16	18,101,282	135,859	<b>20.53</b>	3,119,598	145,687
fib12.txt	377	214.92	118,647,206	406,448	<b>32.45</b>	12,363,254	428,235
thue_morse5.txt	32	<b>0.40</b>	9,562	948	0.70	11,199	1,398
thue_morse6.txt	64	<b>3.48</b>	73,015	4,339	4.83	71,349	5,563
thue_morse7.txt	128	<b>8.15</b>	656,099	19,933	8.18	532,394	23,740
thue_morse8.txt	256	12.60	6,977,803	90,647	<b>11.61</b>	3,825,669	102,260
period_doubling5.txt	32	<b>0.49</b>	11,682	1,135	0.79	11,962	1,585
period_doubling6.txt	64	<b>4.36</b>	109,928	5,586	4.80	75,780	6,810
period_doubling7.txt	128	7.60	1,281,958	27,219	<b>7.31</b>	557,642	31,026
period_doubling8.txt	256	28.23	17,372,984	130,050	<b>12.41</b>	3,951,407	141,663
paperfold4.txt	32	<b>0.23</b>	9,078	880	0.55	10,869	1,330
paperfold5.txt	64	<b>1.66</b>	67,650	3,924	3.71	69,736	5,148
paperfold6.txt	128	<b>6.14</b>	598,774	17,613	8.22	523,842	21,420
paperfold7.txt	256	14.17	6,280,334	78,705	<b>12.92</b>	3,782,468	90,318